JOÃO LUÍS AGUIAR MARTINS NETO

CAG repeat instability in Huntington's disease: insights from HD patients and mouse models.

Tese de Candidatura ao grau de Doutor em Biologia Básica e Aplicada submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto.

Orientador – Prof. Doutora Isabel da Conceição Moreira Pereira Alonso
Categoria – Professora Afiliada
Afiliação – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto (ICBAS).

Coorientadora – Vanessa C. Wheeler
Categoria – Principal Investigator, Associate Professor
Afiliação – Center for Genomic Medicine (CGM), Massachusetts General Hospital/Harvard Medical School (HMS).

Este trabalho foi financiado pela

Fundação para a Ciência e Tecnologia

SFRH / BD / 51705 / 2011

Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

Every great wizard in history has started out as nothing more than what we are now, students. If they can do it, why not us?

— J.K. Rowling

Aos meus pais

Acknowledgments

A doctoral thesis is far from being an individual effort toward an academic degree, thankfully I had many people supporting me throughout this journey.

First, I would like to thank my supervisors Professor Vanessa Wheeler and Professor Isabel Alonso, who guided me through this adventure, and were always available in times of need. It is also very inspiring to get a new exciting result and hear back “this is cool”! Thank you for all the patience and time invested in me.

I would like to thank all the present and past members of the Wheeler lab at the Center for Genomic Medicine, but in special to Ricardo Mouro-Pinto, who was always helped me in large and small tasks, answered whatever “silly” questions I had, and me feel less alone during late nights, as he appeared as a reflection in the window (private joke). A big thank you to Marina Kovalenko, who had infinite patience while teaching me whatever I asked, including the long time alongside me in the beginning of my PhD to guide me through brain tissue dissections which thanks to her I can now almost do blindfolded.

A big thank you to all the members of UnIGENe, Carolina, São, Sara, Diana, Mariana, Joana, Marlene e Miguel, who always kept in touch when I was far away, and have been of tremendous support these last few hard months. I couldn’t have made it without all of you!

Many other people in the present and past members of the (then CHGR) now CGM also helped me in numerous things, regarding different projects, I owe a huge thank you to all of them, Marisa, Marta, Randy, Lakshmi, Kawther, Jon-Min, Mike, Tammy, Jacob, Marcy, Jim. I’m sure I’m missing so many people… but count yourself included!

I also thank the Portuguese crew in Boston, for being my support network when most needed. In special Catarina, Teresa, Nelma and Mariana, who for some reason I cannot explain, have no problems with me bugging regarding whatever, whenever. Thank you for being great friends.

A *huge* thank you to Rita, Estrela and Diogo, for all the support, and for always making the effort of keeping in touch and not minding my complaining, even when at one point we were split among four countries in three time zones. True friendship is really forever.

A big thank you to Ana Rita and Diana, who shared a time zone with me for a while, and helped whenever I needed to vent.

A big thank you to Emanuela and Carlos the best roommates anyone could have, it's very important to know that when you get home on the "bad days" there are people that will listen and cheer you up.

A shout out to all the students of my GABBA edition (go 15th!) with whom I could also share all the ups and downs of this journey.

Finally, the most important, an infinite thank you to my parents, who are always loving and caring no matter what, and completely understood my decision of going to do research across the ocean, even though the distance hurt a bit. Thank you.

Table of Contents

Publications

This thesis includes the following publication as an integral part of its contents:

João Luís Neto, Jong-Min Lee, Ali Afridi, Tammy Gillis, Jolene R. Guide, Stephani Dempsey, Brenda Lager, Isabel Alonso, Vanessa C. Wheeler and Ricardo Mouro Pinto. *Genetics* February 1, 2017 vol. 205 no. 2 503-516;

Abbreviation list

CAG - cytosine-adenine-guanine

CGM - Center for Genomic Medicine

GWAS - genome-wide association study

HD - Huntington's disease

iPSCs - induced pluripotent stem cells

JAX - The Jackson Laboratory

LCLs - lymphoblastoid cell lines

MGH - Massachusetts General Hospital

PCR - polymerase chain reaction

polyQ - polyglutamine

RFLP - restriction fragment-length polymorphism

SNPs - single nucleotide polymorphisms

TNR - trinucleotide repeat

Abstract

Huntington's disease (HD) is a rare, neurodegenerative, progressive, autosomal dominant disorder caused by an expanded cytosine-adenine-guanine (CAG) repeat in the *HTT* gene located at the 4p16.3 *locus*. Repeat size is variable and known to influence several aspects of the disease such as penetrance and age-of-onset. The repeat is unstable across generations and in patient tissues. Large size changes toward expansions in intergenerational transmissions contribute to the anticipation observed in the disorder, sometimes leading to extremely early disease onset. Somatic changes in repeat size, also seem to contribute toward the modulation of age-of-onset. Understanding the factors that contribute to CAG repeat instability is of the utmost importance for the discovery and application of possible therapies, aimed either at preventing CAG repeat size expansions or inducing CAG contractions. There is currently no cure for the disease, so, targeting its gene is a good approach that hits the cause of the disease. Therefore, all the studies presented here have different specific goals, but share the overall objective of finding and understanding the factors that influence CAG repeat instability, using HD patient cells and mouse models for the disorder.

Instability was characterized in lymphoblastoid cell lines (LCLs), from individuals belonging to a nuclear family in which the father presented a high frequency of intergenerationally unstable transmissions in order to examine the behavior of the CAG repeat overtime in culture. These analyses did not highlight any particularly high instability in the LCLs of the father, suggesting that the high levels of instability that appear to be present in his germline may not be recapitulated in his somatic cells. These analyses also provided the opportunity to assess the feasibility of using LCLs as a cell-based model for conducting screens of instability altering factors. The results suggested that LCLs likely do not provide good models for this purpose due either to very modest repeat size changes in cells with lower repeat lengths, or highly heterogeneous and variable repeat size populations in cells with high repeat lengths, severely hindering interpretation.

Despite the apparent absence of an obvious connection between high intergenerational and somatic instability in the father of the nuclear family above,

through the evaluation of instability in LCLs and sperm samples across multiple individuals with expanded repeats, it is shown that germline and LCL instability, particularly expansions, are correlated. This indicates that contributors to somatic and germline instability may be shared.

Possible correlations between instability measures and the expression levels of a group of genes involved in DNA replication and repair processes in a different set of HD patient and control LCLs, were studied as a hypothesis generating tool. This resulted in the identification of several significant and/or interesting results, indicating that genes such as *NTHL1, POLD1, TP73, FAN1* and *LIG1* might have a role in somatic instability.

Intergenerational instability was studied using the largest breeding datasets of HD knock-in mouse models available to date: one contained data from several lines over a wide range of repeat sizes (Htt^{Q20}, Htt^{Q50}, Htt^{Q80}, Htt^{Q92}, Htt^{Q111}, Htt^{Q140}, and Htt^{Q175}) in the B6J background; the other contained data from a more limited set of alleles across six mouse strains (B6J, CD1, FVB, DBA, 129 and B6N). Several interesting conclusions came out of this study. We saw that the presence of a knock-in allele does not skew the proportion of heterozygous versus wild-type mice independently of repeat size. We observed that parental sex influences the frequency of repeat changes, but sex-of-offspring does not seem to have any effect on instability. CAG repeat length influenced the frequency and magnitude of expansions but only the magnitude of contractions. A minor role of paternal age on expansion magnitude was also found. Genetic background (*trans*) effects were identified as modifiers of frequency and magnitude of changes when comparing the different strains, and *Lig1* and *Spata31* may be involved in these differences. Consequences of *cis*-effectors and were also found, specifically through the potential expansion protection effects when comparing lines with and without a neomycin resistance cassette (*neo*) sequence upstream of the CAG repeat.

This same *neo* sequence also appeared to have a mild but significant effect in altering instability and diminish the somatic expansions, specifically in the liver of Htt^{Q175} mice.

Overall, this work yielded multiple possibilities to be explored, and a better understanding of factors that modulate CAG repeat size, adding to the state of the art regarding instability of HD's CAG repeat, bringing us closer to the possibility of targeting the repeat therapeutically and contributing to one day help patients and individuals at-risk for the disease.

Resumo

A doença de Huntington (HD), é uma doença rara, neurdegenerativa, progressiva, autossómica dominante causada por uma repetição de citosina-adenina-guanina (CAG) no gene *HTT* situado no *locus* 4p16.3. O tamanho desta repetição é variável e sabe-se que afeta vários aspetos da doença tal como a sua penetrância e a idade de início dos sintomas. A repetição é instável de geração para geração e em entre diferentes tecidos. Grandes expansões em transmissões intergeracionais contribuem para a anticipação observada na doença, conduzindo por vezes a uma idade de inicio extremamente precoce. Alterações somáticas no tamanho da repetição também contribuem para a modulação da idade de inicio. Perceber que fatores contribuem para a instabilidade da repetição é da maior importância para a descoberta e aplicação de terapias com o objetivo de prevenir expansões ou induzir contrações da sequência repetitiva. Atualmente não existe uma cura para a HD, portanto, procurar modular diretamente a causa genética da doença será uma abordagem apropriada. Todos os projetos aqui apresentados partilham um objetivo global, o de encontrar e compreender fatores que influenciam a instabilidade da repetição, nomeadamente com a utilização de células de doentes e modelos de murganho de HD.

A instabilidade da repetição foi caracterizada em células linfoblastóides (LCLs) de indivíduos de uma família com HD em que o pai afetado apresenta uma alta frequência de transmissões instáveis intergeracionalmente, para perceber de que forma se comporta o tamanho da repetição nestas células ao longo do tempo em cultura. Este estudo não demonstrou uma particular instabilidade nas células somáticas do pai afectado, indicando que a instabilidade da sua linha germinativa não parece estar presente nas suas células somáticas. Neste projeto também se avaliou a potencial viabilidade de utilizar as LCLs como modelo para a procura de factores que modificam a instabilidade. Os resultados obtidos indicam-nos que este não será um modelo apropriado, uma vez que as linhas com repetições mais curtas apresentam alterações muito modestas no tamanho do alelo principal, enquanto que LCLs com repetições muito longas apresentam populações heterogéneas e de comportamento complexo, impedindo qualquer possível interpretação.

Apesar de não se ter observado uma correlação entre a instabilidade somática e instabilidade intergeracional no indivíduo acima referido, quando é avaliada a instabilidade em LCLs e amostras de esperma de um conjunto de doentes, demonstra-se que efectivamente existirá alguma relação entre a instabilidade somática e da linha germinativa, particularmente no que toca a expansões. Isto indica que ambas partilharão alguns elementos que contribuem para a instabilidade da repetição.

Possíveis correlações entre instabilidade e a expressão de genes envolvidos em processos de replicação e reparação de DNA foram avaliadas, num conjunto de indivíduos, com o objetivo de gerar novas hipóteses relativamente a modificadores de instabilidade da repetição. Este estudo resultou na identificação de vários genes, nomeadamente *NTHL1*, *POLD1*, *TP73*, *FAN1* e *LIG1*, como potenciais moduladores da instabilidade somática.

A instabilidade intergeracional também foi estudada utilizando os maiores conjuntos de dados disponíveis até à data relativamente à manutenção de modelos de murganho de HD. Um destes conjuntos de dados incluia várias linhas (Htt^{Q20}, Htt^{Q50}, Htt^{Q80}, Htt^{Q92}, Htt^{Q111}, Htt^{Q140}, Htt^{Q175}) contendo um vasto intervalo de tamanhos de repetição e o outro contendo um intervalo de repetições mais curto, mas pertencendo a estirpes geneticamente distintas (129, CD1, FVB, DBA, 129 e B6N). Várias conclusões interessantes foram retiradas deste estudo. Observamos que a presença de um alelo *knock-in* não influencia a proporção de ratinhos heterozigóticos e *wild-type*, independentemente do tamanho da repetição. Concluímos que o sexo do progenitor transmissor da repetição influência a instabilidade, mas o sexo da cria não. O tamanho da repetição transmitida afeta a frequência e magnitude de expansões, mas apenas a magnitude no caso das contrações. Também se observou que a idade paternal terá algum papel na magnitude das expansões. Efeitos relativos ao *background* genético (efeitos em *trans*) também foram identificados, através da avaliação de alterações na instabilidade nas diferentes estirpes dos modelos de HD, sendo que genes como o *Lig1* e o *Spata31*, poderão estar envolvidos neste processo. Também efeitos em cis foram identificados, nomeadamente a presença de uma cassete de resistência á neomicina (*neo*) a montante da repetição.

A mesma sequência *neo*, parece também ter um efeito moderado, mas significativo na alteração da instabilidade e na diminuição do indice de expansão somática em modelos *Htt*^{Q175}.

Este trabalho gera uma multitude de possibilidades a explorar, e conduz a uma melhor compreensão dos fatores que modificam a instabilidade da repetição de CAGs, contribuindo para o estado-da-arte relativamente à instabilidade da repetição na HD, aproximando a probabilidade de utilização da modulação da repetição como alvo terapêutico, podendo no futuro ajudar doentes e indivíduos em risco para a doença.

1. General Introduction

In his paper *On Chorea*[1] published in 1872, Dr. George Huntington characterized a choreic disorder that, as he put it, "is peculiar in itself and seems to obey certain fixed laws", providing the first comprehensive description of the disorder – that he simply termed *hereditary chorea* – now known as Huntington's disease (HD).

While there are possible earlier descriptions of the disease, Huntington's paper concisely described most of the features present in the typical presentation of HD: its hereditary nature; the presence of cognitive and/or psychiatric symptoms ("tendency to insanity and suicide"); and its manifestation in adult life[1]. It does not seem that Huntington came across the juvenile form of the disorder, which constitutes only a small relative amount of HD patients[2–4] and presents clinical features distinct from the adult-onset disease, notably lacking chorea[2].

1.1. Epidemiology

Prevalence of HD has been estimated for many decades across the globe showing high heterogeneity among populations, with differences up to several fold. Lower prevalence is estimated in African and Asian populations (≤1 per 100,000), while Western populations show higher prevalence, namely, 2-5 per 100,000 in Portugal and ~7.5 per 100,000 in North America[5,6].

1.2. Huntington's disease clinical symptoms, disease progression and diagnosis

Most of HD's symptoms may be classified in three categories: progressive motor, cognitive and psychiatric features.

1.2.1. Motor symptoms

Numerous motor symptoms are observed in HD patients. As stated previously, chorea is a very prominent occurrence, specially early in the disease, and consists of short-lived, excessive involuntary movements present whenever the patients are in an awake state[2], other dyskinetic symptoms, for instance, intermittent and stereotyped movements ("tics"), such as head jerking, sniffing or blinking, are usually also present[2,3,7]. Later in the disease, the hyperkinesia tends to lessen, but dystonia (involuntary, uncontrolled muscle contractions and tone), akinesia (hindrance in commencing movement), bradykinesia (slowness of movement) and rigidity appear more prevalently[2,3,7]. Juvenile-onset HD also frequently presents myoclonic seizures[2,4] along with dystonia, bradykinesia, and spasticity[4]. External phenomena and psychiatric aspects (*e.g.* infections, anxiety, stress) may lead to transitory worsening of these symptoms[2].

1.2.2. Cognitive symptoms

High-level processing and executive functioning are altered in HD, causing slowing of thought processing, impairment of organizational skills, and planning, problems in initiating actions, concentration and multitasking[3,7]. Short-term memory and visuospatial and perceptual skills are also affected[7]. Some of these features occur early in the disease process and together, all of these deficits might severely hamper every-day life, independently of difficulties caused by the motor symptoms[2,3,7]. The juvenile form of the disorder also presents additional symptoms, namely in the form of early learning difficulties or disability[7].

1.2.3. Psychiatric symptoms

Like cognitive features, psychiatric symptoms usually precede motor phenotypes and are an early occurrence in HD. Some of these symptoms might be hard to, at first, link with the disorder as they might be attributed to other underlying causes[2]. Depression and anxiety are the most common mental disturbances in HD. Suicidal ideation is also fairly prevalent, a fact that George Huntington also noticed in his early description of the disorder[1]. Apathy, irritability and aggression are also present. Some of these psychiatric features often are dealt with symptomatically, as overall treatment for HD is not available as of yet[3].

1.2.4. Additional HD symptoms

The sleep-wake phases are disrupted causing sleep disturbance which might lead to somnolence during daytime[3]. Metabolic problems are also present, leading to a catabolic state which causes severe weight loss[2,3,7]. Skeletal muscle atrophy is also observed[2]. Other conditions are also present in advanced stages of the disorder, namely, muteness which usually hampers the patients' ability to communicate[7], and dysphagia, which hinders the ability to provide proper nutrition to patients, who need a high caloric intake due to the catabolic phenotype[2,3,7]. Cardiac failure is also common in HD affected individuals and is the leading cause of death in roughly one third of patients[8].

1.3. Pathology

1.3.1. Neuropathology

HD's motor, cognitive and psychiatric symptoms have been related to its neuropathological changes[9], such as neurodegeneration and neuronal loss[9–12]. Brain structural irregularities precede HD clinical symptoms, and affect to some extent the whole organ[12], and might end in a ~20% loss of brain volume[11] and up to 20-30% of total brain weight[9,10,12]. Nevertheless, particular areas of the brain are more altered then others. The basal ganglia is preferentially affected; more specifically, the striatum (caudate nucleus and putamen) shows a large loss of neuron numbers as well as astrogliosis, which worsen with disease progression resulting therefore in a gross atrophy and high reduction in total volume (~60%) of this structure in late stages of the disorder[11,12]. Within the striatum medium spiny neurons are the most vulnerable to degeneration[10,11,13,14]. HD neuropathology is categorized according to macro- and microscopic changes in the striatum, within – the Vonsattel Grading System[15] – one of five grades (0-4) according to the degree of severity, from lowest to highest, that usually correlate with the degree of clinical disability[10–12].

While the striatal atrophy is by far considered the main neuropathological hallmark of HD, detailed descriptions of degeneration in other brain regions (thalamus, cerebral cortex, hippocampus, among many others) have been thoroughly characterized and assessed[9–12].

1.3.2. Peripheral pathology

Peripheral consequences are also present but might be less apparent in individuals with HD, among them: increased peripheral inflammation and dysfunction of blood-derived cells[8,16]; reduced number of germ cells are present in testis and male patients have reduced levels of testosterone[8]; decrease in insulin secretion and sensitivity[8]; and cardiac failure[8,16].

While peripheral tissues do not seem to be the main ones affected in HD they are more easily accessible and substantially contribute to our understanding of the disorder[8].

1.4. Diagnosis

Currently, formal diagnosis of the disorder is made clinically and requires the presence of extrapyramidal motor symptoms, with genetic testing providing a confirmation of the diagnosis[2,17].

1.5. Heredity and genetics

The heritable nature of HD has been one of its most noticeable characteristics, with both the adult- and juvenile-onset forms being characterized in multi-generational families[1,18]. George Huntington even described its “pattern of inheritance”, referring to the observation that if people affected by the disorder have offspring, one or more will be affected, while individuals in these families that go through life unaffected by the disease are sure to have only non-affected descendants[1], matching the well described autosomal dominant inheritance pattern of HD[19].

The search and discovery of the gene underlying HD also resulted, in part, from thoroughly acquired data from HD families – one very large Venezuelan family from the region of Lake Maracaibo, where the disease originated from a single founder, and a reasonably large American family from Ohio. Initial efforts, using restriction fragment-length polymorphism (RFLP) technology, mapped the gene to chromosome 4, an important breakthrough and the first fruitful linkage analyses using polymorphic markers in humans[20,21]. Furthermore, at the time, assuming HD’s genetic heterogeneity was non-existent, the marker shown to segregate with the disease could be tested in individuals at risk for the disorder[20], and was a first step of the utmost importance in the gene discovery[19].

With molecular technology advancements, higher densities of genetic markers were achieved, eventually leading to a candidate region for the gene responsible for HD, located at the 4p16.3 locus, with 2.2Mb and a ~500kb region showing the highest overall linkage disequilibrium[19,22,23]. Exon trapping was used in this smaller region, where a sequence initially dubbed “interesting transcript 15” (IT15) was found to have a polymorphic cytosine-adenine-guanine (CAG) repeat, which in the non-affected population mostly presented 11 to 24 repeats (with a very small percentage between 25 and 34 CAGs), while all HD chromosomes possessed over 42 repeats[19]. The genetic cause of HD was hence found, consisting of this expanded trinucleotide repeat (TNR) in the – historically termed *IT15,* eventually renamed *HD*, and now called – *HTT* gene.

The repeat is located in exon 1 of *HTT* and encodes a polyglutamine (polyQ) stretch in the N-terminal region of the huntingtin protein[13,18,19,24]. *HTT* is expressed ubiquitously, with highest expression in the brain[8,25–28], although expression levels in specific brain regions do not overlap with severity of neuropathological changes[29].

Huntingtin does not share an extensive homology with other proteins, and therefore at the time of the gene’s discovery its function was unknown[19]. Even though nowadays we know huntingtin is involved in a myriad of cellular processes and pathways (such as transcription, cell division, vesicular trafficking, and autophagy, among others)[24,28], there still is no clear idea of “wild-type” huntingtin’s cellular function[28].

Many possible pathogenic pathways have been described and studied thoroughly are thought to be primarily a consequence of a toxic gain-of-function in the mutant huntingtin[24], although there might be a smaller contribution from loss-of-function due to lower levels of non-expanded protein[28]. Furthermore, expanded polyQ leads to the formation of large intracellular inclusions and aggregates[10,12,24,28], and although these do not seem to correlate with cytotoxicity[30–32] it is unclear if oligomeric precursors, such as amyloid fibrils might play a pathogenic role associated with the toxic gain-of-function[24].

Nonetheless, the overall upstream culprit of the disorder is the expanded CAG repeat which modulates many aspects of HD, such as disease penetrance and age-at-onset.

1.6. CAG repeat length and disease penetrance

Shortly after the discovery of the causal mutation, different groups started quantifying repeat sizes in large HD cohorts[33–38], in order to “fine tune” the intervals that would result in developing the disorder or not.

Currently, *HTT* CAG repeat sizes are divided into four categories (Figure 1): 1) Normal alleles, which have TNRs with less than or equal to 26 repeats (the most frequent repeat sizes are between 17 and 19, but might go as low as 6 CAGs). These do not lead to disease, and are stable in over 99% of meiosis; 2) High normal alleles – commonly referred to as intermediate alleles – which possess between 27 and 35 repeats, are carried by ~2% of the population, and there is an overall consensus that they do not lead to disease. In this range, there are no reports of unstable maternal meiosis, although in paternal transmissions these alleles might expand to disease-causing alleles; 3) Reduced penetrance alleles, where 36 to 39 CAGs are present, have been associated with HD’s clinical and neuropathological symptoms, however, non-affected elderly individuals with the aforementioned repeat sizes have been reported, showing that disease penetrance is not complete in this range. These alleles are unstable meiotically; 4) Full penetrance alleles, which possess 40 or more repeats (that might be up to ~250 CAGs) and also are meiotically unstable[39,40].

Unaffected individuals		Affected individuals	
Normal	Intermediate	Reduced penetrance	Full Penetrance
≤26 CAGs	27-35 CAGs	36-39 CAGs	≥40 CAGs
Meiotically stable	Meiotically unstable paternally	Meiotically unstable	Meiotically unstable

Figure 1 – Classification of alleles depending on *HTT*’s CAG repeat size.

There have been reports of possible HD diagnosis in individuals with repeats in the higher-end of the unaffected range (*i.e.* large intermediate alleles)[41,42], but most of them do not show convincing evidence that the phenotype observed is consequence of *HTT*’s CAG repeat size[41]. There is also discussion if a single reported case for a specific allele size is enough to lead toward changes in classification[43]. Additionally, the existence of HD phenocopies is known, and while for several of them genetic causes have been described[44–48], a majority has no known causal gene[49] as of yet, possibly explaining some of these rare cases.

Nonetheless, standardized guidelines are of extreme importance, as they assure that individuals with comparable repeat sizes get homogenous genetic counseling and disease risk assessment[39,40,43].

1.7. CAG repeat length and HD age-of-onset

The mean age-of-onset for the disorder is 45 years of age, but onset might also occur very early or very late in life[24]. Concomitantly with the discovery of the gene, a potential inverse correlation of repeat size with HD age-of-onset was observed, with large repeats resulting in the juvenile form of the disorder (Figure 2)[19].

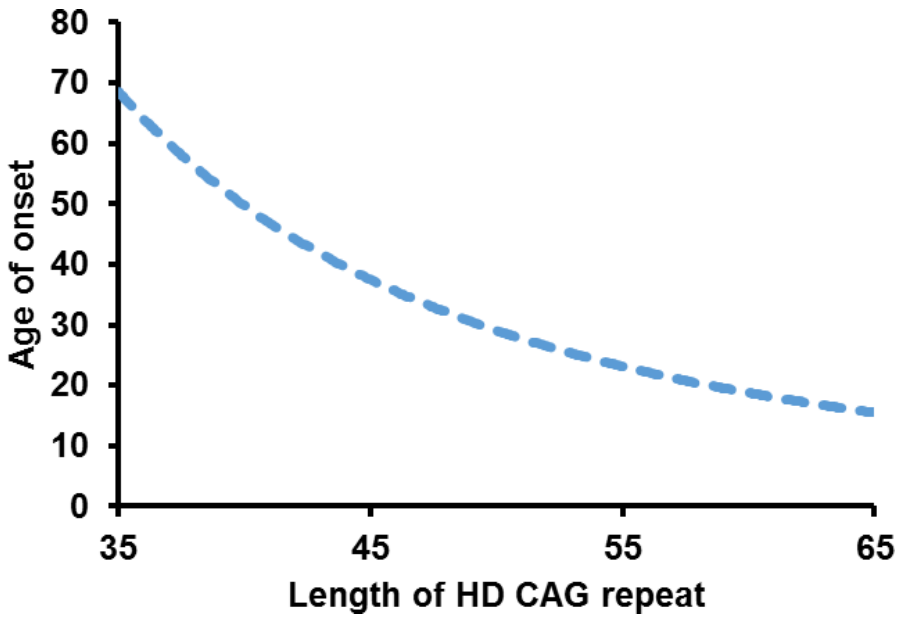


Figure 2 – Relationship between CAG repeat size and age-of-onset. Adapted from [14].

This correlation has been confirmed multiple times in several cohorts[14,24,33,35,50,51]. CAG repeat size accounts for ~56% of age-of-onset variation in HD[14,24,33–35,51,52]. A considerable part of the remaining variability is heritable, implicating a role for other genetic contributors, although environmental factors also play a part[16,18,51,53].

Early on, the search for genetic components contributing to the residual variation of age-of-onset relied on candidate regions or genes, including the size of the normal allele in heterozygous individuals, or genes related to neuronal pathways and neurodegeneration (such as *GRIK2* and *PPARGC1A*), but this approach has not provided results that convincingly withstand rigorous statistical scrutiny[24,51]. Recently, using an unbiased approach through a genome-wide association study (GWAS). Three main distinct *loci* were identified as modifiers of residual age of motor onset, contributing to hasten or delay the phenotype, implicating different clusters of pathways, among them DNA repair[54] leading toward the search of modifier genes using a more data-driven approach.

1.8. Intergenerational instability of the CAG repeat in humans

Changes in the number of repeats in parent-offspring transmissions were identified immediately upon the discovery of the expanded CAG repeat in *HTT* as the genetic cause for HD[19]. Intergenerational instability of the repeat is partly responsible for the earlier onset in successive generations – anticipation – of the disorder[7,55,56]. It is also responsible for changes from intermediate repeat sizes into disease associated alleles, as well as changes from reduced penetrance CAG sizes into fully penetrant alleles[37,57,58]. The latter also explains some “sporadic” HD cases (where there was no previous familial history of the disorder) where these *de novo* mutations happened due to allele expansion in an unaffected parent[19] that may amount to up to 8% of the diagnosed cases[24].

Among individuals with expanded repeats, ~70-80% of transmissions are intergenerationally unstable[33,38,56,59,60], with maternally transmitted repeats being stable or leaning towards contractions, while transmissions through the paternal germline are less stable – including at lower repeat sizes as briefly mentioned previously (Figure 1) – and usually tend towards expansions[33,55,56,59–64]. Furthermore, most very large alleles, associated with juvenile-onset HD, usually emerge from paternal line inheritance[4,65]. Besides parental sex differences, repeat size also influences intergenerational instability, with longer CAG repeats having a tendency toward a higher frequency and larger TNR changes[33,56,60,63,64].

HD families show clustering of transmitted repeat size changes, indicating the presence of possible genetic modifiers of intergenerational instability[60,64]. Regarding possible *cis*-modifiers, a specific 4p16.3 haplogroup of 22 single nucleotide polymorphisms (SNPs), has been associated with the change of intermediate-sized alleles into disease range alleles[66], though later studies with a larger cohort showed no relationship between haplotypes (related to the haplogroup) with the size of the expanded CAG or with intergenerational instability[50,67]. Segregation of instability within the very large Venezuelan pedigree, which has its origin in a single founder and shares a single 4p16.3 *HTT* haplotype also raises the possibility of *trans*-effecting modifiers altering intergenerational instability[60]. Furthermore, other possible modifiers – such as parental age, offspring sex, and repeat size of the normal allele in heterozygotes – of TNR length changes across generations have been proposed, with mixed results regarding an actual modifying role[52,56,60,63,68].

1.9. Somatic instability of the CAG repeat in humans

Somatic instability of the CAG repeat was early identified in multiple tissues from patients, with most brain areas showing a higher degree of instability when compared with peripheral tissues[69]. Moreover, among brain regions, the striatum presented the largest changes – noticeably trending toward expansions – while the cerebellum showed the highest stability[69]. Early on, this lead to speculation that glial cells, very prominent in most brain regions when compared to the densely neuron-packed cerebellum, might be more prone to repeat changes and partially explain these differences[69]. Later observations comparing grey matter (neuron enriched) and white matter (glia enriched) showed the opposite, with predominantly neuronal matter showing larger repeat changes[70]. Laser capture microdissection of individual cell types confirmed this higher susceptibility of neurons to larger expansions[70].

Using techniques such as small-pool PCR, that capture CAG size changes at higher resolution, even more frequent and larger changes in TNR size were discovered in patients' brain tissue, namely in the striatum of individuals showing the earliest stages of the neuropathological process[71]. These differences in instability were lost when looking at striatal tissues in brains at higher grades of pathology. This indicates that the aforementioned changes may be present early in the disorder and might not be observed later due to earlier neuronal death and striatal atrophy caused by these large expansions; notwithstanding these alterations are still measurable in other structures such as the cortex[70,71]. Taking into account that large changes are already present in brain tissue at early stages of the disorder, there is a possibility they may precede and even affect HD symptoms[70,71]. In agreement with this hypothesis, the evaluation of a cohort of individuals presenting either very early or very late ages-of-onset, matched for inherited allele sizes, showed distinctively different distributions of expanded repeats, with individuals with lower ages-of-onset presenting higher maximum expansion values and overall higher relative frequency of large changes[72]. These expansion-trending profiles might therefore be precipitating disease and potentially explaining the earlier than expected onset[24,72].

Other cell types have also been studied in order to better understand and characterize somatic instability[34,73,74]. Buccal cell DNA from individuals with expanded alleles showed that repeat size affects expansion frequency[73] similarly to what was initially proposed in brain tissue[69]. While originally, lymphoblasts from patients were reported to show no instability[75], later studies have reported a limited but existent instability after a determined TNR threshold[74].

1.10. Model systems of HD

Many model systems have been created and used to study HD, among them, yeast, roundworm, fruit fly, zebrafish, numerous mice, rat, pig, sheep, rhesus monkeys, and human derived cell models, such as lymphoblastoid cell lines and induced pluripotent stem cells (iPSCs). Different models have their own limitations and advantages and the most appropriate model is dependent on the goal and object of study[76].

By far, most models developed for HD are murine models, and were created by a plethora of techniques and encompass wide-ranging repeat sizes[76]. Among the most utilized models for HD are knock-in models (Htt^{Q20}-Htt^{Q175})[77–80] and transgenic mice expressing N-terminal fragments of huntingtin with expanded polyQ tracts (R6)[81].

Knock-in mice are genetically accurate replicas of the human mutation, presenting one wild type allele (of mouse *Htt*) and one humanized allele in its proper sequence context[77–80]. Htt^{Q111}, Htt^{Q140} and Htt^{Q175} have been thoroughly characterized in terms of behavioral changes, showing reduced activity in the dark phase, coordination and motor learning deficits, gait impairment, and reduced sensitivity to odors, among other phenotypes[79,80,82–84]. Neuronal inclusions, neurodegeneration and gliosis are also observed in knock models at high repeat sizes[85–87]. While N-terminal truncated models tend to have an extensive degenerative phenotype with a fast onset of motor, cognitive and behavioral symptoms and a shortened lifespan[81]. Both knock-in and transgenic models show intergenerational repeat sizes changes as well as somatic instability, making them useful tools in characterizing CAG repeat instability.

1.11. Intergenerational instability in mouse models

Alterations in CAG repeat size across generations is also observed in many mouse models of HD[78–81] and many factors have been observed to influence this intergenerational instability.

Parental sex is a big determinant of changes, with paternal transmissions preferentially expanding while maternal transmissions most commonly show contractions[78,88,89]. Offspring sex was also proposed to influence intergenerational repeat changes, and while instability was similar in embryos of both sexes, male embryos showed higher frequency of changes towards increases in repeat size, while female embryos tended toward contractions[89]. Parental age does not appear to influence intergenerational TNR size alterations in knock-in models[78]. Larger parental repeats have been shown to result in more frequent repeat length changes[78]. *Cis*-acting factors and genomic context are also thought to play a role in intergenerational instability as transgenic mice with distinct insertion sites for the repeat show very observable differences in parent-to-offspring stability[81]. Htt^{Q111} knock-in mice with different background strains suggest a role for genetic background and *trans*-acting factors in parent-offspring repeat size changes, and indeed genes have been shown to modulate intergenerational instability, namely *Msh2*, *Msh3*, *Msh6*, and *Neil1*[90–92].

1.12. Somatic instability in mouse models

Somatic variation in CAG repeat was identified in the commonly used HD mouse models[78,81]. The tissue specificity of instability is concordant in most models, with high instability in striatum (similarly to what was observed in patient post-mortem brains) and liver, while other organs and tissues such as tail, heart, spleen and cerebellum showed higher stability[78,81,91,93–96].

Somatic instability was shown to depend on repeat size, with longer repeats showing increased instability[78]. Somatic repeat size changes highly correlate with age, with changes prone to larger and more frequent expansions[96–98]. Age dependence has also been observed in fully differentiated cells indicating non-replication pathways are involved in instability[99,100]. Transgenic mice models (R6) with distinct insertion sites for the repeat have shown that genomic context alters repeat stability independently of the tissue analyzed[81]. Genetic background effects have been described to alter somatic instability with knock-in mice in the 129 strain showing decreased striatal instability when compared to mice in B6 and FVB strains[91,101]. Modifiers of somatic CAG repeat instability have been identified using mouse models for the disorder[90,92,102,103]. Linkage mapping lead to the identification of *Mlh1* as the modifier responsible for the differences in striatal instability between B6 and 129 strains. This discovery was validated through crossing knock-in *Htt*^{Q111} mice with knock-out *Mlh1* mice, where all expansions were eliminated in the striatum and liver[102]; and crosses with *Mlh3* knock-outs showed the same consequences[102]. Crossing *Htt*^{Q111} with expanded repeats had also previously identified other modifiers of striatal instability, namely *Msh2*[90] and *Msh3*[104]. Crosses with R6/1 mice have been used for the same purpose, identifying *Neil1* and *Ogg1* as modifiers[92,105]. Functional polymorphisms in the *Mlh1* and *Msh3* may mediate the role of this machinery in repeat instability[102,103]. Furthermore, other approaches have been used to identify candidates of instability modulation, namely through the evaluation of genes differentially expressed in stable versus unstable tissues (such as cerebellum versus striatum)[106,107].

Having all of the previously mentioned information in mind, the projects presented in this thesis have different specific goals, and have focused on different HD models, containing both experimental and data analysis components, but share the global objective of finding factors that influence CAG repeat instability, so that this biological phenomenon might be better understood, and might one day be a pathway toward a therapeutical possibility for patients and carriers of the expanded *HTT* CAG repeat.

2. CAG repeat instability in human lymphoblastoid cell lines (LCLs) and germ line

2.1. Introduction

While a typical presentation of HD starts in adult life, with a mean onset at 45 years of age, the disorder's age-at-onset is highly variable and may start in early childhood or late in life[14,52]. CAG length has a strong relationship with age-at-onset and may explain up to 70% of variance regarding disease onset age[14]. Given this relationship, it is important to consider the role of parent-to-offspring changes in repeat size, which have been observed to occur more frequently in paternal transmissions and trend toward expansions[56,60,63,64,75], partially explaining the genetic anticipation seen in the disorder[55,62].

Nonetheless, factors other than inherited repeat size are at play, as even considering two individuals with a similar CAG size, their age-at-onset might vary up to 20 to 25 years[51], although this variation also has heritable components, indicating the presence of modifier genes[14]. Taking this into account, efforts have been made to identify these modifiers; notably a large GWAS studying modulators of this residual age of motor onset (variation of age-at-onset non-dependent on CAG size effect)[54]. This study identified *loci* associated with residual age-at-onset, some of them containing genes related to DNA repair machinery, such as *MLH1* and *FAN1.* Furthermore, association analyses among pathways were also performed and three stood out oxido-reductase activity, mitochondrial fission, and DNA repair[54].

Interestingly, the modifiers of somatic instability identified in mouse models of HD are DNA repair genes[90,102–104], and in patients it was shown that higher levels of somatic instability (namely somatic expansions) have been associated with younger age of disease onset, raising the possibility of somatic instability being an intermediary in the relationship between age-at-onset modulation and the results mentioned above. Additionally, natural occurring SNPs between mouse strains (namely in *Msh3*, and *Mlh1*) have been proposed to be responsible for this instability modulation[102,103].

Mouse models have also given us some clues about possible relationships between somatic and intergenerational instability. Increasing CAG repeat sizes in knock-in models show both higher levels of intergenerational and somatic changes[78–80]. In early evaluations within the transgenic R6 models presenting no somatic instability also appeared to show more limited intergenerational changes[81]. Regarding *trans*-acting factors most, but not all, seem to modulate both intergenerational and tissue CAG changes[90,102,104]. There is considerable evidence of possible overlap between somatic and intergenerational instability in mice models but in humans the relationship is of yet unknown, but would be important to assess as one might be able to assess if shared mechanisms are present, and therefore find ways to modulate/prevent both types of instability at once.

While mouse models have been of great importance to study the disorder in general, instability and instability modifiers in particular, human models for instability assessment and characterization are still limited, with some efforts made with lymphoblastoid cell lines (LCLs)[74] and iPSC models have been successfully used in other repeat disorders[108].

With all of this in mind this chapter contains several projects related to LCL instability and intergenerational/germline instability: 1) LCLs derived from four patients within a nuclear family presenting an unusually high level of intergenerational instability were studied in order to assess if there might be an unusually high level of somatic instability in the transmitting parent, as well as to characterize if/how cell culture and passages might influence repeat size, and evaluate the potential of these cell lines as human-derived models for instability studies; 2) assess possible relationships between somatic and germline instability through the evaluation of instability, expansion and contraction indexes in LCLs and sperm of individuals within a specific CAG size range; 3) analyze the genetic association of a *MLH1* marker (rs1799977), situated in one of the candidate regions identified in the age-at-onset modifier GWAS, with the three measures of instability in LCLs and sperm samples; 4) using instability data from a series of LCLs spanning a large range of CAG repeat sizes, and RNA-seq expression data of DNA repair machinery genes, provide data-driven and hypothesis-generating candidate instability modifiers based on correlations of instability values and expression of DNA repair genes.

2.2. Methods

2.2.1. Cell culture

The LCL culture here described, pertains only to the nuclear family poject mentioned above and were grown as part of the core facility at the Center for Genomic Medicine (CGM), Massachusetts General Hospital (MGH).

Lymphoblastoid cell lines (LCLs) were cultured in Gibco™ RPMI Medium 1640 (ThermoFisher Scientific) with 10% penicillin-streptomycin (10,000 U/mL, ThermoFisher Scientific) at 37 °C. Media was replaced on average every 2-3 days, and cells were passaged on average every 5-7 days, and in every passage one pellet was collected and frozen at -80 °C for subsequent DNA extraction. Cultures were maintained for a total of 2-3 months and, in total, pellets from thirteen passages were collected.

2.2.2. LCL DNA extraction

DNA was extracted using a modified version of the 5 PRIME (Fisher Scientific) Manual ArchivePure DNA Purification methodology followed by a standard phenol-chloroform extraction. Pellets were thawed on ice and 300µL of Cell Lysis Solution (5 PRIME) and 5µg of Proteinase K were added, followed by mechanical homogenization of the solution and a 3-4h incubation at 50 °C. Afterward, in a fume hood, 300µL of a phenol:chloroform:isoamyl alcohol (25:24:1, v/v) were added, followed by a vortex solution homogenization, and samples were then centrifuged at 13000 rpm for 5 min. The aqueous top layer was transferred to a new tube and ~1 volume of chloroform was added followed by a 2 min centrifugation at 13000 rpm. The top layer was again transferred to a new tube, sodium acetate was added to a final concentration of 0.3M, followed by 600µL of absolute ethanol and homogenization was performed by manual inversion (15 to 20 times) and centrifuged at 4 °C, 15000 rpm, for 30 min for DNA precipitation. Supernatant was discarded and DNA pellets were washed two times with 70% ethanol followed by a 2 min centrifugation at 13000 rpm. Ethanol supernatants were discarded. Samples were left to air dry for 20-30 minutes and were then solubilized in TE buffer.

DNA concentration was determined through spectrophotometry using a Nanodrop 1000 and afterward diluted to 40ng/µL for downstream use.

LCL samples and all the sperm samples utilized in the somatic and germ line instability study were available at a DNA solution stage, and had been sized with a standardized assay for CAG length, consequence of previous large collection of samples by the Gusella and MacDonald groups at the CGM and access to these samples was kindly provided to us.

2.2.3. Fragment analysis and instability calculation

Fragment analysis was performed to evaluate instability, starting with a polymerase chain reaction (PCR) of the sequence containing the expanded repeat. PCRs were performed using the *Taq* PCR Core Kit (Qiagen) where each reaction contained 1x PCR buffer (Qiagen), 20% Q-solution (Qiagen), 16 nmol of HU3 (5' GGCGGCTGAGGAAGCTGAGGA 3') and labelled CAG1 (5' 6-FAM-ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC 3') primers, 4 nmol of dNTPs and 0.5 units of *Taq* in a final volume of 20µL. Cycling conditions were as follows, 95 °C 5min, 30 cycles of 94 °C 30sec ,65 °C 30 sec, 72 °C 90sec, followed by 10 min at 72 °C. Products were then subjected to capillary electrophoresis in an ABI 3730 (Applied Biosystems) sequencer. Electropherograms were analyzed with GeneMapper Software v5.0 (Applied Biosystems) undergoing rigorous quality control and validation.

Instability was determined using a method previously described[93], in short: the PCR reaction generates multiple-sized products that when analyzed in GeneMapper appear as a cluster of peaks differing from each other by one CAG repeat unit, and peak height is assumed to be proportional to the number of alleles containing a specific repeat size; the peak containing the highest signal is considered the “main allele”; alleles to the right are considered expansions and alleles to the left contractions; heights for all alleles of interest are extracted and individual peak heights are normalized to the sum of all heights; normalized peak heights are then multiplied by a factor determined by the change/distance to the main allele (*e.g.* for an allele 1 CAG smaller than the main allele this factor is -1, for an allele 2 CAGs larger the factor is 2, and so forth); the values stemming from this calculation are then added together generating the instability index. Expansion and contraction indexes can also be determined, adding only the peaks to the right or left of the main allele.

In this study, slight changes to the methodology were applied; instead of adding a relative threshold (*e.g.* 20% height of the main allele) as described in the original method[93], an absolute threshold of 50 for minimum peak height was applied, and alternatively to the instability index, an absolute instability index – consisting of the sum of the expansion index and the absolute value of the contraction index – was calculated.

2.2.4. RNA-seq data

RNA-seq data was acquired from a collaboration with other groups at the CGM and was acquired as described in [109].

