




# Systematic microsatellite repeat expansion cloning and validation

Kushal J. Rohilla<sup>1</sup> · Katy N. Ovington<sup>1</sup> · Adrian A. Pater<sup>2</sup> · Maria Barton<sup>1</sup> · Anthony J. Henke<sup>2</sup> · Keith T. Gagnon<sup>1,2</sup> 

Received: 7 January 2020 / Accepted: 4 April 2020  
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## Abstract

Approximately 3% of the human genome is composed of short tandem repeat (STR) DNA sequence known as microsatellites, which can be found in both coding and non-coding regions. When associated with genic regions, expansion of microsatellite repeats beyond a critical threshold causes dozens of neurological repeat expansion disorders. To better understand the molecular pathology of repeat expansion disorders, precise cloning of microsatellite repeat sequence and expansion size is highly valuable. Unfortunately, cloning repeat expansions is often challenging and presents a significant bottleneck to practical investigation. Here, we describe a clear method for seamless and systematic cloning of practically any microsatellite repeat expansion. We use cloning and expansion of GGGGCC repeats, which are the leading genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD), as an example. We employ a recursive directional ligation (RDL) technique to build multiple GGGGCC repeat-containing vectors. We describe methods to validate repeat expansion cloning, including diagnostic restriction digestion, PCR across the repeat, and next-generation long-read MinION nanopore sequencing. Validated cloning of microsatellite repeats beyond the critical expansion threshold can facilitate step-by-step characterization of disease mechanisms at the cellular and molecular level.

## Introduction

The Human Genome Project was completed in April 2003 and resulted in sequence for the vast majority of the genome, including about 99% of genic regions. However, a surprising fraction of the genome still remains unsequenced or unmapped due to highly repetitive regions. These uncharacterized regions are primarily composed of repetitive DNA sequences including short tandem repeats (STRs), or microsatellites (Miga et al. 2015). Microsatellites consist of simple sequence motifs of one to six nucleotides repeated at least 5–15 times at a genetic locus (Ellegren 2004; Rohilla and Gagnon 2017). Microsatellites account for a significant source of genetic diversity and are the basis for DNA fingerprinting techniques (Jeffreys et al. 1985; Roewer 2013;

Weischenfeldt et al. 2013). With respect to disease, STRs can be unstable and undergo expansion that leads to pathogenicity in succeeding generations (Brouwer et al. 2009; Paulson 2018). These expansions are the leading cause of a growing list of more than 30 neurological disorders including Huntington's disease (HD), numerous spinocerebellar ataxias (SCA disorders), fragile X syndrome (FXS) and fragile X-associated tremor/ataxia syndrome (FXTAS), myotonic dystrophy type 1 (DM1) and type 2 (DM2), amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD) (Mirkin 2007; Rohilla and Gagnon 2017; Weischenfeldt et al. 2013; Zhao and Usdin 2015). Improved methods to clone and sequence long, uninterrupted repetitive DNA sequences, especially microsatellites, will improve understanding of their contribution to genome evolution and disease etiology and advance therapeutic strategies.

## Molecular disease mechanisms in microsatellite repeat expansion disorders

Microsatellite repeat expansions are known to cause disease through three common pathogenic mechanisms. These are loss-of-function for a gene, gain-of-function for the repetitive sequence at the RNA level, or gain-of-function at the protein level. Loss-of-function can occur when gene

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00439-020-02165-z>) contains supplementary material, which is available to authorized users.

✉ Keith T. Gagnon  
ktgagnon@siu.edu

<sup>1</sup> Department of Biochemistry and Molecular Biology, School of Medicine, Southern Illinois University, Carbondale, USA

<sup>2</sup> Department of Chemistry and Biochemistry, Southern Illinois University, Carbondale, USA

expression is silenced at the transcriptional level, such as by epigenetic modifications, or at the translational or protein function level (Evans-Galea et al. 2013; He and Todd 2011). For example, in FXS expansion of CGG repeats in the *fragile X mental retardation 1 (FMR1)* gene leads to hypermethylation of the promoter, thereby silencing expression of the gene product fragile X mental retardation protein (FMRP) (Verkerk et al. 1991). Loss of gene function at the protein level can occur due to insertion of an abnormal number of repetitive amino acids in the translation product. Examples include HD, several SCA disorders (SCA1, 2, 3, 6, 7, and 17), and spinal and bulbar muscular atrophy (SBMA) (Labbadia and Morimoto 2013; Shoubridge and Gecz 2012).

RNA-mediated gain-of-function arises when repeat expansions are transcribed into expanded tandem repeat-containing RNAs, which we refer to as xtrRNAs, that can exert a variety of toxic effects. These include sequestration and functional depletion of RNA-binding proteins or translation into aberrant and unnatural repetitive polypeptides (Chen et al. 2009; DeJesus-Hernandez et al. 2011; Green et al. 2016; Zu et al. 2011). For example, in DM1, the CTG repeat expansion in the *myotonic dystrophy protein kinase (DMPK)* gene causes sequestration of essential RNA-binding proteins and forms RNA foci in the nucleus of muscle cells (Mankodi et al. 2000). Translation of expanded repeat RNA embedded in coding regions can lead to aggregation of the host proteins, thereby disrupting normal gene functions (Labbadia and Morimoto 2013; Zhao et al. 2016). Alternatively, independent translation of repeats, such as in non-coding regions, can lead to a protein-mediated gain-of-function where translation often occurs through repeat-associated non-AUG (RAN) translation. This overcomes the requirement of a canonical AUG start codon. RAN translation was discovered in patient SCA8 cells in 2011 (Zu et al. 2011) and is now known to be common in other repeat expansion disorders (Cleary et al. 2018). The products of RAN translation are implicated in several distinct pathological mechanisms that primarily involve protein aggregation or disruption of membrane-free organelle formation and function (Cleary and Ranum 2017; Green et al. 2017; Nguyen et al. 2019).

### Bottlenecks in understanding microsatellite repeat expansion disorder mechanisms

Understanding disease mechanisms and testing potential therapeutic strategies can benefit tremendously from simplified experimental systems that involve vectors containing long microsatellite repeat tracts. However, this necessitates cloning and validation of large repeat expansions. Without satisfactory methods to do so, experimental systems have been developed where the repeat length is not precisely known, where the repeat expansions are interrupted by other

sequences, or where the repeat number is possibly below the pathogenic repeat number (Batra and Lee 2017; de Haro et al. 2006; Mankodi et al. 2000; Seznec et al. 2001; Wen et al. 2014).

Some methods for cloning uninterrupted repeat expansions have been described but possess various shortcomings. Most methods involve PCR-based amplification steps or oligo annealing tricks to generate a starting pool of clones with different repeat numbers (Jiang et al. 1996; Laccone et al. 1999; Matsuura and Ashizawa 2002; Ohshima et al. 1996; Ordway and Detloff 1996; Thys and Wang 2015; Wen et al. 2014). However, PCR is usually not efficient across large, highly GC-rich sequences (Mamedov et al. 2008). Complementary repeat oligonucleotides have been used to generate repeat lengths of different sizes but longer repeats tend to form secondary structures like hairpins or cause strand slippage that produces instability (Grabczyk and Usdin 1999; Thys and Wang 2015). In these methods, control of repeat length and orientation may rely on trial and error or ratios of oligonucleotides.

Here, we describe a PCR-free approach to generate clones containing desirable repeat lengths without sequence interruptions. Cloning begins with a short, synthetic repeat-containing duplex DNA and subsequent rounds of cloning consecutively use the vectors from previous rounds as a source for the repeat sequence. The number of repeats can be easily controlled by selecting the insert to be used at each round. We chose to clone the GGGGCC hexanucleotide repeat expansion, because it exhibits pure GC-content and the protocol developed with this repeat forms the basis of cloning and validating other disease-associated STRs. The method we use for cloning GGGGCC repeat expansions is referred to as recursive directional ligation (RDL) (Meyer and Chilkoti 2002; Mizielinska et al. 2014) and is similar to the early approach used by Usdin and co-workers (Grabczyk and Usdin 1999). We then demonstrate how to validate cloning success by restriction enzyme digestion and PCR, which can measure the length of the repeat expansion and serve as a quick evaluation of cloning success. The precise repeat expansion length and sequence is then unambiguously determined using an accessible next-generation long-read nanopore sequencing workflow with the MinION sequencer from Oxford Nanopore Technologies (Ebbert et al. 2018; George et al. 2017).

### C9ORF72-associated FTD and ALS

A GGGGCC hexanucleotide repeat expansion in the *chromosome 9 open reading frame 72 (C9ORF72)* gene was discovered to be the leading genetic cause of FTD and ALS (DeJesus-Hernandez 2011; Renton et al. 2011). FTD is one of the most common forms of dementia after Alzheimer's disease (AD) for patients under age 60 (Hodges and Piguet

2018). The disease progression is highly variable, with approximately 7–9 years on average from the start of the symptoms, and death usually occurs from respiratory complications (Hodges and Piguet 2018). Approximately 20% of familial FTD cases are caused by the *C9ORF72* mutation (DeJesus-Hernandez et al. 2011; Renton et al. 2011). ALS is a progressive and a fatal neurodegenerative disease that primarily affects the upper and lower motor neurons of the brain and spinal cord (Mitchell and Borasio 2007; Rowland and Shneider 2001). It is characterized by rapid loss of voluntary muscle control, muscle weakness, and paralysis leading to a premature death due to respiratory failure within 3–5 years (Hardiman et al. 2011; Mitchell and Borasio 2007). Familial ALS (fALS) constitutes about 5–10% of ALS cases with gene mutations in the family (Zarei et al. 2015).

The combined *C9ORF72* repeat-associated FTD and ALS diseases are now commonly referred to as C9FTD/ALS. Healthy individuals appear to have less than 24 GGG GCC repeats in the *C9ORF72* genetic locus, whereas in patients this number ranges from 25 to 2000+ repeats (DeJesus-Hernandez et al. 2011; Iacoangeli et al. 2019; Renton et al. 2011). This disease also exhibits classic pathological features where the repeat-containing xtrRNA forms focal aggregates in the cell nuclei (Rohilla and Gagnon 2017). Repeat-containing RNA that escapes into the cytoplasm can also undergo RAN translation into repetitive poly-dipeptides (Cleary et al. 2018; Green et al. 2016; Zu et al. 2011).

