

# Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data

Rick M. Tankard,<sup>1,2,3</sup> Mark F. Bennett,<sup>1,2,4</sup> Peter Degorski,<sup>1,2</sup> Martin B. Delatycki,<sup>5,6,7</sup> Paul J. Lockhart,<sup>5,7,8</sup> and Melanie Bahlo<sup>1,2,8,\*</sup>

Repeat expansions cause more than 30 inherited disorders, predominantly neurogenetic. These can present with overlapping clinical phenotypes, making molecular diagnosis challenging. Single-gene or small-panel PCR-based methods can help to identify the precise genetic cause, but they can be slow and costly and often yield no result. Researchers are increasingly performing genomic analysis via whole-exome and whole-genome sequencing (WES and WGS) to diagnose genetic disorders. However, until recently, analysis protocols could not identify repeat expansions in these datasets. We developed exSTRa (expanded short tandem repeat algorithm), a method that uses either WES or WGS to identify repeat expansions. Performance of exSTRa was assessed in a simulation study. In addition, four retrospective cohorts of individuals with eleven different known repeat-expansion disorders were analyzed with exSTRa. We assessed results by comparing the findings to known disease status. Performance was also compared to three other analysis methods (ExpansionHunter, STRetch, and TREDPARSE), which were developed specifically for WGS data. Expansions in the assessed STR loci were successfully identified in WES and WGS datasets by all four methods with high specificity and sensitivity. Overall, exSTRa demonstrated more robust and superior performance for WES data than did the other three methods. We demonstrate that exSTRa can be effectively utilized as a screening tool for detecting repeat expansions in WES and WGS data, although the best performance would be produced by consensus calling, wherein at least two out of the four currently available screening methods call an expansion.

## 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 (bp) in size. STRs were used as genetic markers for linkage mapping in human studies for many years; they continue to be used, but primarily in non-human studies. They are underrepresented in the coding regions of the human genome,<sup>1</sup> despite the fact that the vast majority are population polymorphisms of no, or very little, phenotypic consequence. A subset of STRs can, however, cause disease. These diseases are known as repeat-expansion disorders. Pathogenic STRs have either one or two alleles, depending on the genetic model, that exceed some threshold for biological tolerance. The abnormal STR allele(s) might affect gene expression levels, cause premature truncation of the protein, or result in aberrant protein folding.<sup>2</sup> Repeat expansions at different STR loci share biological consequences. Common disease mechanisms mediated by repeat-expansion disorders include repeat-associated non-AUG translation and MBNL spliceosome interference, caused by, for example, CUG expansions in myotonic dystrophy type 1 (DM1, MIM: 160900). These mechanisms are reviewed in Hannan.<sup>3</sup>

Repeat expansions cause ~30 inherited germline human disorders, predominantly neurogenetic diseases that most

often present with ataxia as a clinical feature. The size of a pathogenic allele varies from ~60 repeats, observed in the gene encoding the Calcium Voltage-Gated Channel Subunit Alpha1 A (*CACNA1A*), to several thousand repeats, observed in the gene encoding the guanine nucleotide exchange C9orf72 (*C9orf72*) (Table 1). 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 (MIM: 300624) (a repeat in the 5'-UTR of *FMRI*); Huntington disease (MIM: 143100) (a repeat in exon 1 of *HTT*); myotonic dystrophy (MIM: 160900, MIM: 602668) (repeats in *DMPK* and *ZNF9*); fronto-temporal dementia and amyotrophic lateral sclerosis 1 (MIM: 105550) (a 6-mer repeat in *C9orf72*); and Unverricht-Lundborg disease, a severe myoclonic epilepsy (MIM: 254800) (in *CSTB*). The genetic mode of inheritance encompasses autosomal dominant (e.g., SCA1, MIM: 164400) and recessive (e.g., Friedreich ataxia, MIM: 229300), as well as X-linked recessive (e.g., fragile X syndrome, MIM: 300624). Pathogenic alleles underlying repeat-expansion disorders continue to be discovered; the two most recently described STRs are pentamer repeats.<sup>4,5</sup> A selected list of repeat-expansion disorders is shown in Table 1.

Many repeat-expansion disorders show anticipation: a phenomenon whereby younger generations are affected by an earlier age of onset than are preceding generations.

<sup>1</sup>Population Health and Immunity Division, the Walter and Eliza Hall Institute of Medical Research, Parkville 3052, VIC, Australia; <sup>2</sup>Department of Medical Biology, The University of Melbourne, Melbourne 3010, VIC, Australia; <sup>3</sup>Mathematics and Statistics, Murdoch University, Murdoch 6150, WA, Australia; <sup>4</sup>Epilepsy Research Centre, Department of Medicine, The University of Melbourne, Austin Health, Heidelberg 3084, VIC, Australia; <sup>5</sup>Bruce Lefroy Centre for Genetic Health Research, Murdoch Children's Research Institute, Royal Children's Hospital, Parkville 3052, VIC, Australia; <sup>6</sup>Victorian Clinical Genetics Services, Parkville 3052, VIC, Australia; <sup>7</sup>Department of Paediatrics, University of Melbourne, Parkville 3058, VIC, Australia

<sup>8</sup>These authors contributed equally to this work

\*Correspondence: bahlo@wehi.edu.au

<https://doi.org/10.1016/j.ajhg.2018.10.015>

© 2018 American Society of Human Genetics.



**Table 1. Information about Short Tandem-Repeat Loci for STRs That Cause Neurogenetic Disorders**

Disease	Symbol	OMIM	Inheritance	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	<i>HTT</i>	4p16.3	Coding	CAG	6–34	36–100+	+	3,076,604	21.3	96	0	64
Kennedy disease	SBMA	313200	X	<i>AR</i>	Xq12	Coding	CAG	9–35	38–62	+	66,765,159	33.3	86	9	103
Spinocerebellar ataxia 1	SCA1	164400	AD	<i>ATXN1</i>	6p23	Coding	CAG	6–38	39–82	–	16,327,865	30.3	95	0	91
Spinocerebellar ataxia 2	SCA2	183090	AD	<i>ATXN2</i>	12q24	Coding	CAG	15–24	32–200	–	112,036,754	23.3	97	0	70
Machado-Joseph disease	SCA3	109150	AD	<i>ATXN3</i>	14q32.1	Coding	CAG	13–36	61–84	–	92,537,355	14	84	0	42
Spinocerebellar ataxia 6	SCA6	183086	AD	<i>CACNA1A</i>	19p13	Coding	CAG	4–7	21–33	–	13,318,673	13.3	100	0	40
Spinocerebellar ataxia 7	SCA7	164500	AD	<i>ATXN7</i>	3p14.1	Coding	CAG	4–35	37–306	+	63,898,361	10.7	100	0	32
Spinocerebellar ataxia 17	SCA17	607136	AD	<i>TBP</i>	6q27	Coding	CAG	25–42	47–63	+	170,870,995	37	94	0	111
Dentatorubral-pallidoluysian atrophy	DRPLA	125370	AD	<i>DRPLA/ATN1</i>	12p13.31	Coding	CAG	7–34	49–88	+	7,045,880	19.7	92	0	59
Huntington disease-like 2	HDL2	606438	AD	<i>JPH3</i>	16q24.3	Exon	CTG	7–28	66–78	+	87,637,889	15.3	95	4	47
Fragile-X site A	FRAXA	300624	X	<i>FMR1</i>	Xq27.3	5' UTR	CGG	6–54	200–1000+	+	146,993,555	25	90	5	75
Fragile-X site E	FRAXE	309548	X	<i>FMR2</i>	Xq28	5' UTR	CCG	4–39	200–900	+	147,582,159	15.3	100	0	46
Myotonic dystrophy 1	DM1	160900	AD	<i>DMPK</i>	19q13	3' UTR	CTG	5–37	50–10000	–	46,273,463	20.7	100	0	62
Friedreich ataxia	FRDA	229300	AR	<i>FXN</i>	9q13	Intron	GAA	6–32	200–1700	+	71,652,201	6.7	100	0	20
Myotonic dystrophy 2	DM2	602668	AD	<i>ZNF9/CNBP</i>	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	<i>C9orf72</i>	9p21	Intron	GGGGCC	2–19	250–1600	–	27,573,483	10.8	74	8	62

(Continued on next page)