2.2.5. Statistical analyses

Ordinary least squares regressions for evaluation of relationships between CAG repeat size and instability measures (absolute instability, expansion index and contraction index) were performed in Microsoft Excel 2013. Possible correlations between somatic (LCL) instability and germ line (sperm) instability were also analyzed through the same regression method and software. Genotypic association analysis of instability was evaluated through linear regression using SPSS Statistics 20 (IBM). Residual values of absolute instability, expansion and contraction indexes were also calculated with SPSS Statistics 20. Pearson correlation values between residuals and gene expression values were also calculated in the same software. To control for multiple testing the Bonferroni correction was applied to significance levels when appropriate.

2.3. Results

2.3.1. LCL instability in a nuclear family presenting intergenerationally unstable *HTT* CAG repeat transmissions

The nuclear family, of American origin, analyzed in this study consisted of an affected father, originally determined to have a repeat with 42 CAGs, an allele on the low end of fully penetrant sizes, while his offspring showed much larger CAG repeat sizes in the range of 80 to 139 repeats (Figure 3). Therefore, he showed a very high propensity for large intergenerational repeat size changes.

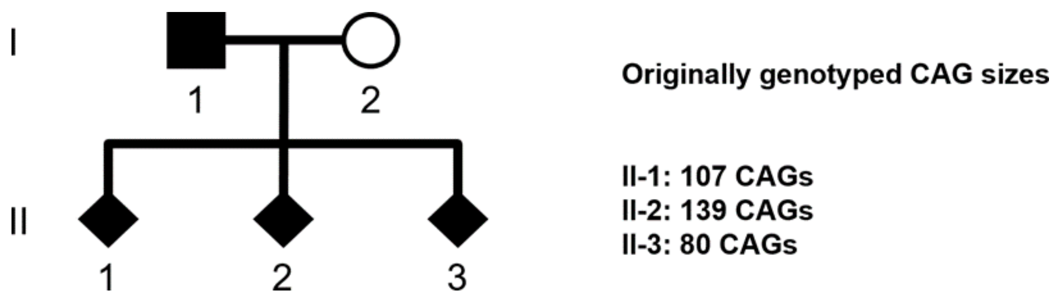


Figure 3 – Pedigree and *HTT* CAG repeat information in the nuclear family in study

Detailed bar-plots depicting relative allele frequency for all passages and for all lines are presented in Figure S1 to Figure S4. For added clarity of interpretation they are shown in increasing order for the individuals' main allele size (*i.e.* I-1, II-3, II-1, II-2). For a simplified view of repeat length changes heat maps depicting the relative abundance of each allele size per passage are presented (Figure 4 to Figure 7), following the same order mentioned above, for consistency and clarity.

CAG repeat size

Passage number	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
P1		0.18	0.22	0.25	0.27	0.31	0.43	1.03	2.82	10.04	27.91	36.64	10.01	5.86	3.92	0.50	0.25	0.21	0.18	0.18	0.08
P2		0.25	0.29	0.29	0.34	0.38	0.53	1.08	3.14	10.16	29.70	34.41	9.42	5.13	4.07	0.49	0.18	0.13	0.11	0.10	
P3			0.17	0.22	0.17	0.20	0.36	0.96	3.01	10.13	28.40	43.32	6.69	3.31	2.86	0.36					
P4			0.23	0.28	0.27	0.23	0.38	0.99	3.08	10.54	28.73	45.44	6.25	2.68	2.26	0.33					
P5	0.17	0.14	0.14	0.17	0.19	0.26	0.43	1.09	3.33	11.15	29.83	46.22	5.07	1.31	0.74	0.15					
P6			0.19	0.20	0.17	0.20	0.35	1.09	3.47	12.13	33.05	43.43	4.99	0.74	0.26						
P7					0.15	0.19	0.38	0.99	3.53	11.62	30.95	46.58	4.97	0.59							
P8																					
P9				0.28	0.31	0.34	0.46	1.13	3.70	11.86	33.07	42.38	6.07	0.86	0.16						
P10			0.13	0.21	0.26	0.28	0.44	1.12	3.69	11.91	33.24	41.19	6.39	0.76	0.22						
P11			0.11	0.23	0.26	0.22	0.45	1.20	3.75	12.77	33.23	40.05	7.05	0.84							
P12							0.34	1.01	3.53	11.21	29.90	47.48	6.17	0.47							
P13							0.36	0.95	3.18	10.36	28.99	47.21	6.43	1.96	0.65						

Figure 4 – Relative signal height (%) per allele in each passage in individual's I-1's LCLs. Underlined values show the highest abundance allele within a population. Green indicates a low abundance, yellow and orange intermediate and red highest for each passage.

CAG repeat size

Passage number	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
P1							0.33	0.71	1.38	1.88	1.50	1.53	2.20	5.30	12.34	23.51
P2			0.21	0.25	0.25	0.35	0.60	1.21	2.31	3.10	1.69	1.28	1.85	4.10	10.21	20.52
P3				0.24	0.22	0.34	0.65	1.38	2.65	3.65	1.99	1.29	1.83	4.19	9.47	18.60
P4		0.19	0.21	0.25	0.28	0.42	0.71	1.35	2.62	3.63	2.07	1.28	1.76	3.83	8.79	17.06
P5					0.22	0.26	0.51	1.20	2.33	3.31	2.09	1.10	1.56	3.56	8.29	16.52
P6					0.22	0.27	0.54	1.22	2.51	3.74	2.87	1.39	1.38	3.08	7.34	15.28
P7			0.22	0.27	0.33	0.43	0.79	1.53	3.12	5.08	4.80	2.47	1.61	2.79	6.22	12.66
P8			0.25	0.28	0.37	0.44	0.74	1.62	3.54	6.52	7.34	3.89	1.90	2.68	5.87	12.18
P9			0.28	0.34	0.38	0.41	0.63	1.19	2.39	4.45	5.75	3.25	2.39	3.82	8.54	16.98
P10			0.22	0.27	0.34	0.29	0.39	0.68	1.25	2.16	2.79	2.18	2.36	4.72	10.74	21.25
P11			0.26	0.37	0.48	0.33	0.41	0.54	0.88	1.20	1.23	1.29	2.12	4.89	11.17	22.50
P12			0.25	0.40	0.47	0.27	0.20	0.24	0.38	0.52	0.65	0.89	1.74	4.43	10.91	22.52
P13	0.21	0.19	0.24	0.43	0.50	0.32	0.27	0.30	0.40	0.50	0.65	0.87	1.67	4.27	10.51	21.40

Passage number	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
P1	29.17	14.34	4.37	1.11	0.34	0.27									
P2	27.83	16.13	5.99	1.74	0.57	0.26									
P3	26.43	16.99	7.06	2.28	0.75	0.24	0.21								
P4	25.18	17.76	8.19	3.01	0.97	0.31	0.19								
P5	24.12	18.99	9.95	4.14	1.37	0.40									
P6	22.38	19.49	11.30	5.09	1.81	0.52									
P7	19.78	17.79	10.95	5.63	2.15	0.64	0.25	0.17							
P8	18.49	15.28	9.64	5.41	2.10	0.73	0.29	0.18	0.16	0.15	0.13	0.15	0.16	0.14	0.14
P9	23.13	13.22	6.82	3.49	1.38	0.52	0.24	0.16	0.18	0.16	0.16				
P10	27.45	13.63	5.72	2.45	0.95	0.36	0.18								
P11	28.42	15.06	5.68	1.98	0.74	0.27	0.20								
P12	29.76	16.88	6.86	2.13	0.68	0.27									
P13	29.13	17.76	7.19	2.31	0.73	0.27									

Figure 5 – Relative signal height (%) per allele in each passage in individual II-3's LCLs. Underlined values show the highest abundance allele within a population. Green indicates a low abundance, yellow and orange intermediate and red highest for each passage.

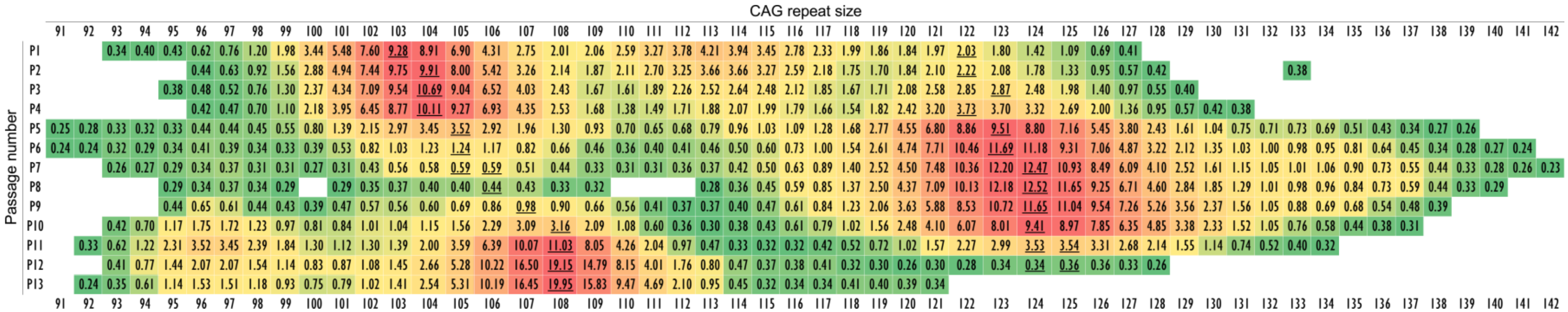


Figure 6 – Relative signal height (%) per allele in each passage in individual II-1's LCLs. Underlined values show the highest abundance allele within a population. Green indicates a low abundance, yellow and orange intermediate and red highest for each passage.

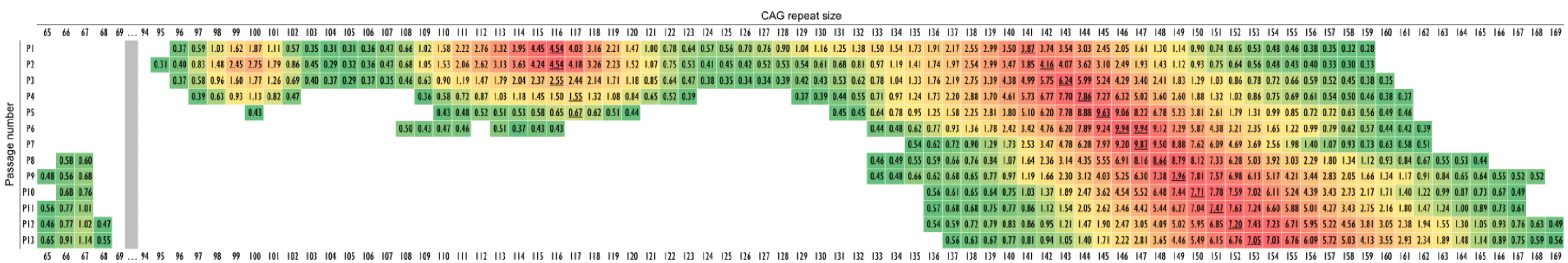


Figure 7 – Relative signal height (%) per allele in each passage in individual II-2's LCLs. Underlined values show the highest abundance allele within a population. Green indicates a low abundance, yellow and orange intermediate and red highest for each passage. Green indicates a low abundance, yellow and orange intermediate and red highest for each passage. Grey bar replaces the 70-93 CAG interval, where no alleles were found.

Overall, mosaicism is observed in all four LCLs studied, although to very different extents.

The affected father (individual I-1), showed the lowest CAG repeat size – (always) consistent at 51 CAGs – and the tightest range of allele sizes, seemingly becoming more constricted as the number of cell passages increased (Figure 4). The offspring bearing the lowest CAG repeat size (individual II-3, 81 CAGs, Figure 5), showed the same tendency for the main allele size to stay consistent throughout passages. Nonetheless, the presence of a second allele population (Figure S2, P2-P10) is noticeable in early passages but eventually dwindles in later passages (Figure S2, P11-P13).

LCLs of the two other offspring show much more complex behavior. LCLs from individual II-1 initially showed allele sizes in three main sets (~103, ~112, ~122 CAGs, Figure 6), probably representing 3 cell populations bearing those sizes. Throughout passaging, the LCLs show an intricate pattern of CAG size distribution, with the population with the highest size showing an increasing prevalence during most passages (Figure 6, Figure S3 P1-P8), with a steep decrease, to complete non-representation in the last studied passage (Figure 6, Figure S3 P13), where the most common allele size is 108 CAGs. Interestingly, if we follow the most prevalent allele per “population” and per passage (Figure 6, underlined), we see a mostly steady trend moving toward larger allele sizes. In II-2’s LCLs this trend is clearly observed in the largest allele size population whose main allele starts at 141 CAGs (Figure 7, Figure S4) and in the last observed passage stands at 153 CAGs (a ~1 CAG average increase per passage in main allele size). This LCL also started with 3 main sets of alleles (at ~100 and 116 CAGs in addition to 141, Figure 7, Figure S4), but the smaller sizes populations were not observed after approximately half the passages (P6). Of note is the overall minor but increasing presence of an allele population at ~67 CAGs (starting at P8, Figure 7, Figure S4) in a range that was not observed in any of the previous passages and much smaller than the any of the alleles initially present.

2.3.2. Relationship between LCL and germline instability

To better assess a possible relation between somatic and germ line instability, DNA samples from LCL and sperm samples of a set of 19 individuals showing expanded alleles within a restricted CAG repeat size range (41-45 CAGs) was evaluated.

After absolute instability, expansion and contractions indexes were calculated, possible correlations with CAG repeat size were evaluated (Figure S5 to Figure S7), as previous studies have shown dependency of instability with CAG repeat size[69,73].

In LCLs, no correlation was apparent (absolute instability: $R^2 = 0.0473$; expansion: $R^2 = 0.0601$; contraction $R^2 = 0.0015$, Figure S5), while sperm samples appear to show a higher correlation (absolute instability: $R^2 = 0.2223$; expansion: $R^2 = 0.2078$; contraction $R^2 = 0.2721$, Figure S6). One of two possibilities might explain this difference; either higher CAG sizes are driving higher instability, or instability levels are so widespread that the most frequent allele (taken as the main allele), might no longer match the inherited repeat size and that change is driving the correlation. Since we are using a set of samples where we also have repeat sizing in LCLs, which are more stable than male germ line cells, and may be a better indicator of the modal allele in the population of PCR products, these correlation levels were re-assessed using the CAG size in the LCLs. In this scenario no correlations were observed (absolute instability: $R^2 = 0.0113$; expansion: $R^2 = 0.0094$; contraction $R^2 = 0.0330$, Figure S7), probably indicating that the previous observation stemmed from high instability levels changing main allele size driving the correlation levels up in germ line cells.

Therefore, considering that in this range the effect of CAG on the measures of instability we are using is minimal or non-existent, and that LCL and sperm samples compared come from the same individual, we went on and evaluated possible correlations between instability indexes (Figure 8 to Figure 10).

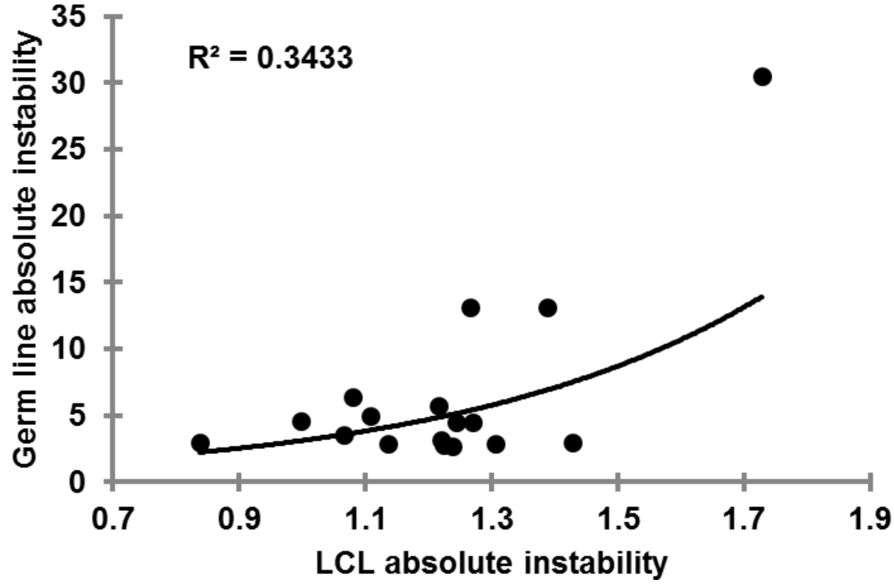


Figure 8 – Absolute instability in LCLs and germ line

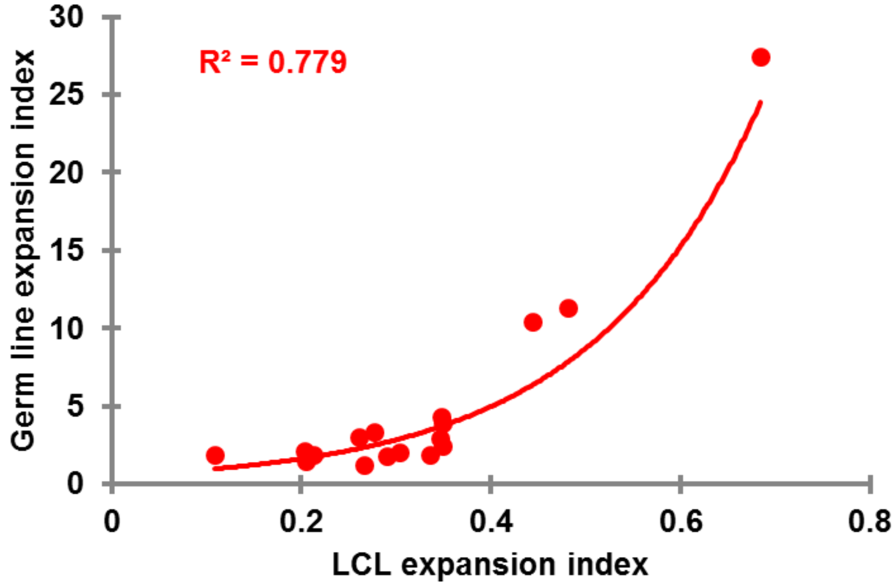


Figure 9 – Expansion index in LCLs and germ line

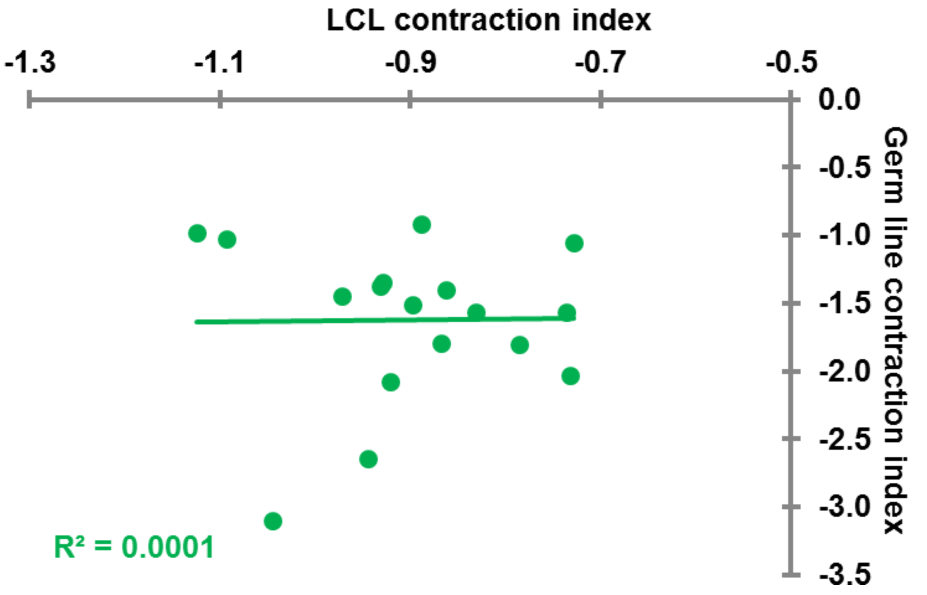


Figure 10 – Contraction index in LCLs and germ line

Interestingly, these preliminary analyses show that LCLs and sperm cells present a correlation of absolute instability ($R^2 = 0.3433$), apparently driven by a highly correlated expansion index ($R^2 = 0.779$), while contraction indexes seem to show no relationship ($R^2 = 0.0001$).

2.3.3. Testing the association of rs1799977 (*MLH1*) with instability in LCL and germline samples

Using the samples from the project mentioned above (as the 19 individuals had been selected having into account their rs1799977 genotype, obtained as part of the previously mentioned GWAS, for this segment of the project), we also evaluated their instability's relationship with the individuals' genotype for a marker present in one of the main regions identified to influence residual age-at-onset.

Association was evaluated through regression analysis for both additive and dominant effects. Representations of absolute instability evaluation in LCLs using either the additive and dominant models are represented in Figure 11.

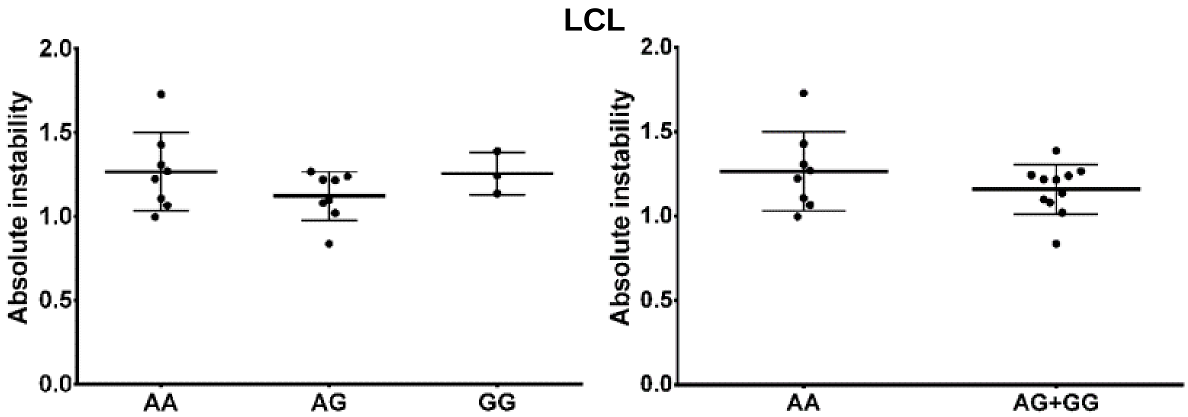


Figure 11 – Absolute instability per rs1799977 genotype considering an additive effect model (left) and a dominant effect model (right).

The evaluation of possible associations between rs1799977 genotype and absolute instability, expansion and contraction indexes showed no significant associations (Table 1).

Table 1 – Significance values for association analysis between rs1799977 genotypes and absolute instability, expansion and contraction indexes.

		Regression p-value	
		Additive effect model	**Dominant effect model**
LCL	**absolute instability**	0.582	0.236
	expansion index	0.886	0.705
	contraction index	0.526	0.199
Sperm	**absolute instability**	0.829	0.684
	expansion index	0.769	0.626
	contraction index	0.507	0.582

Therefore, thus far we may suggest that rs1799977's genotype does not seem to be responsible for alterations in instability levels in LCLs or germ line, in a very limited sample that might not possess enough statistical power or be fully representative a possible effect in a larger number of patients, and should therefore be expanded to include a larger number of individuals.

2.3.4. Search for instability modifiers using human LCLs

In a separate set of LCLs from 24 individuals with a range of CAGs (pertaining to the longer allele size) spanning all the way from normal alleles to very large alleles generally associated with the juvenile form of the disorder (17 to 82 CAGs), instability assessments were performed and indexes calculated for each line. Taking advantage of RNA-seq data from the same lines also being available, expression data from 118 genes (Table S1) involved in numerous pathways related to DNA damage response, repair, and recombination, potentially involved in instability modulation, were analyzed to assess and prioritize possible candidates in a human derived cell model.

Once again, we first evaluated possible correlations between main CAG repeat size and instability levels. We observed that this expanded range absolute instability and the contraction index present high correlation coefficients with CAG size (absolute instability: $R^2 = 0.8375$, Figure 12; contraction index: 0.8701, Figure 14), while expansion index presents a milder correlation value ($R^2 = 0.69$, Figure 13).

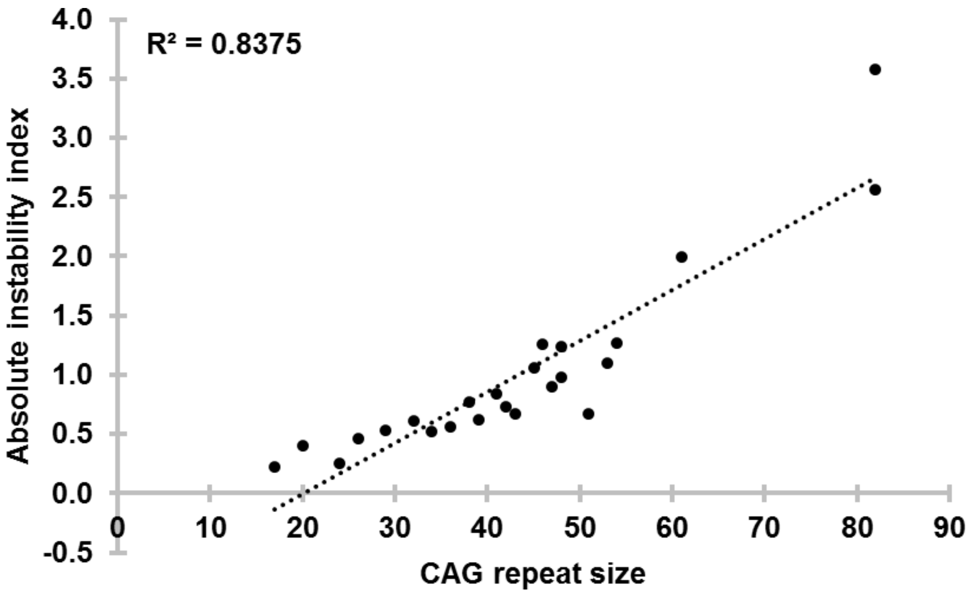


Figure 12 – Absolute instability variation with CAG size in LCLs

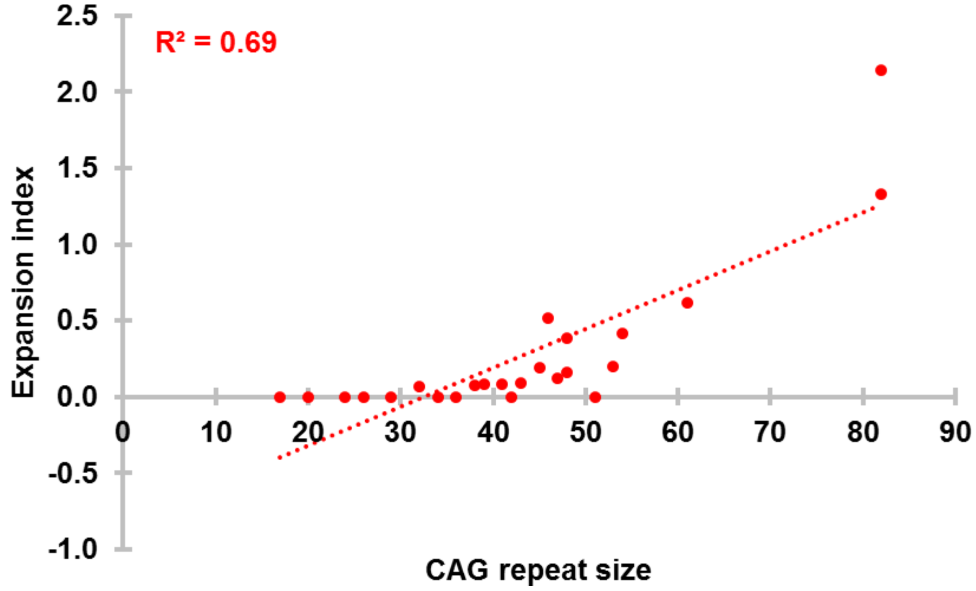


Figure 13 – Expansion index variation with CAG size in LCLs

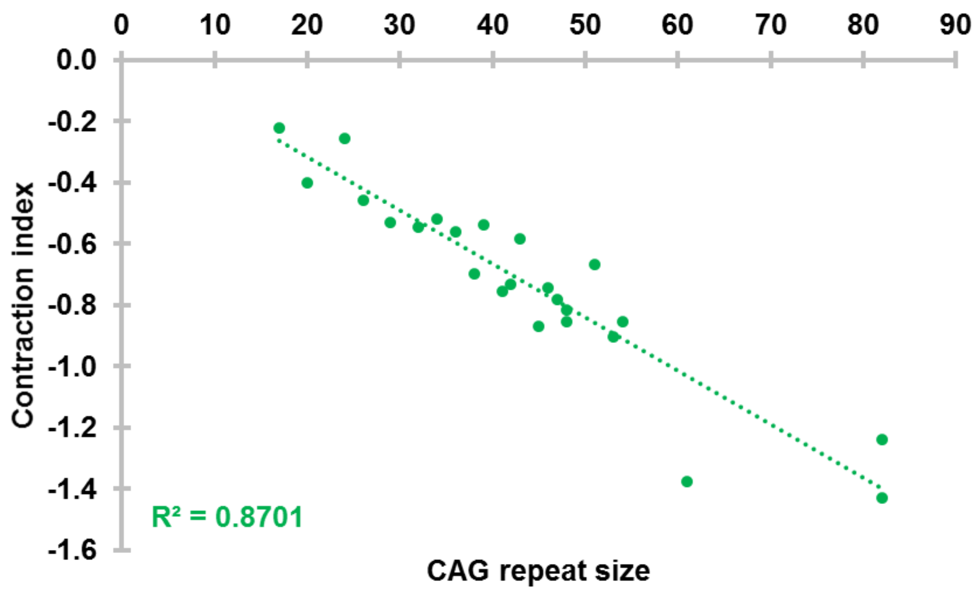


Figure 14 – Contraction index variation with CAG size in LCLs

This is an important observation, as it indicates that direct analyses of instability indexes, could be heavily confounded and/or determined by CAG repeat size.

Indeed, if we took these analyses forward as is, testing for significant correlations between gene expression and CAG size, absolute instability, expansion or contraction indexes, we see that the expression of 30 genes has a nominal significant correlation with CAG size, while 15 have a relationship with absolute instability, 8 with the expansion index and 28 with contraction levels (Table S2; Table S3), but all genes related to absolute instability and expansion indexes are included in the CAG correlated set and only 5 genes correlated with contraction index do not follow this trend (Figure 15).

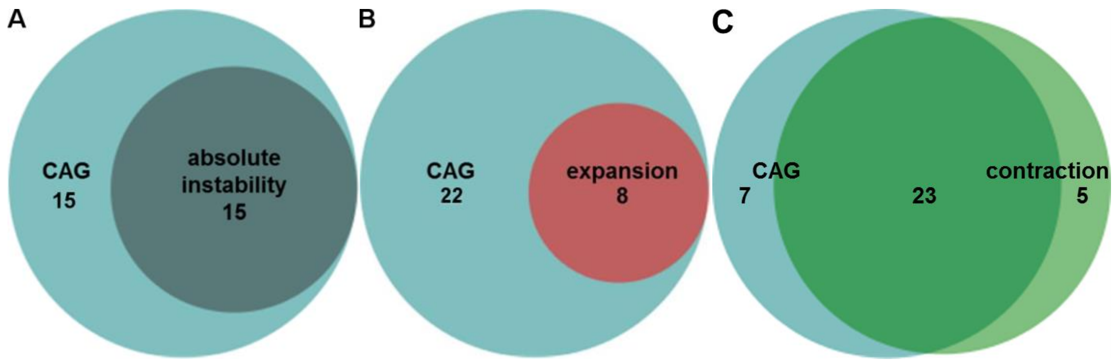


Figure 15 – Overlap of genes correlated with CAG repeat size (teal) and absolute instability (A, grey), expansion index (B, red) and contraction index (C, green)

Therefore, a different approach was taken, calculating linear regressions' residuals from the relationships of CAG size with absolute instability, expansion and contraction indexes, determining values of "CAG-independent instability", and studying their correlation with gene expression.

In these analyses 25 genes showed nominally significant ($p<0.05$) or significant ($p<0.0004$ significance threshold after Bonferroni correction) correlations between their expression and CAG-independent instability/expansion/contraction (Table 2).

Table 2 – Significant and nominally significant correlations between absolute instability, expansion and contraction residual values and gene expression.

Absolute instability residual			Expansion index residual			Contraction index residual		
Gene	**Correlation**	***p***	**Gene**	**Correlation**	***p***	**Gene**	**Correlation**	***p***
TP73	-0.632	0.001	*TP73*	-0.556	0.005	*NTHL1*	0.754	<0.0004
RPA3	-0.431	0.036	*SSBP3*	0.434	0.034	*POLD1*	0.670	<0.0004
SSBP3	0.419	0.042				*RAD23A*	0.627	0.001
						MDM2	-0.590	0.002
						SWSAP1	0.549	0.005
						FAN1	-0.549	0.005
						TDP2	-0.528	0.008
						LIG1	0.506	0.012
						SPEN	0.505	0.012
						LONP1	0.500	0.013
						STRA13	0.490	0.015
						MCM7	0.476	0.019
						DMC1	0.474	0.019
						IGHMBP2	0.469	0.021
						C10orf2	0.452	0.027
						MSH5	0.447	0.028
						SSBP4	0.446	0.029
						NABP1	-0.446	0.029
						POLD2	0.443	0.030
						TREX1	0.443	0.030
						ATM	-0.424	0.039
						ERCC4	-0.421	0.040

Only two comparisons were significant after multiple testing correction; *NTHL1 and POLD1,* both as modulators of contraction index CAG-independent instability, with lower levels of expression relating to more prominent contractions. Among other nominally significant correlations there are some interesting results, such as *TP73*, which shows significance with both absolute instability and expansion index, with higher levels of gene expression protecting against expansions. *TP73* retains its significance in absolute instability when only individuals with 35 CAGs and over are considered (Table S4). Among the genes related to alterations in contraction index, two others, *FAN1* and *LIG1*, are of special interest, with *LIG1* expression showing a behavior not unlike *NTHL1 and POLD1*, while higher *FAN1* levels seem to lean toward larger contractions (Figure 16).

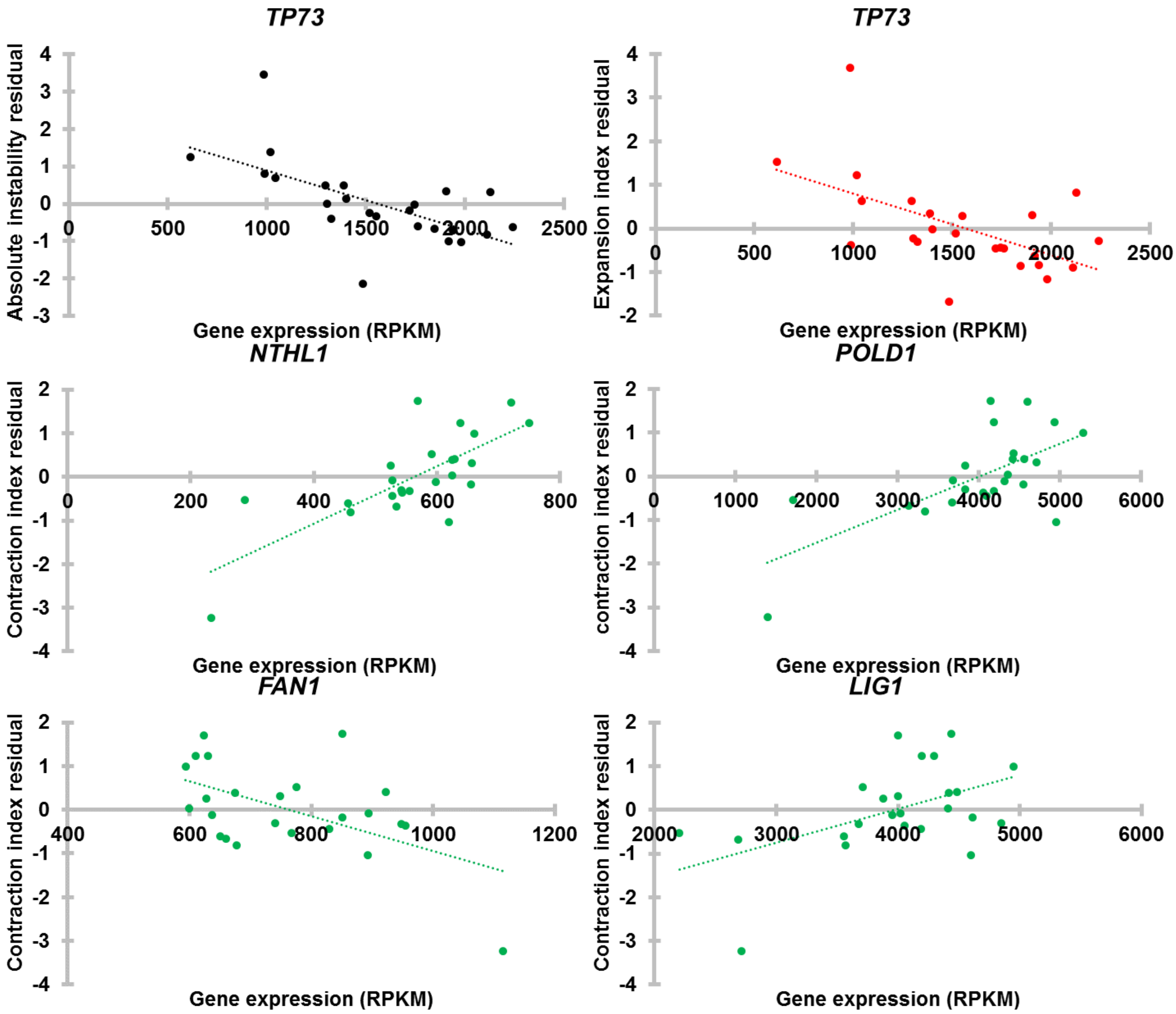


Figure 16 – Relationship between absolute instability (top left) and expansion index (top right) residuals and *TP73* expression levels, as well as, correlations between contraction index residuals and expression levels for *NTHL1* (middle left), *POLD1* (middle right), *FAN1* (bottom left), *LIG1* (bottom right).

2.4. Discussion

In this chapter many characteristics of somatic instability in LCLs, and some features of germline instability have been evaluated. These studies were mainly exploratory using different ways of evaluating LCL and germ line instability, meant to be hypothesis generating tools geared toward guiding and refining further analyses in the field.

Initially, an evaluation of LCL instability in a nuclear family where the father (individual I-1) presented a very strong trend to intergenerationally unstable transmissions, with changes reaching a difference of almost 100 repeats, and his three affected children was performed. This study had two main goals; 1) to work as a preliminary indicator of possible higher somatic instability when high levels of intergenerational instability are present; 2) to assess how LCLs with very distinct main CAG allele sizes behave regarding repeat instability with maintenance and passaging, in order to assess the suitability of these cells as human models to search for instability modulators (*e.g*. being used in – possibly high throughput – screening of small molecules).

The somatic cell line of individual I-1 seems to present a fairly high stability, with the maintenance of modal allele size. Extended culturing of the LCL seems to actually narrow the population variation in CAG repeat size. Considering that the individual showed a high propensity for intergenerational instability but a limited LCL instability one might argue thus far that there is no relationship between them.

Regarding the potential to use LCLs as human models for instability and instability modulation, efforts have been made previously[74] where very low or no instability was reported; even over a longer culturing time-frame than the one(s) presented in this project, only repeats 60 or larger presented repeat size changes[74]. Here, we see limited mosaicism at 51 CAGs, in agreement with the previously described behavior of LCLs in this repeat length range. Even though we see a larger overall spread of allele sizes, we still observe a consistent main allele size in the cell line harboring 81 repeats (individual II-3).

Nevertheless, we do observe interesting behaviors of relative distributions of CAG repeat sizes and instability in LCLs, when analyzing the cell lines from the offspring with over ~100 CAGs (individuals II-1 and II-2), showing that there is indeed a dependence of instability with repeat size.

The two lines with higher CAG sizes show some common and some disparate characteristics. Both of the lines appear to show high levels of mosaicism, with three different populations of alleles, probably due to the presence of three main cell populations in early passages. In individual II-2's LCLs the two smaller populations (at ~100 and 116 CAGs) dwindle and eventually disappear, possibly indicating that lymphoblasts with larger CAG repeat sizes may have some growth advantage in culture when compared to others with smaller CAGs. In individual II-1 we actually start seeing a similar trend but it appears the case is not as straightforward because the two smaller populations never actually disappear and come to overtake prominence at P11, while the larger size population quickly fades from the culture.

Another remarkable shared feature can be observed, very clearly in II-2's LCLs, and prominently, after careful evaluation, in II-1's cell line, which is the slow and steady increase in CAG size within a specific population, regardless of whether the population itself is becoming more prevalent overall or if it is declining. This shows that LCLs with long repeats exhibit a trend toward higher repeat sizes within a population, but not necessarily when considering distinct populations.

The feasibility of using LCLs as human derived cell models to study instability seems highly unlikely given the scenarios observed in this project, due to several aspects: 1) the cell line with the repeat size closer to the most common sizes in patients did not display enough instability to serve as a proper instability model, with no apparent repeat changes over thirteen passages; 2) even the relatively larger repeat size at 81 CAGs did not show a noticeable change in main allele repeat size, and while a second set of alleles seems to be arising in that line we cannot conclude it is due to instability and repeat size change, it might just be an effect of a pre-existing population with low initial representation that rises for a limited time-frame, confounding the results if any other treatment or screening were to be performed at the time; 3) noticeable instability was only observed when LCLs with very rare, very long, alleles were present and even then there isn't a unifying behavior that would allow a proper screening for modulators, as a multitude of populations bearing distinct repeat sizes are present, and do not seem to act similarly, with peculiar changes in the dominant population and modal allele size as well as possible large jumps, that are hard to attribute to anything other than random/stochastic effects hard to control for. Solutions for most of these problems are possible, but cumulatively require a multi-level, high effort, controlled environment, including: the initial "seeding" of a very low number of cells, insuring only one main and matching population of alleles is present at the start for both treated and control cultures, which would be very difficult to achieve; the insurance of similar growth levels in both scenarios, which would implicate increased handling and disturbance of the cell cultures; strictly paired technical aspects such as matched feeding, culturing, splitting, pellet collection, and feeding stages in order to minimize variability between treatment and controls would be required. All of this put together would defeat the overall purpose of using LCLs which would have been to acquire an easy and quick to use human-derived cell model to evaluate somatic instability.

