5-2016

Trinucleotide Heterogeneity Expansions: In vitro Molecule Amplification & Purification

Gennaro DelliCarpini
University at Albany, State University of New York

Trinucleotide Heterogeneity Expansions: In vitro Molecule Amplification & Purification

An Honors Thesis Presented to

The Department of Biology

University At Albany, State University of New York

In partial fulfillment of the requirements

For graduation with honors in Biology

And

Graduation from the honors college

Gennaro DelliCarpini

Research Mentor: Dr. Pan Li

Research Advisor: Cara Pager

Abstract

Simple sequence repeats (SSRs) are mutation-prone DNA tracts composed of tandem repetitions of relatively short motifs (Kashi & King, 2006). SSRs are found scattered throughout the human genome in both coding and non-coding regions. Present under the category of SSRs are Trinucleotide repeats, which may vary in number and, when expanded upon, provide the basis of at least nine neurodegenerative diseases. Large degrees of structural heterogeneity are found in CAG trinucleotide repeat DNA. DNA slippage leads these hetero-structures to form, since certain repeats slip out of the origin of replication and are carried into the next generation of replication. This slippage during replication has only been seen to occur in scattered segments of the genome where repeated nucleotides reside. This can lead to diseased states involving increased amounts of repeats than in a normal-disease free individual. Current research into trinucleotide repeats involved repeat numbers ranging from 10-35 per molecule (Ning et al., 2015). This is due to the difficulty with replication of high repeat molecules. Here, an attempt to solve this problem is taken, by replication of various amplicons, ranging from 10-150, which are similar repeat numbers ranging from disease free genotypes to the most severely diseased states. This replication is to be completed by a real-time PCR assay and Evagreen binding dye as the end goal. By use of PCR with long stable primers, we are able to see amplification occur in a repeat-only manner. In order to deal with the structural heterogeneity of trinucleotide repeats alone, smaller primers were used to replicate the various lengths of DNA. Future collaboration with Nanopore technologies' new device, MinIon, will hopefully reveal insights into the sequence of these repeats, especially at higher length, using mid ranged primers. We hope to elucidate this replication issue, with the possible application of this method to DNA repeats associated with other known diseases.

Acknowledgements

I would like to thank Dr. Pan Li, for all the countless hours I have spent in the lab and his help throughout.

I would like to thank William Stephenson, a future collaborator and past graduate student in the Pan T.X. Li laboratory that initially began this project.

I couldn't have completed this project without my lab mates and friends Kidane Tewolde and Botros Toro, they have always been receptive to all ideas and have helped me during my time in the lab.

I would like to thank Cara Pager for being receptive and helpful for ideas and statements presented in my research.

I would lastly like to thank my family and friends, of which without their support would not have been able to complete my education at UAlbany.

Contents

Introduction:

Background

Short tandem repeats, also known as simple sequence repeats (SSRs), are short tandemly repeated DNA sequences that include a repetitive sequence unit of 1-6 base pairs in length (Fan & Chu, 2007). Repeated sequences are contained within 50% of the human genome, while SSRs are found in about 3% of the genome, scattered throughout (Fan & Chu, 2007; Lander et al., 2001). Given the location differences, the molecular cause of SSR diseases is expected to vary (Sinden et al., 2002). Most of these repeats have been thought of to be "junk" DNA with no biological use at all. SSRs can result in a variety of effects on the affected individual depending on their location within the genome, along with the type of repeat present (Kashi & King, 2006; Usdin, 2008). Abundance in the coding regions can be a result of length variation, as a result of high mutation rate in the TNR (trinucleotide repeat) (Borstnik & Pumpernik, 2002; Sobczak et al., 2010). In response to environmental change, SSRs may provide an evolutionary advantage by mediating change at the genome level (Li, Korol, Fahima, & Nevo, 2004). An aspect of both bacterial and eukaryotic genomes is that they are mutations due to slipped strand mispairing (Moxon, Bayliss, & Hood, 2006).

Repeat lengths have been found to vary depending on both the gene and the location. In the normal population these SSRs are kept in check and are present in low numbers, under 35 repeats, while 35 or more repeats usually represents a diseased individual (Ning et al., 2015). While it is true that the normal function of these normal SSRs is poorly understood, it has been found that abnormally expanded repeats of certain types trigger pathogenesis in several human genetic diseases (Galka-Marciniak, Urbanek, & Krzyzosiak, 2012). Here, we focus on trinucleotide repeats, which are found scattered

throughout the genome. What makes these repeats unique is their abundance in intergenic regions of the genome, introns, and untranslated, as well as translated portions of the genome.

Various lengths of the repeat CAG/CTG are present in 9 diseases such as Huntington's, myotonic dystrophy and multiple forms of spinocerebellar ataxias (Cooper, 2009; Gatchel & Zoghbi, 2005; Orr & Zoghbi, 2007). In these disorders, the presence of expanded CAG trinucleotide repeats in coding (PolyQ diseases) and untranslated regions (UTR) is toxic to the cell (Tsoi, Lau, Tsang, Lau, & Chan, 2012). Repeat expansion is a major cause of a diseased state, at both the RNA and DNA levels. In Huntington's disease, expanded CAG repeats directly relate to the severity of the affected individual. The CAG is translated into a PolyQ stretch and, when it exceeds a critical length of 37 glutamines, a disease state is manifested (Möncke-Buchner et al., 2002). CAG repeats are found predominantly in the coding regions of human genes, which offers their functional significance (Kiliszek, Kierzek, Krzyzosiak, & Rypniewski, 2010). These diseases increase in severity with earlier onset and currently there is no cure (Amrane et al., 2005). Long CAG repeats are mostly unstable during transmission to future generations, leading to potential increased copy number of progeny (Koob et al., 1998; Santillan, Moye, Mittelman, & Wilson, 2014).

Cause

These diseases mainly result from gain of function alleles, which then goes on to create dominant negative effects, leading to the formation of deleterious genes within an affected individual (Tsoi et al., 2012). Occasionally, a segment of the DNA slips out of the nascent strand during DNA replication (Michael J. Hartenstine, Goodman, & Petruska, 2000). This slippage, when expanded upon in further amplifications, provides a basis of the repeat

associated disease. The dynamic aspect of the copy number leads to variable expressivity in families where the diseased state is present (Orr & Zoghbi, 2007). SSR polymorphisms are derived mainly from variability in length, rather than primary sequence (Ellegren, 2004).