## Results and discussion

### Vector considerations for cloning and expression

Criteria for plasmid vector selection during traditional cloning usually includes factors such as high copy number plasmids and a suitable multiple cloning site (MCS) (Al-Allaf et al. 2013; Corchero and Villaverde 1998). However, care needs to be taken for cloning sequences that are prone to deletion and rearrangements due to highly repetitive sequence (usually also high GC content), intrinsic folding propensities (such as G-quadruplex formation) or encoding of toxic protein products. Propagation of plasmids containing unstable or toxic DNA inside cells can often lead to deletion or truncation of the insert sequence (Godiska et al. 2010). One method to reduce potential rearrangements or toxic genetic burden is to use moderate- or low-copy plasmids. Alternatively, highly repetitive sequences can be more stably propagated in transcription-free linear vectors. Vector-driven transcription into the insert sequence can accelerate deletion of the repeat sequence (Stueber and Bujard 1982). Also, transcription due to the presence of active promoter sites in the cloned insert into the vector backbone

can interfere with the plasmid replication machinery thereby causing deletion of the repeat sequence (Stueber and Bujard 1982). Thus, for very troublesome repeat cloning projects, such vectors may be valuable.

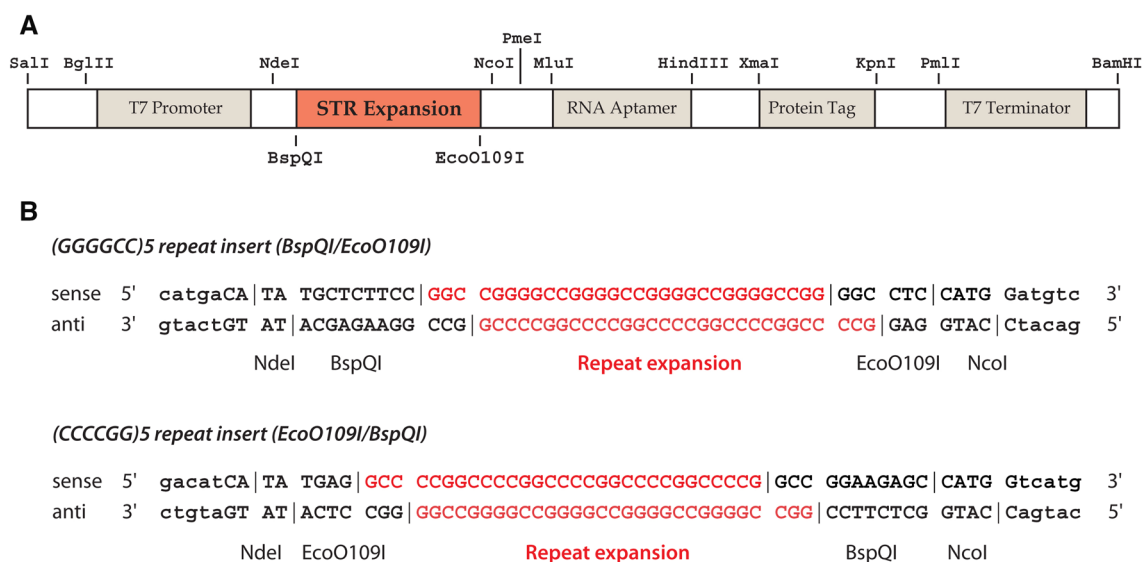
In our method, we focus on commonly used mammalian expression vectors, typically high-copy number, but utilize a tetracycline/doxycycline-inducible promoter system (Gossen and Bujard 1992; Gossen et al. 1995). Inducible promoters may minimize toxicity and can facilitate downstream studies in mammalian model cells, which may be an important consideration for expression of disease-associated xtrRNAs and repetitive polypeptides (Cheng et al. 2018). Cloning of repeat expansions by our method involves the introduction and use of type IIS restriction endonucleases (Szybalski et al. 1991). These are incorporated into the MCS by an initial round of cloning to prepare a parent vector, then used for multiple rounds of STR expansion cloning. Unfortunately, the recognition sequences for type IIS enzymes can often occur redundantly at other positions within a plasmid. Careful selection of the plasmid vector for cloning or site-directed mutagenesis (SDM) can resolve this issue.

### Introducing custom restriction sites for recursive direction ligation (RDL)

The vector we selected for custom restriction site engineering was pTRE3G from Clontech. pTRE3G contains a hybrid CMV promoter driven by an inducible Tet-ON system (Baron and Bujard 2000). The CMV promoter is silent due to insertion of several tetracycline repressor protein (TetR)-binding sites. TetR itself cannot bind the promoter elements unless it is bound to doxycycline, a synthetic tetracycline derivative (Berens and Hillen 2003). The parent plasmid was created by modifying pTRE3G with a custom MCS that contains two new sites for microsatellite repeat expansion cloning. Our new MCS was ordered as a gBlock from Integrated DNA Technologies (IDT) and inserted using the SalI and BamHI restriction sites of pTRE3G. We included a T7 promoter upstream of the repeat cloning site and an RNA aptamer, encoded peptide tags, and a T7 terminator downstream (Fig. 1a). These genetic elements are specific to our applications and contain unique restriction sites between them to enable straightforward addition or removal. The strategy of creating a custom MCS is optional but may be valuable for other researchers. We named our new vector plasmid for inducible non-coding repeat RNA expression with a Tet-On 3G promoter, or pINC3G.

### Restriction endonuclease selection for repeat cloning by RDL

The selection of restriction enzymes for STR cloning is essential to recursive or iterative cloning of expanded



**Fig. 1** Design of a custom MCS for pINC3G vector generation and an insert for the first round of GGGGCC/CCCCGG repeat cloning by RDL. **a** A custom MCS was designed to contain genetic expression elements with restriction sites for facile addition or removal. The synthetic MCS replaced the standard MCS of pTRE3G using SalI and

BamHI restriction sites to generate a new plasmid called pINC3G for GGGGCC repeat expansion cloning. **b** Design of the insert containing five G-rich (GGGGCC) or five C-rich (CCCCGG) repeats used for the first round of RDL cloning

repeats. Fortunately, there are many type IIS restriction endonucleases, and even some traditional restriction enzymes, to choose from. Type IIS enzymes cut “at a distance” from their recognition sequence (Pingoud and Jeltsch 2001; Szybalski et al. 1991). Thus, there is no need for a recognition sequence at the ends of the repetitive insert, only compatible sticky ends. Once an insert is successfully cloned, cleavage with type IIS restriction enzymes can release the insert with no recognition sites but with defined sticky ends (Engler et al. 2008, 2009). The cut site lies anywhere from 0 to about 20 bases away from the recognition sequence, depending on the enzyme, and creates staggered ends of one to several bases (Pingoud and Jeltsch 2001). Proper selection of two distinct type IIS enzymes, or a combination of a type IIS and type IIP (“traditional”) restriction enzyme, can produce compatible staggered sticky ends that enable directional ligation during multiple rounds of expansion. By this strategy, any STR can conceivably be cloned by the RDL method.

For our studies, we chose to clone GGGGCC repeat expansions associated with C9FTD/ALS. The GGGGCC repeats have been cloned by others in various ways. While exploring different STR cloning methods, we found that the method reported by Isaacs and colleagues (Mizielinska et al. 2014), which utilized RDL, was the most reliable and systematic. This method makes use of one traditional type IIP restriction enzyme, EcoO109I (also known as DraII), and one type IIS restriction enzyme, BspQI, and is very similar to that described by Usdin and colleagues much earlier

(Grabczyk and Usdin 1999). BspQI recognizes the asymmetric sequence GCTCTTC and cleaves one base to the right on the sense or “top” strand and three bases further down on the antisense or “bottom” strand yielding a three-base overhang at the 5' end, written as GCTCTTC (1/4) (Fig. 1b). SapI is an isoschizomer of BspQI with an optimal reaction temperature of 37 °C, which is compatible with that of EcoO109I and helpful for one-step double-digest reactions. EcoO109I recognizes the palindromic sequence RGGNCCY and cuts after RG to produce a 5' GNC overhang.

### Designing the initial repeat insert for the first round of RDL

We designed a small 64 base-pair double-stranded DNA oligonucleotide for the first round of cloning (Fig. 1b). The oligo comprised 5 GGGGCC repeats flanked by BspQI and EcoO109I restriction sites at the 5' and 3' ends, respectively, of the sense “top” strand. Adjacent to these restriction sites were external restriction sites for NdeI (located upstream of the BspQI site) and NcoI (located downstream of the EcoO109I site) for insertion into the custom MCS, which did not contain the necessary BspQI and EcoO109I sites when originally designed. This insert design should allow cloning into most vectors without the need for introducing a custom MCS, which is an optional step that we included. We have used a similar insert design and the same restriction enzyme cut sites for also cloning the antisense CCCCCG repeats into pINC3G (Fig. 1b).



## Selection of competent *E. coli* host cells and growth conditions

Special *Escherichia coli* strains have been created that maximize the stability and propagation efficiency of GC-rich and highly repetitive sequences for molecular cloning. Most of these strains carry *recA1* or *recA13* mutations for reduced recombination of the cloned DNA with the host chromosome by rendering the DNA repair enzymes inactive. They also carry *endA* mutations to eliminate the non-specific activity of endonuclease I and thereby generate high quality plasmid preparations. We experimented with different chemically competent *E. coli* strains, including DH5 $\alpha$ , NEB Stable (New England Biolabs, NEB), Max Efficiency Stbl2 (Thermo Fisher Scientific), and One Shot Stbl3 (Thermo Fisher Scientific) cells to check for stable propagation of GGGGCC-containing plasmids (Table 1). Stbl3 cells are commonly used for cloning lentiviral vectors containing direct repeat sequences. While we did not test TOP10 or Stbl4 competent cells (Thermo Fisher Scientific), both are common for cloning. Stbl4 electrocompetent cells are commonly used for generating genomic and cDNA libraries and TOP10 cells are generally used in routine cloning experiments, similar to DH5 $\alpha$ , to propagate high copy number plasmids. Plasmids containing 20 GGGGCC repeats or less could be stably replicated in DH5 $\alpha$  cells. However, as the repeat expansion size increased cloning experiments failed to produce colonies with the desired number of repeats. On the other hand, we did not observe any significant differences in the repeat cloning efficiency when using NEB Stable, Stbl2, or Stbl3 cells (data not shown). Thus, in general, NEB Stable cells were used when > 20 GGGGCC repeats were being cloned or propagated.

Lower growth temperatures have been known to improve retention and propagation of certain plasmids in *E. coli* (Liao 1991). Cultures are generally grown at temperatures below the usual 37 °C for carrying plasmids with high GC content (Godiska et al. 2010). The growth temperature can range from 23 to 30 °C during the transformation recovery phase and between 16 and 30 °C for multiple days of culture.

