Long-read sequencing across the *C9orf72* 'GGGGCC' repeat expansion: implications for genetic discovery efforts in human disease

Mark T. W. Ebbert[1,2], Stefan Farrugia[1], Jonathon Sens[1], Karen Jansen-West[1], Tania F. Gendron[1,2], Mercedes Prudencio[1,2], Mariely DeJesus-Hernandez[1], Patricia H Brown[1], Dennis W. Dickson[1,2], Marka van Blitterswijk[1,2], Rosa Rademakers[1,2], Leonard Petrucelli[1,2], John D. Fryer[1,2]

[1]Department of Neuroscience, Mayo Clinic College of Medicine, Jacksonville, FL 32224; [2]Mayo Graduate School, Mayo Clinic, Rochester, MN 55905, USA

Corresponding authors:
Leonard Petrucelli
904.953.2855
petrucelli.leonard@mayo.edu

John D. Fryer
904.953.2317
fryer.john@mayo.edu

Abstract

Many neurodegenerative diseases are caused by nucleotide repeat expansions, but most expansions, like the *C9orf72* 'GGGGCC' (G_4C_2) repeat that causes approximately 5-7% of all amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) cases, are too long to fully sequence using short-read sequencing technologies. It is unclear whether long-read sequencing technologies can traverse these long, challenging repeat expansions. Here, we demonstrate that two long-read sequencing technologies, Pacific Biosciences' (PacBio) and Oxford Nanopore Technologies', can sequence through disease-causing repeats cloned into plasmids, including the FTD/ALS-causing G_4C_2 repeat expansion. We also report the first long-read sequencing data characterizing the *C9orf72* G_4C_2 repeat expansion at the nucleotide level in a symptomatic expansion carrier. Our findings indicate that long-read sequencing is well suited to characterizing known disease-causing repeat expansions, and for discovering new disease-causing or disease-modifying repeat expansions that have gone undetected with conventional short-read sequencing. Larger and deeper long-read sequencing studies in *C9orf72* expansion carriers will be important to determine heterogeneity and whether the repeats are interrupted by non-G_4C_2 content, potentially mitigating or modifying the disease course or age of onset, as interruptions are known to do in other repeat-expansion disorders. These results have broad implications across all diseases where the genetic etiology remains unclear.

Introduction

Many neurodegenerative diseases, including Huntington's disease (1), spinocerebellar ataxias (2, 3), frontotemporal dementia (FTD) (4), and amyotrophic lateral sclerosis (ALS) (4) are caused by nucleotide repeat expansions (2) that are historically challenging to sequence (5, 6). Repeat expansions are a specific multi-nucleotide DNA sequence that is repeated (i.e., expanded) significantly more times than normal. In 2011, a *C9orf72* 'GGGGCC' (G_4C_2) repeat expansion was discovered (4, 7) that causes approximately 34% and 26% of familial ALS and FTD cases, respectively (8). This finding genetically linked ALS and FTD, generating an exciting opportunity to better understand the etiology of both diseases, and potentially develop a therapeutic approach. Individuals with ALS and FTD caused by the G_4C_2 expansion generally have hundreds to thousands of G_4C_2 repeats (9), while healthy individuals typically have between 2 and 30 G_4C_2 repeats (7, 10), though a precise cutoff for pathogenicity is unclear (10). Additional diseases caused by repeat expansions include Fuch's disease (11), myotonic dystrophy (12), Friedreich's ataxia (13), and Fragile X syndrome (14), among others, demonstrating the breadth of diseases caused by such expansions. Revealing the underlying etiology of these diseases, and discovering additional repeat expansions causing disease will likely require long-read sequencing technologies capable of characterizing at least major portions of the repeat; characterizing these repeats at the nucleotide level will help determine, for example, whether the repeat is interrupted and whether such interruptions mitigate disease, as in other neurodegenerative disorders (15–18).

It is unclear whether third-generation long-read sequencing platforms such as Pacific Biosciences' (PacBio; RS II and Sequel) and Oxford Nanopore Technologies' (ONT; MinION) can traverse these challenging disease-causing repeats, nor is there a report of nucleotide-level sequencing data in a *C9orf72* repeat expansion carrier. Likewise, it is unclear whether the *C9orf72* repeat expansion is pure G_4C_2 repeat in affected carriers, or whether it is interrupted by non-G_4C_2 sequence. The *C9orf72* G_4C_2 expansion may be the most challenging

repeat to sequence, given its extreme length, "pure" GC content (5, 6), and propensity to form G-quadruplexes in both RNA (19, 20) and DNA (20).

Here, we demonstrate that both PacBio and ONT sequencing platforms can sequence through disease-causing repeats cloned into plasmids, including the spinocerebellar ataxia type 36 (SCA36) disease-causing 'GGCCTG' repeat expansion (21) and the FTD- and ALS-causing G_4C_2 repeat expansion, and we report long-read sequencing data from the *C9orf72* G_4C_2 repeat expansion at the nucleotide level in a symptomatic expansion carrier using the PacBio Sequel. Our findings indicate that long-read sequencing is well suited to characterizing known disease-causing repeat expansions and that this technology has great promise for accelerating future genetic discovery efforts across a broad range of diseases that may involve repeat expansions. Structural mutations, and repeat expansions specifically, are challenging for short-read technologies. Thus, long-read sequencing technologies are ideal for discovering new disease-causing or modifying repeat expansions that have escaped detection with conventional short-read sequencing.

Results

RS II and MinION sequence through repeats cloned into plasmids

To generally assess the PacBio and ONT sequencing platforms, we cloned the SCA36 'GGCCTG' (Figure 1b) and *C9orf72* G_4C_2 (Figures 1c and 1d) repeat expansions into plasmids and sequenced them on the RS II and MinION (Figure 2). We also included an EGFP-containing plasmid without a repeat expansion, as a control (Figure 1a). The SCA36 plasmid was included as a secondary control because it has $1/6^{th}$ lower GC content than the *C9orf72* G_4C_2 repeat; we anticipated that the *C9orf72* G_4C_2 may be too challenging for these technologies, as GC-rich sequencing has historically been challenging for any technology (5, 6). We aimed to construct a plasmid containing 62 'GGCCTG' repeats, and two plasmids containing approximately 423 (C9-423)

17 and 774 (C9-774) G_4C_2 repeats, respectively. Because these very long repeat sequences are unstable, most
18 bacterial colonies contained fewer than the targeted number of repeats (Supplemental figure 1).
19
20 Of the reads that aligned to the respective plasmid sequences, there were 46213, 67339, 9012, and 11535 RS
21 II reads for EGFP, SCA36, C9-423, and C9-774, respectively (Figure 3). Similarly, there were 26736, 39059,
22 8276, and 8720 MinION reads for the same plasmids, respectively. The read length distributions for MinION
23 were much tighter than those from the RS II, particularly for the C9-423 and C9-774 plasmids. In general, both
24 platforms had a single mode (i.e., peak) and had similar median read lengths (Figure 3), with the exception of
25 the RS II C9-774 read length distribution. RS II median read lengths for EGFP, SCA36, C9-423, and C9-774 were
26 5854, 5676, 6396, and 5714, respectively, while MinION median read lengths were 5879, 5660, 7024, and
27 7313, respectively (Figure 3). Expected maximum read lengths for each plasmid were 6080 (EGFP), 5984
28 (SCA36), 8813 (C9-423), and 9731 (C9-774) (Figures 1, 3). Approximately 30.2% (EGFP), 19.2% (SCA36), 6.9%
29 (C9-423), and 3.2% (C9-774) RS II reads were longer than the expected maximum length, where only
30 approximately 0.5%, 0.3%, 0.3%, and 0.1% MinION reads were longer than the expected maximum length.
31
32 We also compared how well the platforms sequenced specifically through the repeat regions, where we first
33 compared repeat length distributions. Both platforms produced highly similar distributions for all plasmids
34 (Figure 4), but the repeat lengths varied widely within each plasmid, as expected based on gel intensity curves
35 (Supplemental figure 1). The C9-423 repeat length distribution is more variable than even the C9-774, perhaps
36 because the C9-774 plasmid backbone is more tolerant of the repeat. The median repeat lengths are highly
37 similar for the RS II and MinION (Figure 4), where the median repeat lengths (measured by repeat number, not
38 bases) for the RS II were 35, 148, and 395 for SCA36, C9-423, and C9-774, respectively, while median repeat
39 lengths for the MinION were 37, 172, and 406, respectively. Repeat lengths for the SCA36 plasmid were
40 confirmed by Sanger sequencing, estimating approximately 37 repeats, before sequence traces became

