ARTICLE

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

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

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,[1] 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.[2] 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.[3]

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 *FMR1*); 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., Freidreich 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.[4,5] 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.

[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] Mathematics and Statistics, Murdoch University, Murdoch 6150, WA, Australia; [4] Epilepsy Research Centre, Department of Medicine, The University of Melbourne, Austin Health, Heidelberg 3084, VIC, Australia; [5] Bruce Lefroy Centre for Genetic Health Research, Murdoch Children's Research Institute, Royal Children's Hospital, Parkville 3052, VIC, Australia; [6] Victorian Clinical Genetics Services, Parkville 3052, VIC, Australia; [7] Department of Paediatrics, University of Melbourne, Parkville 3058, VIC, Australia
[8] These authors contributed equally to this work
*Correspondence: bahlo@wehi.edu.au

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

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	*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

(*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	*NOP56*	20p13	Intron	GGCCTG	3–8	1500–2500	+	2,633,379	7.2	97	0	43
Spinocerebellar ataxia 10	SCA10	603516	AD	*ATXN10*	22q13.31	Intron	ATTCT	10–20	500–4500	+	46,191,235	14	100	0	70
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
Spinocerebellar ataxia 31	SCA31	117210	AD	*BEAN1/ TK2*	16q21	Intron	TGGAA[a]	0	2.5–3.8kb[b]	+	66,524,302	0	N/A	N/A	N/A
Spinocerebellar ataxia 37	SCA37	615945	AD	*DAB1*	1p32.3	Intron	ATTTC[a]	0	31–75	–	57,832,716[c]	0	N/A	N/A	N/A
Familial adult myoclonic epilepsy 1[e]	FAME1/ BAFME1	601068	AD	*SAMD12*	8q24	Intron	TTTCA[a]	0	440–3,680[f]	–	119,379,055[d]	0	N/A	N/A	N/A

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

[a] These repeat expansions are insertions and thus not represented in the reference genome at their respective locations.

[b] SCA31 is caused by the insertion of a complex repeat containing $(TGGAA)_n$; hence, the length is given as the length of the expanded repeats in base pairs, instead of the repeat number.

[c] The SCA37 physical map location is given at the reference $(ATTTT)_n$ repeat, where affected individuals have the pathogenic $(ATTTC)_n$ inserted.

[d] The FAME1 physical map location is given as the position of the reference $(TTTTA)_n$ repeat, where affected individuals have $(TTTCA)_n$ inserted.

[e] Ishiura et al. identified similar expansions associated with FAME6 and FAME7, in the genes TNRC6A and RAPGEF2, respectively, but only in single families.[5] These have not been listed.

[f] The FAME1 repeat size is the estimated size of the combined expanded $(TTTCA)_n$ and the $(TTTTA)_n$ 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,[6] 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.[7,8] 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.[9,10] 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).[11] 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.[12–16] 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,[17] STRetch[18] and TREDPARSE,[19] reviewed in Bahlo et al.[20] 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).[17] 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.[17] All sequencing data were aligned to the hg19 human genome reference using the Bowtie 2 aligner[21] 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)[a]	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.[17]

[a]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,[22] 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[17] 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[18] 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[19] 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.[17] 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 1st and 3rd 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.[23] 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).