When we tested different growth temperatures we found that growth at 16 °C on day 1, 18 °C on day 2 and 23 °C on day 3–4 did not differ substantially from continuous growth at 23 °C for up to 4–5 days; similar plasmid yield and quality were obtained for pINC3G plasmids containing > 20 repeats (data not shown). We recommend using transformed colonies within 4 days and harvesting liquid cultures for mini-prep and midi-prep before they enter stationary phase. Due to extended incubation periods, we also recommend the use of carbenicillin instead of ampicillin.

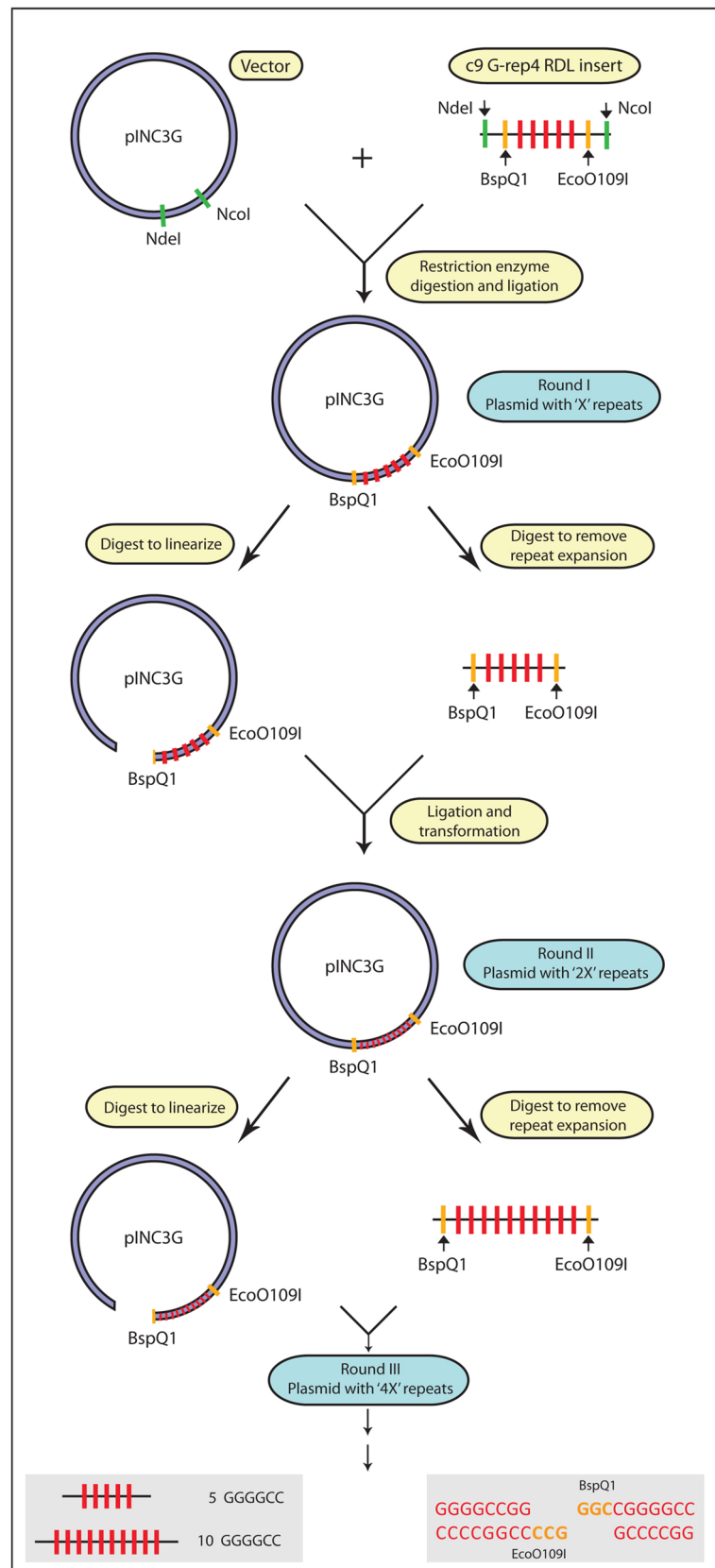
## RDL for cloning of GGGGCC repeats and other microsatellite repeat expansions

Cloning of a small number of microsatellite repeats may be accomplished by relatively routine molecular cloning methods. However, obtaining larger STR expansions in pathogenic ranges with specific defined sizes and without sequence interruptions is difficult. We have found that recursive directional ligation (RDL) makes seamless cloning of large repeat expansions feasible and reliable. The core principles of RDL for STR expansion cloning are (1) the use of at least one type IIS restriction enzyme, (2) a second restriction enzyme that creates compatible but asymmetric sticky ends for directional ligation, and (3) recursive or iterative removal and religation of repeat-containing fragments (Fig. 2). In our case of GGGGCC repeat cloning, BspQI creates a 5' GGC overhang on the insert, while EcoO109I creates a 5' GCC overhang. These compatible ends allow for iterative cloning of the released repeat-containing DNA from a previous round of cloning. The fragment will seamlessly and directionally insert into a single BspQI cut site. The EcoO109I site of the released fragment is lost upon integration and only the very terminal downstream EcoO109I site originally present in the vector is maintained. Because BspQI is a type IIS enzyme and its recognition site resides on the plasmid side of insertion (upstream of the repeat in our case), its recognition sequence is always maintained. After each round of cloning, there is only one BspQI site and one EcoO109I site.

**Table 1** Comparison of commercially available and commonly used competent *E. coli* cell lines for cloning unstable repeat sequences

Strain	Transformation efficiency (cfu/ $\mu$ g)	Uses		
		Routine cloning	Cloning unstable repeats	Transformation of methylated DNA
DH5alpha	> 1 $\times$ 10 <sup>9</sup>	✓	×	✓
One Shot™/ MultiShot™ TOP10	> 1 $\times$ 10 <sup>9</sup>	✓	×	✓
NEB <sup>R</sup> Stable	1–3 $\times$ 10 <sup>9</sup>	✓	✓	✓
Max Efficiency™ Stbl2™	> 1 $\times$ 10 <sup>9</sup>	✓	✓	✓
One Shot™/ MultiShot™ Stbl3™	> 1 $\times$ 10 <sup>8</sup>	✓	✓	✓
ElectroMAX™ Stbl4™	> 5 $\times$ 10 <sup>9</sup>	✓	✓	✓

**Fig. 2** Recursive direction ligation (RDL) strategy for cloning GGGGCC repeats and other STRs. The step-by-step protocol is shown using BspQ1 and EcoO109I restriction enzymes and an initial insert bearing 5 GGGGCC repeats. This strategy is applicable to other STRs with the correct choice of restriction enzymes



A key to RDL design is identifying restriction enzymes that can produce compatible but asymmetric sticky ends. The insert should have two 5' overhangs (or two 3' overhangs) that produce complementary overhangs for directional pairing. For example, a G overhang on sense strand and C on the antisense strand will ensure directionality through specific G–C pairing between the insert and vector but will only allow ligation in one orientation since C–C and G–G pairs are incompatible. Any STR repeat can be conceivably cloned with the selection of type IIS restriction enzymes that only generate a single nucleotide overhang. Indeed, the GGGGCC repeat we have selected could have been created using two different type IIS enzymes, such as AlwI with BccI, BciVI with BmrI, or BspMI with BbsI (Figure S1A). Another example is cloning of trinucleotide GAA repeats with BseRI and BsgI by Usdin and colleagues, which used a similar strategy but involved controlling repeat size by capped and uncapped insert ratios (Grabczyk and Usdin 1999). Both BseRI and BsgI could create dinucleotide 3' overhangs on the insert that can be compatible and directional with proper design, such as 3' CT and GA overhangs.

Even though microsatellite sequences are diverse, a number of sets of type IIS restriction enzymes can be chosen that generate one, two, three, and four nucleotide overhangs. In supplemental Figure S1B, we provide examples of type IIS restriction enzyme selections and suggest model synthetic inserts for cloning CAG, CTG, CGG, CCTG, and ATTCT microsatellite expansions that are all associated with repeat expansion disorders (Rohilla and Gagnon 2017). It should be noted that type IIS enzymes with longer gaps between recognition and cleavage sites tend to be less accurate (Pingoud et al. 2014). Thus, enzymes with shorter gaps, such as (0/2) or (1/4), would generally be preferable over longer gaps like (10/12) or (16/18). Also, although concatemerization of inserts is possible during cloning, it is generally rare and can be controlled or exploited by altering insert and vector ratios (Grabczyk and Usdin 1999; Mizielinska et al. 2014).

In the first round of cloning, only the synthetic insert is used. Once integrated into the parent plasmid, the repeat sequence can be released and used for multiple rounds of cloning. Until the repeat sequence reaches a critical size, it is usually preferred to use the synthetic insert. For example, we used the synthetic insert for two rounds of cloning. This is largely because gel-purification is difficult for very small repeat-containing inserts released from plasmid and poor staining necessitates digestion of several micrograms of plasmid. For our GGGGCC STR expansion cloning, once we reached ten repeats, we began using plasmid-derived insert to expand to 20, 40, 80 and beyond. At any point an insert from earlier rounds of cloning can be inserted. To obtain 50 repeats, 5 could be added to 20, then that insert doubled. Alternatively, 10 could be added to 40. By this method, quite precise repeat sizes can be achieved.

