Detection of long repeat expansions from PCR-free whole-genome sequence data

Egor Dolzhenko[1,#], Joke J.F.A. van Vugt[2,#], Richard J. Shaw[3,4], Mitchell A. Bekritsky[3], Marka van Blitterswijk[5], Zoya Kingsbury[3], Sean J. Humphray[3], Raymond D. Schellevis[2], William J. Brands[2], Matt Baker[5], Rosa Rademakers[5], Maarten Kooyman[6], Gijs H.P. Tazelaar[2], Michael A. van Es[2], Russell McLaughlin[7,8], William Sproviero[9], Aleksey Shatunov[9], Ashley Jones[9], Ahmad Al Khleifat[9], Alan Pittman[10], Sarah Morgan[10], Orla Hardiman[7,8], Ammar Al-Chalabi[9], Leonard H. van den Berg[2], David R. Bentley[3], Michael A. Eberle[1,#,*] and Jan H. Veldink[2,#,*]

[1]Illumina Inc., 5200 Illumina Way, San Diego, CA, USA
[2]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands.
[3]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, UK
[4]Repositive Ltd., Future Business Centre, Kings Hedges Rd, Cambridge, UK
[5]Department of Neuroscience, Mayo Clinic, Jacksonville, FL, USA
[6]SURFsara, Science Park 140, Amsterdam, The Netherlands
[7]Academic Unit of Neurology, Trinity College Dublin, Trinity Biomedical Sciences Institute, Dublin, Republic of Ireland.
[8]Department of Neurology, Beaumont Hospital, Dublin, Republic of Ireland
[9]Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[10]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK

[#]These authors contributed equally
[*]To who correspondence should be addressed

Abstract

Accurately identifying large repeat expansions including those that cause amyotrophic lateral sclerosis (ALS) and Fragile X syndrome is challenging for short-read (100-150bp) whole genome sequencing (WGS) data. A solution to this problem is an important step towards integrating WGS into precision medicine. We have developed a research tool called ExpansionHunter that, using PCR-free WGS data, can identify repeat expansions at the locus of interest, even if the expansion is larger than the read length. We applied our algorithm to WGS data from 3,001 ALS patients who have been tested for the presence of the *C9orf72* repeat expansion with repeat-primed PCR (RP-PCR). Southern blot and fragment length analysis were applied on a subset of samples to confirm the presence or absence of the repeat expansion. Compared to the RP-PCR results, our WGS-based method identified pathogenic repeat expansions (>30 GGCCCC repeats) with 98.1% sensitivity and 99.7% specificity. Further inspection identified that 11 of the 12

conflicting calls were resolved as errors in the original RP-PCR results. Compared against this updated result, ExpansionHunter correctly classified 99.5% (212/213) of the expanded samples and all (2,788/2,788) of the wild type samples. The targeted repeat expansion caller we describe here marks a significant step towards a single whole genome medical test that includes detection of other pathogenic repeat expansions in WGS.

Introduction

For small variants such as single nucleotide polymorphisms and small insertions or deletions, variant callers typically require multiple reads to completely span the full length of the non-reference allele[1, 2]. For variants that deviate significantly from the reference, alternative methods such as *de novo* assembly can be employed if the variant is not highly repetitive[3-5]. Because high-throughput whole-genome sequencing (WGS) technologies are currently limited to ~150 base pair read lengths, variant-calling methods are subsequently limited to repeat lengths less than 150 bases[6]. Many pathogenic repeat expansions have repeats spanning hundreds to thousands of base pairs[7], so it has been assumed that short-read sequencing technologies may not be able to accurately detect repeat expansions of pathogenic lengths[8, 9]. A notable exception is the CAG repeat in the *HTT* gene in which individuals with at least 40 repeat units (120bp) are at risk of developing Huntington's disease[9]. For this repeat expansion, 150bp reads would be sufficient to identify the smallest pathogenic repeat expansions.

A recently discovered hexamer (GGCCCC) repeat expansion in the *C9orf72* locus is a major cause of both ALS and frontotemporal dementia[10, 11]. In particular, the pathogenic repeat length (>30 repeats; >180bp) is present in ~10% of all ALS patients including ~40% of familial ALS cases and ~6-8% of sporadic ALS cases in some populations[10-12]. The most widely used method to detect *C9orf72* repeat expansions is repeat-primed PCR (RP-PCR) together with fragment length analysis[13]. If a saw-tooth-like pattern with more than 30 "teeth" is present in RP-PCR and one or no fragments are detected in the fragment length analysis then the pathogenic *C9orf72* repeat expansion is considered present in the sample. Interpretation of these PCR results can be challenging due to indels in the flanking regions of the repeat, and can lead to both false positives and false negatives[13]. In addition, these PCR techniques do not provide an estimate of the length of the repeat expansions. Southern blotting is the current gold standard for identifying repeat length, but *C9orf72* Southern blotting is very challenging to set up; requiring a significant amount of input DNA (generally 10 micrograms) and suffers from imprecise size estimates due to low DNA quality and somatic heterogeneity[14]. As WGS is widely

adopted for use in precision medicine initiatives[9, 15, 16] and large scale research projects, a reliable method that can identify the presence or absence of repeat expansions in WGS data and also determine their approximate length without additional tests is needed.