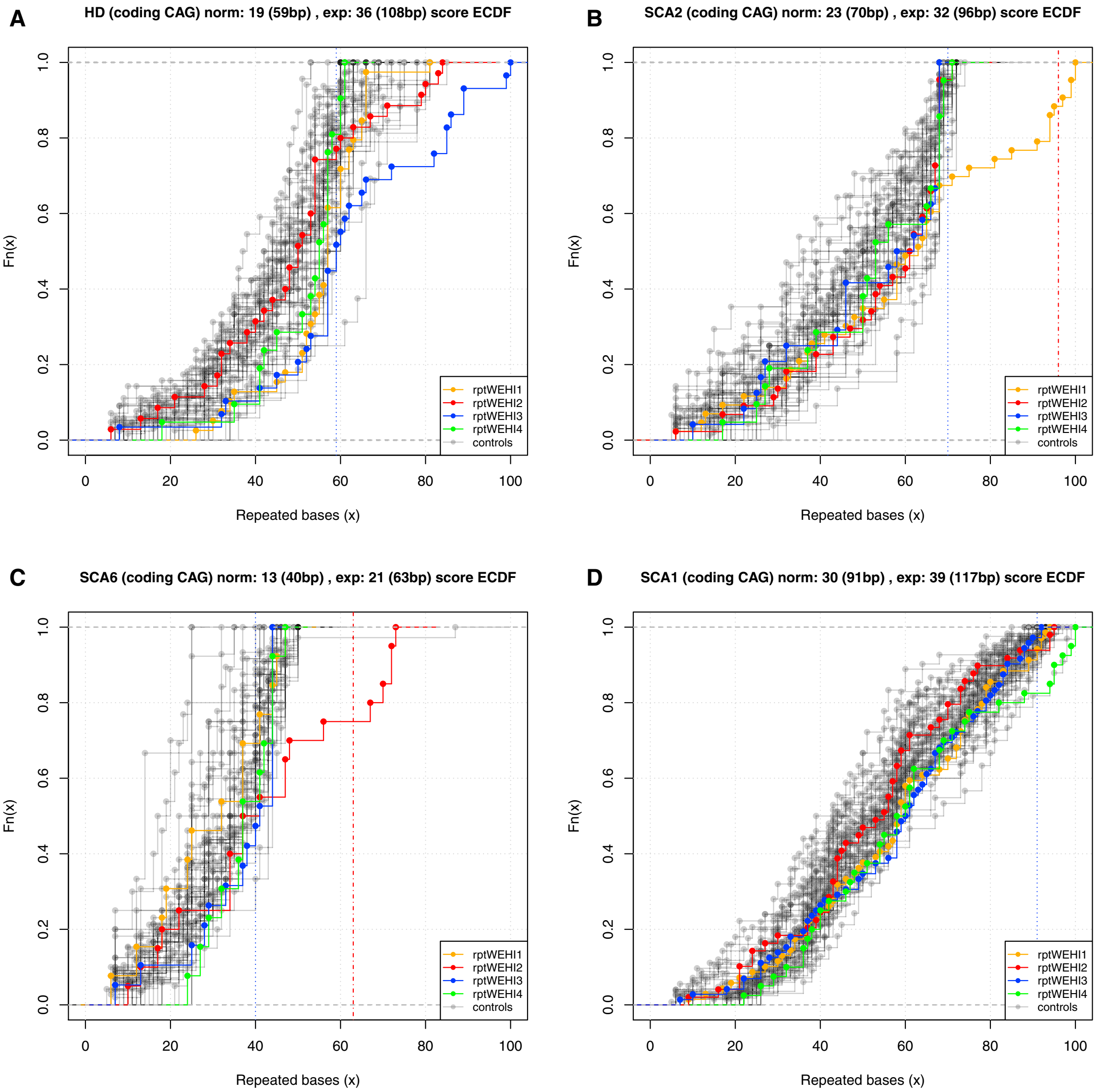


Figure 1. ECDF of Repeat-Expansion Composition of Reads from the WES Cohort
Four different known repeat-expansion disorders captured by WES are shown: (A) HD, (B) SCA2, (C) SCA6, and (D) SCA1. Sample rptWEHI3 (blue) is from an individual with a known HD repeat expansion. The expanded allele size is not known. Sample rptWEHI1 (yellow) is a known SCA2 repeat expansion of length 42 repeats, sample rptWEHI2 (red) is a known SCA6 repeat expansion of length 22 repeats, and sample rptWEHI4 (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