Even though in the previous project, the hypothesized relationship between intergenerational and somatic instability (in LCLs, which are relatively stable) did not seem apparent, a single observation in one individual is insufficient to draw strong conclusions on their relationship. Therefore instability in LCLs and germline (sperm) cells of 19 individuals bearing a very restricted range of CAG repeat sizes was compared. Firstly, we saw that in the strict CAG repeat size range considered, correlation with instability was non-existent in LCLs and germline, therefore no CAG-driven effects should be present in the analysis. This time a modest but noticeable correlation between absolute instability in LCLs and sperm was observed, which seems to be driven by a strong correlation of expansion levels, while contraction indexes seem to be unrelated to each other. This is an interesting observation, which may indicate that specific genetic background and/or shared mechanisms may underlie both somatic and intergenerational expansion. As nineteen individuals might be a limited sample to take broad conclusions from, the analysis will be re-performed in a larger set of over hundred individuals, in order to validate the relationship observed here. Additionally, a previous study has also reported that higher constitutive CAG repeat size from LCLs correlates positively with modal sperm CAG size[60]

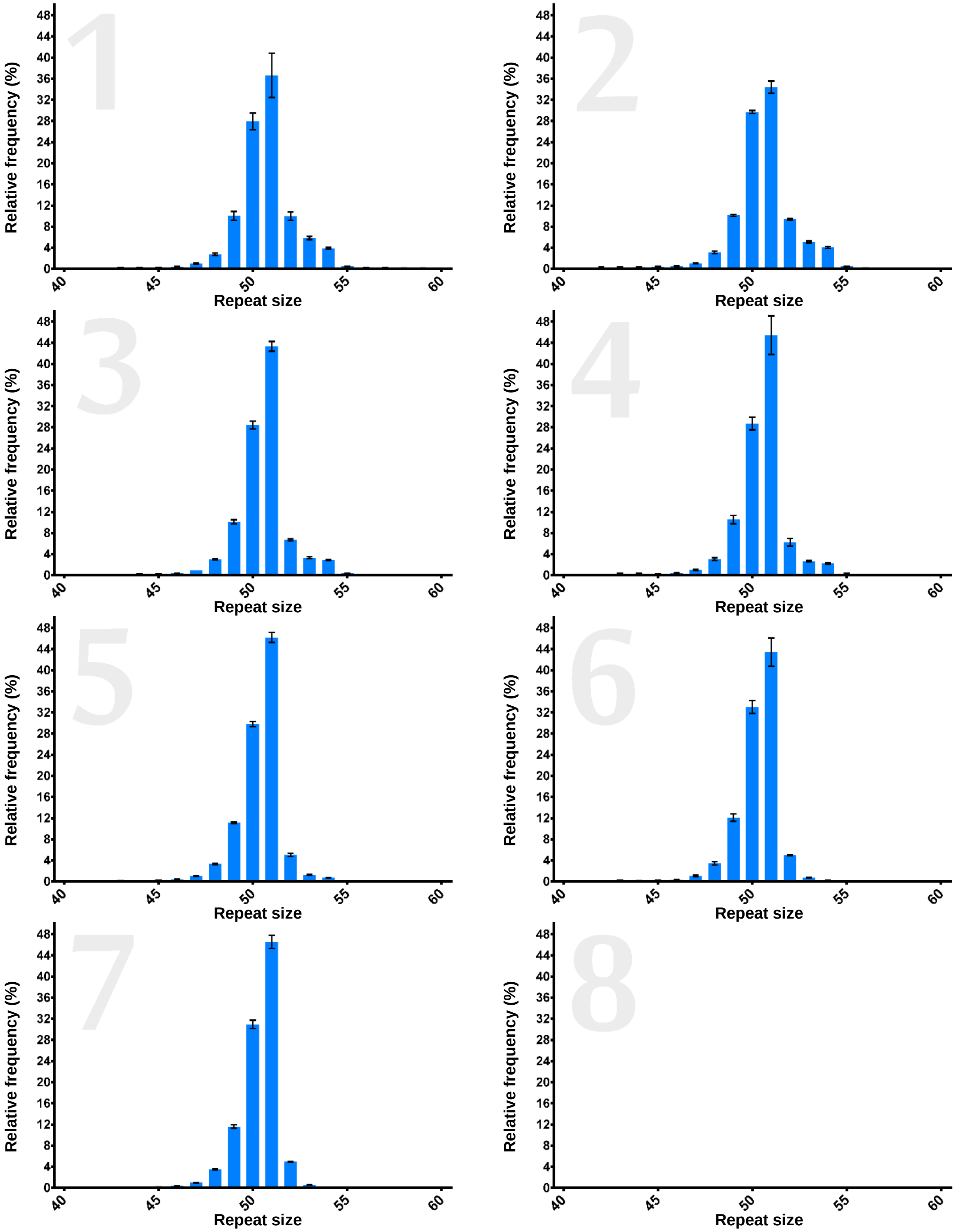
Using the same somatic and germline instability data as above, we evaluated the possible role of a *MLH1* SNP in these measures. *MLH1* was chosen over genes in the other significant *loci,* as mouse model data already indicated that common variation in *Mlh1* might be responsible for differences in instability between two different mouse strains[102]. This marker is part of a block of SNPs identified as one of the main candidate regions in modulating residual age-at-onset in HD, and therefore it would be interesting to evaluate a possible influence in instability. In this limited number of individuals, there does not seem to be a significant correlation of either LCL or germline instability with rs1799977 genotype. Again, as a very limited sample it is likely not to have sufficient power to identify actual differences and this analysis will be re-assessed in a larger sample. Nonetheless, LCLs might not be the best somatic tissue to check for instability as a driver for alterations in HD age-at-onset, and studies on a more relevant tissue for the disease (*e.g.* brain), might possibly show a different and more accurate relationship.

The last project focused on the search for genes involved in DNA damage, metabolism, and repair whose expression might be related to instability levels observed in a set of 24 LCLs, spanning a wide range of CAGs. Our first observation was related to a noticeable relation of CAG size and absolute instability, expansion and contraction indices, which would result in a high level of multicollinearity, which is problematic in regression analyses, and would make overall conclusions hard to interpret. Therefore we controlled for this relationship, by extracting the correlation's residuals as they would show the "CAG-independent" instability measures, and analyzed their correlation with gene expression. This analysis yielded only two significant results after multiple testing correction, namely the genes *NTHL1* and *POLD1* as modulators of contraction. *NTHL1* codes for endonuclease III-like protein 1, a protein involved in oxidative damage repair through base excision repair, and *POLD1* encodes the catalytic subunit of DNA polymerase delta and is therefore involved in DNA replication. Overall, no experimental data so far seems to relate the functions of these two proteins, with only a faint hint of a relationship given by putative homologs being co-expressed in other species, therefore their effects are probably independent of each other. Looking at other possible modulators within the nominal significance values, *TP73* is the most significant player related to absolute instability, as well as expansions. *TP73* codes p73, a protein from the same family as p53, participating in apoptotic response to DNA damage, and unlike the modifier genes found in mice whose knock-outs ablate expansions, with *TP73* higher expression levels protecting against expansions. Among the genes showing nominally significant association with contractions, two are important to mention: 1) *FAN1*, which is present in one of the *loci* described in the previously mentioned GWAS shown to modify HD age-at-onset. This might indicate that instability may be the mechanism interconnecting this association with the change in age-at-onset, in accordance with the observation that individuals bearing similar CAG repeat sizes but very disparate ages-at-onset also show very different levels of instability[72]; 2) *LIG1*, which encodes DNA ligase 1, involved in DNA replication and also responsible for the ligation step (the last one) in the MMR pathway, where other modifiers such as *MLH1*, *MLH3*, *MSH2* and *MSH3* act[90,102–104]. Furthermore the mouse homolog of *LIG1* has been linked to contraction modulation in the CTG repeat in the mouse model of DM1[110].

One also has to address that even though all the instability modifiers already found were evaluated in this analysis, none of them came out as significant, therefore this approach should be further scrutinized to try and understand this disparity, and control for other factors such as patient age at collection. Nonetheless, the results of this study yielded some interesting results in their “hypothesis generating” capacity, giving the genes identified a possible priority position when looking for more instability modifying genes in HD models.

2.5. Supplementary material

2.5.1. Supplementary figures



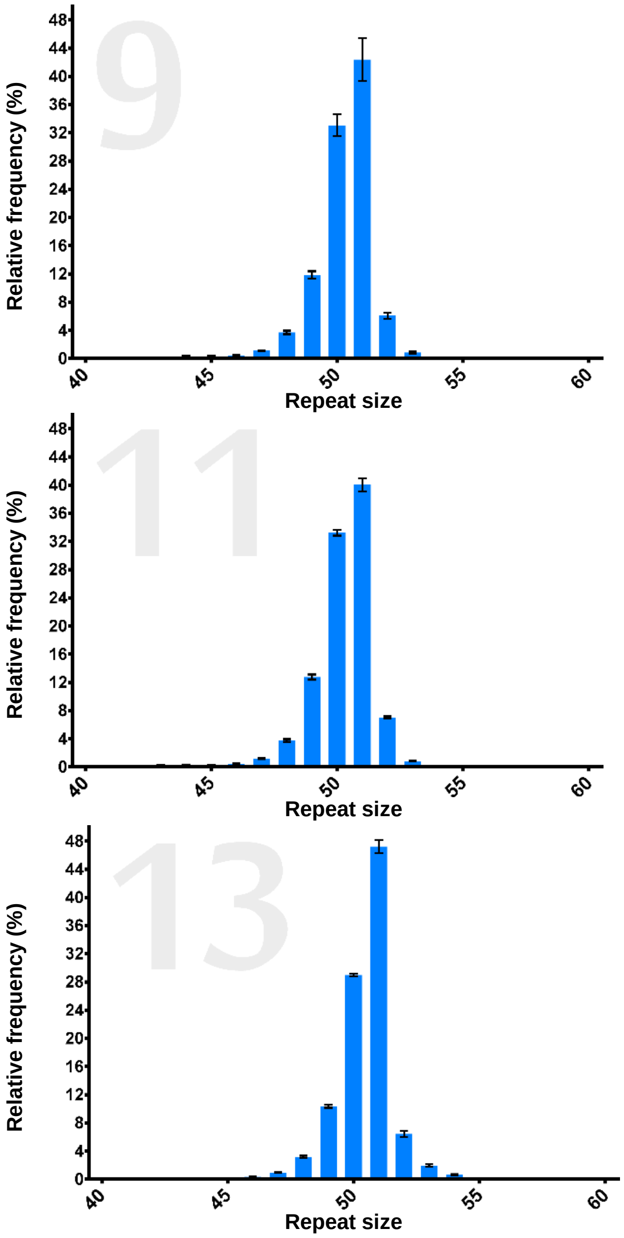
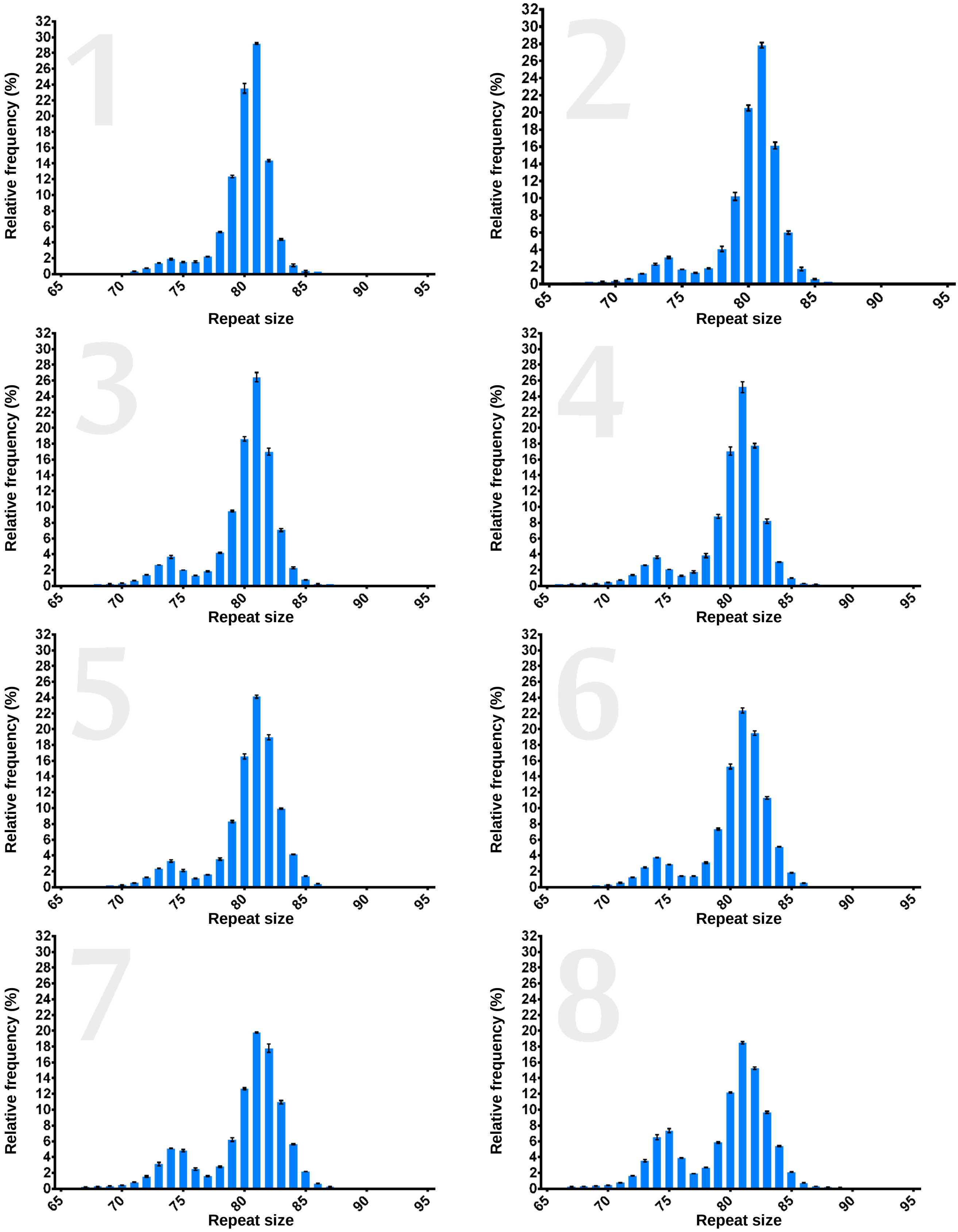


Figure S1 – Bar-plot of detailed relative frequency (%) of each allele size per passage in individual's I-1's (main allele 51) LCLs. Passage number is depicted on the top left quadrant. Column height represents the median relative frequency, error bars represent standard deviation.


1
2
3
4
5
6
7
8
Relative frequency (%)
Repeat size
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
65 70 75 80 85 90 95

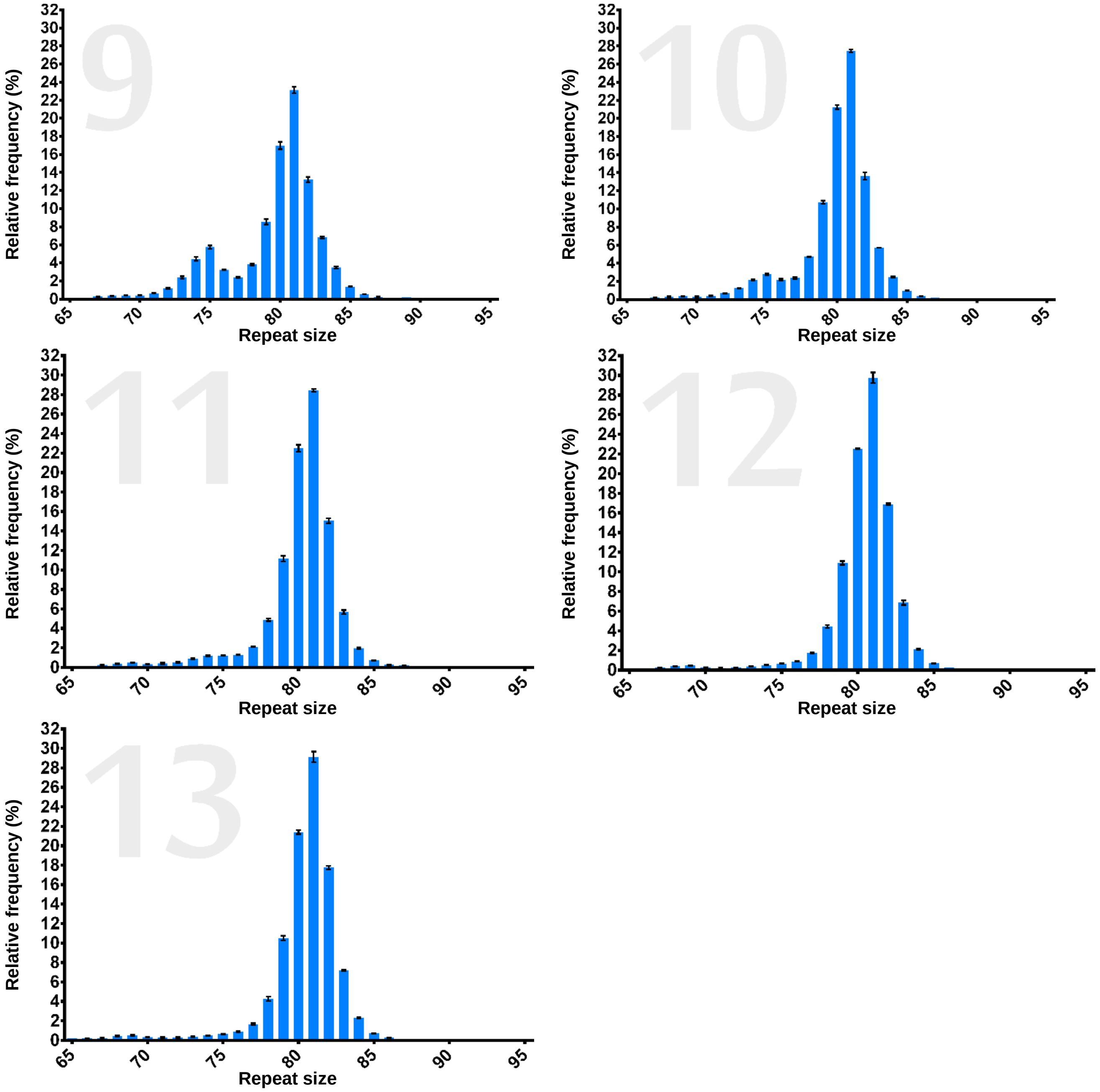
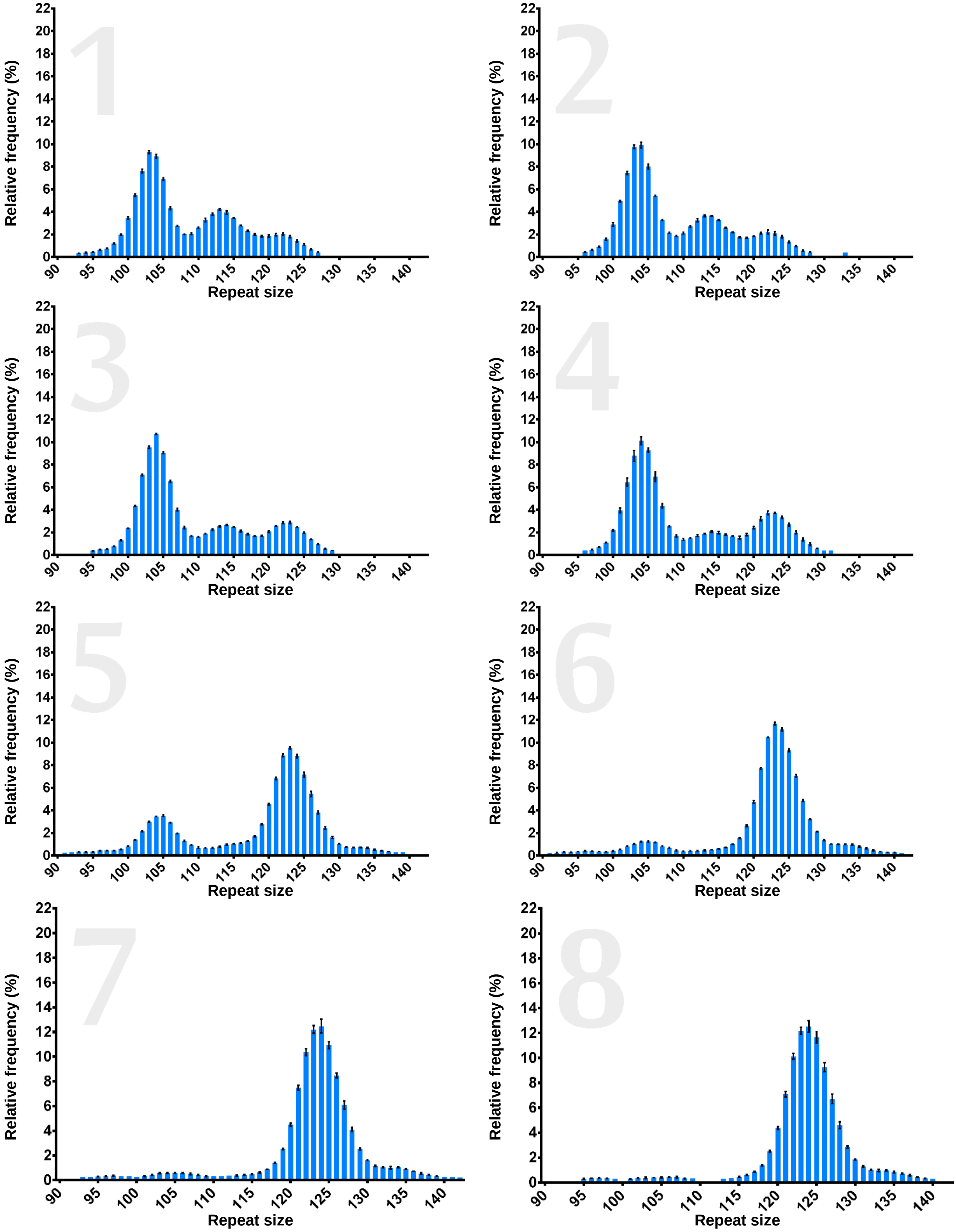


Figure S2 – Bar-plot of detailed relative frequency (%) of each allele size per passage in individual's II-3's (main allele 51) LCLs. Passage number is depicted on the top left quadrant. Column height represents the median relative frequency, error bars represent standard deviation.


1
2
3
4
5
6
7
8
Relative frequency (%)
Repeat size
0
2
4
6
8
10
12
14
16
18
20
22
90
95
100
105
110
115
120
125
130
135
140

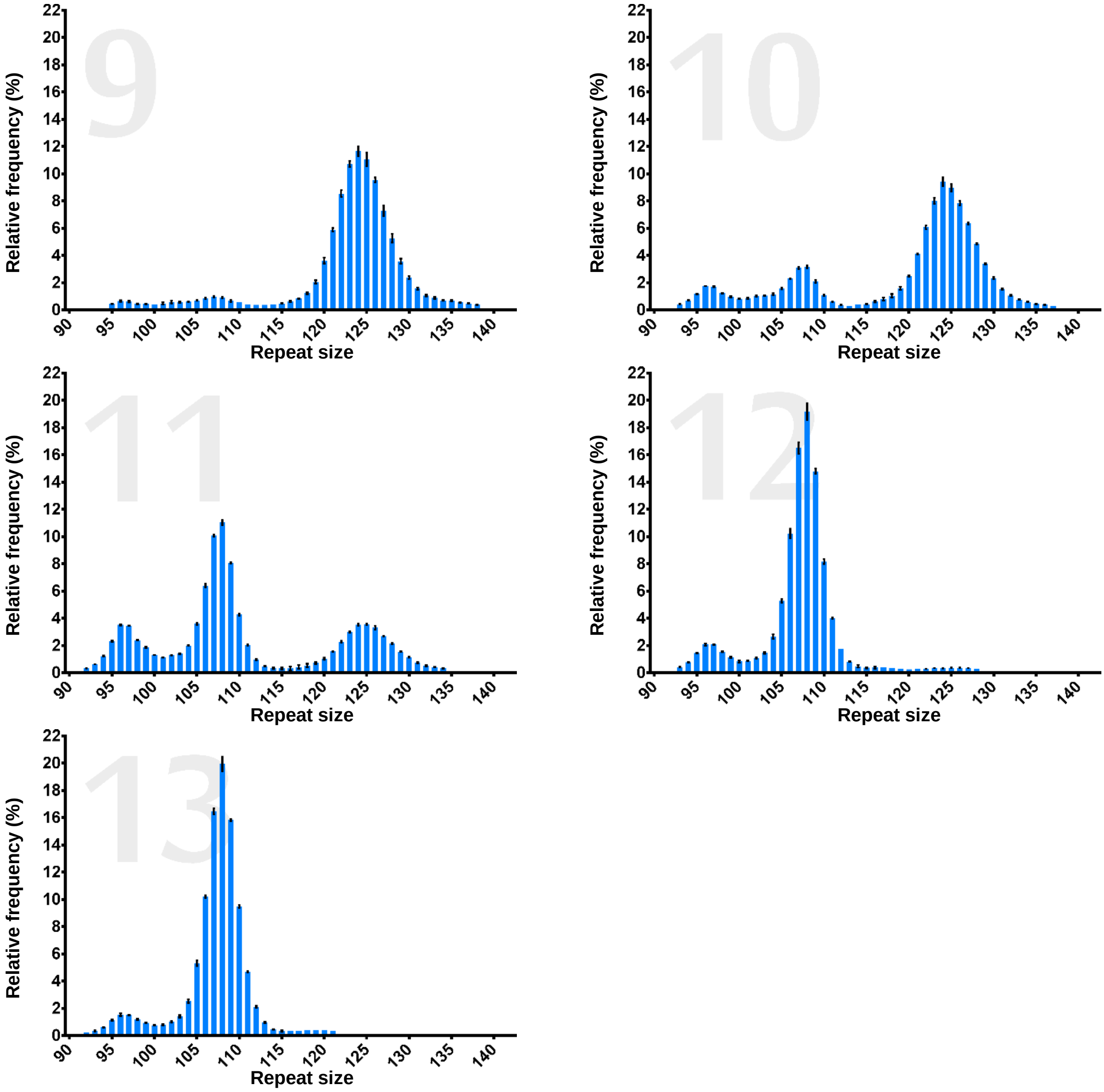
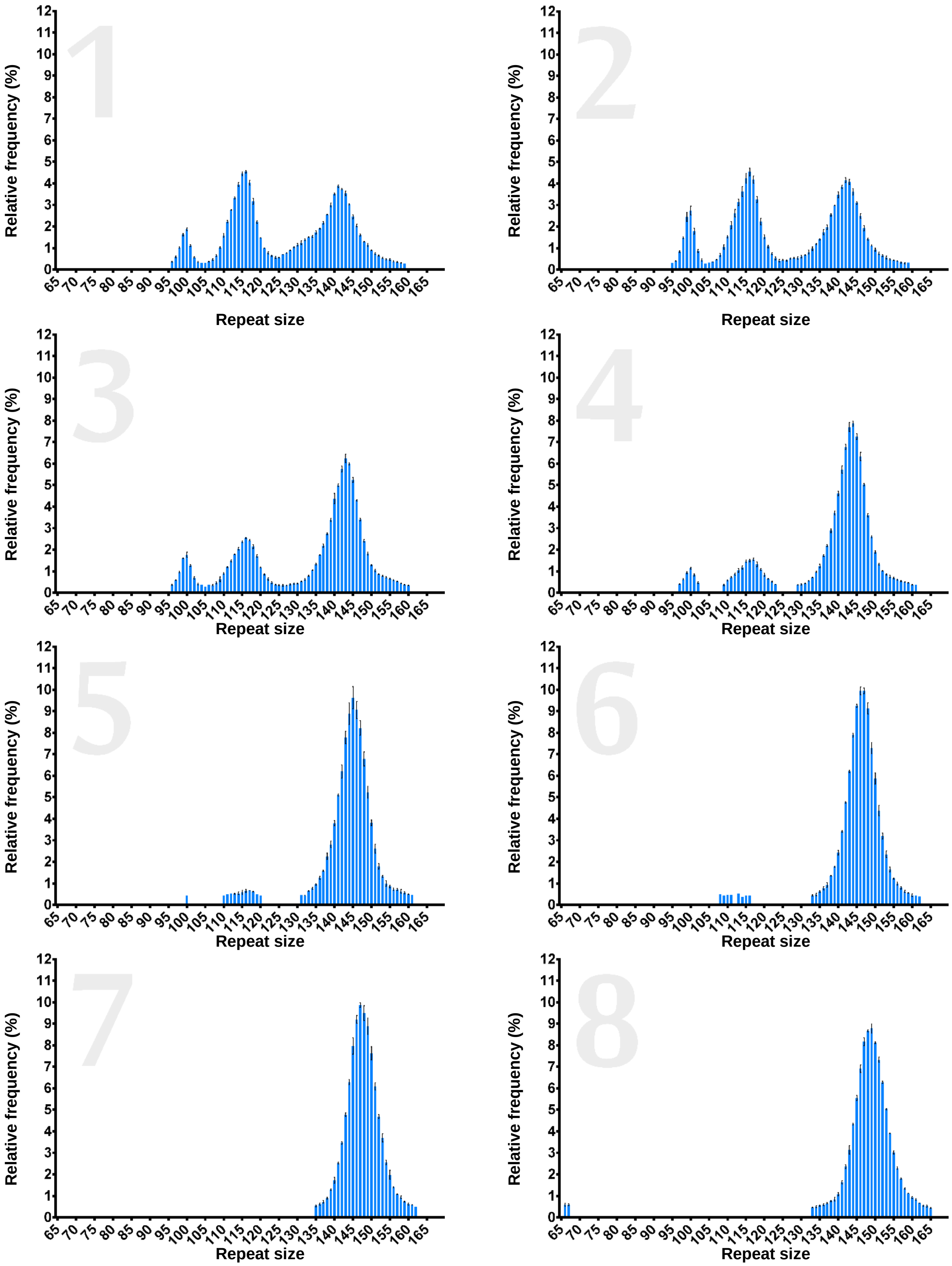


Figure S3 – Bar-plot of detailed relative frequency (%) of each allele size per passage in individual's II-1's (highest alleles in P1: ~103, ~112, ~122 CAGs) LCLs. Passage number is depicted on the top left quadrant. Column height represents the median relative frequency, error bars represent standard deviation.


1
2
3
4
5
6
7
8
Relative frequency (%)
Repeat size
0 1 2 3 4 5 6 7 8 9 10 11 12
65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165

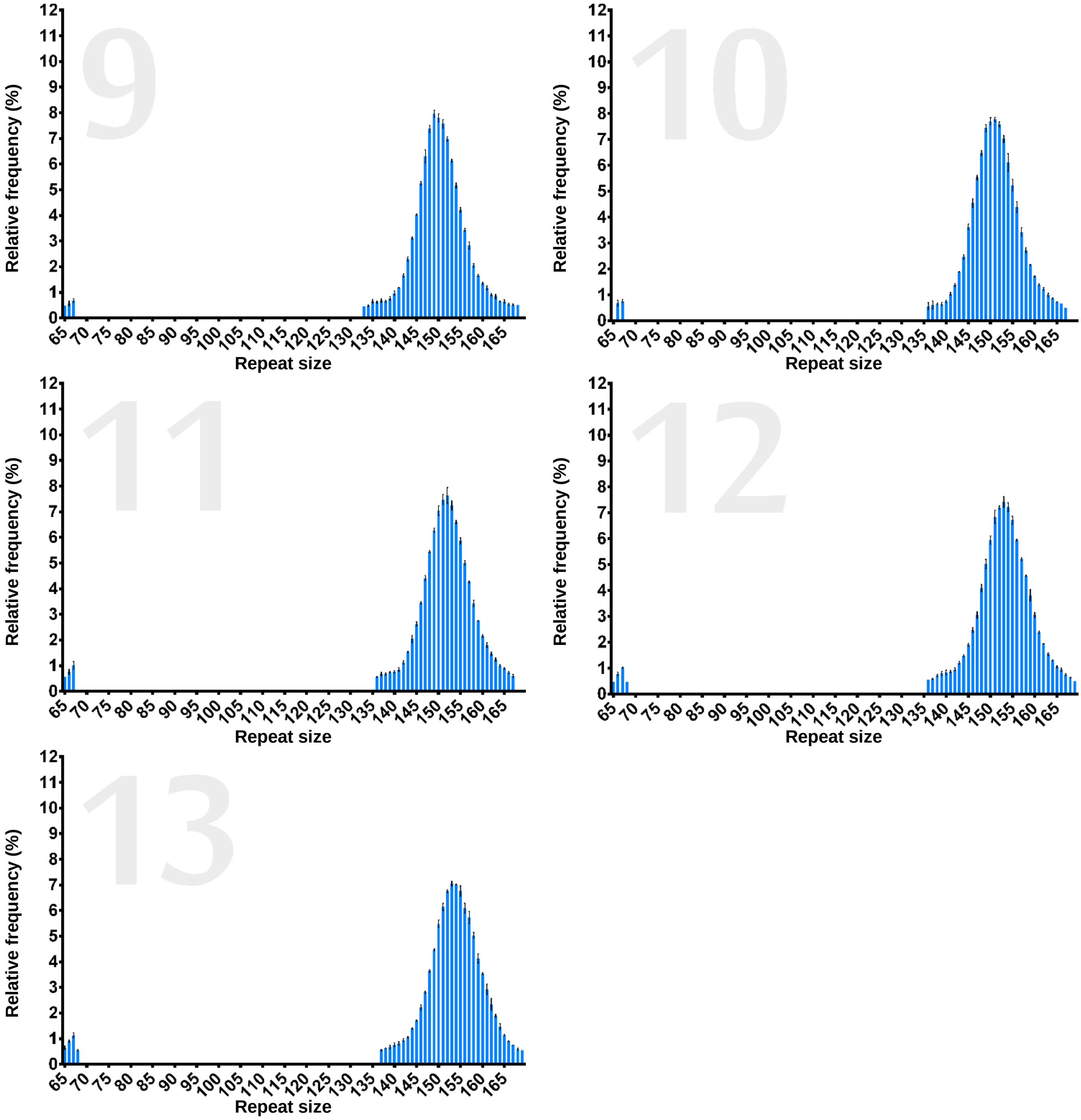


Figure S4 – Bar-plot of detailed relative frequency (%) of each allele size per passage in individual's II-2's (highest alleles in P1: ~100, ~116, ~141 CAGs) LCLs. Passage number is depicted on the top left quadrant. Column height represents the median relative frequency, error bars represent standard deviation.

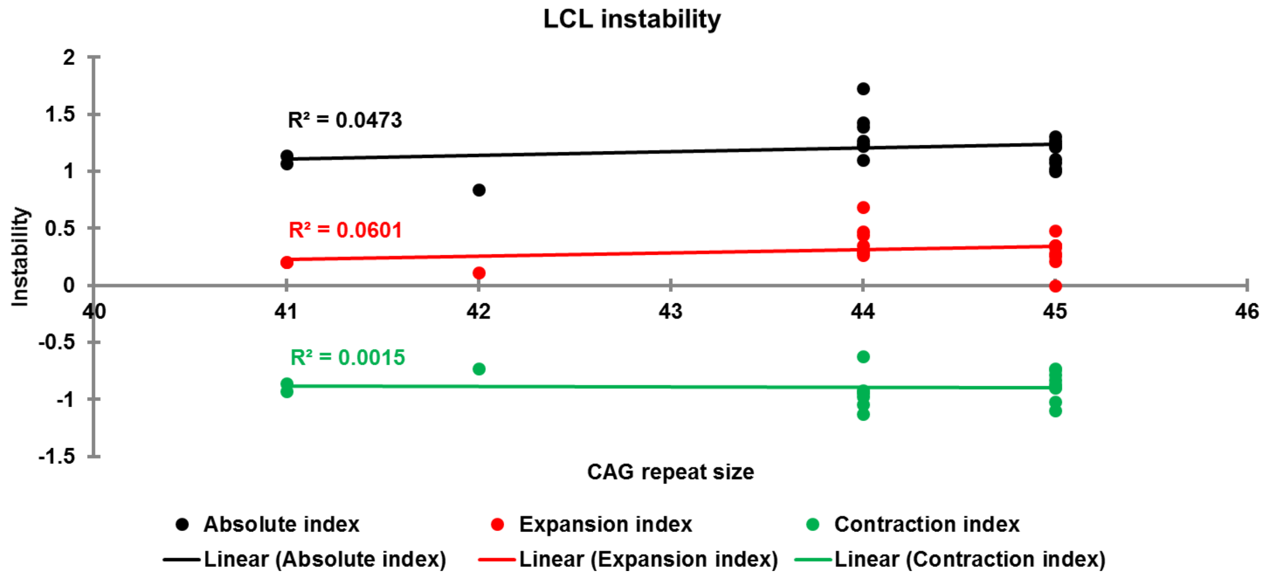


Figure S5 – LCL instability measures (absolute instability, expansion and contraction indexes) by CAG repeat size in the 19 individuals in study.

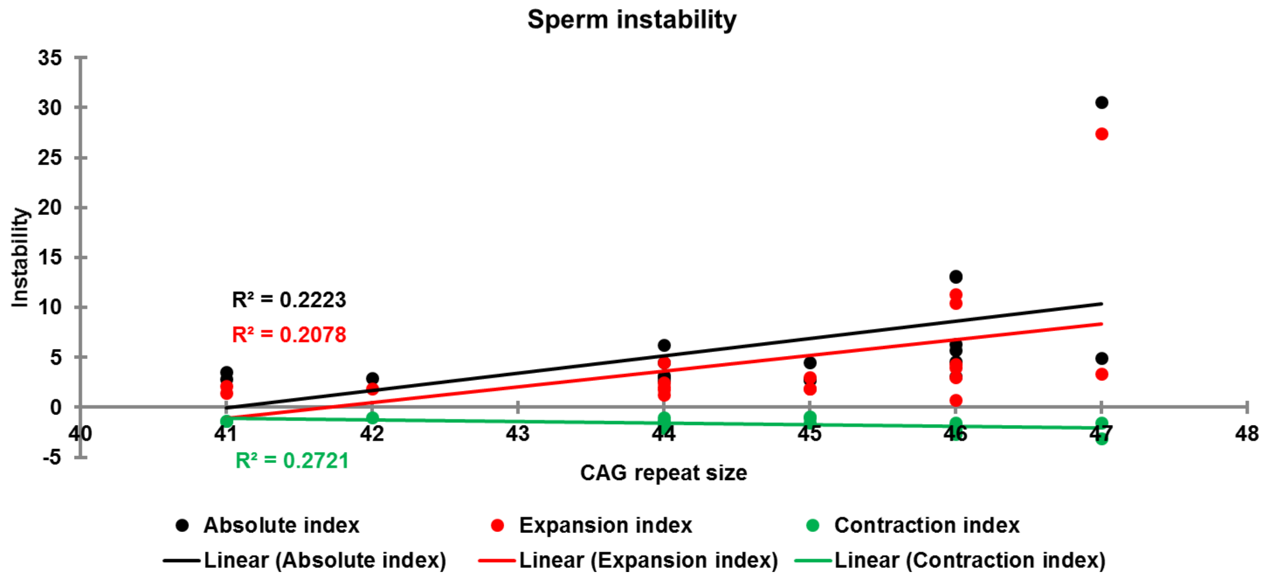


Figure S6 – Sperm instability measures (absolute instability, expansion and contraction indexes) by CAG repeat size in the 19 individuals in study.

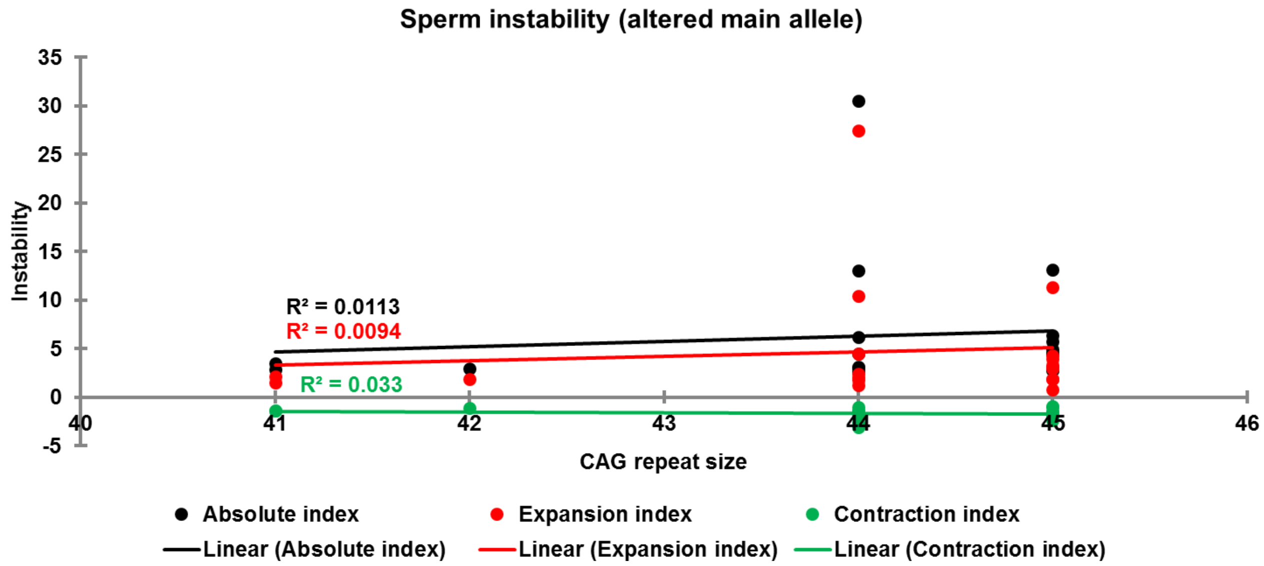


Figure S7 – Sperm instability measures (absolute instability, expansion and contraction indexes) by corrected CAG repeat size (taken from the matching LCL) in the 19 individuals in study.

2.5.2. Supplementary tables

Table S1 – Genes tested for correlation of expression with instability measures

Genes tested					
ABL1	*HMGB1*	*MSH2*	*POLD4*	*RFC1*	*SWSAP1*
APITD1	*HMGB2*	*MSH3*	*POLG2*	*RFC2*	*SYCE3*
ATM	*HNRNPA1*	*MSH4*	*POLL*	*RFC3*	*SYCP1*
ATR	*HNRNPA2B1*	*MSH5*	*POLM*	*RFC4*	*TDG*
BLM	*HNRNPDL*	*MSH6*	*POT1*	*RFC5*	*TDP1*
BRCA2	*HNRNPK*	*MTMR10*	*PURA*	*RNF212*	*TDP2*
C10orf2	*HSPD1*	*MUTYH*	*PURB*	*RPA1*	*TEX11*
CNBP	*IGHMBP2*	*NABP1*	*RAD21*	*RPA2*	*TOP2A*
CRY2	*KLHDC3*	*NABP2*	*RAD23A*	*RPA3*	*TOP2B*
CTC1	*LIG1*	*NEIL3*	*RAD23B*	*RPA4*	*TP73*
DMC1	*LIG3*	*NME1*	*RAD50*	*RTF1*	*TREX1*
ERCC1	*LONP1*	*NTHL1*	*RAD51*	*SPEN*	*TRIP13*
ERCC2	*LRPPRC*	*OBFC1*	*RAD51AP1*	*SPO11*	*UNG*
ERCC3	*MCM4*	*PCBP1*	*RAD51B*	*SSBP1*	*WBP11*
ERCC4	*MCM6*	*PCNA*	*RAD51C*	*SSBP2*	*WRN*
ERCC5	*MCM7*	*PMS1*	*RAD51D*	*SSBP3*	*XPC*
EXO1	*MDM2*	*PMS2*	*RAD52*	*SSBP4*	*YBX1*
FAN1	*MLH1*	*POLD1*	*RAD54B*	*STRA13*	*YBX3*
FANCM	*MLH3*	*POLD2*	*RBMS1*	*STRA8*	
FUBP1	*MRE11A*	*POLD3*	*REC8*	*SUB1*	

Table S2 – Genes whose expression show correlation with CAG repeat size.

CAG repeat size		
Gene	**Correlation**	***p***
MTMR10	0.701	<0.0004
RFC3	0.639	0.001
PURA	0.604	0.002
POLD4	-0.573	0.003
HMGB1	0.570	0.004
HMGB2	0.565	0.004
RAD21	0.557	0.005
MRE11A	0.508	0.011
SUB1	0.488	0.015
TOP2A	0.486	0.016
MUTYH	-0.482	0.017
POLD3	0.466	0.022
SSBP2	0.463	0.023
HNRNPA2B1	0.456	0.025
RTF1	0.449	0.028
RPA4	0.449	0.028
TOP2B	0.448	0.028
BLM	0.440	0.032
ERCC1	-0.438	0.032
TDG	0.436	0.033
CNBP	0.429	0.036
C10orf2	0.428	0.037
HNRNPK	0.426	0.038
ERCC4	0.421	0.040
SSBP3	0.421	0.041
RAD51C	0.411	0.046
MSH6	0.409	0.047
ERCC2	-0.408	0.048
RAD51	0.407	0.048
PURB	0.406	0.049

Table S3 – Genes whose expression show correlation with absolute instability, expansion and contraction index values.