Here, we present a method to detect pathogenic repeat expansions from PCR-free, WGS data implemented in a software package named ExpansionHunter (EH). To quantify the performance of this method we estimated the repeat lengths of two cohorts of ALS patients all of whom were independently assessed for the presence of the pathogenic *C9orf72* repeat expansion using RP-PCR with a subset also assessed using Southern blotting. Along with assessing longer-than-read-length-repeats, we also demonstrated the improved accuracy of this method compared to an existing method (lobSTR) on 865 samples for which the size of one non-pathogenic allele had been determined.

Results

We performed paired-end, PCR-free, WGS at an average depth of 45x using an Illumina HiSeq 2000 (100bp reads) and an Illumina HiSeq X (150bp reads) on two cohorts of patients with ALS (Supplementary Table 1, Supplementary Information 2.1 & 2.2). The first cohort of 2,559 patients was used for the initial methods development and the second cohort of 442 patients was used for blinded validation. All 3,001 samples were tested for presence of the *C9orf72* repeat expansion with RP-PCR (Supplementary Information 2.3, Supplementary Table 2) for a total of 2,377 wild type and 182 repeat-expanded samples in the first cohort and 416 wild type and 26 expanded samples in the second cohort using a combination of RP-PCR and fragment length analysis. A second RP-PCR test using a different primer set, fragment length analysis and Southern blotting was performed on 68 samples from the initial cohort of which 52 had a pathogenic *C9orf72* repeat according to the first RP-PCR (Supplementary Table 2, Supplementary information 2.4 and 4.1).

To quantify repeat lengths in the targeted region from WGS data we developed an algorithm that identifies reads that either: 1) span the repeat (spanning reads), or 2) include the repeat and the flanking sequence on one side of the repeat (flanking reads), or 3) are fully within the repeat, so-called "in-repeat" reads (IRRs, Figure 1). For short repeats (less than the read length of the sequence data) we calculated the repeat length using spanning and flanking reads (Figure 1). To estimate the lengths of repeats that are longer than the read length we identify and count the IRRs. There are three main hurdles associated with using IRRs to accurately identify repeat expansions that exceed read lengths: 1) identifying IRRs comprised of a potentially error-prone repeat motif, 2) identifying regions in the genome where IRR pairs are systematically (and possibly incorrectly) placed by the aligner, and 3) estimating the repeat length based on the total

number of IRRs identified. Here, we describe how we solve these problems and are able to accurately identify and characterize expanded repeats even when the repeats are longer than the reads.

