Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders

Rick Tankard, BSc Hons, DipA 1,2
Martin B Delatycki MD 3,4,5
Paul J Lockhart PhD 3,5[†]
Melanie Bahlo PhD 1,2* [†]

1 Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, Parkville 3052, VIC, Australia
2 Department of Medical Biology, The University of Melbourne, Melbourne 3010, VIC, Australia
3 Bruce Lefroy Centre for Genetic Health Research, Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia.
4 Victorian Clinical Genetics Services, Parkville, Victoria, Australia.
5 Department of Paediatrics, University of Melbourne, Parkville, Victoria, Australia.

[†] Joint senior authors
* Corresponding Author: Melanie Bahlo, bahlo@wehi.edu.au

Summary
Background
Repeat expansions cause over 20 neurogenetic disorders that can present with overlapping clinical phenotypes, making molecular diagnosis challenging. Single gene or small panel PCR-based methods are employed to identify the precise genetic cause, but can be slow and costly, and often yield no result. Genomic analysis via whole exome and whole genome sequencing (WES and WGS) is being increasingly performed to diagnose genetic disorders. However, until recently analysis protocols could not identify repeat expansions in these datasets.
Methods
A new method for the identification of repeat expansions using either WES or WGS was developed. Four retrospective cohorts of individuals with eight different known repeat expansion disorders were analysed with the new method. Results were assessed by comparing to the known disease status. Performance was also compared to a recently published genotyping-based method, ExpansionHunter.
Findings
Expansion repeats were successfully identified in WES and WGS datasets. The new method demonstrated very high predictive capabilities, achieving a median area under the curve (AUC) of 0.9. The new robust method achieved a median specificity and sensitivity of 0.99 and 0.75 respectively, compared to ExpansionHunter, a recently published genotyping-based method (median specificity = 0.99, median sensitivity = 0.56).
Interpretation
The new method, called exSTRa (expanded STR algorithm), is available from https://github.com/bahlolab/exSTRa. It can be applied to existing WES or WGS data to identify likely repeat expansions. We demonstrate that exSTRa can be effectively utilized as a screening tool to interrogate WES and WGS sequencing data generated with PCR-based library preparations which can then be followed up with specific diagnostic tests.
Funding
The Australian National Health and Medical Research Council, the Australian Research Council, The Murdoch Children's Research Institute, The Walter and Eliza Hall Institute of Medical Research, and the Victorian Government.

Introduction
Thousands of short tandem repeats (STRs), also called microsatellites, are scattered throughout the human genome. STRs vary in size but are commonly defined as having a repeat motif 2-6 base pairs (bps) in size. They are underrepresented in the coding regions of the human genome[1], despite the vast majority being population polymorphisms with no, or very little, phenotypic consequence. STRs were used as genetic markers for linkage mapping for human studies for many years, and continue to be used, but primarily for non-human studies. A subset of STRs can however cause disease. Pathogenic STRs have either one or two alleles, depending on the genetic model, that exceed some threshold for biological tolerance. These diseases are known as repeat expansion disorders. The abnormal STR allele(s), may affect gene expression levels, cause premature truncation of the protein, and result in aberrant protein folding[2].

Repeat expansions cause at least 22 inherited germline human disorders, most often presenting with ataxia as a clinical feature. The size of pathogenic repeats varies from ~60 (*CACNA1A*) to several thousand (*C9orf72*) base pairs and cause predominantly neurogenetic diseases, such as ataxias. Remarkably 12 repeat expansions have now been identified as causing dominant forms of spinocerebellar ataxias. Other disorders caused by repeat expansions include fragile X syndrome (a repeat in the 5'UTR of *FMR1*), Huntington Disease (a repeat in exon 1 of *HTT*), myotonic dystrophy (repeats in *DMPK* and *ZNF9*), fronto-temporal dementia and amyotrophic lateral sclerosis 1 (a 6-mer repeat in *C9orf72*) and Unverricht-Lundborg disease, a severe myoclonic epilepsy (*CSTB*). The genetic mode of inheritance encompasses autosomal dominant (e.g. SCA1) and recessive (e.g. Freidreich ataxia), as well as X-linked recessive (e.g. fragile X syndrome). Many repeat expansion disorders show anticipation, a phenomenon whereby younger generations are affected by earlier age of onset. Anticipation is usually caused by an increase in repeat size between generations. When anticipation is observed it indicates that a search for repeat expansions as the cause of disease is warranted.

Friedreich ataxia is the most common of the recessive repeat expansion disorders, with a disease prevalence of 3 to 4/100,000 but with a carrier frequency of 1/100[4]. Fragile X syndrome is the most

common cause of inherited intellectual disability and affects ~1:5000 individuals[5,6]. Hence these diseases as a whole contribute significantly to the overall Mendelian disease burden in human populations.

Diagnostic identification of repeat expansions can be time consuming and costly. Current medical diagnosis consists of precise PCR or Southern blot assay, which require diagnostic laboratories that have refined these assays for each different repeat expansion. The clinician has to determine which repeat expansions are most likely to be relevant and submit the patient's DNA to appropriate laboratories. This can be difficult, given the phenotypic overlap between the different STRs, the potential heterogeneity in the symptoms and the variation in penetrance and age of onset, which is also dependent on the size of the allele and effect of modifier genes[7,8]. In addition, up to 50% of individuals with a diagnosis of ataxia may be due to other mutation types, such as single nucleotide variants (SNVs) and short insertion/deletions (indels)[9]. Therefore, molecular diagnosis of these disorders often also requires conventional sequencing of candidate genes, either by Sanger, targeted panel or Next Generation Sequencing (NGS) methods.

Short-read NGS data, such as that generated by the Illumina sequencing, is currently predominant in both research and clinical diagnostic applications. Moreover, Whole Genome Sequencing (WGS) is now an affordable technology, gradually replacing whole exome sequencing (WES) for clinical genomics. Illumina's HiSeq X platform is currently the most commonly used platform for the generation of human WGS data and in particular clinical human genome sequencing with its low error rates and well-documented, consistent, performance.