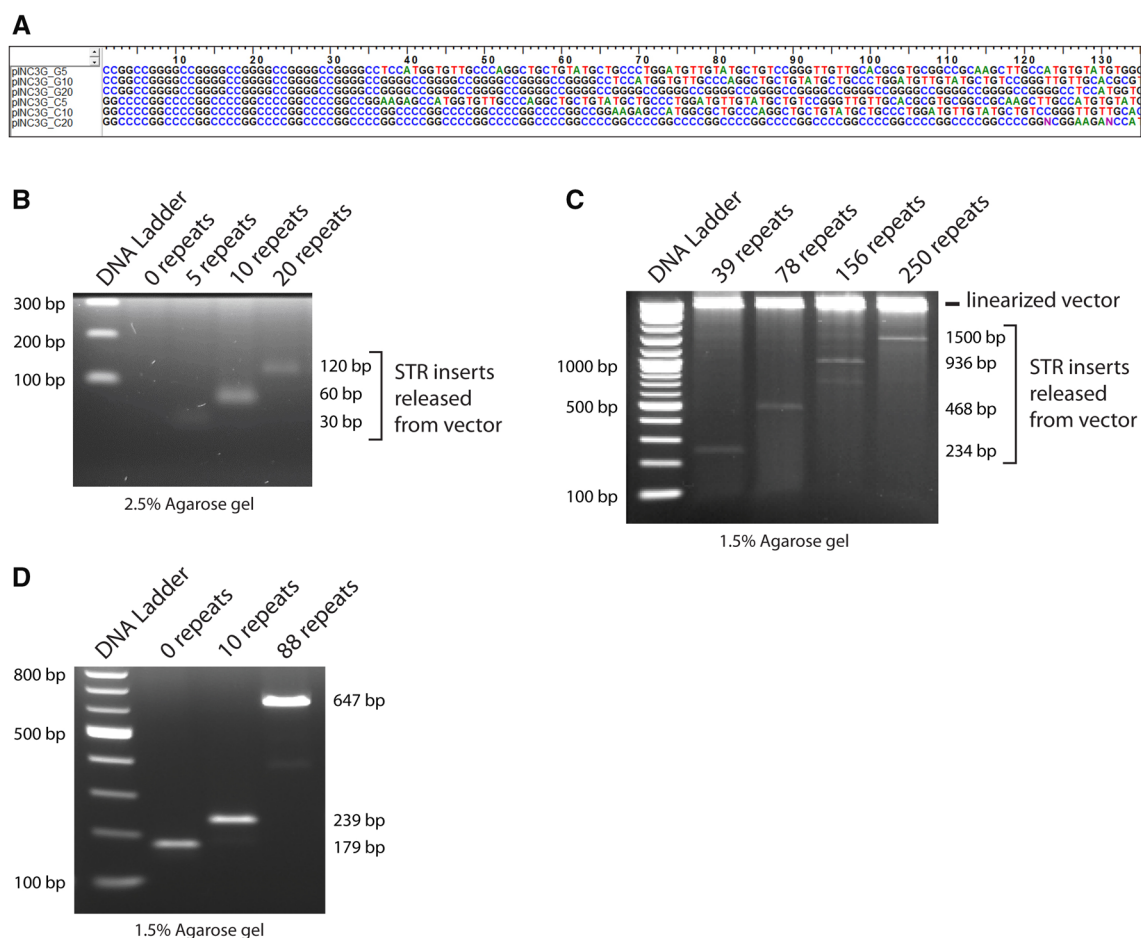
When performing our initial cloning, we obtained 39 repeats rather than the target of 40 repeats. We decided to expand this repeat to 78 GGGGCC repeats, and then further to 156 repeats. In another pINC3G vector, we obtained 88 repeats by adding 78 repeats to a 10-repeat construct. As described below, we validated our cloning in each round with a variety of methods. Along with Sanger sequencing, this included diagnostic restriction enzyme digestions, PCR amplification across the repeat expansion, and next-generation long-read MinION nanopore sequencing to unambiguously characterize repeat expansion size and sequence.

### Sanger sequencing of microsatellite repeat expansions

Because of the possibility of contraction or expansion of repeats, a minimum of eight colonies are typically sent for initial Sanger sequencing when STRs are 30–40 repeats or less. In our cloning experiments, a library of clones containing different repeat numbers were obtained after the iterative rounds of recursive directional ligation. These included pINC3G vectors with 5, 10 and 20 GGGGCC and CCCC GG repeats (Fig. 3a). Successful sequencing was achieved for clones containing up to 39 repeats, but our pINC3G-G39, for example, could only be accurately sequenced once. When plasmid sequences are validated, we recommend storage at  $-80^{\circ}\text{C}$  with minimal freeze–thaw cycles, preferably as precipitated or dry stocks, to avoid potential structural instability of the repetitive insert sequences.

### Diagnostic restriction enzyme digestion to identify cloned repeat expansions

During RDL cloning, ligation of insert and vector as well propagation in *E. coli* host cells can generate a variety of insert types, including no insert, multiple inserts, or possibly inserts in the wrong orientation. One way to identify positive clones of the appropriate size is to exploit the use of restriction enzymes that flank the insert. The restriction enzymes will release the insert, which can then be analyzed by agarose gel electrophoresis to determine if it is the expected size. If so, it can be cautiously assumed that the cloning may have proceeded correctly. For our studies, plasmid was extracted from selected clones and then subjected to digestions with BspQI (or SapI) and EcoO109I. The digested plasmid products were loaded on 2.5% agarose gel to check for inserts of sizes smaller than 120 bp (Fig. 3b) and on a 1.5% agarose gel to verify repeats greater than 120 bp (Fig. 3c). The amount of plasmid used for restriction enzyme digestion depends on the size of the insert under verification. For example, to visualize the repeat plasmid containing 5 GGGGCC repeats (30 bp), 5  $\mu\text{g}$  of pINC3G-G5 was digested. On the other hand, to visualize plasmid containing 78 repeats



**Fig. 3** Sanger sequencing, diagnostic restriction digestion, and PCR to validate GGGGCC/CCCCGG microsatellite repeat expansion cloning. **a** Sanger sequencing results for 5, 10, and 20 GGGGCC and CCCCCG repeats cloned by RDL. **b** Diagnostic restriction enzyme digestion for plasmid vectors containing 0–20 GGGGCC STRs. Reaction products were resolved on a 2.5% agarose gel. **c** Diagnostic

digests for vectors containing 39, 78, 156, and putatively 250 GGGGCC STRs. Reaction products were resolved on a 1.5% agarose gel. **d** PCR amplification across 0, 10 and 88 repeat GGGGCC STRs in pINC3G vectors. Primers flanked the repeat expansion region and reaction products were resolved on a 1.5% agarose gel

(468 bp), 500 ng of pINC3G-G78 plasmid was digested. It is not uncommon to observe expansions and contractions of the repeat during cloning; despite the precautions we have outlined. Bands of varying sizes were seen upon performing diagnostic restriction enzyme digestion on multiple clones expected to contain 156 repeats (Figure S2A). While expanding pINC3G-G78 to create clones with 156 repeats, a clone with an apparent number of ~250 repeats was identified by diagnostic digestion (Fig. 3c).

### PCR amplification across microsatellite repeat expansions

PCR analysis is another method to determine the insertion of correct sequence into the plasmid. PCR can be performed on purified plasmids or possibly from colonies or overnight cultures. To achieve PCR across larger

repeat expansions, polymerase enzymes or buffer systems optimized for high GC content are recommended. The PCR thermal cycling conditions may need to be adjusted depending on the length of the expected PCR amplicon. Also, potential secondary structures generated due to the presence of GC-rich regions can cause stalling of the DNA polymerases and can result in non-specific or incomplete amplicon products. Addition of organic solvents like DMSO or use of modified dNTPs, such as 7-deaza-dNTP, may improve amplification of GC-rich sequences (Musso et al. 2006). For our study, PCR primers flanking the pINC3G repeat cloning site were designed to amplify the sequence containing GGGGCC repeats. PCR was then performed on the parent plasmid pINC3G containing zero repeats, pINC3G-G10, and pINC3G-G88 (Fig. 3d). PCR amplicons were then resolved on an agarose gel to visualize the size of the repeat-containing products.



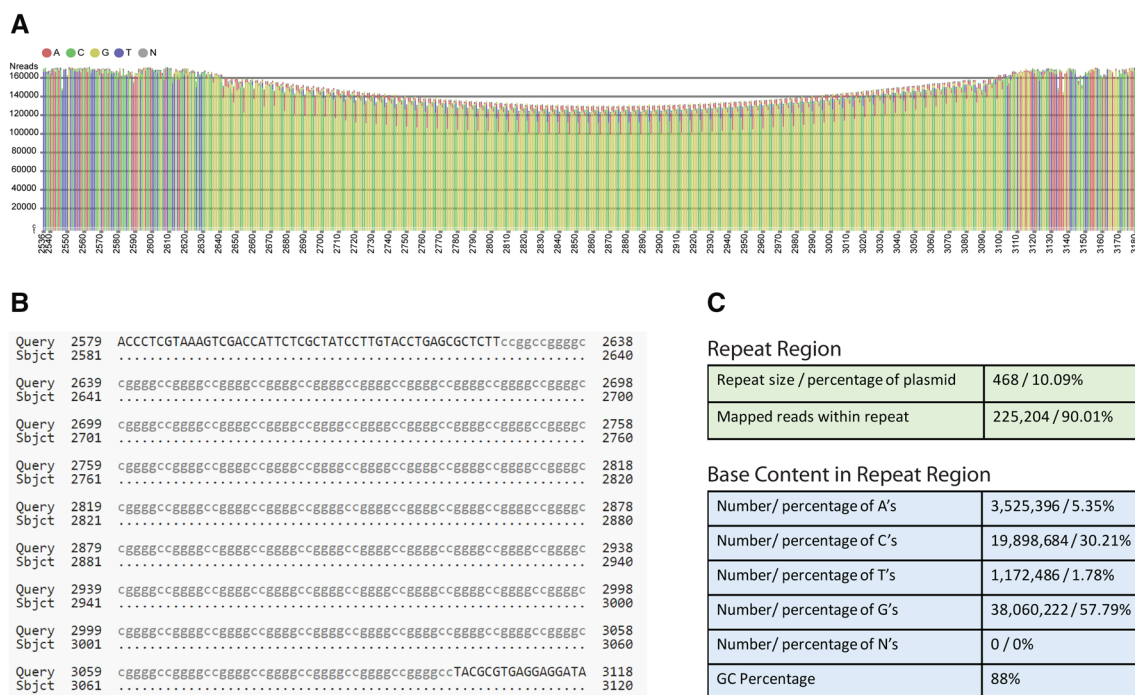
## MinION nanopore sequencing of expanded microsatellite repeat sequences

The most widely used next-generation sequencing technology, the Illumina Sequencing platform, has many advantages, such as a relatively low error-rate, but its read lengths are limited and require bioinformatic read reconstruction to generate longer sequences. The relatively short read lengths generated by Illumina (<250 bp) can sequence repetitive sequence but are unable to place them in the context of a larger repeat since the distance to unique flanking sequence is unknown. Thus, Illumina sequencing is not suitable for characterization of long repetitive sequences (Bahlo et al. 2018). Other platforms such as Nanopore sequencing from Oxford Nanopore Technologies (ONT) and the Sequel and RSII platforms from Pacific Biosystems (PacBio), which utilize the Single Molecule Real-Time Sequencing (SMRT) method, have significantly longer read length (usually 10–40 kbp), which are more suitable to characterizing repeat expansions (Ameur et al. 2019; Ebbert et al. 2018). In our study, we developed and validated a simple workflow that uses the portable MinION nanopore sequencer (ONT) (Figure S3). The MinION sequencer can generate reads as long

as 100 kb (George et al. 2017). Flow cells are reusable to a certain extent and the MinION sequencer is free with the purchase of a few necessary reagents and flow cells. Low cost, portability, and a growing community with software, users and support make the MinION an attractive choice for routine laboratory sequencing of repetitive vector sequences.