Abnormal number of repeats found in these diseases are a result of a process named 'repeat instability,' where repeats are subject to a change in copy number (Salinas-Rios, Belotserkovskii, & Hanawalt, 2011). Expansion of these trinucleotide repeats continues in affected individuals, which then increases the severity of the disease, as well as decreasing the age of onset for subjects who inherit the disease (Mitas, 1997). An affected individual usually contains greater than 50 base pairs of this repeat in a given gene, while a normal gene would contain less than 30. The changes in the number of these repeats are usually large, localized to specific genes, and do not involve any sort of mismatch repair system. This expansion is due to the ability of certain repeats to form unusual complexes, such as a hairpin conformation.

Replication

Duplex B-DNA replication occurs when the two parental strands separate at the replication fork. This process is catalyzed by eukaryotic enzymes in DNA replication, repair, or recombination (M. J. Hartenstine, Goodman, & Petruska, 2002). During this period of replication, there is an opportunity for DNA slippage, where trinucleotide repeats can be incorporated into the DNA itself (Mitas, 1997). The model upon DNA slippage occurs is a simple concept on the surface, where structures containing the repetitive sequence are allowed to slip out of the replication pathway and form small hairpins (Figueroa, Cattie, & Delaney, 2011).

Formation of non-B DNA (normal DNA) conformations, such as a hairpins, can led to increased trinucleotide repeat (TNR) expansion (Figueroa et al., 2011). Looping out of one

or more repeats then converts this loop into expansions during the second round of replication. This then continues throughout replication, the amount of TNR is increased. Prevalence of these hairpins in disease has not been well defined up to this point, however it is known that they occur by two mechanisms: loop-loop interactions found in kissing hairpins or stem-stem interactions forming a cross shape (Figueroa & Delaney, 2010).

Past Work

There has been work done in the past on repetitive sequences in the drosophila model organism, since 1/3 of the species genome consists of repetitive elements, Along with these repetitive elements, constant mutation occurs as a result. Here, simple repetitive satellite sequences were looked at through the usage of qPCR by which it was possible to monitor the replication of the repetitive sequences over time (Aldrich & Maggert, 2014). It is also possible to examine the PCR product of various DNAs by thermal melting studies, which in the past there has mainly been work done with UV melting instead (Amrane et al., 2005; Michael J. Hartenstine et al., 2000).

There has also been an issue in past work to replicate large CAG repeats in vitro, over 100 base pairs. Most recent studies of these CAG repeats only study repeats of lengths between CAG10 and CAG30, which is the normal amount that does not cause any disease (Ning et al., 2015). These studies also tested the length polymorphism of TNRs, by use of a fluorescence assay, in order to see the exact binding of dsDNA to the fluorophores.

Goals:

Immediate Goals

In order to determine the length dependent heterogeneity of CAG repeats, PCR will be performed. By using longer more stable primers, we will be able to see PCR occur and then perform double digestion in order to see if the replication of the repeats occurs in a manner of just the repeats alone. By seeing just the CAG repeats present without primer, shows that just the CAG repeats are being amplified during the reaction and nothing else. Next by PCR with medium sized primers, named G primers, we will hope to have successful amplification of each CAG length and then send them off for collaboration with Nanopore technologies. With their new device, MinIon, we will be able to identify the exact sequence of the CAG PCR product. This will inform us of whether or not the replication of repeats is occurring, and also if there are any differing lengths of CAG present in the product. This will lead us to confirm that the PCR of higher number repeats is highly polymorphic in length. Next by amplicification of all repeats with small primers, we will then use this PCR product as stock for the next part of the reaction. We want to use smaller primers in order to have a small handle on the 5' and 3' ends of the DNA, in order to test just how the CAG lengths replicate in real time, without interference of primers.

Ultimate Goal

The ultimate goal of this honors thesis is to develop an assay to monitor trinucleotide repeat amplification, by real-time PCR. The fluorescent dsDNA binding dye Evagreen will be used in a qPCR reaction again with all eight repeats ($CAG_{10\text{-}150}$). This will lead us to a conclusion about what exactly is happening with normal PCR replication of higher number repeats, such as CAG150. Real-time monitoring of the CAG amplification will

lead us to insights on where exactly the replication is going astray. Fluorescence by Evagreen will should increase on a logarithmic scale, and any miscellaneous products of the reaction will be found by melt curve analysis on each repeat. This will lead to insights on whether there are any heterogenic products created during the amplification itself and also if any sort of primer-dimer complexes form during the reaction. We hope to then be able apply this technique to other repetitive sequences in vitro, in order to apply this information to how exactly these repeats amplify in affected individuals.

Materials and Methods:

Long PCR Schematic of CAG_n

In order to achieve stable replication of the CAG repeats, PCR using longer primers was completed (Figure 1). Once denatured in vitro, these longer primers are annealed to either end of the ssDNA, and replication of a nascent DNA strand is completed. Amplification is then completed, where you have multiple strands of DNA presented, annealed to the 3kb primers. Next double digestion is represented, using restriction enzymes specific for the site between the primer handle and the repeat. After double digestion you will end up with solely the CAG repeat itself.

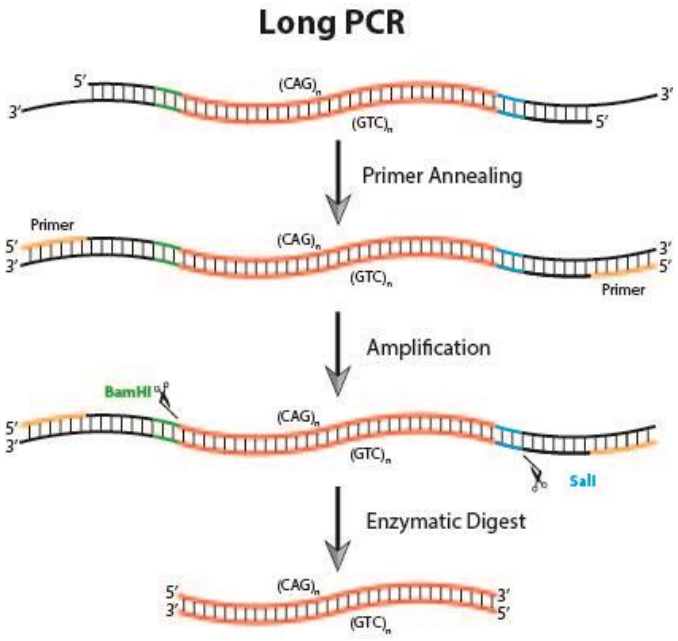

Figure 1: The methods for purifying CAG/DNA samples- represented in a linear plasmid. Restriction enzymes rid of handles used to amplify the DNA. Size of sample was confirmed on an agarose gel.