Absolute instability			Expansion index			Contraction index		
Gene	**Correlation**	***p***	**Gene**	**Correlation**	***p***	**Gene**	**Correlation**	***p***
MTMR10	0.649	0.001	*MTMR10*	0.582	0.003	*MTMR10*	-0.673	<0.0004
PURA	0.585	0.003	*SSBP3*	0.578	0.003	*PURA*	-0.624	0.001
RFC3	0.565	0.004	*RFC3*	0.522	0.009	*RAD21*	-0.582	0.003
SSBP3	0.544	0.006	*PURA*	0.514	0.010	*RFC3*	-0.562	0.004
MRE11A	0.474	0.019	*RPA4*	0.498	0.013	*TOP2B*	-0.557	0.005
RPA4	0.472	0.020	*RTF1*	0.448	0.028	*ERCC4*	-0.542	0.006
TOP2B	0.468	0.021	*HMGB2*	0.444	0.030	*MRE11A*	-0.542	0.006
RAD21	0.463	0.023	*C10orf2*	0.420	0.041	*MUTYH*	0.537	0.007
HMGB2	0.462	0.023				*ERCC1*	0.531	0.008
RTF1	0.459	0.024				*SWSAP1*	0.530	0.008
TDG	0.458	0.024				*TDP2*	-0.512	0.011
HMGB1	0.440	0.031				*POLD4*	0.511	0.011
POLD4	-0.436	0.033				*ERCC2*	0.500	0.013
ERCC1	-0.418	0.042				*TDG*	-0.499	0.013
MUTYH	-0.411	0.046				*HMGB1*	-0.498	0.013
						POLL	0.483	0.017
						POT1	-0.478	0.018
						TOP2A	-0.463	0.023
						HNRNPK	-0.457	0.025
						MSH3	-0.454	0.026
						SSBP2	-0.454	0.026
						CNBP	-0.452	0.027
						HNRNPA2B1	-0.451	0.027
						SUB1	-0.432	0.035
						HMGB2	-0.432	0.035
						RAD51C	-0.421	0.040
						RTF1	-0.417	0.043
						SSBP3	-0.416	0.043

Table S4 – Genes whose expression show correlation with absolute instability, and contraction index residuals considering only individuals with over 35 CAGs.

Absolute instability residual (>35 CAGs)			Contraction index residual (>35 CAGs)		
Gene	**Correlation**	***p***	**Gene**	**Correlation**	***p***
TP73	-0.507	0.038	*NTHL1*	0.815	<0.0004
			SWSAP1	0.761	<0.0004
			POLD1	0.745	0.001
			TDP2	-0.698	0.002
			POLL	0.690	0.002
			ERCC1	0.617	0.008
			FAN1	-0.613	0.009
			MDM2	-0.612	0.009
			RAD23A	0.607	0.010
			LONP1	0.605	0.010
			MCM7	0.593	0.012
			IGHMBP2	0.577	0.015
			ERCC4	-0.571	0.017
			LIG1	0.538	0.026
			SSBP4	0.534	0.027
			NABP1	-0.522	0.032
			TDG	-0.520	0.033
			SPEN	0.511	0.036
			STRA13	0.495	0.044
			DMC1	0.490	0.046

3. Genetic contributors to intergenerational CAG repeat instability in Huntington's disease knock-in mice

This chapter has been published as:

João Luís Neto, Jong-Min Lee, Ali Afridi, Tammy Gillis, Jolene R. Guide, Stephani Dempsey, Brenda Lager, Isabel Alonso, Vanessa C. Wheeler and Ricardo Mouro Pinto. *Genetics* February 1, 2017 vol. 205 no. 2 503-51

3.1. Introduction

Huntington's disease (HD) is a progressive, degenerative, autosomal dominant disorder caused by the expansion of a CAG repeat located in exon 1 of the *HTT* gene (4p16.3), producing an extended polyglutamine tract in the huntingtin protein[19]. Alleles over 35 repeats are associated with disease, with 36-39 CAGs showing reduced penetrance and 40 or more repeats being fully penetrant[19,111]. The length of the expanded repeat is also a major modifier of the age of disease onset[33,36,52,55,61,62]. Underlying the variation in inherited CAG repeat length between individuals are high rates (~70-80%) of intergenerationally unstable transmissions[33,56,59,60,69]. Intergenerational instability of the *HTT* CAG repeat accounts for genetic anticipation seen in HD families[55,56], as well as for changes from high normal alleles (27-35 CAGs) to disease alleles (*i.e. de novo* mutations) and transitions from alleles associated with incomplete penetrance to those causing completely penetrant disease[37,57,58].

Mechanisms underlying intergenerational instability are unclear but important to understand in order to refine variable estimates of new mutation rates for genetic counseling[57,58,112], and for the potential to suppress expansions and/or induce contractions. Studies of intergenerational instability in HD families have shown that the *HTT* CAG repeat is strongly biased towards expansions in transmission from fathers, while transmissions from mothers have a higher tendency to be stable or contract[33,55,56,59–64]. In addition to parent sex, CAG repeat length strongly influences intergenerational repeat instability, with longer repeats being more susceptible to larger changes[33,56,60,63,64]. Minor effects of offspring sex, size of the normal CAG repeat and parental age have also been documented in some but not all HD cohorts examined[56,60,63,64,68].

Clustering of transmitted repeat length changes among HD families implicates genetic modifiers of intergenerational instability[60,64]. Familial segregation of instability in a large HD Venezuelan pedigree that shares a 4p16.3 (*HTT*) haplotype provides evidence for other genes that can modify instability (*trans*-acting)[60]. A 4p16.3 predisposing haplotype (*cis*-modifier) has been proposed to underlie the expansion of high normal length *HTT* CAG repeats into the disease range[66].

However, this haplotype was not associated with the length of the expanded CAG repeat[50] or with its intergenerational instability[67].

Transgenic and knock-in mouse models of HD recapitulate many aspects of intergenerational CAG repeat instability seen in patients[78,81,88,89,91], notably: CAG repeat length-dependent instability at similar high frequency (>65%) for long alleles (>~100 CAGs)[78] as well as paternal expansion and maternal contraction biases[78,88,89]. Offspring sex was also found to contribute to instability in *HTT* exon 1 (R6/1) transgenic mice[89]. No obvious role of parental age was discernible in knock-in models[78]. *Cis*-modifiers of intergenerational instability were suggested based on differential instability in *HTT* exon 1 transgenic (R6) models with similar CAG repeat sizes but distinct transgene insertion sites[81]. DNA repair genes *Msh2*, *Msh3*, *Msh6* and *Neil1* have been identified as *trans*-acting modifiers of intergenerational repeat instability[90,91,96], and a comparison of Htt^{Q111} knock-in lines on different genetic backgrounds indicated the presence of *trans*-factors that drive strain-specific intergenerational instability[91].

Here, we have taken advantage of two large breeding datasets comprising thousands of transmissions from allelic series of *Htt* CAG knock-in mice that differ in CAG length, genetic background and the presence of a *cis*-element – a neomycin resistance cassette (*neo*) – upstream of the CAG repeat, to perform a comprehensive assessment of the factors that drive intergenerational instability of expanded CAG repeat at the mouse *Htt* locus. These analyses confirm major modifiers of instability, distinguish more subtle effects and provide novel insight into potential *trans* and *cis*-mediated effects.

3.2. Methods

3.2.1. Mouse lines

Breeding data were obtained from The Jackson Laboratory (JAX) for the following lines of *Htt* (formerly *Hdh*) CAG knock-in mice on a C57BL6/J (B6J) background: *Htt*^{Q20}, *Htt*^{Q50}, *Htt*^{Q80}, *Htt*^{Q92}, *Htt*^{Q111}, *Htt*^{Q140} and *Htt*^{Q175}. The *Htt*^{Q20}, *Htt*^{Q50}, *Htt*^{Q80}, *Htt*^{Q92} and *Htt*^{Q111} lines were originally derived in the MacDonald laboratory[77,78], with *Htt*^{Q80} being a derivative of the original *Htt*^{Q92} line obtained by selective breeding to smaller CAG repeats. The *Htt*^{Q140} line was originally derived in the Zeitlin laboratory[79], with *Htt*^{Q175} being a derivative of *Htt*^{Q140}, obtained by selective breeding to obtain longer CAG repeats[80]. *Htt*^{Q20}, *Htt*^{Q50}, *Htt*^{Q80}, *Htt*^{Q92} and *Htt*^{Q111} lines represented in the JAX dataset do not contain the upstream *neo* cassette used in targeting these alleles, however, the *Htt*^{Q140} and *Htt*^{Q175} lines retain a *neo* cassette (Figure S8). Together, these lines form part of an “allelic series” of *Htt* CAG knock-in mice for analyses of repeat length-dependent phenotypes[82,113]. Subsequent excision of the *neo* cassette using Cre-mediated recombination resulted in a new line of B6J *Htt*$^{Q175neo-}$ mice (B6.129S1-Htt<tm1.1Mfc>/190ChdiJ). Here, for simplicity we refer to the *Htt*$^{Q175neo+}$ mice that form part of the allelic series simply as *Htt*^{Q175}, unless we specifically need to distinguish them from their *neo*- counterpart, in which case we refer to the lines as *Htt*$^{Q175neo+}$ and *Htt*$^{Q175neo-}$.

Separately, in-house (CHGR) breeding data were obtained for the *Htt*^{Q111} line on six background strains: CD1, B6J, C57BL/6NCrl (B6N), 129S2/SvPasCrlf (129), FVB/NCrl (FVB) and DBA/2J (DBA); as well as for *Htt*^{Q80} and *Htt*^{Q92} on CD1 and B6J backgrounds[78,91,102]. Data from CD1 or B6J lines (*Htt*^{Q80}, *Htt*^{Q92}*, and Htt*^{Q111}) were combined to provide a broad CAG repeat range for each of these background strains. None of the lines used for comparisons of instability across different genetic backgrounds had a *neo* cassette. We also analyzed breeding data from CD1.*Htt*^{Q111} mice that included an upstream *neo* cassette[114], which we compared against the set of CD1 mice that were missing the *neo* cassette. In this comparison, we refer to these mice as CD1^{neo+} and CD1^{neo-} respectively.

3.2.2. Mouse breeding, husbandry and genotyping

This study was carried out in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals, NRC (2010). All animal procedures were carried out to minimize pain and discomfort, under approved IACUC protocols of the Massachusetts General Hospital or The Jackson Laboratory.

Data collected from JAX-maintained *Htt* CAG knock-in lines followed breeding and husbandry conditions described in [113], with genotyping and CAG length determination performed in tail DNA at weaning by Laragen Inc. CHGR breeding was performed as described in previous work[102,104], and genotyping and CAG length determination were performed in tail DNA at weaning as previously described[81,104].

3.2.3. Intergenerational transmission data

Both JAX and CHGR breeding data records were quality controlled to eliminate entries with obvious and systematic errors (*e.g.* typographical), or from crosses with inconclusive assignment of parental CAG repeat length (*e.g.* crosses between two heterozygous parents and harem breedings). In CHGR's strain-specific data, only mice that were at least sixth-generation backcross progeny (F6, >98% congenic) were included, except for B6J.Htt^{Q111}, where speed congenics were utilized to generate the line[97] and generations F4 (~95% congenic) and after were included.

Following quality control, the JAX dataset comprised 44,378 pups from crosses between a heterozygous knock-in and a wild-type parent. Of these, 22,063 pups carried a mutant allele, allowing us to determine CAG repeat length change upon transmission. In the JAX dataset, accurate parental age at which the pups were born could not be determined. After quality control, the CHGR dataset comprised 1,981 pups carrying a mutant allele from crosses between heterozygous knock-in sires and wild-type dams. In these data we were able to assign parental age at the time of birth of the pups unambiguously.

Repeat length change was determined by subtracting the CAG repeat length in the heterozygous knock-in progeny from the CAG repeat length in the respective heterozygous knock-in parent. It should be noted that for both the JAX and CHGR data, breeding records were accumulated over periods of months/years, and that parent and progeny genotyping were performed at separate times. While standardized CAG repeat genotyping assays are used in both cases, there may be some small degree of error in the determination of repeat length change, however, this is likely to have a negligible impact given the large size of the datasets used in this study.

3.2.4. Frequency modeling

To control for possible confounding effects of parental CAG size in the frequency of unstable transmissions between different strains/lines, we implemented a modeling methodology that allowed a pairwise comparison of the rates of expansions, contractions and unchanged transmissions between test strains and a reference strain.

This modeling procedure is represented in Appendix. In essence, 1) weighted linear regressions for relative frequencies of expansions, contractions and unchanged transmissions versus parental CAG size were determined for the reference strain using PASW Statistics 18; 2) a modeled dataset was generated through random number generation and allocation as expansion, contraction or stable transmission based on the frequency intervals of these events (per paternal CAG size) established from the weighted linear regression lines; 3) this was repeated 1,000 times, the modeled datasets were averaged, and a dataset based on the average values was used for comparative and statistical analyses. To validate this methodology, we randomly divided the CHGR B6J set into two subsets with comparable number of transmissions: reference and test subsets (n=354 and n=353, respectively). Based on the reference dataset, we modeled the frequencies for the test subset, and after comparison against the observed values we confirmed that the expected frequencies of the test dataset could be predicted with this process (Figure S9). We then applied this methodology to compare transmission frequencies: 1) in CD1, B6N, 129, DBA and FVB strains against B6J (reference strain); and 2) in *neo*+ versus *neo*- mice, where either CD1^{neo-} or Q175^{neo+} were the reference strains (Appendix).

3.2.5. Statistical analyses

Frequency analyses were performed using Microsoft Excel 2007 and PASW Statistics 18 (IBM). χ^2 tests of independence, unpaired Student's t-test for mean comparison, Pearson correlation analyses, and z-tests for column proportions were carried out using PASW Statistics 18 (IBM) – actual z-scores and p-values were determined using Microsoft Excel 2007 or the on-line Z-Score calculator for 2 population proportions (http://www.socscistatistics.com/tests/ztest/Default2.aspx). To determine the effect of strain background on the magnitude of repeat length changes mixed effect model analyses were performed using "nlme" packages in R program (v3.2.2). Briefly, CAG repeat size changes (expansions or contractions) were modeled as a function of paternal CAG and mouse strain as main effects, with random intercepts for sire. Kruskal-Wallis one-way analysis of variance and Dunn's post-test were performed with GraphPad Prism 6. Bonferroni correction was applied when multiple testing was performed (p-value thresholds stated in text/figure legends). Weighted regression lines for number of transmissions per CAG size were calculated with PASW Statistics 18.

3.2.6. Data availability

Datasets can be provided by the authors upon request.

3.3. Results

To gain insight into factors that influence the intergenerational instability of the *Htt* CAG repeat, we first analyzed a very large breeding dataset from JAX comprising >44,000 offspring from heterozygous parents of an allelic series of HD knock-in mice – Htt^{Q20}, Htt^{Q50}, Htt^{Q80}, Htt^{Q92}, Htt^{Q111}, Htt^{Q140} and Htt^{Q175} – on a B6J genetic background (Table 3). The vast majority of transmissions were from heterozygous knock-in sires, with the Htt^{Q80} and Htt^{Q92} lines also harboring transmissions from mutant dams.

Table 3 – Heterozygous and wild-type progeny from *Htt* knock-in lines in the JAX dataset.

Line	Paternal transmissions			Maternal transmissions		
	Total	Heterozygous progeny	Wild-type progeny	Total	Heterozygous progeny	Wild-type progeny
	N	N (%)	N (%)	N	N (%)	N (%)
Htt^{Q20}	7189	3595 (50.0)	3594 (50.0)	-	-	-
Htt^{Q50}	1025	511 (49.9)	514 (50.1)	-	-	-
Htt^{Q80}	4992	2465 (49.4)	2527 (50.6)	263	129 (49.0)	134 (51.0)
Htt^{Q92}	5970	2883 (48.3)	3087 (51.7)	412	200 (48.5)	212 (51.5)
Htt^{Q111}	3098	1550 (50.0)	1548 (50.0)	-	-	-
Htt^{Q140}	3109	1558 (50.1)	1551 (49.9)	-	-	-
Htt^{Q175}	18288	9172 (50.2)	9116 (49.8)	-	-	-

N, number of transmissions. Htt^{Q20}, Htt^{Q50}, Htt^{Q80}, Htt^{Q92}, Htt^{Q111}: *neo*-. Htt^{Q140}, Htt^{Q175}: *neo*+

3.3.1. Segregation of *Htt* CAG knock-in alleles studied follows Mendelian ratios and is independent of CAG length

A 1:1 Mendelian ratio of heterozygous versus wild-type progeny was broadly observed among the lines, with the exception of paternal transmissions in Htt^{Q92} where a significant difference (two-proportion z-test, $p<0.001$) of fairly small effect (1.7% lower than expected frequency of heterozygotes) was observed. As this effect was unique to this line and not seen in lines with larger repeat sizes, overall, these results indicate the lack of an obvious effect of the CAG mutation on the transmission of the

Htt allele over a large range of repeat lengths among a very high number of transmissions.

3.3.2. Parental sex influences the direction of repeat length changes but does not have a major impact on magnitude

In human *HTT* mutation carriers, parental sex is a major determinant of intergenerational repeat instability[33,55,56,60–64]. We therefore compared frequencies of stable and unstable transmissions (expansions and contractions), as well as the magnitude of these alterations between paternal and maternal transmissions of the Htt^{Q80} and Htt^{Q92} alleles (Table 4 and Table 5).

Table 4 – Paternal CAG size and transmission information for the different lines in the JAX breeding data

	Paternal transmissions										
	Parent CAG			Total	Stable	Contractions			Expansions		
Line	**Min.**	**Max.**	**Mean**	**N**	**N (%)**	**N (%)**	**Max.**	**Mean**	**N (%)**	**Max.**	**Mean**
Htt^{Q20}	18	18	18	3595	3591 (99.89)	1 (0.03)	1	1	3 (0.08)	1	1
Htt^{Q50}	46	48	46.6	511	460 (90.0)	29 (5.7)	2	1.1	22 (4.3)	1	1.0
Htt^{Q80}	80	86	83.1	2465	921 (37.4)	330 (13.4)	15	1.2	1214 (49.2)	29	1.7
Htt^{Q92}	91	98	95.8	2883	1010 (35.0)	268 (9.3)	12	1.3	1605 (55.7)	50	1.9
Htt^{Q111}	106	121	113.9	1550	300 (19.5)	248 (16.0)	28	2.1	983 (64.5)	14	2.3
Htt^{Q140}	130	150	137.4	1558	232 (14.9)	172 (11.0)	18	2.2	1154 (74.1)	65	3.2
Htt^{Q175}	178	199	188.9	9172	592 (6.5)	927 (10.1)	128	5.1	7653 (83.4)	153	5.2

N, number. Htt^{Q20}, Htt^{Q50}, Htt^{Q80}, Htt^{Q92}, Htt^{Q111}: *neo*-. Htt^{Q140}, Htt^{Q175}: *neo+*

Table 5 – Maternal CAG size and transmission information for the different lines in JAX's breeding data.

	Maternal transmissions										
	Parent CAG			Total	Stable	Contractions			Expansions		
Line	**Min.**	**Max.**	**Mean**	**N**	**N (%)**	**N (%)**	**Max.**	**Mean**	**N (%)**	**Max.**	**Mean**
Htt^{Q80}	83	84	83.2	129	59 (45.7)	59 (45.7)	5	1.4	11 (8.5)	3	1.3
Htt^{Q92}	94	98	96.8	200	70 (35.0)	82 (41.0)	5	1.6	48 (24.0)	20	2.2

N, number. Htt^{Q80}, Htt^{Q92}: *neo*-.

For both lines, parent-of-origin determined the frequency distribution of repeat length changes (*Htt*^{Q80}: χ^2=130.86, 2 df, p<0.001; *Htt*^{Q92}: χ^2=200.58, 2 df, p<0.001), with paternal transmissions showing a higher occurrence of expansions, and maternal transmissions showing a higher occurrence of contractions (two-proportion z-test, p<0.01) but no significant differences in the frequencies of stable transmissions (Figure 17).

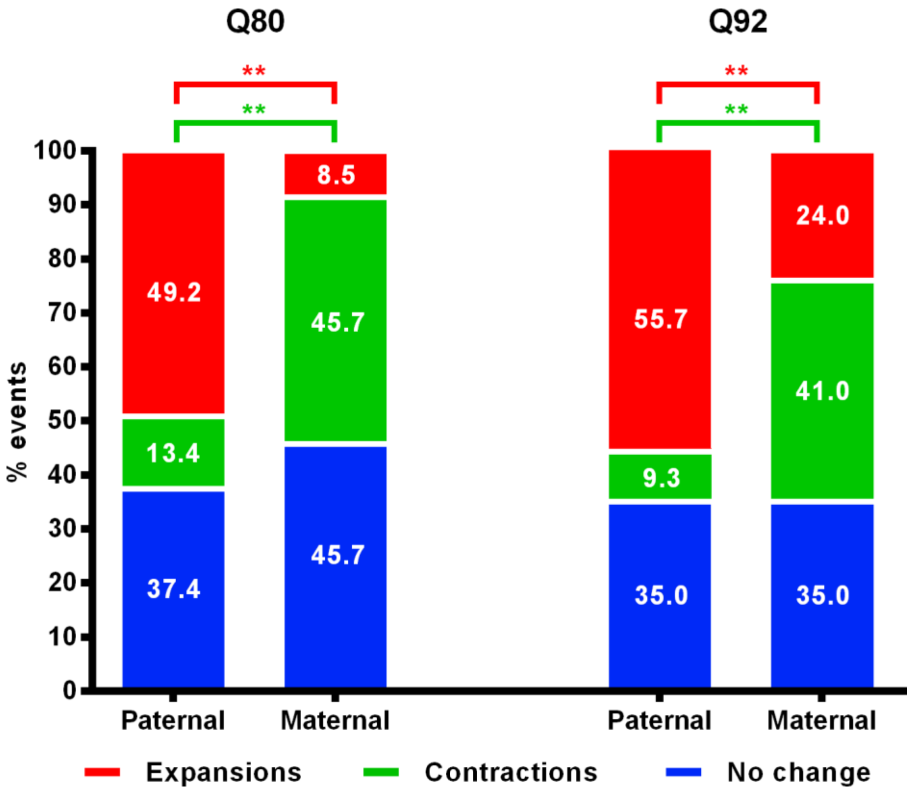


Figure 17 - Frequency of stable transmissions, contractions and expansions in paternal and maternal transmissions of the *Htt*^{Q80} and *Htt*^{Q92} alleles in the JAX transmission data. **p<0.01.

Larger maximum expansions and contractions were observed in paternal transmissions compared to maternal transmissions, probably partly driven by the greater total number of paternal transmissions (Table 4; Table 5). However, parental sex did not significantly alter the mean magnitude of the changes in *Htt*^{Q80} mice (contractions: unpaired t-test, p=0.235; expansions: unpaired t-test, p=0.312) or the mean magnitude of expansions in *Htt*^{Q92} mice (unpaired t-test, p=0.467), though the mean magnitude of contractions was significantly increased in maternal transmissions of the *Htt*^{Q92} line (unpaired t-test, mean difference=0.348 CAGs, p=0.003;Table 4; Table 5; Figure S10). The significance of this is unclear in the absence of additional maternal transmission data from longer alleles, but may indicate a differential sensitivity to repeat length of the mechanisms that mediate contractions in the male and female germline.

Overall, in *Htt* CAG knock-in mice, parent-of-origin mainly influences the relative frequencies of expansions and contractions, with a minor impact on the magnitude of repeat contractions.

3.3.3. Offspring sex does not influence intergenerational instability in *Htt* CAG knock-in mice

Previously, effects of offspring sex on intergenerational CAG repeat instability were identified in human *HTT* mutation carriers and R6/1 mouse models of the disorder[60,89]. We took advantage of this large breeding dataset to determine whether this effect might be recapitulated in transmissions from *Htt* CAG knock-in mice. We examined the frequency and magnitude of repeat length changes inherited by male and female progeny in the expanded *Htt* CAG knock-in lines with available data, analyzing paternal and maternal transmissions separately (Table S5).

For all lines, both in paternal and maternal transmissions, offspring sex did not significantly influence either the relative frequencies of contractions, expansions or stable alleles (χ^2, Table S6, Figure S11), or the magnitude of expansions or contractions (unpaired t-tests, $p>0.05$, Figure S12). Thus, offspring sex is not a major determinant of intergenerational instability in this allelic series of HD knock-in mouse models.

3.3.4. Distinct effects of paternal CAG repeat length on the frequency and magnitude of changes

Intergenerational repeat instability is strongly determined by parental CAG repeat length in *HTT* mutation carriers[56,60,63,64,115]. The wide range of CAG repeat lengths afforded by the *Htt* knock-in allelic series allowed us to perform a comprehensive analysis of the effect of parental CAG repeat length on various measures of intergenerational repeat instability.

Given the parent-of-origin effects described above, the influence of parental CAG size on repeat transmissions was assessed in paternal transmissions only – which constitute the majority of the breeding data available and encompass the widest range of allele sizes (Table 4). The *Htt*^{Q20} line showed ~99.9% stable transmissions. The high stability of this normal CAG length allele is to be expected, with the very rare repeat length changes consistent with occasional unstable normal alleles in humans[67]. Note that *Htt*^{Q20} was not included in subsequent analyses due to the extremely low number of unstable events.

Frequency distributions of expansions, contractions and unchanged alleles for the expanded *Htt* CAG alleles are shown in Figure 18. In *Htt*^{Q50} repeat length was unchanged in the vast majority of transmissions (90%), while the longer *Htt* alleles showed considerable levels of instability. Mouse line significantly predicted the frequency distribution of repeat length change (χ^2=3774.23, 10 df, $p<0.001$) and proportion comparisons revealed that longer repeat lines had significantly higher expansion frequencies and significantly lower frequencies of unchanged alleles when compared to lines of shorter repeat length (two-proportion z-tests, $p<0.003$ – Bonferroni corrected; Figure 18B). For contractions, significant differences between some of the lines were observed but they did not follow any continuous CAG length dependence. To better understand these trends, we determined frequencies of expansions, contractions and unchanged alleles among the more unstable *Htt*^{Q80}, *Htt*^{Q92}, *Htt*^{Q111}, *Htt*^{Q140} and *Htt*^{Q175} lines, taking parental CAG size as a continuous variable and weighting trend lines by the number of transmissions for each CAG length (Figure S13). This confirmed that expansion frequency is positively correlated with CAG repeat length (slope=0.298; R^2=0.728), that the frequency of unchanged alleles is negatively correlated with CAG repeat length (slope=-0.280; R^2=0.809), and highlights a clearer picture of a fairly constant (~10%-12%) contraction frequency throughout the broad CAG range being studied (slope=-0.018 R^2=0.029).

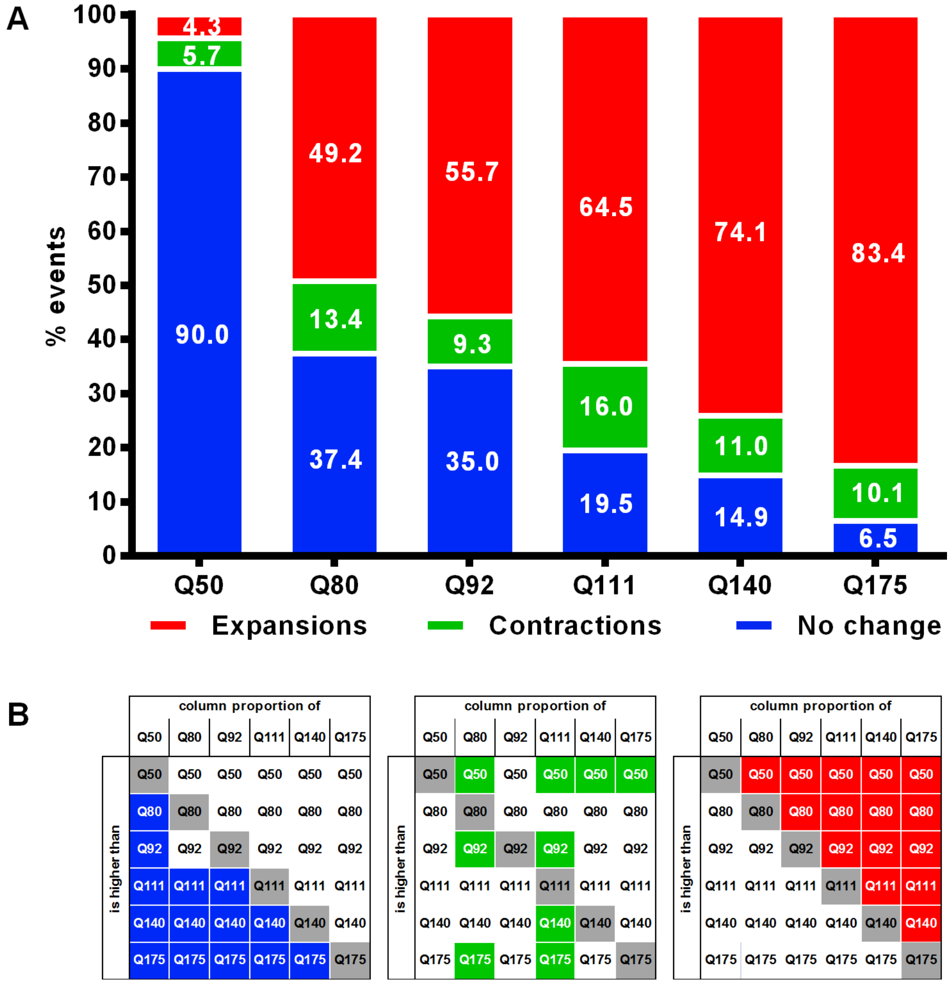


Figure 18 - Relative frequency and significant differences of stable and unstable paternal transmissions in JAX's dataset. (A) Breakdown of transmission frequency by expansions, contractions and stable transmissions. (B) Significant differences ($p<0.003$ threshold after Bonferroni correction) between proportions of the different lines are highlighted for expansions (right), contractions (center) and stable transmissions (left)

We also analyzed the effect of CAG repeat length on the magnitude of expansions and contractions. The large number of transmissions available for analysis allowed us to capture a wide distribution of CAG changes across the different lines (Table 4, Figure 19, Figure 20), undetected in previous analyses of intergenerational instability in mouse models[78,91]. For most lines, very large changes (>20 CAGs) were observed, albeit at low frequencies, with the largest repeat size change being an expansion of 153 CAGs in *Htt*^{Q175} (Figure S12). Mouse line was found to significantly predict the magnitude of both expansions and contractions (Kruskal-Wallis test, $p<0.0001$). Dunn's multiple comparisons test showed significantly higher mean expansions and contractions for most lines when compared to others with lower repeat lengths ($p<0.05$, multiplicity adjusted [116]; Figure 20B). Expansions appeared more sensitive to CAG repeat length, with expansions up to 20 CAGs apparent by ~80 CAGs (*Htt*^{Q80}, Figure S12) and contractions of the same magnitude only apparent by ~106 CAGs (*Htt*^{Q111}, Figure S12). Notably, in *Htt*^{Q50} transmissions, repeat length changes varied only from -2 to +1 CAGs, demonstrating the relatively high stability of this repeat size in these mice.

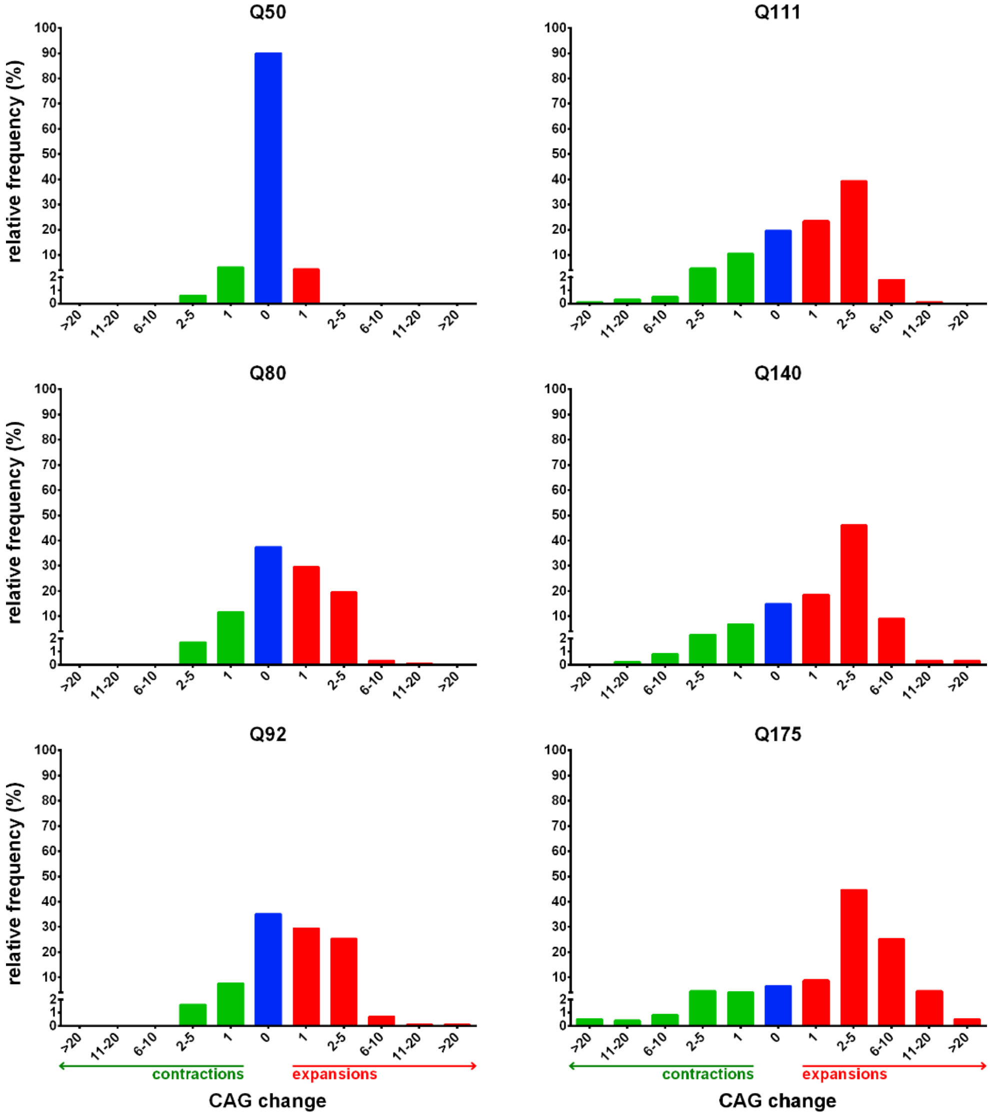


Figure 19 - Relative frequency of CAG size changes within specific ranges across all lines. Stable transmissions (blue), contractions (green), expansions (red).

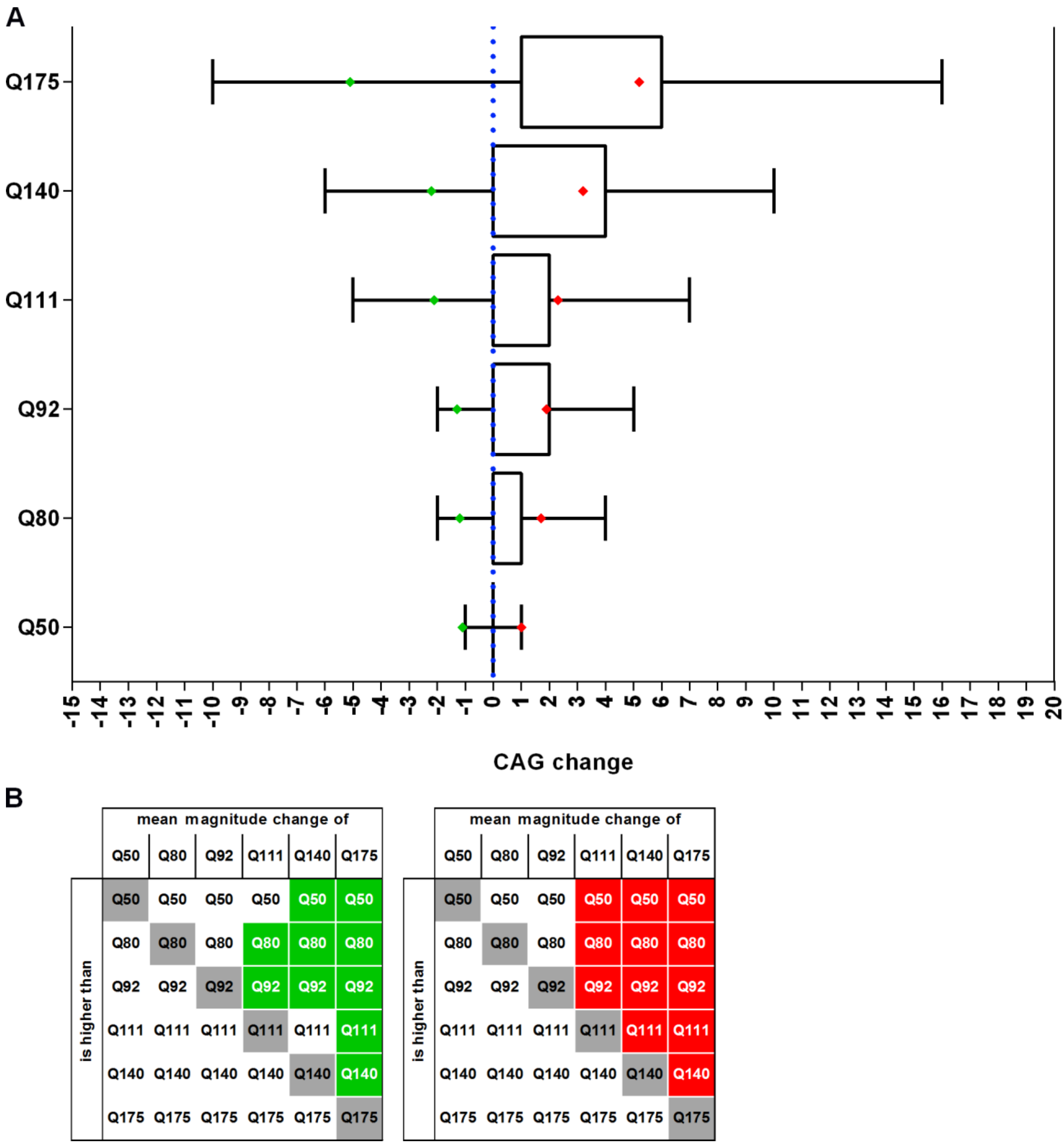


Figure 20 - Magnitude of change and significant differences in mean expansion and contraction among paternal transmissions in the JAX transmission data. (A) Boxplot representation of CAG change in paternal transmissions for each line (boxes: interquartile range of transmissions; whiskers: 1-99 percentile of transmissions; red diamonds: average expansion size; green diamonds: average contraction size).

Overall, while CAG repeat length in transmitting fathers is a major driver of intergenerational CAG instability, our analyses have distinguished CAG length-dependent effects on the frequency and magnitude of expansion and contraction events. Thus, while longer CAG lengths are associated with larger expansions and contractions, they do not impact the likelihood of contraction events, but only increase the frequency of expansions at the expense of unchanged alleles.

3.3.5. Paternal age has a minor impact on the magnitude of CAG repeat expansions

In an effort to probe for additional factors that might contribute to intergenerational CAG instability we analyzed a second dataset of intergenerational repeat length transmissions generated from CHGR's breeding of *Htt* CAG knock-in mice. This dataset is composed of ~1800 paternal transmissions of heterozygous *Htt* knock-in alleles, on six different genetic backgrounds – 129, CD1, FVB, DBA, B6N, and B6J – spanning a range of 81 to 153 CAGs (Table 6, Figure S14B).

Table 6 – Paternal age at offspring birth, CAG size and transmission information for the different lines in CHGR's breeding data.

BG	Parent							Transmissions							
	N	Age at offspring birth (weeks)			CAG			N	Stable	Contractions			Expansions		
		Min.	Max.	Mean	Min.	Max.	Mean		N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean
129	32	9	40	21.4	105	114	109.7	213	72 (33.8)	63 (29.6)	8	1.6	78 (36.6)	9	1.7
CD1	59	7	42	18.6	81	139	105.7	439	116 (26.4)	51 (11.6)	8	1.7	272 (62.0)	12	1.8
FVB	24	9	41	18.9	111	147	132.0	180	32 (17.8)	25 (13.9)	5	2.2	123 (68.3)	9	2.7
DBA	12	9	38	17.3	115	129	123.0	64	8 (12.5)	8 (12.5)	4	2.8	48 (75.0)	6	2.4
B6N	38	9	74	26.1	109	138	119.2	226	40 (17.7)	51 (22.6)	11	1.8	135 (59.7)	6	2.3
B6J	55	8	62	28.0	83	153	128.3	707	73 (10.3)	74 (10.5)	19	2.9	560 (79.2)	21	3.5

BG, background; N, number. All strains: *neo-*

This dataset included paternal age at offspring birth, therefore allowing us to evaluate its potential contribution to intergenerational instability. Given that the B6J strain had both the largest number of transmissions and a broad paternal age range, we analyzed the effect of paternal age on a subset of B6J transmissions (parental age: 8 to 62 weeks; Htt^{Q111}, CAG range: 113-153; N=690). In this dataset and select repeat length range there was no relationship between paternal age at offspring birth and paternal CAG (Figure S15). Therefore, we used this dataset to determine whether paternal age influenced intergenerational instability.

Paternal age did not significantly alter the frequency of expansions, contractions or unchanged alleles (Figure S16), showed no correlation with the magnitude of repeat contractions (Pearson correlation=0.031, p=0.8), but showed a modest though statistically significant correlation with the magnitude of the expansions (Pearson correlation=0.253, p<0.001; Figure 21).

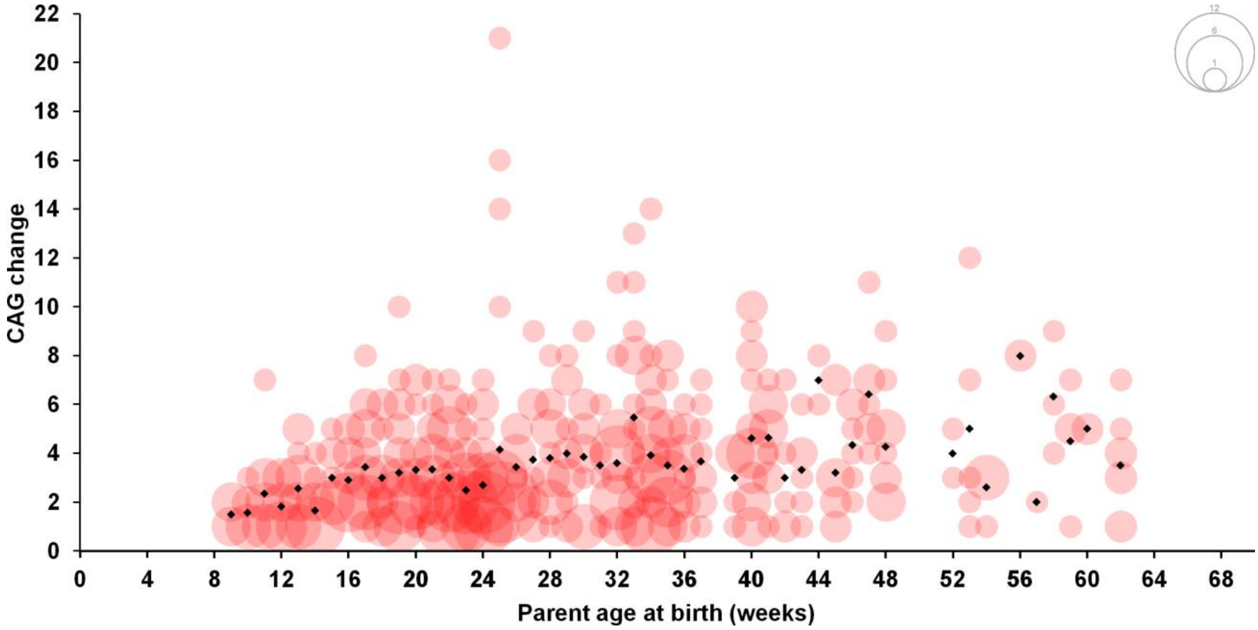


Figure 21 - CAG expansions by paternal age at offspring birth (weeks) in CHGR's B6J.*Htt*^{Q111}. Bubble size is proportional to the total number of observed events (expansions), black diamonds represent average expansion per discrete age (N=690).

3.3.6. Multiple background strains alter intergenerational CAG repeat instability.

We previously analyzed, using a much smaller transmission dataset, strain-specific differences in paternal intergenerational instability of the CAG repeat in *Htt* CAG knock-in mice using three inbred mouse strains (B6N, FVB, 129) and found a significantly greater frequency of intergenerational changes in repeat length (combined expansions and contractions) in B6N mice compared to 129 mice [91]. Here, afforded by a greatly increased number of transmissions and an expanded set of mouse strains, we aimed to investigate further potential effects of genetic background on intergenerational instability upon paternal CAG repeat transmission.