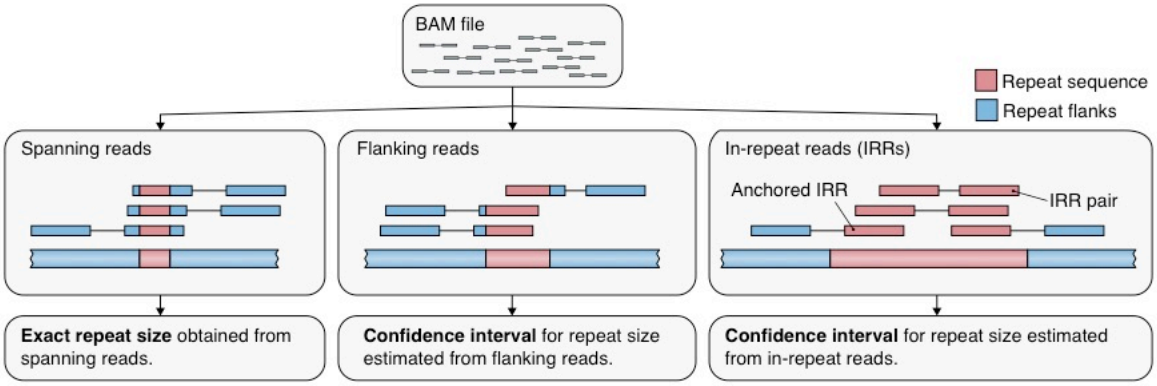

Figure 1. An outline of how EH catalogs reads associated with the repeat locus of interest and estimates repeat lengths starting from a binary alignment/map (BAM) file. The leftmost panels illustrate that short repeats are identified from spanning reads that include the repeat sequence (red) and flanking sequence on either side of the repeat (blue). For these reads the exact size of the repeat can be identified and each repeat size that is identified from spanning reads is reported. The middle panels illustrate that for repeats with a length that is close to the read length, reads may be identified with (part of) the repeat sequence and flanking sequence from one side of the repeat. If the length of the repeat is longer than the maximum size identified from the spanning reads it is assumed that these reads represent longer repeats and if no anchored IRRs are observed the repeat size is estimated with a confidence interval based on the number of flanking reads. The rightmost panels illustrate the reads that may be identified when the repeat is longer than the read length. When at least one anchored IRR is found, EH searches the unaligned reads for IRRs and through pre-defined off-target regions for pairs of IRRs. When IRRs are identified, a single repeat length is estimated based on the number of IRRs. EH outputs all repeat lengths for which it finds evidence.

On-target IRRs

Identifying reads in highly repetitive regions can be difficult because sequencing error rates are higher in low complexity regions such as homopolymers and short tandem repeats (STRs)[17], so we implemented a weighted measure that penalizes base mismatches at low quality bases less than mismatches at high quality bases (Supplementary information 3.1). To limit our analysis to IRRs that occur within the *C9orf72* repeat we extracted all the read pairs where one read is an IRR and the other read aligns with high accuracy (mapping quality (MAPQ)=60) within 1kb of the *C9orf72* repeat locus. We call these anchored IRRs and because the mate aligns nearby we are confident that the IRRs come from the repeat of interest. For repeats shorter than the fragment length but longer than the read length, the number of anchored IRRs can be used to estimate the repeat

length, while for repeats longer than the fragment length the number of anchored IRRs provides a lower bound of the repeat length.

Off-target IRRs

The library preparation used for these sequencing experiments had a mean fragment size of ~350-400bp but the *C9orf72* repeat expansion can be more than 10 kb in length[18]. This means that in addition to anchored IRRs, connected pairs of IRRs could be present in samples with the *C9orf72* repeat expansion (Figure 1). Because the expanded repeat is not present in the reference, these IRR pairs may not align to the *C9orf72* repeat locus and could either not align at all or misalign to a different locus in the genome[19]. To identify possible systematically misaligned reads, we tested every poorly mapped (MAPQ=0) read in all 182 ALS samples of the first cohort identified by RP-PCR as having the *C9orf72* repeat expansion. These 182 samples contained 29,443 IRR pairs altogether, 99% of which resided in 29 loci (Supplementary Figure 3c and c' and Supplementary information 3.2). In the 178 samples where we observed at least one IRR pair, the samples had on average 16.7 anchored IRRs while only one sample had less than four anchored IRRs. Conversely, when we performed the same analysis on 182 random samples without the *C9orf72* repeat expansion according to RP-PCR we did not find IRR pairs in any genomic locus.

We next analyzed all 2,559 samples from cohort one and identified all positions where the mate of any IRR was confidently aligned (MAPQ=60). For each sample we collated the number of anchored IRRs and then grouped IRRs anchored within 500bp of one another. The *C9orf72* repeat locus (chr9:27573490) had many anchored IRRs in nearly all samples with a pathogenic repeat expansion (177 samples had 5 or more anchored IRRs and 152 had 10 or more) indicating that the repeat exceeds the read length in these samples. Besides the *C9orf72* repeat locus, four genomic loci had two or three anchored IRRs in at least one sample (Figure 2). However, the samples in which these other-than-*C9orf72*-repeat-locus anchored IRRs were found did not have IRR pairs, neither did they have anchored IRRs in the *C9orf72* repeat locus. We therefore considered all IRR pairs to originate from the *C9orf72* repeat locus and included them in the size estimation of this repeat.