Short PCR Schematic of CAG_n

Shorter primers are used to replicate CAG repeats from the PCR stock of previous longer PCRs. These shorter primers are shown in Figure 2, and will be used in future experiments in order to test the replication of the repeats so that almost all that will be tested is the repeat itself.

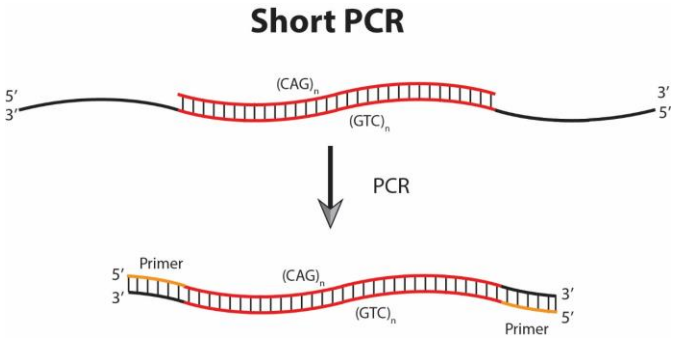

Figure 2: Design of the short PCR using M-primers. After PCR reaction and purification, these amplicons will be used as stock for future reactions.

PCR Reaction Protocol

The general formula for all PCR reactions is represented in Table 1a. This is the general amount of each component of the reaction that has been used and has helped undergo successful replication. Overall procedure for the PCR reaction is represented in Table 1b. The amplification portion of the initial

500µL PCR	Single RXN	MM per CAG Length
Template (10ng/µL)	2	10
Primer 1 pUCT7f (100 µM)	1	6
Primer 2 pUCBr (100µM)	1	6
dNTP (25 mM)	0.8	4.8
Buffer	10	60
H2O	84	504
Taq (10-100) Q5 (150)	0.8	4.8
Total (no Template)	97.6	585.6

Table 1a: Reagents used in order to perform 500µL PCR on seven CAG lengths. CAG10, 20, 30, 40, 60, 100, and 150.

replication schematic follows the Thermal Cycler procedure in order to induce PCR.

Methods

CAG sequence's, of lengths 10-150, were inserted into a plasmid, and provided initially by Dr. Pan Li. Next Polymerase chain reaction (PCR) was completed in order to replicate and create multiple copies of this DNA sequence containing the CAG sequence of choice (Figure 1). The timing of each cycle is critical, since the DNA must be able to denature at higher temperatures and re-anneal at lower temperatures. The 2720 Model Thermal cycler provided by Applied Biosystems was used to complete replication. By use of a protocol named p53-handle, the CAG repeats were able to undergo the usual steps of PCR, including denaturation and re-annealing for about 30 cycles (Figure 1). Lastly an elongation step is implemented in at around 72 degrees, which is completed after all cycles are finished. For CAG repeats of length 10-100, taq polymerase is used, with Q5 polymerase used for CAG150, at least for longer PCR. In order to see the difference in replication with either polymerase enzyme, Q5 was used because of its known higher efficiency. Q5 has an error rate of 100 times less than Taq polymerase, hence why it is used for the more difficult replication target (NEB.com).

Thermal Cycler Procedure:
1. p53 handle program selected
A: Initialization of PCR Amplification Steps
1. Heat to 95°C: Strand Denaturation
2. Cool to 55°C: Primers Anneal
3. Heat to 75°C: Taq/Q5 Begins Amplification
4. Repeat Steps 1, 2 and 3 for 30 Cycles
5. Cool to 72°C: Elongation step of DNA replication
5. Cool to 4°C: Amplification Halt and Refrigeration

Table 1b: PCR procedure. Repeated for 30 cycles to obtain maximum amount of product.

The replicated CAG samples were then obtained and analyzed by non-denaturing gel electrophoresis using an agarose gels of various percentages, which has been used in past studies of DNA amplification by PCR (Amrane et al., 2005). In order to produce the agarose gel, 30mL of 10% Tris buffer was combined agarose powder, heated and cast into a frame to solidify. The samples were loaded, along with a midsized DNA ladder, by using SYBR Green. This dye binds to the target product and emits fluorescence, later detected by the Typhoon Tri gel scanner. A Mid-Range Plus DNA ladder, provided by Fisher BioReagents, ranging from 100bp -5000bp was added to the first lane, and aiding in the representation of the expected size relationship.

After the samples were determined to have amplified properly, each of the seven DNA samples were then purified to obtain a product comprised of only DNA. The GeneJet PCR purification kit, provided by ThermoFisher Scientific was used (Appendix B), where pre-made binding buffer was added to the PCR product, and then the DNA was filtered out, with the remaining PCR byproducts disposed of. 50 μL of purified DNA was obtained from each of the original PCR product mixtures. One microliter of each sample was then taken and applied to the Nanodrop ND-1000 apparatus, provided by Thermo Scientific, and concentrations (in ng/μL) for each CAG sample were taken (Appendix C). The obtained values were all within acceptable ranges, as found before by Dr. Pan Li, showing that the samples are rich in DNA. The seven samples were then run by non-denaturing gel electrophoresis on an agarose gel to confirm the correct product was obtained. In order to scan all gels, the Typhoon Trio was used, at various voltages, in order to detect the fluorescence given off by each of the repeats.

Results and Discussion:

PCR and Purification of $CAG_{10\text{-}150}$ Using Long Primers

In order to see if $CAG_{10\text{-}150}$ repeats contain differential mobility when amplified, 3kb primers were used to amplify various CAG length. The PCR products from each respective CAG repeat run on a non- denaturing agarose gel. According to the Mid-Range Plus DNA ladder in the left most column, increased molecular weight results in decreased travel length for each band. There is a trend present throughout the gel, where the CAG with the lowest repeat number, CAG10, also has the lowest molecular weight. These bands are also expected to be around the 3 kb size that is presented by the DNA ladder. This trend in mobility, where the higher CAG number is less mobile, is what we would expect from amplification. However, this is not seen with CAG150, as it is more mobile than the smaller CAG lengths, which is due to the heterogeneous product present after amplification.