A comparison of the frequencies of expansions, contractions and stable transmissions across the six strains revealed clear differences in instability (Table 6, Figure S14A), however, the different paternal CAG repeat length distributions for each of the different strains suggest that CAG repeat length (Table 6, Figure S14B) may, in part, contribute to the differences in instability between the strains. To control for this, we modeled the frequencies of expansions, contractions and unchanged alleles as a function of CAG repeat length in the B6J strain that has the broadest range of parental CAGs as well as the most transmissions. We then compared the actual transmission frequencies observed in each of the other five strains to transmission frequencies predicted from the B6J model (Appendix and Methods). These results are represented in Figure 22 as pairwise comparisons between the observed (*e.g.* 129) and the expected transmission distributions determined by B6J modeling (*e.g.* ^{m}129).

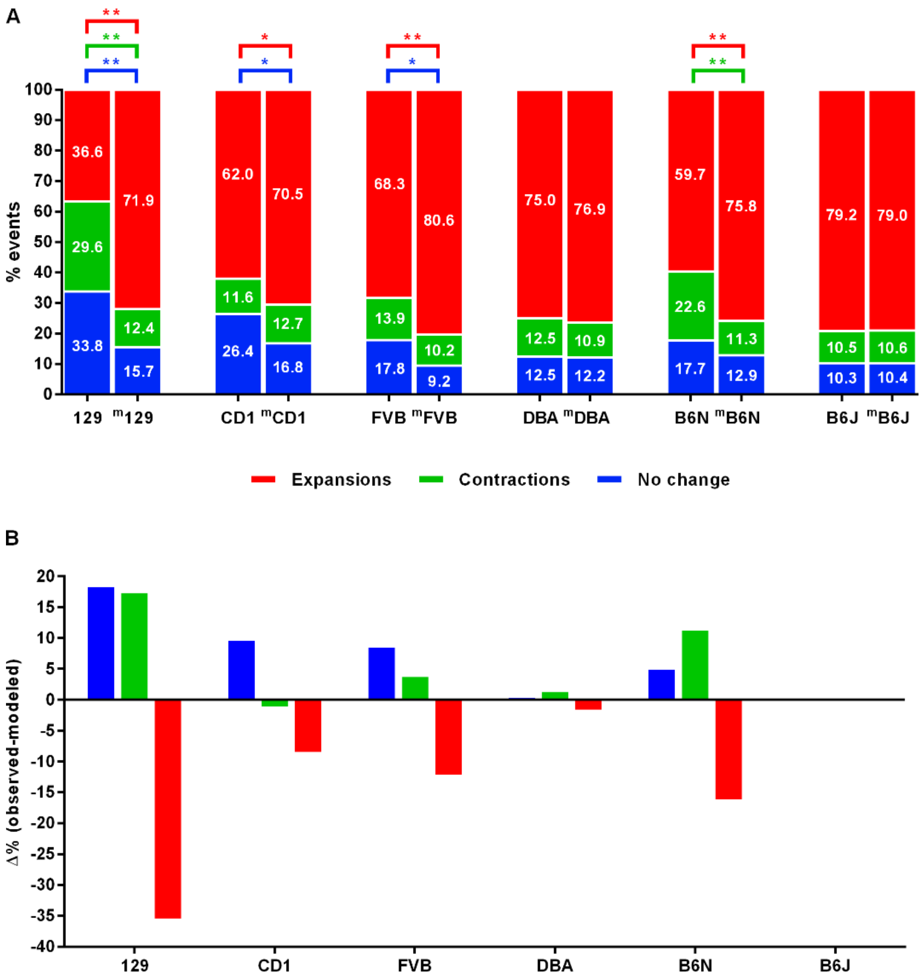


Figure 22 - Comparison of expansions, contractions and stable transmissions frequencies in various strain backgrounds with B6J-modeled distributions. (A) Observed and expected (^{m}) frequencies for each strain. (B) Percent difference between observed and expected frequencies for expansions, contractions and unchanged transmissions. **p<0.01, *p<0.05

The data show that frequencies of expansions, contractions and unchanged alleles in the DBA strain do not differ significantly from those in B6J. In contrast, 129 is the most dissimilar to B6J, with 35% fewer expansions (two-proportion z-test, p<0.01) and 17-18% more contractions and stable transmissions (two-proportion z-tests, p<0.01). CD1 and FVB both show significantly reduced frequencies of expansions (two-proportion z-test; CD1, p<0.05; FVB, p<0.01) and significantly increased frequencies of unchanged alleles (two-proportion z-test, p<0.05) compared to B6J. Interestingly, despite being the most highly related to B6J, B6N exhibited a 16.1% decrease in expansion frequency (two-proportion z-test, p<0.01) and an 11.2% increase in the frequency of contractions (two-proportion z-test, p<0.01).

It is notable that the results obtained based on modeling in B6J as a means to overcome paternal repeat length differences are mostly consistent with the initially observed differences in instability across the strains (Figure S14A), suggesting that, at least within the range of CAG repeats encompassed by these strains, genetic background, rather than CAG repeat length, is the more significant determinant of repeat instability, with distinct strains differentially modifying the frequencies of expansions, contractions and stable alleles.

The effect of background strain on the magnitude of repeat length change was also assessed using B6J as a reference and controlling for CAG size effects through mixed model analyses. Strain background did not influence the size of the contractions (data not shown). However, analysis of the strain effect on magnitude of expansions revealed significantly smaller changes compared to B6J in all strains except DBA, with 129 and CD1 being the most dissimilar to B6J and having the greatest “protective” effect (Table 7).

Table 7 – Effect of background strain on the magnitude of CAG expansions.

Strain	Intercept	P-value
B6J	1.09	NA
B6N	0.27	0.001
129	-0.17	<0.001
CD1	-0.09	<0.0001
DBA	0.33	0.054
FVB	0.31	0.005

CAG changes in offspring were modeled, as a function of paternal CAG repeat length and strain background, in a mixed effect model context with random intercept. Paternal CAG repeat significantly explained variance in size of CAG expansions, with an estimated effect size of 0.017 increase in expansion magnitude per paternal CAG repeat unit increment (p=0.0182). P-values represent significance when compared to the reference strain B6J.

3.3.7. The presence of a *neo* cassette upstream of *Htt* reduces the CAG expansion frequency

Previous data have indicated a role for *cis*-acting modifiers of CAG repeat instability in HD transgenic mice[81,117]. To explore the role of *cis* elements that might influence CAG repeat instability at the *Htt* locus, we have taken advantage of *Htt* knock-in lines that differ by the presence or absence of an upstream neomycin (*neo*) resistance cassette (Figure S8, Table S7). We compared intergenerational instability in paternal transmissions from the CD1 *Htt* CAG knock-in mice described above, which do not contain a *neo* cassette ($CD1^{neo-}$), with intergenerational instability in paternal transmissions from CD1 *Htt* CAG knock-in mice harboring an upstream *neo* cassette ($CD1^{neo+}$)[77] (Table S7).

Direct comparison between these two strains indicated the absence of any significant differences in frequencies of expansions, contractions or stable alleles (Figure 23A, two leftmost bars). However, the significantly higher mean paternal CAG size (unpaired t-test, $p<0.001$) and range of CAG repeats in the $CD1^{neo+}$ mice (Figure 7B, two leftmost bars) suggested that we may be underestimating instability in the $CD1^{neo-}$ mice relative to that in the $CD1^{neo+}$ mice due to repeat size effects. We dealt with the CAG size discrepancies using two approaches (Figure 23A and B): 1) Adjusting the $CD1^{neo-}$ dataset to only include transmissions from a paternal CAG range equivalent to the $CD1^{neo+}$ set ($^{adj}CD1^{neo-}$); 2) Modeling frequencies of events in the $CD1^{neo+}$ mice based on data from the $CD1^{neo-}$ mice ($^{m}CD1^{neo+}$) – as previously performed for the strain background analyses; see Methods and Appendix for details. When controlling for paternal CAG size by either of these two approaches we find that $CD1^{neo+}$ sires are approximately 10% less prone to expansions, with nominal significance (two-proportion z-tests, $p<0.05$) that did not withstand multiple test correction (Bonferroni significance threshold set at $p=0.0167$), with minor and not statistically significant differences in frequencies of contractions or stable transmissions (Figure 23A).

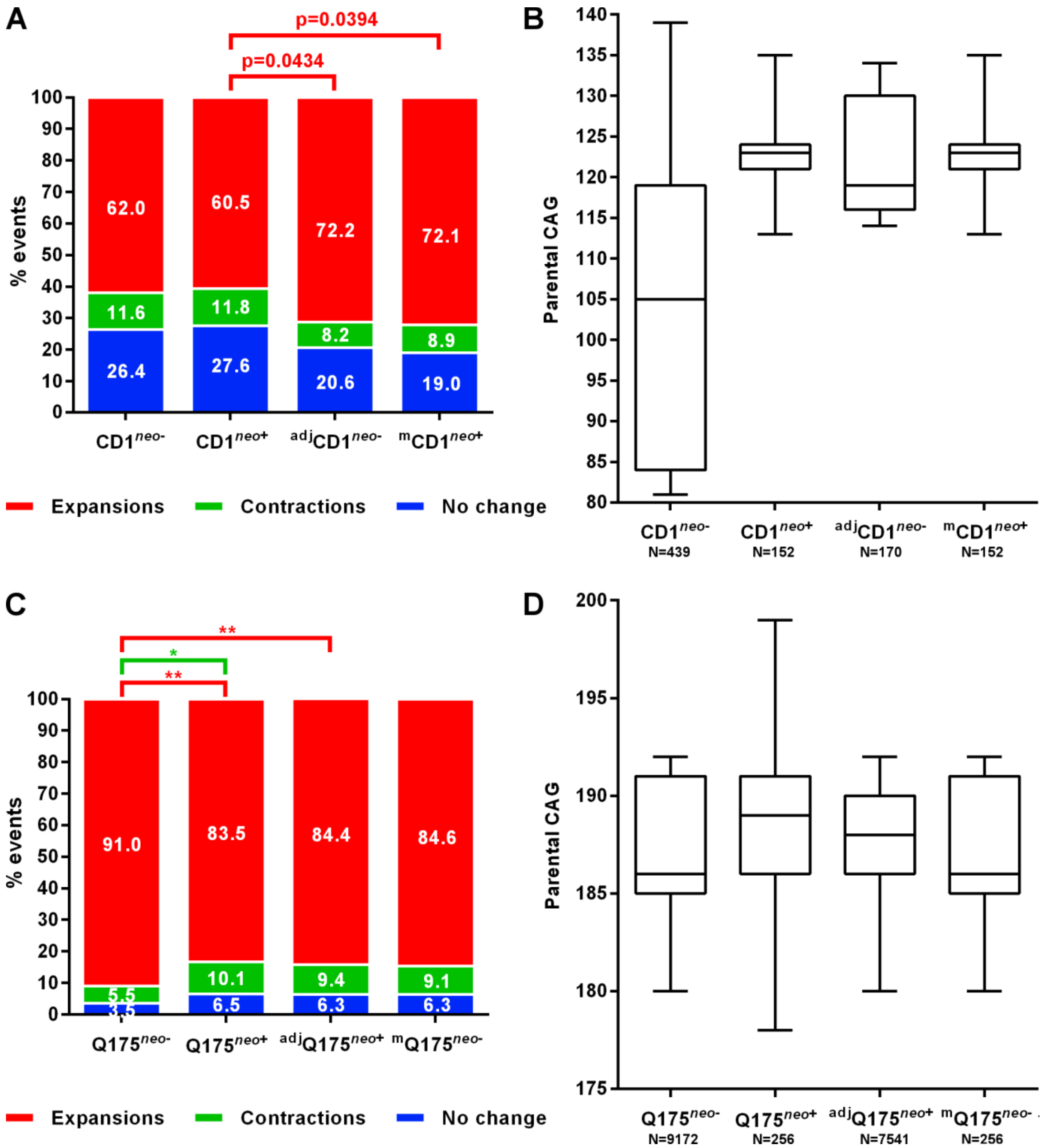


Figure 23 – Effect of an upstream neo cassette on the frequency of stable and unstable transmissions. (A) Frequency of changes in CD1 background lines either containing a neo cassette upstream of the Htt gene (*neo*+), or not (*neo*-), as well as adjusted and modeled corrections for parental CAG size effects (see Table S7). (B) Parental CAG range for the different CD1 lines, as well as adjusted and modeled subsets. (C) Frequency of changes in *Htt*^{Q175} B6J background lines containing (*neo*+) or not (*neo*-) a neo cassette upstream of the repeat, as well as adjusted and modeled corrections for parental CAG size effects (see Table S7). (D) Parental CAG range for the different HttQ175 (B6J) lines, as well as adjusted and modeled subsets. ***p*<0.01, **p*<0.0167 (Bonferroni corrected).

Differences in the magnitude of changes were evaluated by comparing mean changes between the CD1^{neo+} and the ^{adj}CD1^{neo-} set. No significant differences were observed in either mean expansion size (CD1^{neo+}: 2.1 CAGs; ^{adj}CD1^{neo-}: 2.1 CAGs;

unpaired t-test, p=0.961) or contraction size (CD1^{neo+}: 1.6 CAGs; ^{adj}CD1^{neo-}: 1.4 CAGs; unpaired t-test, p=0.610).

As these results indicated *cis* effects on the frequency of unstable transmissions we also investigated a distinct breeding set from the JAX that comprised transmissions from *Htt*$^{Q175neo-}$ parents on a B6J background (Table S7). We compared the instability in paternal transmissions of this allele with those from the *Htt*$^{Q175neo+}$ allele (B6J), which formed part of the allelic series described above.

Direct comparison between the two lines revealed significantly lower expansion frequency in *Htt*$^{Q175neo+}$ mice (two-proportion z-test, p<0.01; Figure 23C), despite a higher mean paternal CAG repeat size (unpaired t-test, p<0.001; Figure 23D), as well as a significantly increased contraction frequency (two-proportion z-test, p<0.0167), and a small non-significant increase in the frequency of stable transmissions (two-proportion z-test, p=0.057 – significance threshold set at p=0.0167).

To control for CAG repeat length we employed the two approaches described above: 1) Using a subset of the *HttQ*$^{175neo+}$ mice with parental CAGs in the same range as those in the *Htt*$^{Q175neo-}$ mice (^{adj}*Htt*$^{Q175neo+}$*;* Table S7) and 2) Modeling frequencies of events in the *Htt*$^{Q175neo-}$ line based on data from the *Htt*$^{Q175neo+}$ mice (^{m}*Htt*$^{Q175neo-}$; see Methods and Appendix). When controlling for CAG size using the adjustment method, we found a significantly reduced expansion frequency in *Htt*$^{Q175neo+}$ (two-proportion z-test, p<0.01), while the effect was only nominally significant when adjusting through the modeling methodology (two-proportion z-test, p=0.03 – Bonferroni corrected significance threshold p=0.0167), likely due to the reduced sample size enforced by this methodology (Figure 23D and Table S7). We did not find any significant differences in the mean magnitude of expansions (*Htt*$^{Q175neo-}$: 5.7 CAGs, ^{adj}*Htt*$^{Q175neo+}$: 5.2 CAGs; unpaired t-test, p=0.565) or contractions (*Htt*$^{Q175neo-}$: 9.07 CAGs, ^{adj}*Htt*$^{Q175neo+}$: 5.42 CAGs; unpaired t-test, p=0.112) in a comparison between *Htt*$^{Q175neo-}$ and the paternal CAG-adjusted *Htt*$^{Q175neo-}$ mice.

Taken together these analyses indicate that the existence of a *neo* cassette upstream of the repeat seems to be a protective factor, reducing the frequency of expansions (by ~6.5-11.5%), in two different HD knock-in mouse models, but having no discernible effects on the magnitude of the repeat length changes.

3.4. Discussion

The length of the expanded CAG repeat plays a critical role in HD, influencing both penetrance and age of onset[19,33,35,52,55,61,111]. Underlying the variation in inherited CAG repeat length between individuals are high rates (~70-80%) of intergenerational instability[33,56,59,60,69]. Here, we have analyzed thousands of intergenerational transmissions across multiple lines of *Htt* CAG knock-in mice, in order to gain further insight into factors that may influence instability.

The availability of an allelic series of mice differing by CAG repeat tract length provided the opportunity to investigate CAG-dependent aspects of repeat transmissions. We first determined whether there was any evidence for segregation distortion of the *Htt* allele. This has previously been suggested for a number of CAG/CTG repeat diseases, although the data in support of this phenomenon are conflicting[118–121]. The majority of the allelic series lines showed the expected 1:1 Mendelian ratio of heterozygous and wild-type mice in transmissions from heterozygous parents. The only exception occurred in Htt^{Q92} paternal transmissions which showed a relatively small decrease in the number of heterozygous progeny (~1.7% less than expected). Given that this was not seen in lines with shorter or longer repeat lengths this minor deviation from the expected 1:1 ratio is likely a stochastic effect, clearly not driven by CAG repeat length. Overall, our transmission analyses of a broad range of expanded CAG repeats provide evidence for no segregation distortion of the mutant *Htt* allele in mouse, supporting data derived from single sperm genotyping in HD individuals[122], and suggesting that locus-specific effects rather than repeat length, primarily drive any potential segregation distortion seen in other diseases[118–121].

For the lines with both paternal and maternal transmissions (Htt^{Q80} and Htt^{Q92}, B6J background), we confirmed previous results in HD patients and in knock-in mice on a different (CD1) genetic background[78] showing a strong expansion bias in male transmissions. We also confirmed a contraction bias in maternal transmissions observed previously in CD1 knock-in mice[78], differing somewhat from maternal transmissions in HD patients that show approximately equal expansion and contraction frequencies[60,63,64]. Limited studies carried-out to date indicate that repeat expansions in the male germline can occur at multiple stages during spermatogenesis[123,124]. Though the male expansion bias may be related to the large number of mitotic cell divisions of pre-meiotic spermatogonia, in the present study, and in line with previous data in HD patients[60], we found only a minor effect of paternal age in determining the magnitude of CAG expansions despite our analyses of a large number of transmissions over a broad age range. As increasing paternal age is expected to correlate with increased accumulation of repeat expansions in continually replicating spermatogonia, the mouse and human data overall do not strongly support spermatogonial cell divisions as the major source of CAG expansions, consistent with the lack of positive correlation of cell division rate with repeat instability[93,125]. Further understanding of the sex- and cell-type-specific processes that drive the generation of repeat expansions and contractions would be of considerable interest.

A small effect of offspring sex on intergenerational instability has been previously observed in a large Venezuelan HD pedigree, implying a role for postzygotic factors that influence instability[60]. Embryo sex also influenced instability in the R6/1 transgenic mouse model[89]. Here, analyzing large numbers of transmissions, we observed no significant effects of offspring sex on the frequency or magnitude of repeat length changes in either paternal or maternal transmissions in any of the lines studied. Even though a minor effect of offspring sex may be present in humans and transgenic mouse models it is not observable in the knock-in.

Parental CAG repeat length has been proposed as a major contributor to intergenerational repeat instability in patients[56,60,63,64]. Here we have examined paternally transmitted CAG repeats over a large range of repeat lengths (18 to ~200) and found that CAG repeat length determines the magnitude of both expansions and contractions. With the number of transmissions analyzed in this study, we were able to detect large (>10-20 CAGs) repeat length changes that are found, typically in paternal transmissions in HD patients that push the repeat into the range associated with juvenile-onset disease. Notably, however, such repeat length changes are rare in the mouse and were not detected in *Htt*^{Q50} mice harboring CAG repeat lengths typical of adult-onset HD in patients. What underlies the apparent stability of the *Htt* CAG repeat in the mouse compared to that in humans is unclear. This may simply be related to the distinct time-spans of gametogenesis in the two species, or alternatively, may reflect underlying differences in the mechanisms that drive repeat length changes. However, the high frequency of small repeat length changes and the detection of larger changes, albeit at low frequency, in mice harboring longer CAG repeats would suggest that the mechanisms that generate such changes are conserved in the mouse.

We also found that longer CAG repeat lengths were associated with a higher frequency of repeat expansions. Interestingly, this was mirrored by a decrease in the frequency of unchanged alleles, with no effect on the frequency of contractions, despite the association of repeat length with the magnitude of the contractions. This may reflect different mechanisms underlying expansions and contractions. In this scenario, mechanism(s) driving expansions are engaged in a repeat length-dependent manner; in contrast, mechanism(s) driving contractions are engaged regardless of repeat length, but once engaged, longer CAG lengths are more likely to drive larger contractions. Previous observations in which paternal transmissions of *Htt*^{Q111} *Msh2* knockout mice exclusively exhibited contractions, in contrast to the predominant expansions in *Htt*^{Q111} mice wild-type for *Msh2* [90], support separate expansion and contraction mechanisms that are, respectively, dependent on and independent of *Msh2*. Different mechanisms of intergenerational expansion and contraction of CAG/CTG repeats are supported by several additional studies[104,110,126,127].

Expanding on previous work[91], we analyzed intergenerational changes across six genetic backgrounds – 129, CD1, FVB, DBA, B6N and B6J – and in expanded datasets that afford the power to distinguish strain-specific effects on both the frequency and magnitude of changes as well as on expansions and contractions.

Overall, B6J and DBA were the most unstable strains, possessing similar frequencies and magnitudes of repeat length changes, while 129 was the most stable strain, with the low frequency and magnitude of changes in this strain consistent with previous data[91]. More specifically, we observed that different strains variably altered the relative frequencies of expansions, contractions and stable transmissions (Figure 22). Strain background also modified the magnitude of the expansions but not the magnitude of the contractions, though this may be in part due to the lower number of contraction events. As all of these strains were derived from an initial *Htt*^{Q111} line, generated by targeting in 129 embryonic stem cells[78], all the *Htt* knock-in alleles are on a local 129 haplotype, and any differences in instability can likely be attributed to *trans*-effects, mediated by variation in other genes.

Previous analyses point to *Mlh1* genetic variation underlying differences in *Htt* CAG somatic instability between 129 and B6N strains[102]. Given overlapping roles of mismatch repair (MMR) genes in somatic and intergenerational instability in HD [90,104] and other repeat disorders[126,128], it seems likely that *Mlh1* genetic variation underlies the 129 versus B6 (B6N and B6J) differences in intergenerational instability. Interestingly, the reduced frequency of expansions in the 129 strain was accompanied by an increased contraction frequency, reminiscent of the impact of loss of *Msh2*[90] and further suggesting that genetic variation in 129 might impact the same pathway(s). Additional, unbiased genetic analyses would be needed to uncover the modifier(s) responsible for the reduced intergenerational instability in 129 mice, as well as in other strains. Interestingly, although B6J and B6N are closely related strains, B6N showed an increase in the frequency of contractions and a decrease in frequency and magnitude of expansions relative to B6J. While these strains do not have any coding or obvious regulatory region SNPs in MMR genes (data not shown), the limited genetic variation between the two strains may provide an opportunity to uncover, new modifiers that shift the balance of repeat length changes from expansions toward contractions.

In addition to genetic background strain effects attributable to *trans*-acting modifiers, we also examined potential *cis*-effects by comparing intergenerational instability in strains with and without a *neo* cassette upstream of the knock-in repeat. The presence of the *neo* cassette was associated with reduced expansion frequency in paternal transmissions, without having an effect on the magnitude. We observed this effect independently of paternal CAG size effects, in two different background strains (CD1 and B6J) in the context of two knock-in lines (*Htt*^{Q111} and *Htt*^{Q175}) with different sites of *neo* insertion (Figure S8).

Considerable data support a role for *cis*-acting modifiers of trinucleotide repeat instability[117,129,130]. At the human *HTT* locus itself, an instability-promoting haplogroup has been proposed to drive expansion from the high normal range[66], however *HTT* haplotype does not modify the intergenerational instability of expanded CAG repeats[50]. Regardless, *cis*-modifiers of *HTT* CAG instability in model systems may provide insight into underlying mechanisms. Reduced instability in the presence of the neo-insertions may be a consequence of chromatin structural changes, and/or alteration of *Htt* transcription levels during germ cell generation. Interestingly, the orientation of the *neo* cassette, in relation to the *Htt* knock-in allele, is different between the Htt^{Q111} and Htt^{Q175} lines (sense and antisense, respectively), and although the *neo* insertion in the Htt^{Q111} allele dramatically reduces its transcription resulting in a “hypomorphic” allele[114] the *neo* insertion in the Htt^{Q175} allele does not have the same impact on *Htt* expression[82]. This suggests that altered transcription may not be the major contributor to reduced instability in the *neo*+ mice, but rather that local sequence structure/chromatin configuration may impact CAG instability via other mechanisms.

While this study was not specifically geared towards providing a comprehensive understanding of the molecular mechanisms underlying CAG repeat instability, our results provide some general insights that may help to direct future research in this area. Our observation that age is not a major determinant of intergenerational instability implies a major role for DNA repair processes that are not directly linked to DNA replication. We suggest that a minor component of intergenerational instability is driven by processes directly linked to DNA replication, where MMR proteins may act at the level of post-replicative MMR, or may play a direct role at the replication fork[127,131]. We also provide further evidence that mechanisms of intergenerational expansion and contraction can be distinguished, perhaps in part reflecting different cell-types in which these events may predominate. Thus, further efforts to understand mechanisms underlying repeat contraction are warranted to provide opportunities for therapeutic strategies aimed at reducing repeat length.

In summary, our comprehensive analyses of intergenerational transmissions of *Htt* CAG repeats in HD knock-in mice confirms parent-of-origin and CAG repeat length as the major modifiers of intergenerational instability, as in HD patients. The large datasets have also allowed us for the first time to discern more subtle effects on instability, *e.g.* distinguishing CAG-dependent effects on frequency and magnitude of expansions and contractions, and to identify large repeat size jumps seen in HD patients, the latter suggesting that fundamental mechanisms of CAG instability are shared between human and mouse. Evidence for both *cis*- and *trans*-modifiers of instability provides a starting point to uncover the underlying modifying factors in the mouse, which will provide further insight into intergenerational instability in patients.

3.5. Supplementary material

3.5.1. Supplementary figures

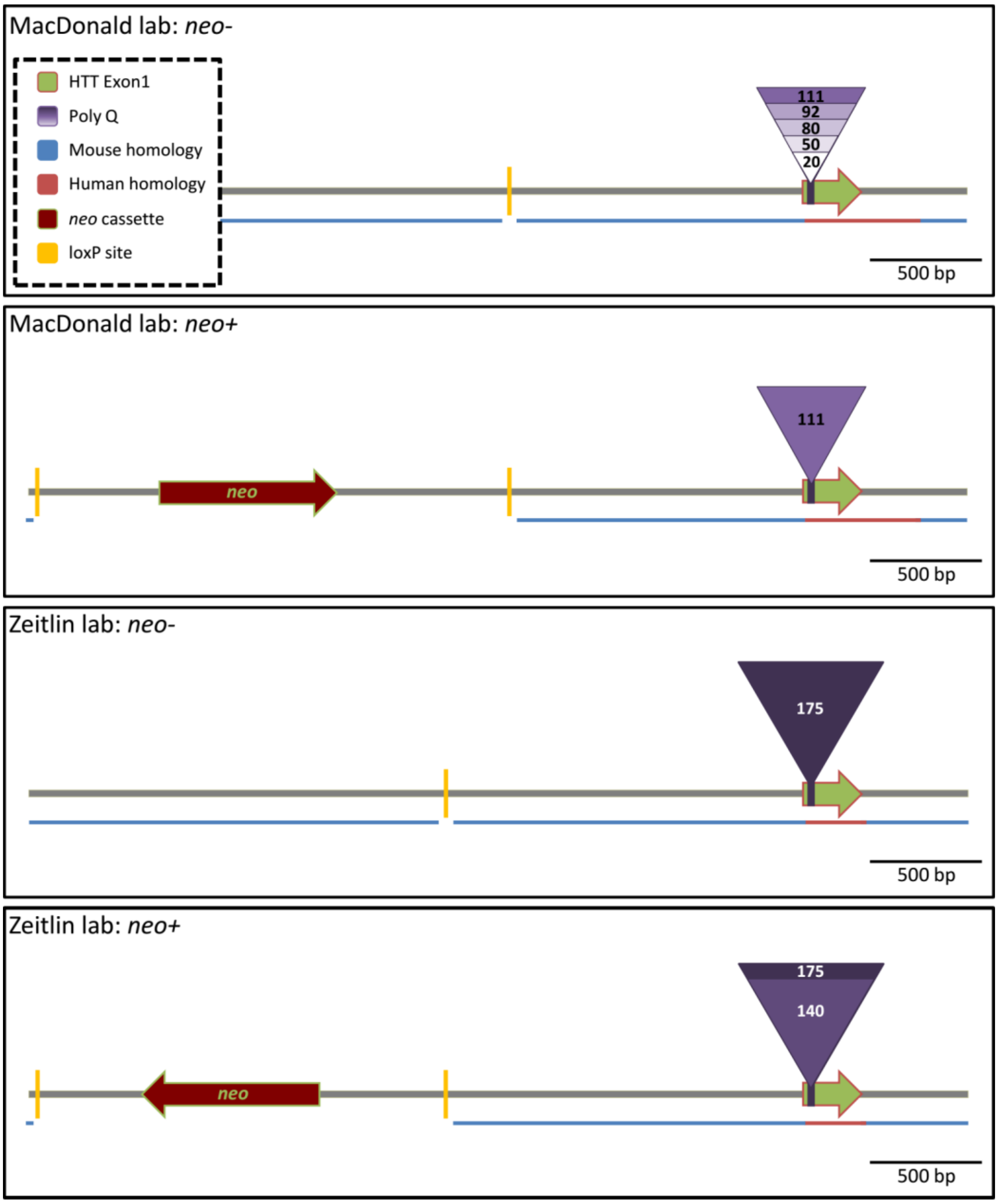


Figure S8 – Schematic representation of the *Htt* locus representing the different lines used in this study as well as the relative location and direction of the neo cassette and other locus elements.

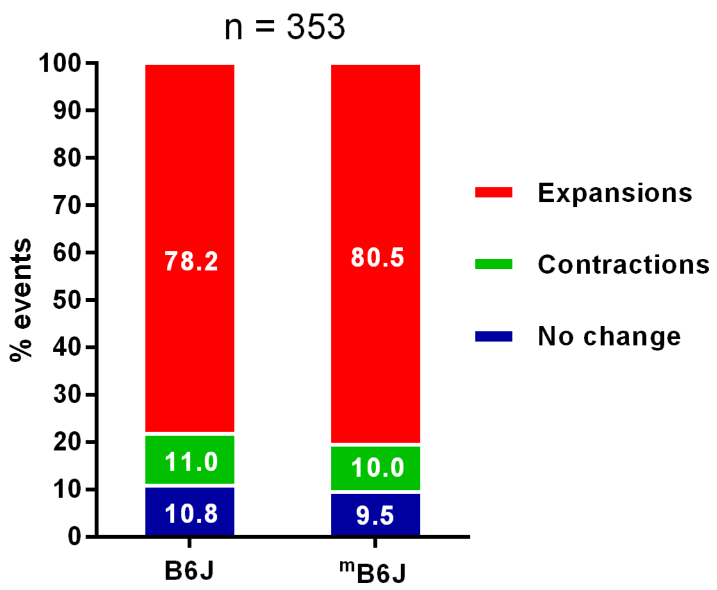


Figure S9 – Validation of the frequency modeling methodology for comparison of intergenerational instability across different datasets. Left bar represents the frequencies observed within 50% of the B6J dataset (n=353), while the right bar represents the same dataset with frequencies modeled based on the remaining 50% (n=354) which was used as the reference dataset. No significant differences were observed.

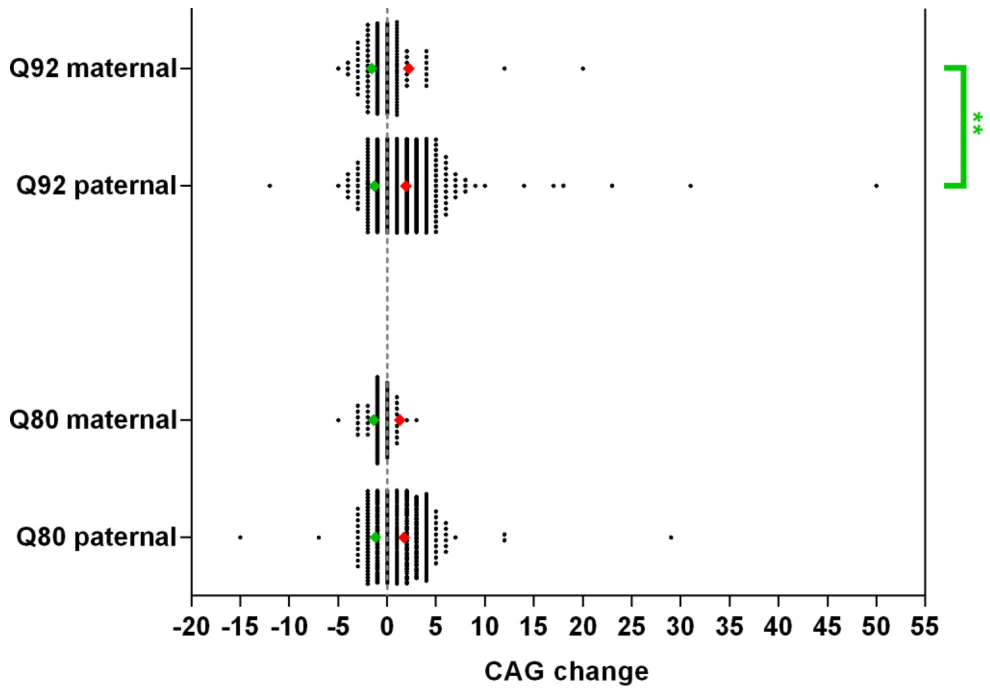


Figure S10 – Magnitude of changes in paternal and maternal transmissions (*Htt*^{Q80} and *Htt*^{Q92}) in JAX's transmission data. Representation of CAG change for all maternal and paternal transmissions observed (dotted, black), in the *Htt*^{Q80} and *Htt*^{Q92} lines, as well as mean expansion (diamonds, red) and mean contraction (diamonds, green) values (see also Table 4; Table 5). **p<0.01

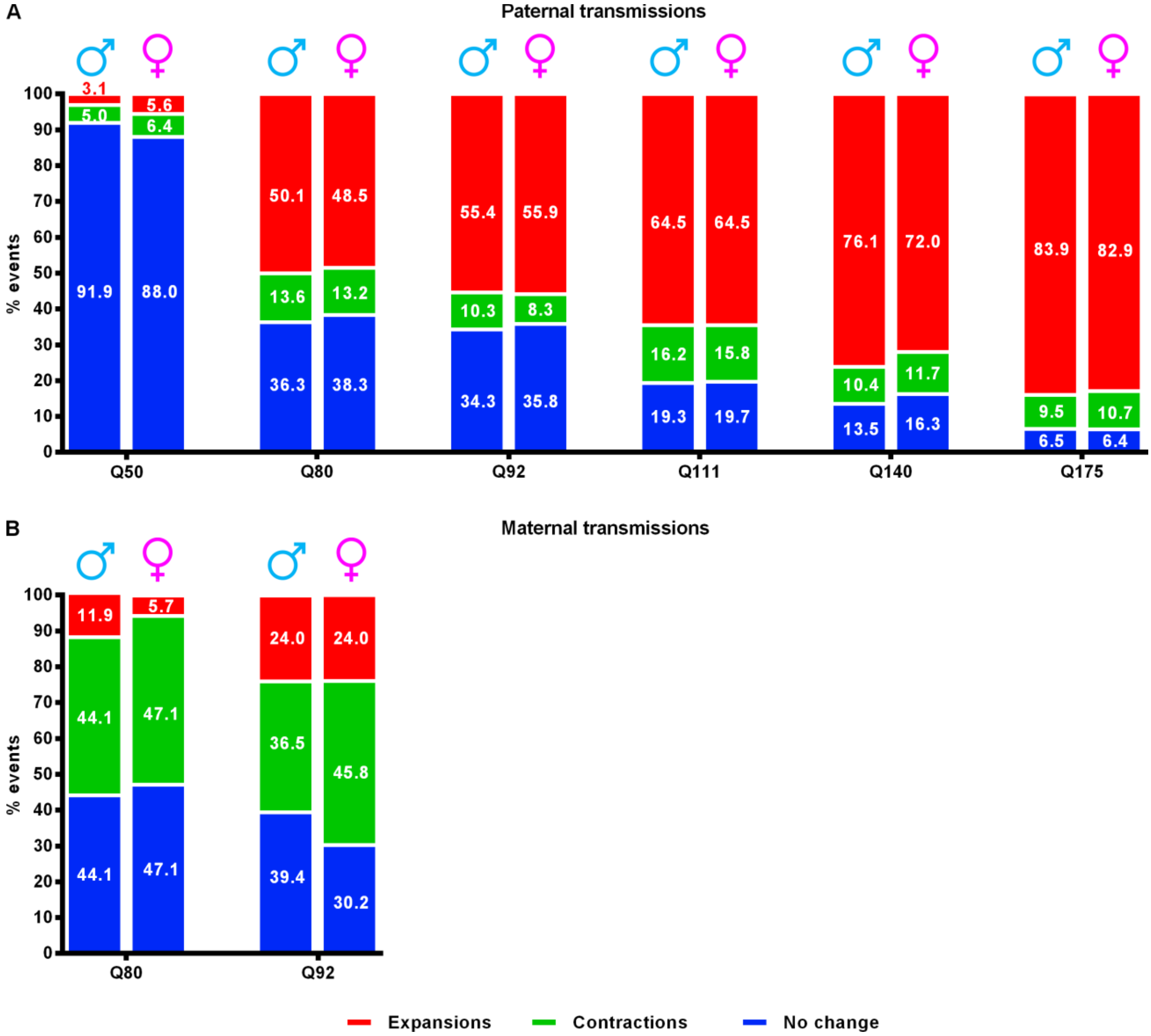


Figure S11 – Relative frequency of stable and unstable transmissions by offspring sex in the JAX *Htt* CAG knock-in dataset. Breakdown of transmission frequency by expansions (red), contractions (green) and stable transmissions (blue) separated by offspring sex among paternal (A) and maternal (B) transmissions.

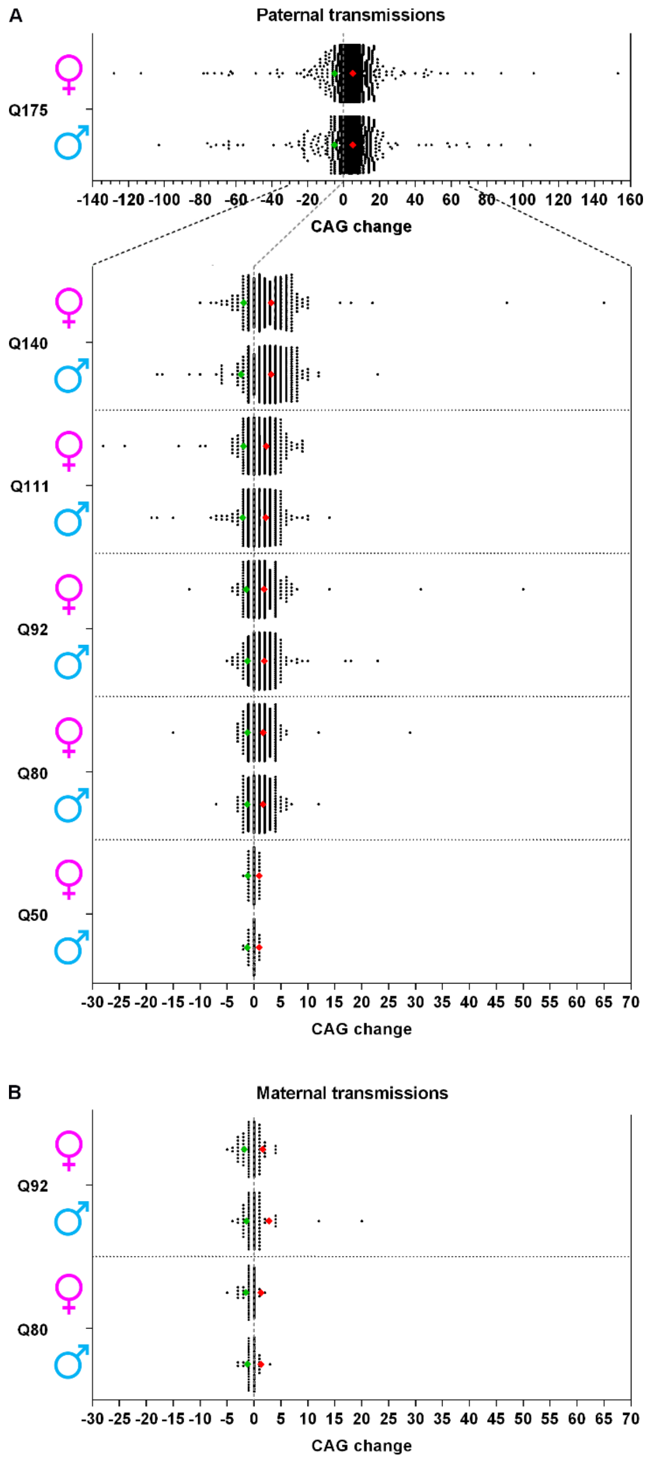


Figure S12 – Magnitude of changes segregated by offspring sex in paternal and maternal transmissions in JAX's transmission data. Representation of CAG changes for all paternal (A) and maternal (B) transmissions observed (dotted, black) in the available lines, as well as mean expansions (diamonds, red) and mean contractions (diamonds, green).

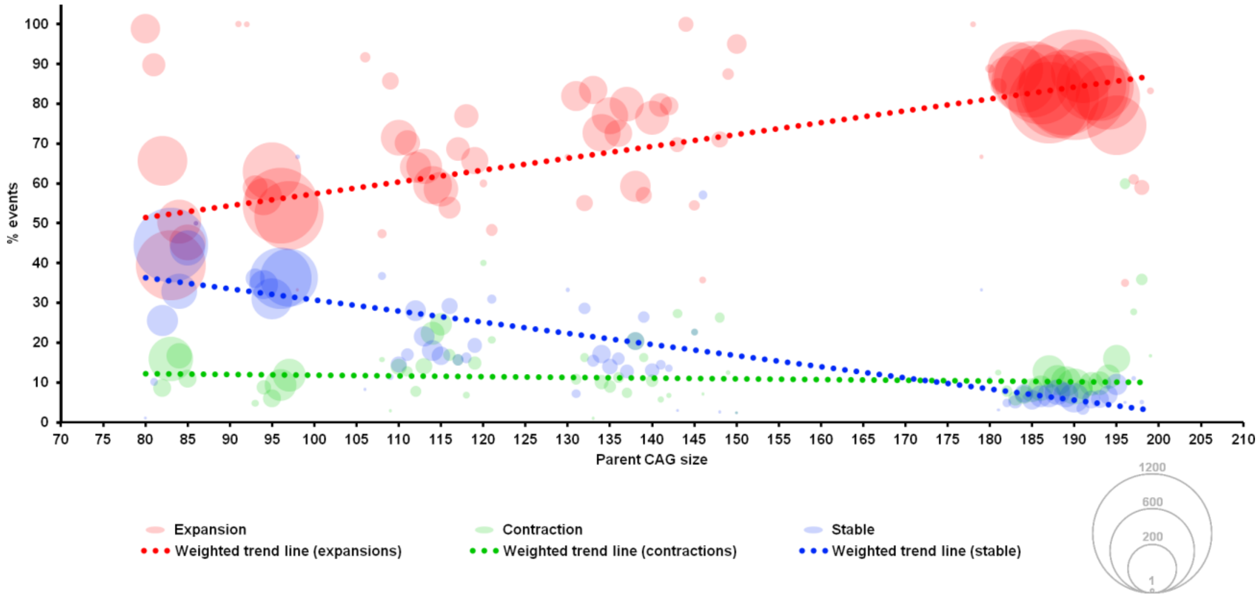


Figure S13 – Relative frequency of stable and unstable (expansions, contractions) transmissions by paternal CAG size in JAX's dataset. Breakdown of transmission frequency by expansions, contractions and stable transmissions using paternal CAG size as a continuous variable. Trend lines weighed by the total number of observed transmissions for each parental CAG length are represented as dotted lines. Bubble size is proportional to the total number of observed events. Events with null frequency (N=0) are considered for trend line weighing but are not depicted as bubbles.