**Table 1. Continued**

Disease	Symbol	OMIM	Inheritance	Gene	Cytogenetic Location	Type	Repeat Motif	Normal Range	Expansion Range	Strand	Start hg19	Reference Repeat Number	TRF Match (%)	TRF Indel (%)	Reference STR Size (bp)
Spinocerebellar ataxia 36	SCA36	614153	AD	<i>NOP56</i>	20p13	Intron	GGCCTG	3–8	1500–2500	+	2,633,379	7.2	97	0	43
Spinocerebellar ataxia 10	SCA10	603516	AD	<i>ATXN10</i>	22q13.31	Intron	ATTCT	10–20	500–4500	+	46,191,235	14	100	0	70
Myoclonic epilepsy of Unverricht and Lundborg	EPM1	254800	AR	<i>CSTB</i>	21q22.3	Promoter	CCCCGCC CCGCG	2–3	40–80	–	45,196,324	3.1	100	0	37
Spinocerebellar ataxia 12	SCA12	604326	AD	<i>PPP2R2B</i>	5q32	Promoter	CAG	7–45	55–78	–	146,258,291	10.7	100	0	32
Spinocerebellar ataxia 8	SCA8	608768	AD	<i>ATXN8OS/ ATXN8</i>	13q21	utRNA	CTG	16–34	74+	+	70,713,516	15.3	100	0	46
Spinocerebellar ataxia 31	SCA31	117210	AD	<i>BEAN1/ TK2</i>	16q21	Intron	TGGAA <sup>a</sup>	0	2.5–3.8kb <sup>b</sup>	+	66,524,302	0	N/A	N/A	N/A
Spinocerebellar ataxia 37	SCA37	615945	AD	<i>DAB1</i>	1p32.3	Intron	ATTTC <sup>c</sup>	0	31–75	–	57,832,716 <sup>c</sup>	0	N/A	N/A	N/A
Familial adult myoclonic epilepsy 1 <sup>e</sup>	FAME1/ BAFME1	601068	AD	<i>SAMD12</i>	8q24	Intron	TTTCA <sup>d</sup>	0	440–3,680 <sup>f</sup>	–	119,379,055 <sup>d</sup>	0	N/A	N/A	N/A

Abbreviations are as follows: TRF = Tandem Repeats Finder (Benson et al., 1999).<sup>1</sup> TRF match and TRF indel describe the purity of the repeat. AD = autosomal dominant; X = X-linked; AR = autosomal recessive; UTR = untranslated region.

<sup>a</sup>These repeat expansions are insertions and thus not represented in the reference genome at their respective locations.

<sup>b</sup>SCA31 is caused by the insertion of a complex repeat containing (TGGA)<sub>n</sub>; hence, the length is given as the length of the expanded repeats in base pairs, instead of the repeat number.

<sup>c</sup>The SCA37 physical map location is given at the reference (ATTT)<sub>n</sub> repeat, where affected individuals have the pathogenic (ATTTC)<sub>n</sub> inserted.

<sup>d</sup>The FAME1 physical map location is given as the position of the reference (TTTTA)<sub>n</sub> repeat, where affected individuals have (TTTCA)<sub>n</sub> inserted.

<sup>e</sup>Ishiyama et al. identified similar expansions associated with FAME6 and FAME7, in the genes TNRC6A and RAPGEF2, respectively, but only in single families.<sup>5</sup> These have not been listed.

<sup>f</sup>The FAME1 repeat size is the estimated size of the combined expanded (TTTCA)<sub>n</sub> and the (TTTTA)<sub>n</sub> reference repeats.

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.

With a disease prevalence of 3 to 4/100,000 and a carrier frequency of 1/100,<sup>6</sup> Friedreich ataxia is the most common of the recessive repeat-expansion disorders. Fragile X syndrome is the most common cause of inherited intellectual disability and affects ~1/5,000 individuals.<sup>7,8</sup> Hence, repeat-expansion disorders 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. Currently, a medical diagnosis is based on precise PCR or Southern-blot assays, 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 individual'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.<sup>9,10</sup> In addition, in up to 50% of individuals with a diagnosis of ataxia, the ataxia might be due to other mutation types, such as single-nucleotide variants (SNVs) or short insertions or deletions (indels).<sup>11</sup> Therefore, molecular diagnosis of these disorders often also requires conventional sequencing of candidate genes, either by Sanger sequencing, targeted panel sequencing, or next-generation sequencing (NGS) methods.

Short-read NGS data, such as those generated by the Illumina sequencing platform, are 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 and NovaSeq platforms are currently the most commonly used platforms for the generation of human WGS data and in particular, for clinical human genome sequencing with low error rates and well-documented, consistent performance.

Illumina HiSeq X data reads are paired, and the DNA insert is sequenced 150 bp inwards from each end; a small gap in the template DNA is predominantly not sequenced between reads. This gap can vary in size, but standard library preparation methodologies generate insertion fragment lengths of ~350 bp, resulting in a gap of ~50 bp.

Standard clinical diagnostic pipelines focus on the identification of SNVs and indels. Bioinformatic tools have been developed for genotyping STRs, but they are almost entirely confined to those STR alleles that are spanned by reads.<sup>12–16</sup> Pathogenic repeat expansions are usually significantly longer than the reads generated by short-read sequencing platforms such as Illumina and can be longer than the library insertion fragments. Therefore, the short

reads cannot span many pathogenic repeat-expansion alleles, such as those that cause SCA2 (MIM: 183090) or SCA7 (MIM: 164500, [Table 1](#)). Furthermore, some of these reads are not mapped, or are poorly mapped, to the STR allele as a result of sequencing bias and alignment issues such as (1) the repetitive nature of the repeat itself, such that the expanded alleles require alignments of additional repetitive bases; (2) multiple occurrences of the same repeat throughout the genome, leading to multi-mapping reads; and (3) GC bias. Despite this, these data do still carry information about the expanded allele; more reads map to the STR for an expanded allele than would be expected on the basis of the reference STR allele lengths.

Several methods now describe the detection of repeat expansions in short-read NGS data. These include ExpansionHunter,<sup>17</sup> STRetch<sup>18</sup> and TREDPARSE,<sup>19</sup> reviewed in Bahlo et al.<sup>20</sup> These methods are focused on the detection of repeat expansions in WGS data; there is a bias toward PCR-free library protocols. ExpansionHunter and TREDPARSE use pre-determined thresholds to determine whether an individual has an expansion; however, TREDPARSE also has a likelihood ratio test, which uses a framework that determines the genetic model and the likelihood of expansion. STRetch uses a genome reference augmented with decoy chromosomes, consisting of long stretches of all 1–6 bp repeat expansions to competitively attract long repeats. None of these methods has been assessed for performance in comparison to each other or to WES data.

Here, we describe the development of the STR repeat-expansion-calling algorithm, exSTRa (expanded STR algorithm), which detects expanded repeat expansion allele(s) at repeat expansion loci, specified by the user, in cohorts of sequenced individuals. We demonstrate the utility of the method with 12 different verified repeat-expansion disorders. exSTRa is designed to be applied to cohorts of individuals without requiring a set of controls. This is because exSTRa is designed as an outlier detection test, wherein the majority of individuals (>85%) are assumed to have normal-length alleles at a particular repeat expansion locus. This assumption is robust for the majority of disease cohorts, even those with spinocerebellar ataxias. exSTRa also generates unique empirical cumulative distribution function (ECDF) plots of individuals' repeat-motif distributions, plotted for all individuals in a cohort; this facilitates QC for batch effects and validity of assumptions. We demonstrate that repeat-expansion detection is possible with WES data and further demonstrate on additional STR loci that WGS data resulting from the preparation of PCR-based libraries, although inferior to that resulting from the preparation of PCR-free libraries, can be used for confident interrogation of most known STR loci. Additionally, we show that a consensus call based on the detection of a repeat expansion at a particular STR locus by at least two out of the four repeat-expansion tools increases sensitivity and specificity. This will enable researchers to interrogate the



**Table 2. Repeat Type, Genetic Model, Diseases, Sample Names, and which Cohorts Samples Appear in**

Class	MOI	Diagnosis	Allele sizes	Gender	WES_PCR	WGS_PCR_1	WGS_PCR_2
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
5' UTR	X	FRAXA	613-1680	male			WGSrpt_19
5' UTR	X	FRAXA (pre)	~100	female			WGSrpt_21
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

Allele sizes are derived from standard laboratory tests for repeat expansions. Some individuals were not tested (not sized), or the data were not available (not recorded). MOI = mode of inheritance, AD = autosomal dominant, X = X-linked, AR = autosomal recessive. Only the total number of controls is given, denoted by (controls). See [Table S5](#) for further sample details.

thousands of existing NGS datasets for repeat expansions at known repeat loci or any other loci they wish to investigate.

## Subjects and Methods

### Study Cohorts and Generation of Next-Generation Sequencing Data

Individuals with already-diagnosed repeat-expansion disorders were recruited for this study. The repeat-expansion status was verified via standard diagnostic 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 because 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).