41 indeterminate (Supplemental figure 2). We also assessed the percentage of reads that extend through the
42 repeats to assess whether the repeats' innate characteristics affected sequencing performance.
43 Approximately 95.9%, 66.8%, and 43.8% of RS II reads successfully extended through the repeat for SCA36, C9-
44 423, and C9-774, respectively, while 99.5%, 97.7%, and 83.5% of MinION reads extended through,
45 respectively.
46
47 Base calling accuracy through the C9-774 repeat region was dramatically different between the RS II and
48 MinION for the *C9orf72* repeat. After aligning the repeat region for individual reads to the expected repeat
49 sequence of the same length, using the global Needleman-Wunsch algorithm (22), the median RS II error rate
50 for C9-774 was 7.4%, while the median MinION error rate for C9-774 was 47.3%. The RS II consensus sequence
51 (Supplemental data 1) contained 774 repeats in the C9-774 plasmid, attaining approximately 99.8% accuracy.
52 The MinION consensus sequence (Supplemental data 2) also contained 774 repeats in the C9-774 plasmid, but
53 was only approximately 26.6% accurate, because many guanines and cytosines were erroneously called as
54 adenine. Thus, in the MinION consensus sequence, guanines and cytosines were represented as mixed
55 nucleotides in the consensus sequence (e.g., R representing G or A, and M representing C or A). Exactly 553
56 (71.4%) of the 774 MinION repeats were represented as either RRRRCM or RRRRMC.
57

58 PacBio Sequel successfully sequenced through the *C9orf72* repeat expansion in an affected
59 carrier

60 Sequencing repeat expansions cloned into plasmids demonstrated that both platforms are capable of
61 sequencing through these challenging repeats, but demonstrating on a human expansion carrier is essential to
62 determining whether long-read technologies can characterize the repeats in their entirety at the nucleotide
63 level, and also to determine whether the technologies are ideal for discovering new disease-causing or
64 disease-modifying repeat expansions. We identified and confirmed a *C9orf72* G_4C_2 repeat expansion carrier by

first performing fluorescent PCR fragment analysis using cerebellar tissue, showing the individual carried 8 repeats in the non-mutant allele (Figure 5a). We then determined the individual was an expansion carrier through repeat-primed PCR (23) (Figure 5b), and estimated the expansion size by Southern blot (Figure 5c). The Southern blot's primary intensity band indicates a repeat size of approximately 1933 repeats (13.9kb, including flanking sequence), though the entire smear ranged from approximately 9.3kb to 23kb.

We next purified high molecular-weight cerebellar DNA for the same *C9orf72* G_4C_2 repeat expansion carrier and generated approximately 7x median genome-wide coverage, and 8x coverage across the *C9orf72* repeat locus from five Pacbio Sequel SMRT cells. We sequenced the same sample on the ONT MinION, generating approximately 3x median coverage, and 2x across the *C9orf72* repeat locus from 15 flow cells. All MinION flow cells passed quality control before loading the library, with >1000 active pores.

Of the two MinION reads, neither covered an expanded allele, and the repeat region for only one of the reads could be clearly defined. We excluded the read for which we could not clearly define the repeat region. Where the human reference genome (hg38) contains three G_4C_2 repeats (Figure 6a), the read for which we could clearly define the repeat region had a total of 41 nucleotides within the repeat region (gain of 25 nucleotides compared to hg38); this equates to approximately seven total repeats (Supplemental figure 3b; Supplemental data 3). While hg38 contains three repeats, this does not accurately represent what is observed in the population, as the most common non-pathogenic allele has two repeats followed by eight repeats (4). An allele with three repeats was not observed in the population (4). The MinION's measurement of seven repeats closely resembles the eight repeats measured by our fluorescent PCR fragment analysis (Figure 5a). The MinION was only able to sequence the non-mutant allele with 15 flow cells in this study.

Sequel reads identify repeat expansion and characterize repeat length

To date, researchers have relied on Southern blots to measure an individual's repeat expansion size, but have been unable to characterize nucleotide content, which may have critical implications on disease etiology, age of onset, duration, and other critical clinical phenotypes. The Sequel reads enabled us to generally characterize repeat length and nucleotide content for this case's G_4C_2 repeat expansion.

Four of the eight Sequel reads capturing the *C9orf72* repeat locus were clearly expanded and four were not. Repeat lengths for the four reads capturing the wild-type (non-pathogenic) allele ranged from 46 to 50 nucleotides, where two measured exactly 48 nucleotides (gain of 30 compared to hg38), equating to eight total repeats (Supplemental figure 3c; Figure 6b); these results matched our fragment analysis (Figure 5a). Of the four reads capturing an expanded allele, three did not bridge the entire repeat region, where one captured 178 nucleotides in the repeat region (approximately 30 repeats; Figure 6c; Supplemental data 5), another captured 419 nucleotides (approximately 69 repeats; Figure 6c; Supplemental data 6), and the third captured 5471 nucleotides (approximately 912 repeats; Figure 6c; Supplemental data 7). It is possible that the read capturing 419 nucleotides bridged the repeat because the end of the read closely matches the sequence adjacent to the repeat region (Supplemental figure 3d). The fourth read, however, spanned the entire repeat with 7941 nucleotides (approximately 1324 repeats; Figure 6c; Supplemental figure 3e; Supplemental data 8), which falls easily within the Southern blot's range (Figure 5c). Mean GC content across the case's repeat region was 87.7% compared to 97.0% for the C9-774 plasmid repeat region.

Discussion

Here, we showed that both PacBio and ONT technologies can sequence through the SCA36 'GGCCTG' and the *C9orf72* G_4C_2 repeat expansions in relatively controlled repeats in plasmids, and that the PacBio can sequence through a human *C9orf72* repeat expansion, in its entirety, depending on length. These results demonstrate the potential these technologies offer in future structural mutation genetic discovery efforts—including those

involving challenging repeat expansions. For example, the *C9orf72* G_4C_2 repeat expansion could have been discovered years earlier if long-read sequencing technologies had been available. Through great effort using the best approaches available at the time, the G_4C_2 repeat was discovered in 2011 (4, 7), approximately five years after chromosome 9p was initially implicated in both ALS and FTD (24, 25).

We found that both platforms are fully capable of sequencing through challenging repeats like the SCA36 'GGCCTG' and *C9orf72* G_4C_2 repeat expansions when cloned into plasmids. It is unclear why the read length distributions for MinION were much tighter than those from the RS II, but it shows promise for future MinION applications. Additionally, while median read lengths were highly similar between the RS II and MinION, the MinION had a higher percentage of reads that extended all the way through the repeat regions for all three repeat plasmids, particularly the C9-423 and C9-774 plasmids.

