UNIL | Université de Lausanne

Faculté de biologie et de médecine

Ecole de biologie

RINGS: A NOVEL METHOD TO MEASURE RATES OF TRINUCLEOTIDE REPEAT INSTABILITY USING NEXT GENERATION SEQUENCING

Travail de Maîtrise universitaire ès Sciences en Sciences moléculaires du vivant
Master Thesis of Science in Molecular Life Sciences

par

Evgeniya TROFIMENKO

Directeurs : Dr. Keith Harshman et Prof. Vincent Dion
Expert : Prof. Ioannis Xenarios
Centre Intégratif de Génomique (CIG)

Abstract

The expansion of trinucleotide repeats (TNR) is the cause of at least 19 neurological disorders. TNR are unstable during the lifetime of a patient in his or her somatic tissues. The instability is biased towards expansions, which is thought to precipitate pathogenesis. Therefore understanding repeat instability is of paramount importance. One major roadblock is that the current methods to assess variation in the number of TNRs are tedious, biased against expansions, not quantitative and/or lack resolution. The goal of this project is to develop an assay, Repeat Instability with Next Generation Sequencing (RINGS), that would increase the throughput and resolution significantly. To do so, I used Single Molecule sequencing in Real Time (SMRT) method developed by Pacific Biosciences, which consists of using a rolling circle replication of single-molecules, thereby creating several large reads (14-30kb), in combination with single DNA molecule amplification. PacBio sequencing approach was rendered more time and cost efficient through multiplexing, meaning that several samples were sequenced in parallel. Based on the results RINGS works in principle, but will require some optimisation. This approach will allow a more precise look at instability within cells with a range of TNR lengths at a larger scale than the current methods.

Résumé

L'expansion des trinucléotides répétés (TNR) est à l'origine d'au moins 19 maladies neurologiques. Les TNR sont instables dans les tissues somatiques d'un patient au cours de sa vie. L'instabilité est biaisé en faveur des expansions, qui induirait la pathogenèse. Par conséquent, la compréhension de l'instabilité des répétitions est d'une importance primordiale. Un obstacle majeur est que les méthodes actuelles pour évaluer la variation de l'instabilité des TNR sont fastidieux, biaisés en faveur des expansions, non quantitatives et/ou manquent de résolution. Le but de ce projet est de développer un essai, Repeat Instability with Next Generation Sequencing (RINGS), qui permettrait d'accroître le débit et la résolution de manière significative. Pour ce fait, j'ai utilisé en premier lieux l'amplification des molécules d'ADN uniques puis j'ai suivie par la méthode Single Molecule sequencing in Real Time (SMRT) développé par Pacific Biosciences, qui consiste à utiliser une réplication circulaire donnant molécules simples, créant ainsi plusieurs grandes séquences (14-30kb). De plus cette méthode de séquençage a été rendu plus efficace grâce au multiplex qui est un moyen d'analyser plusieurs échantillons en même temps ce qui a donné un gain de temps et d'argent. Basé sur les résultats, RINGS fonctionne mais a besoin d'une optimisation. Cette approche permettra un regard plus précis sur l'instabilité dans les cellules et donnera plus de données sur les différentes longueurs du TNR.

Table of content

1. Abbreviations

APRT	Adenine phosphoribosyltransferase
BER	Base excision repair
CCS	Circular consensus sequences
CRISPR-Cas9	Clustered regularly interspaced short palindromic repeats
ePCR	Emulsion PCR
gDNA	Genomic DNA
GFP	Green fluorescent protein
HD	Huntington's Diseases
HPRT	Hypoxanthine–guanine phosphoribosyl- transferase
Indels	Insertions, deletions
MMR	Mismatch repair
Msh2	MutS homologue 2
NGS	Next generation sequencing
oPCR	Open PCR
PacBio	Pacific Biosciences
PAM	Protospacer-adjacent motif
RINGS	Repeat instability with next generation sequencing
SCA10	Spinocerebellar ataxia type 10
siRNA	Small interfering RNA
SMRT	Single molecule sequencing in real time
SP-PCR	Small pool PCR
STR	Short tandem repeats
SV40	Simian virus 40
TC-NER	Transcription coupled nucleotide excision repair
TNR	Trinucleotide repeats
URA3	Orotidine-5'-phosphate decarboxylase
XPA	Xeroderma pigmentosum, complementation group A

2. Introduction

2.1 Short tandem repeats and genomic instability

Over 30% of the human genome is made up of repeated sequences (Kass & Batzer 2001). The majority of these repeats are long terminal repeats and transposable elements. However, shorter repeat units have also been identified. These include mini and microsatellites. Minisatellites can be up to 20 kb in length with repeat units between 10 and 60 bp. This type of repeats is commonly found at chromosome ends. In contrast, microsatellites or short tandem repeats (STR) are dispersed throughout the genome. The unit size of STR is considerably shorter, with 1-6 bp repeat units, yet they can reach several kilobases in total length.

STRs of at least 12 bp in length are found in many organisms such as plants (Sureshkumar et al. 2009), yeast (Vinces et al. 2009), and mammals (Tóth et al. 2000). The genomic distribution and repeat unit content is characteristic of the taxonomic groups, such as mammal, rodentia, embryophyta and fungi (Tóth et al. 2000). STRs are present in the genome of every person, and can exist without provoking any deleterious effects. However, variation in the number of units of some tandem repeats cause over 30 human diseases (Cleary & Pearson 2005), including several neurological and neurodegenerative disorders (Orr & Zoghbi 2007) (Table 1).

A number of these diseases are associated with an increase in the number of units in the repeat tract (a phenomenon referred to repeat expansions). In each disease, a specific locus and a specific tandem repeat is implicated. The most common expansions are those of trinucleotide repeats (TNR). For example, Huntington's disease (HD) is an autosomal dominant neurodegenerative disorder which is associated with CAG/CTG expansions beyond 35 units, within the coding region of the huntigntin gene (Htt)(Group 1993) (Table 1). The mutated huntingtin protein accumulate in the nucleus of striatal neurons and progressively forms insoluble, amyloid-like structures (Scherzinger et al. 1997). Symptoms of this disorder include involuntary movements and loss of cognitive functions. In addition, HD exhibits anticipation, which is the transmission of the disorder from one generation to the next, and is often related to earlier age of onset of symptoms and increased severity of the disorder. This phenomenon can be explained by the size of the initial repeat tract affecting the age of onset of pathogenesis, as well as the pattern and the rate of instability in the organism (Kennedy et al. 2003). After transmission of the disorder to offspring, the expansions continue in somatic cells. This type of expansions participates even further to disease pathogenesis. In an HD mouse model, the age of symptoms onset can be delayed through suppression of somatic expansions (Budworth et al. 2015). In addition, the pathogenesis of the disorder is aggravated with normal age associated decline, such as oxidative damage. In fact, by interfering with removal of oxidized bases decreases somatic expansions rates in HD mouse model (Kovtun et al. 2007). Taken together, pathogenesis of such disorders is complex and is determined not only by the degree of instability of the repeats, but also by factors that may influence the expansions.

TNRs that are associated with diseases and that are unstable invariably form unusual, non-B-DNA structures during replication, recombination and repair (Figure 1). The exception to this is SCA10, in which the AT-rich pentanucleotide repeat unwinds itself instead of forming stable secondary structures (Potaman et al. 2003). The formation of secondary structures is favoured by single stranded DNA, such as those occurring during DNA repair, recombination, and lagging- and leading-strand replication. Additionally, polymerase slippage or misalignment during replication on either leading or lagging stand may result in secondary structure formation. An *in vitro* study showed that human polymerase β is slowed down in the repeat tract (Kang et al. 1995), which is not the case for the helicase. This is resolved by polymerase 'skipping' a segment within the repeat tract, leading to hairpin formation and to variation in the number of TNRs. The stability of the secondary structures depends of the composition of the repeat tract. For example, hairpins formed by CTG repeats are stronger than those formed by CAG repeats. This suggests that some secondary structures can have more effect on instability. That is because these aberrant secondary structures also stall the progression of the RNA and DNA polymerases. Furthermore, secondary structures interfere with DNA repair mechanisms (López Castel et al. 2010; McMurray 2010). In fact, hairpins formed at repeat regions are recognised as damaged DNA and misrepaired, leading to either expansions or contractions.

Table 1: Unstable repeat disorders. Adapted from Orr and Zoghbi 2007, McMurray 2010.

Disease	Gene product	Repeat	Expansion range	Symptoms
ALS/FTD	C90RF72	GGGGCC	22-90	Muscle weakness, atrophy, dysphagia, dysarthria, behavioural changes, speech and language problems, memory difficulties
DM1	Dystrophia myotonica - protein kinase	CTG	50–10,000	Myotonia, weakness, cardiac conduction defects, insulin resistance, cataracts, testicular atrophy, and mental retardation in congenital form
DM2	Zinc finger 9	CCTG	75–11,000	Similar to DM1 but no congenital form
DRPLA	Atrophin	CAG	49-88	Ataxia, seizures, choreoathetosis, dementia
FRAXA	Fragile site mental retardation 1	CGC	>200 (full mutation)	Mental retardation, macroorchidism, connective tissue defects, behavioral abnormalities
FRAXE	Fragile site mental retardation 2	CCG	200–900	Mental retardation
FRDA	Frataxin	GAA	200–1700	Sensory ataxia, cardiomyopathy, diabetes
FCD	Transcription factor 4	CTG	>50	Blurred and distorted vision, reduced contrast perception, blisters on cornea surface
FXTAS	Fragile site mental retardation 1	CGG	60–200 (premutation)	Ataxia, tremor, Parkinsonism and dementia
HD	Huntingtin	CAG	36–121	Chorea, dystonia, cognitive deficits, psychiatric problems
HDL2	Junctophilin	CTG	66–78	Similar to HD
SBMA	Androgen	CAG	38–62	Motor weakness, swallowing, gynecomastia, decreased fertility
SCA1	Ataxin 1	CAG	39–82	Ataxia, slurred speech, spasticity, cognitive impairments
SCA2	Ataxin 2	CAG	32–200	Ataxia, polyneuropathy, decreased reflexes, infantile variant with retinopathy
SCA3	Ataxin 3	CAG	61–84	Ataxia, parkinsonism, spasticity
SCA6	CACNA1A	CAG	10–33	Ataxia, dysarthria, nystagmus, tremors
SCA7	Ataxin 7	CAG	37–306	Ataxia, blindness, cardiac failure in infantile form
SCA8	C10orf2	CTG	>74	Ataxia, slurred speech, nystagmus
SCA10	Ataxin 10	ATTCT	500–4500	Ataxia, tremor, dementia
SCA12	Protein phosphatase 2 beta	CAG	55–78	Ataxia and seizures
SCA17	TATA box binding protein	CAG	47–63	Ataxia, cognitive decline, seizures, and psychiatric problems

In fact, nearly all DNA repair pathways are involved in instability. In a HD mouse model lack of both Msh2 alleles, leads to disruption the first step of mismatch repair pathway (MMR), which is the recognition of a DNA loop by MutS complexes containing Msh2. This results in significant decrease in expansions in comparison to mice with wild type Msh2 expression, suggesting its involvement in somatic expansions (Manley et al. 1999). siRNA knock down of XPA that is involved in transcription coupled nucleotide

excision repair (TC-NER), leads to reduced contractions in human cells (Lin et al. 2006). In the same study, by targeting both Msh2 and XPA together and separately, lead to similar contractions frequency, which suggests interaction between MMR and TC-NER pathways (Lin et al. 2006). On the other hand, interference with base excision repair (BER) pathway through loss of OGG1, which has a major role in removal of oxidized bases, results in suppression of somatic expansions in a mouse model for HD (Kovtun et al. 2007). Taken together, deficiencies of various repair pathways have been shown either to prevent or induce expansions or contractions.

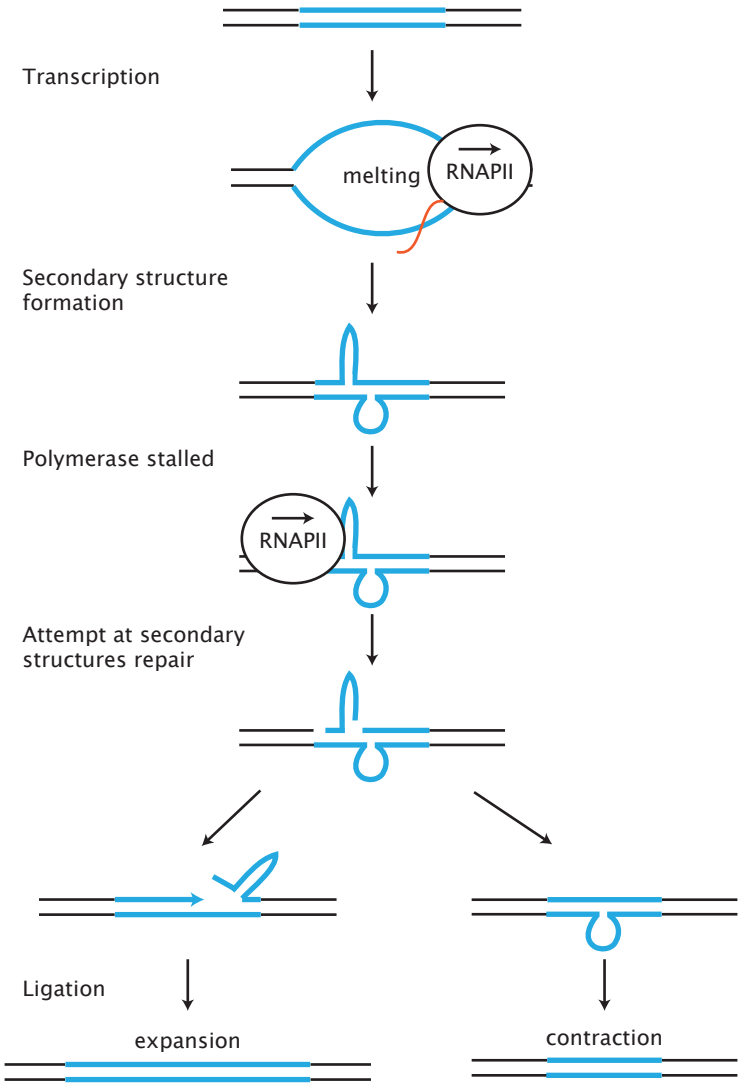


Figure 1: Illustration of common pathway leading to instability. Adapted from (Dion 2014).

Many aspects of the mechanisms by which the instability occurs can be targets for potential treatments of the associated disease. It has been shown that the severity of the repeat expansion associated disorders depends in large part on the length of the expanding repeat tract (McMurray 2010). One can hypothesise that inducing specifically contractions within the repeat tract could alleviate the symptoms. Therefore, being able to determine precisely the rate of trinucleotide instability is key to predicting the development of the disorder and the efficiency of various treatments.

2.2 Current methods applied to study repeat instability

The methods of measuring instability are critical when deciphering the mechanism of repeat instability. Unfortunately, the existing methods are often slow and tedious. For instance, the classic way to look at TNR expansions is to perform a Southern Blot with a probe specific for the repeat region. In this case, each sample takes several days to process and the instability appears as a smeary band that is difficult to quantify.

An easier way to quantify instability is by amplifying the TNRs from genomic DNA (approximately 100ng) with a labelled oligonucleotide. The product is then analysed by capillary electrophoresis and the signal is quantified using the software GeneScan (Mangiarini et al. 1996). This approach is especially prone to PCR artefacts including the preferential amplification of shorter alleles.

One way to prevent PCR biases is to combine Southern Blotting with small pool PCR (SP-PCR) (Monckton & Caskey 1995). The goal of SP-PCR is to conduct in parallel separate reactions amplifying a few DNA molecules. To achieve this, DNA is diluted to just a few amplifiable genomes per tube. The product from such low amounts of DNA is not visible on an agarose gel and is therefore transferred to a membrane and detected with a labeled probe against the expected PCR product. However this method is time consuming, it takes up to two weeks of tedious work per sample and is therefore impossible to perform on a large scale.