Four cohorts underwent different types of NGS, and some individuals were sequenced multiple times. Individuals were sequenced with (1) WES with the Agilent V5+UTR capture platform (58 controls and four repeat-expansion individuals with four different expansion disorders); (2) WGS with the TruSeq Nano protocol, which includes a PCR step to increase sequencing material (16 controls and 17 repeat-expansion individuals with eight different expansion disorders); or (3) WGS with the PCR-free cohort consisting of 118 individuals (52 females and 66

males). Most individuals in this cohort were either carriers for or affected with one of these repeat-expansion disorders: FRAXA (16 expanded, 18 intermediate), FRDA (25), DM1 (17), HD (13), SCA1 (3), DRPLA (2), SBMA (1), and SCA3 (1). 22 individuals were relatives with no known expansion. All cell lines were sourced from the Coriell resource by Illumina, who performed the sequencing, as described in Dolzhenko et al (2017).<sup>17</sup> The WES cohort is designated as WES\_PCR. Two different cohorts were sequenced with protocol (2). These are designated as WGS\_PCR\_1 and WGS\_PCR\_2. The PCR-free WGS cohort was designated as WGS\_PF. These cohorts are described in [Table 2](#) and [Table 3](#).

### Generation of Sequencing Data

WGS data with PCR (WGS\_PCR\_1 and WGS\_PCR\_2) were generated by the Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, Australia, and sequenced on a HiSeq X Ten sequencer. The WES data (WES\_PCR) were generated by the Australian Genome Research Facility, Melbourne, Australia, and sequenced on a HiSeq 2500 sequencer. All WGS\_PF samples were sequenced on the Illumina HiSeq X sequencing platform at Illumina, La Jolla, California, USA. Further details can be found in Dolzhenko et al.<sup>17</sup> All sequencing data were aligned to the hg19 human genome reference using the Bowtie 2 aligner<sup>21</sup> in local alignment mode.

### Definition of Repeat Expansion Loci

[Table 1](#) defines the chromosomal location, physical map location, disease, genetic disease model, repeat motif, and normal and

**Table 3. Repeat Type, Genetic Model, and Number of Samples Associated with Each Diagnosis for WGS\_PF Cohort**

Class	MOI	Diagnosis	Expanded	Affected	Not Expanded
PolyQ	AD	HD	13	13	105
PolyQ	AD	SCA1	3	3	115
PolyQ	AD	SCA3	1	1	117
PolyQ	AD	DRPLA	2	2	116
PolyQ	AD	SBMA	1	1	117
5' UTR	X	FRAXA	16	16	102
5' UTR	X	FRAXA (pre)	34	21	84
3' UTR	AD	DM1	17	17	101
Intron	AR	FRDA	25	14	93
		Total (FRAXA) <sup>a</sup>	78	66	40
		Total (FRAXA pre)	96	72	22
		Total (no FRAXA)	62	51	56

The WGS\_PF cohort consists of 118 individuals sequenced with Illumina PCR-free library preparation. Only the total number of samples is listed, rather than actual samples. Details of samples are available in Dolzhenko et al., 2017.<sup>17</sup>

<sup>a</sup>Total only includes FXS individuals, and no intermediate pre-expansions. The “expanded” and “not expanded” entries add to 118 for each row.

repeat expansion sizes for 24 repeat-expansion loci, which cause neurological disorders. For the analyses in this paper, we examined 21 of these STR loci, but we excluded the more recently discovered SCA37 and FAME1 loci. We also excluded the SCA31 locus because the inserted repeat is not in the reference sequence. This focused the analysis on the expansion loci currently most likely to be tested, and, in particular, concentrated on the repeat-expansion loci associated with spinocerebellar ataxia.

### Data Extraction for Repeat Expansions

We developed a two-step analysis method called exSTRa, detailed in the [Supplemental Data](#), to identify individuals likely to have a repeat expansion at a particular STR locus. For each read, the analysis method extracts STR repeat content information that stems from a particular individual and 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 who is to be tested within a cohort of affected individuals and controls. All N individuals within a cohort are examined in turn at each of the 21 known pathogenic repeat-expansion loci via comparison of each individual in turn to all N individuals in each cohort. This generates 21N test statistics per cohort. The empirical p values of the test statistics were determined via 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 were visualized via empirical cumulative distribution functions (ECDFs), which display as a step function the distribution, from smallest to largest, of the amount of STR repeat motif found in each read. 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 when

the individual with the repeat expansion is compared to individuals with normal alleles.

### Simulation Study

We conducted a simulation study by using the NGS data simulation package ART,<sup>22</sup> which simulates NGS data with realistic error profiles based on supplied reference genomes. Alleles at STR loci were simulated from reference genomes in which alleles (normal, intermediate, and expanded) had been inserted. Some STR loci do not have an intermediate range, or they have only a very narrow intermediate range. One such example is the HD repeat expansion, which has no intermediate range.

We extensively searched the literature to determine pathogenic and non-pathogenic ranges of STR length alleles. We only used the “overall” distribution and ignored any ethnic specificity for these loci. Applying a stutter model in the simulations was not feasible because of ART’s constraints. We simulated data for 20 STR loci (excluding FAME1, SCA31, SCA36, and SCA37), for 200 controls and 30 affected individuals: ten normal range, ten intermediate range, and ten expanded range. These 30 affected individuals were tested for expansions. The STR genotype for the controls was randomly chosen on the basis of the distributions of these as described in the literature ([Table S3](#)). The ten alleles simulated for each of the normal, intermediate, and expanded alleles were chosen on the basis of uniform distances between alleles; distances covered the known normal, intermediate, and expanded allele ranges as described in the literature ([Table S3](#)). For autosomal-dominant loci, the second allele was chosen randomly by the same method as for the controls. For the recessive STR loci EPM1 and FRDA, we sampled two expanded alleles for individuals with the diseases. To allow for STR loci assessment on the X chromosome (FRAXA, FRAXE, and SBMA), we generated half of the samples as male and the other half as female; the males had a single X chromosome and hence a single STR allele. For the X chromosome STR loci, we only investigated

the male individuals. To investigate the effect of control sample size on detection with exSTRA, we sub-sampled the control cohort at intervals of 50; control cohort sizes ranged from 50 to 200 individuals. The ART command used for generating the simulated data was as follows:

```
art_illumina -i $file -p -na -l 150 -f 50 -m 450 -s 50 -o
$outfile/$base -1 $profiles/HiSeqXPCRfreeL150R1.txt -2
$profiles/HiSeqXPCRfreeL150R2.txt
```

## Performance Evaluation

For exSTRA, we labelled individuals as being normal or expanded on the basis of the Bonferroni multiple-testing-corrected p values derived from our empirical p values. The number of Bonferroni corrections performed was based on the 21 STRs tested per individual for the WGS cohorts but was based on the ten STRs tested for the WES cohort. Repeat-expansion calls were compared to the known disease status. We evaluated performance of all four methods by examining the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), as well as sensitivity, which is defined as  $TP/(TP+FN)$ , and specificity, which is defined as  $TN/(TN+FP)$ , at each STR; these were then summarized across the STR loci within cohorts.

## Comparison to ExpansionHunter, STRetch, and TREDPARSE

ExpansionHunter<sup>17</sup> estimates the repeat size by using a parametric model but does not attempt to call repeat expansions in a probabilistic framework. ExpansionHunter was used for determining whether alleles were larger than the currently known smallest disease-causing repeat-expansion alleles. STRetch<sup>18</sup> was used for detecting the presence of repeat expansions via its statistical test, which is also an outlier-detection test. Bonferroni corrections were calculated as per the exSTRA analysis. TREDPARSE<sup>19</sup> was used both for estimating the repeat size and for detecting the presence of an expansion on the basis of its likelihood model. Bonferroni corrections were applied in the same way as for exSTRA.

## Results

### Results of the Simulation Study

The simulation study of the 20 STR loci provides evidence of the validity and robustness of the exSTRA test statistic with respect to control-cohort size, repeat-expansion size, and known expansion status. Decreasing the control cohort in exSTRA showed that results were robust as the control sample size decreased (Figure S5). exSTRA also showed consistent results when the size of the repeat-expansion allele varied; longer expansion alleles achieved smaller p values (Figure S4). Overall, exSTRA p values showed adequate type 1 error and good discriminatory ability between expansion and non-expansion individuals (Figures S3 and S4). The ECDF plots, which are unique to exSTRA, show the effect of increasing expansion size in all STRs through commensurate right shifts of the distributions. The ECDFs also allow heuristic determination of the genetic model, showing larger shifts to the

right for the recessive FRDA STR and the X-linked STRs (FRAXA, FRAXE, and SBMA). Dominant loci only show the shift in the upper half of the ECDF plots (Figure S2). All STR loci, including FRAXA and FRAXE, performed well for repeat-expansion detection in the simulation studies.