Both the RS II and MinION correctly identified the maximal expected number of repeats in the C9-774 plasmid, based on their consensus sequences, but base calling accuracy in the repeat regions was dramatically higher for the RS II. The RS II attained approximately 99.8% consensus accuracy, while the MinION consensus sequence was only 26.6% accurate because of the mixed nucleotides in the consensus sequence. We believe this will be a relatively easy fix in the ONT base calling algorithms because it appears systematic based on the RRRRCM and RRRRMC repeats in the consensus sequences, which demonstrates the same base calling errors occur consistently.

The smear within the Southern blot for the symptomatic carrier, and those from previous studies (4, 9), demonstrates the heterogeneity (i.e., mosaicism) of this repeat expansion's size within just a small piece of tissue, and emphasizes the need for extensive long-read sequencing studies to characterize repeat expansions at the nucleotide level. We were unable to attain sufficient coverage using the MinION, but we attained 8x

coverage from the Sequel across the *C9orf72* repeat region, where four reads covered the individual's expected eight-repeat (unexpanded) allele, three reads that ended 30, 69, and 912 repeats into the expansion, respectively, and one read that fully spanned an expanded repeat region with 1324 repeats. Additional studies will be necessary to determine the maximum repeat size that these technologies can sequence. Our study used cerebellar tissue, which is known to generally harbor shorter expansions than other brain regions in the same patient (9, 24, 25). Additionally, the read spanning 1324 repeats is on the lower end of the Southern blot. A deeper sequencing study will clarify the upper limit.

There was an 8.5% difference in mean GC content between the C9-774 plasmid repeat region and the human case's, but a larger, deeper sequencing study will be necessary to determine whether the difference in GC content should be credited to sequencing error or whether the individual's repeat is interrupted by non-GC content. There were greater than 11000 C9-774 plasmid reads, which provided significantly more data to estimate GC content, so it is possible the reads from the human case were uncharacteristically low quality. On the other hand, GC content from the C9-774 plasmid reads (mean = 97.0%) should represent expected G/C to A/T sequencing error through this GC-rich repeat, suggesting that at least some of the human case's non-GC content may be legitimate.

Given that several studies have shown the repeat expansion is variable across tissues within a given patient (9, 25–27), we suggest that a large, deep long-read sequencing study across the *C9orf72* repeat is important to better understand how repeat content affects disease onset and progression. Repeat interruptions are known to mitigate disease in other neurodegenerative disorders (15–18). Fully characterizing the repeat at the nucleotide level may have critical implications on our understanding of disease etiology, development, duration, and on future therapy. A large, long-read sequencing study of affected *C9orf72* G_4C_2 repeat expansion carriers would also allow us to characterize mosaicism within individuals; there may be expansion

sub-species that explain the more aggressive forms of ALS and FTD, that are not measurable through traditional methods, such as Southern blotting.

Cost is a limiting factor for long-read sequencing technologies, making it impractical for extremely large studies or for diagnostic use. Because of these limitations, researcher have made great efforts to maximize the utility of short-read sequencing technologies, employing the large amount of short-read sequencing data already generated across nearly every disease currently studied. An excellent example is the effort to identify repeat expansions based on evidence in existing short-read data (28, 29). These efforts offer researchers and physicians that have already generated short-read data for patients or study participants the ability to determine whether an individual has a repeat expansion, but only if the repeat expansion and its location are already known. The approaches are also generally unable to estimate the repeat size. The limitations in these approaches reflect the limitations of short-read sequencing, because short reads cannot span the repeat. Long-read sequencing, while having a much higher error rate, overcomes these limitations, and may be more amenable to regular use in the future. For now, long-read sequencing may be ideal for small familial studies or for smaller studies intent on identifying repeat expansions that exist among a small cohort of cases. Researchers could then follow up with more cost-effective methods such as repeat-primed PCR or Southern blotting.

Knowing PacBio and ONT long-read sequencing technologies are fully capable of sequencing through challenging disease-causing repeats, such as the SCA36 'GGCCTG' and *C9orf72* 'GGGGCC' repeats, lays important ground work for future sequencing studies to understand the nucleotide-level nature of all repeat-expansion disorders. It also demonstrates that long-read sequencing technologies are ideal for future repeat expansion discovery efforts. Further utilizing these technologies in larger studies will be critical to properly

84 characterizing known repeats (e.g., *C9orf72*) and their allelic distributions (size and content) on the nucleotide
85 level to better understand how they contribute to disease.
86

87 # Materials and Methods

88 ## Study participants

89 Cerebellar samples included in this were obtained from the Mayo Clinic Brain Bank, following Mayo Clinic's IRB
90 protocols. The individual was male, and was diagnosed with FTD with an age of onset and duration of 69.96
91 and 2.4 years, respectively. Fluorescent and repeat-primed PCR (4), and our Southern blot (23) techniques
92 were previously described.
93

94 ## Repeat plasmids

95 To generate the G_4C_2 repeat 423 and 774 expression vectors, we used muscle or spleen DNA from an affected
96 *C9orf72* expansion carrier as a template in a nested PCR strategy. We used ThermalAce DNA Polymerase
97 (Invitrogen) to amplify the 66 G_4C_2 repeat region of a previously constructed G_4C_2 repeat plasmid, including
98 113 and 99 nucleotides of 5' and 3' flanking sequence, respectively. We then used these intermediate
99 plasmids to construct PCR products for the 423 and 774 sequence that were subsequently cloned into the
100 pAG3 and pcDNA6 expression vectors containing 3 different C-terminal tags in alternate frames. The *EGFP*
101 gene and SCA36 'GGCCTG' 66 repeat were cloned into pAAV expression vectors. The SCA36 clone was Sanger
102 sequenced to determine the number of repeats, but the others were too long for Sanger sequencing.
103

104 ## Sequencing

105 Plasmid sequencing libraries were prepared by first linearizing plasmids individually and then pooling in equal
106 concentration based on NanoDrop (ThermoFisher Scientific) measurements. Plasmids were linearized using

the restriction site that provided the most non-repeat sequence up and downstream of the repeat, itself. EGFP, SCA36, and C9-774 plasmids were linearized using the AvrII restriction enzyme, while C9-423 was linearized using MluI (Figure 1). DNA was then purified using Agencourt AMPure XP beads, per the manufacturer's recommended protocol, and split equally for MinION and RS II library preparation.

We used the MinION "1D Genomic DNA by ligation" kit and protocol (SQK-LSK108) for MinION library preparation and sequencing, skipping optional DNA fragmentation and DNA repair steps to maximize DNA size and quantity, respectively. Briefly, we began with end repair and dA-tailing using all recommended reagents and steps, which include mixing the DNA, Ultra II end-prep reaction buffer, Ultra II end-prep enzyme mix, and nuclease-free water, and incubating at 20°C for five minutes and 65°C for five minutes. DNA was then purified using Agencourt AMPure XP beads, and eluted in nuclease-free water. The 1D sequencing adapter was then ligated by mixing the DNA, 1D adapter mix (AMX1D), and Blunt/TA ligation master mix and incubating for 10 minutes. DNA was then purified again using the Agencourt AMPure XP beads, but washing with the ONT Adapter Bead Binding (ABB) buffer and eluting in ONT's elution buffer (ELB). Flow cells were primed and loaded per recommended procedure and sequenced using the 48-hour sequencing protocol in MinKNOW. We performed the recommended quality control run on all flow cells prior to priming and loading the library and only used flow cells that had >1000 active pores and did not have existing air bubbles in the Application-Specific Integrated Circuit (ASIC) upon opening.