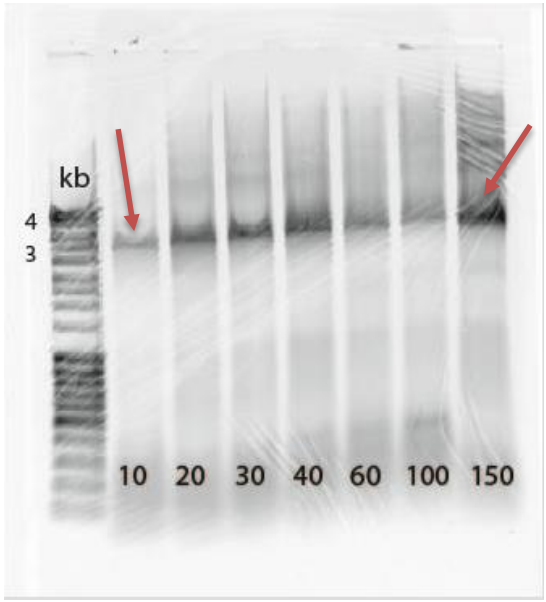

Figure 3: 0.8% Non-Denaturing Agarose Gel- Ran at 120V for 40 minutes with a DNA ladder loaded in the first lane. CAG samples are labeled and loaded sequentially in each lane.

~3Kb Primer Purification and Concentration by NanoDrop

Once purification of each PCR sample was completed, 1 µL of each of the samples was obtained and the DNA content of each was verified (Table 2). For most of the repeats we found concentration values around 200 ng/µL. For the CAG 150 repeat, there was more DNA than most. This is likely due to the larger amount of CAG repeats in the DNA itself, and is attributed to heterogeneity of the products produced from the PCR reaction of this repeat itself.

CAG Repeat	Concentration (ng/ µL)	260nm Wavelength 10mm Path
10	407.9	8.159
20	248	4.96
30	227.1	4.541
40	272.1	5.442
60	227.1	4.541
100	196.3	3.926
150	394	7.879

Table 2: Nanodrop values after purification. The DNA concentration is represented for a 1-µL sample of each CAG repeat.

*CAG10 PCR purification was repeated after failure.

The concentration of CAG10 came out to be higher than normal. This may be because it was replicated on its own, since it came out with a smeared look and low concentration (95.1ng/ µL) during the first replication. The overall expectation would be that CAG150 would have most prominent band represented on the gel, since it has the highest molecular weight and theoretically the most product present after amplification. However, due to the suspected heterogeneity of CAG amplification, this is not the case, as seen with CAG10 having a much darker band than CAG150, indicating that the PCR went well for the smaller product versus the larger.

There is seen to be a stepwise nature to the CAG repeat mobility represented after obtaining the purified PCR product (Figure 4). Now the bands are much more concise and ran farther down the gel, showing a decrease in weight for each CAG repeat. This is due to getting rid of all the other byproducts to PCR itself, such as the dNTP, buffer and water. There is a strong band present for CAG10, which is most likely due to the decreased error and heterogeneity present with amplification. All of the other repeats are presented as expected, besides CAG150. Although CAG150 was replicated with Q5, there still may have been error during the replication process. Because of the large amount of CAG present with higher number of repeats, the polymerase enzyme is surmised to skip over many repeats, forming different products than what is expected from replication.

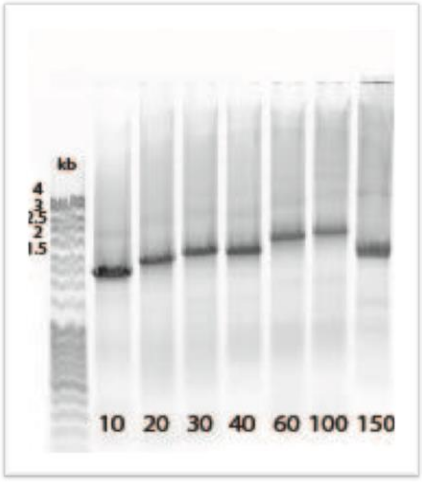

Figure 4: Agarose gel run by electrophoresis at 120V for 40 minutes. Each of the 7 purified CAG repeats lengths is present, along with the DNA ladder.

Double Digestion of Long CAG_n PCR Products

After purified PCR product for each amplicon was obtained, double digestion of each repeat was performed. Restriction enzymes used are BamHI and SalI. BamHI is found to cleave specifically at a target site, which in this case is right after the 5’ Guanine on each strand containing the target sequence of 5'-GGATCC-3' (NEB). SalI also cleaves at a specific site, except the enzyme recognizes a different sequence, 5'-GTCGAC-3'. The resulting double digestion shows the exact heterogeneity present in high number repeat amplification. All bands hover around the 50bp mark, makes sense since the repeats without handles should

all be around that size, for example CAG10 is 40 bp in length, which is close to where the band presents itself. This gel is scanned at such a high intensity in order to see the CAG insert after the primers are taken off both the 5' and 3' ends. By completing the double digestion of CAG150, there is a smear present as seen on previous gels. The smear presented is due to the heterogeneity of the amplification of CAG150. This again proves that the replication of higher number repeats is a difficult task that increases in difficulty with size. However, CAG60 and CAG100 have solid bands closer to 100bp on the gel, which shows they replicated well. Overall the digestion of these repeats settles the issue that the replication is occurring in such a manner that the repeats alone are able to amplify.

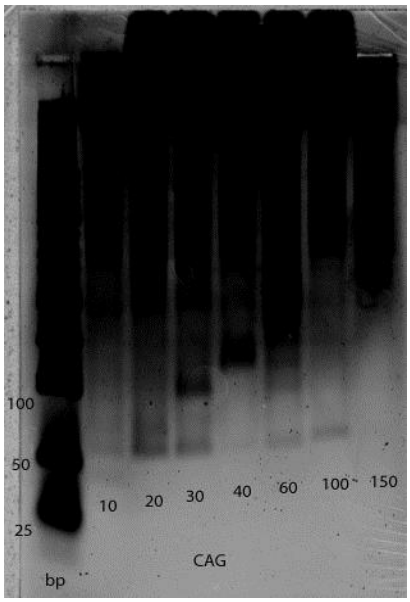

Figure 5: 3% agarose gel of the purified $(CAG)_{10-150}$ PCR product Double Digestion, scanned at 700pmt.