Illumina HiSeq X data reads are 150 bp in length and are designed so that the reads are transcribed facing each other, where the template DNA has a small gap between the reads that is not sequenced. This gap can vary in size, but standard library preparation methodologies generate insert fragment lengths of ~350 bps, resulting in a gap of ~50bp.

Standard clinical diagnostic pipelines focus on the identification of SNVs and indels. Bioinformatic tools have been developed to genotype STRs, but are almost entirely confined to those STR alleles that are spanned by reads[10-14]. Pathogenic repeat expansions are usually significantly longer than the reads generated by short-read sequencing platforms such as Illumina, and may be longer than the library insert fragments lengths. Therefore, the short reads cannot span many pathogenic repeat expansion alleles, such as those that cause SCA2, or SCA7 (Table 1). Furthermore some of these reads are not mapped, or poorly mapped, to the STR allele, due to sequencing bias and alignment issues such as: (i) the repetitive nature of the repeat itself where the expanded alleles require alignments of additional repetitive bases, (ii) multiple occurrences of the same repeat throughout the genome, leading to multi-mapping reads, and (iii) GC bias. Despite this, these data do still carry information about the expanded allele with a larger number of reads mapping to the STR for an expanded allele than expected, based on the reference STR allele lengths. The key to effectively utilizing this data is to make use of the uniquely mapped mate pairs of the paired-end read, rather than analyzing only the sequence that maps into the STR.

A recent preprint described ExpansionHunter, an algorithm that facilitates detection of expansion repeats in WGS data[15] and analysed NGS data at two repeat expansion disorder loci (*FMR1* and *C9orf72*). ExpansionHunter is designed to determine the allele size of a repeat expansion, without any statistical evaluation framework to determine whether a particular individual is likely to harbour a repeat expansion at a given locus. For validation the method analysed WGS data generated using a recent library protocol which is 'PCR-free', meaning that one of the additional sources of STR stutter[16] is not present, thus making calling of alleles simpler.

Here we present the STR repeat expansion-calling algorithm, exSTRa (**ex**panded **STR a**lgorithm), which detects expanded repeat expansion allele(s) at known repeat expansion loci for neurogenetics disorders. We demonstrate the utility of our method with eight different verified repeat expansion disorders. We also demonstrate for the first time, that repeat expansion detection is possible with WES data.

Methods

Study cohorts and next-generation sequencing data generation

Individuals with already diagnosed repeat expansion disorders were recruited to the study. The repeat expansion status was verified via standard STR-specific PCR-based assays. Individuals affected by neurogenetics disorders not due to known repeat expansions were recruited as controls. These individuals were not tested for any of the known repeat expansion loci with standard methods as none of them are affected by symptoms that are typical of expansion disorders such as ataxia. All individuals were recruited at the Murdoch Children's Research Institute, and provided written informed consent (Human Research Ethics Committee #28097, #25043 and #22073).

The four cohorts underwent different types of NGS, with some individuals being sequenced multiple times. Individuals were sequenced with either: (i) WES with the Agilent V5+UTR capture platform (4 repeat expansion patients, with 4 different expansion disorders, 58 controls), (ii) WGS with the TruSeq Nano protocol, which includes a PCR step to increase sequencing material (17 repeat expansion patients, with 8 different expansion disorders, 16 controls), or (iii) WGS with the KAPA Hyper PCR-free protocol (three repeat expansion patients, with three different expansion disorders, 5 controls). Two different cohorts were sequenced with protocol (ii). These are designated as WGS_PCR_1 and WGS_PCR_2. These cohorts are described in Table 2.

Sequencing Data generation

WGS data was generated by the Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, Australia with HiSeq X Ten. The WES data was generated by the Australian Genome Research Facility, Melbourne, Australia, and sequenced on a HiSeq 2500 sequencer. All sequencing data was aligned to the hg19 human genome reference using the Bowtie 2 aligner[17] in local alignment mode.

Definition of Repeat Expansion Loci

Table 1 defines the chromosomal location, physical map location, disease, genetic disease model and repeat motif, normal and repeat expansion size for all 21 neurogenetic repeat expansion loci that were analysed.

Data extraction for repeat expansions

We developed a two-step analysis method, called exSTRa, detailed in the Supplementary Data, to identify individuals likely to have a repeat expansion at a particular STR locus. The analysis method extracts STR repeat content information for each read, stemming from a particular individual, which has been identified as mapping to one of the 21 STR loci. We designed a statistical test that captures the differences between an individual to be tested within a cohort of cases and controls. All N individuals within a cohort are examined in turn at each of the 21 known pathogenic repeat expansion loci by comparing each individual in turn to all N other individuals in each cohort. This generates 21xN test statistics per cohort. The empirical p-value of the test statistic was determined using a simulation method. All p-values over all STR loci for all individuals within each cohort were assessed for approximate uniform distribution with histograms and Quantile-Quantile (Q-Q) plots.

Raw data was visualized using empirical cumulative distribution functions (ECDFs), which display the distribution of the amount of STR repeat motif found in each read, ordered from smallest to largest content amount, as a step function. This allows comparison of the distributions, regardless of sequencing depth. Reads generated from expanded alleles have increased numbers of repeat motifs in their reads compared to reads stemming from normal alleles. This produces a shift of the read repeat motif distribution to the right for the individual with the repeat expansion, in comparison to reads from individuals with normal alleles. The shape of the ECDF is affected by additional factors such as: the genetic model (dominant, recessive or X-linked), GC content and capture efficiency (for WES). PCR stutter and somatic changes in STR length are also a factor, but are not as prominent. The analysis methods are available from http://github.com/bahlolab/exSTRa.