Reporter-based assays can also be used to investigate TNR instability. Two types of assays have been developed for use in mammalian cells, one plasmid-based and the other chromosomal. The plasmid-based assays take advantage of tissue culture followed by transformation in either yeast or bacteria (Cleary et al. 2001; Pelletier et al. 2005; Farrell & Lahue 2006). In these assays, a shuttle vector containing TNRs is transfected into mammalian cultured cells. Due to presence encoded within the shuttle vector of the SV40 viral origin of replication and the large T antigen, the vector can replicate in cultured cells. After allowing for the DNA replication to take place over a few days, the cells are lysed and the shuttle vector is recovered. The shuttle vector is then transformed into yeast, where TNR contractions can be detected through the TNR-URA3 reporter that is also contained within the vector. Alternatively, the DNA recovered from mammalian cells is transformed into bacteria where plasmid preparation can be made and the size of the repeat analyzed using restriction enzymes. A limitation of these assays is that the plasmids do not necessarily reflect the state of the endogenous chromatin. Moreover, the assays cannot detect both expansions and contractions with the same reporter construct – requiring multiple reports to do this, and making the experiments yet more tedious.

Several chromosomal reporter assays have been developed that offer the advantage of investigation TNR stability under conditions that are closer to physiological. These assays take advantage of the fact that repeat tracts inserted within an intronic region disrupt the expression of reporter genes. In early versions of this assay, CAGs were integrated within an intronic sequence of the *APRT* or *HPRT* genes of mammalian cells. The CA motifs within the expanded repeats act as splicing enchancers (Blencowe 2000). As the number of the repeats increases, splicing becomes more efficient and the CAG repeat tract becomes an alternative exon with an additional 38 bp downstream of TNRs. This results in the 2nd part of the reporter gene being out of frame, disrupting APRT or HPRT protein expression (Gorbunova et al. 2003). Clones with less than 35 repeats synthesize enough APRT or HPRT protein to survive selection and appear as wild type, making this assay capable of selecting clones specifically with contractions through reporter gene analysis. Further optimization of the HPRT reporter, for example by incorporating a TetON inducible promoter, makes it possible to test the correlation between the repeat expansions and transcription efficiency through the tract (Lin et al. 2006). These assays are useful because they can be coupled to siRNA knockdown and pharmacological treatments. They are less labour intensive than the plasmid-based approaches but they are slow, requiring that the single colonies grow from single-cell. In addition, they are restricted to investigating contractions and are blind to expansions.

A more recent example of a chromosomal reporter in mammalian cells is described by Santillan and colleagues (Santillan et al. 2014). In this assay a GFP (green fluorescent protein) reporter, under the control of an inducible promoter, is split into two parts by an intron containing a different number of TNR units. The resulting cell line is designated GFP$(CAG)_X$, where X is the number of repeats present within the GFP intron. The principle here is the same as in the APRT and HPRT-based assays described above where the CAG repeat interferes with splicing in a length-dependent manner. A major advantage of this assay is that it does not require an all or none selection because the GFP intensities scale with the size of the repeat. Repeat instability in this system will affect the change of the GFP fluorescence. This assay has the major advantage of being fast, although it measures only indirectly the repeat size. Whereas changes in repeat length will cause a change in GFP levels, changes in GFP will not necessarily reflect repeat size changes. This is because, for example, loss of Tet repressor will give rise to highly GFP fluorescent cells, which can be mis-assigned as having contractions while loss of reporter gene would give rise to non-fluorescent cells.

Even though there are many approaches available to assess variation in the number of TNRs, all of the currently available methods are tedious, not quantitative, are indirect, and/or lack resolution. There is, therefore, a great need for improved assays to measure repeat instability efficiently.

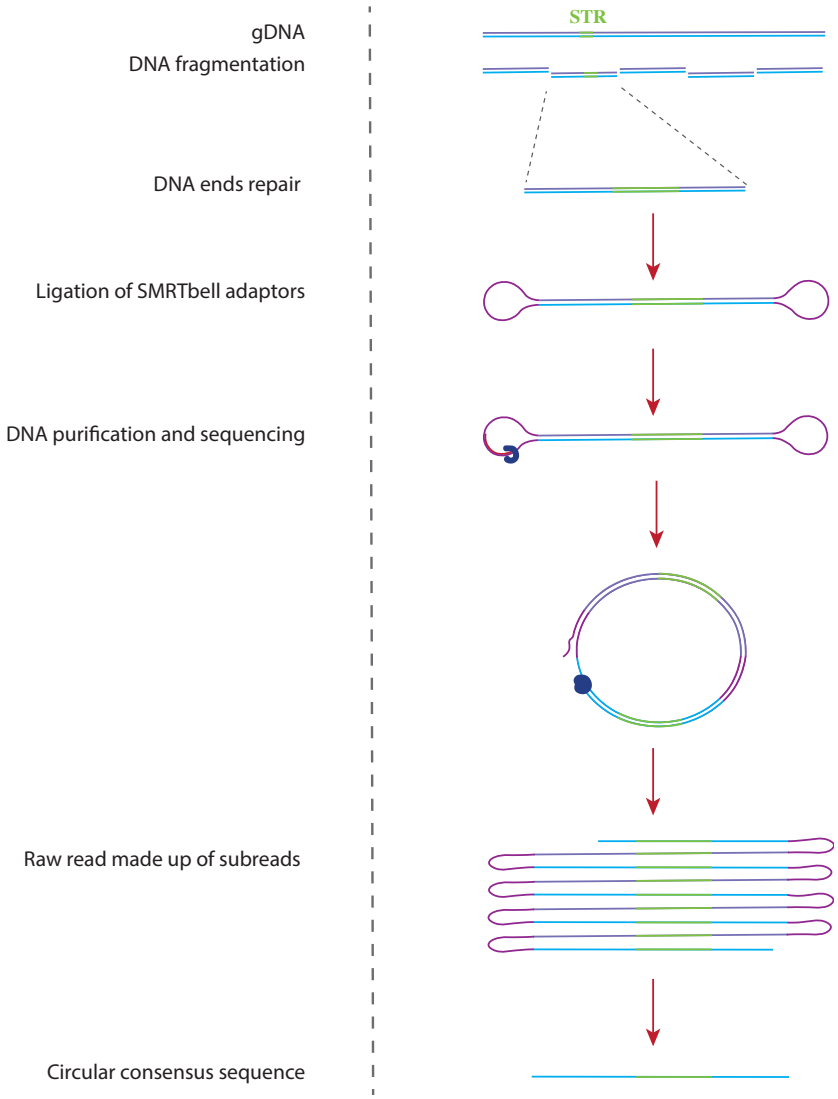


Figure 2: Major steps of SMRT sequencing. Adapted from (Travers et al. 2010).

2.3 Improving the 'gold standard'

SP-PCR combined with Southern Blot is considered the gold standard to measure the rates of trinucleotide instability. However, this method is far from ideal and at least two aspects of this approach can be improved. First, is to set up a more efficient method for single DNA molecule amplification and second, to have a faster, more precise and more quantitative approach to quantify the instability. The goal of this thesis was to improve both of these limitations by developing a method that combines emulsion PCR (ePCR) and a high throughput, single molecule sequencing method.

2.3.1 Emulsion polymerase chain reaction

The major benefit of the gold standard method is that it prevents the biased amplification of short repeats over longer ones. However, to make SP-PCR efficient, a large number of separate PCRs must be run in parallel, which is tedious and time consuming. A way to improve on this approach is to take advantage of the

ePCR method described by Dressman and colleagues and then simplified by Gloekler's group (Dressman et al. 2003; Schütze et al. 2011).

The goal of emulsion PCR is the same as that of SP-PCR, which is to conduct individual reactions, each amplifying single DNA molecules. With ePCR this can be done in a single tube (500 ul reaction). ePCR mix is made up of two liquid phases: oil and aqueous. Mixing the two phases together, leads to formation of droplets (approx. 1 x 10^9 droplets per reaction (Schütze et al. 2011)). The DNA amplification occurs only within the aqueous phase droplets, each of which should contain approximately 1 genome, which is expected to reduce the PCR bias. ePCR has successfully been used for amplification of DNA libraries in aptamer selection (Shao et al. 2011) and for quantification of lung cancer related microRNA (Wang et al. 2015), but has not yet been tried for TNR instability.

2.3.2 Next generation sequencing

Thanks to recent technologies the most straight forward way to gain single-molecule view of a large number of DNA templates is using Next Generation Sequencing (NGS) methods developed by Illumina, Pacific Biosciences (PacBio) and other companies. These methods have been used to look at polymorphisms and tandem repeats (Loomis et al. 2013; Doi et al. 2014).

Illumina sequencing is based on a sequencing cluster of molecules, which is generated by template amplification. The main disadvantage of this method is that it generates only short reads of maximum 125 nucleotides for Illumina HiSeq and 300 nucleotides for Illumina MiSeq. These methods are useful to identify changes in the length of microsatellites between different chromosomal alleles, for example. Assembly and precise determination of the length of pathogenic alleles would be impossible with such short reads, as seen with Illumina sequencing of patients with over 100 TNRs.

The Single Molecule sequencing in Real Time (SMRT) method developed by PacBio allows one to obtain reads reaching up to 40 kb in length (Eid et al. 2009). The DNA sequencing polymerase can make multiple passes around the circular sequencing template thereby allowing the generation of a consensus sequence of the sequencing template molecule (Figure 2). Although this sequencing system has a relatively high error rate for any single sequencing pass, the consensus sequence generated from these multiple passes – called the Circular Consensus Sequence (CCS) - is highly accurate. Therefore, because of its long read length and highly accurate CCS, the PacBio sequencing method appears to be well adapted to studies involving short tandem repeats. Indeed, the PacBio technology has been recently shown to work with CGG repeats. Here I aim to adapt and improve this method for CAG repeats and to measure repeat instability rather than simply the length of the repeat.

3. Research plan

Here, I propose to use a novel approach, Repeat Instability with Next Generation Sequencing (RINGS), to measure TNR instability on a larger scale. This approach combine single DNA molecule amplification by SP-PCR with NGS method developed by PacBio, as described above. The PacBio sequencing approach can be rendered more time and cost efficient through multiplexing, meaning that several samples can be sequenced in parallel. By using sample specific barcodes, demultiplexing can be done during the post sequencing analysis. Here, I attempt to sequence 4-5 DNA samples in parallel. To determine sensitivity and resolution of RINGS, I will use GFP(CAG) cells with approximately 15, 50, 97 and 270 CAG repeats and to determine whether multiplexing barcodes have an effect on sequencing output and sample integrity.

Taken together this study combines optimisation steps of the RINGS and its use on human cells instability model.

I address the following questions:
1. Can RINGS method be used to measure the rates of TNR instability?
2. Which optimisation steps can be taken to improve the efficiency of RINGS?
3. Can SP-PCR be replaced by ePCR?

A similar study, which combines PCR of CGG repeats on large amounts of DNA and SMRT sequencing approach was done by Loomis and colleagues in 2013 to use it as a diagnostics tool for patients with fragile X syndrome (Loomis et al. 2013). As a proof of concept sequencing through plasmids with known number of repeats, showed that sample heterogeneity arises from the repeat region and not the flanking sequences. The patient sample was estimated at 720 CGG units on average. However, this sample is much more heterogeneous, which reflects not only TNR instability within patient cells, but also PCR associated artefacts.

Although sequencing through the full mutation allele that is incredibly GC-rich, could be done using bulk PCR amplified template, few reads per sequencing unit were of sufficient quality to assess repeat size. Therefore, the sequencing data was pooled from several SMRT cells together, making the approach expansive and precludes multiplexing. Since PacBio sequencing is error-prone and the polymerase is expected to stall at secondary structures within the template, having a large number of reads is useful for averaging out the sequencing data and having a higher quality final result. In addition, a severe length bias during the loading of the flow cell for sequencing was detected. This means that shorter DNA fragments are more likely to reach the DNA polymerase at the bottom of SMRT cell than the longer fragments, skewing the instability assessment. This bias needs to be corrected by using an algorithm along with the loading of fragments of known sizes.

Prior to RINGS analysis the cell will be submitted to two types of treatment. First, long term passaging of human cells and second, CRISPR-Cas9 nickase genome editing technology. All of the analysis are done in the GFP(CAG) HEK293 cells supplied with an inducible promoter, as described above (Santillan et al. 2014). The two replicates of long term passaging samples were used to confirm that multiplexing is possible, meaning that barcodes do not impact on the sequencing reaction and that a larger spread of repeat size can be measured from a single heterogeneous sample. The expected result from long term treatment experiment is that cells should display a significantly larger range of genomic instability specific to the CAG/CTG integrated repeat locus, which can be measured with RINGS.

A treatment known to influence instability is a genome editing technology derived from bacterial adaptive immunity and based on clustered regularly interspaced palindromic repeats (CRISPR). CRISPR-associated protein 9 (Cas9) is an endonuclease. When complexed with a single 20 bp, guide RNA that recognises a sequence of interest within the genome, Cas9 can cut the target DNA by inducing a double strand break. The targeting specificity is increased, by the presence of protospacer-adjecent motif (PAM). PAM is a short three nucleotide sequence, NGG, where N can be any of four nucleotides. However, recent study showed that NAG can also be used as PAM (Hsu et al. 2013). A modification to D10A catalytic site of Cas9 has been made, which turns the nuclease into a nickase, generating single-strand breaks rather than double stand break (Cong et al. 2013). In this study I will use the modified CRISPR-Cas9 nickase, PAM CAG recognition sequence and guide RNA targeting CTG repeats within GFP(CAG)101 cell line. The results from flow cytometry analysis and SP-PCR-Southern bloting suggest that the Cas9 nickase treatment induces preferentially contractions of CAG/CTG repeats within GFP mini gene (Cinesi et al. submitted). Here, I

would like to determine whether contractions induced by the CRISPR-Cas9 nickase can be detected with RINGS.

Another goal of this project is to set up a protocol for ePCR and to determine whether it can replace the time consuming SP-PCR. I will start with plasmid DNA amplification and then I will move on to genomic DNA from same samples that will be used for RINGS.

4. Results

4.1 Sample and library preparation

As an initial proof of concept for RINGS, I used samples with CAG repeat length estimated with Sanger sequencing at 15, 50, 101 and over 270 repeats. These samples were first used to evaluate sensitivity and resolution of RINGS in comparison to Sanger sequencing.

The initial SP-PCR was performed using primers located 294 base pairs (bp) upstream and 230 bp or 1341 bp downstream of the repeat tract for the first and second libraries, respectively (Figure 3A and B). The template preparation was done in an identical fashion for all the samples (Figure 3A). In each primer pair, one of the primers either on 3' or 5' end has an additional 16 bp sequence, a barcode, specific to each sample (see Material and Methods). According to Fragment Analyzer the prepared samples were of sufficient quality and of expected sizes (see Annex 9.1). Any fragment of below 524 bp or 1.5 kb, depending on the reverse primer used, are the result of unspecific amplification and primer dimers. These fragments were removed during the library preparation, using a PacBio kit with 500 bp or 2 kb cutoff, respectively for library 1 and 2 (Figure 3C and E).

However, after library preparation for the first library the recovery rate was only 5% with the optimum being between 30% and 40%. The main hypothesis to explain this, was that the DNA repair step of the library preparation kit contains enzymes that could remove molecules containing stable secondary structures in both the repeat tract and the flanking sequences. This hypothesis was tested with UNA fold tool from Integrated DNA Technologies and by preparing an additional library without the Repair DNA Damage step of the kit. Analysis of flanking sequences with UNA tool yielded negative values, which suggest that secondary structure formation is favourable within the template, including the flanking sequences (Figure 3D). On the other hand, the recovery rate of the second library made with the Repair DNA Damage step was 23.2% and without 28.7%. Indeed, the secondary structures and the DNA repair step appears to have an effect on the recovery rate. However, it does not explain fully the 5% recovery rate obtained with the first library. Therefore, other factors should be considered, such as other steps of the library preparation, as well the quality of the DNA template and the amount of primer dimers that were eliminated. Nevertheless, both libraries were deemed of sufficient qualities for sequencing.

Table 2: RINGS productivity comparison for Library 1 and 2 according to type of loading and presence of DNA damage repair step during library preparation

	Loading	DNA damage repair	Productivity %		
			0	1	2
Library 1	Diffusion	yes	70.08	2.07	27.85
	MagBeads	yes	93.16	5.58	1.26
Library 2	MagBeads	yes	11.68	63.45	24.88
	MagBeads	no	8.53	56.31	35.17

4.2 Quality control post sequencing

The output of SMRT cell sequencing is a several kilobases (kb) long raw read for each coordinate on the SMRT cell sequencing unit. The raw read contains the template amplified in forward and reverse, separated by the SMRT bell adaptors (Figure 2). These raw reads are analysed with with SmrtPipe using SMRTportal through the analysis steps including kinetics, base calling, adaptor trimming and filtering, and finally processing into a consensus sequences (CCS). The library quality control post sequencing is also done using SMRTportal.