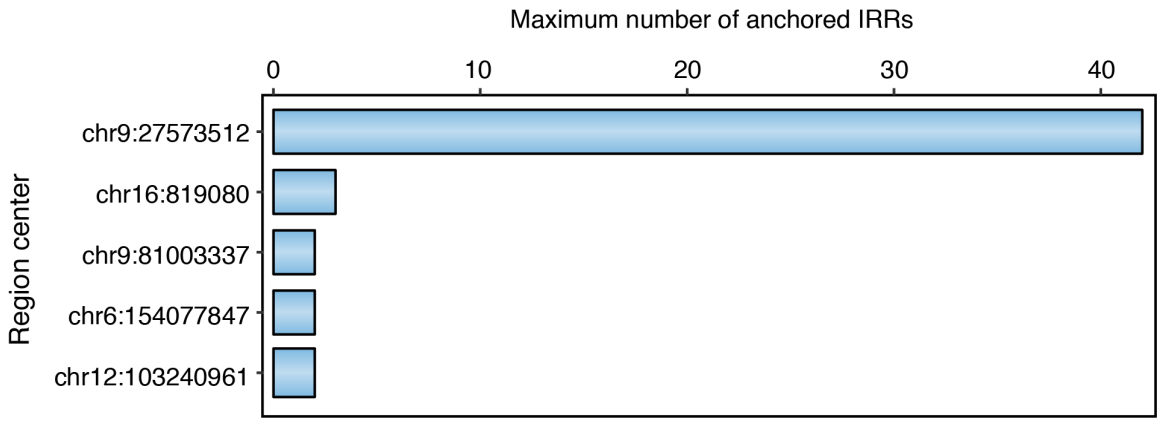

Figure 2. The maximum number of anchored IRRs observed in any of the 2,559 samples from cohort one for the genomic loci with at least two anchored IRRs (Supplementary information 3.2).

Repeat size estimation

Developments such as PCR-free WGS minimize the GC bias inherent in WGS data compared to PCR-based WGS data[20], as illustrated by the improved coverage of high GC regions such as the *FMR1* repeat (Supplementary Figure 5). Building on these improvements, we used the Binomial expectation to estimate the repeat length of expansions longer than the read length based on the number of IRRs identified (Supplementary information 3.3). A 95% confidence interval for the repeat size was estimated based on the number of IRRs, the read length and average genomic read depth using the parametric bootstrap method[21]. All IRRs of a single sample were considered to originate from the same haplotype.

For shorter alleles, the sizes of repeats are determined using spanning reads (Figure 1 and Supplementary information 3.4). For repeats that are close to the read length the repeat may be too long to identify spanning reads but too short to produce IRRs. In addition to spanning reads and IRRs, EH also uses flanking reads (Figure 1) to estimate the repeat size (Supplementary information 3.3). In the 3,001 samples analyzed here 2.2% (68) of the samples had a repeat size estimated based on flanking reads (Supplementary Table 2).

Pathogenic *C9orf72* repeat expansion determination

The *C9orf72* repeat sizes for both ALS cohorts were determined with EH and compared to the original RP-PCR results (Supplementary Table 2). Cases where the estimated confidence interval overlapped the pathogenic *C9orf72* repeat size, i.e. the lower bound was less than 30 repeats and the upper bound was greater than 30 repeats, were defined as "grey". Using the RP-PCR calls as the truth, the overall sensitivity and specificity for the WGS-based calls was 98.1% and 99.7%, respectively. There were 12 samples with a discrepant classification between our method and the RP-PCR in both cohorts, of which 8

were “false EH positive” and 4 “false EH negative” calls relative to RP-PCR. However, each of the “false EH positives” had at least 13 anchored IRRs and between 233 and 875 total IRRs per sample, which indicates strong supporting evidence for a pathogenic repeat expansion in these samples (Supplementary Table 4). Predicting the repeat length using only the anchored reads also provided evidence to support the existence of the pathogenic repeat expansion in all 8 “false EH positive” samples. Two of the 4 “false EH negative” samples had evidence supporting their negative status: i.e. they contained two alleles each with less than 30 repeats. Specifically, one sample had two different length alleles (2 and 5 repeats) each of which was supported by 10 spanning reads, indicating a heterozygous variant; while another sample had just one allele identified (2 repeats) but the number of spanning reads (38) was consistent with the read depth (mean depth = 44x) in this sample supporting a homozygous, non-pathogenic variant (Supplementary Table 4). The third and fourth “false EH negative” samples had a size estimate of just under the pathogenic cutoff (25 to 29 and 18 to 24 repeat units). In practice calls this close to pathogenic would likely be candidates for follow-up analysis.

	1st cohort		2nd cohort	
	RP-PCR ‘long’	RP-PCR ‘wt’	RP-PCR ‘long’	RP-PCR ‘wt’
True Positive	176	-	24	-
True Negative	-	2,372	-	412
Grey	3	0	1	1
False Positive	-	5	-	3
False Negative	3	-	1	-
Sensitivity	98.4%	-	96.2%	-
Specificity	-	99.8%	-	99.0%

Table 1. *C9orf72* repeat expansion sensitivity and specificity of EH on the ALS samples as compared to the RP-PCR results. RP-PCR ‘long’ samples reveal a pathogenic *C9orf72* repeat size pattern (>30 repeats), whereas RP-PCR ‘wt’ samples do not (<30 repeats). Grey calls have a confidence interval with a lower bound less than 30 repeats and an upper bound greater than 30 repeats. Grey calls were considered expanded when calculating the specificity and sensitivity.