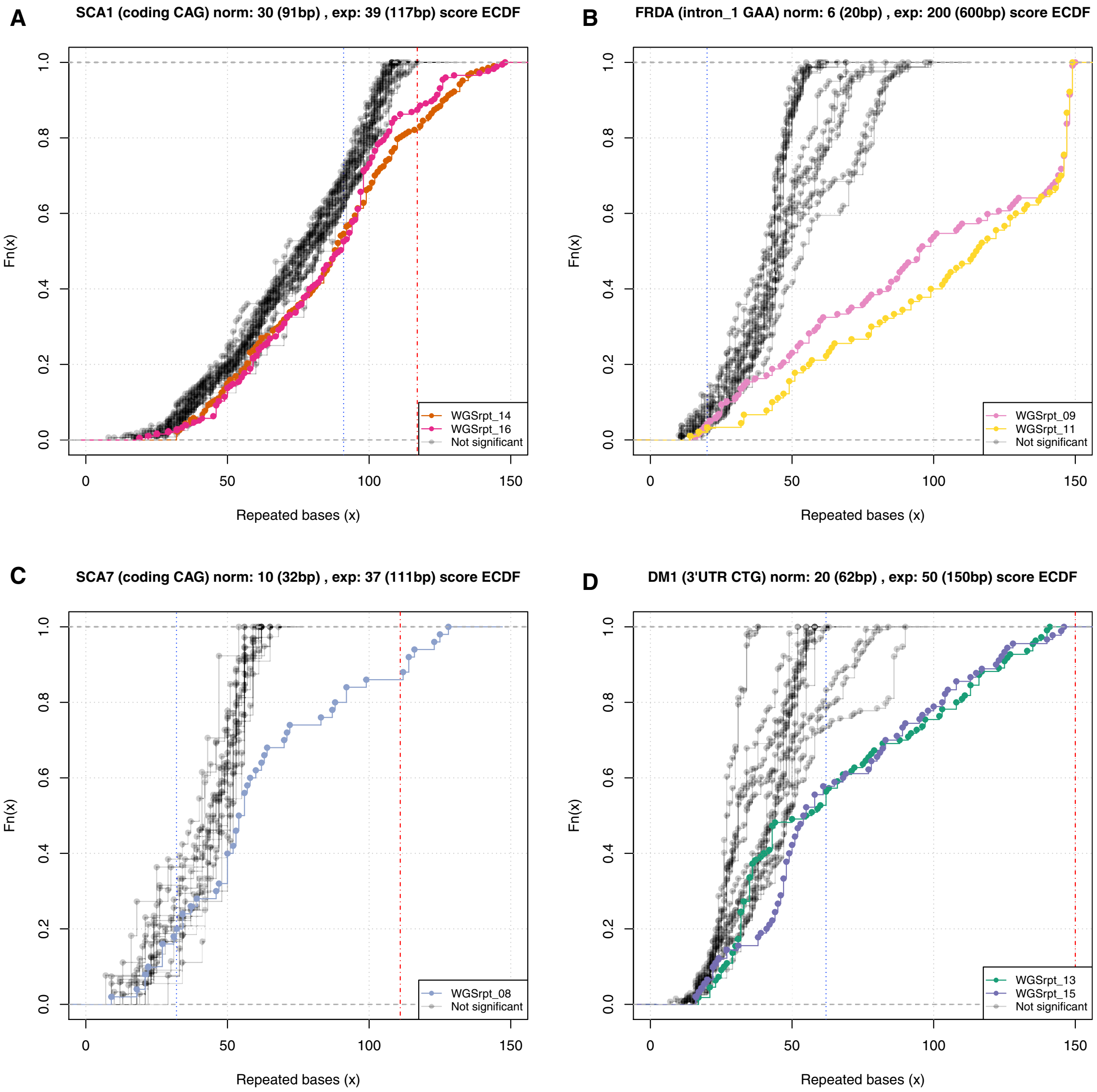


Figure 2. ECDFs of Repeat-Expansion Composition of Reads from the WGS_PCR_2 Cohort
Four different STR loci are shown: (A) SCA1 (lengths of the expanded alleles are 52 and 45 repeats); (B) FRDA (lengths of the expanded alleles are 320 and 788 repeats); (C) SCA7 (length of the expanded allele is 39 repeats); 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 subcohorts. 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[a]	Method	TP	FN	TN	FP	Sensitivity	Specificity
WES_PCR[b]	4	58	exSTRa	4	0	607	9	1	0.99
			ExpansionHunter	2	2	616	0	0.5	1
			STRetch[c]	3	1	613	3	0.75	1
			TREDPARSE-T[d]	4	0	585	31	1	0.95
			TREDPARSE-L[d]	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[c]	1	2	336	1	0.33	1
			TREDPARSE-T[d]	3	0	354	0	1	1
			TREDPARSE-L[d]	3	0	354	0	1	1
WGS_PCR_2[e]	16	2	exSTRa	13	3	352	10	0.81	0.97
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch[c]	11	5	338	6	0.69	0.98
			TREDPARSE-T[d]	12	4	362	0	0.75	1
			TREDPARSE-L[d]	11	5	362	0	0.69	1
WGS_PCR_2_30X_1[e]	16	2	exSTRa	13	3	357	5	0.81	0.99
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch[c]	11	5	340	4	0.69	0.99
			TREDPARSE-T[d]	13	3	362	0	0.81	1
			TREDPARSE-L[d]	9	7	362	0	0.56	1
WGS_PCR_2_30X_2[e]	16	2	exSTRa	12	4	354	8	0.75	0.98
			ExpansionHunter	8	8	362	0	0.5	1
			STRetch[c]	11	5	336	8	0.69	0.98
			TREDPARSE-T[d]	13	3	362	0	0.81	1
			TREDPARSE-L[d]	10	6	362	0	0.62	1
WGS_PF	78	40	exSTRa	60	18	2329	71	0.77	0.97
			ExpansionHunter[f]	62	16	2394	6	0.79	1
			STRetch[c]	62	16	2206	76	0.79	0.97
			TREDPARSE-T[d]	52	26	2383	17	0.67	0.99