One of initial indications of efficiency of the sequencing run are the productivity values. The results are separated into three productivity groups: 0, 1 and 2 (Table 2). Productivity 0 corresponds to the wells within

SMRT cell, where sequencing did not occur. Productivity 2 indicates the percentage of wells that contained more than one DNA molecule, which means that the signal coming from these wells cannot be attributed with certainty to only one DNA molecule. Wells with Productivity 1 have only one DNA molecule and

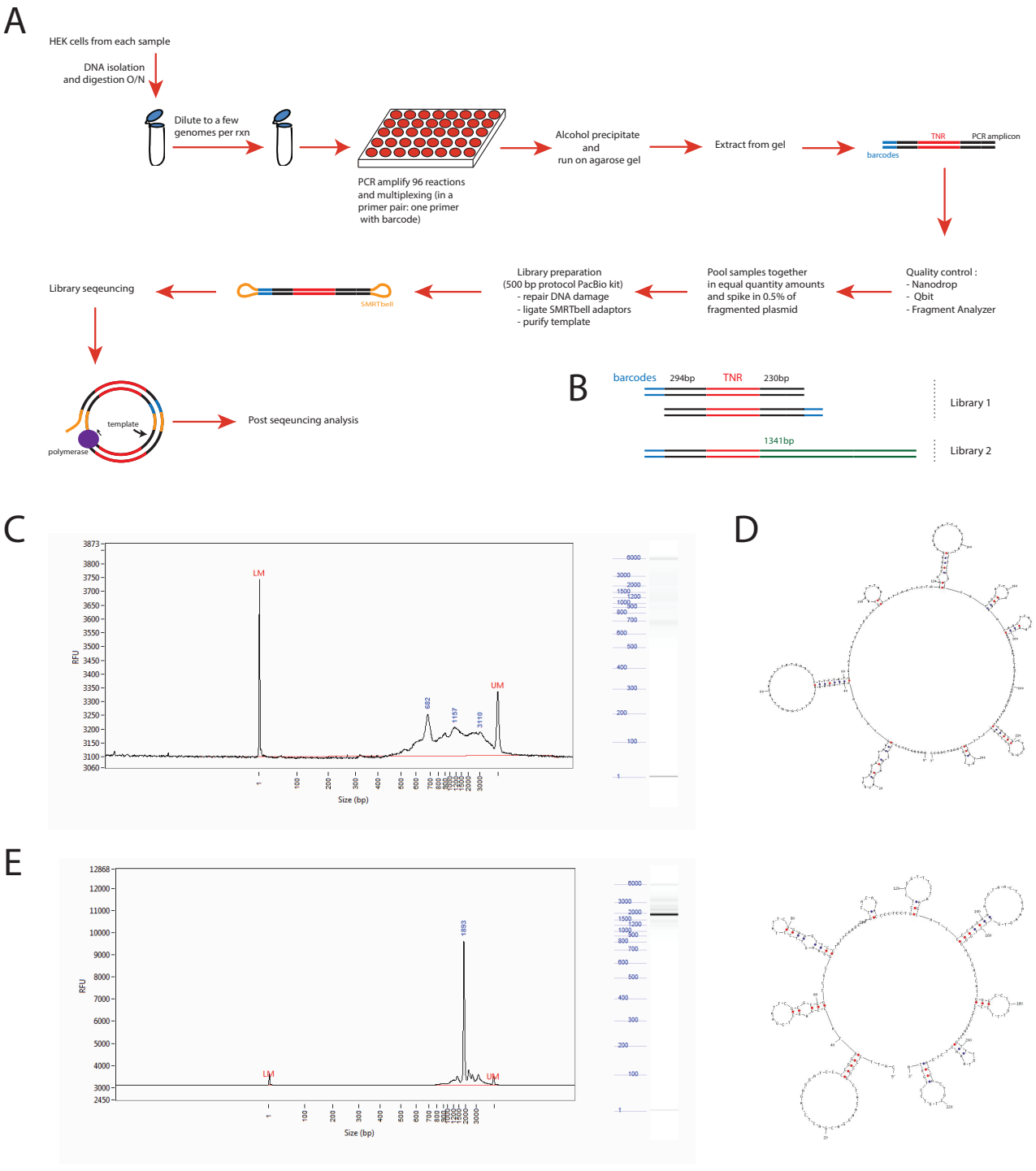


Figure 3: Pre-sequencing template analysis
A: Step by step template preparation
B: Scheme of the final template for each library
C: Fragment analyzer result after lirary preparation for Library 1. Right panel: acrylamide gel. Left panel: spectrum of intensities of bands detectable on the gel.
D: Prediction of possible DNA secondary structure using UNA fold tool. Left panel: 5' end of the sequence flancking the repeats. Right panel: 3'end of flancking sequence following the repeats.
E: Fragment analyzer result after lirary preparation for Library 2. Right panel: acrylamide gel. Left panel: spectrum of intensities of bands detectable on the gel.

polymerase per well, therefore the sequence content can be precisely determined. Only sequences with productivity 1 and minimum quality of 90% are used in the subsequent manual steps of the data analysis (Annex 9.4). For Library 1, the Productivity 1 values are much lower than the optimum 40%. However, for Library 2, all the productivity values were in the range of what is to be expected, with productivity 1 value being over 50%. In fact, for all of the sequencing output characteristics including number of reads and the number of CCS have increased over 10 times for the second library (Table 3). At the moment, there is no specific hypothesis to explain this observation.

Coverage results, which represent the number of subreads per raw read, from both libraries suggest that most of the molecules are sequenced once (Figure 4A and B). Nevertheless, there are enough molecules to get CCS, most of which display a high number of passes, which refers to the number of subreads per CCS. On average there are 23 and 10 passes per CCS for library 1 and 2, respectively (Figure 4C and D). This explains the high CCS quality rate, which is approximately 98% instead of 90% expected.

Taken together, it can be concluded that the template containing various number of TNR units can be sequenced through with high accuracy using PacBio and yields enough data to proceed with manual analysis and demultiplexing.

Table 3: Sequencing characteristics output for Library 1 and Library 2

	Library 1		**Library 2**	
	Diffusion Loading	MagBead Loading	with DNA damage repair	without DNA damage repair
Number of reads	3114	8378	190696	169237
Mean read length (bases)	5607	2532	12221	12876
Mean read quality	81%	83%	98.17%	98.08%
Number of subreads	18467	27227	1253607	1154262
Mean subread length (bases)	754	748	1819	1847
Number of CCS	408	630	89767	79299
Mean CCS length	716	845	1505	1444

4.3 RINGS can be used to assess the number of CAG units within GFP(CAG)$_{101}$ cells

Currently there are two available ways to load the DNA samples onto the SMRT cell sequencing unit. The first is diffusion loading, where the sample is allowed to passively enter the wells within the sequencing unit. MagBeads loading is done with the use of paramagnetic beads coated with dT oligo. The template after library preparation has a polyA fragment attached to the SMRTbell adaptors, which can bind to polyT on the surface of MagBeads. Bound magnetic beads roll on top of the SMRT cell wells and only the template complexes of certain length can reach the bottom of the wells and remain there for sequencing (Pacific Biosciences 2012). Therefore MagBeads provide minimum length-selective loading of template. The first library was sequenced using either diffusion or MagBeads loading. From the point of view of the quality control and data output there is only a slight improvement for the samples sequenced after MagBead loading (Table 2 and 3).

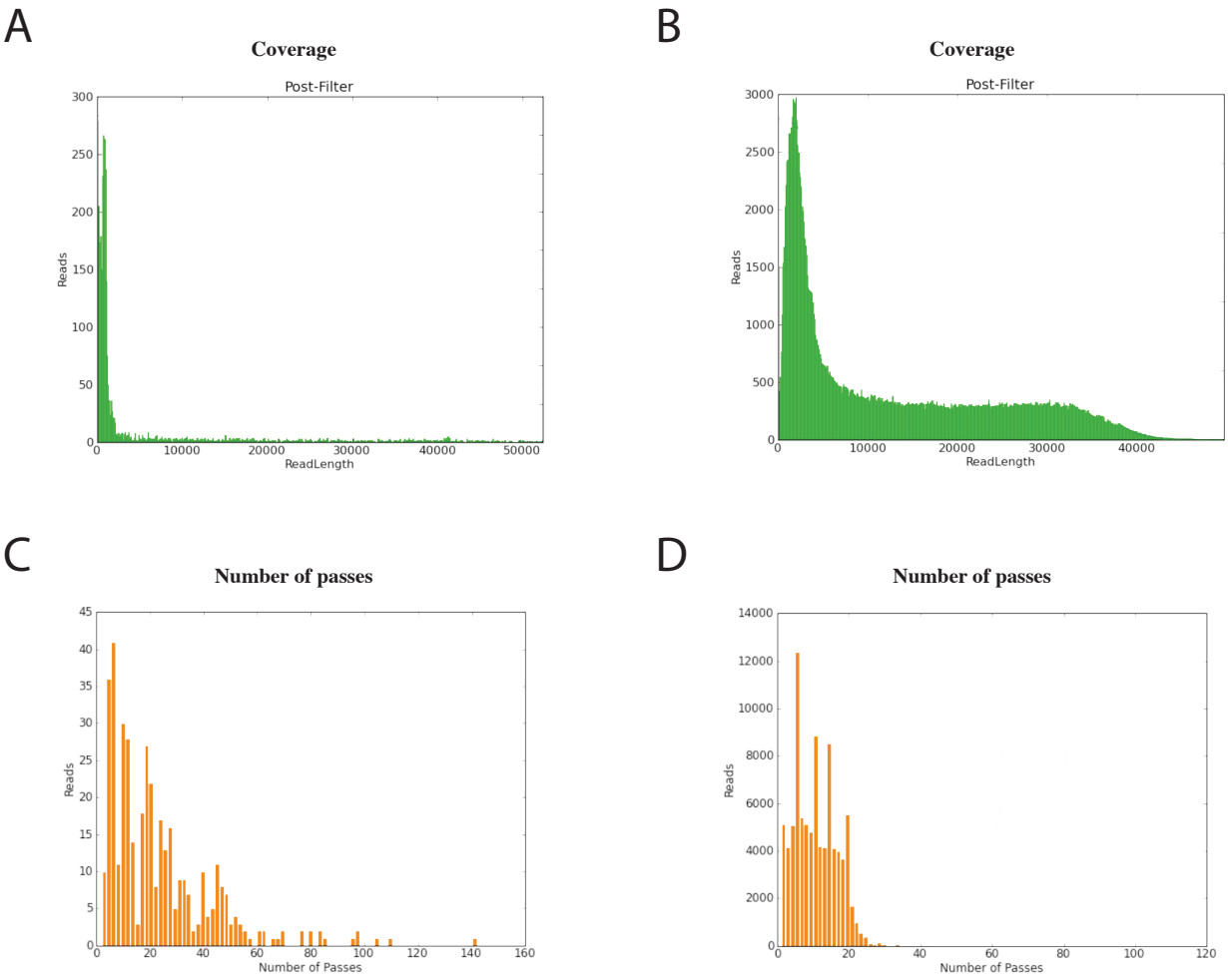


Figure 4: Post sequencing library quality control
A: Coverage - distribution of the number of subreads that make up raw reads for Library 1
B: Coverage - distribution of the number of subreads that make up raw reads for Library 2
C: CCS coverage distribution within the sequenced first library
D: CCS coverage distribution within the sequenced second library

The CCS within the first library from both sequencing rounds are made up of sequences with or without TNR repeats (Figure 5A). BLAST analysis was done to determine the composition of CCS without repeats. Plasmid sequences correspond to 11.3% and 14.6% of CCS without TNRs from diffusion or MagBeads loading, respectively. The rest of the sequences map to random chromosomes on the human genome, as well as, to mouse chromosome X. Sequences mapping to the human genome most likely correspond to the template DNA being carried over during the library preparation or poor primer specificity. And the mouse chromosome X sequences map specifically to the Pem1 intron. This can be explained by the fact that in the GFP(CAG) cell line the CAG repeats are embedded within the Pem1 intron sequence. These sequences were considered artifacts and were excluded from further analysis.

Initially the plasmid was spiked in at 0.5% of total quantity of the initial library. From the data analysis it is apparent that the plasmid sequences are overrepresented, and even more so with MagBeads loading. We considered that one explanation may be that the plasmid fragments do not contain any TNRs and therefore the polymerase can go through these sequences more easily. This would predict that the number of subread per CCS coverage for the plasmid sequences should be higher than for sequences containing the CAG repeats. However, the CCS coverage profile appears to be similar between sequences with and without the repeats (Figure 5B). Therefore, the presence or length of the CAG repeat does not explain plasmid sequences overrepresentation.

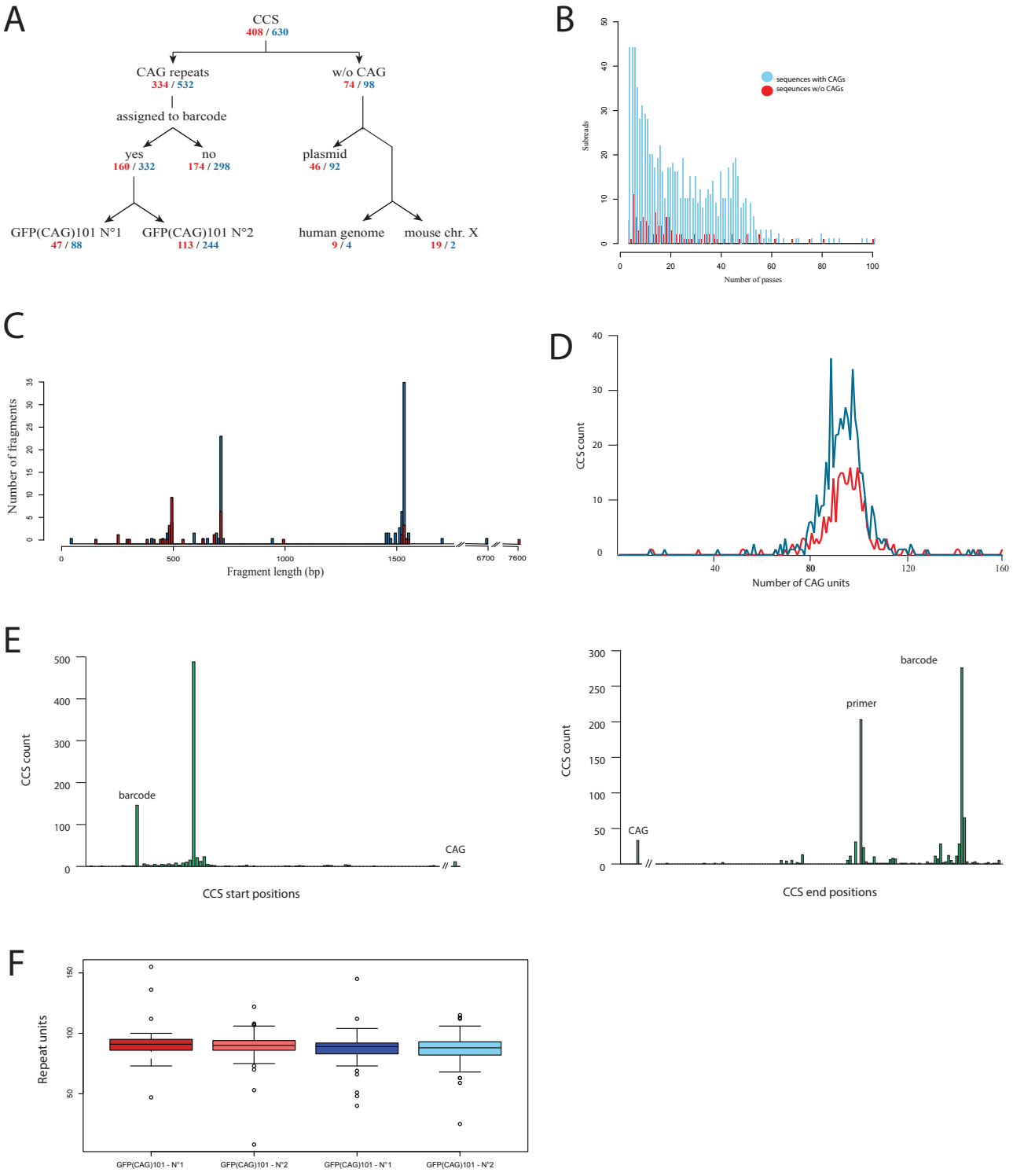


Figure 5: Post sequencing comparison manual data analysis between diffusion and MagBeads loading
Note: red corresponds to diffusion loading and blue to MagBeads loading
A: Full library CCS content
B: Number of passes for the full library, but distribution separated into sequences with and without the CAG repeats
C: Fragment size distribution for the digested plasmid
D: Distribution of CAG/CTG repeats within the full library
E:Distribution of CCS start (5'end) and end (3'end) positions
F: Distribution of number of TNRs within GFP$(CAG)_{101}$ samples treated with DOX or DMSO over 12 months. Wilcoxon test: not significant; p-value=0.5987 and 0.8229 for diffusion and MagBeads loading approaches respectively

Nevertheless, the size distribution of plasmid fragments is as expected for both loading approaches with the main peaks at 1.6 kb, 0.8 kb and 0.5 kb (Figure 5C). However, in comparison to the diffusion loading method with MagBeads there is a shift in the length of sequenced plasmid fragments towards the larger size, which is expected, due to minimum size selection. The size distribution of the digested plasmid fragments appears to form a hill-like profile, reflecting the small count of sequences below 500 bp and above 2 kb in length. The lack of large fragments latter is an indication of loading bias that is not specific to the plasmid fragments.