CAG_{10-150} Replication by ~500bp Primers

After replication and digestion of longer repeats, intermediate length G-primers

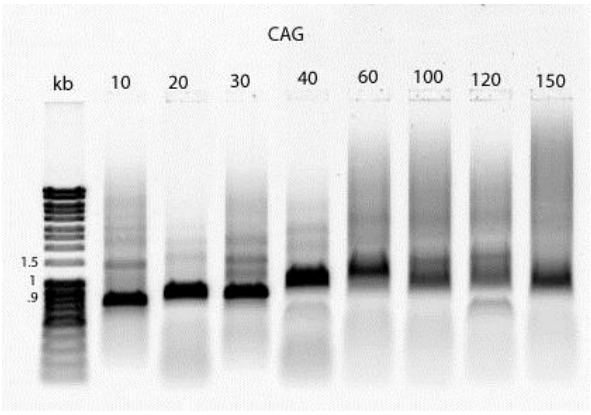

Figure 6: 0.8% agarose gel used to run the PCR product of the G-primer amplicons. All samples present with similar step-wise nature, with increasing numbers of CAG repeats. CAG120 is

(of ~500bp in length) were used to replicate the PCR product made from the longer-primer PCR. Figure 6 shows the results from a 0.5mL PCR of G-primer repeats, which presents the same decrease in mobility we have surveyed with the ~3kb primer amplification. The purification gel, Figure 7, shows more concise bands on the agarose gel. One of the main points of the intermediate sized primers is to visualize the gap between CAG100 and 150. So far only those repeats have been amplified, which presents a huge gap in the number of repeats between CAG100 and 150. By including CAG120, you can see that it has partially amplified and partially acts like CAG150, meaning that this number repeat also produces heterogeneous product instead containing all CAG120. The G-Primer concentrations generally increased as the repeat number of CAG increased, which shows the DNA amplified and is present in the product (Table 3). Using

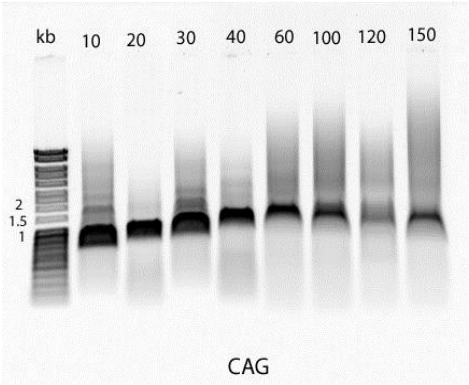

Figure 7: 0.5% agarose gel used to run the purified PCR product of the G-primer amplicons. Purification of the PCR product results in more concise bands, to view where exactly the DNA is migrating.

$(CAG)_n$	ng/uL
10	458.3
20	355.5
30	470.2
40	379.7
60	486.8
100	577.2
120	478
150	587.2

Table 3: $(CAG)_n$ G-primer nanodrop values. Concentrations for each repeat presented.

Nanopore technology we hope to sequence and observe the length variation in PCR, namely in the higher repeat numbers.

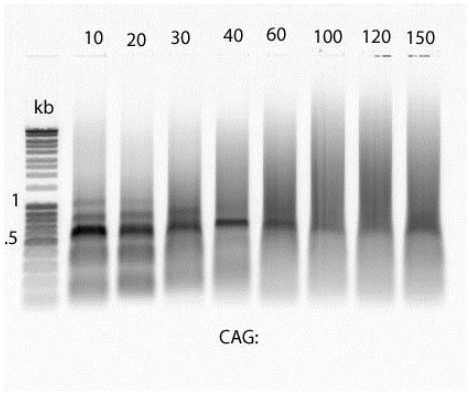

Figure 8: Purified M-primer amplicons ran on a .5% agarose gel. The m-primers are much shorter and weigh less than the G-primer series. Amplification of each repeat is shown to occur for each CAG length, with CAG150 producing a complete smear. Fainter bands appear as the repeat number is increased.

CAG Replication Using ~40bp Primers

Replication of CAG repeats using smaller M-primers was completed using the same PCR formula as replication with G-primers. Purified CAG PCR product, using G-primers, was used as the starting vector to which the M-primers were then annealed. Because of the smaller size of the primers, fainter bands appeared for each of the CAG lengths as opposed to the more prominent bands seen with the long PCR and G-primer PCR. Still the same trend of each repeat, seen in figure 8, appeared in the agarose gel of the M-primer purified amplicons. By using the smaller primers to amplify all CAG lengths, we will be able to visualize how the CAG insert amplifies with use of real-time PCR.

$(CAG)_n$ M Primer Concentrations

$(CAG)_n$	ng/uL
10	306.58
20	315.86
30	346.48
40	243.13
60	382.42
100	314.73
120	406.10
150	373.84

Table 4: $(CAG)_n$ M-primer NanoDrop values. Concentrations for each repeat were recorded as well as their absorbance value at 260 and 280nm.

Real-Time PCR Assay

An assay for qPCR was devised, in which we can monitor the amplification of all amplicons in real time. Here, CAG10 and CAG40 were used, along with CD4 non-repetitive DNA. All three amplicons replicated along with evagreen dsDNA binding dye, so that amplification of dsDNA correlates with the relative fluorescence units given off by the Evagreen-dsDNA binding complex. In order to see how repetitive sequences are amplified, CD4 DNA was used as a positive control to see how exactly a non-repetitive sequence is amplified in real time. You can see from the graph that CAG10 reaches its maximum concentration of DNA within 10 cycles of amplification. CAG40, however, takes a much longer time to amplify and reach its maximum DNA concentration. Also, there is much less CAG40 present after amplification, which is not expected if the replication were to be performed without any heterogeneous product present. With future real time amplification of larger sized amplicons, we will be able to see how exactly other CAG lengths replicate.