	66	52	TREDPARSE-L[d]	34	32	2396	16	0.52	0.99
WGS_PF (FRAXA pre) [g]	96	22	exSTRa	63	33	2314	68	0.66	0.97
			ExpansionHunter[f]	95	1	2374	8	0.99	1
			STRetch[c]	62	34	2188	76	0.65	0.97
			TREDPARSE-T[d]	72	24	2364	18	0.75	0.99
	72	46	TREDPARSE-L[d]	48	24	2383	23	0.67	0.99
WGS_PF (no FRAXA)[g]	62	56	exSTRa	52	10	2231	67	0.84	0.97
			ExpansionHunter[f]	61	1	2292	6	0.98	1
			STRetch[c]	62	0	2104	76	1	0.97

(*Continued on next page*)

Table 4. *Continued*

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

[a] 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).
[b] This WES cohort was only assessed over ten STR loci in the capture design.
[c] STRetch was Bonferroni corrected for the same number of tests as the other methods and not genome-wide corrected.
[d] 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.
[e] 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.
[f] For the WGS_PF cohort, the original ExpansionHunter results from Dolzhenko et al.[17] were used, which make use of reads aligned with a different aligner.
[g] For the WGS_PF cohort, we computed additional results for "WGS_PF (FRAXA pre)," using the premutation 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 premutation *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"[17,20], 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 (≥0.97). The specificity drops to ≥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

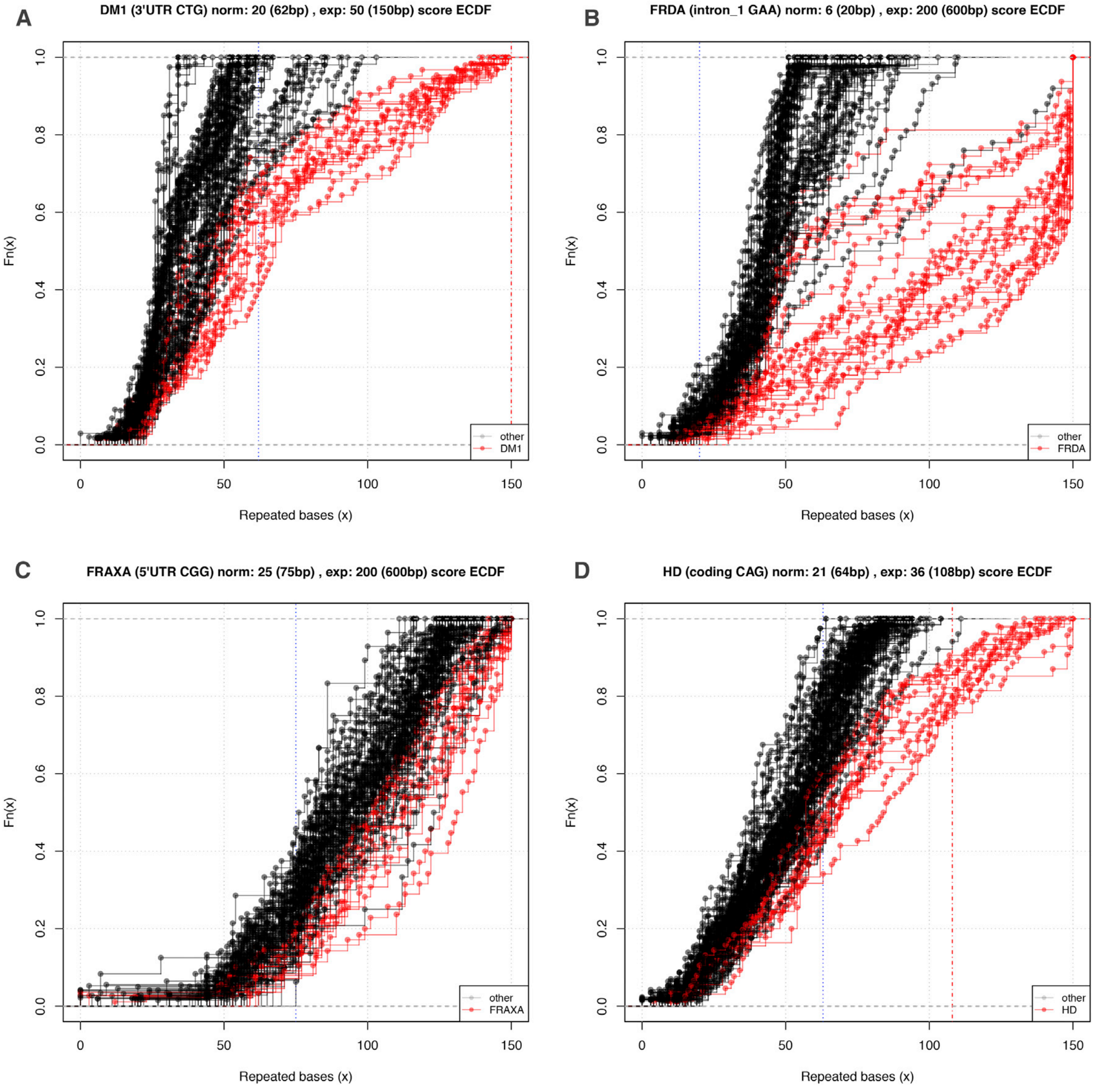


Figure 3. ECDFs of Repeat-Expansion Composition of Reads from the WGS_PF Cohort

(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.[20]