We used the MinION nanopore sequencer to sequence the entire 4.7 kb pINC3G-G78 plasmid. Only a few micrograms of plasmid and a short 4 h run time were needed to obtain high-quality sequence. The sequencing run generated 490,211 reads with a mean Phred score of 12.02. Read length distribution was mostly found within the predicted plasmid length of ~4681 bp (Figure S2B). We generated a consensus sequence from the reads, as well as mapped them to a reference plasmid sequence for pINC3G-G78. When building a consensus sequence, a specific position is assigned to the nucleotide with the highest count at that position. The consensus sequence results were visualized and revealed high accuracy base-calling with only a shallow dip in coverage in the middle of the repeat region (Fig. 4a).

To determine if reduced sequence coverage in the repeat region was due to the repeat expansion sequence itself, a sequence of the same size (468 bp) was chosen elsewhere



**Fig. 4** Sequencing of a 78 GGGGCC STR expansion cloned into pINC3G using next-generation long-read MinION Nanopore sequencing. **a** Nucleotide plot of the 78 GGGGCC repeat region of pINC3G-G78 with 100 bp of flanking sequence illustrates the distribution of nucleotides from all reads mapped at each position of the consensus. **b** BLASTn alignment of the generated consensus sequence (Sbjct) against the expected pINC3G-G78 sequence (Query) reference. Nucleotides identical to the reference (Query)

are indicated by dots (Sbjct). Perfect base-calling and consensus sequence for alignment across the repeat is demonstrated. **c** Tables summarizing the mapping statistics for the repeat-containing region. The number of adenine (A) bases assigned was found to be higher than for thymine (T) within the pure GC repeat region. Adenine (A) was most often mis-called for guanine (G) whereas thymine (T) was most often mis-called for cytosine (C)

from the plasmid as a control. We found that the mean coverage was much higher and more consistent across the control region as compared to the coverage in the repeat-containing region (Figure S2C). These results suggest an inability of a small fraction of nanopores to read all the way through the repeat expansion sequence from both sense and antisense strands. Despite technological progress, there continues to be base-calling errors within homo-polymeric sequence regions (Wick et al. 2019). Nonetheless, a strong consensus can clearly be observed, resulting in a perfect base-calling and sequence assignment of exactly 78 GGGGCC repeats as expected (Fig. 4b). Upon investigation of mis-called bases, we found that a total of 5.25% adenine (A) and 1.78% thymine (T) were mis-called within the mapped repeat region (Fig. 4c). Based on these results, much larger repeat expansions should be easily sequenced. To offset reduced coverage across very large repeat expansions, run time of the MinION sequencer can be increased. Thus, we recommend using MinION sequencing for accurate and unambiguous verification of cloned microsatellite/STR repeat expansions.

## Materials and methods

### General molecular cloning

Standard molecular cloning practices were employed unless otherwise noted. Enzyme manufacturer's recommended protocols were followed unless otherwise noted. Bacterial culture to propagate plasmid vectors generally used Stbl2 (Thermo Fisher Scientific) or NEB Stable (NEB) cells and Luria–Bertani (LB) agar (petri dish) and liquid LB culture growth at 30 °C. Standard commercial mini-prep or midi-prep kits (Omega Bio-Tek) were used for plasmid vector extraction and purification. Sanger sequencing was performed by MCLab using standard T7 promoter primers or custom sequencing primers.

### Site-directed mutagenesis (SDM) for removal of redundant BspQI and EcoO109I sites

The parent vector pTRE3G contained one redundant site for both BspQI and EcoO109I. Site-directed mutagenesis (SDM) was performed on pTRE3G with primer pairs designed to disrupt the redundant restriction sites outside of the MCS. *PfuTurbo* DNA polymerase and Quick-Change Site-Directed Mutagenesis Kit were used (Agilent). 100–200 ng of template along with 10 pmol primer pair were added to a 40 µl SDM PCR reaction containing *PfuTurbo* polymerase with the cycling conditions set according to the type of mutation desired and following the manufacturer's recommended protocol (Agilent). Following the reaction, the PCR product was treated with DpnI before purification by

agarose gel electrophoresis. Finally, the purified PCR product was transformed into 50 µl of DH5α competent cells. After transformation, plasmid was isolated from colonies grown on agar plates containing ampicillin and verified by Sanger sequencing.

### Recursive directional ligation (RDL) of GGGGCC and CCCC GG repeat expansions

In the first round of RDL cloning, 5 µg of synthetic repeat-containing insert was digested with NdeI (NEB) and NcoI (NEB) using the manufacturer's recommended protocols. At the same time, 2 µg of pINC3G parent plasmid was also subjected to NdeI and NcoI digestion, followed by dephosphorylation with FastAP (Thermo Fisher Scientific) also using the manufacturer's recommended protocol. Both reactions were phenol–chloroform extracted and precipitated. After resuspension in water, the digested oligonucleotide insert and pINC3G vector were ligated with T4 DNA ligase (Thermo Fisher Scientific) at room temperature for 1 h. A total of 20 ng of plasmid was used with three different vector:insert ratios of 1:1, 1:3 and 1:5. Some of the ligation reaction (5 µL) was then transformed into chemically competent DH5α cells (50 µL) by heat shock at 42 °C for 30 s followed by incubation on ice for 10 min. Cells were recovered after heat shock by mixing with 1 mL of SOC media and rotation at 37 °C for 1 h. Cells in media (100 µL) were then plated on LB-agar plates containing 100 µg/mL of carbenicillin (pINC3G harbors an ampicillin resistance gene) and incubated at 30 °C for 24 h. At least eight individual colonies were picked from the plate, cultured in 5 mL of LB at 30 °C for 24 h, then purified by mini-prep kit for validation via Sanger sequencing. Once sequence was confirmed, midi-preps were performed to generate larger stocks of plasmid. Sanger sequencing was repeated to ensure repeat size had not changed. This first round of cloning created pINC3G-G5.

In the second round of RDL, the same synthetic repeat-containing insert was digested and used again to generate a ten repeat-containing plasmid, pINC3G-G10. Digestion of the insert was carried out with both enzymes in CutSmart Buffer (NEB). The first digestion with BspQI was performed at 50 °C for 6 h followed by addition of EcoO109I and incubation at 37 °C for another 6 h. The order of addition for each enzyme does not matter in this case; digestion with EcoO109I can be performed first followed by BspQI. In all reactions, 5–10 units of each enzyme were used for every 1 µg of plasmid. In a second, separate reaction tube the pINC3G-G5 plasmid was digested with only BspQI to linearize the vector at the 5' (upstream) end of the repeat region. The vector was further treated with FastAP to dephosphorylate the free 5' ends of the vector and reduce self-ligation. Ligation of the insert was performed with 20 ng of linearized

plasmid and vector:insert ratio of 1:1, 1:3 and 1:5. Transformation, plasmid purification, and sequencing were all performed as described for the first round of cloning.

In additional rounds of cloning, the same protocol as described for the second round above was followed. However, from this point forward inserts from previous rounds were used. pINC3G-G10 (8 µg) was digested with both BspQI (NEB) and EcoO109I (NEB) to release the repeat-containing insert. In later rounds of cloning, to enable rapid double-digests, we alternatively used SapI (NEB) in place of BspQI. Both SapI and EcoO109I are incubated together at 37 °C in CutSmart Buffer overnight. As the repeat-containing insert increased in size, less vector was required to release sufficient insert for cloning. The reaction products from double-digestion were resolved on a 0.8% agarose gel and the ten repeat-containing DNA fragment was gel-purified using an E.Z.N.A Gel Extraction Kit (Omega Bio-Tek). The pINC3G-G10 was also linearized with SapI (instead of BspQI since the incubation temperature is lower) in a separate reaction as described above in second round cloning. Ligation, transformation, plasmid purification and sequencing were all performed as described above except that NEB Stable cells were used, stable outgrowth media (NEB) was used during transformation recovery, and all growth temperatures were reduced to room temperature (23 °C) for longer duration. After obtaining pINC3G-G20, Sanger sequencing became less reliable and we turned primarily to diagnostic restriction digestions and PCR across the repeat region to confirm successful cloning.

### Diagnostic restriction enzyme digestions for cloned GGGGCC repeats

For diagnostic restriction enzyme digestions, 500 ng of plasmid for several clones was digested using ten units of SapI and EcoO109I at 37 °C for overnight in CutSmart Buffer (NEB). Reactions were ethanol precipitated then resuspended in 10 µL of water and 2 µL of 4× loading dye (50% glycerol, 1× Tris–borate EDTA buffer, 0.1% orange G dye). Reaction products were then resolved on a 2.5% agarose gel for plasmids containing less than 20 repeats and a 1.5% agarose gel for plasmids containing more than 20 repeats. Agarose gels contained ethidium bromide and were imaged by UV illumination.

### PCR amplification across GGGGCC repeat expansion regions

PCR across repeat expansion regions in the pINC3G vectors was performed using PrimeSTAR GXL DNA Polymerase (Takara). 10 ng of plasmid DNA and a final primer concentration of 0.2 µM was used in the manufacturer's recommended short PCR protocol. PCR primers were designed

by standard methods to possess melting temperatures of approximately 60 °C. PCR thermal cycling conditions were as follows: initial denaturation at 98 °C for 2 min, 30 cycles at 98 °C for 10 s, 60 °C for 15 s, then 68 °C for 10 s, followed by a final extension at 68 °C for 5 min. PCR amplicons were resolved on a 1.5% agarose gel containing ethidium bromide to visualize the repeat-containing PCR products by UV illumination.

### Plasmid and library preparation for MinION sequencing of cloned GGGGCC repeat expansions

To prepare pINC3G-G78 plasmid for MinION sequencing, 5 µg was digested with PciI (NEB) in a 200 µL reaction at 37 °C for 4 h. Following digestion, 200 ng of plasmid was resolved on a 0.8% agarose gel to verify successful linearization. The recognition site for PciI is centered at base pair 2287 on the plasmid, resulting in flanking sequences of approximately equal length around the expected GGGGCC repeat expansion of pINC3G plasmids. The restriction enzyme digestion reaction was phenol–chloroform extracted, ethanol precipitated, and the DNA pellet resuspended in 0.5X Tris–EDTA buffer to a final concentration of 106 ng/µL.