Samples with initial 97, 101 and 270 repeats were multiplexed during the first library preparation. The CCS sequences with the CAG repeats represent over 80% of the full library. To achieve a measure of instability comparable to SP-PCR, but with better resolution, the optimum number of CCS per sample with RINGS should be between 500 and 1000. Even though, the CCS count is much lower for the first library, data analysis can still be done. Preliminary analysis reveals that the distribution of the CAG units within the library is similar independently of the loading method (Figure 5D). The number of CAG repeats ranges from 0 to 160 units, which suggests that either the GFP$(CAG)_{270}$ DNA fragments were not sequenced or the actual number of the repeats within GFP$(CAG)_{270}$ sample is lower then expected.

To distinguish between different samples that were sequenced they must be demultiplexed. Unfortunately, none of the standard automated methods fastx_barcode_splitter, blat UCSC or bwa from github proved useful, therefore manual demultiplexing was done using a perl script (Annex 9.3). This script allows to scan through the CCS sequences and to look for four different barcodes, regardless of their position, considering one insertion, deletion (indel) or a mismatch. However, only 50-60% of CCS with CAG repeats were assigned to a barcode. The main hypothesis to explain this is that the script is too stringent. Nevertheless, to determine whether there are additional explanations for low demultiplexing efficiency, first 50 bp at 5' or 3' end of CCS with repeats were aligned using SeaView and muscle alignment tool. From the distribution of both start and end positions of CCS, it is apparent that most of the sequences start and end with a primer or a barcode sequence, as expected (Figure 5E). However, there is some variation within primer and barcode positions. This explains the difficulty to assign the sequenced samples to their barcodes, since a portion of different length can be missing from the 16 bp barcode sequence.

CCS with an assigned barcode correspond to two replicates of GFP$(CAG)_{101}$ long term passaging samples. As expected, the samples display a range of instability. Here the number of TNRs varies between 8 and 155 repeat units (Figure 5F). The results are similar between the two replicates, as well as, between the two loading methods. This is reassuring since exactly the same library was sequenced., which confirms that different barcodes and loading methods do not interfere with the template integrity.

From the data analysis it can be concluded that RINGS with multiplexing can be used to effectively measure the rates of TNR instability by sequencing through the full locus containing the repeats and that more efficiently with MagBeads loading. Nevertheless, optimisation steps can be taken to further increase the efficiency of this method.

4.4 RINGS calibration using GFP(CAG) 15, 50, 101 and 270

To improve RINGS efficiency another round of sequencing was done using DNA extracted from cell lines with estimated number of TNRs at the same locus. According the results from library 1, the main steps of RINGS to improve are the template quality, recovery rate, increase fragment length to make library more adapted to MagBead loading, perform sequencing in two SMRT cells per sample and use wider range of TNR number. Samples multiplexed during this library preparation are estimated to contain 15, 50, 101 and 270 CAG/CTG repeats. The sample 101 TNRs was treated with either CRISPR-Cas9 nickase targeting CTG repeats or an empty gRNA vector. In fact, the optimisation steps together increased the quality of raw sequencing output for the second library in comparison to the first, with the total number of CCS even higher than the optimum (Table 3).

However, the number of CCS per sample varies drastically from close to 50 thousand CCS assigned to sample with estimated 15 CAGs to 17 CCS for sample with 270 repeats (Figure 6A). This phenomenon is observed in both libraries with and without the DNA damage repair treatment. By excluding the samples with below 100 TNRs, the number CCS obtained is similar between the first and second library. One can hypothesise that as the number of repeats within a locus increases, secondary structures become more frequent and more likely, therefore stalling DNA polymerase, which would result in low quality and low

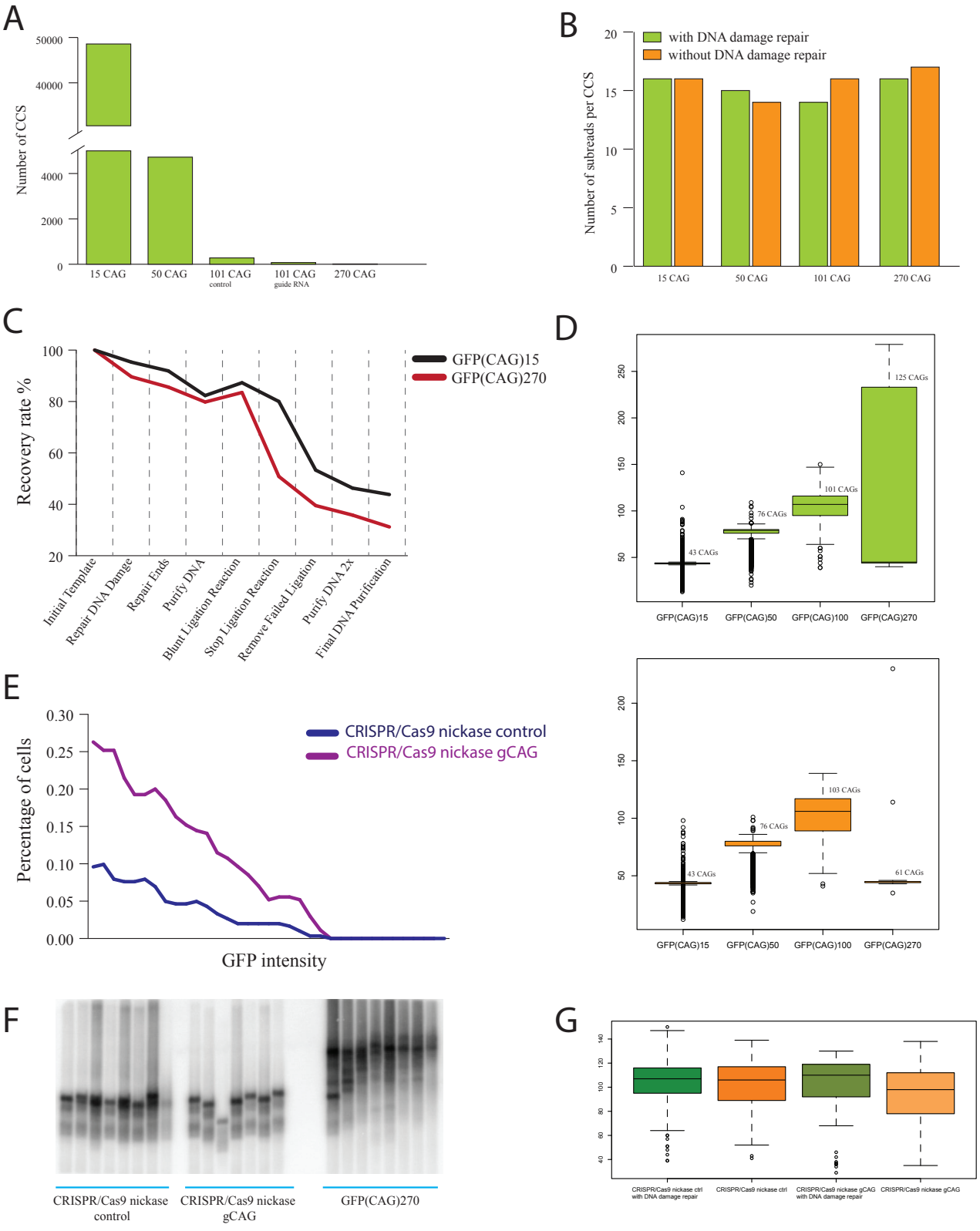


Figure 6: Analysis of distribution of TNRs within Library 2

A: Distribution of number of CCS per sample for library with DNA damage repair treatment. Note: the profile is similar for the library without the DNA damage repair step of library preparation

B: Distribution of number of subreads per CCS on average for each sample with and without the DNA damage repair treatment

C: Recovery rates for each step of library preparation for GFP$(CAG)_{15}$ and GFP$(CAG)_{270}$ libraries without multiplexing

D: Distribution of TNRs per sample after sequencing. Top: library with DNA damage repair, bottom: without DNA damage repair

E: Flow cytometry analysis for the percentage of cells within the 1% of brightest GFP cells treated with CRISPR/Cas9 nickase

F: SP-PCR for GFP$(CAG)_{101}$ CRISPR/Cas9 nickase treated samples and GFP$(CAG)_{270}$

G: Distribution of TNRs within GFP$(CAG)_{101}$ samples treated with CRISPR/Cas9 nickase with or without DNA damage repair

number of passes for such template. This in turn would lead to the observed low number of CCS detected with over 100 repeats. This hypothesis, however, is not supported by the data. In the second library the number of subreads per CCS is on average equally distributed between the samples (Figure 6B). This suggests that the template containing relatively large number of repeats is lost in the steps leading up to sequencing. The loss of material would most likely occur during loading or library preparation.

To determine whether fragments with a large number of repeats are lost during a specific step of library preparation two additional libraries were made with either GFP$(CAG)_{15}$ or GFP$(CAG)_{270}$ and the recovery rate was calculated based on Qubit measured concentration values. During library preparation steps the recovery rates are similar between the GFP$(CAG)_{15}$ and GFP$(CAG)_{270}$ samples and the losses are in large part due to exonuclease treatment that removes failed ligation products (Figure 6C). However there is a 1.5 fold decrease recovery for the library with 270 TNRs after ligase inactivation step. During this step the template is exposed to 65°C for 10 minutes. At this time there is no specific explanation for material loss during the ligase inactivation step, however it is unlikely that it is due to the formation of secondary structures. Additional material loss occurs during AMPMagBead DNA purification steps, but this is to be expected. At the moment there is no specific explanation for the GFP$(CAG)_{270}$ specific drop in recovery rate after ligase inactivation step, however one could imagine that by adding larger proportions of fragments with longer repeat tracts during multiplexing library preparation, the distribution of CCS per sample could potentially be equalised.

Regardless of the number of CCS per sample the distribution of the number of the repeat within each sample can be visualised and appears to be similar between the samples subjected to DNA damage repair step or not (Figure 6D). Surprisingly, the estimation of the number of the repeats on average with Sanger sequencing did not correspond to the results obtained with RINGS, except for GFP$(CAG)_{100}$ (Figure 6D). In fact, for GFP$(CAG)_{15}$ and GFP$(CAG)_{50}$ the number of repeats appears to be over represented with RINGS. The sample with the estimated 270 repeats is made up, according to RINGS, of two distinct populations with approximately 60 and over 200 repeats. To exclude the possibility that this observation is an artefact of template preparation, the samples, as they were prior to library preparation, were sequenced separately with the Sanger method. The additional sequencing of the library confirmed the previous Sanger sequencing estimation. This could be the result of Sanger sequencing bias towards shorter fragments, as well as, the need for optimisation of RINGS bioinformatics data analysis post sequencing.

Taken together these data indicates that DNA damage repair step of library preparation appears to have a slight effect on the sequencing output, especially for the GFP$(CAG)_{270}$ sample. However, the importance of other steps on sequencing output, such as sample loading prior to sequencing and data analysis post sequencing should also be taken into consideration. In addition, it is evidenced in the results that the length of the repeat tract is of great importance, when it comes to RINGS sequencing output.

4.5 RINGS can be used to detect CRISPR/Cas9 Nickase-induced contractions

Within the second library two samples were treated with CRISPR/Cas9 nickase. The goal was to measure the instability by inducing nicks in one of the DNA stands containing the repeat locus. Based on previous research from the lab, such treatment induces preferentially contractions within the GFP(CAG)101 cell line (Cinesi et al. submitted). I repeated this result: flow cytometry analysis showed that the number of cells within the brightest 1% of GFP fluorescence and therefore, potentially shortest number of repeats, increases by 3 fold after CRISPR/Cas9 nickase and gCAG treatments in comparison to the control with gRNA empty vector (Figure 6E). I further confirmed this using SP-PCR (Figure 6F). For most of the SP-PCRs the distribution in the number of the repeats is similar between the CRISPR/Cas9 nickase control and gCAG samples, except for the lane 12, where there is evidence of contractions. However, there are not enough reactions for quantification of such contraction events.

With RINGS, the number of the repeats on average is very similar between the control and CTG repeats targeted samples with and without the DNA damage repair treatment around 100 TNRs (Figure 6G). However, the analysis of variance between these samples indicates that CRISPR/Cas9 nickase gCAG treatment has a significant effect (multifactor ANOVA: p-value=0.03) on the distribution of the number of TNRs within GFP$(CAG)_{101}$ sample. This analysis also confirms the previous observations that the DNA damage repair step of library preparation does not influence the sample integrity (multifactor ANOVA: p-value = 0.11).

It can be concluded that the contractions caused by the CRISPR/Cas9 nickase treatment can be detected with RINGS.

4.6 Deletions at the junction of the repeats are not detected

From previous research it is known that deletions occur in the sequences flanking the repeat tract after cutting the repeat with ZFNs, which reflect contractions within the TNRs (Mittelman et al. 2009). To test whether RINGS can be used to detect such deletions and to determine whether this phenomenon is observed within the two replicates of long term passaging, the full CCS, without the repeat tract from two sequencing rounds were aligned using the muscle alignment tool. Both insertions and deletions (indels) can be seen from the alignment data (Figure 7A). The length of indels is estimated through comparison to the expected template length. The range of indels is between 0 and 377 basepairs. With SMRT technology the expected error rate is approximately 10%, which fits with most of the CCS analysed. The deletions appear to have a wider length distribution and are predominant over insertions.