Plasmid libraries for the RS II were generated following PacBio's protocol for the SMRTbell Template Prep Kit 1.0 (Part #100-259-100) and PacBio's "Procedure & Checklist—10 kb Template Preparation and Sequencing". Briefly, DNA damage repair, end repair, blunt end hairpin adapter ligation, and final exonuclease treatment were performed using 5µg of intact, non-sheared pooled plasmid DNA. AMPure PB magnetic beads (Pacific Biosciences) were used for all purification steps. Qualitative and quantitative analysis were performed using

Advanced Analytical Fragment Analyzer (AATI) and Qubit fluorometer with Quant-iT dsDNA BR Assay Kits (Invitrogen). SMRT bell templates were annealed to v2 sequencing primers then bound to DNA polymerase P6 following PacBio's protocol using the DNA/Polymerase Binding Kit P6 (part #100-356-300), as directed using Binding Calculator version 2.3.1.1. Polymerase-template complexes were purified per manufacturer's protocol (PacBio) using Pacific Biosciences Magbead Binding Kit (part #100-133-600) and setting up sample reaction as directed using Binding Calculator. Sequencing was carried out on the PacBio RS II (SMRT) sequencer, equipped with MagBead Station upgrade, using C4 DNA Sequencing Kit 2.0 (Part #: 100-216-400) reagents. The sample was loaded onto single SMRT cell v3 (Part #: 100-171-800) and the movie length was 360 min. Secondary Analysis was performed using Pacific Biosciences SMRT Portal, using SMRT Analysis System software (v2.3.0) for Sub-read filtering.

When sequencing the affected *C9orf72* repeat expansion carrier, we extracted DNA from the affected *C9orf72* repeat expansion carrier using the Agilent RecoverEase DNA Isolation Kit (Agilent Technologies) and a standard, previously-described isolation protocol [34](9). We followed the same MinION "1D Genomic DNA by ligation" kit and protocol used for the plasmids, using 15 total flow cells. For PacBio sequencing, we sequenced the DNA using the Sequel instead of the RS II because the Sequel provides greater throughput. We prepared the Sequel library per PacBio's recommended protocol "Procedure & Checklist—20kb Template Preparation Using BluePippin Size-Selection System", using the Megaruptor 2 (Diagenode, Denville, NJ, USA) for shearing and the Fragment Analyzer (Advanced Analytical, Ankeny, IA, USA) to size the DNA.

Analyses

Base calling and alignment

All MinION reads were base called using Albacore version 1.2.1 and PacBio reads were base called using SMRT Analysis version 2.3 for the RS II data and SMRT Link version 4.0 for the Sequel data. After base calling, plasmid

reads were separated by sequence unique to each plasmid backbone (Supplemental data 9) using BLAST version 2.6.0. Reads were first separated by backbone to avoid mixing reads from plasmids with the same repeat (e.g., C9-423 and C9-774 had the same repeat, but different backbones). EGFP and SCA36 were then separated by their unique sequences, the *EGFP* gene and the 'GGCCTG' repeat, respectively. Separated reads were then aligned to their respective reference sequences using graphmap version 0.5.1 (30) with default parameters. All reads from the affected *C9or72* expansion carrier were aligned to human reference genome hg38 using graphmap version 0.5.1 with default parameters. Reads across the *C9orf72* repeat expansion locus for the affected expansion carrier were hand curated, using the original graphmap alignment as a guide, where specific sequence "landmarks" were used to define the repeat region. The final hand-curated alignments are included in Supplemental figure 3.

Statistical analyses

Total read counts, reads extending through the repeat, and repeat lengths for each plasmid were calculated using pysam version 0.12, which is a wrapper for samtools (31). To calculate reads extending through the repeats, we required the read to have sequence within 100 nucleotides before and after the repeat region, based on its alignment (i.e., the read could not start or end within the repeat). Repeat lengths were calculated by extracting all bases that aligned within the reference sequence's repeat region, removing all gaps, and counting remaining nucleotides. Summary statistics, including median read lengths, and plots were generated in R version 3.4.1 (32) and ggplot2 version 2.2.1 (33). Coverage across the *C9orf72* repeat expansion carrier's genome was calculated using bedtools genomecov (34).

To calculate error rates for MinION and RS II in the C9-774 plasmid repeat region, we extracted individual repeat sequence from each read, as previously explained, and aligned them to the reference repeat sequence

78 equal to the measured repeat length of the given read. We used the global Needleman-Wunsch algorithm (22)
79 for these alignments, for the most accurate result.
80

81 Funding

82 This work was supported by the PhRMA Foundation [RSGTMT17 to M.E.]; the Ed and Ethel Moore Alzheimer's
83 Disease Research Program of Florida Department of Health [8AZ10 to M.E. and 6AZ06 to J.F.]; the National
84 Institutes of Health [NS094137 to J.F., AG047327 to J.F, AG049992 to J.F., R35NS097273 to L.P., R21NS084528
85 to L.P., P01NS084974 to L.P., P01NS099114 to L.P., R01NS088689 to L.P., R01NS093865 to L.P.]; Department
86 of Defense [ALSRP AL130125 to L.P.]; Mayo Clinic Foundation (L.P. and J.F.); Mayo Clinic Center for
87 Individualized Medicine (L.P. and J.F.); Amyotrophic Lateral Sclerosis Association (L.P., M.P.); Robert Packard
88 Center for ALS Research at Johns Hopkins (L.P.) Target ALS (L.P.); Association for Frontotemporal Degeneration
89 (L.P.); GHR Foundation (J.F.); and the Mayo Clinic Gerstner Family Career Development Award (J.F.).
90

91 Acknowledgements

92 We gratefully acknowledge contributions by the Brigham Young University DNA Sequencing Center (DNASC),
93 and the Mayo Clinic Molecular Biology Core who provided valuable input and services. We also recognize the
94 Mayo Clinic Research Computing Services who provided the requisite computational resources for all analyses
95 performed in this study.
96

97 Conflict of Interest Statement

98 All authors declare that they have no conflicts of interest.
99

100

References

1. MacDonald,M.E., Ambrose,C.M., Duyao,M.P., Myers,R.H., Lin,C., Srinidhi,L., Barnes,G., Taylor,S.A., James,M., Groot,N., *et al.* (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.

2. La Spada,A.R. and Taylor,J.P. (2010) Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.*, **11**, 247–258.

3. Orr,H.T., Chung,M., Banfi,S., Kwiatkowski,T.J., Servadio,A., Beaudet,A.L., McCall,A.E., Duvick,L.A., Ranum,L.P.W. and Zoghbi,H.Y. (1993) Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **4**, 221–226.

4. DeJesus-Hernandez,M., Mackenzie,I.R., Boeve,B.F., Boxer,A.L., Baker,M., Rutherford,N.J., Nicholson,A.M., Finch,N.A., Gilmer,H.F., Adamson,J., *et al.* (2011) Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron*, **72**, 245–256.

5. Kieleczawa,J. (2006) Fundamentals of Sequencing of Difficult Templates—An Overview. *J. Biomol. Tech. JBT*, **17**, 207–217.

6. Zhao,X., Haqqi,T. and Yadav,S.P. (2000) Sequencing telomeric DNA template with short tandem repeats using dye terminator cycle sequencing. *J. Biomol. Tech. JBT*, **11**, 111–121.

7. Renton,A.E., Majounie,E., Waite,A., Simón-Sánchez,J., Rollinson,S., Gibbs,J.R., Schymick,J.C., Laaksovirta,H., van Swieten,J.C., Myllykangas,L., *et al.* (2011) A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron*, **72**, 257–268.

8. van Blitterswijk,M., DeJesus-Hernandez,M. and Rademakers,R. (2012) How do C9ORF72 repeat expansions cause ALS and FTD: can we learn from other non-coding repeat expansion disorders? *Curr. Opin. Neurol.*, **25**, 689–700.