Library preparation was performed using 1D Genomic DNA Ligation Sequencing Kit SQK-LSK109 (ONT) according to the manufacturer's protocol. T4 DNA Ligase from Thermo Fisher Scientific was substituted for Quick T4 DNA Ligase recommended in the protocol. Concentration and quality of the plasmid DNA were estimated by absorbance at 260 nm on a NanoDrop spectrophotometer (Thermo Fisher Scientific). DNA fragmentation and repair steps were eliminated to retain long reads and improve DNA yield, respectively. Briefly, 1 µg (51 µL) of linearized plasmid DNA was mixed with 1 µL DNA CS, 3 µL Ultra II End-prep enzyme mix, and 6 µL Ultra II End-prep reaction buffer (NEB). The reaction was incubated at 20 °C for 5 min followed by 65 °C for 5 min using a thermal cycler. The prepared DNA library was then purified using AMPure XP beads (Beckman Coulter) following the manufacturer's recommended protocol and quantified by absorbance at 260 nm on a NanoDrop. Adapter ligation was performed by mixing the reaction from the previous step with 25 µL of Ligation Buffer, 3.33 µL of T4 DNA Ligase HC (30 units/µL) (Thermo Fisher Scientific), and 5 µL of Adapter Mix. The reaction was incubated for 10 min at room temperature followed by purification with Ampure XP Beads and quantification by NanoDrop.

### MinION sequencing and data analysis

Sequencing was performed on a new R9.4 FLO-MIN106 Spot-On flow cell (ONT). Prior to sequencing, the number of active flow cells was determined by Platform QC in

MinKNOW (v3.4.8). The flow cell was primed using Flow Cell Priming Kit EXP-FLP002 (ONT) as described in the Nanopore Sequencing Kit SQK-LSK109 (ONT) protocol. Briefly, 800  $\mu$ L of priming buffer was added through the priming port and incubated for 5 min at room temperature. During the incubation period, the library was prepared for loading by mixing 37.5  $\mu$ L of Sequencing Buffer (SB), 25.5  $\mu$ L of Loading Beads (LB) and 12  $\mu$ L of DNA Library at the concentration of 50 fmol as specified in the manufacturer's protocol. To complete flow cell priming, 200  $\mu$ L of priming mix was loaded through the priming port with the SpotON port open. Immediately, 75  $\mu$ L of the prepared sample was added via the SpotON sample port in a dropwise fashion. Sequencing was performed using MinION MK1B (ONT) on MinKNOW software (v3.4.8) for 4 h. The FAST5 files were base called using Guppy (v3.2.4) with minimum q-score parameter set to 8. Post-sequencing quality control was determined by ToulligQC (v1.1). Reads flagged as passed were concatenated and assembled using Nanopipe with minimum query sequence length set to 400 bp and all other parameters left at default. Nanopipe uses LAST (v9.2.3) to assemble a consensus sequence based on input FAST5 files. The BAM and BAM index files were extracted and visualized on Tablet (v1.19.09.03). The consensus sequence was extracted from Nanopipe and aligned against the reference sequence using BLASTn. Statistics for total mapping and repeat-containing region were completed using Qualimap 2 (v2.2.1).

## Conclusions

Using recursive direction ligation methods, microsatellite repeat expansions can be systematically cloned with precision to sizes above the pathological threshold for most known repeat expansion disorders. This method can conceivably clone any microsatellite repeat expansion when simple design rules described here are followed. Validation of cloning can be achieved by more traditional methods described here, including Sanger sequencing, restriction digestion, and PCR. However, next-generation sequencing using the MinION nanopore sequencer enables unrivaled characterization of repeat expansion size and sequence in cloning vectors. Together, these methods should accelerate microsatellite repeat expansion studies to better understand disease mechanisms as well as identify and evaluate new therapeutic strategies.

**Acknowledgements** We thank T.R. Murphy of the Hamilton-Brehm laboratory (SIU) for assistance in MinION data analysis.

**Author contributions** KJR designed and executed RDL cloning experiments, performed data analyses, and wrote the manuscript. KNO designed and executed RDL cloning experiments, performed data

analyses, and assisted in manuscript preparation. AAP performed and analyzed MinION sequencing, developed the bioinformatics workflow, and assisted in manuscript preparation. MB executed RDL cloning experiments. AJH assisted with RDL cloning and PCR experiments. KTG designed, planned and supervised all experiments, interpreted results, and wrote the manuscript.

**Funding** This work was supported by an ALS Association grant to K.T.G., a Judith and Jean Pape Adams Charitable Foundation grant to K.T.G., and a Department of Defense ALSRP grant to K.T.G.

**Availability of data and material** The datasets generated and analyzed during this study will be made available on a public database to be determined or by request to the corresponding author.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no competing interests.

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

## References

- Al-Allaf FA, Tolmachov OE, Zambetti LP, Tchetchelnitski V, Mehmet H (2013) Remarkable stability of an instability-prone lentiviral vector plasmid in *Escherichia coli* Stbl3. 3 Biotech 3:61–70. <https://doi.org/10.1007/s13205-012-0070-8>
- Ameur A, Kloosterman WP, Hestand MS (2019) Single-molecule sequencing: towards clinical applications. Trends Biotechnol 37:72–85. <https://doi.org/10.1016/j.tibtech.2018.07.013>
- Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ (2018) Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Res. <https://doi.org/10.12688/f1000research.13980.1>
- Baron U, Bujard H (2000) Tet repressor-based system for regulated gene expression in eukaryotic cells: principles and advances. Methods Enzymol 327:401–421. [https://doi.org/10.1016/s0076-6879\(00\)27292-3](https://doi.org/10.1016/s0076-6879(00)27292-3)
- Batra R, Lee CW (2017) Mouse models of C9orf72 hexanucleotide repeat expansion in amyotrophic lateral sclerosis/frontotemporal dementia. Front Cell Neurosci 11:196. <https://doi.org/10.3389/fncel.2017.00196>
- Berens C, Hillen W (2003) Gene regulation by tetracyclines. Constraints of resistance regulation in bacteria shape TetR for application in eukaryotes. Eur J Biochem 270:3109–3121. <https://doi.org/10.1046/j.1432-1033.2003.03694.x>
- Brouwer JR, Willemsen R, Oostra BA (2009) Microsatellite repeat instability and neurological disease. BioEssays 31:71–83. <https://doi.org/10.1002/bies.080122>
- Chen IC et al (2009) Spinocerebellar ataxia type 8 larger triplet expansion histone modification and induces RNA foci. BMC Mol Biol 10:9. <https://doi.org/10.1186/1471-2199-10-9>
- Cheng W et al (2018) C9ORF72 GGGGCC repeat-associated non-AUG translation is upregulated by stress through eIF2alpha phosphorylation. Nat Commun 9:51. <https://doi.org/10.1038/s41467-017-02495-z>
- Cleary JD, Ranum LP (2017) New developments in RAN translation: insights from multiple diseases. Curr Opin Genet Dev 44:125–134. <https://doi.org/10.1016/j.gde.2017.03.006>