However, the detected indels are not exclusive to the 5' and 3' extremities of the template (Figure 5E). Surprisingly, at the repeat junction there are more sequences with deletions ahead of the repeats tract than downstream of it (Figure 7B and C). For the library 2 samples (Figure 7C) this phenomenon is less pronounced than for library 1 (Figure 7B). The most likely hypothesis to explain this high number of deletions at the repeat junction, is the sequence content. Immediately upstream of the CAG tract, there are 5 consecutive cytosines. To determine whether the observed deletions at the repeat junction is specific to the repeat tract or a PCR and sequencing artefact, additional template fragments can be considered that comprise of mononucleotide repeats. From the distribution of the number of deletions observed at two additional 5

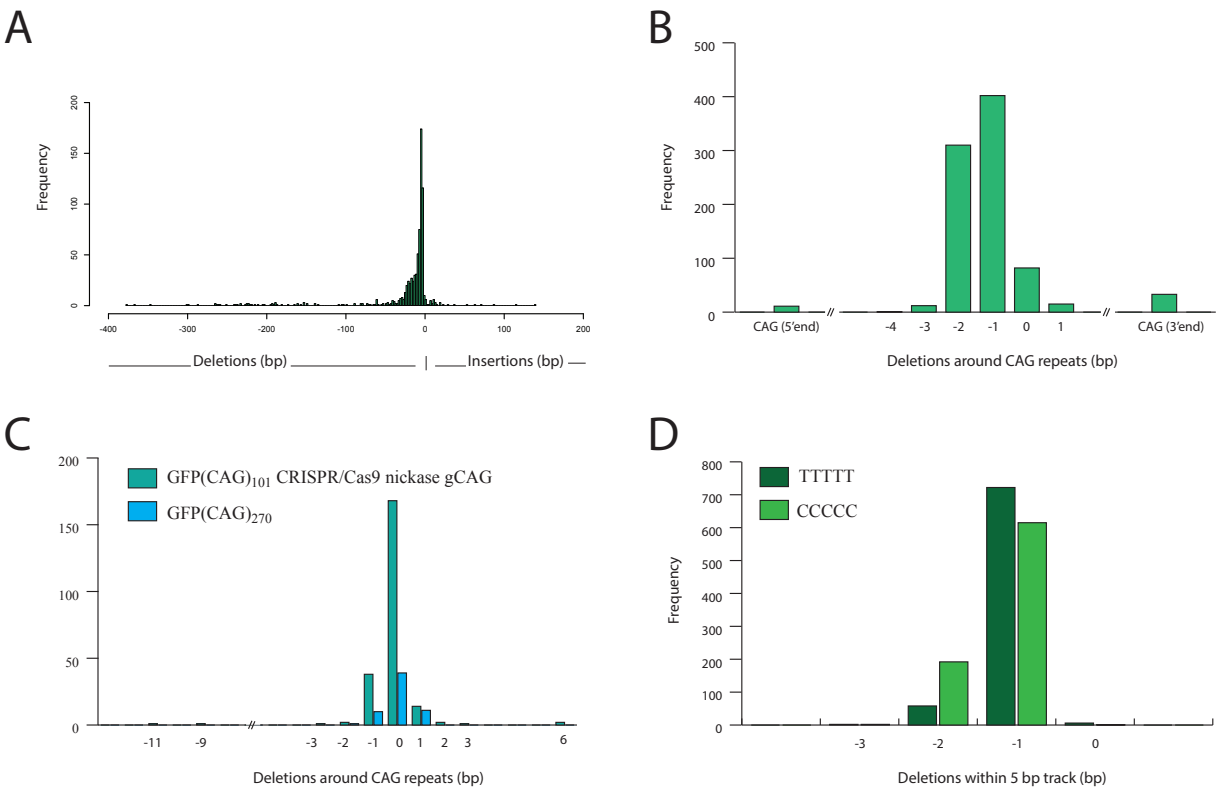


Figure 7: Insertions / deletions within CCS from two sequencing rounds of the same library
A: Size distribution of indels per full length CCS
B: Distribution of deletions around the TNR tract for full Library 1. Zero corresponds to no deletions, -1 to 1 bp deletion aat 5'end and 1 to 1 bp deletion at 3' end of the repeat tract
C: Same as B but for Library 2 samples: $GFP(CAG)_{270}$ and $GFP(CAG)_{101}$ treated with CRISPR/Cas9 nickase gCAG
D: Distribution of deletions within mononucleotide repeat tract

mononucleotide stretches, it can be concluded the the profile of indels is similar between the three mono nucleotide repeat tracts (Figure 7D). This suggests that the observed indels are PCR and sequencing artefacts and are not specific to the repeat tract.

There are larger deletion events that do not occur at mononucleotide repeats. In fact, deletions of over 230 bp can be attributed to the sequences starting or ending with the CAG repeats. Since the template that was used for sequencing is PCR based, all of the CCS are expected to have both primers at the extremities. The sequences that do not are artefacts of library preparation or of the standard PacBio bioinformatics pipeline. In fact, CCS with large deletions are missing one primer. This suggests that large deletions can also be considered as artefacts.

In conclusion, the majority of the deletions observed within the analysed template fit within the expected PacBio sequencing error rate and the rest are the artefacts of template preparation, sequencing or post sequencing bioinformatics data analysis.

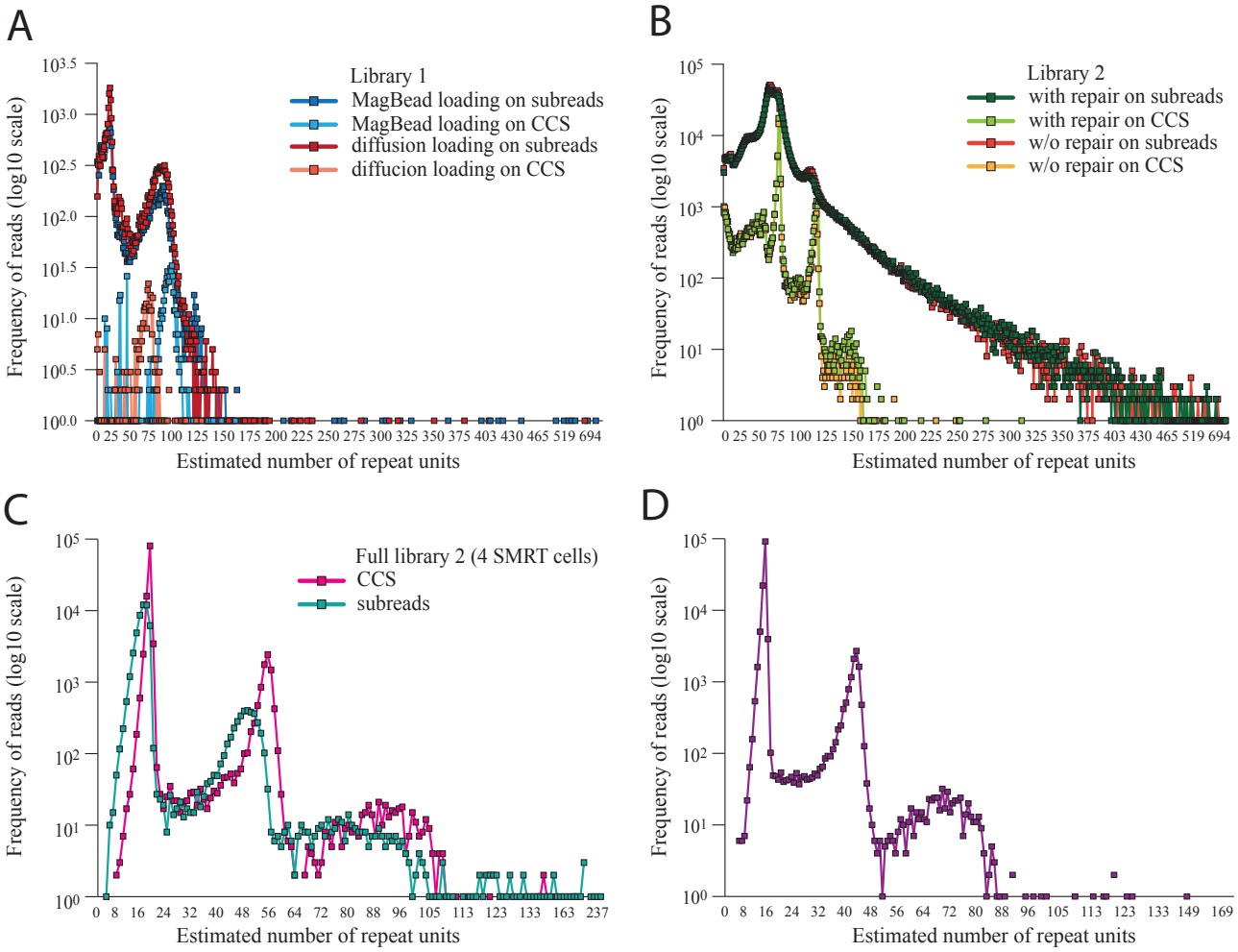


Figure 8: Alternative TNR distribution on subreads and CCS
A: Estimation of TNR number in library 1 on subreads or CCS, by looking for relaxed CAG motif
B: Estimation of TNR number in library 2 on subreads or CCS, by looking for relaxed CAG motif
C: Estimation of TNR number on CCS and subreads in full library 2 through selection of genomic regions flancking the repeats
D: Estimation of TNR number on CCS in full library 2 by looking for deteriorated barcodes

4.7 'We love the smell of the raw data' - Improving the post sequencing data analysis

When performing post sequencing data analysis there is a number of assumptions that are made and are considered true in the consecutive steps. Therefore, it can sometimes be hard to distinguish the real data from artefacts. This also applies to making CCS. In fact, the algorithm, which is used for subreads processing into CCS is intolerant towards polymerase stuttering. Unfortunately, it is a major issue with TNRs. Therefore, the default stringent algorithm for making CCS can reduce the amount of data available for analysis, which in

turn can falsify the estimation of the range of instability within the analysed samples. To verify whether this concept applies to the dataset obtained from libraries 1 and 2, the analysis of the distribution of the number of TNRs was done on subreads. In this analysis, each repeat tract is considered to be made up of 1 to 3 Cs, followed by 1 to 3 As and same for Gs. Any of these combinations are counted as one repeat.

For library 1 there are three major peaks for subreads, as well as, for CCS at approximately 20, 100 and 130 TNRs, instead of the expected 97, 101 and 270 repeats (Figure 8A). However, the major peaks for the second library are in accordance with the expectations and the previous analysis of TNR distribution within CCS (Figure 8B). For both libraries the distribution of the repeat sizes is much larger than with the previous analysis, from 1 to 2060 units observed with subreads. However there is still a decrease within the number of sequences with large number of repeats, which supports the hypothesis of material loss in the steps leading to sequencing (Figure 8A and B). However, upon closer analysis the estimated number of the repeats does not always correspond to the sequence content, especially for those over 500 TNRs. This can be explained by the fact that this analysis allows for a large number of variations within what is considered a repeat unit.

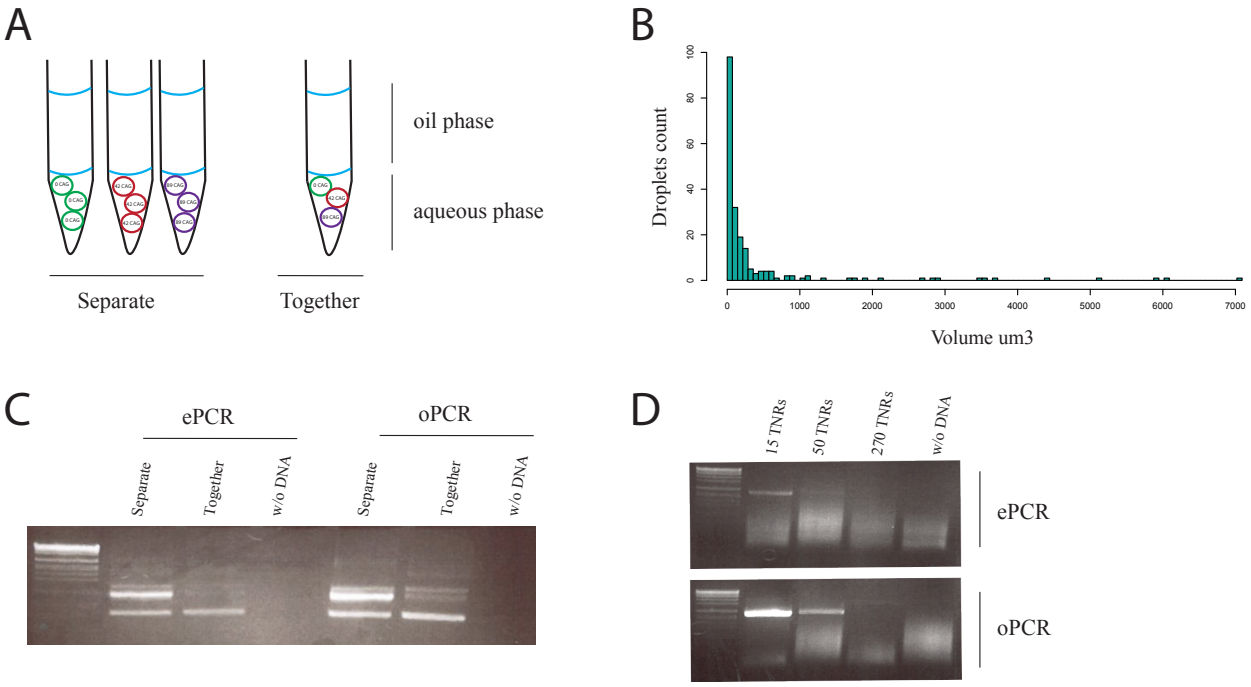


Figure 9: Insertions / deletions within CCS from two sequencing rounds of the same library
A: Illustration of different mixes of plasmids tested with ePCR. Note: the principle also applies to the genomic DNA samples
B: Droplet size estimation
C: ePCR vs oPCR on plasmids with 0, 42 and 89 TNRs
D: ePCR vs oPCR on gDNA with 15, 50 and 270 TNRs

The model for analysis can be improved by applying it to sequences that contain fragments of genomic DNA flanking the repeat tracts. By performing this analysis on the full second library, again three major peaks can be detected at approximately 15, 50 and 85 repeat units (Figure 8C). This estimation of the repeat number corresponds more closely to the initial expectations. The observed 85 repeats on average instead of 101, could reflect the contractions within one of the samples treated with CRISPR/Cas9 nickase targeting CTG repeats, which will have to be confirmed with demultiplexing on this model. Another attempt at improving post sequencing analysis model is to consider deteriorated 20 bp primers by running megablast on CCS of the full second library. The distribution of the TNR number is almost identical to what was observed with the analysis considering flanking regions (Figure 8D). However, for both models there is a very low number of sequences above 200 estimated repeats that can be explained by the fact that the new model for analysis requires perfect match within the genomic DNA regions and primers (Figure 8C and D). Therefore, the output number of sequences from this model decreased substantially in comparison to the previous, more relaxed model. The observed trend is similar between the analysis done on CCS and subreads. However, it appears that there are more sequences that were selected by considering CCS than subreads (Figure 8C and D). This too can be explained by the fact that regions flanking the repeats and primer sequences are expected to be perfect match by the model, which is a more reasonable assumption for the CCS.

In addition, this analysis supports prior results that there is a bias towards shorter fragments with diffusion loading in comparison to MagBead loading, which is reflected in the slightly higher number of subreads with shorter repeat tract (Figure 8A). From the distribution of the number of the repeat units within the second library it can be confirmed that the DNA damage repair step of library preparation does not have an effect on following steps of RINGS approach.

In conclusion, the obtain data suggests that this relaxed model for data analysis, which takes into account polymerase stalling over the repeat tract does not distort the data and taking into consideration the flanking regions is more adapted towards estimation of TNR number in the full library.

4.8 Assessing the use of ePCR as a replacement for SP-PCR

Further optimisation of RINGS can be done on early steps of template preparation. With that in mind, I assessed whether ePCR (Figure 9A) could be used instead of the SP-PCR approach used so far. To determine the optimal amount of DNA needed per ePCR, and to achieve the ratio on average of one molecule per droplet, the number of droplets was estimated under a microscope using a fluorescent dye (Figure 9B, see Materials and methods). I found that on average there are 2.4 x 10^8 droplets per ePCR, with the average volume of 411 um^3 per droplet (Figure 9B).

As a proof of concept, ePCRs and open PCRs (oPCR) were conducted on plasmids with various length of TNRs to assess PCR bias towards smaller alleles. The experiments were done in parallel on the mixture of plasmids with 0, 42 or 89 TNRs, estimated with Sanger sequencing, or on the three samples separately (Figure 9A). The minimal total amount of DNA tested for the plasmid DNA is 0.5 ng per reaction. At that amount of DNA, most of the droplets are empty, essentially achieving the same thing as using a small number of template molecule per reaction. Nevertheless, PCR bias was apparent in the oPCR samples (Figure 9C). The same is true for an experiment done with 50 ng genomic DNA (gDNA) samples with 15, 50 and 270 repeats (Figure 9D). As expected ePCR is less efficient than open PCR, and it appears to be even less efficient on genomic DNA.

In conclusion, ePCR is promising, but more optimisation with regards to aqueous-to-oil phase ratio, the amount of DNA per reaction and aqueous phase mixture is needed to make the reactions more efficient in terms of bias and output quantity.