For the 12 conflicting calls, we re-evaluated the original RP-PCR calls and performed an additional RP-PCR and fragment length analysis when our re-assessment of the original RP-PCR call was not conclusive. For 10 of the 12 conflicting calls, we determined that the original RP-PCR call was incorrect and were therefore no longer conflicting with the EH results (Supplementary Table 4). Of the two samples that were still in conflict with EH, one had an EH repeat size estimate of 25-29 repeats, being close to the repeat size number estimated from RP-PCR (29-32). The second sample produced consistent results with the EH calls when an additional RP-PCR was performed on these samples with different primers (Supplementary information 2.4). After modifying our truth calls to incorporate this additional assessment, there was only one false negative sample, though with a similar repeat-size estimate.

Repeats shorter than the read length

To quantify the accuracy of EH for alleles shorter than the read length, we compared our results to those obtained on 865 samples for which one allele size was estimated using fragment length analysis (Supplementary Table 2). In addition, we also analyzed these samples using the STR calling tool lobSTR[22]. It should be noted that lobSTR is designed for more general genome-wide STR calling based on spanning reads and is limited to calling repeat lengths shorter than the read length so it may fail to make a call for longer repeats. In this comparison, the EH calls agreed with the fragment length analysis in 829 (95.8%) of the samples and the lobSTR calls agreed with the fragment length analysis in 737 (85.2%) of the samples. Of the 36 EH calls that did not agree with the fragment length analysis, 15 (42%) were in agreement with the lobSTR calls and the remaining were all predicted to be longer repeats (spanning 9 or more repeat units) where lobSTR is less likely to make a call (Supplementary Table 2 and 5). In the remaining 21 samples, 10 were only off by one repeat unit and in two EH detected a significant expansion.

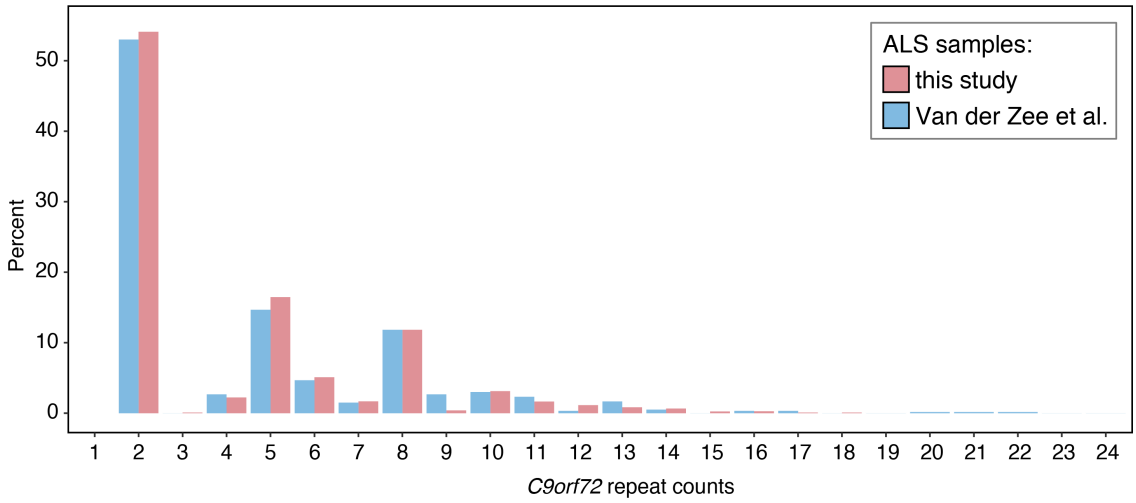

Figure 3. Distribution of allele sizes of the *C9orf72* repeat supported by multiple spanning reads in the 1,770 samples with 150bp reads from cohort one and two, compared with those of the FTLD cohort of 318 samples from a previous study[23].

Next, we analyzed the 1,770 samples that were sequenced with 2x150bp reads to get the distribution of the repeat lengths identified from spanning reads in the *C9orf72* repeat. The distribution achieved by this analysis is very similar to the results obtained in a previous study[23] (Figure 3) indicating that we are able to resolve the repeat lengths accurately based on spanning reads.