Comparison with ExpansionHunter

ExpansionHunter[15] estimates the repeat size using a parametric model but does not attempt to call repeat expansions in a probabilistic framework. ExpansionHunter was used to determine if alleles were larger than currently known smallest disease-causing repeat expansion alleles. This resulted in binary calls of normal/expanded for each of the 21 STR loci for each of the individuals in the cohorts.

Performance evaluation

We called individuals as being normal or expanded based on the Bonferroni multiple testing corrected p-values derived from our empirical p-values (see Supplementary Methods). These results were compared to the known disease status. Performance of exSTRa and ExpansionHunter were evaluated for sensitivity and specificity. Area under the curve (AUC) was also calculated for exSTRa.

Results

Coverage and Alignment Results

Full coverage and alignment results are in Supplementary Table 2. Genome-wide sample specific coverage variability, as measured by the median IQR of the mean coverage library size corrected samples, was very similar between all three WGS_cohorts (WGS_PCR_1 median IQR = 8, WGS_PCR_2 median IQR = 5.7, WGS_PF_3 median IQR = 7.1). In contrast the WES data showed substantial variability (median IQR = 22.3).

We also examined the 21 STR loci for coverage in our four study cohorts. As expected WES_PCR only achieved reasonable coverage for repeat expansion detection in a subset of the STR loci. However, this included many of the known repeat expansion STRs located in coding regions (8 out of 10) (Figure 1 and Supplementary Figure 1). SCA6 (*CACNA1A*) and SCA7 (*ATXN7)* are poorly covered. Despite the use of the Agilent SureSelect V5+UTR capture platform which incorporates UTRs we achieved no, or very low coverage, for the known repeat expansion loci located in the UTR, such as FRAXA, FRAXE and DM1. DM1 and SCA7 are not captured by the Agilent enrichment platform (Supplementary Table 3), however both FRAXA and FRAXE are targeted and should be captured. In general WGS data outperformed WES over all STR loci, with one exception, SCA3, located in the coding region of *ATXN3*. The reason for this is currently unknown.

Visualizations of repeat motif distributions

ECDF curves of selected loci are shown for each cohort to illustrate the data. Full results for all 21 loci, for all four cohorts are given in Supplementary Figures 3-8. STR loci varied in their coverage with several loci consistently poorly captured. These were usually loci that are rich in GC content. Short read NGS data has a known GC bias with a GC content of 40-55% maximizing sequencing yield, depending on sequencing platform[18].

The STR loci also showed differences in variability with regards to STR motif lengths. Some STR loci, such as SCA17 and HDL2, showed little variability in STR allele distributions, regardless of NGS platform. Identification of outliers is easier for these loci, with low background variability. Those repeat expansion disorders that are autosomal recessive or X-linked recessive (in males), also show much clearer outlier distributions (Figure 2, top right panel). This is due to the outlier distribution deviating for either both alleles, or, in the case of the X-chromosome, and only males, just the one allele having to be examined (not performed in this analysis).

Statistical test results

Test statistics were generated for all 21 loci for all N individuals for all four cohorts with exSTRa. Combined p-values over all STR loci for all individuals within each cohort showed approximate uniform distribution with histograms (Supplementary Figure 9) and Q-Q plots (Supplementary Figure 10), albeit with some inflation of p-values at both tails. Our study cohorts had very small numbers of control individuals for some of the cohorts.

Expansion call results

Expansion call results are presented in summary form in Table 3 and at the individual level in Supplementary table 3. In the WGS_PCR_2 cohort we achieved an AUC of 0.98, with a sensitivity of 0.75 and specificity of 0.98. WES_PCR, WGS_PCR_1 and WGS_PCR_2 achieved AUCs≥0.9 (Table 3). We also divided the WGS_PCR_2 cohort data into two sub-cohorts, where each sample's data comes from a single flow cell lane that has ~30X coverage. This allowed an investigation of reproducibility, and assessment at the more standard 30X coverage. Results were highly reproducible between the two 30X replicates, with only one sample changing its call between the two sequencing runs. We also observed very little change in performance between the 60X and 30X data with virtually identical AUCs of ~0.9.

Across all cohorts exSTRa called 42 expansions compared to ExpansionHunter, which called 32 expansions out of 58 samples with expansions. Notably, of the eight expansion loci assessed, exSTRa correctly identified expanded repeats at seven. Both ExpansionHunter and exSTRa performed poorly

when analyzing samples with an *FMR1* expansion. In the WGS_PCR_1 and 2 cohorts this is due to poor coverage at the FRAXA and FRAXE loci and GC bias issues (Supplementary Figure 1). In the WGS_PF data, the coverage was much higher for FRAXA when compared to the WGS_PCR cohorts (compare Figure 3 top two panels). However, the performance of exSTRa was compromised by a lack of controls. Although there was a clear right shift of both the full mutation and premutation *FMR1* samples (Figure 3 top right panel), this was not statistically significant.

Discussion

Genomic medicine, which uses genomic information about an individual as part of their clinical care, promises better patient outcomes and a more efficient health system through rapid diagnosis, early intervention, prevention and targeted therapy[19,20]. A single affordable front-line test that is able to comprehensively detect the genetic basis of human disease is the ultimate goal of diagnostics for genomic medicine and represents the logical way forward in an era of personalised medicine. Screening tests will play a major role in the implementation of preventative medicine.

Currently, the diagnostic pathway for suspected repeat expansion disorders utilizes single gene tests or small target panels, employing a condition-by-condition approach. This method is cost effective when the clinical diagnosis is straightforward. However, for some disorders, such as spinocerebellar ataxias, the 'right' test is not immediately obvious[9]. Many families remain unsolved, even after extensive genetic studies encompassing both gene sequencing and expansion repeat testing[9]. The implementation of a single NGS-based test that could identify causal point mutations, indels and expanded STRs is likely to be cost effective in this context. Expected outcomes would include both increased diagnostic yield and a reduction in the diagnostic odyssey for many affected individuals.