- Cleary JD, Pattamatta A, Ranum LPW (2018) Repeat associated non-ATG (RAN) translation. *J Biol Chem*. <https://doi.org/10.1074/jbc.R118.003237>
- Corchero JL, Villaverde A (1998) Plasmid maintenance in *Escherichia coli* recombinant cultures is dramatically, steadily, and specifically influenced by features of the encoded proteins. *Biotechnol Bioeng* 58:625–632
- de Haro M et al (2006) MBNL1 and CUGBP1 modify expanded CUG-induced toxicity in a *Drosophila* model of myotonic dystrophy type 1. *Hum Mol Genet* 15:2138–2145. <https://doi.org/10.1093/hmg/ddl137>
- DeJesus-Hernandez M et al (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* 72:245–256. <https://doi.org/10.1016/j.neuron.2011.09.011>
- Ebbert MTW et al (2018) Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: implications for clinical use and genetic discovery efforts in human disease. *Mol Neurodegener* 13:46. <https://doi.org/10.1186/s13024-018-0274-4>
- Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 5:435–445. <https://doi.org/10.1038/nrg1348>
- Engler C, Kandzia R, Marillonnet S (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* 3:e3647. <https://doi.org/10.1371/journal.pone.0003647>
- Engler C, Gruetzner R, Kandzia R, Marillonnet S (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type II restriction enzymes. *PLoS ONE* 4:e5553. <https://doi.org/10.1371/journal.pone.0005553>
- Evans-Galea MV, Hannan AJ, Carroddus N, Delatycki MB, Saffery R (2013) Epigenetic modifications in trinucleotide repeat diseases. *Trends Mol Med* 19:655–663. <https://doi.org/10.1016/j.molmed.2013.07.007>
- George S, Pankhurst L, Hubbard A, Votintseva A, Stoesser N, Sheppard AE, Mathers A, Norris R, Navickaite I, Eaton C, Iqbal Z, Crook DW, Phan HTT (2017) Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microb Genom* 3:e000118. <https://doi.org/10.1099/mgen.0.000118>
- Godiska R et al (2010) Linear plasmid vector for cloning of repetitive or unstable sequences in *Escherichia coli*. *Nucleic Acids Res* 38:e88. <https://doi.org/10.1093/nar/gkp1181>
- Gossen M, Bujard H (1992) Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc Natl Acad Sci USA* 89:5547–5551. <https://doi.org/10.1073/pnas.89.12.5547>
- Gossen M, Freundlieb S, Bender G, Muller G, Hillen W, Bujard H (1995) Transcriptional activation by tetracyclines in mammalian cells. *Science* 268:1766–1769. <https://doi.org/10.1126/science.7792603>
- Grabczyk E, Usdin K (1999) Generation of microgram quantities of trinucleotide repeat tracts of defined length, interspersed pattern, and orientation. *Anal Biochem* 267:241–243. <https://doi.org/10.1006/abio.1998.2962>
- Green KM et al (2017) RAN translation at C9orf72-associated repeat expansions is selectively enhanced by the integrated stress response. *Nat Commun* 8:2005. <https://doi.org/10.1038/s41467-017-02200-0>
- Green KM, Linsalata AE, Todd PK (2016) RAN translation—what makes it run? *Brain Res* 1647:30–42. <https://doi.org/10.1016/j.brainres.2016.04.003>
- Hardiman O, van den Berg LH, Kiernan MC (2011) Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nat Rev Neurol* 7:639–649. <https://doi.org/10.1038/nrneurol.2011.153>
- He F, Todd PK (2011) Epigenetics in nucleotide repeat expansion disorders. *Semin Neurol* 31:470–483. <https://doi.org/10.1055/s-0031-1299786>
- Hodges JR, Piguet O (2018) Progress and challenges in frontotemporal dementia research: a 20-year review. *J Alzheimers Dis* 62:1467–1480. <https://doi.org/10.3233/JAD-171087>
- Iacoangeli A et al (2019) C9orf72 intermediate expansions of 24–30 repeats are associated with ALS. *Acta Neuropathol Commun* 7:115. <https://doi.org/10.1186/s40478-019-0724-4>
- Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. *Nature* 316:76–79. <https://doi.org/10.1038/316076a0>
- Jiang SW, Trujillo MA, Eberhardt NL (1996) An efficient method for generation and subcloning of tandemly repeated DNA sequences with defined length, orientation and spacing. *Nucleic Acids Res* 24:3278–3279. <https://doi.org/10.1093/nar/24.16.3278>
- Labbadia J, Morimoto RI (2013) Huntington's disease: underlying molecular mechanisms and emerging concepts. *Trends Biochem Sci* 38:378–385. <https://doi.org/10.1016/j.tibs.2013.05.003>
- Laccone F, Maiwald R, Bingemann S (1999) A fast polymerase chain reaction-mediated strategy for introducing repeat expansions into CAG-repeat containing genes. *Hum Mutat* 13:497–502. [https://doi.org/10.1002/\(SICI\)1098-1004\(1999\)13:6%3c497::AID-HUMU10%3e3.0.CO;2-6](https://doi.org/10.1002/(SICI)1098-1004(1999)13:6%3c497::AID-HUMU10%3e3.0.CO;2-6)
- Liao HH (1991) Effect of temperature on the expression of wild-type and thermostable mutants of kanamycin nucleotidyltransferase in *Escherichia coli*. *Protein Expr Purif* 2:43–50
- Mamedov TG et al (2008) A fundamental study of the PCR amplification of GC-rich DNA templates. *Comput Biol Chem* 32:452–457. <https://doi.org/10.1016/j.compbiolchem.2008.07.021>
- Mankodi A et al (2000) Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. *Science* 289:1769–1773
- Matsuura T, Ashizawa T (2002) Polymerase chain reaction amplification of expanded ATTCT repeat in spinocerebellar ataxia type 10. *Ann Neurol* 51:271–272. <https://doi.org/10.1002/ana.10049>
- Meyer DE, Chilkoti A (2002) Genetically encoded synthesis of protein-based polymers with precisely specified molecular weight and sequence by recursive directional ligation: examples from the elastin-like polypeptide system. *Biomacromol* 3:357–367
- Miga KH, Eisenhart C, Kent WJ (2015) Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* 43:e133. <https://doi.org/10.1093/nar/gkv671>
- Mirkin SM (2007) Expandable DNA repeats and human disease. *Nature* 447:932–940. <https://doi.org/10.1038/nature05977>
- Mitchell JD, Borasio GD (2007) Amyotrophic lateral sclerosis. *Lancet* 369:2031–2041. [https://doi.org/10.1016/S0140-6736\(07\)60944-1](https://doi.org/10.1016/S0140-6736(07)60944-1)
- Mizielinska S et al (2014) C9orf72 repeat expansions cause neurodegeneration in *Drosophila* through arginine-rich proteins. *Science* 345:1192–1194. <https://doi.org/10.1126/science.1256800>
- Musso M, Boccardi R, Parodi S, Ravazzolo R, Ceccherini I (2006) Betaine, dimethyl sulfoxide, and 7-deaza-dGTP, a powerful mixture for amplification of GC-rich DNA sequences. *J Mol Diagn* 8:544–550. <https://doi.org/10.2353/jmoldx.2006.060058>
- Nguyen L, Cleary JD, Ranum LPW (2019) Repeat-associated non-ATG translation: molecular mechanisms and contribution to neurological disease. *Annu Rev Neurosci* 42:227–247. <https://doi.org/10.1146/annurev-neuro-070918-050405>
- Ohshima K, Kang S, Larson JE, Wells RD (1996) Cloning, characterization, and properties of seven triplet repeat DNA sequences. *J Biol Chem* 271:16773–16783. <https://doi.org/10.1074/jbc.271.28.16773>
- Ordway JM, Detloff PJ (1996) *in vitro* synthesis and cloning of long CAG repeats. *Biotechniques* 21(609–610):612. <https://doi.org/10.2144/96214bm08>
- Paulson H (2018) Repeat expansion diseases. *Handb Clin Neurol* 147:105–123. <https://doi.org/10.1016/B978-0-444-63233-3.00009-9>



- Pingoud A, Jeltsch A (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res* 29:3705–3727. <https://doi.org/10.1093/nar/29.18.3705>
- Pingoud A, Wilson GG, Wende W (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res* 42:7489–7527. <https://doi.org/10.1093/nar/gku447>
- Renton AE et al (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* 72:257–268. <https://doi.org/10.1016/j.neuron.2011.09.010>
- Roewer L (2013) DNA fingerprinting in forensics: past, present, future. *Investig Genet* 4:22. <https://doi.org/10.1186/2041-2223-4-22>
- Rohilla KJ, Gagnon KT (2017) RNA biology of disease-associated microsatellite repeat expansions. *Acta Neuropathol Commun* 5:63. <https://doi.org/10.1186/s40478-017-0468-y>
- Rowland LP, Shneider NA (2001) Amyotrophic lateral sclerosis. *N Engl J Med* 344:1688–1700. <https://doi.org/10.1056/NEJM200105313442207>
- Seznec H et al (2001) Mice transgenic for the human myotonic dystrophy region with expanded CTG repeats display muscular and brain abnormalities. *Hum Mol Genet* 10:2717–2726. <https://doi.org/10.1093/hmg/10.23.2717>
- Shoubridge C, Gecz J (2012) Polyalanine tract disorders and neurocognitive phenotypes. *Adv Exp Med Biol* 769:185–203
- Stueber D, Bujard H (1982) Transcription from efficient promoters can interfere with plasmid replication and diminish expression of plasmid specified genes. *EMBO J* 1:1399–1404
- Szybalski W, Kim SC, Hasan N, Podhajaska AJ (1991) Class-II restriction enzymes—a review. *Gene* 100:13–26. [https://doi.org/10.1016/0378-1119\(91\)90345-c](https://doi.org/10.1016/0378-1119(91)90345-c)
- Thys RG, Wang YH (2015) DNA replication dynamics of the GGG GCC repeat of the C9orf72. *Gene J Biol Chem* 290:28953–28962. <https://doi.org/10.1074/jbc.M115.660324>
- Verkerk AJ et al (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* 65:905–914. [https://doi.org/10.1016/0092-8674\(91\)90397-h](https://doi.org/10.1016/0092-8674(91)90397-h)
- Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 14:125–138. <https://doi.org/10.1038/nrg3373>
- Wen X et al (2014) Antisense proline-arginine RAN dipeptides linked to C9ORF72-ALS/FTD form toxic nuclear aggregates that initiate in vitro and in vivo neuronal death. *Neuron* 84:1213–1225. <https://doi.org/10.1016/j.neuron.2014.12.010>
- Wick RR, Judd LM, Holt KE (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* 20:129. <https://doi.org/10.1186/s13059-019-1727-y>
- Zarei S et al (2015) A comprehensive review of amyotrophic lateral sclerosis. *Surg Neurol Int* 6:171. <https://doi.org/10.4103/2152-7806.169561>
- Zhao XN, Usdin K (2015) The repeat expansion diseases: The dark side of DNA repair. *DNA Repair (Amst)* 32:96–105. <https://doi.org/10.1016/j.dnarep.2015.04.019>
- Zhao T, Hong Y, Li XJ, Li SH (2016) Subcellular clearance and accumulation of huntington disease protein: a mini-review. *Front Mol Neurosci* 9:27. <https://doi.org/10.3389/fnmol.2016.00027>
- Zu T et al (2011) Non-ATG-initiated translation directed by microsatellite expansions. *Proc Natl Acad Sci USA* 108:260–265. <https://doi.org/10.1073/pnas.1013343108>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# A

## *(GGGGCC)5 repeat insert (SapI/EarI)*

sense 5' NNNNNN GCTCTTC N | GGC CGGGGCCGGGGCCGGGGCCGGGGCCGG | GGC N GAAGAG NNNNNN 3'  
anti 3' NNNNNN CGAGAAG N CCG | GCCCGGCCCGGCCCGGCCCGGCCCGGCC CCG | N CTTCTC NNNNNN 5'  
SapI Repeat expansion EarI

## *(GGGGCC)5 repeat insert (AlwI/Bccl)*

sense 5' NNNNNN GGATC NNNN | G GGGCCGGGGCCGGGGCCGGGGCCGGGGCC | G NNNN GATGG NNNNNN 3'  
anti 3' NNNNNN CCTAG NNNN C | CCCGGCCCCGGCCCCGGCCCCGGCCCCGG C | NNNN CTACC NNNNNN 5'  
AlwI Repeat expansion Bccl

## *(CCCCGG)5 repeat insert (BciVI/Bmrl)*

sense 5' NNNNNN GTATCC NNNNN C | CCCGGCCCCGGCCCCGGCCCCGGCCCCGG C | NNNN CCCAGT NNNNNN 3'  
anti 3' NNNNNN CATAGG NNNNN | G GGGCCGGGGCCGGGGCCGGGGCCGGGGCC | G NNNN GGGTCA NNNNNN 5'  
BciVI Repeat expansion Bmrl

# B

## *(CAG)10 repeat insert (BsrDI/BtsCI)*

sense 5' NNNNNN GCAATG CA | GCAGCAGCAGCAGCAGCAGCAGCAGCAG CA | CATCC NNNNNN 3'  
anti 3' NNNNNN CGTTAC | GT CGTCGTCGTCGTCGTCGTCGTCGTCGTC | GT GTAGG NNNNNN 5'  
BsrDI Repeat expansion BtsCI

## *(CTG)10 repeat insert (BsrDI/BtsCI)*

sense 5' NNNNNN GCAATG CT | GCTGCTGCTGCTGCTGCTGCTGCTGCTG CT | CATCC NNNNNN 3'  
anti 3' NNNNNN CGTTAC | GA CGACGACGACGACGACGACGACGACGAC | GA GTAGG NNNNNN 5'  
BsrDI Repeat expansion BtsCI

## *(CGG)10 repeat insert (SapI/EarI)*

sense 5' NNNNNN GCTCTTC N | CGG CGGCGCGCGCGCGCGCGCGCGCG | CGG N GAAGAG NNNNNN 3'  
anti 3' NNNNNN CGAGAAG N GCC | GCCCGCGCGCGCGCGCGCGCGCGGCC GCC | N CTTCTC NNNNNN 5'  
SapI Repeat expansion EarI