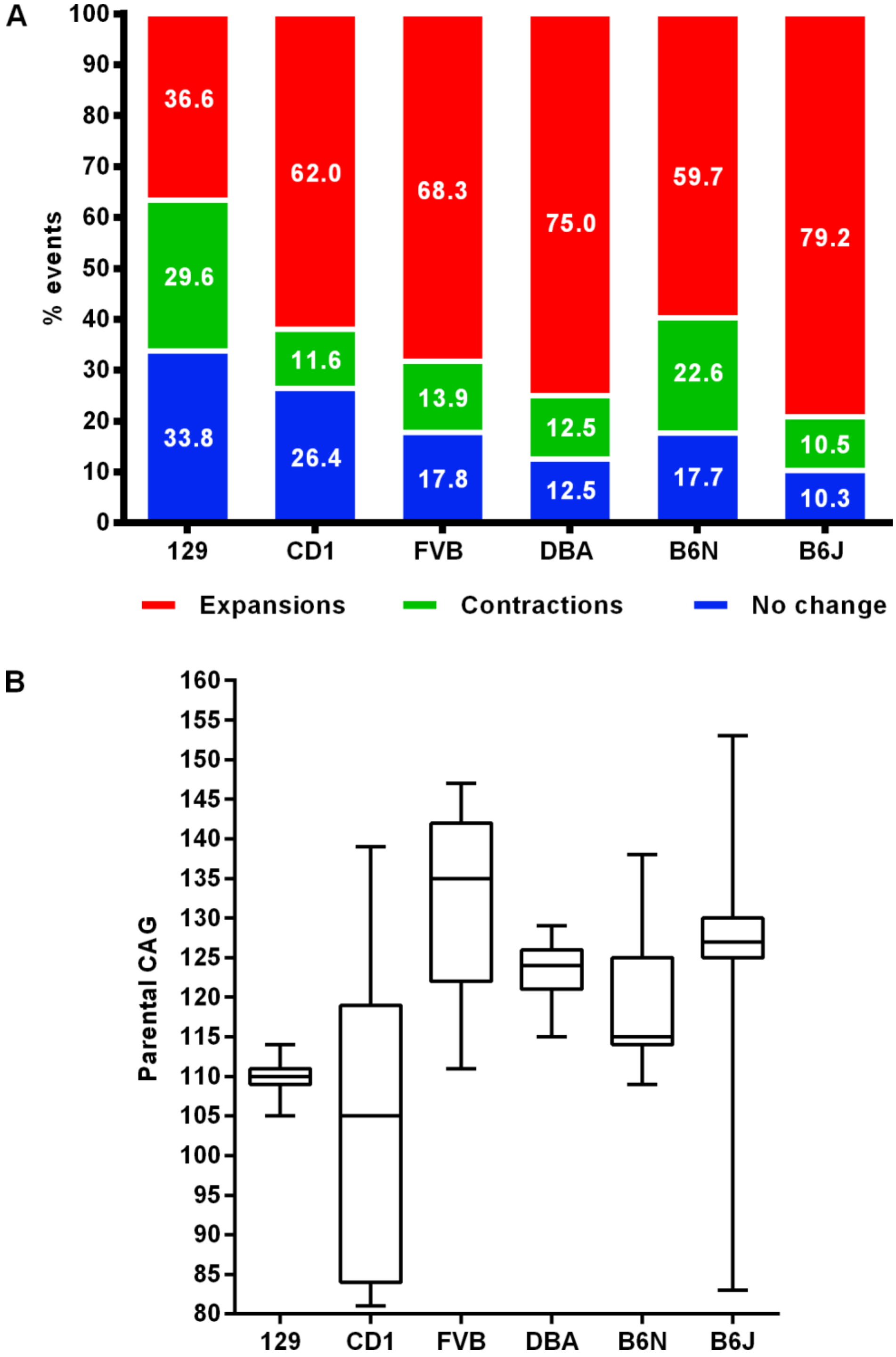


Figure S14 – Frequency of changes and paternal CAG sizes across the six genetic backgrounds. (A) Frequency of expansions, contractions and stable transmissions across different strains in the CHGR breeding dataset (B) Parental CAG range among the six strains (boxes encompass 50% of total transmissions, whiskers represent minimum to maximum size).

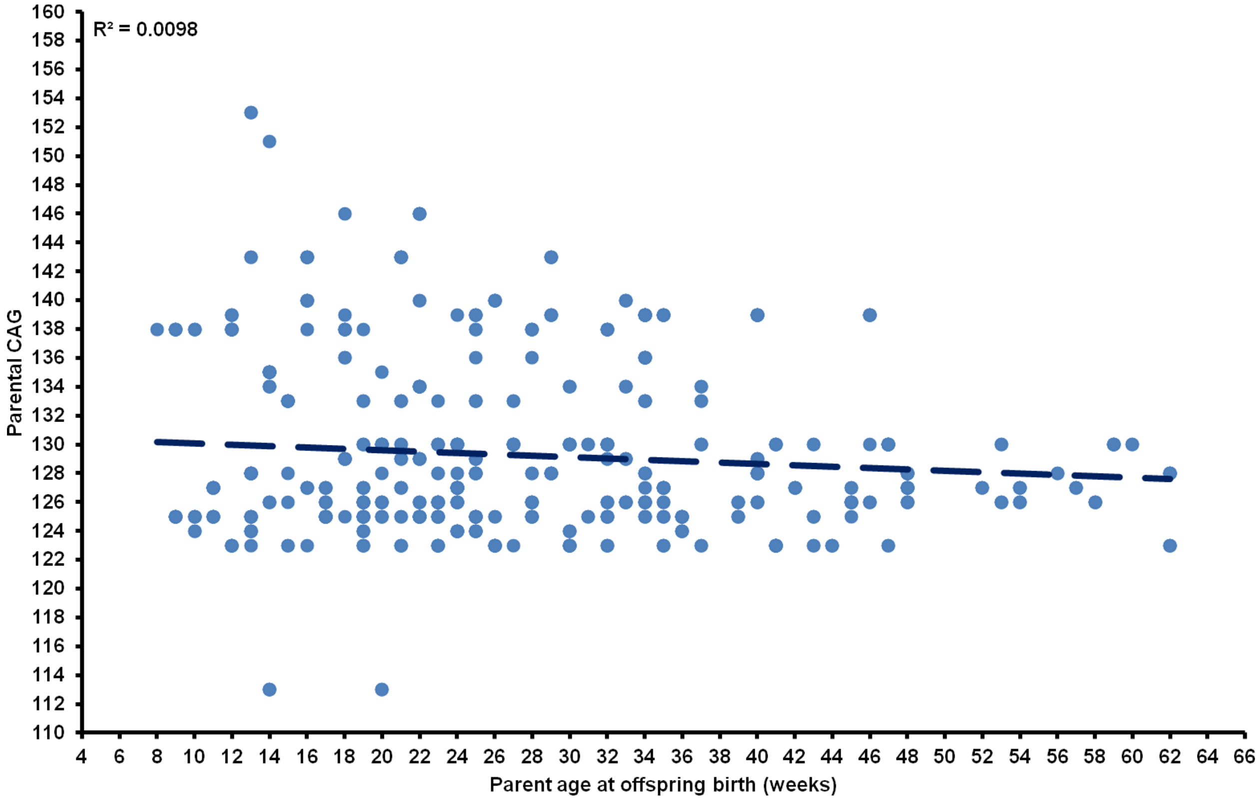


Figure S15 - Distribution of parental CAG repeat size across paternal age at birth in the B6J.*Htt*^{Q111} mice (113-153 CAGs; N=690 transmissions) in CHGR's breeding dataset.

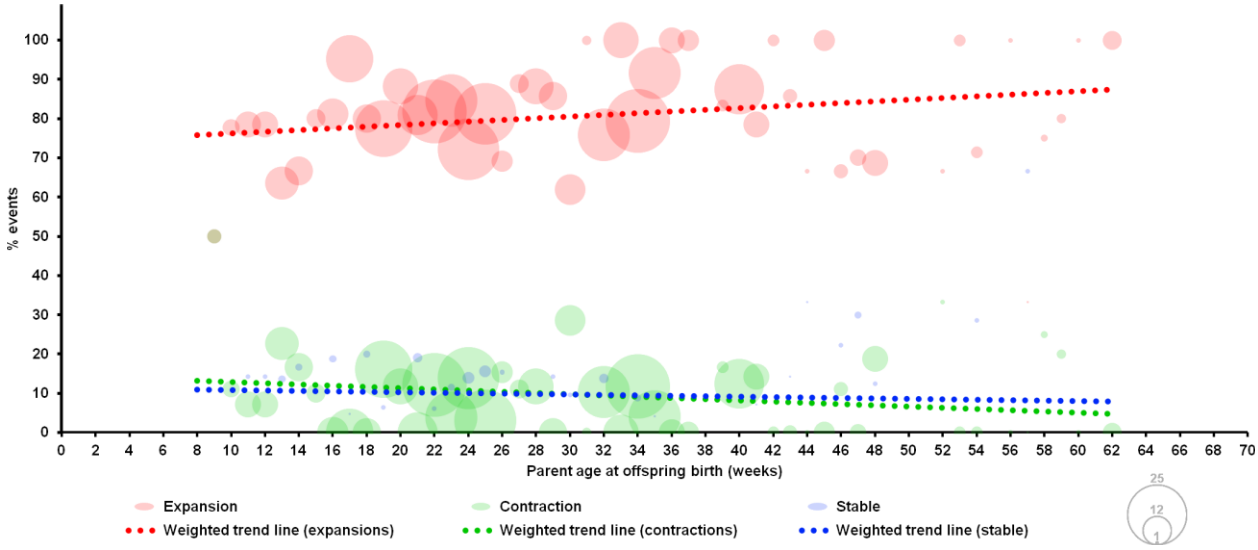


Figure S16 – Frequency of changes by paternal age at offspring birth (B6J background, CHGR dataset). Trend lines weighed by the total number of observed transmissions for each parental age are represented as dotted lines. Bubble size is proportional to the total number of observed events. Events with null frequency (N=0) are considered for trend line weighing but are not depicted as bubbles.

3.5.2. Supplementary tables

Table S5 – Paternal and maternal transmission data for the different lines segregated by offspring sex in JAX's breeding data.

A

	Paternal transmissions																
	to male offspring								to female offspring								
	Total	Stable	Contractions			Expansions			Total	Stable	Contractions			Expansions			
Line	N	N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean	N	N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean	
Q50	260	239 (91.9)	13 (5.0)	2	1.2	8 (3.1)	1	1.0	251	221 (88.0)	16 (6.4)	2	1.1	14 (5.6)	1	1.0	
Q80	1184	430 (36.3)	141 (13.6)	7	1.2	593 (50.1)	12	1.7	1281	491 (38.3)	169 (13.2)	15	1.2	621 (48.5)	29	1.7	
Q92	1427	489 (34.3)	147 (10.3)	5	1.2	791 (55.4)	23	1.9	1456	521 (35.8)	121 (8.3)	12	1.4	814 (55.9)	50	1.9	
Q111	772	149 (19.3)	125 (16.2)	19	2.1	498 (64.5)	14	2.2	778	153 (19.7)	123 (15.8)	28	1.9	502 (64.5)	9	2.3	
Q140	779	105 (13.5)	81 (10.4)	18	2.4	593 (76.1)	23	3.2	779	127 (16.3)	91 (11.7)	10	1.9	561 (72.0)	65	3.2	
Q175	4719	309 (6.5)	450 (9.5)	103	5.1	3960 (83.9)	104	5.2	4453	283 (6.4)	477 (10.7)	128	5.0	3693 (82.9)	153	5.2	

N, number

B

	Maternal transmissions																
	to male offspring								to female offspring								
	Total	Stable	Contractions			Expansions			Total	Stable	Contractions			Expansions			
Line	N	N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean	N	N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean	
Q80	59	26 (44.1)	29 (44.1)	3	1.2	7 (11.9)	3	1.3	70	33 (47.1)	33 (47.1)	5	1.5	4 (5.7)	2	1.3	
Q92	104	41 (39.4)	38 (36.5)	4	1.4	25 (24.0)	20	2.8	98	29 (30.2)	44 (45.8)	5	1.8	23 (24.0)	4	1.6	

Table S6 – χ^2 and p-values for offspring sex effect on the relative frequencies of contractions, expansions and stable alleles in the JAX *Htt* CAG knock-in dataset.

		χ^2 tests of independence regarding offspring gender		
Line	Transmission	χ^2	df	p-value
Htt ^{Q50}	Paternal	2.493	2	0.287
Htt ^{Q80}	Paternal	1.065	2	0.587
	Maternal	1.553	2	0.460
Htt ^{Q92}	Paternal	3.574	2	0.167
	Maternal	2.263	2	0.323
Htt ^{Q111}	Paternal	0.062	2	0.970
Htt ^{Q140}	Paternal	3.555	2	0.169
Htt ^{Q175}	Paternal	3.532	2	0.171

Table S7 – Characteristics of paternal CAG intervals and transmission frequency and magnitude in the *neo*+, *neo*- and adjusted datasets.

Line	Paternal CAG			Transmissions							
				N	Stable	Contractions			Expansions		
	Min.	Max.	Mean		N (%)	N (%)	Max.	Mean	N (%)	Max.	Mean
CD1 $^{neo-}$	81	139	105.7	439	116 (26.4)	51 (11.6)	8	1.7	272 (62.0)	12	1.8
CD1 $^{neo+}$	113	135	122.6	152	42 (27.6)	18 (11.8)	2	1.4	92 (60.5)	8	2.1
^{adj}CD1 $^{neo-}$	114	134	121.5	170	35 (20.6)	14 (8.2)	6	1.6	121 (71.2)	12	2.1
Q175 $^{neo+}$	178	199	188.9	9172	592 (6.5)	927 (10.1)	128	5.1	7653 (83.4)	153	5.2
Q175 $^{neo-}$	180	192	187.1	256	9 (3.5)	14 (5.5)	88	9.1	233 (91.0)	30	5.7
^{adj}Q175 $^{neo+}$	180	192	187.8	7541	473 (6.3)	706 (9.4)	128	5.4	6362 (84.4)	153	5.2

4. Characterization and comparison of somatic repeat instability in *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ HD mouse models

4.1. Introduction

Somatic instability of the repeat is a prevalent feature in mouse models of HD[78,81,97]. Both neuronal and non-neuronal tissues are affected by CAG repeat size changes, with liver and striatum showing the most dramatic changes in *Htt*^{Q111}, R6/1 and R6/2 models[78,97,102]. Instability in *Htt*^{Q175} tissues has been observed, but has of yet not been thoroughly characterized.

Overall, a large amount of evidence regarding *trans*-acting modifiers of somatic instability has been amassed using the aforementioned *Htt*^{Q111} mice as well as R6/1 models[92,102–104,132]. Genetic background of *Htt*^{Q111} mice was shown to influence somatic instability, with B6 and FVB strains showing larger changes in repeat size than the 129 strain. Through linkage mapping the likely culprit of differences in instability in the striatum was identified as *Mlh1*, and validation of this fact was performed through the generation of *Htt*^{Q111}.*Mlh1*$^{-/-}$ (knock-out) mice, which show a high ablation of expansions in striatum and liver. This same strategy crossing HD models with knock-out mice for other MMR genes such as *Msh2*, *Msh3*, *Mlh1*, *Mlh3*, and *Neil1*[90,92,102,104], has been prolific in the identification of somatic instability *trans*-modifiers. In the case of *Mlh1* and *Msh3*, even naturally occurring SNPs have been found to probably be the underlying cause of instability differences amongst distinct strains[102,103].

While mounting evidence has been found for *trans* modifiers of somatic instability, little is known regarding *cis* and short-range modulators in mouse models, the most substantial indicators of *cis*-modifiers relate to early reports of differential instability amongst R6 models, where the expanded repeat showed very distinct profiles of repeat size changes in numerous tissues depending only on a different insertion location and genetic context[81].

In the previous chapter we described a potential role for a specific *cis*-element, a neomycin resistance cassette upstream of the repeat, as a modulator of intergenerational instability in *Htt*^{Q175} (B6J) and *Htt*^{Q111} (CD1) models, namely through a visible effect in the reduction of expansion frequency.

In this work we set out to characterize somatic instability in different tissues of *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ models, assessing differences in tissue instability within each of the lines and searching for possible instability modulation effects of the *neo* sequence upstream of the repeat.

4.2. Animals and methods

4.2.1. Animals and tissues

Htt$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ mice (8 for each line, 4 male, 4 female) were acquired from the The Jackson Laboratory and were sacrificed and perfused with paraformaldehyde for tissue fixation at 7-weeks-old, as previous results from the lab showed that older mice with typical repeat sizes from this line show very high levels of instability hard to quantify through the methodology here applied, and should still be appropriate toward finding possible differences.

Several tissues were collected and subsequently frozen. Part of seven of the collected tissues – tail, heart, spleen, liver, cerebellum, cortex and striatum – was chipped from the frozen samples and utilized for DNA extraction.

4.2.2. DNA extraction

DNA was extracted using the 5 PRIME (Fisher Scientific) Manual ArchivePure DNA Purification methodology. Tissues were incubated overnight at 50 °C, in 300µL of Cell Lysis Solution (5 PRIME) and 5µg of Proteinase K. The “broken-down” tissue solutions were subsequently incubated with RNAse A for 1h at 37 °C. Samples were then moved to room temperature (RT) and 100µL of Protein Precipitation Solution (5 PRIME) was added, followed by vortex homogenization, and samples were subsequently cooled on ice for 20 minutes. At RT, samples were centrifuged for 3 minutes at 13000 rpm, supernatants were transferred to new tubes and pellets were discarded. 300µL of molecular biology grade isopropanol (Fisher Scientific) were added to the new tubes for DNA precipitation and manual homogenization was performed by inversion (15 to 20 times) or until precipitated DNA was visible. Samples were centrifuged once more at 13000 rpm for 3 min. Supernatants were then discarded and DNA pellets were washed with 300µL of a 70% ethanol solution by manual inversion. Samples were subjected to a final centrifugation at 13000 rpm for 1 minute and ethanol supernatants were carefully removed to avoid disturbing DNA pellets. Samples were left to air dry for ~20 minutes and were then solubilized in TE buffer.

DNA concentration was determined through spectrophotometry using a Nanodrop 1000 and afterwards diluted to 40ng/µL for downstream use.

4.2.3. Determination of somatic instability

Fragment analysis was performed to evaluate instability in the tissues of interest, starting with a polymerase chain reaction (PCR) of the sequence containing the expanded repeat. PCRs were performed using the *Taq* PCR Core Kit (Qiagen) where each reaction contained 1x PCR buffer (Qiagen), 20% Q-solution (Qiagen), 16 nmol of HU3 (5' GGCGGCTGAGGAAGCTGAGGA 3') and labelled CAG1 (5' 6-FAM-ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC 3') primers – well characterized and specific to the humanized repeat containing allele –, 4 nmol of dNTPs and 0.5 units of *Taq* in a final volume of 20µL. Cycling conditions were as follows, 95 °C 5min, 30 cycles of 94 °C 30sec ,65 °C 30 sec, 72°C 90sec, followed by 10 min at 72 °C. Products were then subjected to capillary electrophoresis in an ABI3730 (Applied Biosystems) sequencer. Electropherograms were subsequently analyzed with GeneMapper Software v5.0 (Applied Biosystems) undergoing rigorous quality control and validation.

Somatic instability was determined using a method previously described[93]. In short: the PCR reaction generates different products of multiple sizes that when analyzed in GeneMapper appear as a cluster of peaks differing from each other by one CAG repeat unit, and peak height is proportional to the number of alleles containing a specific repeat size that are amplified from the mixture of genomic DNA of different repeat lengths; the peak containing the highest signal is considered the “main allele”; alleles to the right are considered expansions and alleles to the left contractions; heights for all alleles of interest are extracted and individual peak heights are normalized to the sum of all heights; normalized peak heights are then multiplied by a factor determined by the change/distance to the main allele (*e.g.* for an allele 1 CAG smaller than the main allele this factor is -1, for an allele 2 CAGs larger the factor is 2, and so forth); the values stemming from this calculation are then added together generating the instability index. Expansion and contraction indexes can also be determined, adding only the peaks to the right or left of the main allele.

In this study, slight changes to the methodology were applied. Instead of adding a relative threshold (*e.g.* 20% height of the main allele) as described in the original method[93], an absolute threshold of 50 for minimum peak height was applied (without any relative threshold application, as noise levels here were fairly low and this allowed to capture the most amount of expanded alleles), and alternatively to the instability index, an absolute instability index – consisting of the sum of the expansion index and the absolute value of the contraction index – was calculated.

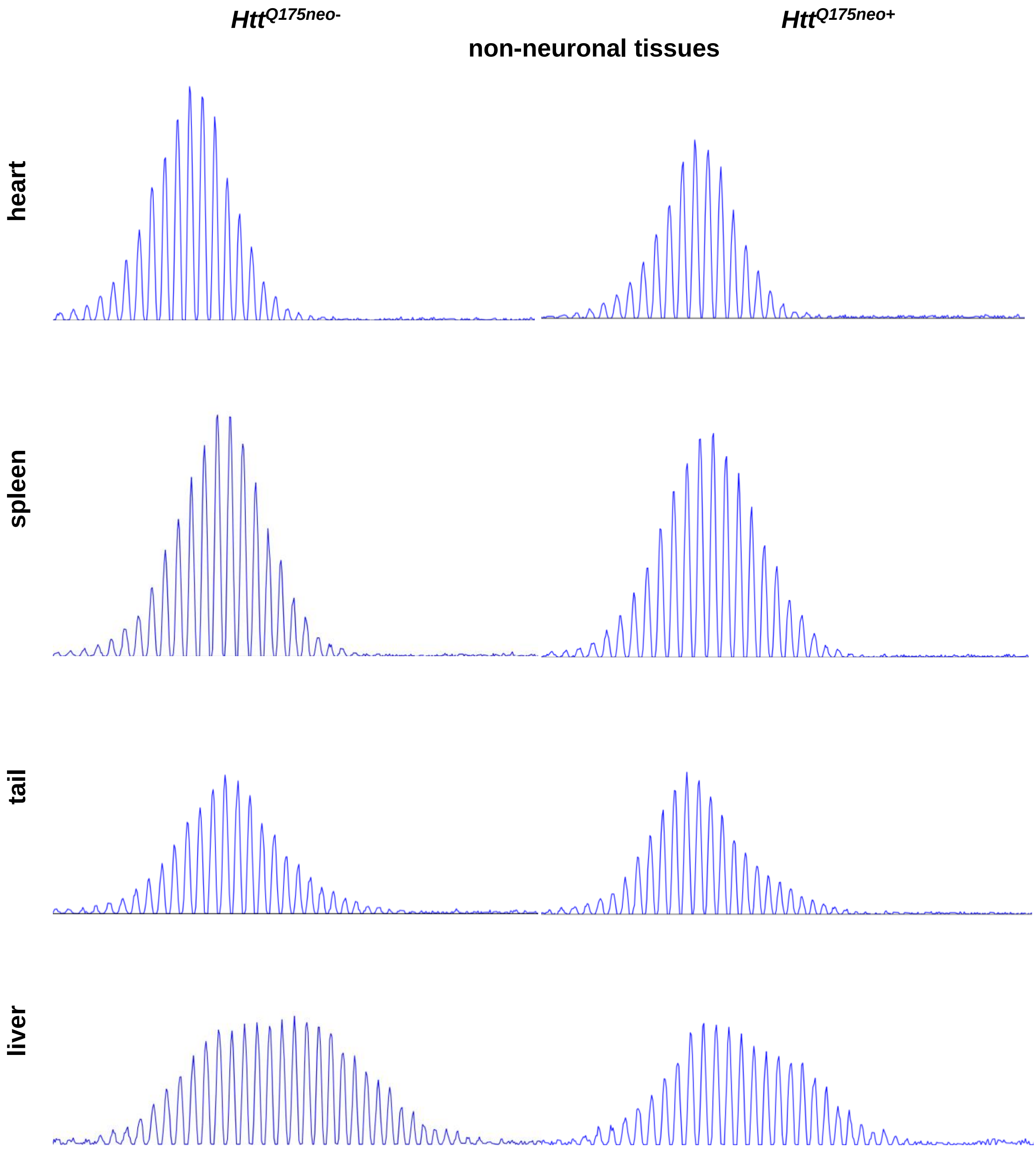
4.3. Statistical analyses

Comparisons between tissues' absolute instability, expansion indexes and contraction indexes within the $Htt^{Q175neo-}$ and $Htt^{Q175neo+}$ lines were evaluated through one-way ANOVA analyses separated in two distinct groups of neuronal and non-neuronal tissues. The Brown-Forsythe test was utilized to verify homogeneity of variances within groups. When variances were significantly different, data transformations were applied – in the $Htt^{Q175neo-}$ line for absolute instability, expansion indexes of non-neuronal tissues, and for contraction indexes of neuronal tissues reciprocal transformations (1/x) were performed, while in the $Htt^{Q175neo-}$ line a power transformation (x^2) was applied – allowing variances' homogenization and applicability of the ANOVA. Specific differences between the tissues were evaluated with the post-hoc Tukey's multiple comparisons test with multiplicity adjusted p-values. Both analyses were performed with Graphpad Prism v6.01.

Differences between the two lines ($Htt^{Q175neo-}$ and $Htt^{Q175neo+}$), including differences in main allele repeat sizes in heart and cerebellum, as well as expansion/contraction indexes and absolute instability in tail, heart, spleen, liver, cerebellum, cortex, and striatum were evaluated with unpaired Student's t-test for mean comparison, also using Graphpad Prism v6.01.

4.4. Results

Representative traces for each tissue in the *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ lines are shown in Figure 24. Overall, in a qualitative analysis liver and striatum stand-out, showing a wide-range distribution of alleles, typically seen in samples with high instability levels while the other tissues show a much more concentrated distribution.



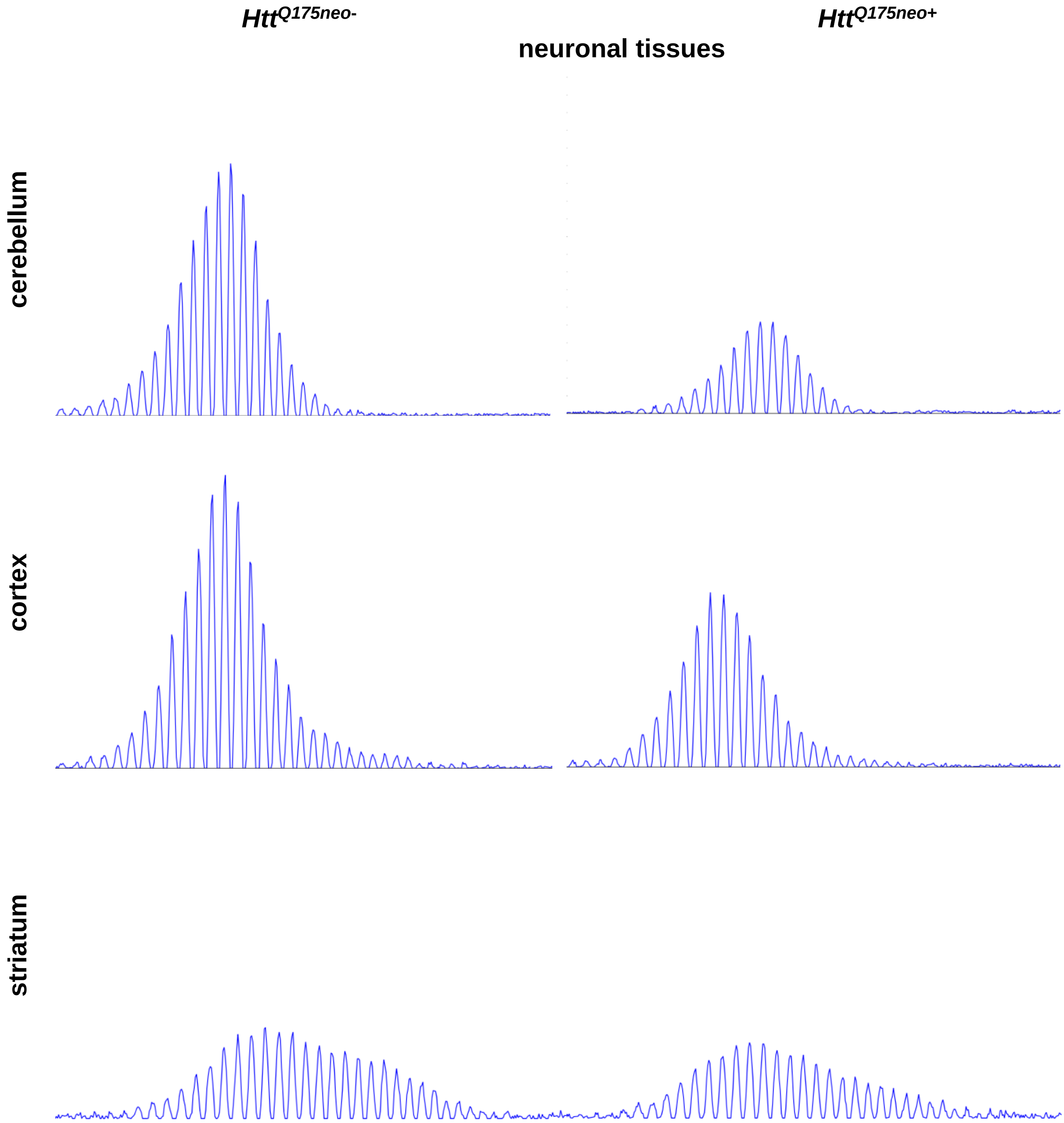


Figure 24 – Representative electropherogram traces of CAG repeat sizing in neuronal and non-neuronal tissues of *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ mice. In the x-axes CAG repeat size is represented, with peaks differing by one CAG unit. Signal height is represented in the y-axes. Liver and striatum traces present a 2× zoom-in in signal height when compared to all other tissues.

Quantitative analyses of instability were performed to determine expansion and contraction indexes as well as absolute instability (detailed information in Methods and Supplementary material).

First, we evaluated specific differences between tissues, separated in two groups: neuronal (cerebellum, cortex and striatum); and non-neuronal (heart, spleen, tail and liver) within both *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ lines.

Expansion indexes and absolute instability show a fairly straightforward picture (Figure 25, Figure 27), where among non-neuronal tissues, liver always shows significantly higher expansion indexes and absolute instability when compared to heart, spleen or tail, which, in this order, seem to show either trending or statistically significant increasing instability in both lines. Among neuronal tissues, results for expansions and instability are also clear, with striatum showing significantly more instability than cerebellum and cortex, while they appear to only differ from each other in *Htt*$^{Q175neo-}$ mice.

Regarding contractions (Figure 26), *Htt*$^{Q175neo-}$ shows fewer differences between tissues than in the other indexes, with only striatum and liver showing significant changes relative to other neuronal and non-neuronal tissues respectively. In *Htt*$^{Q175neo+}$ only heart and liver differ significantly in their contraction index.

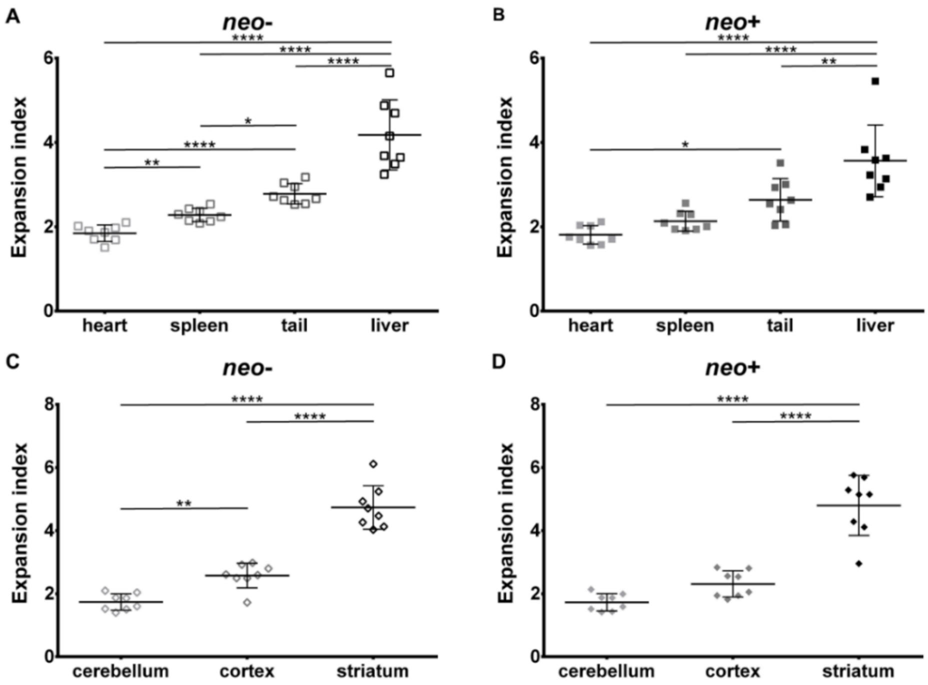


Figure 25 – Expansion indexes in (A, B) non-neuronal and (C, D) neuronal tissues in *HttQ175*$^{neo-}$ and *HttQ175*$^{neo+}$ mice. (*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, multiplicity adjusted)

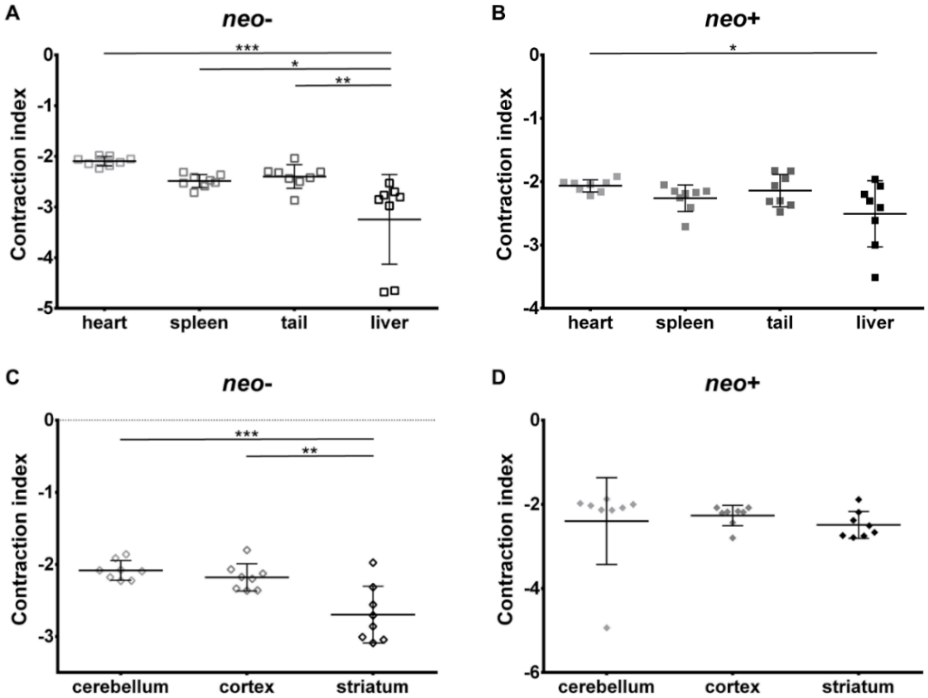


Figure 26 – Contraction indexes in (A, B) non-neuronal and (C, D) neuronal tissues in *HttQ175^{neo-}* and *HttQ175^{neo+}* mice. (*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, multiplicity adjusted)

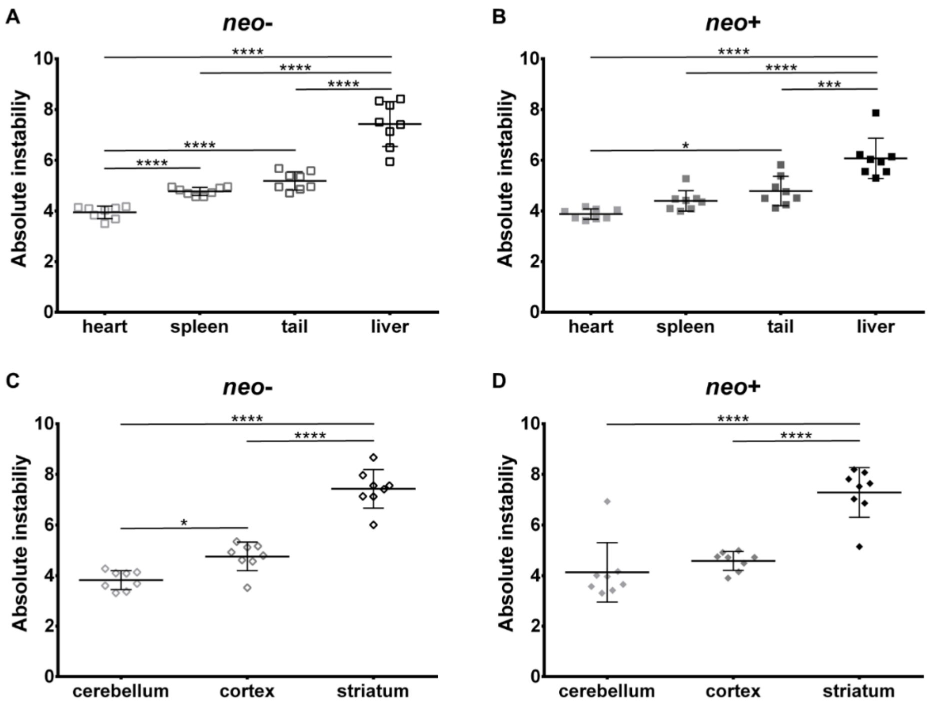


Figure 27 – Absolute instability in (A, B) non-neuronal and (C, D) neuronal tissues in *HttQ175^{neo-}* and *HttQ175^{neo+}* mice. (*p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, multiplicity adjusted)

We then evaluated if there were any differences in expansions, contractions or absolute instability between the two lines, in all tissues, in order to assess specific alterations stemming from the presence/absence of the *neo* cassette.

As stated previously, instability is influenced by CAG repeat size, and therefore differences in repeat length between these series of Q175 mice might influence the comparison (at hand). Therefore we evaluated if there were differences in the CAG size of the main alleles between the groups in their most stable tissues, heart and cerebellum (Figure 28).

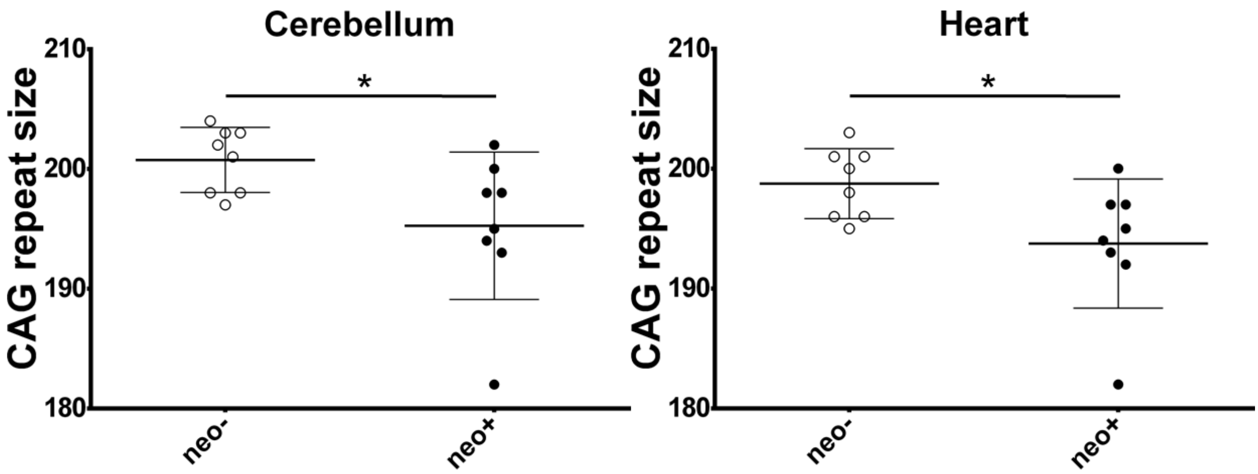


Figure 28 – Comparison of main allele CAG repeat sizes in cerebellum and heart for eight mice in the *HttQ175^{neo-}* and *HttQ175^{neo+}* lines. (*p<0.05)

Indeed, the two groups showed significant differences in both heart and cerebellum (difference between means of ~5 CAG units, heart: p=0.0368, cerebellum: p=0.0365). To balance CAG sizes between the lines, the two mice possessing the highest allele sizes in *Htt$^{Q175neo-}$* (Supplementary tables, mice E and H), as well as the two mice showing the shortest sizes in *Htt$^{Q175neo+}$* (Supplementary tables, mice I and J) were excluded to avoid biases driven by CAG repeat differences.

When the groups, now containing six mice, were compared regarding average repeat size in heart and cerebellum, they showed non-statistically significant differences (≤2 CAG units, heart: p=0.271, cerebellum: p=0.236, Figure 29), therefore these CAG size-controlled groups were considered for the downstream analysis.

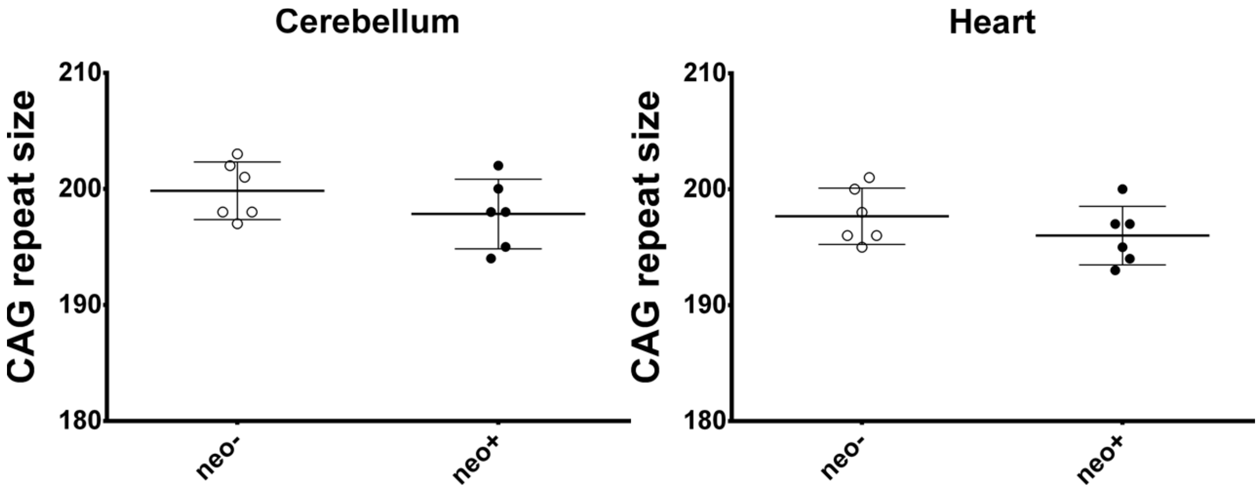


Figure 29 – Comparison of main allele CAG repeat sizes in cerebellum and heart for six mice in the *HttQ175*$^{neo-}$ and *HttQ175*$^{neo+}$ lines.

When comparing either expansion, contraction indexes or absolute instability between the lines in the seven tissues of interest – heart, spleen, tail, liver, cerebellum, cortex and striatum – most comparisons showed non-significant differences (Supplementary figures). Even though not significant, all expansion indices consistently showed lower values in the *Htt*$^{Q175neo+}$ line for all tissues when compared to their *neo*- counterparts.

Nonetheless, the liver's absolute instability showed the only significant difference between the two lines (p=0.0321, Figure 30).