Accuracy of repeat size estimates longer than the read length

We performed Southern blotting to confirm the *C9orf72* repeat length of 52 pathogenic and 16 wild type samples from cohort one (Supplementary information 4.1). Low DNA quality, due to several freeze/thaw cycles, long-term storage in the freezer, shipment, and/or automated DNA extraction, prevented us from obtaining accurate repeat size estimates from the Southern blot experiments, in particular above 450 repeats (Figure 4a). With these caveats, we categorized the repeat sizes into bins (<30, 30-80, 80-450 and >450 repeats) and compared them with the repeat sizes estimated from EH (Figure 4b). This comparison demonstrated a broad overlap in repeat size estimates up to approximately 40 repeats (240 bp) though for the samples that were predicted to have more than 450 repeats many were estimated to have 80-450 repeats by EH.

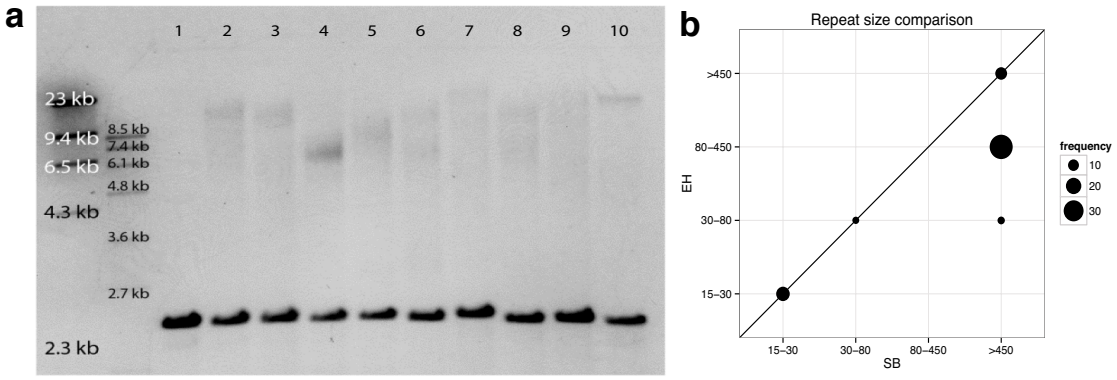

Figure 4. Repeat size comparison between Southern blot and EH. (a) Southern blot of ALS samples without (lane 1) and with (lanes 2-10) a pathogenic *C9orf72* repeat expansion for DNA that had been stored at -80°C for at least a year. The DNA smears above 5 kbp in lanes 2-10 indicate expansion of the *C9orf72* repeat, but DNA degradation and possible mosaicism prevented accurate repeat size estimation. The estimated repeat sizes are indicated in Supplementary Table 2. (b) Comparison of the estimated average repeat sizes from Southern blot (SB) and WGS for 68 samples.

We also measured the *FMR1* repeat expansion in eight samples obtained from the Coriell biorepository (https://catalog.coriell.org). *FMR1* is a pathogenic repeat expansion where diagnosis depends on the repeat size: Tremor Ataxia Syndrome corresponds to a repeat consisting of 55-200 repeats (165-600bp) and Fragile X Syndrome is diagnosed with 200 or more repeats (>600bp). The experimentally determined size of the *FMR1* repeat expansion in these samples ranged from 76 to 645 repeats (228-1,935bp). We identified 20 off-target regions where CGG IRR pairs may misalign. For these samples, EH correctly categorized each sample as either pre-mutated (55-200 repeats) or expanded (>200 repeats), though it underestimated the repeat size in three of these samples (Figure 5). As repeats become longer, even PCR-free sequence data may have biases either due to cluster growth or higher error rates causing these regions of extremely high (in this case 100%) GC content to be under represented in the raw reads. More samples with repeat expansions of known lengths will be needed to correctly quantify biases in the repeat size calculation that may be present for long repeats. While the sizes are underestimated for long, high-GC repeats it is likely that these biases are systematic and the ranking calls will be broadly correct. For example, the longest repeat in Figure 5 was also predicted to be much longer than the other samples by our method.