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.[24,25] 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 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.[11] 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[14] 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.[26]

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[27] or PEDDY,[28] 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 ~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.[29] 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.[30,31] There is now discussion regarding the clinical utility of screening *FMR1* for reproductive and personal healthcare.[32]

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*[4,5] 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.[33] 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 (IRIISS).

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

1. Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.
2. Jones, L., Houlden, H., and Tabrizi, S.J. (2017). DNA repair in the trinucleotide repeat disorders. Lancet Neurol. *16*, 88–96.
3. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298.
4. 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.
5. 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 TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nat. Genet. *50*, 581–590.
6. 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.
7. 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.
8. Tassone, F. (2014). Newborn screening for fragile X syndrome. JAMA Neurol. *71*, 355–359.
9. 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.
10. 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.
11. 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.
12. 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.
13. Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. *22*, 1154–1162.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: A next-generation sequencing read simulator. Bioinformatics *28*, 593–594.
23. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72.
24. 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.

The American Journal of Human Genetics, Volume *103*

Supplemental Data

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

Rick M. Tankard, Mark F. Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo

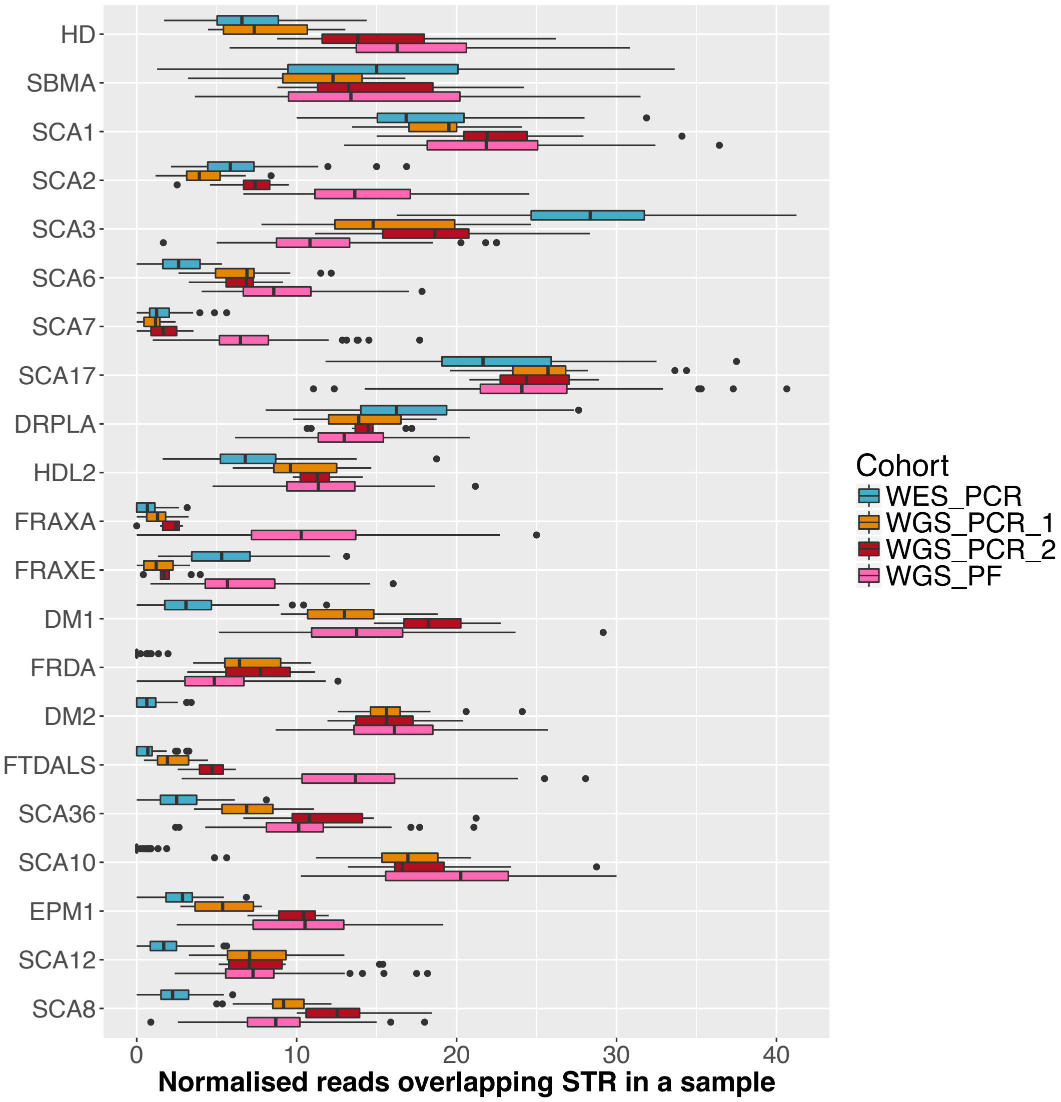


Figure S1 Normalised sequencing coverage comparison between the four sequencing cohorts, split by repeat expansion type.