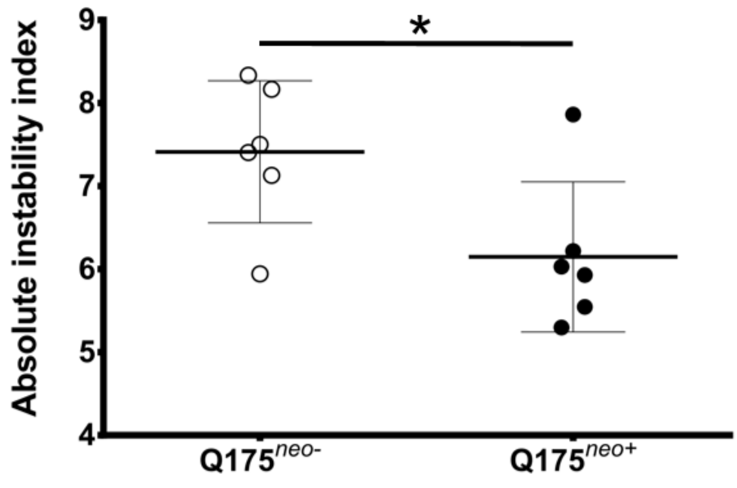


Figure 30 – Liver absolute instability in *HttQ175*$^{neo-}$ and *HttQ175*$^{neo+}$. (*p<0.05)

Keeping in mind that the delineation of the main allele influences the calculations of instability indexes, and that liver and striatum are the most unstable tissues where it is sometimes difficult to definitely assign the main allele (as seen in Figure 24), a different approach was taken to avoid mis-estimation of differences between the lines stemming from this fact. Expansion/contraction indexes and absolute instability were then re-calculated for liver and striatum, where the main allele considered for each mouse, was the highest represented in the most stable tissue of the same category, namely, cerebellum for striatum and heart for liver.

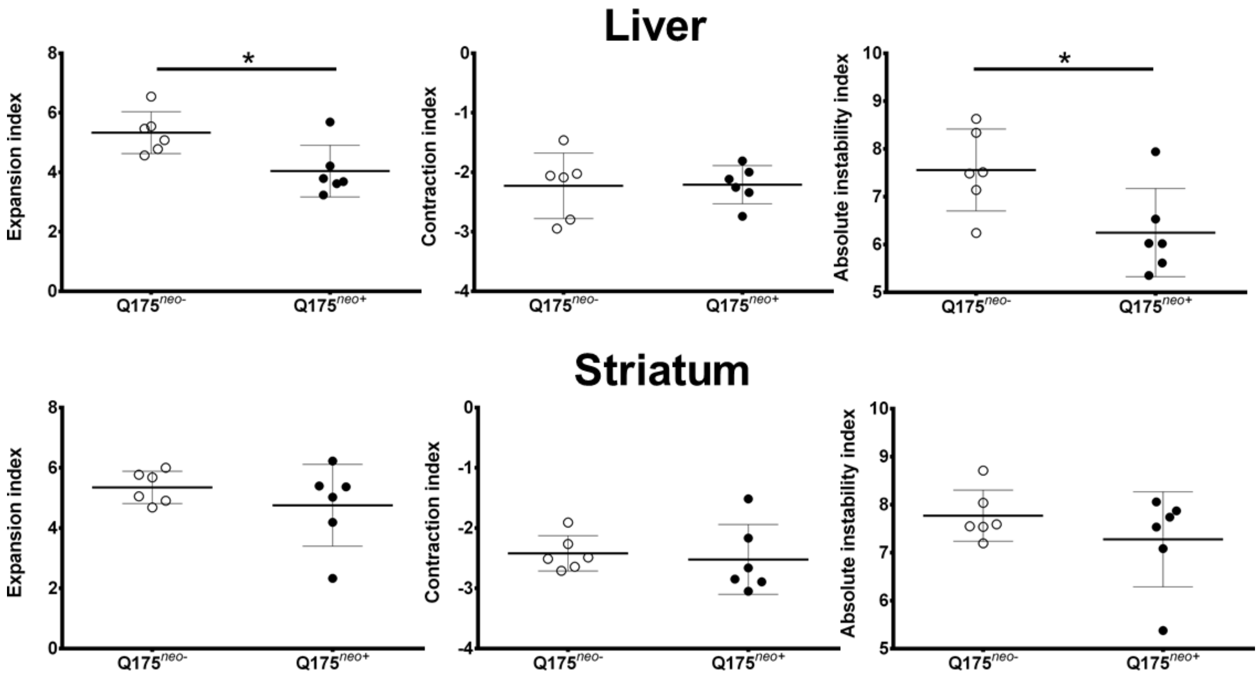


Figure 31 – Absolute instability, expansion and contraction indexes, with correction for main CAG allele size in liver and striatum of *HttQ175*$^{neo-}$ and *HttQ175*$^{neo+}$. (*p<0.05)

These changes showed to a small extent a different picture regarding differences between the *Htt*$^{Q175neo-}$ and *Htt*$^{Q175neo+}$ lines, with liver now showing significant differences regarding the expansion index (p=0.0178), and a still (but slightly more) significant difference in absolute instability (p=0.0289, Figure 31). Striatum still did not show any significant differences (Figure 31).

4.5. Discussion

In this work we, for the first time, thoroughly describe how tissues from *Htt*^{Q175} mice differ in CAG repeat instability and determine whether the presence of a *neo* sequence upstream of the repeat modulates somatic CAG repeat instability.

Regarding differences among tissues, Q175 (both *neo*- and *neo*+) show the most instability in striatum (which is also observed in post-mortem brains of patients) and liver, in agreement with instability levels observed in other models such as *Htt*^{Q111} and R6 mice[81,97,102,132]. All of the other tissues evaluated were far more stable, but surprisingly, tail which is often considered a staple among stable tissues in other models, did not appear the most stable even among non-neuronal tissues, consistently showing significantly higher expansion indexes and absolute instability specially when compared to heart. Among neuronal tissues, as expected, cerebellum was the most stable and cortex showed intermediate levels of instability. Expansion indexes appeared to be more dissimilar between tissues (and therefore driving differences between absolute instabilities) than contraction indexes, and while both lines showed this, differences were more prevalent and displayed more statistical significance in the *neo*- line.

Concerning the effect of the *neo* insert upstream of the repeat (in instability), in the work described in the previous chapter we have shown a “protective” effect of the *neo* sequence in intergenerational instability, where it seemed to play a role in reducing the frequency of expansions in parent-to-offspring transmissions. We therefore decided to evaluate if it might have a similar effect on somatic instability.

In the series of seven tissues evaluated, most showed no significant differences in either expansion index, contraction index or absolute instability between the *neo*+ and *neo*+ mice. Nonetheless, focusing on expansion indexes, we always see lower indices in the *Htt*$^{Q175neo+}$ line when compared with *Htt*$^{Q175neo-}$ consistent with a potential, although tenuous, overall protective effect on expansions.

Liver showed the only significant differences between lines; in a first analysis only the absolute instability was different between the strains, and after correcting for main allele size both expansion index and absolute instability showed significant differences. This correction for the main allele size seems a fairly sensible alteration to make, especially in the more unstable tissues, as their peak spread is large making it easy to confound the “ancestral” main allele with an allele size that became more prevalent exactly due to a high overall instability.

In conclusion, here we show that *cis*-factors, namely this *neo* sequence upstream of the repeat, can indeed modulate somatic instability, through the reduction of expansion indexes which ultimately affect instability levels.

4.6. Supplementary material

4.6.1. Supplementary tables

Table S8 – Instability measures for heart

Mouse line	Mouse ID	heart			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	201	1.905	-2.184	4.089
	B	196	1.704	-1.981	3.685
	C	195	2.104	-2.061	4.164
	D	198	1.511	-1.988	3.499
	E	203	1.690	-2.147	3.837
	F	200	2.016	-2.121	4.137
	G	196	1.869	-2.236	4.106
	H	201	1.975	-2.052	4.027
Htt$^{Q175neo+}$	I	182	2.037	-2.027	4.064
	J	192	1.564	-2.161	3.725
	K	195	2.022	-2.013	4.035
	L	193	1.573	-2.117	3.690
	M	197	1.710	-2.017	3.727
	N	200	1.699	-1.920	3.619
	O	194	1.756	-2.221	3.977
	P	197	2.119	-2.043	4.162

Table S9 – Instability measures for spleen

Mouse line	Mouse ID	spleen			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	202	2.238	-2.714	4.952
	B	198	2.375	-2.312	4.687
	C	197	2.297	-2.527	4.824
	D	199	2.084	-2.474	4.558
	E	203	2.146	-2.411	4.557
	F	201	2.426	-2.517	4.943
	G	198	2.539	-2.366	4.905
	H	203	2.133	-2.587	4.719
Htt$^{Q175neo+}$	I	182	2.297	-2.190	4.487
	J	192	2.009	-2.409	4.417
	K	197	2.562	-2.710	5.272
	L	194	1.940	-2.051	3.991
	M	197	2.093	-2.255	4.348
	N	201	1.944	-2.151	4.095
	O	195	1.909	-2.171	4.080
	P	199	2.315	-2.150	4.466

Table S10 – Instability measures for tail

Mouse line	Mouse ID	Tail			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	202	2.948	-2.437	5.384
	B	197	2.716	-2.868	5.584
	C	197	3.175	-2.495	5.671
	D	201	2.63	-2.32	4.949
	E	203	2.531	-2.422	4.952
	F	201	2.663	-2.039	4.703
	G	198	3.043	-2.3	5.343
	H	204	2.547	-2.308	4.856
Htt$^{Q175neo+}$	I	183	2.412	-1.832	4.244
	J	192	2.632	-2.31	4.943
	K	195	3.514	-2.302	5.817
	L	199	2.052	-2.063	4.115
	M	202	2.931	-1.829	4.76
	N	196	3.006	-2.37	5.375
	O	194	2.546	-1.95	4.496
	P	199	2.031	-2.477	4.509

Table S11 – Instability measures for liver

Mouse line	Mouse ID	liver			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	205	3.489	-4.677	8.166
	B	198	4.152	-2.977	7.129
	C	201	3.683	-4.651	8.334
	D	201	3.241	-2.702	5.943
	E	202	5.651	-2.765	8.416
	F	200	4.695	-2.805	7.500
	G	197	4.875	-2.529	7.404
	H	203	3.645	-2.853	6.498
Htt$^{Q175neo+}$	I	182	3.139	-2.999	6.138
	J	193	2.942	-2.612	5.554
	K	196	3.626	-2.304	5.930
	L	193	3.228	-2.070	5.297
	M	197	3.832	-2.198	6.031
	N	201	5.453	-2.409	7.862
	O	194	3.583	-1.963	5.546
	P	199	2.704	-3.514	6.218

Table S12 – Instability measures for cerebellum

Mouse line	Mouse ID	cerebellum			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	203	1.400	-1.914	3.314
	B	198	1.510	-1.859	3.369
	C	197	1.857	-2.228	4.085
	D	201	1.870	-2.226	4.096
	E	203	1.523	-2.076	3.599
	F	202	2.095	-2.178	4.273
	G	198	2.038	-2.095	4.133
	H	204	1.602	-2.083	3.685
Htt$^{Q175neo+}$	I	182	2.135	-2.031	4.166
	J	193	1.431	-1.877	3.308
	K	198	1.592	-1.978	3.571
	L	195	1.514	-2.134	3.648
	M	198	1.871	-2.139	4.009
	N	202	1.416	-2.002	3.418
	O	194	1.874	-2.082	3.957
	P	200	1.993	-4.939	6.932

Table S13 – Instability measures for cortex

Mouse line	Mouse ID	cortex			
		Average main allele	Expansion index	Contraction index	Absolute instability
Htt$^{Q175neo-}$	A	200	2.919	-2.199	5.118
	B	196	2.490	-2.126	4.616
	C	196	2.980	-2.364	5.344
	D	199	1.721	-1.802	3.523
	E	201	2.612	-2.176	4.788
	F	200	2.591	-2.334	4.925
	G	197	2.494	-2.069	4.563
	H	201	2.800	-2.361	5.161
Htt$^{Q175neo+}$	I	182	2.559	-2.165	4.724
	J	191	1.817	-2.082	3.899
	K	195	2.805	-2.186	4.990
	L	193	2.052	-2.435	4.488
	M	198	1.939	-2.211	4.150
	N	200	2.834	-2.082	4.916
	O	195	2.538	-2.180	4.718
	P	198	1.942	-2.800	4.743

Table S14 – Instability measures for striatum

Mouse line	Mouse ID	striatum			
		Average main allele	Expansion index	Contraction index	Absolute instability
$Htt^{Q175neo-}$	A	204	4.469	-3.090	7.559
	B	198	6.113	-2.559	8.672
	C	198	5.240	-2.316	7.556
	D	202	4.127	-3.008	7.134
	E	203	4.031	-1.977	6.008
	F	203	4.707	-2.708	7.415
	G	200	4.924	-3.045	7.969
	H	205	4.264	-2.860	7.124
$Htt^{Q175neo+}$	I	183	4.109	-2.751	6.861
	J	193	5.685	-2.508	8.192
	K	198	5.148	-2.668	7.817
	L	195	4.285	-2.742	7.027
	M	199	5.142	-2.381	7.523
	N	202	5.286	-2.790	8.076
	O	195	5.760	-1.886	7.645
	P	198	2.955	-2.189	5.144

4.6.2. Supplemental figures

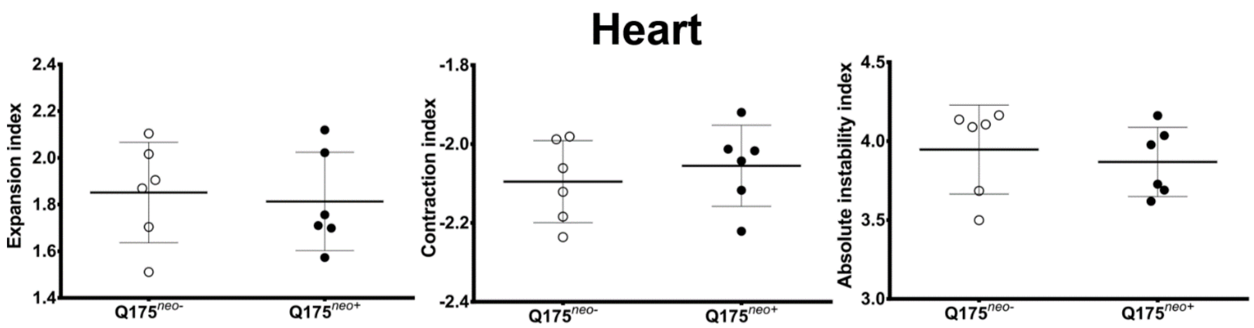


Figure S17 – Comparison of instability measures in heart

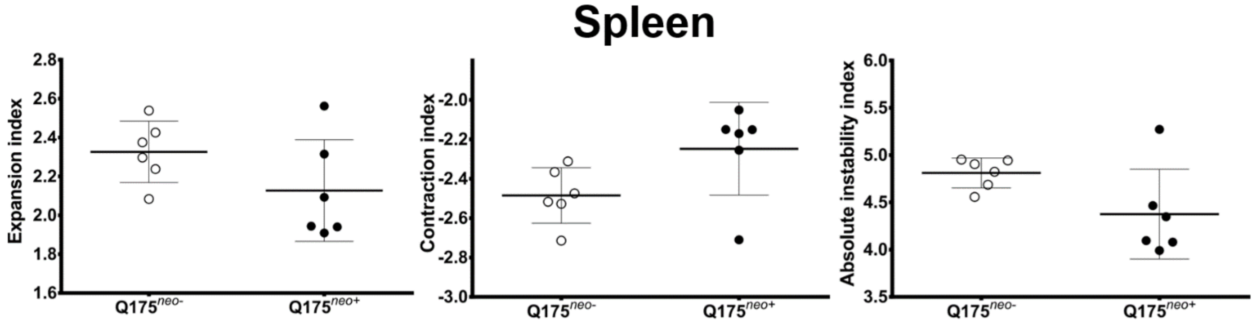


Figure S18 – Comparison of instability measures in spleen

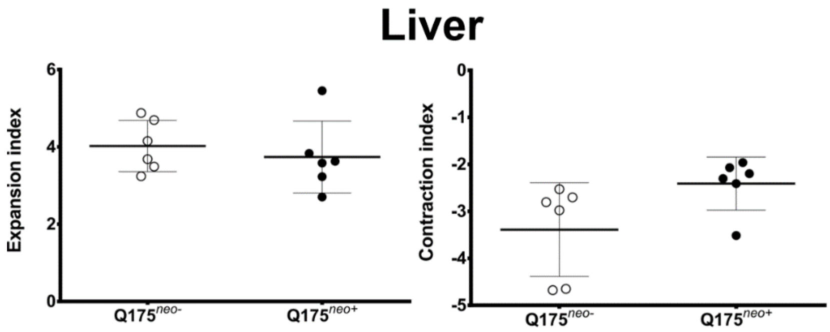


Figure S19 – Comparison of instability measures in liver

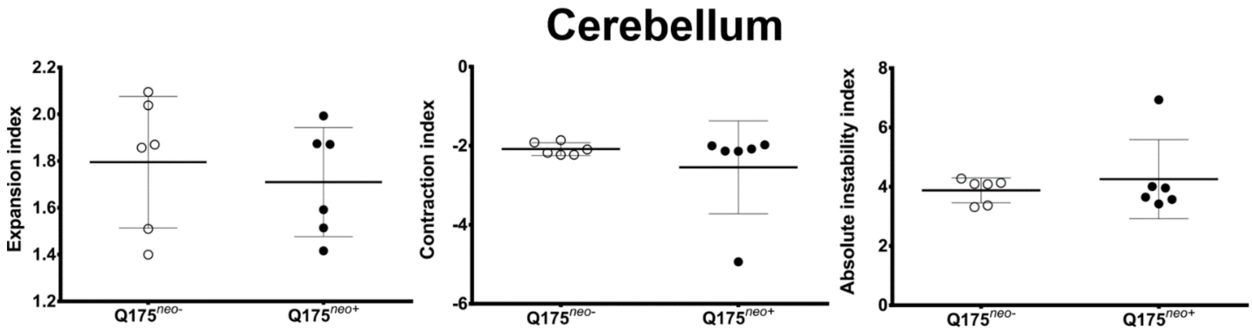


Figure S20 – Comparison of instability measures in cerebellum

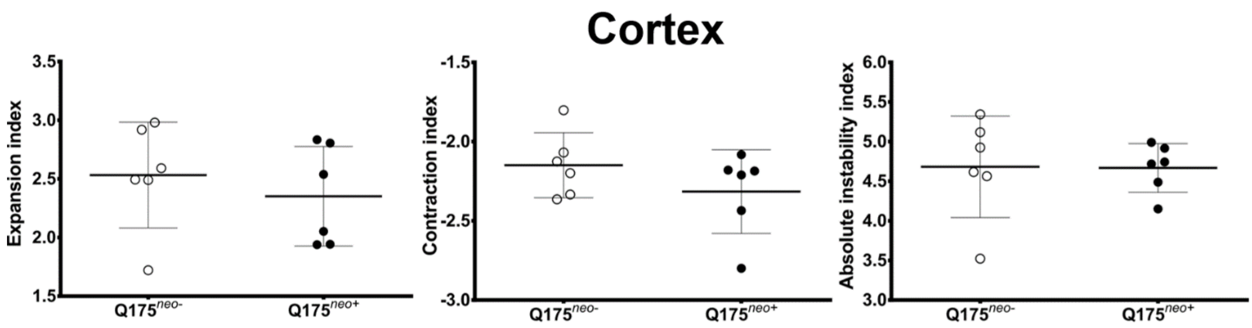


Figure S21 – Comparison of instability measures in Cortex

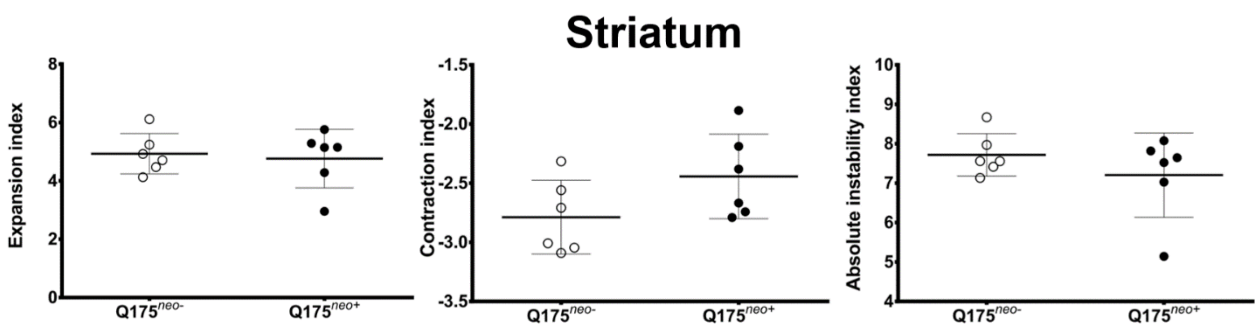


Figure S22 – Comparison of instability measures in striatum

5. General discussion and concluding remarks

HD is an autosomal dominant progressive neurodegenerative disorder caused by a CAG trinucleotide repeat expansion in exon 1 of the *HTT* gene located in 4p16.3[19]. The size of the repeat is of extreme importance as it determines the penetrance of the disorder, and is a main driver of age-at-onset[14,19,24,33,35,51,52]. In a typical presentation of HD one of the features where repeat size seems to have less of an impact is in disease duration, which appears to be mostly constant[52], nonetheless, since higher CAG sizes result in earlier onset, longer CAGs will on average result in lower life expectancy.

Changes in the number of CAGs, also referred to as repeat instability, either from one generation to another – intergenerational instability – or within somatic cells/tissues of an individual – somatic instability – has been identified in both patients and mouse models of the disorder[19,69,78–81]. Knowing the impact repeat size has on aspects of the disease, understanding the factors that may modulate it, either to prevent increases in size – expansions – in order to avoid increased penetrance and earlier onset, or induce changes toward small sizes – contractions – toward delaying the disorder is critical.

There also seems to exist a role for instability itself as a component of onset variation, as individuals with similar inherited repeat sizes but different degrees of instability differ greatly in age-at-onset, with larger somatically expanded repeats being associated with an earlier onset of disease[72]. Furthermore, an extensive GWAS recently showed genomic regions responsible for modulation of age-at-onset where DNA repair was one of the most represented pathways[54]. Members of the mismatch DNA repair pathway have been identified as modifiers of instability in mouse models of HD[90,102–104], conceivably validating the overall importance of CAG repeat instability in the disorder.

With this in mind, all the projects presented in this thesis were performed toward a better understanding of CAG repeat instability in HD, both somatic and intergenerational, which included human-derived cell models, patient germline samples and mouse models. This overall discussion therefore aims to connect the findings in the projects previously presented and how they might contribute to the state of the art regarding instability in HD, and ultimately contribute to help patients or individuals at-risk for the disorder.

Four cell lines from individuals of a nuclear family presenting an unusual level of intergenerational instability, with three offspring presenting large repeat size changes when compared to the affected father were studied with two goals: to assess if the father showed somatic instability with the same propensity for repeat size change; and to characterize the behavior and test the adequacy of using LCLs as easy to handle, human-derived models for high-throughput studies of instability modulators. Both of these goals were achieved to the extent possible.

The father prone to intergenerationally unstable transmissions did not show a particularly high instability in LCLs, possibly indicating that mechanisms of somatic and intergenerational instability might not be shared. However, a more systematic approach with a larger sample number seems to dispel this notion and indicate that absolute instability levels between LCL and sperm cells have a mild association, and expansion indexes show a particularly high correlation. The identification of this relationship using human samples could be especially informative for individuals with repeats in the intermediate and reduced penetrance ranges, as there is general uncertainty regarding the probability of the repeat changing toward the fully penetrant sizes. Somatic evaluation of the expansion index might serve as an indicator of propensity toward germline changes in repeat size.

While the association described above should still be validated in a higher number of subjects, there are previous observations in HD mouse models, namely, among the most somatically unstable lines of R6 models, that also seemed to show the highest intergenerational changes[81]. Further, in studies searching for instability modifiers, most of the identified genes seem to affect both types of instability[90,92,102,103], showing that genetic background effects might be acting in both somatic cells and germline, strengthening our observations.

Considering the observed relationship between somatic and intergenerational instability, there is the need to explain the exception to this trend in the previously mentioned father, with low somatic instability but high intergenerational instability. A possible reason for the exceptional level of intergenerational CAG repeat changes coming from this subject is a high level of mosaicism between different tissues, where spermatogonia, and therefore the entire germline, might possess a substantially higher repeat size number then the lymphocyte-derived cells lines studied, which would explain this trend toward large changes.

A relationship between CAG repeat size and instability was a common feature throughout the projects performed.

It was observed in the LCLs from the family mentioned above, during the effort to characterize their instability behavior that the individuals with lower CAG sizes showed very limited changes in main allele sizes and only the offspring with longer repeats seemed to show appreciable levels of instability, although with a very complex pattern to evaluate. These facts mainly lead to the exclusion of the possibility of LCLs broadly being used as high throughput models for instability, as the “best case scenario”, which would be observing instability at the lower end of pathogenic repeat sizes (40-50 CAGs), and therefore more informative and most directly applicable to a typical affected individual, did not come to fruition. Furthermore, the complicated behavior of the cell lines that do show instability, with multiple populations that vary due to non-controllable effects, show large jumps in repeat size without an explainable cause, and a lack of overall consistency would not make them suitable models. These unstable lines do show one common observable characteristic; a cohesive progression of main allele size with passages within specific populations, meaning that if one were to successfully isolate a single LCL population with complete certainty of homogeneity, strongly control for any systematic changes, and try to diminish the effect and/or account for random variation. This approach might be feasible but extremely technically challenging and still enormously hard to validate. For less, or at most, the same amount of effort one could effectively pursue other avenues including studying instability in patient neuronal cell cultures which may be derived from patient iPSCs which are already available[133], and would be a probably more adequate and important model for the disorder and potentially instability.

Dependence of instability on CAG was also observed when in the set of LCLs spanning a broad range of sizes (17-82 CAGs), with absolute instability, expansion and contraction indexes all showing a high correlation with repeat size.

CAG repeat length also showed a very prominent effect on intergenerational instability in knock-in mouse models of HD, where it affected the frequency of expansions as well as the magnitude of both expansions and contractions, especially when considering lines with underlying larger repeats (Htt^{Q111} and longer).

Therefore great care was undertaken to appropriately evaluate instability, always controlling for this prominent role of CAG size. In the LCL samples spanning the broad repeat range, CAG-independent instability was evaluated through the calculation of residual absolute instability, residual expansion and residual contraction values, highlighting the difference between observed values and the expected instability in case CAG was the only contributing component. In the study of intergenerational changes in mouse models, performing the comparison between distinct strains while adjusting for different average and range of repeat sizes was more technically challenging, and a methodology – that included theoretical modelling of expected frequency distributions contingent to a reference strain – had to be developed in order for a fair amount of confidence to be present in this analysis. In other cases such as the evaluation of the *neo* insert in both intergenerational and somatic instability, ultimately the ranges had to be adapted and some samples ended up being excluded, resulting in the evaluation of a lower number of observations and a consequential loss of statistical power to detect differences, but a greater confidence that the effects found were real and were not by CAG repeat size effects.

A significant effort was performed to try to pinpoint specific genes that correlated with the different measures of instability. Regarding somatic changes, this was evaluated by testing the correlations between CAG-independent instability and the expression levels of genes involved in DNA replication and repair, which, based on previous studies, are the pathways most likely to harbor instability modifier genes. Only two genes, *NTHL1* and *POLD1,* presented a significant correlation after multiple testing correction, providing candidates that modulate contractions. If we also consider nominally significant correlations we do see additional interesting genes notably, *TP73* expression correlates with absolute instability and expansions, and *FAN1* and *LIG1* correlating with contractions.

Concerning intergenerational instability, we identified differences between frequency and magnitude of expansions, contractions or unchanged transmissions among different mouse strains. To accurately pinpoint which specific genetic variants would be responsible for these differences, a thorough linkage mapping effort should be undertaken as has been done for somatic instability in the B6 and 129 strains[102]. However, we also found differences in instability between B6J and B6N strains, the most genetically similar among the background strains compared. It may therefore be easier to pinpoint a genetic causal differences as their overall number is more limited than when equating any other line. As stated previously no coding or regulatory SNPs in MMR genes or genes in other DNA repair pathways that might explain the differences between B6 and B6J strains. However, if the analysis is expanded even further, by the inclusion of not only SNPs but also indel variants, or by evaluating genome-wide differences in the Mouse Phenome Database (http://phenome.jax.org/), two interesting candidates are found: 1) an insertion/deletion of one base in the coding region of *Lig1* resulting in a frameshift mutation; and 2) a SNP in B6N causing a stop codon gain (therefore a non-sense mutation) in *Spata31*, a spermatogenesis associated gene, which is of particular notice since our evaluation of differences between strains was based on paternal transmissions.

So, overall we can use the information here gathered to evaluate the viability of using these genes as targets for the reduction of expansions or induction of contractions, the most eventual useful scenarios for patients and people at-risk for HD.

NTHL1 encodes endonuclease III-like protein 1, a protein that participates in base excision repair of oxidative lesions. We see that low levels of *NTHL1* expression are correlated with contractions. Following this path might be problematic, as diminishing the expression levels should be done with enough precision not to abolish expression and function, as loss-of-function mutations in *NTHL1* are responsible for different types of tumors[134,135]. Knockout models for the mouse ortholog *Nth1* also show other consequences such as increased telomere fragility[136]. Additionally, other studies in mice show that the lack of *Nth1* might reduce protection from oxidative DNA damage, when other DNA repair machinery involved in BER (namely, *Neil1*)[137], or even MMR (*Msh2*)[138] are also lacking. Lack of *Nth1* also diminishes protection from replication oxidative stress in a telomerase null background[136]. Therefore, even if done conservatively, modulating the expression of this gene might result in unplanned and unexpected consequences that might outweigh benefits.

POLD1 codes for the p125 subunit of polymerase δ (Polδ), which is known to be involved in both DNA replication and repair. The contraction modulation scenario is very similar to the described above; low expression levels of *POLD1* also correlate with contraction magnitude, but the disadvantages of wrongfully modulating the expression of this gene might be as severe or worse than *NTHL1*. Polδ is essential for DNA replication, and is extremely important for many of the DNA repair pathways, including MMR[139,140]. Downregulation of *POLD1* has been shown to arrest cell cycle and DNA synthesis and suppress cell proliferation[141], *POLD1* depletion causes genomic instability, observed as DNA breaks and chromosome abnormalities, and total loss of function from the catalytic domain has been associated with different types of cancer[142] and also a multisystem disorder[143]. Consequently, the pros and cons of modulating its expression should be strongly evaluated.

TP73 is responsible for the production of the p73 protein, from the p53 family. While it would have been problematic to reduce *TP73* expression, as knock-out mice have shown a phenotype of neurodegeneration[144], in the case of this gene an increase in expression levels seems to protect against expansions and absolute instability. *TP73* appears to have multiple main functions and important roles, among them, neuronal stem cell maintenance[145], and an anti-apoptotic role during neuronal development[146], both of which probably important in HD. Nevertheless, the inherent biology of *TP73* and p73 appears to be very complex, with multiple promoters and alternative splicing in play[145], where different isoforms seem to have distinct roles, namely in tumor suppression or tumor progression, depending on the isoform considered[147,148]. Thus, before trying to modulate repeat instability with *TP73*, a more detailed investigation would have to be performed, in order to understand if a specific isoform might be driving this expansion protection effect, and if so, confirm that this induced overexpression would not tilt the functional balance toward the oncogenic side.

FAN1 encodes FANCI-associated nuclease 1, whose main functions relate to DNA interstrand crosslink repair and processing of stalled replication forks[149,150]. Larger contractions seem to correlate with higher expression levels of *FAN1*, and while there is phenotypic characterization of the effects of *Fan1* knock-out in mice[151] and nuclease-defective knockins[149], possible harmful effects of overexpression are not clear. Nonetheless, it is important to note that *FAN1* is located in the most significant *locus* associated with residual age-at-onset in the recent GWAS[54] and assuming that instability might be the intermediate in this modulation, it is interesting to hypothesize that a naturally occurring variant might be responsible for increased expression and therefore an increase in contractions, altering age-at-onset.

LIG1 is also an interesting candidate, as several pieces of evidence indicate it may be a useful modifier. From the work here presented, lower levels of *LIG1* expression seem to correlate with contraction magnitude in LCLs. There is a significant difference between the B6J and B6N strains in terms of expansion and contraction frequency and magnitude in intergenerational transmissions, which can be due to a coding functional indel in the mouse ortholog *Lig1*. Furthermore, the literature indicates that *Lig1* expression levels are higher in cerebellum where the CAG repeat shows higher stability when compared to striatum which shows high instability[106]. Additionally, a study using a mouse model of DM1, a CTG repeat expansion-caused neuromuscular disorder that presents instability, identified, a defective *Lig1* (with ~3% of normal activity) that appeared to reduce the frequency of expansions and increase the frequency of contractions in female repeat transmissions (although we should keep in mind that no changes were observed in male transmissions or somatic changes, creating a more complex scenario)[110]. *LIG1* encodes ligase I, a protein involved in both DNA replication and repair. Interestingly, unlike the some of the genes mentioned above, knock-out mice for *Lig1* does not seem to present DNA repair impairment, possibly because of redundant functionality from other ligases. Rather, phenotypes of these mice seem to be related to problems in DNA replication and not repair[152]. With these observations in mind, there is a large opportunity to further explore the potential of *LIG1* as a modulator of instability, specifically in non-replicating cells such as neurons, as there are many factors indicating it might be beneficial without any predictable disadvantages. Specifically, it might be interesting to evaluate somatic expansion and contraction indexes, in replicating and non-replicating tissues, as well as intergenerational changes in HD mouse models expressing the low activity *Lig1* as mentioned in the DM1 study[110] to confirm the observations here stated.

Finally, *Spata31* should also be considered, but especially regarding a role in intergenerational instability among paternal transmissions, as the protein it produces, spermatogenesis-associated protein 31, appears not to be present in any tissue besides the male germline in mice. It is very hard to pinpoint how this factor might be influencing instability as the protein function is not clear. Our *in silico* analyses did not reveal any clear functional domain and studies in mice have only related it to fertility[153]. In humans, there is a large family of *SPATA31* genes with high homology between them, containing multiple isoforms which might have even acquired new functions, since the evolutionary deviation from mice[154], potentially complicating the translation of discoveries made in the mouse to applicability in patients.

Regarding intergenerational instability, a great effort was performed to thoroughly evaluate the largest breeding datasets available to date from HD knock-in mouse models. Beside the findings already mentioned in this discussion – the CAG dependence regarding frequency and magnitude of changes, and the genetic background effects evaluated between strains – due to the overlap with conclusions from other projects, there are other important observations also worth mentioning.

To begin with, we observed that the presence of an expanded CAG allele did not seem to skew the overall proportion of heterozygous and wild-type offspring independently of CAG size, showing different results from what has been observed in other diseases[118–121], but in accordance with previous results from single sperm genotype of HD-affected individuals, suggesting that the repeat size does not seem to be the inherent cause of the distortions observed.

Sex-of-parent effects, that were already thoroughly studied previously[60,63,64,78] were confirmed, namely the significant differences in the frequency of expansions, much higher in paternal transmissions, and frequency of contractions, that were predominant in maternal transmissions. These differences probably stem from intrinsic differences in gamete formation processes. Interestingly, magnitude of changes was not different between paternal and maternal transmissions, indicating that the differences in gametogenesis have a stronger role in determining the occurrence of the event, but does not overly influence the extent of the repeat change.

Sex-of-offspring effects were previously described both in humans and transgenic mouse models of the disorder[60,89], but we observed no such effects in any of the knock-in lines evaluated, regarding either frequency or magnitude of changes in paternal or maternal transmissions, suggesting that in knock-in mouse models, effects other than post-zygotic X or Y-encoded factors influence instability.

One more interesting result was the observation that *cis*-effectors, in this case in the form of a neomycin resistance cassette upstream of the repeat, alter the frequency of changes in intergenerational transmissions, by reducing the relative frequency of expansions, with a slight increase in both contraction and stable transmissions. The presence of *cis*-effects have been previously suggested in transgenic mouse models[81], but have not been described in HD knock-in models.

To complement this observation we also evaluated the impact of the upstream *neo* in seven tissues of Htt^{Q175} models, and only observed significant effects in liver, the most unstable among the peripheral tissues, in the same direction as intergenerational instability, where $Htt^{Q175neo+}$ mice show a protective effect regarding expansions.

A possible explanation might involve changes in DNA conformation stemming from the presence of the *neo* sequence through a short-range interaction with the CAG repeat sequence.

The contents present in this thesis do contribute to the state of the art in the field of CAG repeat instability in Huntington's disease. It shows that patient-derived cell lines, in this case LCLs, are an important and fruitful resource to use when evaluating instability, and even though they are on the easier to culture and maintain side of cell lines they might not be the most suited model to use for screenings (*e.g.* small molecule/drug screenings) as they show inherent complexity of CAG repeat size changes. It also focuses on the importance of checking and controlling for CAG size effects on instability in order to accurately describe different factors influencing changes. It proposes that germ line and somatic instability might share some common ground. It suggests genes to be studied more in depth as instability modifiers based on data gathered from patient derived cells. It includes a very substantial effort to thoroughly evaluate the contributors to intergenerational instability in knock-in mouse models of HD, confirming some previous observations and shedding light into new ones using the largest breeding datasets available to date, after careful curation and validation to ensure the highest possible data quality and results. Finally, it also transposes one of these new observations into somatic instability in an HD model. Overall the results here presented open some new hypothesis and point towards possible directions in future studies of *HTT* CAG repeat instability.

6. References

1. Huntington G. On chorea. Med Surg Rep. 1872;26: 317–321.

2. Ghosh R, Tabrizi SJ. Clinical Aspects of Huntington's Disease. Clin Asp Huntington's Dis. 2013; 289–320. doi:10.1007/7854

3. Novak MJU, Tabrizi SJ. Huntington's disease: Clinical presentation and treatment [Internet]. International Review of Neurobiology. Elsevier Inc.; 2011. doi:10.1016/B978-0-12-381328-2.00013-4

4. Molón LR. Juvenile Huntington ' s disease : a case report and literature review. Salud Ment. 2010;38: 285–294.

5. Costa MDC, Magalhães P, Ferreirinha F, Guimarães L, Januário C, Gaspar I, et al. Molecular diagnosis of Huntington disease in Portugal: implications for genetic counselling and clinical practice. Eur J Hum Genet. 2003;11: 872–878. doi:10.1038/sj.ejhg.5201055

6. Rawlins MD, Wexler NS, Wexler AR, Tabrizi SJ, Douglas I, Evans SJW, et al. The prevalence of huntington's disease. Neuroepidemiology. 2016;46: 144–153. doi:10.1159/000443738

7. Novak MJU, Tabrizi SJ. Huntington's disease. BMJ. 2010;341: 34–40. doi:10.1016/S0140-6736(07)60111-1

8. van der Burg JM, Björkqvist M, Brundin P. Beyond the brain: widespread pathology in Huntington's disease. Lancet Neurol. Elsevier Ltd; 2009;8: 765–774. doi:10.1016/S1474-4422(09)70178-4

9. Waldvogel HJ, Kim EH, Thu DC V, Tippett LJ, Faull RLM. New perspectives on the neuropathology in Huntington's disease in the human brain and its relation to symptom variation. J Huntingtons Dis. 2012;1: 143–153. doi:10.3233/JHD-2012-120018

10. Waldvogel HJ, Kim EH, Tippett LJ, Vonsattel J-PG, Faull RLM. The Neuropathology of Huntington's Disease. In: Nguyen HHP, Cenci MA, editors. Behavioral Neurobiology of Huntington's Disease and Parkinson's Disease. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015. pp. 33–80. doi:10.1007/7854_2014_354

11. Vonsattel JPG, Keller C, Cortes Ramirez EP. Huntington's disease - neuropathology [Internet]. 1st ed. Handbook of Clinical Neurology. Elsevier B.V.; 2011. doi:10.1016/B978-0-444-52014-2.00004-5

12. Reiner A, Dragatsis I, Dietrich P. Genetics and neuropathology of huntington's disease [Internet]. International Review of Neurobiology. Elsevier Inc.; 2011. doi:10.1016/B978-0-12-381328-2.00014-6

13. Walker FO. Huntington's disease. Lancet. 2007;369: 218–228. doi:10.1016/S0140-6736(07)60111-1

14. Gusella JF, MacDonald ME. Huntington's disease: seeing the pathogenic process through a genetic lens. Trends Biochem Sci. 2006;31: 533–540.

doi:10.1016/j.tibs.2006.06.009

15. Vonsattel J-P, Myers RH, Stevens TJ, Ferrante RJ, Bird ED, Richardson EP. Neuropathological Classification of Huntington's Disease. J Neuropathol Exp Neurol. 1985;44: 559–577. doi:10.1097/00005072-198511000-00003

16. Carroll JB, Bates GP, Steffan J, Saft C, Tabrizi SJ. Treating the whole body in Huntington's disease. Lancet Neurol. 2015;14: 1135–1142. doi:10.1016/S1474-4422(15)00177-5

17. Huntington Study Group. Unified Huntington's disease rating scale: Reliability and consistency. Mov Disord. 1996;11: 136–142. doi:10.1002/mds.870110204

18. Bates GP. The molecular genetics of Huntington disease — a history. Nature. 2005;6: 766–773. doi:10.1038/nrg1686

19. The Huntington's Disease Collaborative Research Group, MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72: 971–983. doi:10.1016/0092-8674(93)90585-E

20. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson M a, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306: 234–238. doi:10.1017/CBO9781107415324.004

21. Gusella JF. Genetic linkage of the Huntington's disease to a DNA Marker. Can J Neurol Sci. 1984;11: 421–425.

22. Bućan M, Zimmer M, Whaley WL, Poustka A, Youngman S, Allitto BA, et al. Physical maps of 4p16.3, the area expected to contain the Huntington disease mutation. Genomics. 1990;6: 1–15. doi:10.1016/0888-7543(90)90442-W

23. MacDonald ME, Novelletto A, Lin C, Tagle D, Barnes G, Bates G, et al. The Huntington's disease candidate region exhibits many different haplotypes. Nat Genet. 1992;1: 99–103. doi:10.1038/ng0592-99

24. Bates GP, Dorsey R, Gusella JF, Hayden MR, Kay C, Leavitt BR, et al. Huntington disease. Nat Rev Dis Prim. 2015; 15005. doi:10.1038/nrdp.2015.5

25. Li S-H, Schilling G, Young WS, Li X-., Margolis RL, Stine OC, et al. Huntington's disease gene (IT15) is widely expressed in human and rat tissues. Neuron. 1993;11: 985–993. doi:10.1016/0896-6273(93)90127-D

26. Strong T V, Tagle D a, Valdes JM, Elmer LW, Boehm K, Swaroop M, et al. Widespread expression of the human and rat Huntington's disease gene in brain and nonneural tissues. Nat Genet. 1993;5: 259–265. doi:10.1038/ng1193-259

27. Ambrose CM, Duyao MP, Barnes G, Bates GP, Lin CS, Srinidhi J, et al. Structure and expression of the Huntington's disease gene: Evidence against simple inactivation due to an expanded CAG repeat. Somat Cell Mol Genet. 1994;20: 27–38. doi:10.1007/BF02257483

28. Saudou F, Humbert S. The Biology of Huntingtin. Neuron. 2016;89: 910–926. doi:10.1016/j.neuron.2016.02.003

29. Landwehrmeyer GB, McNeil SM, Dure LS, Ge P, Aizawa H, Huang Q, et al. Huntington's disease gene: Regional and cellular expression in brain of normal and affected individuals. Ann Neurol. 1995;37: 218–230. doi:10.1002/ana.410370213

30. Arrasate M, Mitra S, Schweitzer ES, Segal MR, Finkbeiner S. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. Nature. 2004;431: 805–10. doi:10.1038/nature02998

31. Saudou F, Finkbeiner S, Devys D, Greenberg ME. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. Cell. 1998;95: 55–56. doi:10.1016/S0092-8674(00)81782-1

32. Kim M, Lee HS, LaForet G, McIntyre C, Martin EJ, Chang P, et al. Mutant huntingtin expression in clonal striatal cells: dissociation of inclusion formation and neuronal survival by caspase inhibition. J Neurosci. 1999;19: 964–73. Available: http://www.ncbi.nlm.nih.gov/pubmed/9920660

33. Duyao M, Ambrose C, Myers R, Novelletto a, Persichetti F, Frontali M, et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. Nat Genet. 1993;4: 387–392. doi:10.1038/ng0893-387