## Coverage and Alignment Results for Study Cohorts

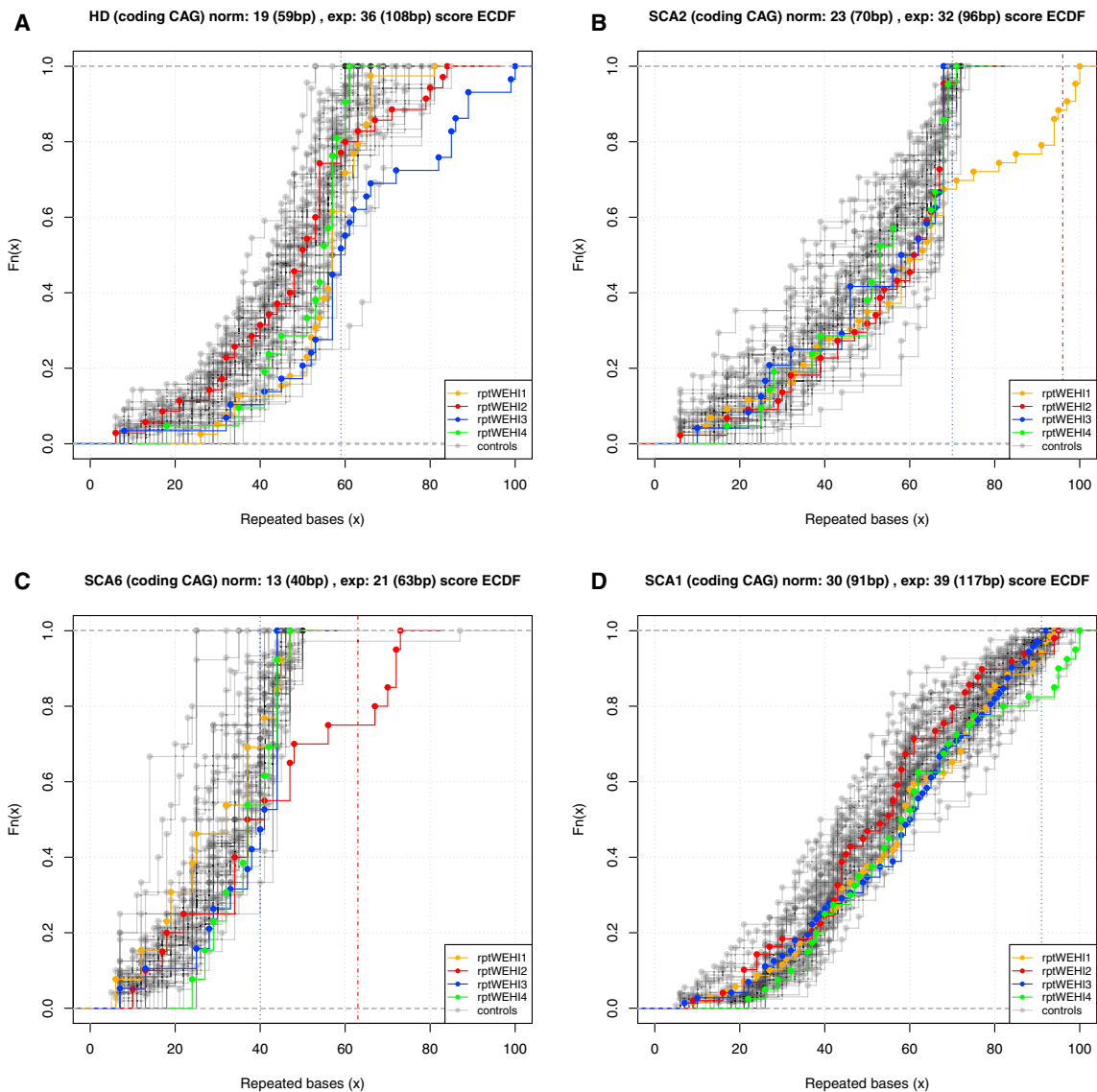
The full coverage and alignment results are in Table S2 for three cohorts, but not for WGS\_PF, which is described in Dolzhenko et al.<sup>17</sup> The median coverage achieved was 44, 66, 82, and 46.3 for cohorts WES, WGS\_PCR\_1, WGS\_PCR\_2, and WGS\_PF, respectively, and there was 1<sup>st</sup> and 3<sup>rd</sup> quartile coverage of (37, 48.25), (49.5, 71), (76.5, 84), and (44.9, 47.9). 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, and WGS\_PF median IQR = 8.3). In contrast, the WES data showed substantial variability (median IQR = 22.3).

## STR Loci Sequencing Coverage Ability

We 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 subset included many of the known repeat-expansion STRs located in coding regions (eight out of ten) (Figure 1 and Figure S1). SCA6 (CACNA1A, MIM: 183086) and SCA7 (ATXN7, MIM: 164500) were 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 UTR repeat-expansion loci, such as FRAXA (MIM: 300624), FRAXE (MIM: 309548), and DM1 (MIM: 160900). DM1 and SCA7 are not captured by the Agilent enrichment platform (Table S4); however, both FRAXA and FRAXE are targeted and therefore should be captured. In general, WGS data outperformed WES data over all STR loci, with one exception: SCA3 (MIM: 109150), located in the coding region of ATXN3. The reason for this is currently unknown.

## Visualizations of Repeat-Motif Distributions

ECDF curves for selected loci are shown for each cohort. Full results for all 21 loci, for all WGS cohorts, and for the ten loci with sufficient read coverage in the WES cohort are given in Figures S6–S11. STR loci varied in their coverage, and several loci were consistently poorly captured. These were usually loci that are rich in GC content. Short-read NGS data have a known GC bias, and a GC content of 40%–55% maximizes sequencing yield, depending on the sequencing platform.<sup>23</sup> The shape of the ECDF is affected by additional factors, such as the genetic model (dominant, recessive, or X-linked) and capture efficiency (for WES).



**Figure 1. ECDF of Repeat-Expansion Composition of Reads from the WES Cohort**

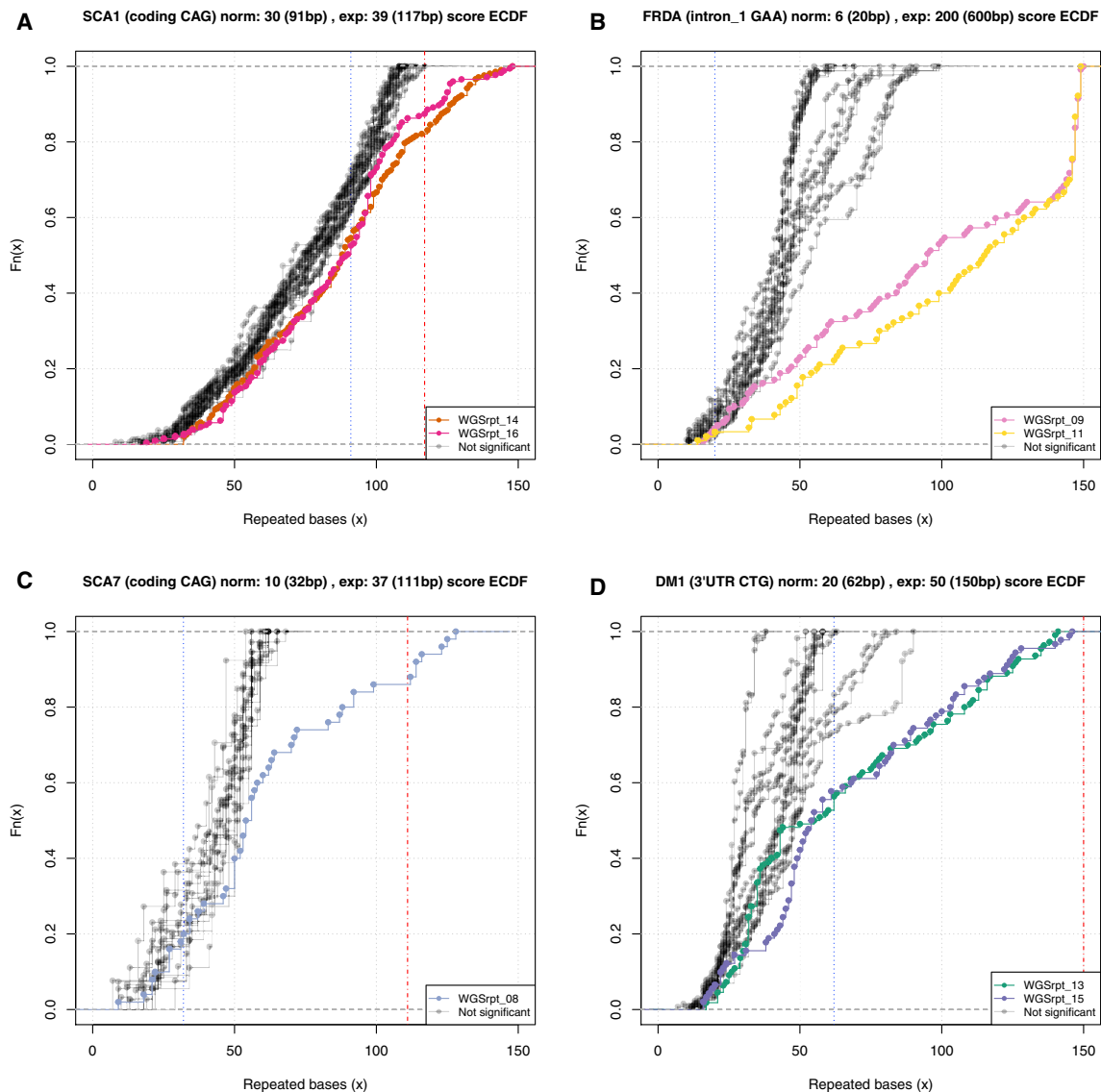
Four different known repeat-expansion disorders captured by WES are shown: (A) HD, (B) SCA2, (C) SCA6, and (D) SCA1. Sample rptWEH3 (blue) is from an individual with a known HD repeat expansion. The expanded allele size is not known. Sample rptWEH1 (yellow) is a known SCA2 repeat expansion of length 42 repeats, sample rptWEH2 (red) is a known SCA6 repeat expansion of length 22 repeats, and sample rptWEH4 (green) is a known SCA1 repeat expansion of length 52 repeats. 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 and the corresponding number of base pairs; and the smallest reported expanded allele in the literature (the corresponding number of base pairs is given in brackets). The blue dashed vertical line in the plot denotes the largest known normal allele and the red dashed vertical line denotes the smallest known expanded allele.