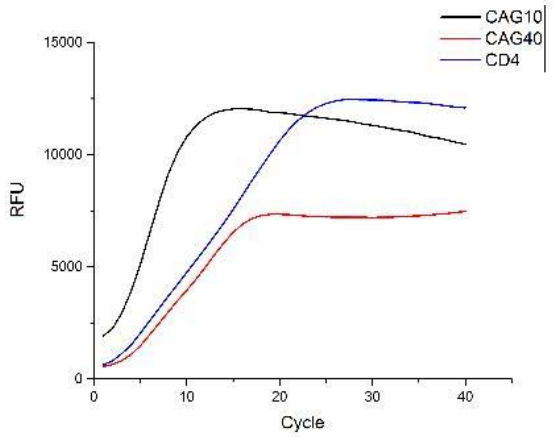

Figure 9: Real time PCR using samples of CAG10, CAG40, and CD4 DNA. CAG10 Amplifies to a maximum RFU in 10 Cycles, while CAG40 takes about 20 cycle to reach its maximum DNA concentration.

Conclusion:

CAG repeats have a medically significant impact on various human diseases. Looking into these CNG repeats provides a promising target for therapeutic treatment of trinucleotide repeat diseases (Broda, Kierzek, Gdaniec, Kulinski, & Kierzek, 2005). Here, we replicated and purified seven CAG repeats, which all varied in size, as do repeat lengths present in multiple diseases such as Huntington's and spinocerebellar ataxia. Purification for all repeats was successful, as the repeats traveled down the gel in a stepwise fashion, attributing to their repeat size. The differences seen in mobility correlates to the molecular weight of each CAG repeat. However, we are still trying to solve what is happening with CAG150 replication, as it should travel the least among all repeat lengths.

We surmised and came to the conclusion that the polymerase enzyme skips over many of the repetitive sequences during amplification. This 'skipping' is mostly present in amplicons with greater amounts of repeats, due to increased error of the polymerase. Most likely attributing to its large repeat size; CAG150 appears as a large band that seems to travel lower on the agarose gel than other repeats on the purification gel. Appearing further down the gel indicates CAG150 would have the lowest DNA size; however, in theory, it should be the heaviest. CAG150 is presented further down the gel due to the heterogeneity of the products formed from its amplification. This provides further proof that the polymerase enzyme skips over many repeats during amplification and creates heterogeneous product.

CAG10 was the only repeat, besides CAG150, that had an abnormally high DNA concentration, presented by both the band on the gel and by the Nano-drop concentration. This concentration also leads us to the conclusion that there is an issue with CAG150 amplification, where there is great heterogeneity of product present after PCR. Also the

CAG150 band is less solid than amplicons with lower numbers of CAG repeats, which shows that something is wrong with amplification and we are not getting the true product.

By use of double digestion, with specific restriction enzymes BamHI and SalHF, of the long PCR product, it is now seen that replication of the repeats occurs in a manner that the repeats alone are able amplify (Figure 5). However, CAG150, as shown by double digestion, still appears as a smear on the agarose gel, since it is not able to amplify completely.

Next, use of intermediate sized G primers were used all CAG lengths, including a new amplicon, CAG120. This new molecule, CAG120, was used in order to fill in the gap between CAG100 and CAG150. CAG120 is seen as a partial smear on the gel, which shows that some of the CAG120 actually amplified, instead of containing completely heterogeneous product, as is seen with CAG150. When these amplicons are then sequenced using the MinIon device, with Nanopore collaboration, this will then let us know what product is being formed from PCR of larger repeats, and will lead us to have greater insight of their heterogeneity.

The stock of G primer amplification was then purified and used as a template in the PCR of even smaller M primer replication. By using the smaller primers, the amplicons are presented almost as the repeat insert themselves. These smaller primers were then used for qPCR, in order to look at the amplification in a repeat-only manner. By qPCR it was found that CAG10 amplifies much faster than a larger CAG insert, CAG40. By future use of other repeats, we will be able to tell what exactly is going on with replication of larger sized molecules, compared to smaller repetitive sequences and non-repetitive sequences.

Future Work:

In order to further explore the length CAG repeats the real time PCR assay must be used in order to monitor the amplification of larger sized repeats. This will be able to inform us on the process of amplification while it is occurring, with the use of the fluorescent dye EvaGreen. Along with amplification monitoring by qPCR, we will be completing thermal melting studies post amplification. These melting studies can only be observed and performed in qPCR, where the fluorescence molecule remains associated with the amplicon itself (Life Technologies, 2012). This melting curve will be able to report the change in fluorescence observed during the conversion of dsDNA to ssDNA due to the release of the dye during this process. The curve is formed by increasing the temperature in small increments and monitor the fluorescence signal at each step (BIO-RAD Laboratories, 2006). As the dsDNA in the reaction denatures, due to increases temperature, the signal will decrease. Fluorescence is plotted against temperature, which the change in both values will be plotted against temperature to obtain a clear graph of melting dynamics (Life Technologies, 2012).

Primer dimers form when two PCR primers, either same-sense primers or same and anti-sense primers, bind to one another instead of the target. During melting curve analysis, these primer dimers can be set apart from the actual amplicon because they have a significantly lower melting temperature. Presence of primer dimers can affect actual template amplification, since they did not anneal to their target. Catching these through analysis can lead us to understand why one product did not amplify well and how it affects the DNA concentration of the sample. A standard curve is able to form from each cycle, since fluorescence increases as the cycle number increases (BIO-RAD Laboratories, 2006). We hope that quantification of fluorescence will be able to report, in a step-wise fashion, what exactly is going on during amplification.

In addition to qPCR, sequencing larger repeats by the use of the MinIon device will lead us to have greater insight of what heterogeneous products are being formed during PCR. The collaboration with Nanopore will be completed using the intermediate sized primers, since the device needs a certain length of nucleic acid in order to have accurate sequencing. After sequencing, we will now be able to see what products are formed and have a greater outlook on what occurs during replication of larger amplicons.

Appendix A: Polymerase Chain Reaction

The polymerase chain reaction (PCR) is a sensitive and specific lab technique, with the ability to address a specific DNA sequence and amplify the sequence to very high copy numbers. It is possible to make this reaction fit any specific sequence with the use of primers that can be developed to bind complementary to the ends of the DNA itself. Choice of DNA depends on what exactly you are trying to study; however, one requirement is that the DNA must be clean and uncontaminated with other DNAs. Also if the target DNA has been replicated using PCR during another reaction, it must be free of all byproducts of the PCR reaction or the current reaction will not be able to work. However, the specificity of primers can sometimes override this contamination, since the primers will still be able to anneal to their complementary target sequence and amplify the correct DNA.