5. Discussion

The results from RINGS analysis are promising. The use of RINGS allows the sequencing and quantifying of the repeats size and heterogeneity in several samples in parallel. The sequencing output is of better quality than expected especially with the second library, which is probably the reflection of higher recovery rate and the inclusion of samples with shorter repeat length. The library preparation steps do not have an effect on RINGS output. On the other hand, MagBead loading appears to lead to more efficient sequencing output. The requirement of longer fragments means that it reduces loading bias. In addition, RINGS allows for base pair resolution of the sequences containing the repeat tract. This would allow detection or rearrangement within the sequences. However, no such modifications have been detected within the sequences in this study, which is contrary to previously reported data (Cinesi et al. submitted; Mittelman et al. 2009). In the previous work only the rare and large contractions or expansions events were considered, which was not the case in this study, where a population of cells was analysed.

The main conclusion of this study is that the length of trinucleotide repeats is of great importance at each stage of RINGS approach protocol. Therefore, additional optimisation steps are necessary. For example, the number of secondary structures formed within repeat tract and flanking regions and their stability can be reduced by adding dimethyl sulfoxide before loading. This might facilitate loading and sequencing of fragments with large repeats tracts, which appear to be underrepresented. Loading efficiency of such fragments may also be increased through sequential loading, starting with large repeats or by increasing the quantity of fragments with large repeats proportionally to short repeats during library preparation. Additionally, sequences with large repeats are even further underrepresented in the CCS. To reduce that effect, optimisation can be done on demultiplexing method. The use of a relaxed model to look for repeats selectively within selected regions proved to be most efficient. However, setting up an analysis pipeline with subsequent demultiplexing steps is necessary to continue with potentially wide usage of RINGS.

RINGS can be tested further on hemizygous mouse samples. Knowing that repeat instability varies with different tissue types (Dion 2014), performing that analysis would allow to optimise RINGS and to gain further insight into tissue specific instability. Additional applications for RINGS could be extended to heterozygous HD patient samples with 50 and 250 repeats, as well as DM1 patient samples with 1000s of TNRs. The patient samples will present an additional challenge due to the presence of the wild type allele that is much easier to PCR amplify than the expanded one. However, if RINGS proves successful, this approach will be a time and cost efficient manner to assess repeat length and instability with applicability as a diagnostic tool as well as in understanding the pathogenesis of expanded TNR diseases.

6. Materials and methods

6.1 Template preparation

6.1.1 Cell culture

GFP(CAG) cell lines were maintained at 37°C and 5% CO2 in Dulbecco's modified Eagle's medium (DMEM) glutamate, supplemented with 10% fetal bovine serum, 15 µg/µl blasticidine and 150 µg/ul hydromycine. Cells at long term treatment were split twice a week and supplemented with 2µg/µl DOX or DMSO. For flow cytometry analysis the cells were maintained in 10% dialysed calf serum with pen-strep.

6.1.2 Transfection

cDNA transfections were done on 600'000 cells in 12-well plates using 0.5 µg of Cas9 Nickase and 0.5 µg of guide RNA in Lipofectamine 2000 (Life Technoogies) per well. The medium was changed 6 hours after transfection and 2µg/µl of DOX, diluted in DMSO was added. 48 hours later medium was changed and again DOX was added. After another 48 hours flow cytometry and DNA extraction was performed.

6.1.3 Flow cytometry

Samples for flow cytometry were washed with phosphate buffered saline (PBS) and then ressuspended in 700 µl of PBS with 1mM EDTA. With LSRII 150'000 events were recorded and analyzed using Flowing software.

6.1.4 DNA extraction

DNA extraction was done using PeqGOLD MicroSpin Tissue DNA kit (peqlab). Quality control post extraction was done with Nanodrop.

6.1.5 Generation of SP-PCR fragments

For samples from Library 1 (Table S1) the PCR was carried out with MangoTaq (Bioline) under the following cycling conditions: 95°C for 2 min, 5 cycles of 95°C for 20 sec, 52°C for 20 sec, 72°C for 1 min, 40 cycles of 95°C for 20 sec, 55°C for 20 sec, 72°C for 1 min, and 72°C for 2 min.

For samples from Library 2 (Table S1) the PCR was carried out with PrimeStar GXL (Clonetech) under the following cycling conditions: 98°C for 2 min, 45 cycles of 98°C for 10 sec, 60°C for 15 sec, 60°C for 2 min 30 sec, and 68°C for 2 min.

Sample specific primer combinations with barcodes were used (Table S1). For each sample, in total 96 reactions were set up in total, with 7 reactions as negative controls and 1 reaction at 10 ng of DNA as positive control. The remaining 88 reactions were run on a few genomes of DNA. After the PCR, the samples were pooled together, alcohol precipitated and extracted from 0.8% agarose gel using the Nucleospin PCR clean up kit (Machery Nagel). The quality control was done with Nanodrop, Qbit and Fragment Analyser.

Table S1: Combinations of primers used for template amplification. Sample-specific barcodes are in bold.

	Sample	Forward Primer	Reverse Primer	Expected fragment size (bp)
Library 1	GFP(CAG)$_{97}$	oVIN895 **AGACTCTACAGAGATA** AAGAGCTTCCCTTTACAC AACG	oVIN437 TACCAGGACAGCAGTGG TCA	834
	GFP(CAG)$_{270}$	oVIN887 **TCTCTCACAGTCGAGC** AAGAGCTTCCCTTTACAC AACG	oVIN437 TACCAGGACAGCAGTGG TCA	1353
	GFP(CAG)$_{101}$ replicate N°1	oVIN883 **ATCTGTGCGAGACTAC** AAGAGCTTCCCTTTACAC AACG	oVIN437 TACCAGGACAGCAGTGG TCA	846
	GFP(CAG)$_{101}$ replicate N°2	oVIN459 AAGAGCTTCCCTTTACAC AACG	oVIN888 **GCTCGACTGTGAGAGA** TACCAGGACA-GCAGTGGTCA	846
Library 2	GFP(CAG)$_{101}$ CRISPR-Cas9 Nickase control	oVIN879 **CGCGCGTGTGTGCGTG** AAGAGCTTCCCTTTACAC AACG	oVIN587 GCATGGACGAGCTGTAC AAGTA	1999
	GFP(CAG)$_{101}$ CRISPR-Cas9 Nickase gCAG	oVIN889 **GTCATCACACATCTCT**A AGAGCTTCCCTTTACACA ACG	oVIN587 GCATGGACGAGCTGTAC AAGTA	1999
	GFP(CAG)$_{15}$	oVIN881 **CACACGCGCGTGCTCG** AAGAGCTTCCCTTTACAC AACG	oVIN587 GCATGGACGAGCTGTAC AAGTA	1741
	GFP(CAG)$_{50}$	oVIN887 **TCTCTCACAGTCGAGC** AAGAGCTTCCCTTTACAC AACG	oVIN587 GCATGGACGAGCTGTAC AAGTA	1846
	GFP(CAG)$_{270}$	oVIN883 **ATCTGTGCGAGACTAC** AAGAGCTTCCCTTTACAC AACG	oVIN587 GCATGGACGAGCTGTAC AAGTA	2506

6.1.6 Generation of plasmid fragments

To assess and correct for potential loading bias, pLenti CMV Neo Dest was digested with a variety of restriction enzymes (Table S2). For library 2, the same plasmid was digested with different restriction enzymes separately, the bands of interest were extracted from 0.8% agarose gel using the Nucleospin PCR clean up kit (Machery Nagel) and pooled together. Quality control was done with Nanodrop, Qbit and Fragment Analyser.

Table S2: Enzymes and conditions used for plasmid fragments generation

	Enzyme	Buffer	Incubation (temperature and time)	Heat Inactivation (temperature and time)	Fragments of interest (kb)
Library 1	PstI	H	37°C, 15 min	65°C, 15 min	6.6/ 1.6 /0.768/ 0.540
Library 2	BamHI	E	37°C, 1 h	65°C, 15 min	2.7/ 6.3
	BspEI	Tango	55°C, 1 h	80°C, 20 min	1.88 / 7.4
	HindIII	E	37°C, 1 h	65°C, 15 min	4.5/ 3.3
	NaeI	CutSmart	37°C, 1 h		1.2/ 1.3/ 7.3
	NotI	D	37°C, 1 h	65°C, 15 min	1.6/ 1.74/ 6.45

6.2 Library preparation

Prior to library preparation the samples for sequencing (Table S1) were pooled together in equal quantity amounts (500 ng for library 1 and 600 ng for library 2) and spiked with 0.5% of digested plasmid. Library preparation was done using Template preparation kit provided by PacBio. For library 1, 500 bp protocol was used. For library 2, library preparation was done with and without the Repair DNA Damage step. 2kb protocol was used for the library with the Repair DNA damage step, according to the instruction of the manufacturer. And for the library without that step, the library preparation was started with Repair Ends step from 250 bp protocol and then continued with first Purify DNA step from 2kb protocol. Quality control post library preparation was done with Qbit and Fragment Analyzer.

6.3 Data Analysis

Data analysis was done using manual commands in perl (see Annex 9.4)

6.3.1 Demultiplexing

Demultiplexing was done manually using perl script written by Emmanuel Beaudoing, GTF (see Annex 9.3).

6.3.2 Alignment

Sequence alignment was done using SeaView software and muscle alignment tool with default settings.

6.3.3 Graphs and statistical analysis

Graphes and statistical analysis were done using R software.

6.4 SP-PCR

SP-PCR was done as described by Dion and colleagues (Dion et al. 2008) using sample specific primers (Table S1). The probe against repeat region was derived from a PCR product from a plasmid containing 40 CAG/CTG repeats.

6.5 ePCR

6.5.1 Preparation of ePCR mix

The preparation was done separately for 400 µl oil and 100 µl aqueous phase and then pooled together and vortex at 4°C for 5 min at maximum speed. The oil phase contained 4.5% of Span80, 0.40% Tween80, Triton X-100 0.05% and mineral oil. The aqueous phase was prepared in identical fashion as for RINGS template preparation with Mango Taq.

6.5.2 Estimation of droplet size

A fluorescent dye Atto488 was added to the aqueous phase during the preparation of ePCR mixture. After the oil and aqueous phase were mixed together, the droplets were visualised under a microscope. The diameter of droplets was estimated using ImageJ.

6.5.3 Generation of ePCR fragments

The ePCR mix was split into 100 µl reactions and PCR was carried out with MangoTaq (Bioline) for, both plasmid and genomic DNA, under the following cycling conditions: 95°C for 2 min, 5 cycles of 95°C for 10 sec, 52°C for 20 sec, 72°C for 2 min 30 sec, 40 cycles of 95°C for 10 sec, 55°C for 20 sec, 72°C for 2 min 30 sec, and 72°C for 2 min.

6.5.4 Breaking the emulsion

To break the emulsion after PCR the samples were pooled together and 1 ml of isobutanol was added, after which the samples were vortexed for approximately 5 seconds and centrifuged for 2 minutes at maximum speed. The organic phase was then discarded and the samples were subjected to PCR clean up kit from Machery-Nagel.

7. Acknowledgements

I am thankful to Vincent Dion, Keith Karshman, Emmanuel Beaudoing, Mélanie Dupasquier, Ioannis Xenarios and Lorene Aeschbach for the helpful discussions and their creative input into this project. I am also thankful to my mother and grandfather for never-ending moral support. This work was financially supported by University of Lausanne (UNIL) and SNF Professorship awarded to Vincent Dion.

8. References

BLENCOWE B.J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci* 25: 106-110.

BUDWORTH H., HARRIS F.R., WILLIAMS P., LEE D.Y., HOLT A., PAHNKE J., SZCZESNY B., ACEVEDO-TORRES K., AYALA-PEÑA S. & MCMURRAY C.T. 2015. Suppression of Somatic Expansion Delays the Onset of Pathophysiology in a Mouse Model of Huntington's Disease. *PLoS Genet* 11: e1005267.

CINESI C., AESCHBACH L., B Y. & V D. Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. *NSMB* submitted:

CLEARY J.D. & PEARSON C.E. 2005. Replication fork dynamics and dynamic mutations: the fork-shift model of repeat instability. *Trends Genet* 21: 272?280.

CLEARY M.A., VAN RAAMSDONK C.D., LEVORSE J., ZHENG B., BRADLEY A. & TILGHMAN S.M. 2001. Disruption of an imprinted gene cluster by a targeted chromosomal translocation in mice. *Nat Genet* 29: 78-82.

CONG L., RAN F.A., COX D., LIN S., BARRETTO R., HABIB N., HSU P.D., WU X., JIANG W., MARRAFFINI L.A. & ZHANG F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339: 819-823.

DION V. 2014. Tissue specificity in DNA repair: lessons from trinucleotide repeat instability. *Trends Genet* 30: 220?229.

DION V., LIN Y., HUBERT L., WATERLAND R.A. & WILSON J.H. 2008. Dnmt1 deficiency promotes CAG repeat expansion in the mouse germline. *Hum Mol Genet* 17: 1306-1317.

DOI K., MONJO T., HOANG P.H., YOSHIMURA J., YURINO H., MITSUI J., ISHIURA H., TAKAHASHI Y., ICHIKAWA Y., GOTO J., TSUJI S. & MORISHITA S. 2014. Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* 30: 815-822.

DRESSMAN D., YAN H., TRAVERSO G., KINZLER K.W. & VOGELSTEIN B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A* 100: 8817-8822.

EID J., FEHR A., GRAY J., LUONG K., LYLE J., OTTO G., PELUSO P., RANK D., BAYBAYAN P., BETTMAN B., BIBILLO A., BJORNSON K., CHAUDHURI B., CHRISTIANS F., CICERO R., CLARK S., DALAL R., DEWINTER A., DIXON J., FOQUET M., GAERTNER A., HARDENBOL P., HEINER C., HESTER K., HOLDEN D., KEARNS G., KONG X., KUSE R., LACROIX Y., LIN S., LUNDQUIST P., MA C., MARKS P., MAXHAM M., MURPHY D., PARK I., PHAM T., PHILLIPS M., ROY J., SEBRA R., SHEN G., SORENSON J., TOMANEY A., TRAVERS K., TRULSON M., VIECELI J., WEGENER J., WU D., YANG A., ZACCARIN D., ZHAO P., ZHONG F., KORLACH J. & TURNER S. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133?138.

FARRELL B.T. & LAHUE R.S. 2006. CAG*CTG repeat instability in cultured human astrocytes. *Nucleic Acids Res* 34: 4495-4505.

GORBUNOVA V., SELUANOV A., DION V., SANDOR Z., MESERVY J.L. & WILSON J.H. 2003. Selectable system for monitoring the instability of CTG/CAG triplet repeats in mammalian cells. *Mol Cell Biol* 23: 4485-4493.

HSU P.D., SCOTT D.A., WEINSTEIN J.A., RAN F.A., KONERMANN S., AGARWALA V., LI Y., FINE E.J., WU X., SHALEM O., CRADICK T.J., MARRAFFINI L.A., BAO G. & ZHANG F. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 31: 827-832.

KANG S., OHSHIMA K., SHIMIZU M., AMIRHAERI S. & WELLS R.D. 1995. Pausing of DNA synthesis in vitro at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. *J Biol Chem* 270: 27014-27021.

KASS D. & BATZER M.A. 2001. Genome Organization/Human. *ENCYCLOPEDIA OF LIFE SCIENCES*

KENNEDY L., EVANS E., CHEN C.M., CRAVEN L., DETLOFF P.J., ENNIS M. & SHELBOURNE P.F. 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Hum Mol Genet* 12: 3359-3367.

KOVTUN I.V., LIU Y., BJORAS M., KLUNGLAND A., WILSON S.H. & MCMURRAY C.T. 2007. OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature* 447: 447-452.

LIN Y., DION V. & WILSON J.H. 2006. Transcription promotes contraction of CAG repeat tracts in human cells. *Nat Struct Mol Biol* 13: 179-180.

LOOMIS E.W., EID J.S., PELUSO P., YIN J., HICKEY L., RANK D., MCCALMON S., HAGERMAN R.J., TASSONE F. & HAGERMAN P.J. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res* 23: 121?128.

LÓPEZ CASTEL A., CLEARY J.D. & PEARSON C.E. 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* 11: 165-170.

MANGIARINI L., SATHASIVAM K., SELLER M., COZENS B., HARPER A., HETHERINGTON C., LAWTON M., TROTTIER Y., LEHRACH H., DAVIES S.W. & BATES G.P. 1996. Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell* 87: 493-506.