The STR loci also showed differences in variability with regard to STR motif lengths. Some STR loci, such as SCA17 (MIM: 607136) and HDL2 (MIM: 606438), showed little variability in STR allele distributions in our cohorts, regardless of NGS platform. Identification of outliers is easier for these loci because of their low normal population variability. Those repeat-expansion disorders that are autosomal recessive or X-linked recessive (in males) also show much clearer outlier distributions (Figure 2B). This is because the outlier distribution deviates for either both alleles or, in the case of the X chromosome in males only,

just the one examined allele (not performed in this analysis).

### Results of exSTRa Statistical Test

Test statistics were generated with exSTRa for all 21 loci for all N individuals in all four cohorts. Combined p values over all STR loci for all individuals within each cohort showed approximately uniform distribution with histograms (Figure S12) and Q-Q plots (Figure S13), albeit with some inflation of p values at both tails. Our study cohorts



**Figure 2. ECDFs of Repeat-Expansion Composition of Reads from the WGS\_PCR\_2 Cohort**

Four different STR loci are shown: (A) SCA1 (lengths of the expanded alleles are 52 and 45 repeats); (B) FRDA (lengths of the expanded alleles are 320 and 788 repeats); (C) SCA7 (length of the expanded allele is 39 repeats); and (D) DM1 (lengths of the expanded alleles are 173 and 83 repeats). Colored samples are those called by exSTRa as repeat expansions at the STR locus. The blue dashed vertical line in the plot denotes the largest known normal allele and the red dashed vertical line denotes the smallest known expanded allele.

had very small numbers of control individuals for some of the cohorts.

### Results of Expansion Calls

Results of expansion calls are presented in summary form in Table 4 and at the individual level in Tables S4 and S5. For the cohorts WES\_PCR, WGS\_PCR\_1, WGS\_PCR\_2, WGS\_PCR\_2\_30X\_1, WGS\_PCR\_2\_30X\_2, and WGS\_PF, exSTRa achieved sensitivities of 1, 0.67, 0.81, 0.81, 0.75, and 0.77, respectively (Table 4), with very high specificity (all cohorts > 0.97). Sensitivity is poorly estimated because the fact that there are a small number of true positives (TPs) in some cohorts leads to large variability. This is particularly the case for WES\_PCR (four cases) and WGS\_PCR\_1 (three cases). This has also resulted in

highly variable results for the other methods. FRAXA was the STR most refractory to analysis and performed poorly regardless of sequencing platform and repeat-expansion detection method. Excluding this locus in the evaluation of WGS\_PF increased the sensitivity from 0.77 to 0.84, but specificity remained unchanged at 0.97.

We divided the WGS\_PCR\_2 cohort data into two sub-cohorts. Each sample's data comes from a single-flow cell lane that has ~30× coverage. This allowed both an investigation of reproducibility and assessment at the more standard 30× coverage. Results were highly reproducible between the two 30× replicates; only one sample generated an alternative call between the two sequencing runs. We also observed very little change in performance

**Table 4. Repeat-Expansion Detection Results for All Four Cohorts**

Cohort	Affected Individuals	Controls <sup>a</sup>	Method	TP	FN	TN	FP	Sensitivity	Specificity
WES_PCR <sup>b</sup>	4	58	exSTRa	4	0	607	9	1	0.99
			ExpansionHunter	2	2	616	0	0.5	1
			STRetch <sup>c</sup>	3	1	613	3	0.75	1
			TREDPARSE-T <sup>d</sup>	4	0	585	31	1	0.95
			TREDPARSE-L <sup>d</sup>	4	0	574	42	1	0.93
WGS_PCR_1	3	14	exSTRa	2	1	343	11	0.67	0.97
			ExpansionHunter	3	0	354	0	1	1
			STRetch <sup>c</sup>	1	2	336	1	0.33	1
			TREDPARSE-T <sup>d</sup>	3	0	354	0	1	1
			TREDPARSE-L <sup>d</sup>	3	0	354	0	1	1
WGS_PCR_2 <sup>e</sup>	16	2	exSTRa	13	3	352	10	0.81	0.97
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch <sup>c</sup>	11	5	338	6	0.69	0.98
			TREDPARSE-T <sup>d</sup>	12	4	362	0	0.75	1
			TREDPARSE-L <sup>d</sup>	11	5	362	0	0.69	1
WGS_PCR_2_30X_1 <sup>e</sup>	16	2	exSTRa	13	3	357	5	0.81	0.99
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch <sup>c</sup>	11	5	340	4	0.69	0.99
			TREDPARSE-T <sup>d</sup>	13	3	362	0	0.81	1
			TREDPARSE-L <sup>d</sup>	9	7	362	0	0.56	1
WGS_PCR_2_30X_2 <sup>e</sup>	16	2	exSTRa	12	4	354	8	0.75	0.98
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch <sup>c</sup>	11	5	336	8	0.69	0.98
			TREDPARSE-T <sup>d</sup>	13	3	362	0	0.81	1
			TREDPARSE-L <sup>d</sup>	10	6	362	0	0.62	1
WGS_PF	78	40	exSTRa	60	18	2329	71	0.77	0.97
			ExpansionHunter <sup>f</sup>	62	16	2394	6	0.79	1
			STRetch <sup>c</sup>	62	16	2206	76	0.79	0.97
			TREDPARSE-T <sup>d</sup>	52	26	2383	17	0.67	0.99
			TREDPARSE-L <sup>d</sup>	34	32	2396	16	0.52	0.99
WGS_PF (FRAXA pre) <sup>g</sup>	96	22	exSTRa	63	33	2314	68	0.66	0.97
			ExpansionHunter <sup>f</sup>	95	1	2374	8	0.99	1
			STRetch <sup>c</sup>	62	34	2188	76	0.65	0.97
			TREDPARSE-T <sup>d</sup>	72	24	2364	18	0.75	0.99
			TREDPARSE-L <sup>d</sup>	48	24	2383	23	0.67	0.99
WGS_PF (no FRAXA) <sup>g</sup>	62	56	exSTRa	52	10	2231	67	0.84	0.97
			ExpansionHunter <sup>f</sup>	61	1	2292	6	0.98	1
			STRetch <sup>c</sup>	62	0	2104	76	1	0.97

*(Continued on next page)*



**Table 4. Continued**

Cohort	Affected Individuals	Controls <sup>a</sup>	Method	TP	FN	TN	FP	Sensitivity	Specificity
			TREDPARSE-T <sup>d</sup>	52	10	2281	17	0.84	0.99
	51	67	TREDPARSE-L <sup>d</sup>	34	17	2293	16	0.67	0.99

<sup>a</sup>Individuals designated as controls have no known repeat expansions. Individuals designated as affected have one known repeat expansion but are controls for all other loci tested. TP = true positive; FN = false negative; N = true negative; FP = false positive; sensitivity = TP/(TP+FN); and specificity = TN/(FP+TN).

<sup>b</sup>This WES cohort was only assessed over ten STR loci in the capture design.

<sup>c</sup>STRetch was Bonferroni corrected for the same number of tests as the other methods and not genome-wide corrected.