Once the proper target sequence is chosen, the basic reaction components are: water, reaction buffer, forward and reverse primers, polymerase enzyme, and dNTPs (to provide the nucleotides found in DNA itself). The reaction conditions, such as temperature during each transition, must be set as it pertains to each of the PCR reaction steps. Based off of the amount of DNA copies needed for future research, a researcher can come up with exactly how many cycles of DNA the original sequence will undergo. The normal amount of cycles to be completed is around 25-35 cycles, depending on the amount of starting DNA you have. If there is less DNA present in the starting solution, one might want to veer toward higher copy number to produce the same amount of replicons in the end product. Depending on cycle number millions to billions of DNA copies can be produced, because of the exponential nature of replication itself, since the new DNA created is used in the next cycle of amplification.

When preparing the eppendorf tubes for the reaction, a mastermix of each respective length is made for the full amount of desired PCR product. The full mix should then be allocated into each PCR tube, with a max reaction volume of 100uL per tube. In a timely fashion, the DNA polymerase should be added to each reaction mixture and then placed in the pre-set thermal cycler. The entire PCR reaction takes about 5 hours to complete, which can be left overnight since it remains in a refrigerated state until the samples are obtained from the apparatus.

The first step of the reaction involves the denaturation of the DNA itself into single strands, which occurs at a high temperature of 95 degrees Celsius and occurs in a duration of 30 seconds (Intergrated DNA Technologies, 2011). The dsDNA must denature into two single complementary strands in order for the primers to anneal to the ends, which occurs in about 30 seconds as well. The primers added are specific to the complementary handles placed at the ends of the DNA molecules. Starting from the primer, DNA polymerase produces a complimentary copy of the target DNA sequence. This occurs at a temperature of 75 degrees Celsius, optimum for the DNA polymerase itself. Depending on the amount of bases present in the original sequence, the DNA polymerase can continue to replicate for an average of five minutes. These steps repeat for the cycles that are preset on the selected replication program. Replication comes to a halt once all cycles are complete. At this point the temperature is lowered to about 4°C in order to renature and preserve the DNA in its common double stranded state. Analysis of the PCR product by non-denaturing gel electrophoresis is then completed to ensure that PCR reaction was completed properly.

Appendix B: Non- Denaturing Gel Electrophoresis

In order to test whether a PCR product is successful, the products must be loaded into an agarose gel. Current is then applied to the loaded gel and DNA is separated by mass only, known as non-denaturing gel electrophoresis. In order to start off the gel, the proper amount of agarose gel powder must be weighed out. In order to obtain a 0.8% gel (Figure 1 & 2), the percentage must be multiplied by the total gel volume, which would be 30mL. Once the proper powder weight is found, it is then placed into a mixture 10% TBE buffer. In order to create this buffer, a 1:9 ratio of 10x TBE and deionized water, respectively, were mixed together. The powder and buffer are then heated in a flask to a boil and, in a timely fashion, cast into a premade gel frame. A gel comb is placed into the frame, so that the wells will be made and hardened for use with loading. The gel is allowed about 30 minutes to solidify, which gives a glossy appearance when finished and ready to use.

Once the gel is prepared it is time to load the gel with sample. First, new eppendorf tubes are labeled, one for each sample, with an extra tube for the DNA ladder to be loaded in the first well. The sample is added first into each respective tube, and then placed directly back in the refrigerator. Next SYBR green is added, in order to see the fluorescence when scanning is completed later. Loading dye and water are added last, for a total loading volume of 12µL for each sample. The loading mix for the DNA ladder tube is slightly different, containing the premade DNA ladder mix, water and SYBR green. The samples are then loaded into the gel, which is then ran for 40 minutes at 120 volts.

Once the electrophoresis is complete, the gel is taken to the typhoon trio scanner, where it is scanned at 500pmt in order to see the DNA, bound to SYBR green, fluoresce.

Appendix C: DNA Purification

Once the proper PCR products are identified, the next step is to purify the DNA itself. The GeneJet purification kit was used for each of the PCR product mixtures. Separate clean eppendorf tubes must be labeled for each CAG replicon. Next a 1:1 volume of binding buffer to PCR product must be combined in an eppendorf tube. The resulting solution should be properly mixed in order to ensure that all DNA present is bound in order to be separated from the other PCR by-products. Next, half of the mixture should be added to an eppendorf tube containing a purification column and centrifuged for about a minute. The flow-through should be discarded and the previous step repeated with the second half of the solution. As per the GeneJet protocol, 700μL of the wash buffer (containing a high amount of ethanol) is added to the purification column and centrifuged for about a minute to help wash out the by-products of PCR.

The column should then be centrifuged by itself in order to rid of excess ethanol and any other artifact remaining, with the flow-through disposed of. The last step is to obtain the remaining purified DNA in the column, by dispensing 50 μL of elution buffer directly to the center of the column (to ensure the maximum amount of DNA is collected from the column itself). The purified DNA samples should then be stored at a temperature of -20°C, in order to prevent denaturation.

Appendix D: Concentration Analysis by Nano-Drop

Once the pure samples are obtained, Nano-drop can be completed in order to check the DNA concentration of each replicon. By using the ND-1000 program, connected to the Nano-drop apparatus itself, it is possible to make this measurement. In order to make sure there is no residual DNA from previous uses, deionized water and Kim wipes are used to clean off the sensor. Next 1µL of deionized water is loaded onto the ND apparatus, and scanned in order to blank the instrument. 1µL of each sample is then loaded onto the sensor, and scanned for the DNA concentration. The 10mm path length was recorded at 260nm wavelength. If there is a good peak present at this wavelength, it shows that the DNA was replicated and pure.

Reference

Aldrich, J. C., & Maggert, K. A. (2014). Simple quantitative PCR approach to reveal naturally occurring and mutation-induced repetitive sequence variation on the Drosophila Y chromosome. *PloS One*, *9*(10), e109906. http://doi.org/10.1371/journal.pone.0109906

Amrane, S., Saccà, B., Mills, M., Chauhan, M., Klump, H. H., & Mergny, J. L. (2005). Length-dependent energetics of (CTG)n and (CAG)n trinucleotide repeats. *Nucleic Acids Research*, *33*(13), 4065–4077. http://doi.org/10.1093/nar/gki716

BIO-RAD Laboratories. (2006). Real-Time PCR: Applications Guide, 2–4.