9. van Blitterswijk,M., DeJesus-Hernandez,M., Niemantsverdriet,E., Murray,M.E., Heckman,M.G., Diehl,N.N., Brown,P.H., Baker,M.C., Finch,N.A., Bauer,P.O., *et al.* (2013) Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol.*, **12**, 978–988.

10. Van Mossevelde,S., van der Zee,J., Cruts,M. and Van Broeckhoven,C. (2017) Relationship between C9orf72 repeat size and clinical phenotype. *Curr. Opin. Genet. Dev.*, **44**, 117–124.

11. Nakano,M., Okumura,N., Nakagawa,H., Koizumi,N., Ikeda,Y., Ueno,M., Yoshii,K., Adachi,H., Aleff,R.A., Butz,M.L., *et al.* (2015) Trinucleotide Repeat Expansion in the TCF4 Gene in Fuchs' Endothelial Corneal Dystrophy in Japanese. *Invest. Ophthalmol. Vis. Sci.*, **56**, 4865–4869.

12. Mahadevan,M., Tsilfidis,C., Sabourin,L., Shutler,G., Amemiya,C., Jansen,G., Neville,C., Narang,M., Barceló,J., O'Hoy,K., *et al.* (1992) Myotonic Dystrophy Mutation: An Unstable CTG Repeat in the 3′ Untranslated Region of the Gene. *Science*, **255**, 1253–1255.

13. Campuzano,V., Montermini,L., Moltò,M.D., Pianese,L., Cossée,M., Cavalcanti,F., Monros,E., Rodius,F., Duclos,F., Monticelli,A., *et al.* (1996) Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, **271**, 1423–1427.

14. Verkerk,A.J.M.H., Pieretti,M., Sutcliffe,J.S., Fu,Y.-H., Kuhl,D.P.A., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F., *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.

15. Kraus-Perrotta,C. and Lagalwar,S. (2016) Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias*, **3**.

16. Matsuura,T., Fang,P., Pearson,C.E., Jayakar,P., Ashizawa,T., Roa,B.B. and Nelson,D.L. (2006) Interruptions in the Expanded ATTCT Repeat of Spinocerebellar Ataxia Type 10: Repeat Purity as a Disease Modifier? *Am. J. Hum. Genet.*, **78**, 125–129.

17. Sakamoto,N., Larson,J.E., Iyer,R.R., Montermini,L., Pandolfo,M. and Wells,R.D. (2001) GGA*TCC-interrupted triplets in long GAA*TTC repeats inhibit the formation of triplex and sticky DNA structures, alleviate transcription inhibition, and reduce genetic instabilities. *J. Biol. Chem.*, **276**, 27178–27187.

18. Stolle,C.A., Frackelton,E.C., McCallum,J., Farmer,J.M., Tsou,A., Wilson,R.B. and Lynch,D.R. (2008) Novel, complex interruptions of the GAA repeat in small, expanded alleles of two affected siblings with late-onset Friedreich ataxia. *Mov. Disord. Off. J. Mov. Disord. Soc.*, **23**, 1303–1306.

19. Fratta,P., Mizielinska,S., Nicoll,A.J., Zloh,M., Fisher,E.M.C., Parkinson,G. and Isaacs,A.M. (2012) *C9orf72* hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. *Sci. Rep.*, **2**, srep01016.

20. Haeusler,A.R., Donnelly,C.J., Periz,G., Simko,E.A.J., Shaw,P.G., Kim,M.-S., Maragakis,N.J., Troncoso,J.C., Pandey,A., Sattler,R., *et al.* (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature*, **507**, 195–200.

21. Kobayashi,H., Abe,K., Matsuura,T., Ikeda,Y., Hitomi,T., Akechi,Y., Habu,T., Liu,W., Okuda,H. and Koizumi,A. (2011) Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.*, **89**, 121–130.

22. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

23. van Blitterswijk,M., Baker,M.C., DeJesus-Hernandez,M., Ghidoni,R., Benussi,L., Finger,E., Hsiung,G.-Y.R., Kelley,B.J., Murray,M.E., Rutherford,N.J., *et al.* (2013) C9ORF72 repeat expansions in cases with previously identified pathogenic mutations. *Neurology*, **81**, 1332–1341.

24. Waite,A.J., Bäumer,D., East,S., Neal,J., Morris,H.R., Ansorge,O. and Blake,D.J. (2014) Reduced C9orf72 protein levels in frontal cortex of amyotrophic lateral sclerosis and frontotemporal degeneration brain with the C9ORF72 hexanucleotide repeat expansion. *Neurobiol. Aging*, **35**, 1779.e5-1779.e13.

25. Suh,E., Lee,E.B., Neal,D., Wood,E.M., Toledo,J.B., Rennert,L., Irwin,D.J., McMillan,C.T., Krock,B., Elman,L.B., *et al.* (2015) Semi-automated quantification of C9orf72 expansion size reveals inverse correlation

between hexanucleotide repeat number and disease duration in frontotemporal degeneration. *Acta Neuropathol. (Berl.)*, **130**, 363–372.

26. Dols-Icardo,O., García-Redondo,A., Rojas-García,R., Sánchez-Valle,R., Noguera,A., Gómez-Tortosa,E., Pastor,P., Hernández,I., Esteban-Pérez,J., Suárez-Calvet,M., *et al.* (2014) Characterization of the repeat expansion size in C9orf72 in amyotrophic lateral sclerosis and frontotemporal dementia. *Hum. Mol. Genet.*, **23**, 749–754.

27. Nordin,A., Akimoto,C., Wuolikainen,A., Alstermark,H., Jonsson,P., Birve,A., Marklund,S.L., Graffmo,K.S., Forsberg,K., Brännström,T., *et al.* (2015) Extensive size variability of the GGGGCC expansion in C9orf72 in both neuronal and non-neuronal tissues in 18 patients with ALS or FTD. *Hum. Mol. Genet.*, **24**, 3133–3142.

28. Dolzhenko,E., Vugt,J.J.F.A. van, Shaw,R.J., Bekritsky,M.A., Blitterswijk,M. van, Narzisi,G., Ajay,S.S., Rajan,V., Lajoie,B., Johnson,N.H., *et al.* (2017) Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, 10.1101/gr.225672.117.

29. Dashnow,H., Lek,M., Phipson,B., Halman,A., Davis,M., Lamont,P., Laing,N., MacArthur,D. and Oshlack,A. (2017) STRetch: detecting and discovering pathogenic short tandem repeats expansions. *bioRxiv*, 10.1101/159228.

30. Sović,I., Šikić,M., Wilm,A., Fenlon,S.N., Chen,S. and Nagarajan,N. (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.*, **7**, ncomms11307.

31. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.

32. R Development Core Team (2011) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

33. ggplot2 - Elegant Graphics for Data Analysis | Hadley Wickham | Springer.

34. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Figures

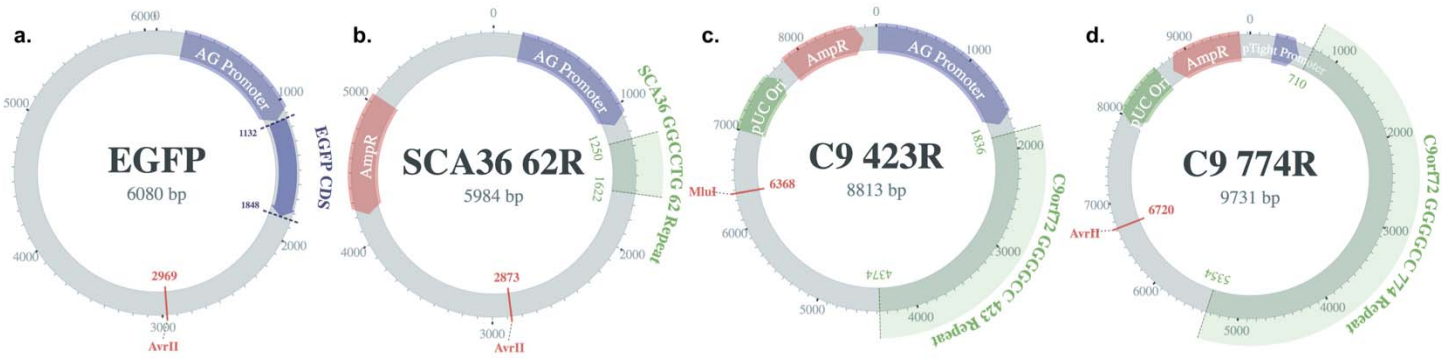


Figure 1. Schematic diagrams for plasmids used to test PacBio and ONT long-read sequencing technologies. To minimize biases when comparing the RS II and MinION, we constructed four plasmids, including three repeat-containing plasmids and a non-repeat-containing plasmid. Each plasmid map identifies estimated plasmid size, and the location and size of the repeat within the plasmid. **(a)** The first plasmid did not contain a repeat, as a control, but instead included the *EGFP* gene. The *EGFP* plasmid was linearized at position 2969 with the AvrII restriction enzyme. **(b)** We also constructed a plasmid with 62 repeats of the spinocerebellar ataxia type 36 (SCA36) 'GGCCTG' repeat, which was linearized at position 2873 with AvrII. **(c)** A third plasmid contained 423 *C9orf72* 'GGGGCC' repeats, and was linearized at position 6368 with MluI to maximize non-repeat sequence both up and downstream of the plasmid, thus avoiding bias against reads in either direction; allowing the repeat to be too close to either end could compromise sequencing or downstream analyses. **(d)** We included an additional plasmid with 774 *C9orf72* 'GGGGCC' repeats to simulate the expansion size found in ALS- or FTD-affected expansion carriers. While 774 repeats is dramatically smaller than the expansion found in many affected carriers, it was the largest we were able to construct reliably, because these repeats are unstable in bacteria. Additionally, while we targeted the number of specified repeats for each plasmid, most colonies contained fewer than the targeted repeats because repeats are generally unstable in bacteria (Supplementary Figure 1). Thus, the targeted number of repeats serves as an estimated maximum number of repeats.

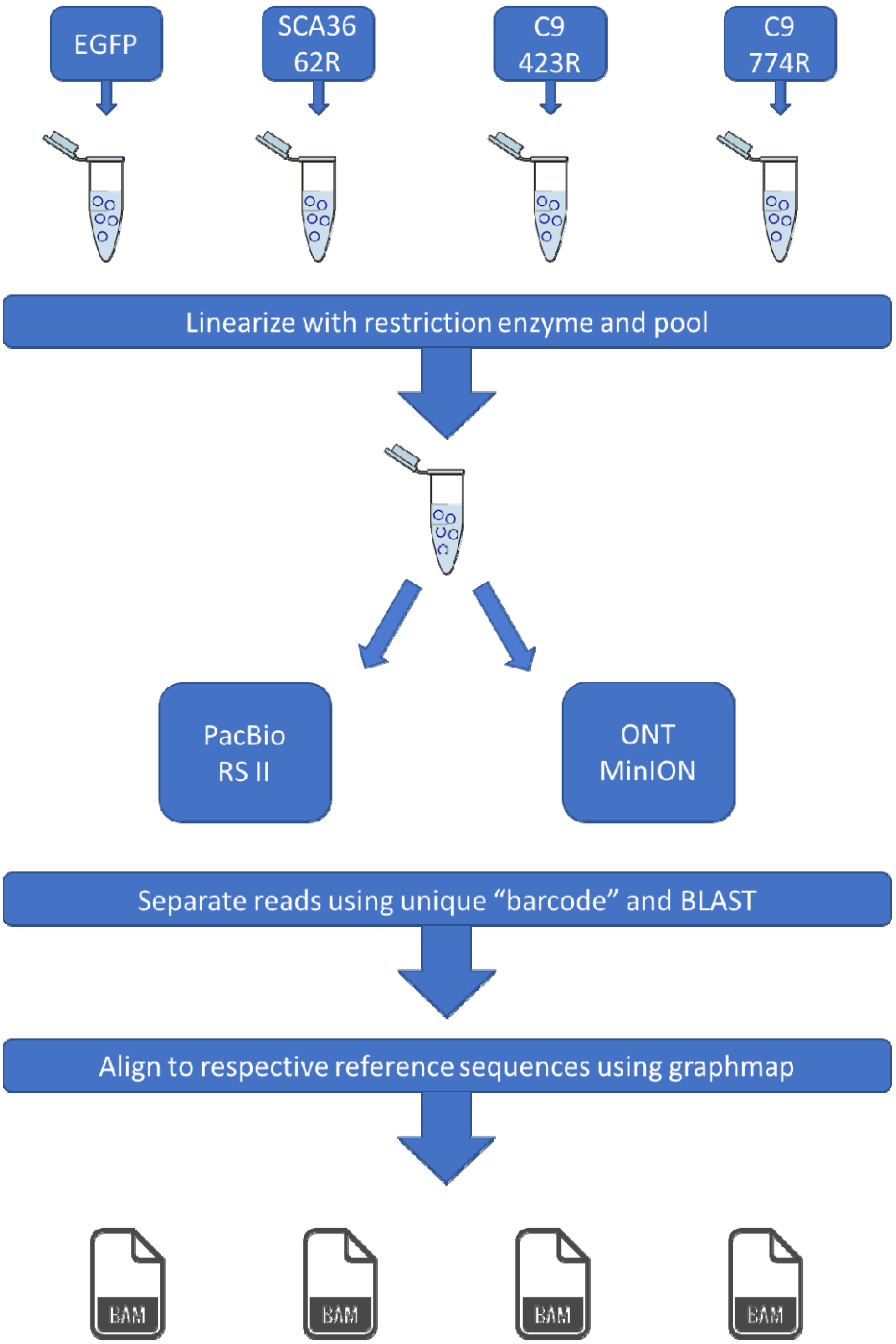


Figure 2. Workflow for linearizing, pooling, and sequencing plasmids on the PacBio RS II and ONT MinION long-read platforms. Each plasmid was cut with the restriction enzyme identified in the respective plasmid maps (Figure 1), and at the specified location. After linearizing each plasmid independently, the plasmids were pooled and cleaned. We then sequenced the same pool on the PacBio RS II and Oxford Nanopore Technologies' (ONT) MinION using their respective library preparation protocols. After sequencing, reads from each plasmid were identified using BLAST and then aligned to their respective reference sequences using graphmap, as preparation for downstream comparisons.