MANLEY K., SHIRLEY T.L., FLAHERTY L. & MESSER A. 1999. Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. *Nat Genet* 23: 471-473.

MCMURRAY C.T. 2010. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* 11: 786-799.

MITTELMAN D., MOYE C., MORTON J., SYKOUDIS K., LIN Y., CARROLL D. & WILSON J.H. 2009. Zinc-finger directed double-strand breaks within CAG repeat tracts promote repeat instability in human cells. *Proc Natl Acad Sci U S A* 106: 9607-9612.

MONCKTON D.G. & CASKEY C.T. 1995. Unstable triplet repeat diseases. *Circulation* 91: 513-520.

ORR H.T. & ZOGHBI H.Y. 2007. Trinucleotide repeat disorders. *Annu Rev Neurosci* 30: 575-621.

BIOSCIENCES P. **Using MagBeads to load SMRTbells in the PacBio RS II**. *http://www.pacb.com/videos/using-magbeads-to-load-smrtbells-in-the-pacbio-rs-ii/*

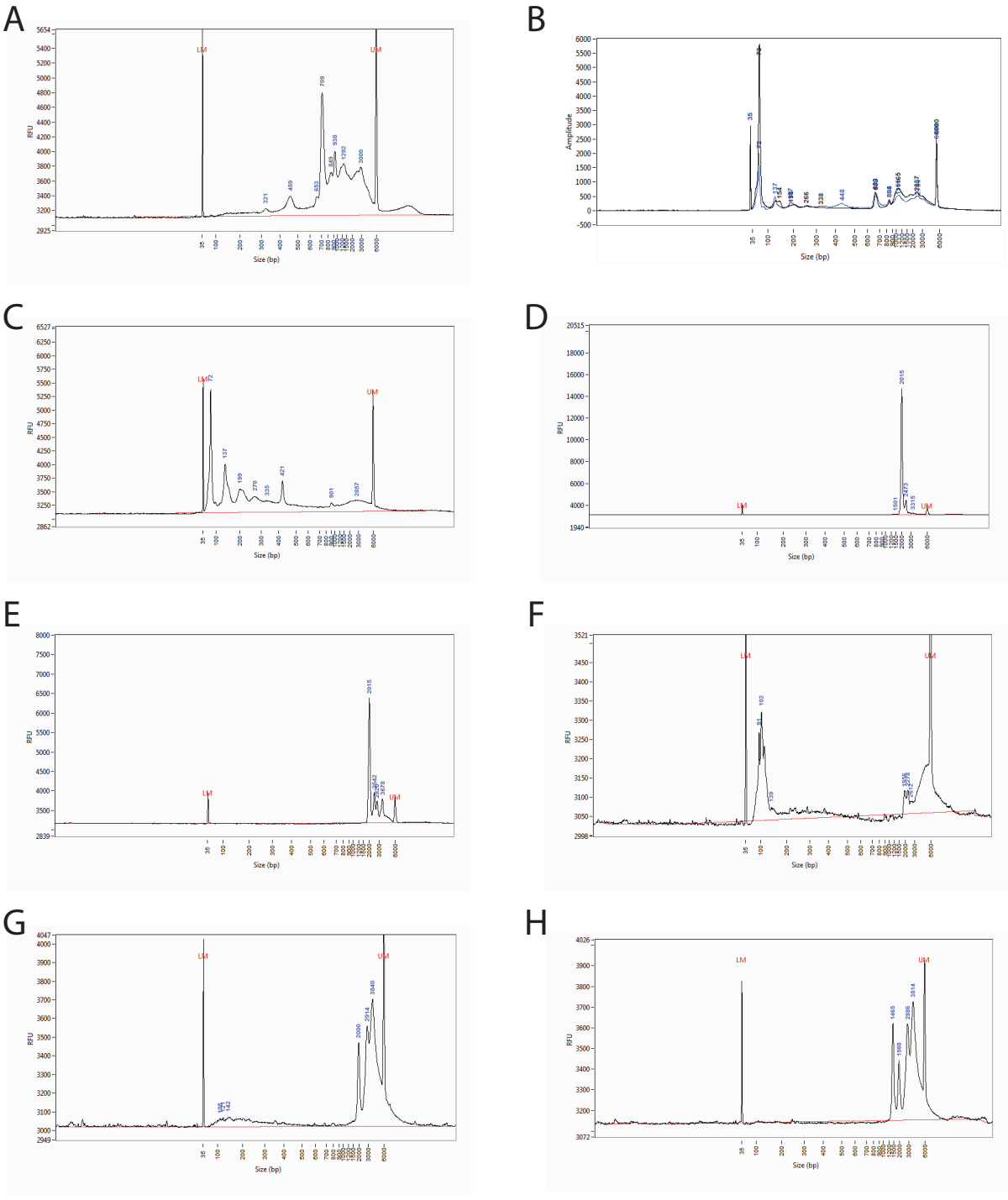
PELLETIER R., FARRELL B.T., MIRET J.J. & LAHUE R.S. 2005. Mechanistic features of CAG*CTG repeat contractions in cultured cells revealed by a novel genetic assay. *Nucleic Acids Res* 33: 5667-5676.

POTAMAN V.N., BISSLER J.J., HASHEM V.I., OUSSATCHEVA E.A., LU L., SHLYAKHTENKO L.S., LYUBCHENKO Y.L., MATSUURA T., ASHIZAWA T., LEFFAK M., BENHAM C.J. & SINDEN R.R. 2003. Unpaired structures in SCA10 (ATTCT)n.(AGAAT)n repeats. *J Mol Biol* 326: 1095-1111.

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: e113952.
SCHERZINGER E., LURZ R., TURMAINE M., MANGIARINI L., HOLLENBACH B., HASENBANK R., BATES G.P., DAVIES S.W., LEHRACH H. & WANKER E.E. 1997. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell* 90: 549-558.
SCHÜTZE T., RUBELT F., REPKOW J., GREINER N., ERDMANN V.A., LEHRACH H., KONTHUR Z. & GLÖKLER J. 2011. A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Anal Biochem* 410: 155-157.
SHAO K., DING W., WANG F., LI H., MA D. & WANG H. 2011. Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PLoS One* 6: e24910.
SURESHKUMAR S., TODESCO M., SCHNEEBERGER K., HARILAL R., BALASUBRAMANIAN S. & WEIGEL D. 2009. A genetic defect caused by a triplet repeat expansion in Arabidopsis thaliana. *Science* 323: 1060-1063.
GROUP T.H.D.C.R. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* 72: 971-983.
TÓTH G., GÁSPÁRI Z. & JURKA J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10: 967-981.
VINCES M.D., LEGENDRE M., CALDARA M., HAGIHARA M. & VERSTREPEN K.J. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324: 1213-1216.
WANG P., JING F., LI G., WU Z., CHENG Z., ZHANG J., ZHANG H., JIA C., JIN Q., MAO H. & ZHAO J. 2015. Absolute quantification of lung cancer related microRNA by droplet digital PCR. *Biosens Bioelectron* 74: 836-842.

9. Annex

9.1 Quality controls



Annex figure 1: Fragment Analyzer results before library preparation
A: Library 1 - GFP$(CAG)_{97}$
B: Library 1 - GFP$(CAG)_{101}$ overlay between two replicates
C: Library 1 - GFP$(CAG)_{270}$
D: Library 2 - GFP$(CAG)_{15}$
E: Library 2 - GFP$(CAG)_{50}$
F: Library 2 - GFP$(CAG)_{270}$
G: Library 2 - CRISPR/Cas9 nickase control
H: Library 2 - CRISPR/Cas9 nickase gCAG

9.2 Protocols

9.2.1 RINGS

SP-PCR combined with Southern Blot is considered the gold standard to measure the rates of trinucleotide instability. However, this method is time consuming. Therefore the goal of RINGS is to have a faster, more precise and more quantitative approach to quantify the instability. This approach takes advantage of Single Molecule sequencing in Real Time (SMRT) method developed by PacBio in combination with SP-PCR. The DNA polymerase goes in circles around the template by recognising the SMRTbell adaptors ligated to the template, thereby creating several fragments in both forward and reverse directions of the template. Although the polymerase introduces a number of mutations as it replicates, one can take advantage of the fact that each single molecule is sequenced several times and thus a consensus sequence averages out the polymerase stuttering and alleviates the errors that occur during sequencing.

References:

LOOMIS E.W., EID J.S., PELUSO P., YIN J., HICKEY L., RANK D., MCCALMON S., HAGERMAN R.J., TASSONE F. & HAGERMAN P.J. 2013. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res* 23: 121-128.

Step 1: DNA extraction

Step 2: DNA digestion with EcoRV (Incubation O/N at 37°C)

	Final concentration	Stock	Volume for 1 reaction from stock
DNA	500 ng		x
Buffer	1x	10x	5 ul
Spermidine	1 mM	10 mM	5 ul
BSA	0.1 mg/ul	10 mg/ml	0.5 ul
EcoRV	50 U	10 U/ul	5 ul
H2O			to 50 ul

Step 3: Dilutions test

Set up PCRs and run the product on 1% agarose gel:

- 1/5 dilutions: 10ng / 2ng / 400 pg/ 80 pg/ 16 pg/ 3.2 pg
- negative control: w/o DNA
- prepare PCR mix for 8 reactions in total, 20 ul each

	Mango Taq Mix	PrimeStar Takara GXL Mix
Buffer	32 ul	32 ul
Mg2+	6.4 ul	-
dNTP	8 ul	12.8 ul
Primer 1	12 ul	12 ul
Primer 2	12 ul	12 ul
DMSO	4 ul	4 ul
Polymerase	4 ul	3.2 ul
DNA	1 ul / rxn	1 ul / rxn
H2O	to 160 ul	to 160 ul

Note :

1. primers diluted 10x from stock

2. for multiplexing one primer should contain a barcode

3. **forward primers** with 16 bp barcode attached to oVIN 459 are: oVIN 879, 881, 883, 885, 887, 889, 893, 895. These primers work well with oVIN 587 - expected size w/o repeats is 1.680 kb. (Don't use in combination with oVIN 467— unspecific bands !)

4. **reverse primers** with 16 bp barcode attached to oVIN 437 are: oVIN 880, 882, 884, 886, 888, 890, 894, 896. Do not use in combination with oVIN 98 - unspecific bands !

5. potentially can have a template with barcodes on 3' and 5' end, see quartzy.com

6. better to have final PCR product of > 1kb, so can use MagBeads loading

	PCR program for Mango Taq Mix:		
	Temperature °C	**Time**	
1	95	2 min	
2	95	10 sec	
3	52	20 sec	
4	72	2 min 30 sec	go to 2 - 5 cycles
5	95	10 sec	
6	55	20 sec	
7	72	2 min 30 sec	go to 5 - 40 cycles
8	72	2 min	
9	15	pause	

	PCR program for PrimeStar Takara GXL Mix:		
	Temperature °C	**Time**	
1	98	2 min	
2	98	10 sec	
3	60	15 sec	
4	68	2 min 30 sec	go to 2 - 45 cycles
5	68	2 min	
6	15	pause	

Step 4: Choose the amount of DNA per reaction for SP-PCR

- Select of the gel from step 3, the lane with the lowest amount of DNA, but where the amplification band is still visible under the UV light.
- Divide that amount of DNA by 2. This is the amount of DNA that is to be used for SP-PCR per reaction

Step 5: SP-PCR

- run reactions on a 96 well plate
- below PCR mix is for 53 rxns, so for a full 96-well plate need to prepare 2 of PCR mixes below

	Mango Taq Mix : 53 rxns	**PrimeStar Takara GXL Mix : 53 rxns**
Buffer	212 ul	212 ul
Mg2+	42.4 ul	-
dNTP	53 ul	84.8 ul
Primer 1	79.5 ul	79.5 ul
Primer 2	79.5 ul	79.5 ul
DMSO	26.5 ul	26.5 ul
Polymerase	26.5 ul	21.2 ul
H2O	487.6 ul	503.5 ul

1. prepare PCR mix for 53 reactions 2x (106 reactions in total) without DNA
2. on a 96 well-plate load 7 reactions with 19 ul of PCR mix and 1 ul H2O per reaction = negative control
3. load 1 reaction with 19 ul of PCR mix and add 1 ul of 10ng/ul concentrated DNA = positive control
4. add diluted DNA to the rest of the mix at the amount selected in the previous step, 1ul/rxn = 88 rxns
5. run PCR program from step 3

Step 6: Alcohol precipitation

	✓	Negative control (7 reactions)	Reactions of interest (88 reactions)	Positive control (1 rxn)
1		pool together	pool together	transfer to a single PCR tube and store at 4°C
2		add 500 ul of isopropanol	add 1 volume of isopropanol	
3		add 50 ul of NaAc pH 5.5 (stock at 3M)	add 1/10 volume of NaAc pH 5.5 (stock at 3M)	
4		spin at max for 3 min	spin at max for 3 min	
5		remove supernatant	remove supernatant	
6		add 500 ul EtOH 70%	add 1 volume EtOH 70%	
7		spin at max for 3 min	spin at max for 3 min	
8		remove EtOH	remove EtOH	
9		spin 30 sec and remove the remaining EtOH	spin 30 sec and remove the remaining EtOH	
10		let dry for 5 min and add 30 ul H2O	let dry for 5 min and add 30 ul H2O	
11		let the remaining EtOH to evaporate for approx. 30 min	let the remaining EtOH to evaporate for approx. 30 min	
12		Run total sample on 0.8% SeaKem agarose gel with SYBR safe DNA stain	Run total sample on 0.8% SeaKem agarose gel with SYBR safe DNA stain	Run total sample on 0.8% SeaKem agarose gel with SYBR safe DNA stain

Step 7: Extraction from agarose gel
1. check if samples look as expected
2. for the pooled 88 rxns cut out the full lane from the gel without the primer dimers
3. for DNA extraction from the gel use PCR clean up kit
4. resuspend sample in 30 ul elution buffer

Step 8: Quality control
1. Nanodrop: concentration and ratios
2. Qbit
3. Fragment Analyzer
4. Run 3 ul of sample on an agarose gel

Note: expect difference in Nanodrop and Qbit concentrations. Aim for 2 times difference max.

Step 9: Plasmid digestion - monitor for loading bias
1. Choose a plasmid without CAG repeats (for example pLenti CMV Neo Dest bVIN-275)
2. Digest the vector with restriction enzymes to generate fragments that are in the range of RINGS template size

Step 10: Library preparation
Use one of the kits provided by PacBio

9.2.2 Emulsion PCR

The goal of emulsion PCR is similar to SP-PCR, which is to conduct separate reactions amplifying single DNA molecules. To achieve this goal with SP-PCR DNA is diluted to just a few amplifiable genomes per tube and at least 98 different reactions are needed. With ePCR it can be done in a single tube (500 ul reaction). ePCR mix is made up of two phases: oil and aqueous. Mixing the two phases together, lead to formation of droplets (approx. 2.4 x 10^8 droplets per reaction). The aim is to use low amounts of DNA in order to have one genome per droplet at maximum.

References:

Dressman, D., et al., Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. PNAS, 2003. 100(15): p. 8817-8822
Schütze, T., et al., A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. Analytical Biochemistry 2011 Mar 1;410(1):155-7.

Step 1: Prepare oil phase

400 ul for 1 reaction (= 1 DNA template)

Solutions	Final concentrations	Stock	Volume for 1 reaction from stock
Span80	4.5%	10%	180 ul
Tween80	0.40%	10%	16 ul
Triton X-100	0.05%	10%	2 ul
Mineral oil			202 ul

Note: pipet slooooowly …….