## *(CCTG)10 repeat insert (BsrDI/BtsCI)*

sense 5' NNNNNN GCAATG CT | GCCTGCCTGCCTGCCTGCCTGCCTGCCTGCCTGC CT | CATCC NNNNNN 3'  
anti 3' NNNNNN CGTTAC | GA CGGACGGACGGACGGACGGACGGACGGACGGACGGACG | GA GTAGG NNNNNN 5'  
BsrDI Repeat expansion BtsCI

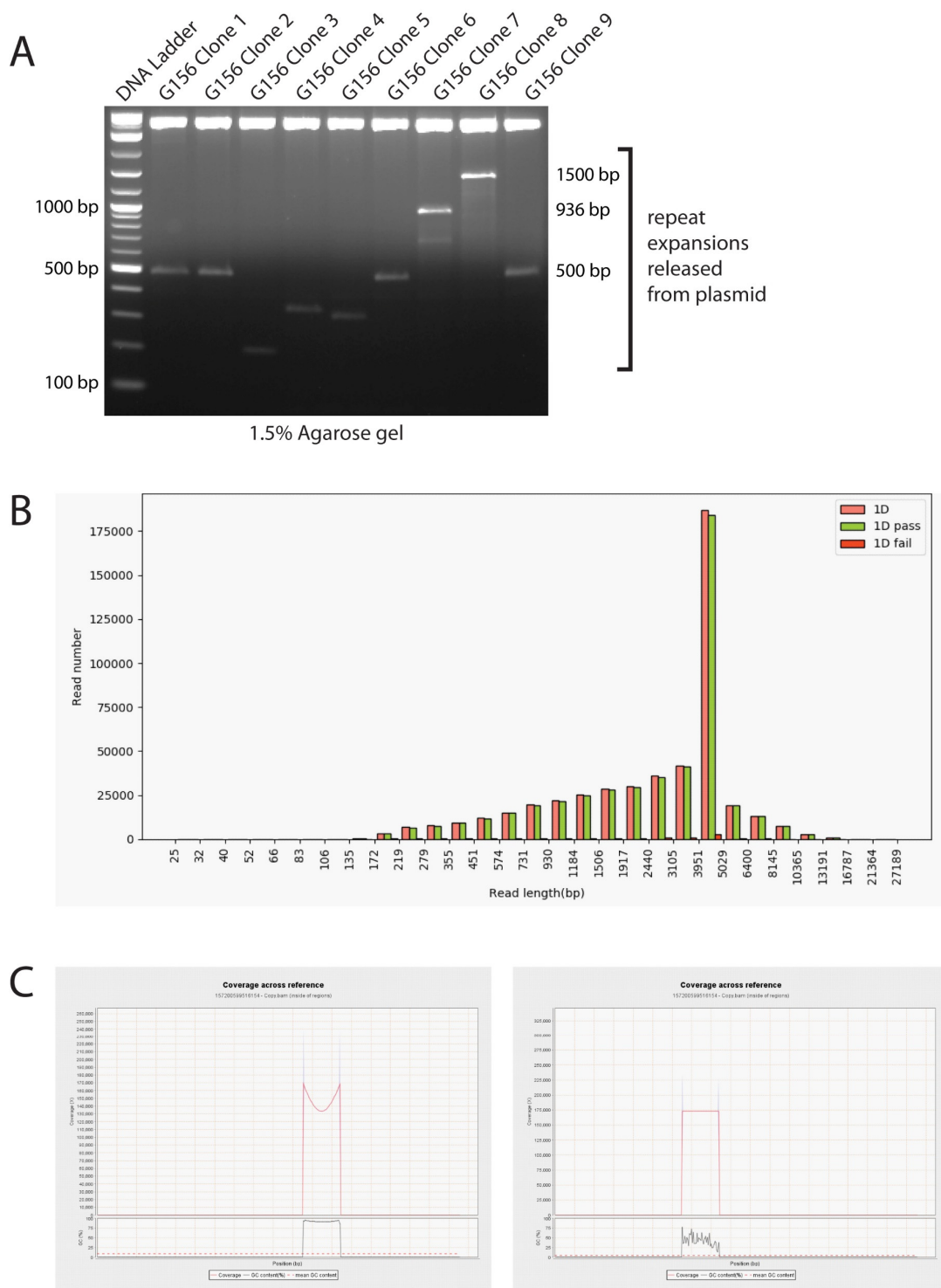
## *(ATTCT)6 repeat insert (BspMI/BbsI)*

sense 5' NNNNNN ACCTGC NNNN | ATTC TATTCTATTCTATTCTATTCTATTCT | ATTC NN GTCTTC NNNNNN 3'  
anti 3' NNNNNN CATAGG NNNN TAAG | ATAAGATAAGATAAGATAAGATAAGA TAAG | NN CAGAAG NNNNNN 5'  
BspMI Repeat expansion BbsI

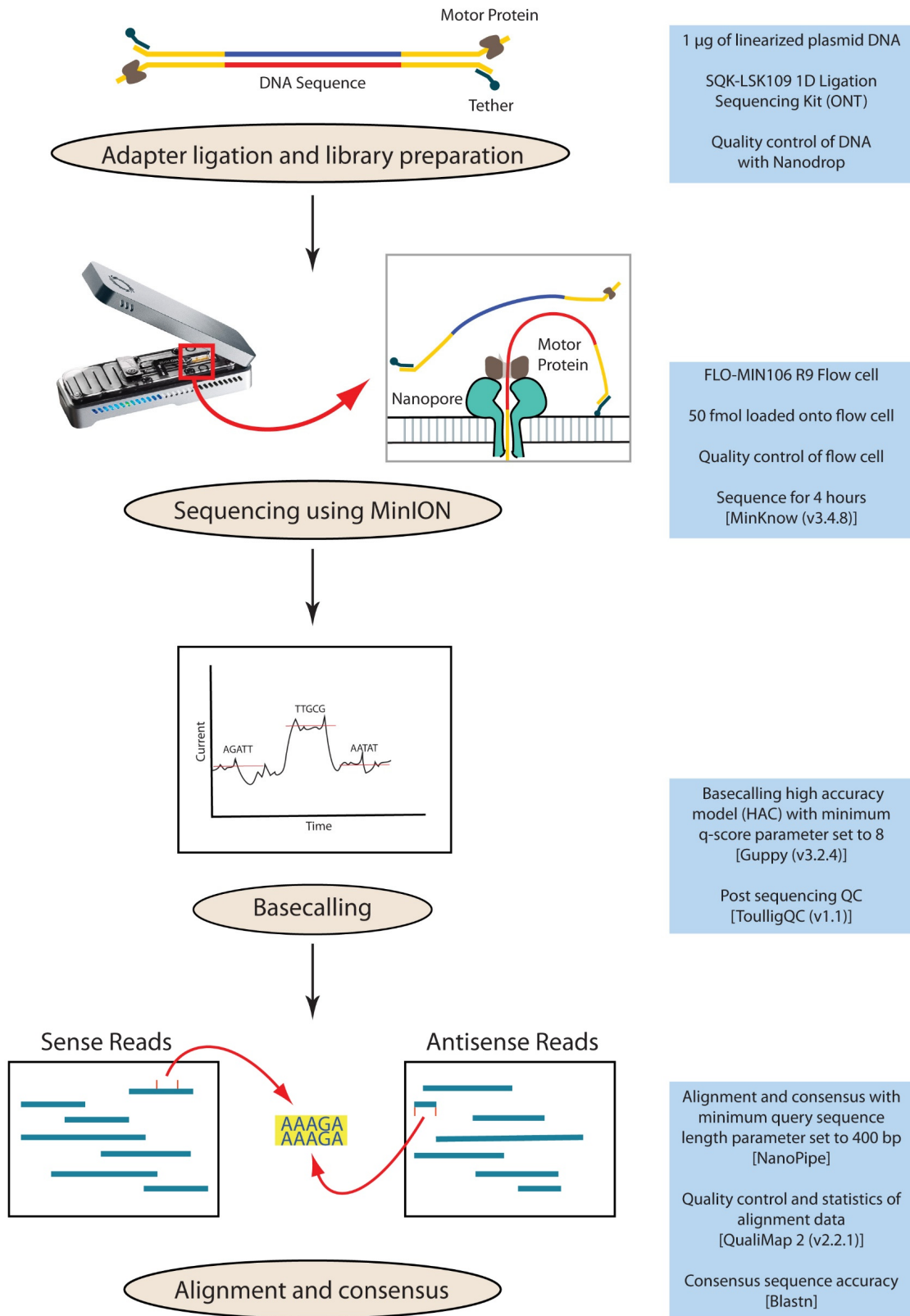
## *(TGGAA)6 repeat insert (BspMI/BbsI)*

sense 5' NNNNNN ACCTGC NNNN | TGGAA ATGGAATGGAATGGAATGGAATGGAATGGA | TGGAA NN GTCTTC NNNNNN 3'  
anti 3' NNNNNN CATAGG NNNN ACCT | TACCTTACCTTACCTTACCTTACCTT ACCT | NN CAGAAG NNNNNN 5'  
BspMI Repeat expansion BbsI

**Figure S1: Examples of synthetic inserts for cloning of several disease-associated microsatellite repeat expansions. (A)** Alternative Type IIS restriction enzymes for GGGGCC and CCCCCG repeat cloning. **(B)** Examples of potential inserts for initial cloning of disease-associated CAG, CTG, CGG, CCTG, ATTCT, or TGGAA repeat expansions. Selection of Type IIS enzyme pairs and cleavage within the repeat expansion should be selected to provide directionality.



**Figure S2:** (A) Expansion and contraction of the repeat region was often seen upon performing diagnostic restriction enzyme digestion on multiple clones expected to contain larger repeats in the range of 156 GGGGCC (doubling of the 78 GGGGCC repeats from pINC3G-G78). (B) MinION-sequenced pINC3G-G78 1D read-length histogram from 490,211 reads. Green bars represent passed reads with a q-score of 8 and above while dark red bars represent failed reads with a q-score below 8. The peak between 3,951–5,029 represents the expected pINC3G-G78 length and indicates plasmids that were sequenced in their entirety. (C) MinION read coverage across the repeat with pINC3G-G78 as reference (left panel). Coverage across a random non-repetitive location (of the same length as the repeat region) is also shown (right panel).



**Figure S3: Workflow and experimental and bioinformatics pipeline for MinION Nanopore sequencing.** This workflow and data analysis pipeline was used to sequence and characterize pINC3G-G78.