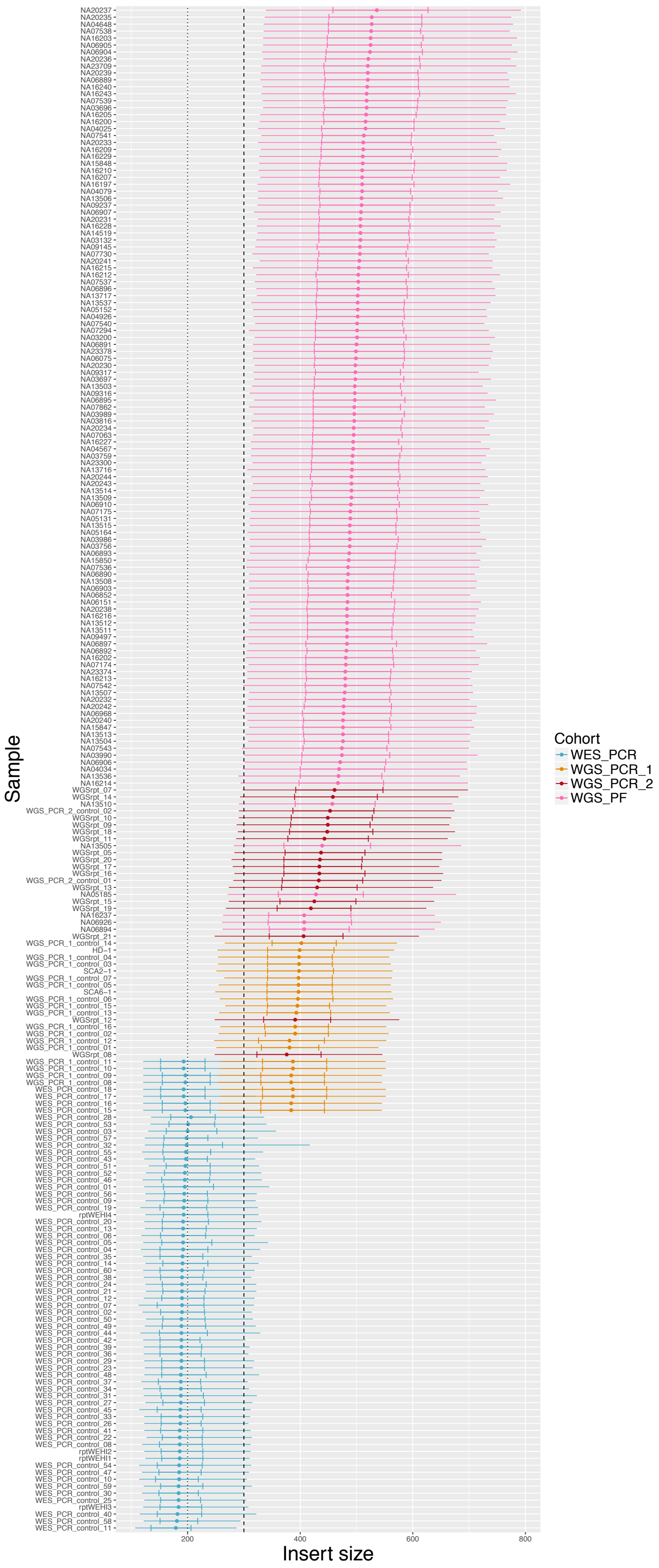
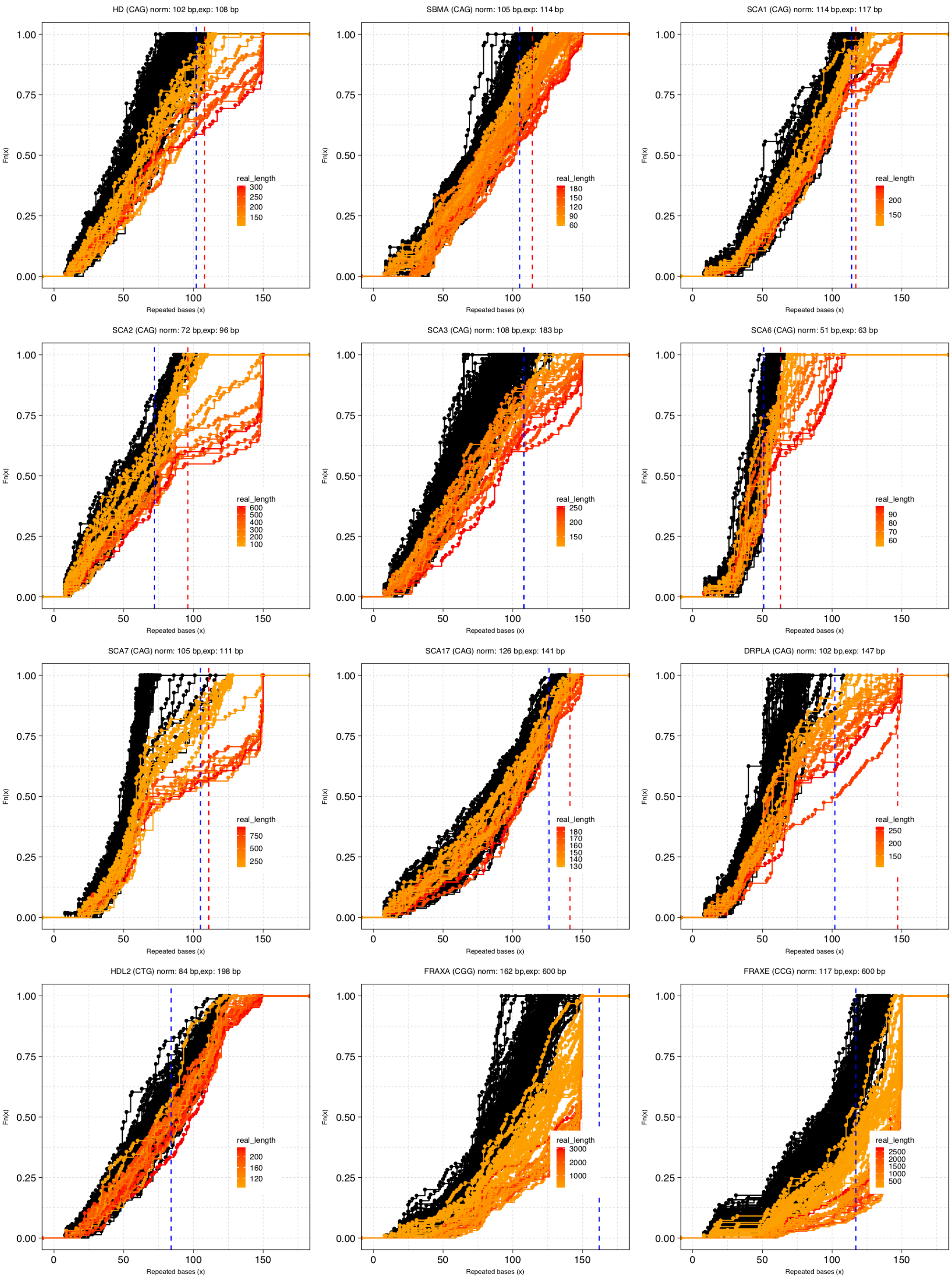