<sup>d</sup>TREDPARSE results are given for the repeat-expansion-size-threshold method (TREDPARSE-T) and for the likelihood-ratio-test-based method (TREDPARSE-L). For STR loci with recessive inheritance, samples with double expansions were designated as cases for TREDPARSE-L, which considers the inheritance model.

<sup>e</sup>WGS\_PCR\_2 was analyzed, then split into two sub-cohorts divided by flow cell lane and designated as WGS\_PCR\_2\_30X\_1 and WGS\_PCR\_2\_30X\_2.

<sup>f</sup>For the WGS\_PF cohort, the original ExpansionHunter results from Dolzhenko et al.<sup>17</sup> were used, which make use of reads aligned with a different aligner.

<sup>g</sup>For the WGS\_PF cohort, we computed additional results for "WGS\_PF (FRAXA pre)," using the permutation threshold to test for FRAXA expansions with ExpansionHunter, TREDPARSE-T, and TREDPARSE-L and to classify which samples had expansions at the FRAXA locus; we also computed results for "WGS\_PF (no FRAXA)," removing FRAXA from the list of loci tested. See Tables S5 and S6 for individual sample results.

between the 60× and 30× data, which had virtually identical sensitivity and specificity (Table 4).

### Comparison to Other Repeat-Expansion Detection Methods

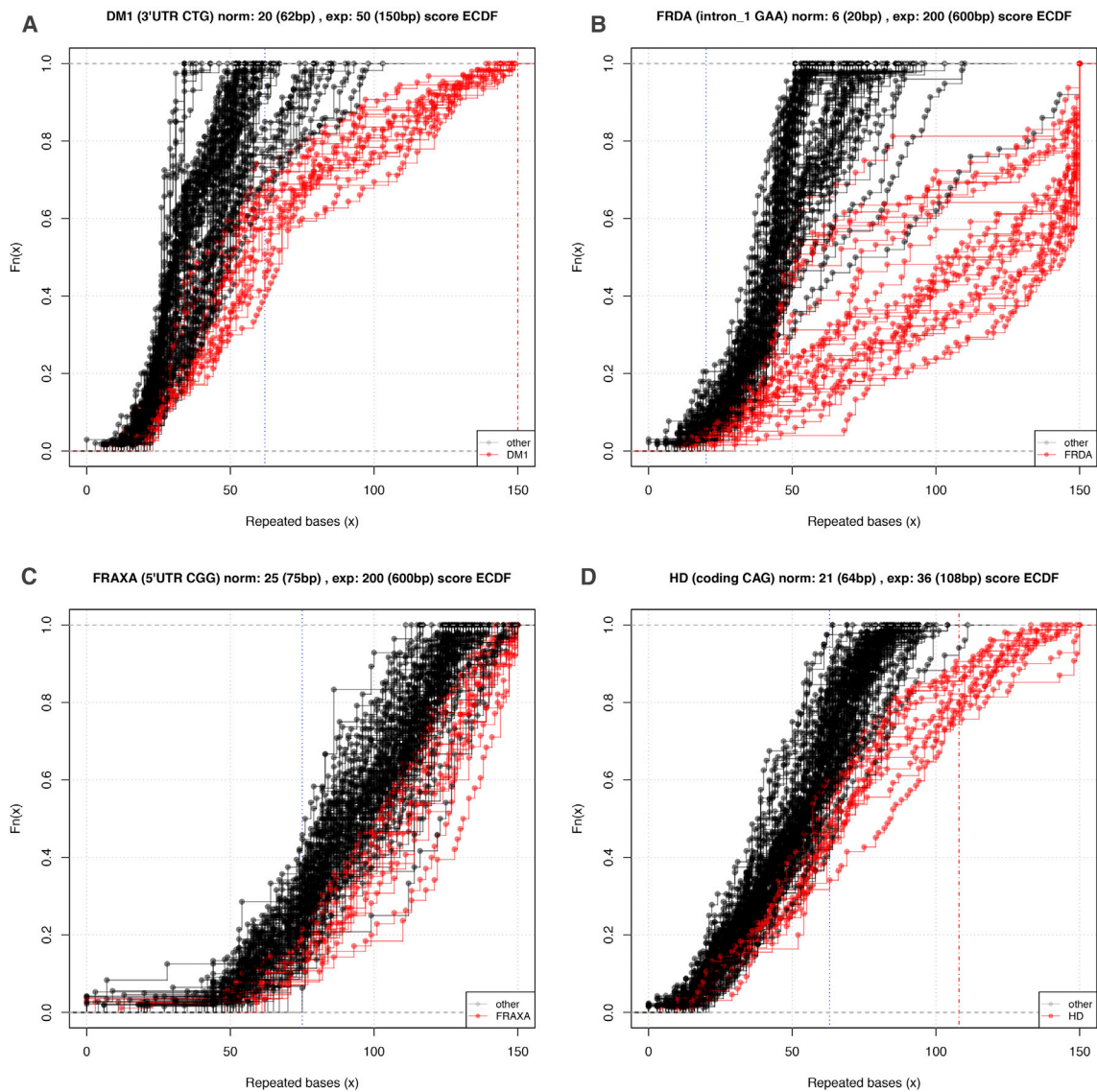
Across all cohorts (WES\_PCR, WGS\_PCR\_1, WGS\_PCR\_2, and WGS\_PF), exSTRa called more expansions (79 out of 101 known expansions) than ExpansionHunter (75 expansions), STRetch (77 expansions), TREDPARSE-L (52 expansions), and TREDPARSE-T (71 expansions), albeit with slightly different results in the REs identified. Excluding FRAXA, exSTRa called 71 out of 82 (87%) expansions, ExpansionHunter called 74 expansions, STRetch called 77 expansions, TREDPARSE-L called 52 expansions, and TREDPARSE-T called 71 expansions. Notably, exSTRa was able to identify expanded repeats at all eleven STR expansions examined, which included FRAXA. STRetch was unable to identify the SCA6 expansions in any cohort ( $n = 2$  in WGS\_PCR\_2,  $n = 1$  in WGS\_PCR\_1, and  $n = 1$  in WES\_PCR). SCA6 is the shortest of all known repeat expansions. These shorter expansions fail to map preferentially to the decoy chromosome for the most part, leading to the inability to call this locus. This will also apply to other short repeat expansion alleles. However, the other methods found most of the SCA6 expansions, regardless of sequencing platform. All four methods performed poorly when analyzing samples with an *FMR1* expansion (FRAXA). In the WGS\_PCR\_1 and 2 cohorts, this is due to poor coverage at the FRAXA and FRAXE loci caused by GC bias issues (Figure S1). Although there was a clear right shift of the exSTRa ECDF plots of both the full mutation and pre-mutation *FMR1* samples (Figure 3C), this was not always statistically significant. The other methods similarly performed poorly with this expansion, often failing to detect it. However, ExpansionHunter and TREDPARSE-T and -L identified pre-mutation alleles for this locus ~75% of the time. exSTRa identified 8 of 16 (50%) FRAXA expansions but STRetch identified none and called three of these as SCA3 expansions instead. STRetch performed equally as well as ExpansionHunter in the WGS\_PF cohort, but was the best performer once FRAXA was ignored, then finding all remaining repeat expansions, albeit with the highest

false positive rate. TREDPARSE and STRetch both perform particularly well for large expansions because their use of "in-repeat reads"<sup>17,20</sup>, or reads that map entirely to the repeat, is highly advantageous. exSTRa does not use this information, and ExpansionHunter only uses it optionally, for large repeats. Remarkably, all four methods called all 13 HD expansions correctly in the WGS\_PF cohort (Table S6), suggesting highly robust detection of HD expansions for WGS data. The four methods also unanimously identified the SBMA expansion and the two DRPLA expansions.

exSTRa performed equally as well as TREDPARSE for the WGS\_PCR cohorts, and it performed best overall for the WES cohort. Overall, all methods performed more poorly in the WES and WGS\_PCR cohorts than in the WGS\_PF cohort. exSTRa performs well for small repeat expansions and for platforms where small read fragments have been preferentially selected (WES\_PCR and WGS\_PCR). Overall, the results indicate that no single method is optimal over this breadth of sequencing library preparations and STR loci. These results suggest that a consensus call that makes use of all existing methods could be advantageous. Concordance with at least one other method, out of the four methods that are currently available for the detection of repeat expansions, will be useful for maximizing detection of expansions, especially because specificity is high in all WGS cohorts across all methods ( $\geq 0.97$ ). The specificity drops to  $\geq 0.93$  for WES data. Using a rule whereby at least two expansion calls are required and at least two calling methods must show concordant results to calculate a consensus call leads to sensitivities of 1 for WES\_1, 1 for WGS\_PCR\_1, 0.81 for WGS\_PCR\_2 (1, if FRAXA is excluded), 0.77 for WGS\_PF, and 0.84 for WGS\_PF (excluding FRAXA) (Tables S4 and S5, last columns).