Step 2: Prepare aqueous phase

100 ul for 1 reaction (= 1 DNA template)

Solutions	**Final concentrations**	**Stock**	**Volume for 1 reaction from stock**
Mango Buffer			20 ul
MgCl2	2 mM	50 mM	4 ul
dNTP	0.5 mM	10 mM	5 ul
Forward primer	0.5 uM	10 uM	5 ul
Reverse primer	0.5 uM	10 uM	5 ul
DMSO			2.5 ul
Mango Taq Polymerase	12.5 units	1000 units	2.5 ul
DNA template (optimal)	0.1-0.5 ng : plasmid 50 ng : gDNA		X
H2O			up to 100 ul

Step 3: Mix oil phase and aqueous phase
- Add total volume of aqueous phase to the oil phase
- Vortex in the cold room for 5 min at max speed

Step 4: PCR
- Divide the obtained emulsion from step 3 into 100 ul aliquots for PCR
- Run PCR as usual
- PCR program for CAG amplification in BAS101 cells

	Temperature °C	Time	
1	95	2 min	
2	95	10 sec	
3	52	20 sec	
4	72	2 min 30 sec	go to 2 - 5 cycles
5	95	10 sec	
6	55	20 sec	
7	72	2 min 30 sec	go to 5 - 40 cycles
8	72	2 min	
9	15	pause	

Step 5: Break the emulsion
- Add 100 ul isobutanol per PCR tube
- Pipet a few times until the PCR tube is clean
- Pool the PCRs together
- Add isobutanol to total volume (PCRs + isobutanol) of 1.5 ml
- Vortex for 5-10 sec
- Centrifuge for 2 minutes at max (14’000 rpm)
- Discard the organic phase

Step 6: PCR clean up kit
- Add 2 volumes of NTI to the remaining sample
- Load the sample onto a column and follow the rest of the instructions from Nucleospin Gel and PCR Clean-up (Machery Nagel)

9.3 Demultiplexing script

```
#################################################################
# Demultiplex sequences looking for 1 error in index sequence   #
#                                                               #
# USAGE : perl scriptname.pl input.fasta                        #
#                                                               #
# GTF / CIG / UNIL - 2015 june 25th - emmanuel.beaudoing@unil.ch #
#                                                               #
#################################################################

# list of barcodes
@list_index=(‘ATCTGTGCGAGACTAC',
             'AGACTCTACAGAGATA',
             'TCTCTCACAGTCGAGC',
             'GCTCGACTGTGAGAGA');

open WHO, $ARGV[0];

while (<WHO>) {

   if (/^>/) {$sequence = $_; $_ = <WHO> ; $sequence .= $_}

   print STDERR "\n\n\n".$c++.">>>>>>>>>$sequence<<<<<<<<\n";

   $seen=0;
   foreach $index (@list_index) {

        if ($sequence =~ /$index/) {
           $demux{$index} .= $sequence;
           print STDERR "=> $index [NO ERROR]\n";
           $seen++;
        }

        else {
           # 1 mismatch generation
           @degenerated_list_index = ();

           (@list_base_index) = split "", $index;

           for ($i = 0; $i <= $#list_base_index; $i++) {

           @neo_list_base = @list_base_index;
           $neo_list_base[$i] = "A";
           $newindexA = "";
           foreach $base (@neo_list_base) {
              $newindexA .= $base
           }
           push @degenerated_list_index,"$newindexA";

           @neo_list_base = @list_base_index;
           $neo_list_base[$i] = "C";
           $newindexC = "";
           foreach $base (@neo_list_base) {
              $newindexC .= $base
           }
           push @degenerated_list_index,"$newindexC";

           @neo_list_base = @list_base_index;
           $neo_list_base[$i] = "G";
           $newindexG = "";
           foreach $base (@neo_list_base) {
              $newindexG .= $base
           }
```

```
            push @degenerated_list_index,"$newindexG";

            @neo_list_base = @list_base_index;
            $neo_list_base[$i] = "T";
            $newindexT = "";
            foreach $base (@neo_list_base) {
               $newindexT .= $base
            }
            push @degenerated_list_index,"$newindexT";
            }

            # consider deletion, not on extremities
            for ($i = 1; $i < $#list_base_index; $i++) {
            @neo_list_base = @list_base_index;
            $neo_list_base[$i] = "";
            $newindexDel = "";
            foreach $base (@neo_list_base) {
               $newindexDel .= $base
            }
            push @degenerated_list_index,"$newindexDel";


            }

            foreach $newindex (@degenerated_list_index) {
            if ($sequence =~ /$newindex/) {
               $demux{$index} .= $sequence;
               print STDERR "=> $index [$newindex]\n";
               $seen++;
            }
            }
         }

   }
   if ($seen == 0) {print STDERR "=> 000 [NO INDEX]\n";}
}

close WHO;

foreach $index (@list_index) {
   open IND, ">$index.CCSlist.fa";
   print IND $demux{$index};
   close IND;
}
```

9.4 Manual data analysis

remove carrier return (\n) : easier to look for motives
perl -ne 'if ($_ !~ />/){chomp;print}else{print "\n",$_}' reads_of_insert.fasta > reads_of_insert_chomped.fasta

do reverse complement
perl -ne 'if (/>/){print}elsif(/CTGCTGCTGCTG/){chomp;tr/[ATGC]/[TACG]/;$n = reverse $_ ;print $n,"\n" }else{print}'
reads_of_insert_chomped.fasta > reads_of_insert_chomped.RC.fasta

#mask replacing CAG by "."
perl -ne 's/CAG/./g;print' reads_of_insert_chomped.RC.fasta > ROI.masked.fa

#displays number of CAG per CCS
perl -ne '$n=0; while (/\./g){$n++};/(\w{10}\.\.\..*\.\.\w{10})/ and print "$n"."cag $1\n"' ROI.masked.fa > ROI.CAG.fa

#displays nonCAG CCS
egrep -v "\.\.\.\.\.\.\.\.\." ROI.masked.fa > ROI.non.CAG.txt

distribution of number of CAG repeats in the complete library (w/o demultiplexing)
cat ROI.masked.fa | grep -v ">"| perl -ne '$n=0; while (/\./g){$n++}; print "$n $_"' | perl -ne '$n=0; while (/\./g){$n++};/(\w{10}\.\.\..*\.\.\w{10})/ and
print "$n"."\n"' | sort -n | uniq -c | sort -n -k2

how many fasta headers (1 header per subread)
grep -vc ">" filtered_subreads.chomped.fasta

remove carrier return (\n) (easier to grep sequence)
perl -ne 'if (/>/){print}elsif(/CTGCTGCTGCTG/){chomp;tr/[ATGC]/[TACG]/;$n = reverse $_ ;print $n,"\n" }else{print}'
filtered_subreads.chomped.fasta > filtered_subreads.chomped.RC.fasta

how many reads contain poly(CAG)
perl -ne '/>/ and s/\n/\t/; print' filtered_subreads.chomped.RC.fasta | grep CAGCAGCAG filtered_subreads.chomped.RC.tab.fasta | cut -f 1 | perl -ne
's/\d+_\d+$// and print' | sort | uniq -c | wc -l

how many CCS sequence
grep -c ">" reads_of_insert_chomped_RC.fasta

how many poly(CAG) in CCS file
grep -c "CAGCAGCAGCAG" reads_of_insert_chomped_RC.fasta

naive count with grep (look only for perfect match)
for d in GCTCGACTGTGAGAGA ATCTGTGCGAGACTAC AGACTCTACAGAGATA TCTCTCACAGTCGAGC ;do echo $d; grep -c $d
reads_of_insert_chomped_RC.fasta; done

naive count with grep (look only for perfect match) (reverse complement barcode sequences)
for d in TCTCTCACAGTCGAGC GTAGTCTCGCACAGAT TATCTCTGTAGAGTCT GCTCGACTGTGAGAGA ;do echo $d; grep -c $d
reads_of_insert_chomped_RC.fasta; done

split ROI (reads of interest/CCS) in CAG/nonCAG
perl -ne 'if (/>/){$head=$_;$_=<>;if (/CAGCAGCAGCAG/){print $head,$_}}' reads_of_insert_chomped_RC.fasta >
reads_of_insert_chomped_RC.CAG.fasta

perl -ne 'if (/>/){$head=$_;$_=<>;if (!/CAGCAGCAGCAG/){print $head,$_}}' reads_of_insert_chomped_RC.fasta >
reads_of_insert_chomped_RC.NOCAG.fasta

grep -c ">" reads_of_insert_chomped_RC.CAG.fasta reads_of_insert_chomped_RC.NOCAG.fasta

what are the pol(CAG) sequences made of?
- plasmid?
formatdb -i plasmid.fa -p F

megablast -i reads_of_insert_chomped_RC.CAG.fasta -d plasmid.fa -F F -o Aln.out
are there any plasmid sequences in NON-pol(CAG) sequences?

```
megablast -i reads_of_insert_chomped_RC.NOCAG.fasta -d plasmid.fa -F F -b 1 -v 1 -m 8 -o Aln.out

-    human genomic DNA?
# are there any human sequences in NON-pol(CAG)?
megablast -i reads_of_insert_chomped_RC.NOCAG.fasta -d  /data/cig/daf/data6/uhts-genomes/GENOMES/H_sapiens_hg19/index/hg19.fa -F F -b 1 -v 1 -m 8 | cut -f 1,2  | uniqm150603_150743_42182_c100829742550000001823181912311505_s1_p0/64272/ccs
```

chart of number of subreads per read distribution (distribution of pass number)

```
grep ">" filtered_subreads.fasta | perl -ne 'if (/\/(\d+)\/(\d+)/) {$list{$1} .=" $2"} if (eof()) {while (($a,$b) = each %list){(@splited)=split " ",$b; print $#splited+1,"\n"}}' | sort -n | uniq -c
```

distribution of start position

```
grep ">" filtered_subreads.fasta | perl -ne 'if (/\/(\d+)\/(\d+)/) {$list{$1} .=" $2"} if (eof()) {while (($a,$b) = each %list){print "$a => $b\n"}}'  | sort -n -k 3 | egrep -c "=>  0"
```

How many sequences are there with polyCAG among the 3’114 reads?

```
perl -ne 'if (/\/(\d+)\/(\d+)/) {$list{$1} .=" $2"; $_=<> ; $seq{$1}=$_} if (eof()) {while (($a,$b) = each %list){print "$a => $b\n$seq{$a}"}}' filtered_subreads.chomped.RC.fasta | grep -c "CAGCAGCAGCAG"
```

Is there any bias in CCS sequences containing poly CAG ou GTC, we expect equilibrium?

```
perl -ne 'if (/\/(\d+)\/(\d+)/) {$list{$1} .=" $2"; $_=<> ; $seq{$1}=$_} if (eof()) {while (($a,$b) = each %list){print "$a => $b\n$seq{$a}"}}' filtered_subreads.chomped.fasta | grep -c "CAGCAGCAGCAG"


perl -ne 'if (/\/(\d+)\/(\d+)/) {$list{$1} .=" $2"; $_=<> ; $seq{$1}=$_} if (eof()) {while (($a,$b) = each %list){print "$a => $b\n$seq{$a}"}}' filtered_subreads.chomped.fasta | grep -c "CTGCTGCTGCTG"
```

Sequences corresponding to plasmid?

```
megablast -i reads_of_insert_chomped_RC.NOCAG.fasta -d plasmid.fa -F F -b 1 -v 1 -m 8 -o Aln.out
```

is there any MOUSE sequence on NON-poly(CAG) sequence?

```
megablast -i reads_of_insert_chomped_RC.NOCAG.fasta -d  /data/cig/daf/data6/uhts-genomes/GENOMES/M_musculus_mm9/Mouse.fasta -F F -b 1 -v 1 -m 8 | cut -f 1,2  | uniq
```

#number of sequences containing a barcode without errors

```
perl demultiplex.01.pl reads_of_insert_chomped_RC.fasta 2> STDERR.txt
grep "=>" STDERR.txt | grep -c "NO ERROR"
```

#number of sequences containing a barcode with one error or deletion

```
grep "=>" STDERR.txt | egrep -v "NO" | wc -l
```

#number of sequences that were not assigned to a barcode

```
grep "=>" STDERR.txt | grep -c "NO INDEX”
```

#number of sequences containing a oVIN459 primer sequence (expected to be present on all sequences with CAG repeats)

```
perl demultiplex.459.pl reads_of_insert_chomped_RC.fasta 2> STDERR.txt ; grep "=>" STDERR.txt | grep -v "NO INDEX" | wc -l
```

#manual demultiplexing on subreads

```
perl -ne 'if (/>/) {/\/(\d+)\// and $list{$1}++ ; if($list{$1} ==1){print;$_=<>;print} }' filtered_subreads.chomped.RC.fasta > filtered_subreads.chomped.RC.FIRSTSUBREADS.fasta

perl demultiplex.01.pl filtered_subreads.chomped.RC.FIRSTSUBREADS.fasta 2> error.txt
```

distribution of TNRs for each barcode

```
perl -ne 's/CAG/./g;print' ATCTGTGCGAGACTAC.CCSlist.fa | perl -ne '$n=0; while (/\./g){$n++};/(\w{10}\.\.\..*\.\.\w{10})/ and print "$n"."cag   $1\n"' | sort | grep -v ">"| perl -ne '$n=0; while (/\./g){$n++}; print "$n $_"' | perl -ne '$n=0; while (/\./g){$n++};/(\w{10}\.\.\..*\.\.\w{10})/ and print "$n"."\n"' | sort -n | uniq -c | sort -n -k2
```

Number of reads per CCS

```
grep ">" filtered_subreads.fasta | perl -ne '/\/(\d+)\// and $list{$1}++; if (eof()) { open ROI, "reads_of_insert.fasta"; while ($line = <ROI>){ if ($line =~ />.+\/(\d+)\//) {chomp $line; print $line,":$list{$1}\n"} else {print $line}  } }' > reads_of_insert.nbsubreads.fasta
```

analysis of 50 bp at 3' and 5' of CCS w/ or w/o CAGs

```
grep -v ">" reads_of_insert.chomped.RC.CAG.fa | perl -ne 'print substr ($_,0,50),
"\n"' | sort | uniq -c | sort -n > reads_of_insert.RC.CAG.5.fa

grep -v ">" reads_of_insert.chomped.RC.NOCAG.fa | perl -ne 'print substr ($_,0,50),
"\n"' | sort | uniq -c | sort -n > reads_of_insert.RC.noCAG.5.fa

grep -v ">" reads_of_insert.chomped.RC.CAG.fa | perl -ne 'print substr ($_,-50),  "\n"'
| sort | uniq -c | sort -n  > reads_of_insert.RC.CAG.3.fa

grep -v ">" reads_of_insert.chomped.RC.NOCAG.fa | perl -ne 'print substr ($_,-50),  "\n"'
| sort | uniq -c | sort -n  > reads_of_insert.RC.noCAG.3.fa
```

Relaxed model: looking for CCCAAAGGG motifs

```
grep '^[ATCGN]*$' *.fastq | perl -ane 'chomp;$i++;print ">withoutRepair_$i\n$_\n";'> subreads.chomped.fasta

grep -v ">" subreads.chomped.fasta | perl -ne 'print;@A=$_=~/(C{1,3}A{1,3}G{1,3})/g; print "===>",@A,"\n" ;printf "count:%d\n
\n",$#A+1' > repeats_count.txt

grep -v ">" subreads.chomped.fasta | perl -ne '@A=$_=~/(C{1,3}A{1,3}G{1,3})/g;printf "%d\n\n",$#A+1'| sort -n | uniq -c >
repeats_count.txt
```

looking for CAG motifs selectively within sequences flanked by genomic DNA

```
grep -v ">" *.chomped.fasta | perl -ne '/TCTGTGATCCCC(.+)CATTCCCGGCTA/ and print $1,"\n"; '| perl -ne '@A=$_=~/(C{1,3}
A{1,3}G{1,3})/g; printf " %d\n",$#A+1' > Acount.txt

grep -v ">" *.chomped.fasta | perl -ne '/TAGCCGGGAATG(.+)GGGGATCACAGA/ and print $1,"\n"; '| perl -ne '@T=$_=~/(C{1,3}
T{1,3}G{1,3})/g; printf " %d\n",$#T+1' > Tcount.txt sort -n Acount.txt Tcount.txt | uniq -c| perl -ne '/(\d+)\s+(\d+)/ and print
"$1;$2\n"' > total.csv

grep -v ">" RING2_606-607_reads_of_insert.chomped.fasta | perl -ne '/TAGCCGGGAATG(.+)GGGGATCACAGA/ and print
$1,"\n"; /TCTGTGATCCCC(.+)CATTCCCGGCTA/ and print $1,"\n""| perl -ne 'print">>>>>>>>>\n",$_,"\n"; @A=$_=~/(C{1,3}
A{1,3}G{1,3})/g; @T=$_=~/(C{1,3}T{1,3}G{1,3})/g; print "===>@T\n" ; printf "count: %d\n\n",$#T+1,"\n" ; print "===>@A
\n" ;printf "count: %d\n\n",$#A+1' | less
```