Figure S2 Insert sizes by sample. Samples are in decreasing order of the median insert size that is indicated by a circle. Bars extend to cover 90% of insert sizes, at the 5th and 95th percentile. The interquartile range (IQR), covering 50% of the data, is indicated by small vertical bars. The dotted and dashed vertical lines indicates the threshold at which our WES and WGS samples respectively will usually have overlapping bases, between the two ends of the read.


HD (CAG) norm: 102 bp,exp: 108 bp
SBMA (CAG) norm: 105 bp,exp: 114 bp
SCA1 (CAG) norm: 114 bp,exp: 117 bp
SCA2 (CAG) norm: 72 bp,exp: 96 bp
SCA3 (CAG) norm: 108 bp,exp: 183 bp
SCA6 (CAG) norm: 51 bp,exp: 63 bp
SCA7 (CAG) norm: 105 bp,exp: 111 bp
SCA17 (CAG) norm: 126 bp,exp: 141 bp
DRPLA (CAG) norm: 102 bp,exp: 147 bp
HDL2 (CTG) norm: 84 bp,exp: 198 bp
FRAXA (CGG) norm: 162 bp,exp: 600 bp
FRAXE (CCG) norm: 117 bp,exp: 600 bp
Fn(x)
0.00
0.25
0.50
0.75
1.00
Repeated bases (x)
0
50
100
150
real_length
300
250
200
150
180
150
120
90
60
200
150
600
500
400
300
200
100
250
200
150
90
80
70
60
750
500
250
180
170
160
150
140
130
250
200
150
200
160
120
3000
2000
1000
2500
2000
1500
1000
500

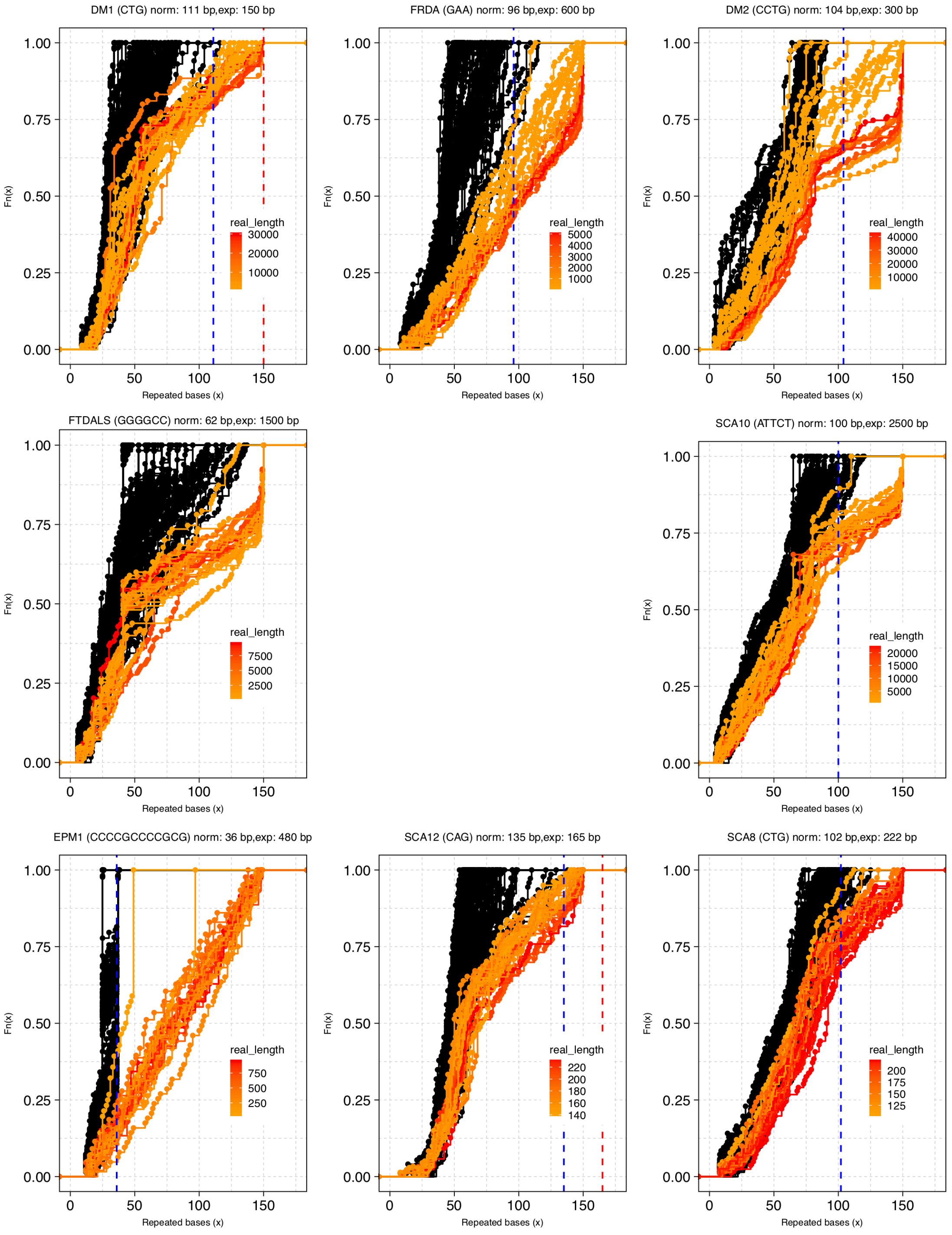
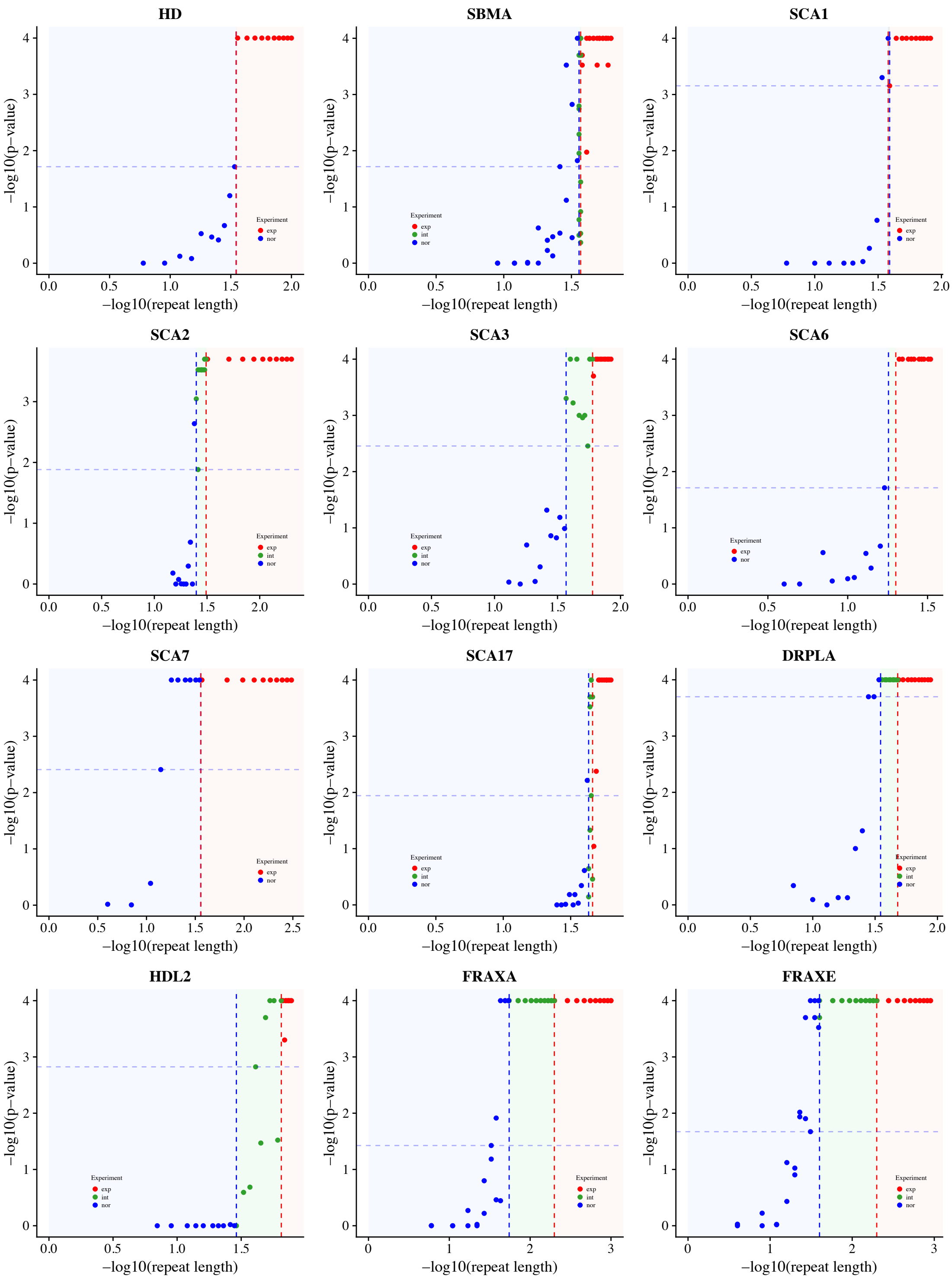