Computational expense varied between the different repeat-expansion tools. For the WGS\_PF cohort comprising 118 samples, running time when 8 CPUs were used was approximately 0.5 hours for exSTRa with  $10^4$  permutations (12.6 hours for  $10^6$  permutations), 0.6 hr for ExpansionHunter, 1.6 hr for TREDPARSE, and 2,300 hr for STRetch. STRetch requires that data be realigned to their custom reference genome, which comprises the majority of computation time, and also creates additional





**Figure 3. ECDFs of Repeat-Expansion Composition of Reads from the WGS\_PF Cohort**

(A) DM1, (B) FRDA, (C) FRAXA, and (D) HD. 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 and the corresponding number of base pairs; and the smallest reported expanded allele in the literature (the corresponding number of base pairs is given in brackets). The blue dashed vertical line in the plot denotes the largest known normal allele and the red dashed vertical line denotes the smallest known expanded allele.

data storage requirements. We summarize the computational costs, capabilities, and limitations of the four repeat-expansion methods in [Table S1](#), which is adapted from Bahlo et al.<sup>20</sup>

## Discussion

Genomic medicine, which uses genomic information about an individual as part of that individual's clinical care, promises better outcomes for individuals and a more efficient health system through rapid diagnosis, early intervention, prevention, and targeted therapy.<sup>24,25</sup> A single, affordable front-line test that is able to comprehensively detect the genetic basis of human disease is the ulti-

mate goal of diagnostics for genomic medicine and represents the logical way forward in an era of personalized 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 and employs a condition-by-condition approach. This method is cost-effective when the clinical diagnosis is straightforward. However, for some disorders, such as the spinocerebellar ataxias, the "right" test is not immediately obvious. The genetic basis of disease remains unsolved in many individuals and families, even after extensive genetic studies encompassing both gene sequencing and repeat-expansion testing.<sup>11</sup> 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. NGS-based tests will act as a screening tool to identify putative expansions, which clinicians then need to follow up on with gold-standard methods such as Southern-blot analysis or repeat-primed PCR. Clinical geneticists will need to determine pathogenicity once the precise make-up of the repeat is determined. SNVs and indels detected in NGS also have to be validated and clinically interpreted. Detecting repeat expansions with NGS-based tests would result in both increased diagnostic yield and a reduction in the diagnostic odyssey for many affected individuals.

Previously described methods such as hipSTR<sup>14</sup> attempt to genotype STRs, i.e., estimate the allele sizes, which renders the methods ineffective when the repeat size exceeds the read length of the sequencing platform. To address this shortcoming, researchers have developed several methods that are designed to specifically call repeat expansions. By examining the performance of these methods on data from >100 individuals known to have repeat expansions and whose conditions spanned twelve different repeat-expansion disorders, we show that exSTRa does not require PCR-free library sequencing protocols, or even WGS, to detect repeat expansions. We show that exSTRa delivers consistent, robust results in simulation studies.

exSTRa analysis can be run in self-contained cohorts with 15 or more individuals. exSTRa does not require any individuals who are known to be unaffected by repeat expansions because it makes use of expanded individuals as “controls” for other loci by using all available data with its robust outlier detection method. exSTRa performs simulations, parameterized with robust estimators, to determine the significance of the outlier test statistic. Hence, the default setting for exSTRa requires that no more than 15% of individuals in the cohort have the same repeat expansion. exSTRa has an adjustable trimming parameter. Trimming too many observations leads to non-robust results. The default setting is 15%, but this can be increased if necessary and assessed for performance with the ECDF plots. This setting was increased to 25% for the WGS\_PF cohort for the loci that had a large number of individuals with the expansions FRAXA (34/118, 29%), FRDA, (25 of 118, 21%) and DM1 (17 of 118, 14%). Real disease cohorts, even those ascertained from individuals with diseases such as spinocerebellar ataxia, which is known to be enriched for repeat expansions, are highly unlikely to have a >15% contribution from one particular repeat expansion, based on known frequencies of such expansions.

We show that exSTRa detected the most repeat expansions across all platforms and STR loci tested. It outperforms other methods at some loci, such as FRAXA, which is the highest-frequency Mendelian cause of autism. exSTRa performs well in cohorts that have sequencing data with more restrictions on size fragments and greater PCR artifacts; such data include those obtained from WES

or WGS with PCR-based library preparations. Other advantages are that it can be run with fewer requirements (no controls necessary, no size thresholds) and that its graphical ECDF representation allows QC and fine-tuning of analysis. The exSTRa input file is easily amended to include further loci beyond the 21 investigated. The user can determine these loci by making use of the Tandem Repeat Finder output in the UCSC Genome Browser. As part of the GitHub exSTRa archive, we also supply an additional input file of STRs, which consists of a genome-wide list of STR loci that are specifically expressed in the brain. The user can amend this file to target specific genomic regions, such as those identified in linkage analysis. In comparison, ExpansionHunter and TREDPARSE (for the threshold model) currently require knowledge of the pathogenic allele size, which will not be known for newly discovered repeat-expansion loci. STRetch investigates all STRs listed in its input file simultaneously and uses its decoy-chromosome method, facilitating genome-wide analysis. However, this requires re-alignment to an augmented chromosome. We also found that the decoy-chromosome method does not perform well with short expansions such as SCA6 because these shorter expanded alleles will preferentially find other sites in the genome, rather than the augmented genome (data not shown). exSTRa does not attempt to call allele sizes, which TREDPARSE, ExpansionHunter, and STRetch infer. However, gold-standard validation with repeat-primed PCR or Southern blot still needs to occur prior to return of the genetic findings, and these methods size alleles more accurately than the NGS-based methods.<sup>26</sup>

We have not investigated the impact of different aligners in detail, but examination of ECDFs that were from the same cohort but that were aligned with BWA and Bowtie, the two most commonly used aligners, show highly concordant results. The ability to use existing alignments is a valuable, time-saving step for STR-expansion analysis. exSTRa’s ECDF plots inform researchers as to whether re-alignment is necessary or not when batches from different cohorts are combined. Combining cohorts across sequencing platforms is not advisable because motif capture, and hence distributions of motif sizes, differ between platforms, and these differences lead to batch effects.

Some expansion alleles show population heterogeneity in allele sizes; this heterogeneity could influence the inference of expansions with exSTRa but will also affect other repeat-expansion detection methods because they also implicitly assume homogeneity of repeat-expansion distributions. One advantage of exSTRa in this context is that the ECDF method allows assessments of the results for such features. If appropriate, population heterogeneity and membership can be assessed with methods such as PLINK<sup>27</sup> or PEDDY,<sup>28</sup> allowing the identification and removal of population outliers or stratification of cohorts. Furthermore, the exSTRa ECDF method allows assessments of the results for such features.

In the context of our results, exSTRa and the other three methods appear to have potential as tools that can screen populations for carrier status. For example, all the methods should be able to identify, with high sensitivity and specificity, carriers for Friedreich's ataxia, which is the most prevalent of the inherited ataxias and has a carrier frequency of  $\sim 1/100$ . More broadly, although the current version of exSTRa performed suboptimally for detection of *FMR1* expansions, we anticipate that these limitations can be resolved with further refinements of exSTRa or similar detection methods. Fragile X syndrome (FXS) is the most common cause of inherited intellectual disability. Approximately 1/300 individuals carry a premutation allele (55–200 repeats) that causes fragile-X-associated tremor ataxia syndrome and fragile X primary ovarian insufficiency.<sup>29</sup> 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.<sup>30,31</sup> There is now discussion regarding the clinical utility of screening *FMR1* for reproductive and personal healthcare.<sup>32</sup>

Given that the genetic basis of disease in many affected individuals currently remains unsolved, even after extensive genetic sequencing, we recommend the introduction of a protocol, such as exSTRa, into any standard sequencing-analysis pipeline, and we suggest that this be run both prospectively and retrospectively. This should identify missed repeat expansions in individuals who have only been tested for a subset of common repeat expansions, which is currently standard clinical practice, and will also expedite the diagnosis of individuals potentially suffering from a repeat-expansion disorder. There are already more than 20 known repeat-expansion loci, but more are probably awaiting discovery. In OMIM there are additional putative SCA loci, such as SCA25 (MIM: 608703, 2p21–p13), whose genetic causes have not yet been identified but which are potentially due to pathogenic repeat expansions.

We anticipate that, with large cohorts and further improvements in methodology, methods such as exSTRa and future developments in technology will facilitate the discovery of new repeat-expansion loci, which in turn will identify the etiology of neurodegenerative disorders in more affected individuals and families. exSTRa enables fast discovery of repeat expansions in next-generation sequencing discovery cohorts, including retrospective cohorts consisting mainly of WES data or WGS PCR-based library-preparation data. An important new challenge lies in the detection of repeat expansions that are *de novo*<sup>4,5</sup> and not represented in the reference set of STRs that all four methods need in order to stipulate the genomic locations at which to test. Addressing this current limitation of all repeat-expansion detection algorithms will require refinement of existing, or the development of new, bioinformatics tools.