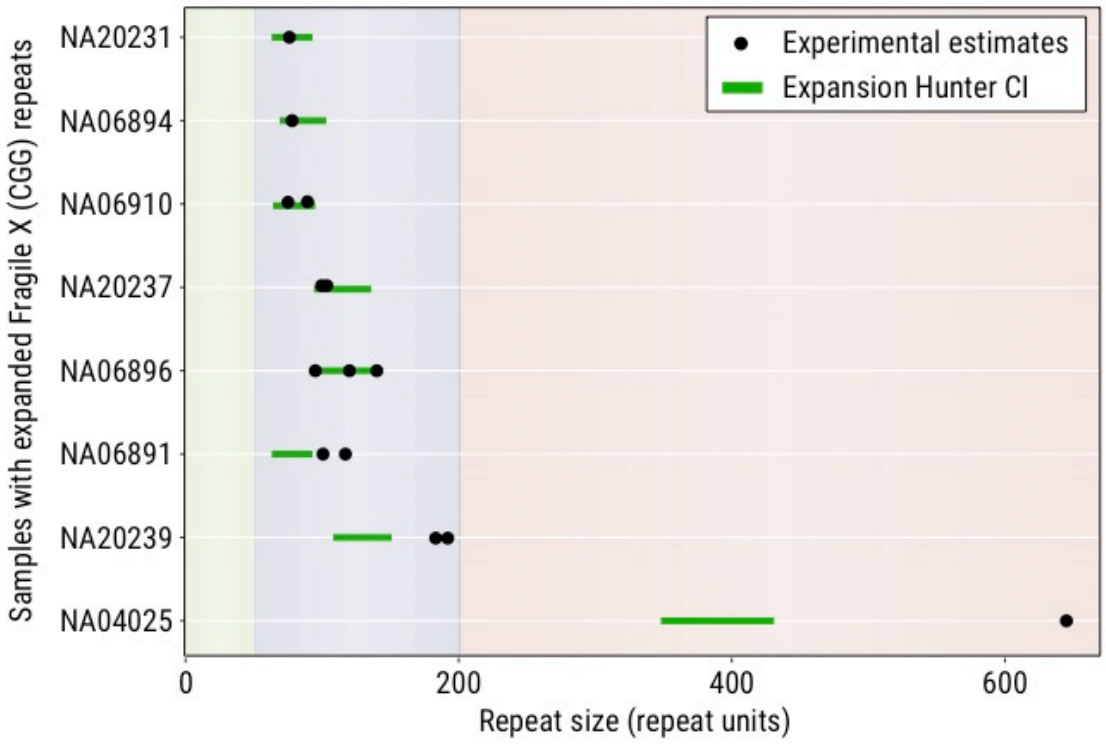

Figure 5. Confidence intervals for repeat sizes produced by EH (bars) and corresponding experimental estimates provided by Coriell (points corresponding to experimental point estimates or ends of size ranges) for expanded *FMR1* repeat alleles. Green shaded region highlights the "normal" range for this repeat; blue shaded region highlights the repeat range associated with Tremor Ataxia Syndrome; red shaded region highlights the repeat length associated with Fragile X Syndrome.

Discussion

We have demonstrated a method that can identify repeat expansions from paired-end PCR-free WGS data. This method will not work with whole-exome sequence data or WGS data that includes a PCR step in the sample preparation, because: 1) many of the repeats of interest are not exonic, 2) many repeat expansions have a high GC content and 3) size estimates for large repeats require a strong prior for the average number of reads per base. Comparing our results from WGS data against the current wetlab protocol for identifying pathogenic repeat expansions in the *C9orf72* locus, only 12 out of 3,001 samples had conflicting calls. Ten of these 12 conflicts were subsequently identified as miscalling of the original RP-PCR results. One conflict was resolved with a RP-PCR with a different primer set, suggesting an indel near the *C9orf72* repeat locus[23]. The remaining conflict had an EH repeat size estimate very close to the repeat size number estimated from RP-PCR, 25-29 and 29-32, respectively.

EH determines both the length of repeats that are shorter than the read length and estimates a repeat size range for repeats longer than the read length. While identifying whether a repeat of interest exceeds its pathogenic size is crucial, accurate repeat length estimates are also important because the repeat length may correlate with the age of onset and phenotype severity. Comparing the size estimates from our WGS method against annotated sizes from wet lab results indicates that our repeat size estimates are accurate up to repeats of ~240-300 bp. These results are based on only a few samples and future work will sequence high quality DNA from many more samples with expanded repeats of known size to provide a better assessment and to improve the accuracy of repeat expansion detection. It should be noted that both the *C9orf72* and *FMR1* repeat expansions analyzed here are 100% GC and thus likely to provide a lower bound for the ability of EH to estimate repeat sizes. We expect to achieve better expansion size estimates for long CAG repeats, such as the 50-1000 repeats in the *DM1* gene associated with Myotonic Dystrophy, both because the base quality will be better and also because the GC bias will be lower.