Methods such as lobSTR[10], attempt to genotype STRs, i.e. estimate the allele sizes, which renders them ineffective when the repeat size exceeds the read length of the sequencing platform. Only one method, the recently described ExpansionHunter[15] is designed to call repeat expansions. By examining performance using multiple individuals with eight different repeat expansion disorders, we show that our method, exSTRa, does not require PCR-free library sequencing protocols, nor even WGS, to detect repeat expansions. We show that it outperforms ExpansionHunter, with a very slight loss of specificity (<0.02). ExpansionHunter currently requires knowledge of the pathogenic allele size, which will not be known for novel repeat expansion loci.

In the context of our results, exSTRa appears to have significant potential as a population screening tool for carrier status. For example, we will be able to identify carriers for Friedreich's ataxia, the most prevalent of the inherited ataxias, with a carrier frequency of ~1/100. More broadly, although the current version of exSTRa performed suboptimally for detection of *FMR1* expansions, we believe these limitations can be resolved with further refinements of exSTRa or similar detection methods. Our study was hampered by a lack of control individuals for the PCR-free cohort that included the two individuals with FRAXA full mutations and the single individual with a FRAXA premutation (WGS_PF). Fragile X syndrome (FXS) is the most common cause of inherited ID. Approximately 1:300 individuals carry a premutation allele (55-200 repeats) which causes fragile X-associated tremor ataxia syndrome and fragile X primary ovarian insufficiency[21]. Currently, newborn/carrier screening is not performed for FXS. Historically, there was no medical advantage to early detection of FXS, although recent targeted treatments have shown potential benefits[22,23]. There is now discussion regarding the clinical utility of screening *FMR1* for reproductive and personal healthcare[24].

Given that many families currently remain unsolved, even after extensive genetic sequencing, we recommend the introduction of a protocol such as ours into any standard sequencing analysis pipeline and that this be run both prospectively and retrospectively. This should not only identify individuals that have escaped repeat expansion detection but will also expedite the diagnosis of individuals potentially suffering from a repeat expansion disorder. There are currently 22 known repeat expansion loci, yet more are likely awaiting discovery. In OMIM there are additional putative SCA loci, such as SCA25 (OMIM #608703, 2p21-p13), with as yet unidentified genetic causes, but which are potentially repeat expansions. Therefore, with large cohorts and further improvements in methodology, we believe methods such as exSTRa will facilitate the discovery of novel repeat expansion loci, which, in turn, will identify the etiology of neurodegenerative disorders in more families.

A repeat expansion call from exSTRa, or ExpansionHunter, should not replace gold standard, locus-specific, PCR-based tests. Firstly, these will remain gold-standard, with higher sensitivity and

specificity than the sequencing-based methods, and secondly, they give much more accurate estimates of the size of the expanded allele(s), which has prognostic implications for the age of onset, disease progression and outcome. This is analogous to the requirement for Sanger sequencing confirmation of SNVs and indels to confirm variants detected by standard NGS analysis pipelines.

We anticipate that we will be able to further improve our method going forward, with a future goal of calling STRs genome-wide. This will require much larger groups of controls. Fortunately this will be the likely sampling framework in cohort studies, with far fewer repeat expansions and a large number of true negatives, which will better describe the null distributions of the normal alleles/genotypes. There are clearly sources of bias that affect certain loci that are contributing to the poor performance at some of the STRs. For instance we observed a GC bias for the repeat expansion alleles for repeat expansion disorders FRAXA, FRAXE and FTDALS1, with far fewer reads able to capture these repeat expansions due to their extreme GC content. Notably FRAXA and FTDALS1 had substantially improved coverage with the PCR-free protocol.

Long read WGS will see further improvements in the detection of repeat expansion alleles, allowing capture of the entire expanded allele in a read fragment, but is currently not cost-effective, being almost 10 times more expensive than the prevailing Illumina HiSeq X sequencing platform. The development of methods such as exSTRa will lead to further improvements in patient care via clinical genomic sequencing. They will also facilitate the pending era of precision/preventative medicine, when screening tests will become much more prevalent. A universal single test will be cost effective in comparison to the array of existing tests currently required to test for all known mutation types.

Contributions
RT wrote the software and performed all the data analysis. MB and RT conceptualized the algorithms. MB, PL and MD designed the sequencing experiments. MD and PL recruited the individuals for the study. MB and RT wrote the first draft of the manuscript. PL and MD reviewed subsequent drafts.

Conflicts of interest
We declare that we have no conflicts of interest.

Acknowledgements
We would like to acknowledge Egor Dolzhenko from Illumina Inc, who generated the STR specification files to allow us to generate the ExpansionHunter results. We thank Leslie Burnett, Ben Lundie, Katie Ayres and Andrew Sinclair for access to control WGS datasets prepared using the PCR-free protocol. We thank Greta Gillies, Kate Pope and Peter Degorski for assistance with recruitment, sample preparation and manuscript preparation.

RT was supported by an Australian Postgraduate Award and funding from the Edith Moffat fund. PJL was supported by NHMRC CDA2 (GNT1032364). MB was supported by NHMRC Program Grant (GNT1054618) and NHMRC Senior Research Fellowship (APP1102971). This work was supported by the Victorian Government's Operational Infrastructure Support Program and Australian Government National Health and Medical Research Council Independent Research Institute Infrastructure Support Scheme (NHMRC IRIISS).