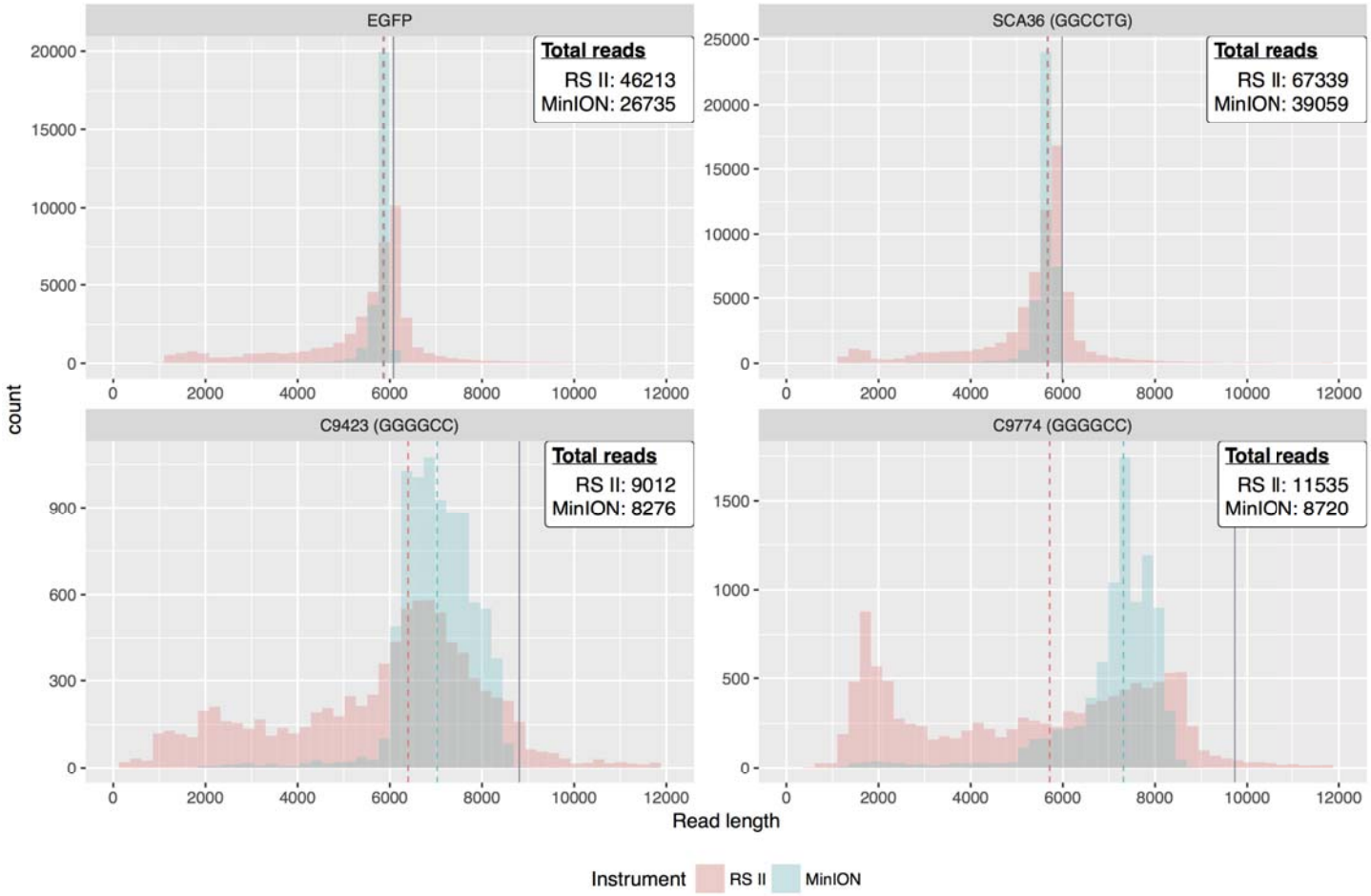
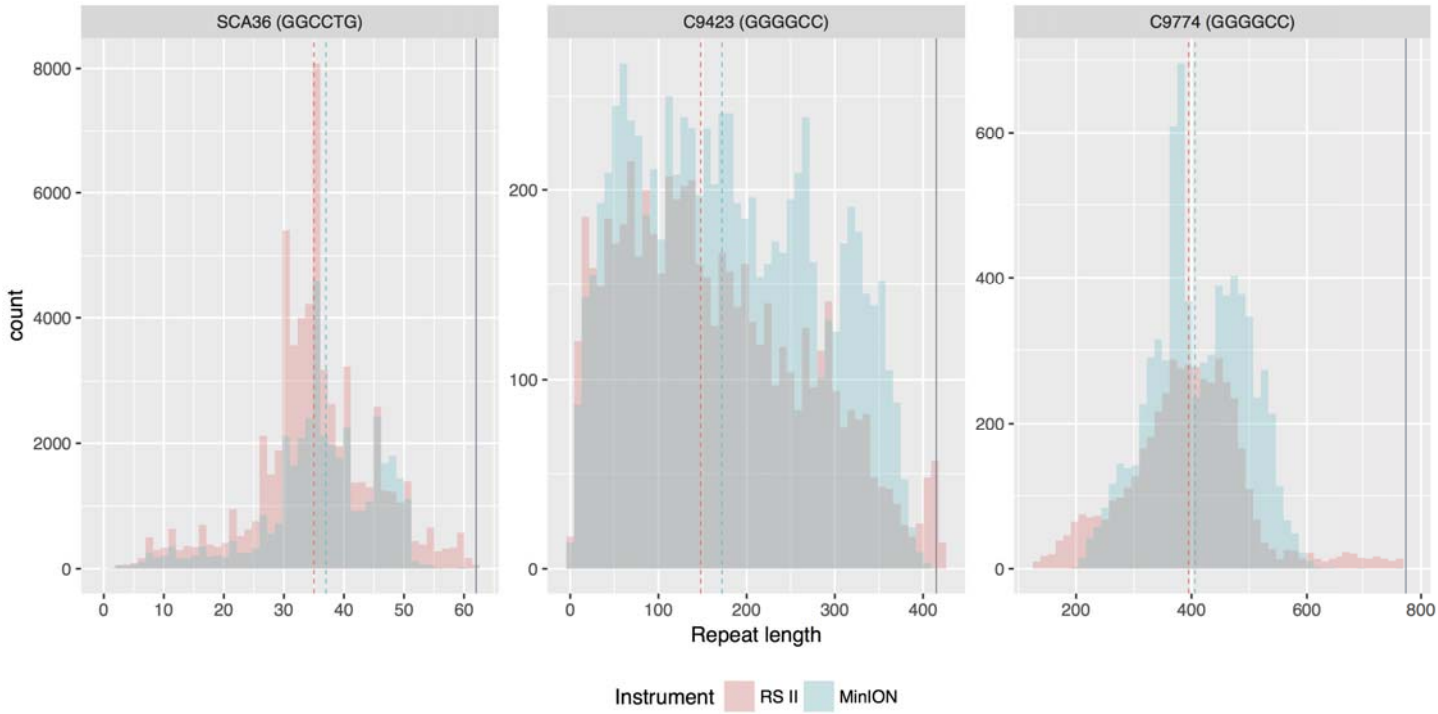


Figure 3. Both the RS II and MinION successfully sequence through repeats, but the RS II had more variable read lengths. After selecting only those reads that could be clearly identified for each plasmid (described in Figure 1), there were 46213, 67339, 9012, and 11535 RS II reads for *EGFP*, SCA36, C9-423, and C9-774, respectively. Likewise, there were 26735, 39059, 8276, and 8720 MinION reads for the same respective plasmids. The RS II generally had more reads, but read length distributions are much tighter for the MinION across all four plasmids, and more closely resemble expected read lengths. The median read length for each instrument is indicated by dashed lines, and the expected maximum read length is indicated by a solid gray line. Expected maximum read lengths for each plasmid were 6080 (*EGFP*), 5984 (SCA36), 8813 (C9-423), and 9731 (C9-774). RS II read variability may be related to library preparation.



46

47 **Figure 4. Repeat length distributions for the RS II and MinION were highly concordant.** Although read length

48 distributions for the RS II and MinION were different, both technologies obtained highly concordant repeat

49 length distributions, suggesting both technologies perform similarly when traversing the repeats. The C9-423

50 repeat length distribution is much more variable than even the C9-774, which may be because the C9-774

51 plasmid backbone is more tolerant of the repeat.