The identification of a potentially pathogenic repeat expansion by using detection methods such as exSTRa

should not replace the current diagnostic, locus-specific, PCR-based tests. First, with higher sensitivity and specificity than the sequencing-based methods, these will remain the gold standard, and second, they give much more accurate estimates of the size of the expanded allele(s) and the makeup of the repeat, including whether there are interruptions, which has prognostic implications for the age of onset, disease progression, and outcome.

We anticipate that there will be further improvements to all of the current methods that identify repeat expansions in NGS data. Clearly, sources of bias affect certain loci and contribute to the poor performance at some of the STRs. For instance, we observed a GC bias for the repeat-expansion alleles underlying FRAXA, FRAXE, and FTDALS1; far fewer reads were able to capture these repeat expansions than captured other expansions as a result of their extreme GC content. Notably, FRAXA and FTDALS1 had substantially improved coverage with the PCR-free protocol; however, the GC bias can also arise from sources other than library preparation.<sup>33</sup> Further investigation of the impact of bias, and potential incorporation of such information to improve detection methods, is beyond the scope of this manuscript. Detection of repeat expansions would also improve with some library-preparation changes, such as increased fragment lengths, longer insertion sizes, and increased read lengths. However, these would come at increased costs and would require re-sequencing or non-standard library preparation.

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 being almost ten times more expensive than the prevailing Illumina HiSeq X sequencing platform, it is currently not cost effective. The development of methods such as exSTRa will lead to further improvements for individuals' care via clinical genomic sequencing. Such methods will also facilitate the pending era of precision and preventative medicine, wherein screening tests will become much more prevalent. A universal single test will be cost and time effective in comparison to the existing stand-alone tests currently required when clinicians wish to screen for known repeat expansions.

## Supplemental Data

Supplemental Data include thirteen figures and six tables and can be found with this article online at <https://doi.org/10.1016/j.ajhg.2018.10.015>.

## Acknowledgments

We would like to thank Egor Dolzhenko and Michael Eberle (Illumina), who produced the STR specification files for ExpansionHunter and who facilitated access to the dataset EGA00001003562 from the European Genome-Phenome Archive. We thank Leslie Burnett, Ben Lundie, Katie Ayres, and Andrew Sinclair for granting access to control datasets. We also thank Kate Pope and Greta Gillies for assistance with recruitment and

sample preparation. R.M.T. was supported by an Australian Postgraduate Award and funding from the Edith Moffat fund. P.J.L. was supported by the Australian National Health and Medical Research Council (NHMRC) Career Development Award Level 2 (GNT1032364). M.B. was supported by an NHMRC Program Grant (GNT1054618) and an NHMRC Senior Research Fellowship (GNT1102971). This work was supported by the Victorian Government's Operational Infrastructure Support Program and the NHMRC Independent Research Institute Infrastructure Support Scheme (IRIIS).

## Declaration of Interests

The authors declare no competing interests.

Received: July 13, 2017

Accepted: October 16, 2018

Published: November 29, 2018

## Web Resources

Coriell, <https://www.coriell.org/>  
ExpansionHunter, <https://github.com/Illumina/ExpansionHunter>  
exSTRa, <https://github.com/bahlolab/exSTRa>  
GATK, IndelAligner <https://software.broadinstitute.org/gatk/>  
Novosort, <http://www.novocraft.com/products/novosort/>  
OMIM, <https://www.omim.org>  
Picard, <http://broadinstitute.github.io/picard/>  
STRetch, <https://github.com/Oshlack/STRetch>  
TREDPARSE, <https://github.com/humanlongevity/tredparse>

## References

- Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* *27*, 573–580.
- Jones, L., Houlden, H., and Tabrizi, S.J. (2017). DNA repair in the trinucleotide repeat disorders. *Lancet Neurol.* *16*, 88–96.
- Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* *19*, 286–298.
- Seixas, A.I., Loureiro, J.R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C.L., Loureiro, J.L., Dhingra, A., Brandão, E., Cruz, V.T., et al. (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. *Am. J. Hum. Genet.* *101*, 87–103.
- Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M.K., Fujiyama, A., Toyoshima, Y., Kakita, A., Takahashi, H., Suzuki, Y., et al. (2018). Expansions of intronic TTTC and TTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* *50*, 581–590.
- Schulz, J.B., Boesch, S., Bürk, K., Dürr, A., Giunti, P., Mariotti, C., Pousset, F., Schöls, L., Vankan, P., and Pandolfo, M. (2009). Diagnosis and treatment of Friedreich ataxia: a European perspective. *Nat. Rev. Neurol.* *5*, 222–234.
- Seltzer, M.M., Baker, M.W., Hong, J., Maenner, M., Greenberg, J., and Mandel, D. (2012). Prevalence of CGG expansions of the FMR1 gene in a US population-based sample. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* *159B*, 589–597.
- Tassone, F. (2014). Newborn screening for fragile X syndrome. *JAMA Neurol.* *71*, 355–359.
- Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium (2015). Identification of genetic factors that modify clinical onset of Huntington's disease. *Cell* *162*, 516–526.
- Bettencourt, C., Hensman-Moss, D., Flower, M., Wiethoff, S., Brice, A., Goizet, C., Stevanin, G., Koutsis, G., Karadima, G., Panas, M., et al.; SPATAX Network (2016). DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Ann. Neurol.* *79*, 983–990.
- Németh, A.H., Kwasniewska, A.C., Lise, S., Parolin Schnekenberg, R., Becker, E.B.E., Bera, K.D., Shanks, M.E., Gregory, L., Buck, D., Zameel Cader, M., et al.; UK Ataxia Consortium (2013). Next generation sequencing for molecular diagnosis of neurological disorders using ataxias as a model. *Brain* *136*, 3106–3118.
- Tae, H., Kim, D.-Y., McCormick, J., Settlage, R.E., and Garner, H.R. (2014). Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. *Bioinformatics* *30*, 652–659.
- Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* *22*, 1154–1162.
- Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* *14*, 590–592.
- Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., and Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.* *41*, e32.
- Cao, M.D., Tasker, E., Willadsen, K., Imelfort, M., Vishwanathan, S., Sureshkumar, S., Balasubramanian, S., and Bodén, M. (2014). Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res.* *42*, e16.
- Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al.; US–Venezuela Collaborative Research Group (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* *27*, 1895–1903.
- Dashnow, H., Lek, M., Phipson, B., Halman, A., Sadedin, S., Lonsdale, A., Davis, M., Lamont, P., Clayton, J.S., Laing, N.G., et al. (2018). STRetch: Detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* *19*, 121.
- Tang, H., Kirkness, E.F., Lippert, C., Biggs, W.H., Fabani, M., Guzman, E., Ramakrishnan, S., Lavrenko, V., Kakaradov, B., Hou, C., et al. (2017). Profiling of short-tandem-repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.* *101*, 700–715.
- Bahlo, M., Bennett, M.F., Degorski, P., Tankard, R.M., Delatycki, M.B., and Lockhart, P.J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res.* *7*, 736.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S., Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* *10*, R25. PubMed.
- Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics* *28*, 593–594.
- Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* *40*, e72.
- Rehm, H.L. (2017). Evolving health care through personal genomics. *Nat. Rev. Genet.* *18*, 259–267.



25. Ashley, E.A. (2015). The precision medicine initiative: A new national effort. *JAMA* 313, 2119–2120.
26. Mousavi, N., Shleizer-Burko, S., and Gymrek, M. (2018). Profiling the genome-wide landscape of tandem repeat expansions. *bioRxiv*. <https://doi.org/10.1101/361162>.
27. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.
28. Pedersen, B.S., and Quinlan, A.R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am. J. Hum. Genet.* 100, 406–413.
29. Hunter, J., Rivero-Arias, O., Angelov, A., Kim, E., Fotheringham, I., and Leal, J. (2014). Epidemiology of fragile X syndrome: A systematic review and meta-analysis. *Am. J. Med. Genet. A.* 164A, 1648–1658.
30. Ligsay, A., and Hagerman, R.J. (2016). Review of targeted treatments in fragile X syndrome. *Intractable Rare Dis. Res.* 5, 158–167.
31. Hagerman, R.J., and Polussa, J. (2015). Treatment of the psychiatric problems associated with fragile X syndrome. *Curr. Opin. Psychiatry* 28, 107–112.
32. Hagerman, R., and Hagerman, P. (2013). Advances in clinical and molecular understanding of the FMR1 premutation and fragile X-associated tremor/ataxia syndrome. *Lancet Neurol.* 12, 786–798.
33. Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 12, R18.