Disease	Symbol	OMIM	Inheri tance	Gene	Cytogenetic Location	Type	Repeat Motif	Normal Range	Expansion Range	Strand	Start hg19	Reference Repeat Number	TRF Match (%)	TRF Indel (%)	Reference STR size (bp)
Huntington disease	HD	143100	AD	*HTT*	4p16.3	coding	CAG	6-34	36-100+	+	3,076,604	21.3	96	0	64
Kennedy disease	SBMA	313200	X	*AR*	Xq12	coding	CAG	9-35	38-62	+	66,765,159	33.3	86	9	103
Spinocerebellar ataxia 1	SCA1	164400	AD	*ATXN1*	6p23	coding	CAG	6-38	39-82	-	16,327,865	30.3	95	0	91
Spinocerebellar ataxia 2	SCA2	183090	AD	*ATXN2*	12q24	coding	CAG	15-24	32-200	-	112,036,754	23.3	97	0	70
Machado-Joseph disease	SCA3	109150	AD	*ATXN3*	14q32.1	coding	CAG	13-36	61-84	-	92,537,355	14	84	0	42
Spinocerebellar ataxia 6	SCA6	183086	AD	*CACNA1A*	19p13	coding	CAG	4-7	21-33	-	13,318,673	13.3	100	0	40
Spinocerebellar ataxia 7	SCA7	164500	AD	*ATXN7*	3p14.1	coding	CAG	4-35	37-306	+	63,898,361	10.7	100	0	32
Spinocerebellar ataxia 17	SCA17	607136	AD	*TBP*	6q27	coding	CAG	25-42	47-63	+	170,870,995	37	94	0	111
Dentatorubral-pallidoluysian atrophy	DRPLA	125370	AD	*DRPLA/ATN1*	12p13.31	coding	CAG	7-34	49-88	+	7,045,880	19.7	92	0	59
Huntington disease-like 2	HDL2	606438	AD	*JPH3*	16q24.3	exon	CTG	7-28	66-78	+	87,637,889	15.3	95	4	47
Fragile-X site A	FRAXA	300624	X	*FMR1*	Xq27.3	5'UTR	CGG	6-54	200-1000+	+	146,993,555	25	90	5	75
Fragile-X site E	FRAXE	309548	X	*FMR2*	Xq28	5'UTR	CCG	4-39	200-900	+	147,582,159	15.3	100	0	46
Myotonic dystrophy 1	DM1	160900	AD	*DMPK*	19q13	3'UTR	CTG	5-37	50-10000	-	46,273,463	20.7	100	0	62
Friedreich ataxia	FRDA	229300	AR	*FXN*	9q13	intron	GAA	6-32	200-1700	+	71,652,201	6.7	100	0	20
Myotonic dystrophy 2	DM2	602668	AD	*ZNF9/CNBP*	3q21.3	intron	CCTG	10-26	75-11000	-	128,891,420	20.8	92	0	83
Frontotemporal dementia and/or amyotrophic lateral sclerosis 1	FTDALS1	105550	AD	*C9orf72*	9p21	intron	GGGGCC	2-19	250-1600	-	27,573,483	10.8	74	8	62

Myoclonic epilepsy of Unverricht and Lundborg	EPM1	254800	AR	*CSTB*	21q22.3	promoter	CCCCGCC CCGCG	2-3	40-80	-	45,196,324	3.1	100	0	37
Spinocerebellar ataxia 12	SCA12	604326	AD	*PPP2R2B*	5q32	promoter	CAG	7-45	55-78	-	146,258,291	10.7	100	0	32
Spinocerebellar ataxia 8	SCA8	608768	AD	*ATXN8OS/ATXN8*	13q21	utRNA	CTG	16-34	74+	+	70,713,516	15.3	100	0	46

Table 1 Detailed STR loci information. TRF = Tandem Repeats Finder (Benson et al, 1999). TRF match and TRF indel describe the purity of the repeat. AD = autosomal dominant, X = X-linked, AR = autosomal recessive

Class	MOI	Diagnosis	Allele sizes	Gender	WES_PCR	WGS_PCR_1	WGS_PCR_2	WGS_PF_3
PolyQ	AD	HD	Not recorded	male	rptWEHI3	HD-1		
PolyQ	AD	HD	17,39	female			WGSrpt_10	
PolyQ	AD	HD	20,42	male			WGSrpt_12	
PolyQ	AD	SCA1	36,52	female	rptWEHI4		WGSrpt_14	
PolyQ	AD	SCA1	30,45	male			WGSrpt_16	
PolyQ	AD	SCA2	21,42	female	rptWEHI1	SCA2-1	WGSrpt_18	
PolyQ	AD	SCA2	23,39	male			WGSrpt_20	
PolyQ	AD	SCA6	11,22	female	rptWEHI2	SCA6-1	WGSrpt_05	
PolyQ	AD	SCA6	10,21	female			WGSrpt_07	
PolyQ	AD	SCA7	13,39	female			WGSrpt_08	
5'UTR	X	FRAXA	Not sized	male			WGSrpt_17	WGSrpt_17_F
5'UTR	X	FRAXA	613-1680	male			WGSrpt_19	WGSrpt_19_F
5'UTR	X	FRAXA (pre)	~100	female			WGSrpt_21	WGSrpt_21_F
3'UTR	AD	DM1	8,173	female			WGSrpt_13	
3'UTR	AD	DM1	13,83	male			WGSrpt_15	
Intron	AR	FRDA	320,320	male			WGSrpt_09	
Intron	AR	FRDA	788,788	male			WGSrpt_11	
		(controls)			58	14	2	5

Table 2. Repeat type, genetic model, diseases, sample names and which cohorts samples appear in. Allele sizes are derived from standard laboratory tests for repeat expansions. Some individuals were not tested (Not sized) or the data was not available (not recorded). MOI = mode of inheritance (AD = autosomal dominant, X = X-linked, AR = autosomal recessive). Only the total number of controls are given denoted by (controls).