52

53

54

55

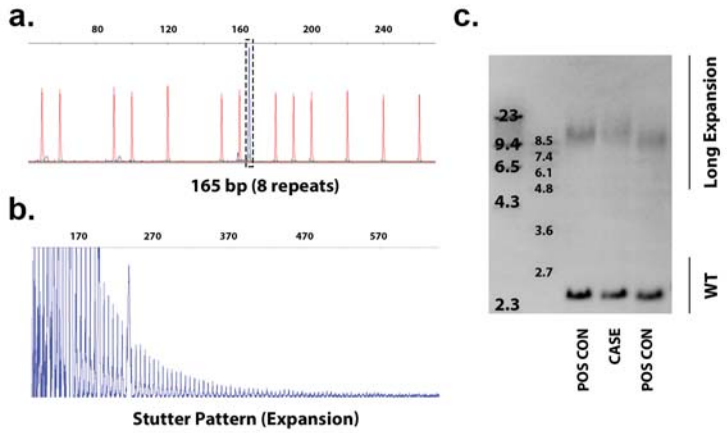


Figure 5. Characterization of the affected *C9orf72* repeat expansion carrier using standard methodologies. **(a)** We first performed fluorescent PCR to determine the individual's non-pathogenic repeat size. Genomic DNA is PCR-amplified with genotyping primers and one fluorescently labeled primer. Fragment length analysis of the PCR product is then performed on an ABI3730 DNA analyzer and visualized using GeneMapper software. A peak is seen at 165 bp, indicating that the non-pathogenic allele contains eight repeats. A single peak also indicates that the individual is either homozygous for the eight-repeat allele, or has an expansion. **(b)** To determine whether the individual had a repeat expansion, we performed a repeat-primed PCR analysis. PCR products of a repeat-primed PCR are separated on an ABI3730 DNA analyzer and visualized by GeneMapper software, showing a stutter amplification characteristic for a *C9orf72* repeat expansion. This does not indicate expansion size, however. **(c)** After determining the individual was an expansion carrier, we performed a Southern blot to estimate the size. The Southern blot reveals a long repeat expansion in three individuals for whom cerebellar tissue was available, including two positive controls (POS CON; lane 1 and 3) and our patient of interest (CASE; lane 2). DIG-labeled DNA Molecular Weight Markers (Roche) are shown to estimate the repeat expansion's size. Measurements were based on four separate Southern blots; for simplicity one representative Southern blot is shown. The most abundant expansion size in our case is estimated around 1933 repeats (13.9 kb). The smear ranged from approximately 9.3kb to 23kb, demonstrating the heterogeneity (i.e., mosaicism) of this repeat expansion within a small tissue chunk. This demonstrates the importance of additional long-read sequencing studies to characterize the repeat at the nucleotide level.

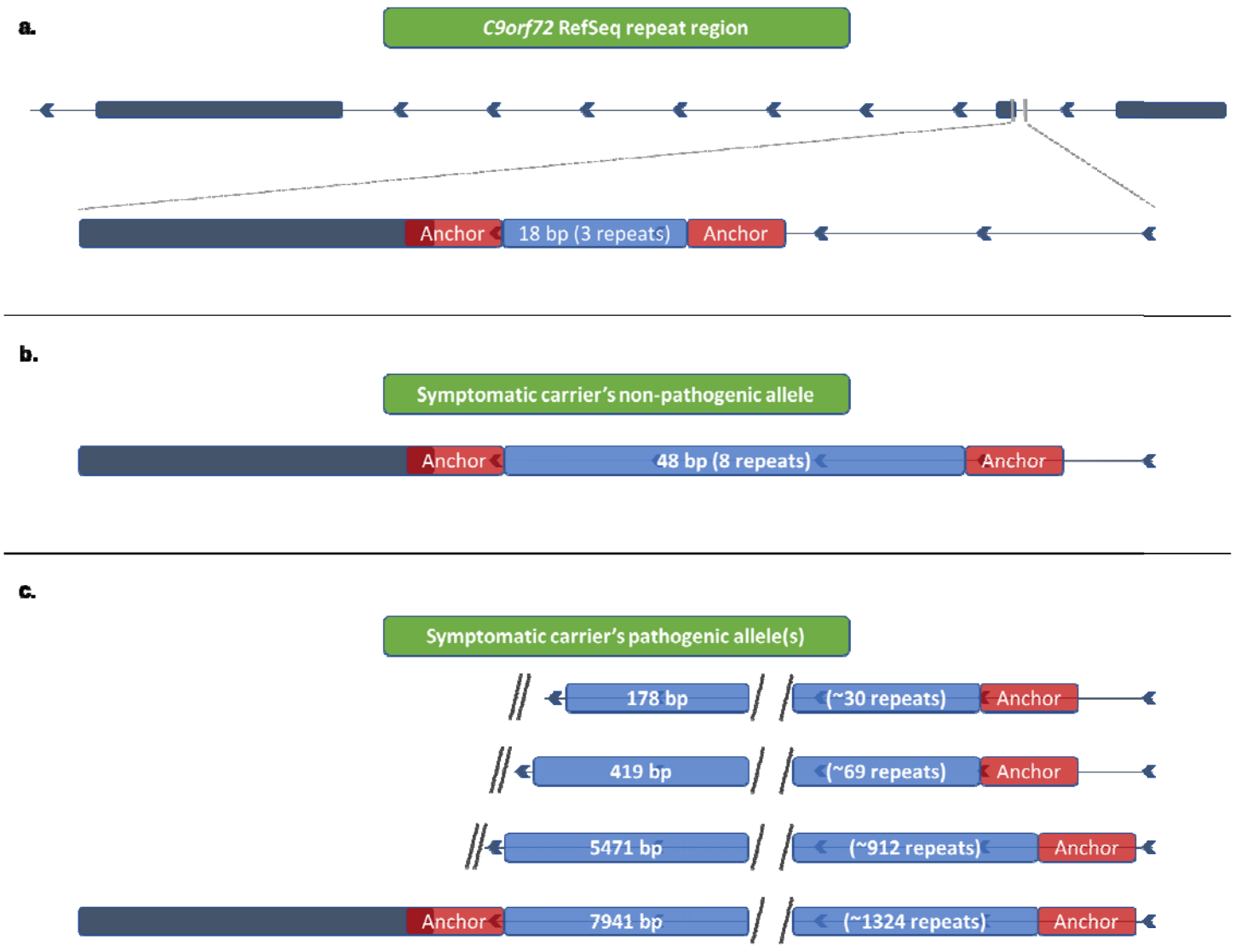


Figure 6. PacBio Sequel reads traverse the repeat region for pathogenic and non-pathogenic alleles. The PacBio Sequel sequenced through both pathogenic and non-pathogenic alleles, demonstrating the platform is capable of characterizing repeat expansions. All of these reads were first aligned by graphmap, and then hand curated to determine the repeat region. **(a)** The human genome reference sequence (hg38) contains three G_4C_2 repeats (18 nucleotides). We identified specific "landmarks" before and after the repeat region in the reference sequence to properly locate the repeat region in the reads, and to hand curate the alignments. Landmarks are identified by red bars adjacent to the repeat region. **(b)** We obtained four Sequel reads covering the eight-repeat sequence, spanning 48 nucleotides. There was a net gain of 29 nucleotides within the defined repeat region, which equates to approximately 5 additional repeats; this concurs with our

fragment analysis (Figure 5a). **(c)** We also obtained four reads that covered an expanded allele, one of which bridged the entire repeat expansion, with approximately 1324 repeats (7941 nucleotides). The other three reads ended before bridging the repeat region, where one captured approximately 30 repeats (178 nucleotides), another captured approximately 69 repeats (419 nucleotides), and the third captured approximately 912 repeats (5471 nucleotides).