34. Snell RG, MacMillan JC, Cheadle JP, Fenton I, Lazarou LP, Davies P, et al. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. Nat Genet. 1993;4: 393–397. doi:10.1038/ng0893-393

35. Read AP. Huntington's disease: testing the test. Nat Genet. 1993;4: 329–330. doi:10.1038/ng0893-329

36. Andrew SE, Goldberg YP, Kremer B, Telenius H, Theilmann J, Adam S, et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet. 1993;4: 398–403. doi:10.1038/ng0893-398

37. Myers RH, MacDonald ME, Koroshetz WJ, Duyao MP, Ambrose CM, Taylor S a, et al. De novo expansion of a (CAG)n repeat in sporadic Huntington's disease. Nat Genet. 1993;5: 168–173. doi:10.1038/ng1093-168

38. Kremer B, Goldberg P, Andrew SE, Theilmann J, Telenius H, Zeisler J, et al. A Worldwide Study of the Huntington's Disease Mutation: The Sensitivity and Specificity of Measuring CAG Repeats. N Engl J Med. 1994;330: 1401–1406. doi:10.1056/nejm199405193302001

39. Bean L, Bayrak-Toydemir P. 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. 2014;16: e2. doi:10.1038/gim.2014.146

40. Losekoot M, van Belzen MJ, Seneca S, Bauer P, Stenhouse S a R, Barton DE.

EMQN/CMGS best practice guidelines for the molecular genetic testing of Huntington disease. Eur J Hum Genet. Nature Publishing Group; 2013;21: 480–486. doi:10.1038/ejhg.2012.200

41. Oosterloo M, Van Belzen MJ, Bijlsma EK., Roos RAC. Is There Convincing Evidence that Intermediate Repeats in the HTT Gene Cause Huntington's Disease? J Huntingtons Dis. 2015;4: 141–148. doi:10.3233/JHD-140120

42. Garcia-ruiz PJ, Garcia-caldentey J, Feliz C, Val J, Herranz A, Martínez-castrillo JC. Late onset Huntington's disease with 29 CAG repeat expansion. J Neurol Sci. Elsevier B.V.; 2016;363: 114–115. doi:10.1016/j.jns.2016.02.030

43. Semaka A, Creighton S, Warby S, Hayden MA. Predictive testing for Huntington disease: Interpretation and significance of intermediate alleles. Clin Genet. 2006;70: 283–294. doi:10.1111/j.1399-0004.2006.00668.x

44. Moore RC, Xiang F, Monaghan J, Han D, Zhang Z, Edström L, et al. Huntington Disease phenocopy is a Familial Prion Disease. Am J Hum Genet. 2001;69: 1385–1388. doi:10.1086/324414

45. Margolis RL, O'Hearn E, Rosenblatt A, Willour V, Holmes SE, Franz ML, et al. A disorder similar to Huntington's disease is associated with a novel CAG repeat expansion. Ann Neurol. 2001;50: 373–380. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11558794%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med4&NEWS=N&AN=11761463

46. Stevanin G, Brice A. Spinocerebellar Ataxia 17 And Huntington's Disease-like 4. Genetic Instabilities and Neurological Diseases, Second Edition. 2006. pp. 475–483. doi:10.1016/B978-012369462-1/50033-8

47. Hensman DJ, Poulter M, Beck J, Hehir J, Polke JM, Campbell T, et al. C9orf72 expansions are the most common genetic cause of Huntington disease phenocopies. Neurology. 2014;82: 292–299. doi:10.1212/WNL.0000000000000061

48. Dupré N, Fcrp C, Rouleau G, Frcp C. The Puzzle of Huntington Disease Phenocopies. JAMA Neurol. 2016;73(9): 1056–8. doi:10.1001/jamaneurol.2016.2215.3

49. Wild EJ, Tabrizi SJ. Huntington's disease phenocopy syndromes. Curr Opin Neurol. 2007;20: 681–687. doi:10.1097/WCO.0b013e3282f12074

50. Lee JM, Gillis T, Mysore JS, Ramos EM, Myers RH, Hayden MR, et al. Common SNP-based haplotype analysis of the 4p16.3 Huntington disease gene region. Am J Hum Genet. The American Society of Human Genetics; 2012;90: 434–444. doi:10.1016/j.ajhg.2012.01.005

51. Gusella JF, MacDonald ME. Genetic modifiers of Huntington's disease. Mov Disord. 2014;29: 1359–1365. doi:10.1002/mds.26001

52. Lee JM, Ramos EM, Lee JH, Gillis T, Mysore JS, Hayden MR, et al. CAG repeat

expansion in Huntington disease determines age at onset in a fully dominant fashion. Neurology. 2012;78: 690–695. doi:10.1212/WNL.0b013e318249f683

53. Wexler NS, Lorimer J, Porter J, Gomez F, Moskowitz C, Shackell E, et al. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. Proc Natl Acad Sci U S A. 2004;101: 3498–3503. doi:10.1073/pnas.0308679101

54. Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. Cell. 2015;162: 516–526. doi:10.1016/j.cell.2015.07.003

55. Ranen NG, Stine OC, Abbott MH, Sherr M, Codori AM, Franz ML, et al. Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. Am J Hum Genet. 1995;57: 593–602. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1801258&tool=pmcentrez&rendertype=abstract

56. Kremer B, Almqvist E, Theilmann J, Spence N, Telenius H, Goldberg YP, et al. Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes. Am J Hum Genet. 1995;57: 343–50. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1801544&tool=pmcentrez&rendertype=abstract

57. Hendricks AE, Latourelle JC, Lunetta KL, Cupples LA, Wheeler V, MacDonald ME, et al. 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 Part A. 2009;149: 1375–1381. doi:10.1002/ajmg.a.32901

58. Sequeiros J, Ramos EM, Cerqueira J, Costa MC, Sousa A, Pinto-Basto J, et al. 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. 2010;78: 381–387. doi:10.1111/j.1399-0004.2010.01388.x

59. Zühlke C, Riess O, Bockel B, Lange H, Thies U. Mitotic stability and meiotic variability of the (CAG)n repeat in the Huntington disease gene. Hum Mol Genet. 1993;2: 2063–7. doi:10.1093/hmg/2.12.2063

60. Wheeler VC, Persichetti F, McNeil SM, Mysore JS, Mysore SS, MacDonald ME, et al. Factors associated with HD CAG repeat instability in Huntington disease. J Med Genet. 2007;44: 695–701. doi:10.1136/jmg.2007.050930

61. Telenius H, Kremer HP, Theilmann J, Andrew SE, Almqvist E, Anvret M, et al. Molecular analysis of juvenile Huntington disease: the major influence on (CAG)n repeat length is the sex of the affected parent. Hum Mol Genet. 1993;2: 1535–40. doi:10.1093/hmg/2.10.1535

62. Trottier Y, Biancalana V, Mandel JL. Instability of CAG repeats in Huntington's disease: relation to parental transmission and age of onset. J Med Genet.

1994;31: 377–82. doi:10.1136/jmg.31.5.377

63. Aziz NA, Van Belzen MJ, Coops ID, Belfroid RDM, Roos RAC. Parent-of-origin differences of mutant HTT CAG repeat instability in Huntington's disease. Eur J Med Genet. Elsevier Masson SAS; 2011;54: e413–e418. doi:10.1016/j.ejmg.2011.04.002

64. Ramos EM, Cerqueira J, Lemos C, Pinto-Basto J, Alonso I, Sequeiros J. Intergenerational Instability in Huntington Disease: Extreme Repeat Changes Among 134 Transmissions. Mov Disord. 2012;27: 580–581. doi:10.1002/mds.24011

65. Myers RH. Huntington's disease genetics. NeuroRx. 2004;1: 255–62. doi:10.1602/neurorx.1.2.255

66. Warby SC, Montpetit A, Hayden AR, Carroll JB, Butland SL, Visscher H, et al. CAG Expansion in the Huntington Disease Gene Is Associated with a Specific and Targetable Predisposing Haplogroup. Am J Hum Genet. The American Society of Human Genetics; 2009;84: 351–366. doi:10.1016/j.ajhg.2009.02.003

67. Ramos EM, Gillis T, Mysore JS, Lee JM, Gögele M, D'Elia Y, et al. Haplotype analysis of the 4p16.3 region in Portuguese families with Huntington's disease. Am J Med Genet Part B Neuropsychiatr Genet. 2015;168: 135–143. doi:10.1002/ajmg.b.32289

68. Leeflang EP, Tavaré S, Marjoram P, Neal COS, Srinidhi J, MacDonald ME, et al. Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. Hum Mol Genet. 1999;8: 173–183. doi:10.1093/hmg/8.2.173

69. Telenius H, Kremer B, Goldberg Y. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. Nat Genet. 1994;6: 409–414. doi:10.1038/ng0494-409

70. Shelbourne PF, Keller-McGandy C, Bi WL, Yoon SR, Dubeau L, Veitch NJ, et al. Triplet repeat mutation length gains correlate with cell-type specific vulnerability in Huntington disease brain. Hum Mol Genet. 2007;16: 1133–1142. doi:10.1093/hmg/ddm054

71. Kennedy L, Evans E, Chen CM, Craven L, Detloff PJ, Ennis M, et al. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. Hum Mol Genet. 2003;12: 3359–3367. doi:10.1093/hmg/ddg352

72. Swami M, Hendricks AE, Gillis T, Massood T, Mysore J, Myers RH, et al. Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. Hum Mol Genet. 2009;18: 3039–3047. doi:10.1093/hmg/ddp242

73. Veitch NJ, Ennis M, McAbney JP, Shelbourne PF, Monckton DG. Inherited CAG·CTG allele length is a major modifier of somatic mutation length variability

in Huntington disease. DNA Repair (Amst). 2007;6: 789–796. doi:10.1016/j.dnarep.2007.01.002

74. Cannella M, Maglione V, Martino T, Ragona G, Frati L, Li G-M, et al. DNA instability in replicating Huntington's disease lymphoblasts. BMC Med Genet. 2009;10: 11. doi:10.1186/1471-2350-10-11

75. MacDonald ME, Barnes G, Srinidhi J, Duyao MP, Ambrose CM, Myers RH, et al. Gametic but not somatic instability of CAG repeat length in Huntington's disease. J Med Genet. 1993;30: 982–6. doi:10.1136/jmg.30.12.982

76. Pouladi M a, Morton a J, Hayden MR. Choosing an animal model for the study of Huntington's disease. Nat Rev Neurosci. Nature Publishing Group; 2013;14: 708–21. doi:10.1038/nrn3570

77. White JK, Auerbach W, Duyao MP, Vonsattel JP, Gusella JF, Joyner AL, et al. Huntingtin is required for neurogenesis and is not impaired by the Huntington's disease CAG expansion. Nat Genet. 1997;17: 404–10. doi:10.1038/ng1297-404

78. Wheeler VC, Auerbach W, White JK, Srinidhi J, Auerbach A, Ryan A, et al. Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. Hum Mol Genet. 1999;8: 115–122. doi:10.1093/hmg/8.1.115

79. Menalled LB, Sison JD, Dragatsis I, Zeitlin S, Chesselet MF. Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. J Comp Neurol. 2003;465: 11–26. doi:10.1002/cne.10776

80. Menalled LB, Kudwa AE, Miller S, Fitzpatrick J, Watson-Johnson J, Keating N, et al. Comprehensive Behavioral and Molecular Characterization of a New Knock-In Mouse Model of Huntington's Disease: ZQ175. PLoS One. 2012;7. doi:10.1371/journal.pone.0049838

81. Mangiarini L, Sathasivam K, Mahal A, Mott R, Seller M, Bates GP. Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. Nat Genet. 1997;15: 197–200. doi:10.1038/ng0297-197

82. Alexandrov V, Brunner D, Menalled LB, Kudwa A, Watson-Johnson J, Mazzella M, et al. Large-scale phenome analysis defines a behavioral signature for Huntington's disease genotype in mice. Nat Biotechnol. 2016; doi:10.1038/nbt.3587

83. Bragg R, Coffey SR, Weston RR, Ament SA, Cantle JP, Minnig S, et al. Motivational, proteostatic and transcriptional deficits precede synapse loss, gliosis and neurodegeneration in the B6.HttQ111/+ model of Huntington's disease. bioRxiv. 2016; 1–17. doi:10.1038/srep41570

84. Hölter SM, Stromberg M, Kovalenko M, Garrett L, Glasl L, Lopez E, et al. A broad phenotypic screen identifies novel phenotypes driven by a single mutant allele in Huntington's disease CAG knock-in mice. PLoS One. 2013;8: 1–19. doi:10.1371/journal.pone.0080923

85. Wheeler VC, White JK, Gutekunst C, Vrbanac V, Weaver M, Li X, et al. Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in Hdh Q92 and Hdh Q111 knock-in mice. 2000;9: 503–514.

86. Wheeler VC, Gutekunst C, Vrbanac V, Lebel L, Schilling G, Hersch S, et al. Early phenotypes that presage late-onset neurodegenerative disease allow testing of modifiers in Hdh CAG knock-in mice. 2002;11: 633–640.

87. Heng MY, Detloff PJ, Albin RL. Rodent genetic models of Huntington disease. Neurobiol Dis. 2008;32: 1–9. doi:10.1016/j.nbd.2008.06.005

88. Shelbourne PF, Killeen N, Hevner RF, Johnston HM, Tecott L, Lewandoski M, et al. A Huntington's disease CAG expansion at the murine Hdh locus is unstable and associated with behavioural abnormalities in mice. Hum Mol Genet. 1999;8: 763–774. doi:10.1093/hmg/8.5.763

89. Kovtun I V, Therneau TM, McMurray CT. Gender of the embryo contributes to CAG instability in transgenic mice containing a Huntington's disease gene. Hum Mol Genet. 2000;9: 2767–75. Available: http://www.ncbi.nlm.nih.gov/pubmed/11063736

90. Wheeler VC, Lebel LA, Vrbanac V, Teed A, te Riele HT, MacDonald ME. Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. Hum Mol Genet. 2003;12: 273–281. doi:10.1093/hmg/ddg056

91. Lloret A, Dragileva E, Teed A, Espinola J, Fossale E, Gillis T, et al. Genetic background modifies nuclear mutant huntingtin accumulation and HD CAG repeat instability in Huntington's disease knock-in mice. Hum Mol Genet. 2006;15: 2015–2024. doi:10.1093/hmg/ddl125

92. Møllersen L, Rowe AD, Illuzzi JL, Hildrestrand GA, Gerhold KJ, Tveterås L, et al. Neil1 is a genetic modifier of somatic and germline CAG trinucleotide repeat instability in R6/1 mice. Hum Mol Genet. 2012;21: 4939–4947. doi:10.1093/hmg/dds337

93. Lee J-M, Zhang J, Su AI, Walker JR, Wiltshire T, Kang K, et al. A novel approach to investigate tissue-specific trinucleotide repeat instability. BMC Syst Biol. 2010;4: 29. doi:10.1186/1752-0509-4-29

94. Dion V. Tissue specificity in DNA repair: Lessons from trinucleotide repeat instability. Trends Genet. Elsevier Ltd; 2014;30: 220–229. doi:10.1016/j.tig.2014.04.005

95. Mason RP, Breda C, Kooner GS, Mallucci GR, Kyriacou CP, Giorgini F. Modeling Huntington Disease in Yeast and Invertebrates [Internet]. Second Edi. Movement Disorders: Genetics and Models: Second Edition. Elsevier Inc.; 2014. doi:10.1016/B978-0-12-405195-9.00033-0

96. Møllersen L, Rowe AD, Larsen E, Rognes T, Klungland A. Continuous and periodic expansion of CAG repeats in huntington's disease R6/1 mice. PLoS

Genet. 2010;6: 1–11. doi:10.1371/journal.pgen.1001242

97. Lee JM, Pinto RM, Gillis T, St. Claire JC, Wheeler VC. Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different expansion dynamics in striatum and liver. PLoS One. 2011;6: 6–13. doi:10.1371/journal.pone.0023647

98. Larson E, Fyfe I, Morton AJ, Monckton DG. Age-, tissue- and length-dependent bidirectional somatic CAG•CTG repeat instability in an allelic series of R6/2 Huntington disease mice. Neurobiol Dis. Elsevier Inc.; 2015;76: 98–111. doi:10.1016/j.nbd.2015.01.004

99. Gonitel R, Moffitt H, Sathasivam K, Woodman B, Detloff PJ, Faull RLM, et al. DNA instability in postmitotic neurons. Proc Natl Acad Sci U S A. 2008;105: 3467–72. doi:10.1073/pnas.0800048105

100. Kovalenko M, Dragileva E, St. Claire J, Gillis T, Guide JR, New J, et al. Msh2 Acts in Medium-Spiny Striatal Neurons as an Enhancer of CAG Instability and Mutant Huntingtin Phenotypes in Huntington's Disease Knock-In Mice. PLoS One. 2012;7: 1–10. doi:10.1371/journal.pone.0044273

101. Ament SA, Pearl JR, Grindeland A, St. Claire J, Earls JC, Kovalenko M, et al. High resolution time-course mapping of early transcriptomic, molecular and cellular phenotypes in Huntington's disease CAG knock-in mice across multiple genetic backgrounds. Hum Mol Genet. 2017;26: 913–922. doi:10.1093/hmg/ddx006

102. Pinto RM, Dragileva E, Kirby A, Lloret A, Lopez E, St. Claire J, et al. Mismatch Repair Genes Mlh1 and Mlh3 Modify CAG Instability in Huntington's Disease Mice: Genome-Wide and Candidate Approaches. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003930

103. Tomé S, Manley K, Simard JP, Clark GW, Slean MM, Swami M, et al. MSH3 Polymorphisms and Protein Levels Affect CAG Repeat Instability in Huntington's Disease Mice. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003280

104. Dragileva E, Hendricks A, Teed A, Gillis T, Lopez ET, Friedberg EC, et al. Intergenerational and striatal CAG repeat instability in Huntington's disease knock-in mice involve different DNA repair genes. Neurobiol Dis. Elsevier Inc.; 2009;33: 37–47. doi:10.1016/j.nbd.2008.09.014

105. Kovtun I V, Liu Y, Bjoras M, Klungland A, Wilson SH, McMurray CT. OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature. 2007;447: 447–52. doi:10.1038/nature05778

106. Mason AG, Tomé S, Simard JP, Libby RT, Bammler TK, Beyer RP, et al. Expression levels of DNA replication and repair genes predict regional somatic repeat instability in the brain but are not altered by polyglutamine disease protein expression or age. Hum Mol Genet. 2014;23: 1606–1618. doi:10.1093/hmg/ddt551

107. Goula AV, Berquist BR, Wilson DM, Wheeler VC, Trottier Y, Merienne K. Stoichiometry of base excision repair proteins correlates with increased somatic CAG instability in striatum over cerebellum in Huntington's disease transgenic mice. PLoS Genet. 2009;5. doi:10.1371/journal.pgen.1000749

108. Watson LM, Wong MMK, Becker EBE. Induced pluripotent stem cell technology for modelling and therapy of cerebellar ataxia. Open Biol. 2015;5: 150056. doi:10.1098/rsob.150056

109. Shin A, Shin B, Shin JW, Kim K-H, Atwal RS, Hope JM, et al. Novel allele-specific quantification methods reveal no effects of adult onset CAG repeats on HTT mRNA and protein levels. Hum Mol Genet. 2017;26: 1258–1267. doi:10.1093/hmg/ddx033

110. Tomé S, Panigrahi GB, Castel AL, Foiry L, Melton DW, Gourdon G, et al. Maternal germline-specific effect of DNA ligase I on CTG/CAG instability. Hum Mol Genet. 2011;20: 2131–2143. doi:10.1093/hmg/ddr099

111. McNeil SM, Novelletto a, Srinidhi J, Barnes G, Kornbluth I, Altherr MR, et al. Reduced penetrance of the Huntington's disease mutation. Hum Mol Genet. 1997;6: 775–779. doi:10.1093/hmg/6.5.775

112. Brocklebank D, Gayán J, Andresen JM, Roberts SA, Young AB, Snodgrass SR, et al. Repeat Instability in the 27-39 CAG Range of the HD Gene in the Venezuelan Kindreds: Counseling Implications. Am J Med Genet Part B Neuropsychiatr Genet. 2009;150: 425–429. doi:10.1002/ajmg.b.30826

113. Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, et al. Integrated genomics and proteomics define huntingtin CAG length–dependent networks in mice. Nat Neurosci. 2016;19: 623–633. doi:10.1038/nn.4256

114. Auerbach W, Hurlbert MS, Hilditch-Maguire P, Wadghiri YZ, Wheeler VC, Cohen SI, et al. The HD mutation causes progressive lethal neurological disease in mice expressing reduced levels of huntingtin. Hum Mol Genet. 2001;10: 2515–2523.

115. Semaka A, Collins JA, Hayden MR. Unstable familial transmissions of huntington disease alleles with 27-35 CAG repeats (intermediate alleles). Am J Med Genet Part B Neuropsychiatr Genet. 2010;153: 314–320. doi:10.1002/ajmg.b.30970

116. Wright SP. Adjusted P-Values for Simultaneous Inference. Biometrics. 1992;48: 1005. doi:10.2307/2532694

117. Goula AV, Stys A, Chan JPK, Trottier Y, Festenstein R, Merienne K. Transcription Elongation and Tissue-Specific Somatic CAG Instability. PLoS Genet. 2012;8. doi:10.1371/journal.pgen.1003051

118. Takiyama Y, Sakoe K, Soutome M, Namekawa M, Ogawa T, Nakano I, et al. Single sperm analysis of the CAG repeats in the gene for Machado-Joseph disease (MJD1): Evidence for non-Mendelian transmission of the MJD1 gene

and for the effect of the intragenic CGG/GGG polymorphism on the intergenerational instability. Hum Mol Genet. 1997;6: 1063–1068. doi:10.1093/hmg/6.7.1063

119. Leeflang EP, McPeek MS, Arnheim N. Analysis of meiotic segregation, using single-sperm typing: meiotic drive at the myotonic dystrophy locus. Am J Hum Genet. 1996;59: 896–904. Available: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m&form=6&dopt=r&uid=8808606

120. Ikeuchi T, Igarashi S, Takiyama Y, Onodera O, Oyake M, Takano H, et al. Non-Mendelian transmission in dentatorubral-pallidoluysian atrophy and Machado-Joseph disease: the mutant allele is preferentially transmitted in male meiosis. Am J Hum Genet. 1996;58: 730–3. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1914673&tool=pmcentrez&rendertype=abstract

121. Carey N, Johnson K, Nokelainen P, Peltonen L, Savontaus M-L, Juvonen V, et al. Meiotic drive at the myotonic dystrophy locus? Nature. 1994;6: 117–118. doi:10.1038/ng1294-340

122. Leeflang EP, Zhang L, Tavare S, Hubert R, Srinidhi J, Macdonald ME, et al. Single Sperm Analysis of the Trinucleotide Repeats in the Huntingtons-Disease Gene - Quantification of the Mutation Frequency-Spectrum. Hum Mol Genet. 1995;4: 1519–1526. Available: //a1995rt12400006

123. Yoon S-R, Dubeau L, de Young M, Wexler NS, Arnheim N. Huntington disease expansion mutations in humans can occur before meiosis is completed. Proc Natl Acad Sci U S A. 2003;100: 8834–8838. doi:10.1073/pnas.1331390100

124. Kovtun I V, McMurray CT. Trinucleotide expansion in haploid germ cells by gap repair. Nat Genet. 2001;27: 407–11. doi:10.1038/86906

125. Gomes-Pereira M, Fortune MT, Monckton DG. Mouse tissue culture models of unstable triplet repeats: in vitro selection for larger alleles, mutational expansion bias and tissue specificity, but no association with cell division rates. Hum Mol Genet. 2001;10: 845–54. Available: http://www.ncbi.nlm.nih.gov/pubmed/11285250

126. Foiry L, Dong L, Savouret C, Hubert L, te Riele H, Junien C, et al. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. Hum Genet. 2006;119: 520–526. doi:10.1007/s00439-006-0164-7

127. Slean MM, Panigrahi GB, Castel AL, Pearson AB, Tomkinson AE, Pearson CE. Absence of MutSβ leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. DNA Repair (Amst). Elsevier B.V.; 2016;42: 107–118. doi:10.1016/j.dnarep.2016.04.002

128. Ezzatizadeh V, Sandi C, Sandi M, Anjomani-Virmouni S, Al-Mahdawi S, Pook MA. MutLα Heterodimers heterodimers modify the molecular phenotype of Friedreich ataxia. PLoS One. 2014;9. doi:10.1371/journal.pone.0100523

129. Libby RT, Hagerman KA, Pineda V V., Lau R, Cho DH, Baccam SL, et al. CTCF cis-regulates trinucleotide repeat instability in an epigenetic manner: A novel basis for mutational hot spot determination. PLoS Genet. 2008;4. doi:10.1371/journal.pgen.1000257

130. Nestor CE, Monckton DG. Correlation of inter-locus polyglutamine toxicity with CAG●CTG triplet repeat expandability and flanking genomic DNA GC content. PLoS One. 2011;6. doi:10.1371/journal.pone.0028260

131. Viterbo D, Michoud G, Mosbach V, Dujon B, Richard GF. Replication stalling and heteroduplex formation within CAG/CTG trinucleotide repeats by mismatch repair. DNA Repair (Amst). Elsevier B.V.; 2016;42: 94–106. doi:10.1016/j.dnarep.2016.03.002

132. MacDonald ME, Gines S, Gusella JF, Wheeler VC. Huntington's disease. NeuroMolecular Med. 2003;4: 7–20. doi:10.1385/NMM:4:1-2:7

133. The HD iPSC Consortium. Induced pluripotent stem cells from patients with huntington's disease show cag-repeat-expansion-associated phenotypes. Cell Stem Cell. 2012;11: 264–278. doi:10.1016/j.stem.2012.04.027

134. Rivera B, Castellsagué E, Bah I, Foulkes WD, van Kempen LC. Biallelic *NTHL1* Mutations in a Woman with Multiple Primary Tumors. N Engl J Med. 2015;373: 1985–1986. doi:10.1056/NEJMc1506878

135. Weren RD a, Ligtenberg MJL, Kets CM, de Voer RM, Verwiel EENTP, Spruijt L, et al. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet. 2015;47: 1–6. doi:10.1038/ng.3287

136. Vallabhaneni H, O'Callaghan N, Sidorova J, Liu Y. Defective Repair of Oxidative Base Lesions by the DNA Glycosylase Nth1 Associates with Multiple Telomere Defects. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003639

137. Chan MK, Ocampo-Hafalla MT, Vartanian V, Jaruga P, Kirkali G, Koenig KL, et al. Targeted deletion of the genes encoding NTH1 and NEIL1 DNA N-glycosylases reveals the existence of novel carcinogenic oxidative damage to DNA. DNA Repair (Amst). 2009;8: 786–794. doi:10.1016/j.dnarep.2009.03.001

138. Cooley N, Elder RH, Povey AC. The effect of Msh2 knockdown on toxicity induced by tert -butyl-hydroperoxide, potassium bromate, and hydrogen peroxide in base excision repair proficient and deficient cells. Biomed Res Int. 2013;2013. doi:10.1155/2013/152909

139. Tumini E, Barroso S, -Calero CP, Aguilera A. Roles of human POLD1 and POLD3 in genome stability. Sci Rep. Nature Publishing Group; 2016;6: 38873. doi:10.1038/srep38873

140. Nicolas E, Golemis EA, Arora S. POLD1: Central mediator of DNA replication and repair, and implication in cancer and other pathologies. Gene. Elsevier B.V.; 2016;590: 128–141. doi:10.1016/j.gene.2016.06.031

141. Song J, Hong P, Liu C, Zhang Y, Wang J, Wang P. Human POLD1 modulates cell cycle progression and DNA damage repair. BMC Biochem. BMC Biochemistry; 2015;16: 14. doi:10.1186/s12858-015-0044-7

142. Palles C, Cazier J-B, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. Nature Publishing Group; 2013;45: 136–44. doi:10.1038/ng.2503

143. Weedon MN, Ellard S, Prindle MJ, Caswell R, Lango Allen H, Oram R, et al. An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy. Nat Genet. Nature Publishing Group; 2013;45: 947–50. doi:10.1038/ng.2670

144. Yang A, Walker N, Bronson R, Kaghad M, Oosterwegel M, Bonnin J, et al. p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. Nature. 2000;25: 99–103.

145. Candi E, Agostini M, Melino G, Bernassola F. How the TP53 Family Proteins TP63 and TP73 contribute to tumorigenesis: Regulators and effectors. Hum Mutat. 2014;35: 702–714. doi:10.1002/humu.22523

146. Pozniak CD, Radinovic S, Yang A, Mckeon F, Kaplan DR, Miller FD. An Anti-Apoptotic Role for the During Developmental Neuron Death. Science (80-). 2000;289: 304–306. doi:10.1126/science.289.5477.304

147. Stiewe T, Pützer BM. Role of p73 in malignancy: tumor suppressor or oncogene? Cell Death Differ. 2002;9: 237–45. doi:10.1038/sj.cdd.4400995

148. Engelmann D, Meier C, Alla V, Pützer BM. A balancing act: orchestrating amino-truncated and full-length p73 variants as decisive factors in cancer progression. Oncogene. 2014; 1–13. doi:10.1038/onc.2014.365

149. Lachaud C, Moreno A, Marchesi F, Toth R, Blow JJ, Rouse J. Ubiquitinated Fancd2 recruits Fan1 to stalled replication forks to prevent genome instability. Science (80-). 2016;351: 846–849. doi:10.1126/science.aad5634

150. Chaudhury I, Stroik DR, Sobeck A. FANCD2-controlled chromatin access of the Fanconi-associated nuclease FAN1 is crucial for the recovery of stalled replication forks. Mol Cell Biol. 2014;34: 3939–54. doi:10.1128/MCB.00457-14

151. Airik R, Schueler M, Airik M, Cho J, Porath JD, Mukherjee E, et al. A FANCD2/FANCI-Associated Nuclease 1-Knockout Model Develops Karyomegalic Interstitial Nephritis. J Am Soc Nephrol. 2016; ASN.2015101108-. doi:10.1681/ASN.2015101108

152. Bentley DJ, Harrison C, Ketchen A-M, Redhead NJ, Samuel K, Waterfall M, et al. DNA ligase I null mouse cells show normal DNA repair activity but altered DNA replication and reduced genome stability. J Cell Sci. 2002;115: 1551–1561.

153. Wu YY, Yang Y, Xu Y De, Yu HL. Targeted disruption of the spermatid-specific gene Spata31 causes male infertility. Mol Reprod Dev. 2015;82: 432–440.

doi:10.1002/mrd.22491

154. Bekpen C, Künzel S, Xie C, Eaaswarkhanth M, Lin Y-L, Gokcumen O, et al. Segmental duplications and evolutionary acquisition of UV damage response in the SPATA31 gene family of primates and humans. BMC Genomics. BMC Genomics; 2017;18: 222. doi:10.1186/s12864-017-3595-8

7. Appendix

Modeling strategy for repeat transmission frequency comparisons accounting for paternal CAG repeat differences

To control for potential confounding effects of paternal CAG size on the frequency of unstable transmissions between distinct strains and/or lines, we compared actual transmission frequencies in test datasets with expected frequencies derived from simulated data based on a reference dataset.

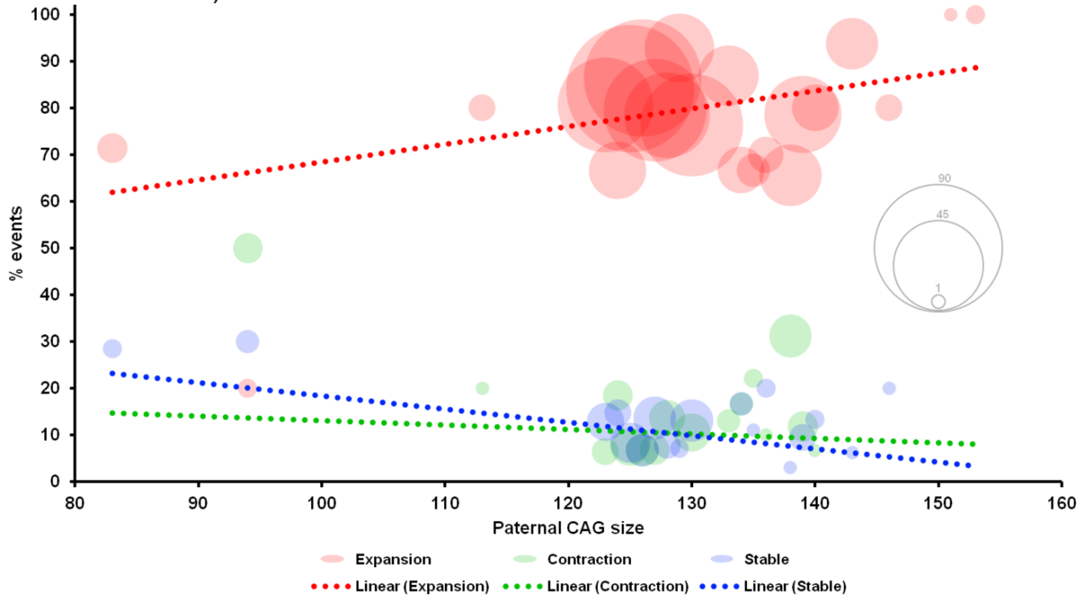
Reference dataset	dataset N	Test dataset(s)	dataset N	Notes
50% B6J (CHGR)	354	50% B6J (CHGR)	353	validation
B6J (CHGR)	707	129 (CHGR) CD1 (CHGR) FVB (CHGR) DBA (CHGR) B6N (CHGR) B6J (CHGR)	213 439 180 64 226 707	strain comparison
$CD1^{neo-}$ (CHGR)	439	$CD1^{neo+}$ (CHGR)	152	neo cassette comparison (I)
$Q175^{neo+}$ (JAX)	9172	$Q175^{neo-}$ (JAX)	256	neo cassette comparison (II)

Notes

Reference datasets always possessed a higher total number of transmissions in order to build optimal frequency vs. paternal CAG models.

Figures in this supplementary file are representations of B6J (CHGR) as the reference dataset and CD1 (CHGR) as the test dataset.

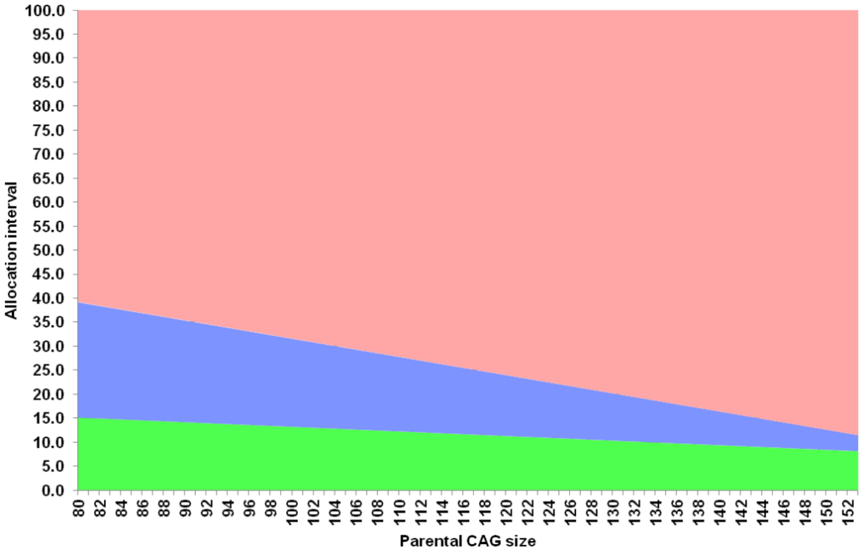
Step 1. Linear weighted trend lines for percent of events (expansions, contractions and stable transmissions) per paternal CAG size were calculated for the reference dataset (e. g. CHGR B6J strain).



Events with null frequency (N=0) are considered for trend line weighing but are not depicted as bubbles.

3

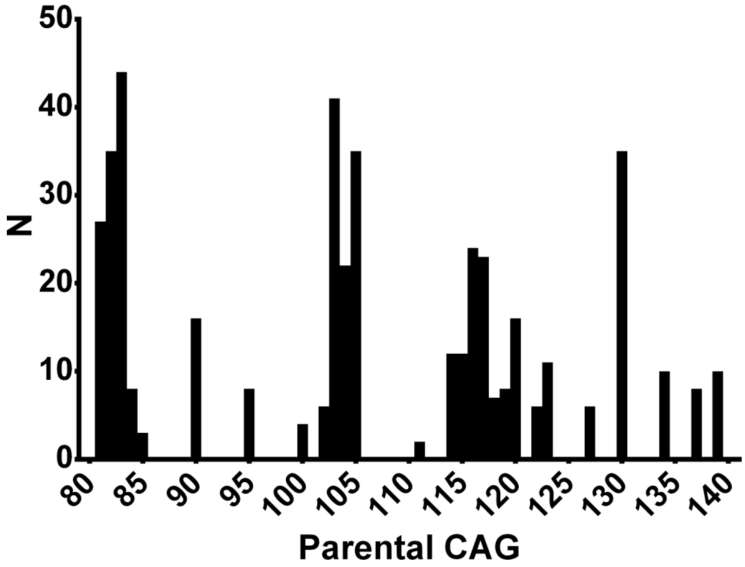
Step 2. Based on the weighted trend lines we determined frequency intervals for each event over the range of CAG sizes in the reference dataset.



Expansion – red; Stable – blue; Contraction – green

4

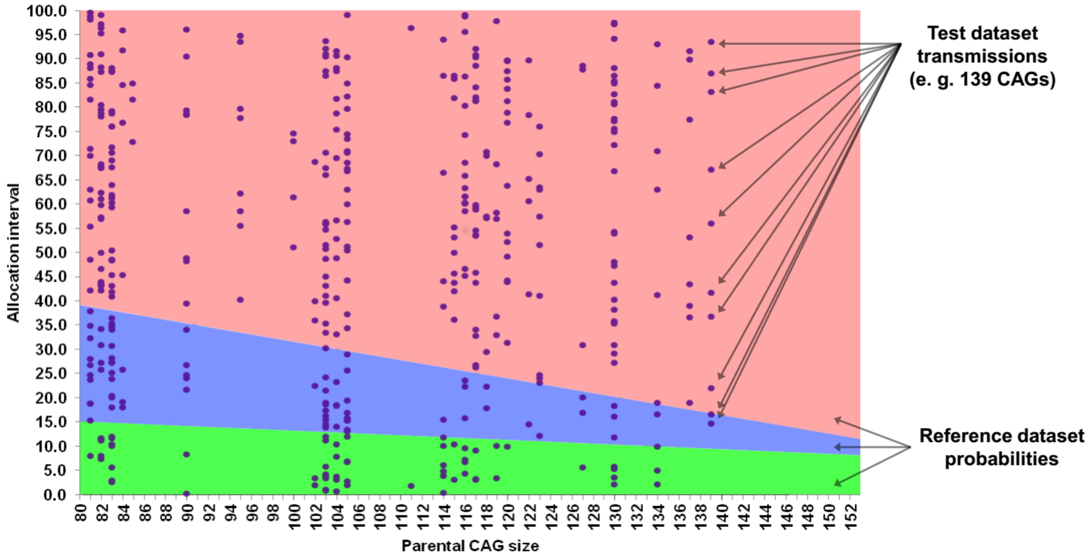
Paternal CAG repeat size distribution in the CHGR CD1 test dataset.



N_{total} = 439

5

Step 3. A random number between 0.0 and 100.0 was generated for each transmission and, based on the paternal CAG length, was allocated to contraction / expansion / stable transmission according to the frequency intervals defined by the reference dataset.



iteration 1

Each dot represents a unique transmission (N=439)

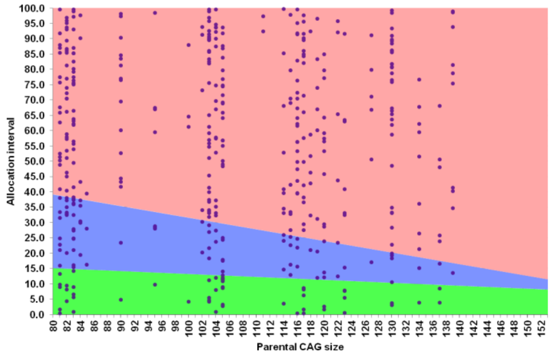
Overall distribution

%Expansion	68.8
%Stable	17.5
%Contraction	13.7

6

Step 4. 1,000 iterations of the random number generation and event allocation were performed.

iteration 2

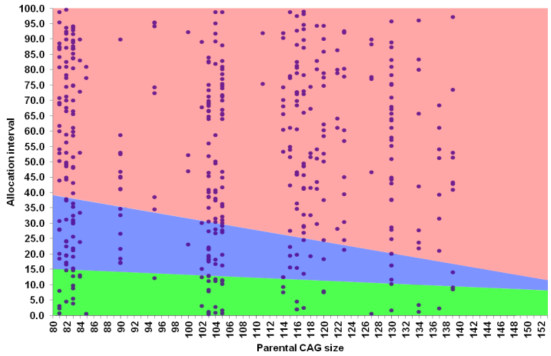


%Expansion 71.5
%Stable 16.9
%Contraction 11.6

…

…

iteration 1000



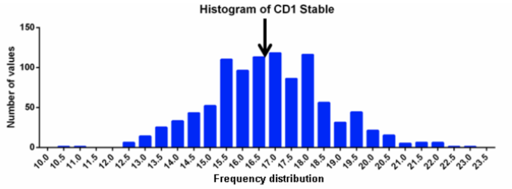
%Expansion 70.8
%Stable 17.8
%Contraction 11.4

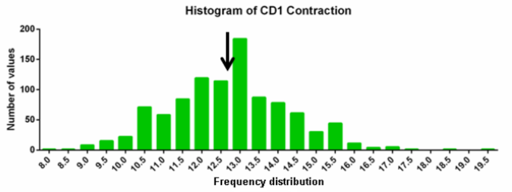
7

Step 5. The average dataset was determined and characterized.

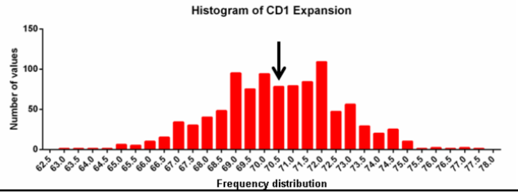
Line	Event	Mean	St. dev.	95% CI
CD1	%Stable	16.8	1.8	0.11
	%Contraction	12.7	1.6	0.10
	%Expansion	70.5	2.2	0.13

Average dataset for the CD1 strain



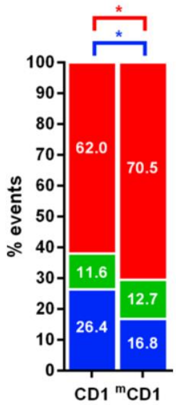


Frequency distributions (binned at 0.5% intervals for figure simplification) for stable transmissions, contractions and expansions across the 1,000 simulated datasets. Arrows indicate the value of the average dataset.



8

Step 6. Statistical analyses were performed to determine significant differences between observed and expected frequencies, as detailed in the Methods section. Validation and results are present in Figure S9, Figure 22A, Figure 23A and C.



Comparison of expansions, contractions and stable transmissions frequencies in the CD1 strain. Observed frequencies (left) and modeled(m)/expected frequencies (right).

(Partial reproduction of Figure 22A)

9

Trend lines determined for the reference datasets:

```
Validation [50% B6J (CHGR), N=354]
%Expansions = 0.214 × Parental CAG + 52.929
%Contractions = 0.053 × Parental CAG + 3.095
%Stable = -0.267 × Parental CAG + 43.976

Strain comparison [B6J (CHGR)]
%Expansions = 0.380 × Parental CAG + 30.423
%Contractions = -0.096 × Parental CAG + 22.77
%Stable = -0.284 × Parental CAG + 46.807

neo cassette comparison (I) [CD1neo- (CHGR)]
%Expansions = 0.596 × Parental CAG - 0.998
%Contractions = -0.156 × Parental CAG + 28.08
%Stable = -0.44 × Parental CAG + 72.918

neo cassette comparison (II) [Q175neo+ (JAX)]
%Expansions = -0.581 × Parental CAG + 193.207
%Contractions = 0.524 × Parental CAG - 88.907
%Stable = 0.057 × Parental CAG - 4.3
```

10