Cohort	Cases	Controls	Method	AUC	TP	FN	Sensitivity	Specificity	Tests
WES_PCR*	4	58	exSTRa	0.99	3	1	0.75	0.99	682
			ExpansionHunter	NA	2	2	0.5	0.99	682
WGS_PCR_1	3	14	exSTRa	0.99	2	1	0.67	0.99	357
			ExpansionHunter	NA	3	0	1	1	357
WGS_PCR_2	21	2	exSTRa	0.89	12	4	0.75	0.98	378
			ExpansionHunter	NA	9	7	0.56	1	378
WGS_PCR_2_30X_1	21	2	exSTRa	0.9	13	3	0.81	0.99	378
			ExpansionHunter	NA	9	7	0.56	1	378
WGS_PCR_2_30X_2	21	2	exSTRa	0.9	12	4	0.75	0.99	378
			ExpansionHunter	NA	9	7	0.56	1	378
WGS_PF	3	5	exSTRa	0.76	0	3	0	0.98	168
			ExpansionHunter	NA	0	3	0	1	168
All	73	83	exSTRa	NA	42	16	NA	NA	NA
	73	83	ExpansionHunter	NA	32	26	NA	NA	NA

Table 3 Repeat Expansion detection results for exSTRa and ExpansionHunter. AUC = area under the receiver operating characteristics (ROC) curve. TP = true positive, FN = false negative, Sensitivity = TP/(TP+FN), Specificity = TN/(FP+TN). NA = not applicable. WES cohort labeled with (*) only assessed over eleven STR loci in the capture design. WGS_PCR_2 was also analysed split into two sub-cohorts, split by flow cell lane, and are designated as WGS_PCR_2_30X_1 and WGS_PCR_2_30X_2.

Figures

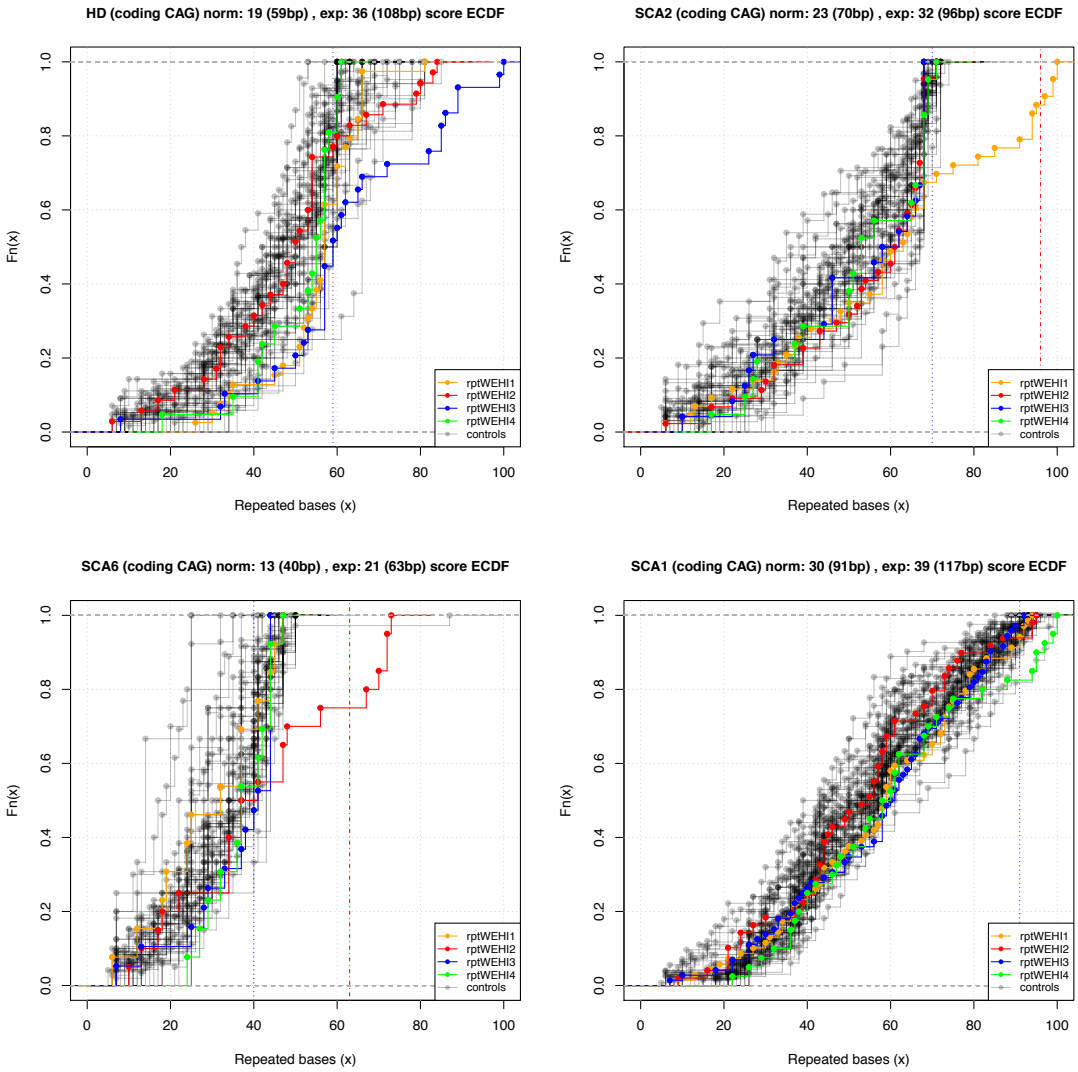

Figure 1 ECDF of repeat expansion composition of reads from the WES cohort, depicting four different known repeat expansion disorders captured by WES (HD, SCA2, SCA6 and SCA1). Sample rptWEHI3 (blue) is a known HD repeat expansion patient, sample rptWEHI1 (yellow) a known SCA2 repeat expansion, sample rptWEHI2 (red) a known SCA6 and sample rptWEHI4 (green) a known SCA1 patient. The title at the top of each individual figure gives the locus being examined, the reference number of repeats in the hg19 human genome reference with the corresponding number of bps, and the smallest reported expanded allele in the literature (with the corresponding number of bps in brackets). The blue dashed vertical line in the plot denotes the largest known normal allele, the red dashed vertical line denotes the smallest known expanded allele.