Borstnik, B., & Pumpernik, D. (2002). Tandem repeats in protein coding regions of primate genes. *Genome Research*, *12*(6), 909–15. http://doi.org/10.1101/gr.138802

Broda, M., Kierzek, E., Gdaniec, Z., Kulinski, T., & Kierzek, R. (2005). Thermodynamic stability of RNA structures formed by CNG trinucleotide repeats. Implication for prediction of RNA structure. *Biochemistry*, *44*(32), 10873–10882. http://doi.org/10.1021/bi0502339

Cooper, T. a. (2009). Chemical reversal of the RNA gain of function in myotonic dystrophy. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(44), 18433–18434. http://doi.org/10.1073/pnas.0910643106

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews. Genetics*, *5*(6), 435–445. http://doi.org/10.1038/nrg1348

Fan, H., & Chu, J.-Y. (2007). A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics & Bioinformatics*. http://doi.org/10.1016/S1672-0229(07)60009-6

Figueroa, a. a., Cattie, D., & Delaney, S. (2011). Structure of even/odd trinucleotide repeat sequences modulates persistence of non-b conformations and conversion to duplex. *Biochemistry*, *50*(21), 4441–4450. http://doi.org/10.1021/bi200397b

Figueroa, a. a., & Delaney, S. (2010). Mechanistic studies of hairpin to duplex conversion for trinucleotide repeat sequences. *Journal of Biological Chemistry*, *285*(19), 14648–14657. http://doi.org/10.1074/jbc.M109.061853

Galka-Marciniak, P., Urbanek, M. O., & Krzyzosiak, W. J. (2012). Triplet repeats in transcripts: Structural insights into RNA toxicity. *Biological Chemistry*. http://doi.org/10.1515/hsz-2012-0218

Gatchel, J. R., & Zoghbi, H. Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nature Reviews. Genetics*, *6*(10), 743–755. http://doi.org/10.1038/nrg1691

Hartenstine, M. J., Goodman, M. F., & Petruska, J. (2000). Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *Journal of Biological Chemistry*, *275*(24), 18382–18390. http://doi.org/10.1074/jbc.275.24.18382

Hartenstine, M. J., Goodman, M. F., & Petruska, J. (2002). Weak Strand Displacement Activity Enables Human DNA Polymerase to Expand CAG/CTG Triplet Repeats at Strand Breaks. *Journal of Biological Chemistry*, *277*(44), 41379–41389. http://doi.org/10.1074/jbc.M207013200

Intergrated DNA Technologies. (2011). A Basic Polymerase Chain Reaction Protocol, 1–4.

Kashi, Y., & King, D. G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, *22*(5), 253–259. http://doi.org/10.1016/j.tig.2006.03.005

Kiliszek, A., Kierzek, R., Krzyzosiak, W. J., & Rypniewski, W. (2010). Atomic resolution structure of CAG RNA repeats: Structural insights and implications for the trinucleotide repeat expansion diseases. *Nucleic Acids Research*, *38*(22), 8370–8376. http://doi.org/10.1093/nar/gkq700

Koob, M. D., Benzow, K. A., Bird, T. D., Day, J. W., Moseley, M. L., & Ranum, L. P. (1998). Rapid cloning of expanded trinucleotide repeat sequences from genomic DNA. *Nature Genetics*, *18*(1), 72–75. http://doi.org/10.1038/ng0198-72

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Li, Y. C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution*, *21*(6), 991–1007. http://doi.org/10.1093/molbev/msh073

Life Technologies. (2012). Life TechnologiesReal-time PCR handbook, 1–20. Retrieved from http://find.lifetechnologies.com/Global/FileLib/qPCR/RealTimePCR_Handbook_Update_FLR.pdf

Mitas, M. (1997). Trinucleotide repeats associated with human disease. *Nucleic Acids Research*, *25*(12), 2245–2253. http://doi.org/10.1093/nar/25.12.2245

Möncke-Buchner, E., Reich, S., Mücke, M., Reuter, M., Messer, W., Wanker, E. E., & Krüger, D. H. (2002). Counting CAG repeats in the Huntington's disease gene by restriction endonuclease EcoP15I cleavage. *Nucleic Acids Research*, *30*(16), e83. http://doi.org/10.1093/nar/gnf082

Moxon, R., Bayliss, C., & Hood, D. (2006). Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics*, *40*, 307–333. http://doi.org/10.1146/annurev.genet.40.110405.090442

Ning, L., Yang, D., Gao, T., Lu, S., Yin, Y., & Li, G. (2015). A new method to evaluate trinucleotide repeats length polymorphism. *Talanta*, *143*, 414–418. http://doi.org/10.1016/j.talanta.2015.05.016

Orr, H. T., & Zoghbi, H. Y. (2007). Trinucleotide Repeat Disorders. *Annual Review of Neuroscience*, *30*(1), 575–621. http://doi.org/10.1146/annurev.neuro.29.051605.113042

Salinas-Rios, V., Belotserkovskii, B. P., & Hanawalt, P. C. (2011). DNA slip-outs cause RNA polymerase II arrest in vitro: Potential implications for genetic instability. *Nucleic Acids Research*, *39*(17), 7444–7455. http://doi.org/10.1093/nar/gkr429

Santillan, B. A., Moye, C., Mittelman, D., & Wilson, J. H. (2014). GFP-Based Fluorescence Assay for CAG Repeat Instability in Cultured Human Cells. *PloS One*, *9*(11), e113952. http://doi.org/10.1371/journal.pone.0113952

Sinden, R. R., Potaman, V. N., Oussatcheva, E. a, Pearson, C. E., Lyubchenko, Y. L., & Shlyakhtenko, L. S. (2002). Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *Journal of Biosciences*, *27*(1 Suppl 1), 53–65. http://doi.org/10.1007/BF02703683

Sobczak, K., Michlewski, G., De Mezer, M., Kierzek, E., Krol, J., Olejniczak, M., … Krzyzosiak, W. J. (2010). Structural diversity of triplet repeat RNAs. *Journal of Biological Chemistry*, *285*(17), 12755–12764. http://doi.org/10.1074/jbc.M109.078790

Tsoi, H., Lau, T. C.-K., Tsang, S.-Y., Lau, K.-F., & Chan, H. Y. E. (2012). CAG expansion induces nucleolar stress in polyglutamine diseases. *Proceedings of the National Academy of Sciences*. http://doi.org/10.1073/pnas.1204089109

Usdin, K. (2008). The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*. http://doi.org/10.1101/gr.070409.107