We analyzed two pathogenic repeat expansion regions with a particular emphasis on the *C9orf72* repeat expansion though further work is in progress to validate calling other

known pathogenic repeat expansions. As described here, EH is primarily designed to identify the longest repeat at a targeted locus of interest and work is ongoing to assess and improve the accuracy of EH as a general genotyping tool for STRs. Additionally, future work will focus on extending this method to identify all possible locations in the genome that could have a repeat expansion based on the observation of anchored IRRs to reduce potential false positives. This study represents the first method that can utilize WGS data to accurately identify long repeat expansions and estimate their lengths. Ultimately, once all known and newly identified pathogenic repeat expansions have been validated, all repeat expansions can be tested from a single PCR-free WGS sequencing run.

Code availability

ExpansionHunter is written in C++ and the source code is freely available at https://github.com/Illumina/ExpansionHunter.

Data access

The raw sequence data used for *C9orf72* repeat expansion detection was deposited at SURFsara (Amsterdam) and is available on request. The raw sequence data of the FMR1 repeat in the Coriell samples is available in the Supplementary Material.

Acknowledgements

This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. This research was supported by NIH/NINDS P01 NS084974 (MvB, RR), R01 NS080882 (MvB, RR), the Thierry Latran foundation (MAvE, JHV, GPT), the Netherlands Organization for Health Research and Development (Veni scheme, MAvE), the ALS Foundation Netherlands, and the MND association (UK) (Project MinE, www.projectmine.com). Research leading to these results has received funding from the European Community's Health Seventh Framework Programme (FP7/2007-2013). This study was supported by ZonMW under the frame of E-Rare-2, the ERA Net for Research on Rare Diseases (PYRAMID). This is an EU Joint Programme–Neurodegenerative Disease Research (JPND) project (STRENGTH, SOPHIA). The project is supported through the following funding organizations under the aegis of JPND: UK, Medical Research Council; Ireland, Health Research Board; Netherlands, ZonMw. AAC receives salary support from the National Institute for Health Research (NIHR) Dementia Biomedical Research Unit at South London and Maudsley NHS Foundation Trust and King's College London. Samples used in this research were in part obtained from the UK National DNA Bank for MND Research, funded by the MND

Association and the Wellcome Trust.

References

1. Raczy, C. et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041-2043 (2013).
2. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
3. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44**, 226-232 (2012).
4. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**, 3694-3696 (2015).
5. Weisenfeld, N.I. et al. Comprehensive variation discovery in single human genomes. *Nat Genet* **46**, 1350-1355 (2014).
6. Narzisi, G. & Schatz, M.C. The challenge of small-scale repeats for indel discovery. *Front Bioeng Biotechnol* **3**, 8 (2015).
7. Gatchel, J.R. & Zoghbi, H.Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet* **6**, 743-755 (2005).
8. Loomis, E.W. et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res* **23**, 121-128 (2013).
9. Ashley, E.A. Towards precision medicine. *Nat Rev Genet* **17**, 507-522 (2016).
10. DeJesus-Hernandez, M. et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245-256 (2011).
11. Renton, A.E. et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257-268 (2011).
12. Gijselinck, I. et al. A C9orf72 promoter repeat expansion in a Flanders-Belgian cohort with disorders of the frontotemporal lobar degeneration-amyotrophic lateral sclerosis spectrum: a gene identification study. *Lancet Neurol* **11**, 54-65 (2012).
13. Akimoto, C. et al. A blinded international study on the reliability of genetic testing for GGGGCC-repeat expansions in C9orf72 reveals marked differences in results among 14 laboratories. *J Med Genet* **51**, 419-424 (2014).
14. Buchman, V.L. et al. Simultaneous and independent detection of C9ORF72 alleles with low and high number of GGGGCC repeats using an optimised protocol of Southern blot hybridisation. *Mol Neurodegener* **8**, 12 (2013).
15. Marx, V. The DNA of a nation. *Nature* **524**, 503-505 (2015).

16. Ashley, E.A. The precision medicine initiative: a new national effort. *JAMA* **313**, 2119-2120 (2015).

17. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012).

18. Gijselinck, I. et al. The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol Psychiatry* **21**, 1112-1124 (2016).

19. Church, D.M. et al. Extending reference assembly models. *Genome Biol* **16**, 13 (2015).

20. Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum Genet* **135**, 359-362 (2016).

21. Rice, J.A. Mathematical statistics and data analysis, Edn. 3rd. (Thomson/Brooks/Cole, Belmont, CA; 2007).

22. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**, 1154-1162 (2012).

23. van der Zee, J. et al. A pan-European study of the C9orf72 repeat associated with FTLD: geographic prevalence, genomic instability, and intermediate repeats. *Hum Mutat* **34**, 363-373 (2013).

24. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

25. Thion, M.S. et al. CAG repeat size in Huntingtin alleles is associated with cancer prognosis. *Eur J Hum Genet* **24**, 1310-1315 (2016).