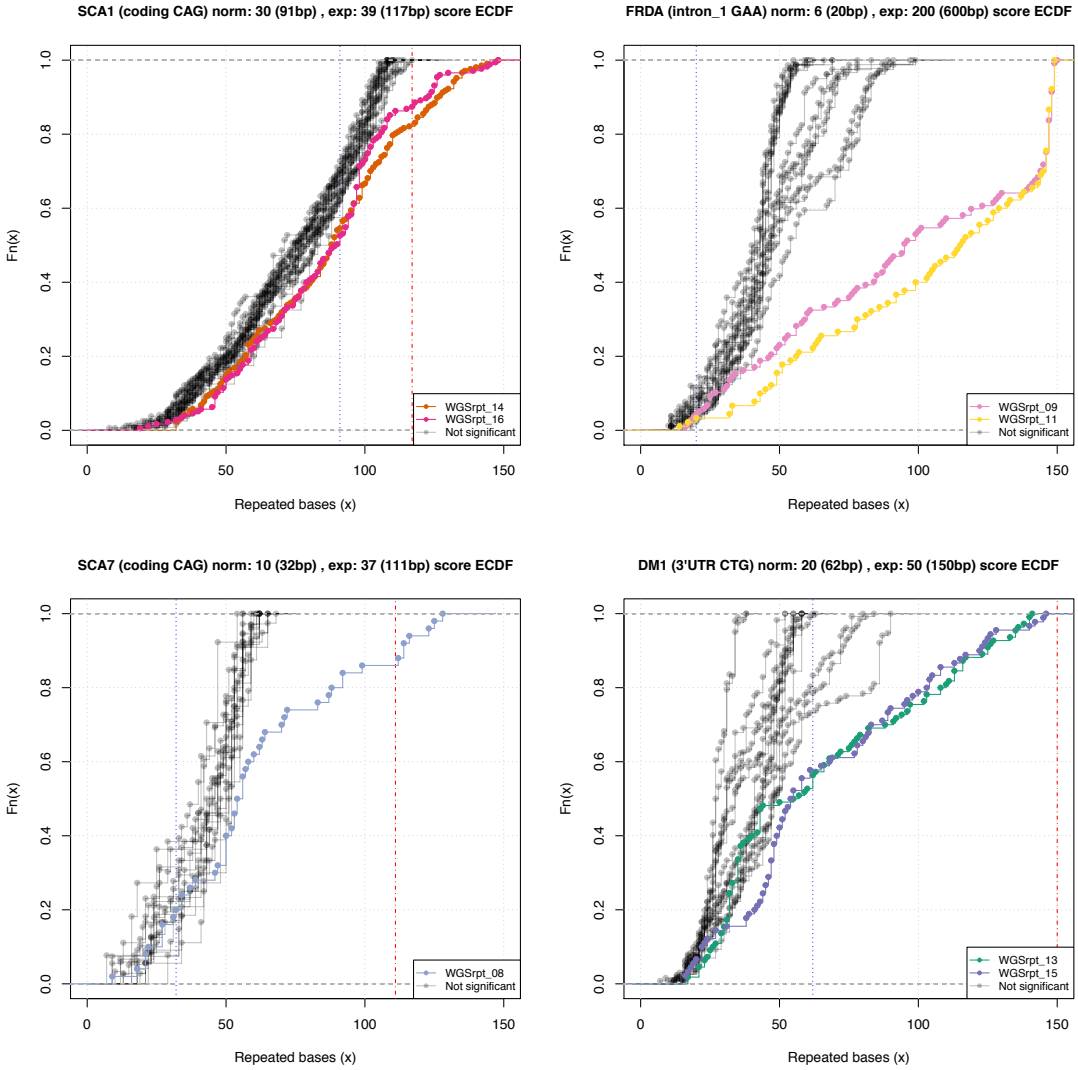

Figure 2 ECDFs of repeat expansion composition of reads from the WGS_PCR_2 cohort, depicting four different STR loci (top left = SCA1, top right = FRDA, bottom left = SCA7, bottom right = DM1). Here coloured samples at each STR indicate those called by exSTRa as repeat expansions at the STR locus. The title at the top of each individual figure gives the locus being examined, the reference number of repeats in the hg19 human genome reference with the corresponding number of bps, and the smallest reported expanded allele in the literature (with the corresponding number of bps in brackets). The blue dashed vertical line in the plot denotes the largest known normal allele, the red dashed vertical line denotes the smallest known expanded allele.

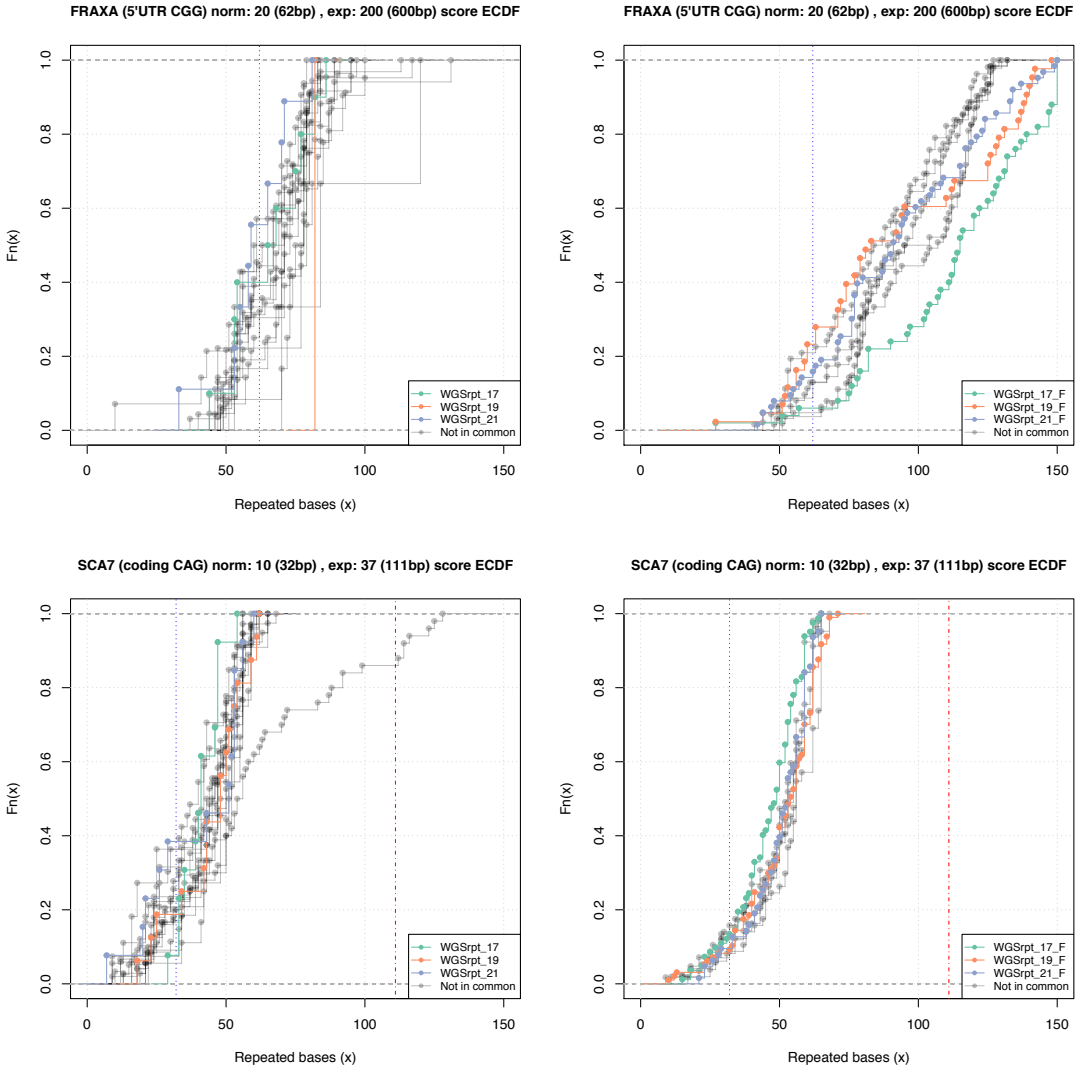

Figure 3 ECDFs comparing repeat expansion composition of reads between WGS from the PCR based library preparation protocol (left panels) and the WGS from the PCR free library preparation protocol (right panels) for two different repeat expansions, FRAXA (top panels) and SCA7 (bottom panels). Individuals in grey are not shared between the two cohorts. WGSrpt_17 (teal) and WGSrpt_19 (orange) are FRAXA repeat expansion patients, and WGSrpt_21 (blue) is a FRAXA pre-expansion patient. The title at the top of each individual figure gives the locus being examined, the reference number of repeats in the hg19 human genome reference with the corresponding number of bps, and the smallest reported expanded allele in the literature (with the corresponding number of bps in brackets). The blue dashed vertical line in the plot denotes the largest known normal allele, the red dashed vertical line denotes the smallest known expanded allele. "_F" in the sample names on the right hand side panels indicate that samples were generated with the PCR free library preparation method.

References

1. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res; 1999. p. 573-80.
2. Jones L, Houlden H, Tabrizi SJ. DNA repair in the trinucleotide repeat disorders. *The Lancet Neurology* 2017; **16**(1): 88-96.
3. Fondon IIIJW, Hammock EAD, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. *Trends in Neurosciences* 2008; **31**(7): 328-34.
4. Schulz JB, Boesch S, Bürk K, et al. Diagnosis and treatment of Friedreich ataxia: a European perspective. *Nat Rev Neurol* 2009; **5**(4): 222-34.

5. Seltzer MM, Baker MW, Hong J, Maenner M, Greenberg J, Mandel D. Prevalence of CGG expansions of the FMR1 gene in a US population-based sample. *Am J Med Genet B Neuropsychiatr Genet* 2012; **159B**(5): 589-97.
6. Tassone F. Newborn screening for fragile X syndrome. *JAMA Neurol* 2014; **71**(3): 355-9.
7. Genetic Modifiers of Huntington's Disease C. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* 2015; **162**(3): 516-26.
8. Bettencourt C, Hensman Moss DJ, Flower M, et al. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann Neurol* 2016.
9. Nemeth AH, Kwasniewska AC, Lise S, et al. Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. Brain; 2013.
10. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res; 2012. p. 1154-62.
11. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* 2017; **14**(6): 590-2.
12. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* 2013; **41**(1): e32.
13. Cao MD, Tasker E, Willadsen K, et al. Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res* 2013.
14. Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. 2014; **30**(5): 652-9.
15. Dolzhenko E, van Vugt J, Shaw RJ, Bekritsky MA. Detection of long repeat expansions from PCR-free whole-genome sequence data. bioRxiv; 2016.
16. Ewen KR, Bahlo M, Treloar SA, Levinson DF. Identification and Analysis of Error Types in High-Throughput Genotyping. *The American Journal of* ... 2000.
17. Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol; 2009. p. R25.
18. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res; 2012. p. e72.
19. Rehm HL. Evolving health care through personal genomics. *Nat Rev Genet* 2017; **18**(4): 259-67.
20. Ashley EA. The precision medicine initiative: a new national effort. *JAMA* 2015; **313**(21): 2119-20.
21. Hunter J, Rivero-Arias O, Angelov A, Kim E, Fotheringham I, Leal J. Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am J Med Genet* 2014; **164A**(7): 1648-58.
22. Ligsay A, Hagerman RJ. Review of targeted treatments in fragile X syndrome. *Intractable Rare Dis Res* 2016; **5**(3): 158-67.
23. Hagerman RJ, Polussa J. Treatment of the psychiatric problems associated with fragile X syndrome. *Curr Opin Psychiatry* 2015; **28**(2): 107-12.
24. Hagerman R, Hagerman P. Advances in clinical and molecular understanding of the FMR1 premutation and fragile X-associated tremor/ataxia syndrome. *The Lancet Neurology* 2013; **12**(8): 786-98.