Figure S3 exSTRa eCDF plots for simulated data labeled by size of repeat expansion. Each panel depicts one STR with 210 controls (black) and 20 intermediate size tandem repeat alleles and 20 expanded repeat alleles, with intermediates and expansions coloured in red, with smallest repeat alleles in yellow and largest in red.


HD
SBMA
SCA1
SCA2
SCA3
SCA6
SCA7
SCA17
DRPLA
HDL2
FRAXA
FRAXE
−log10(p−value)
−log10(repeat length)
Experiment
exp
int
nor
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5

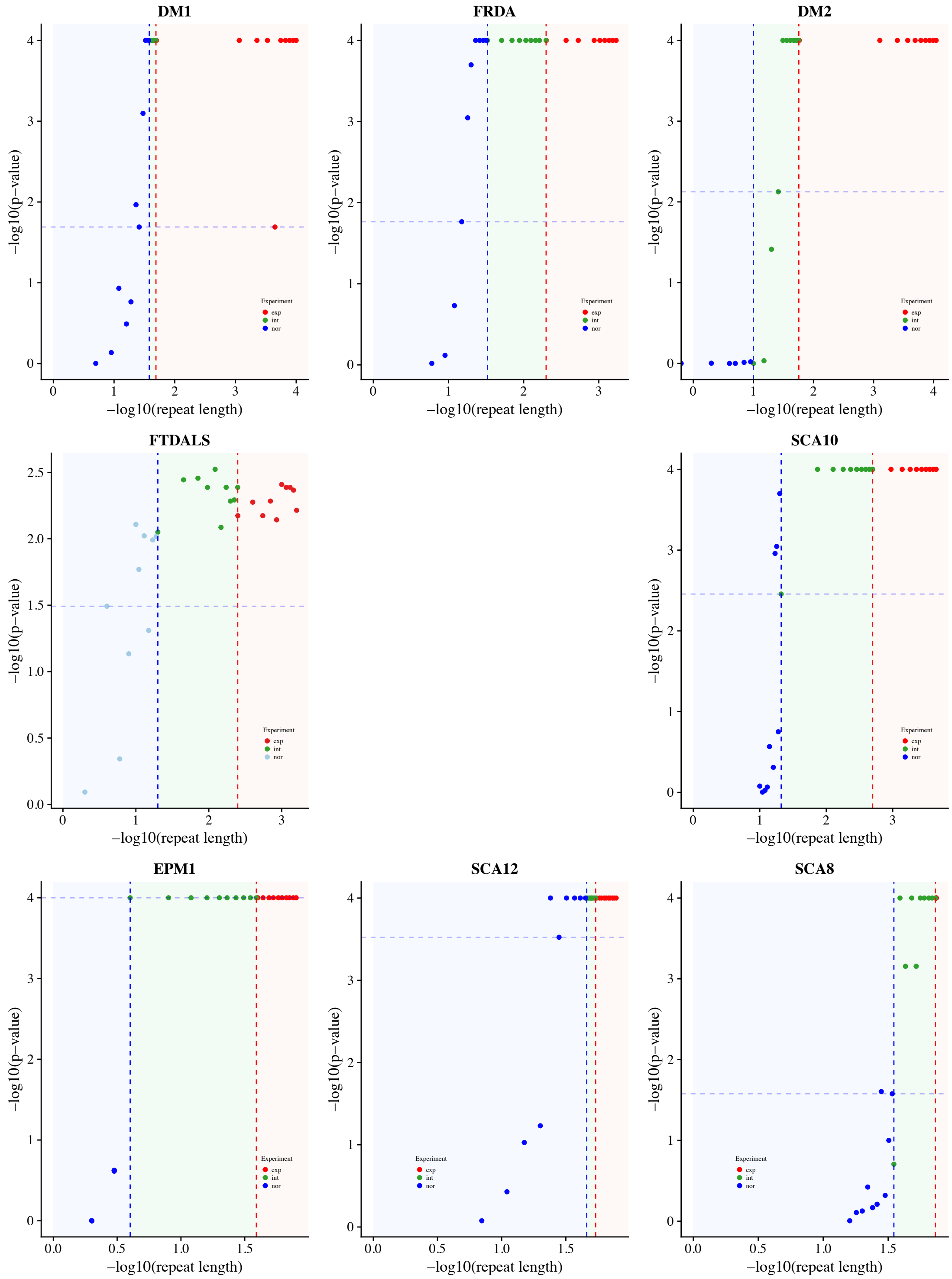
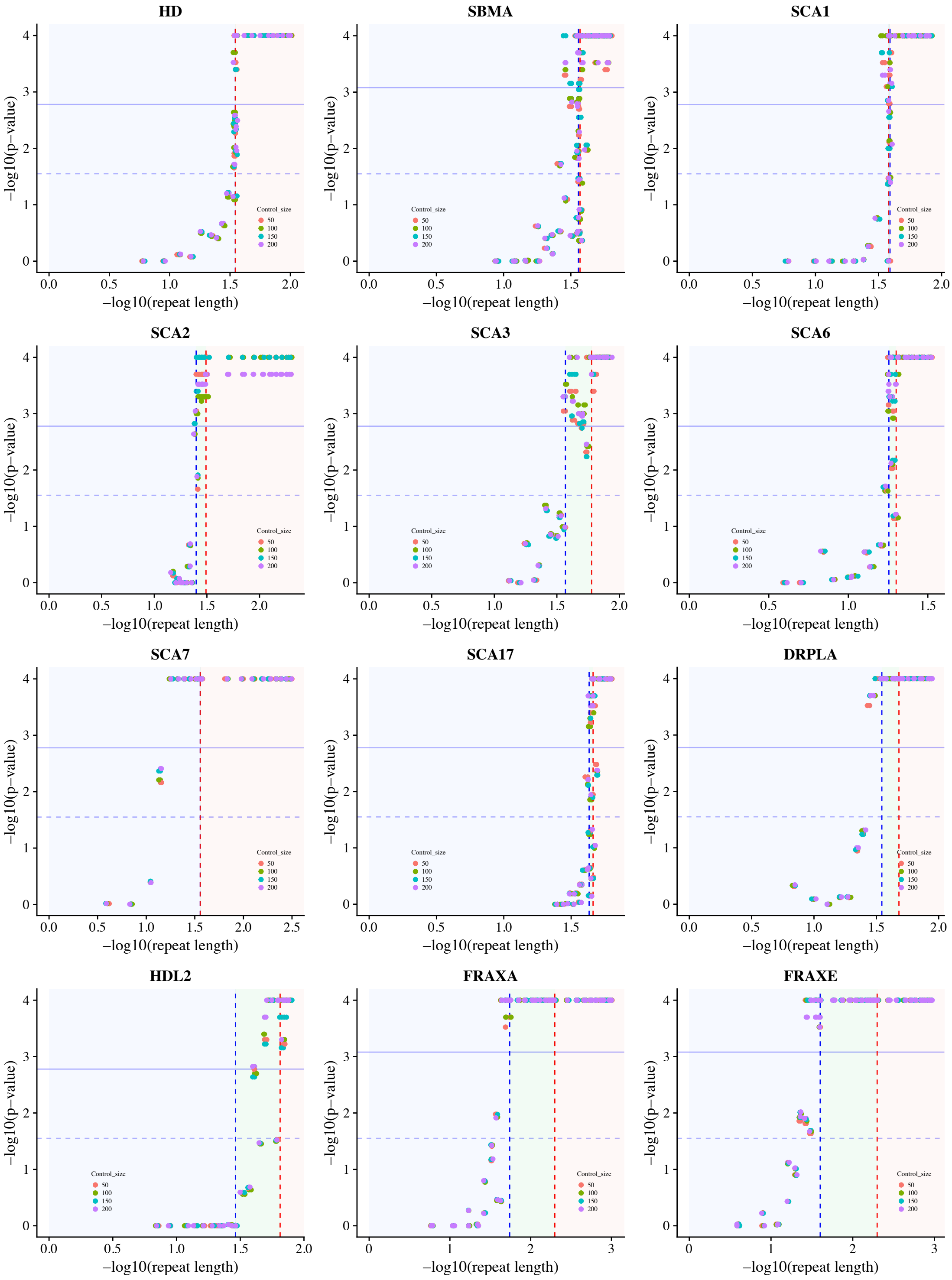


Figure S4 exSTRa p-value behavior with varying repeat length. Regions in light blue are normal ranges, regions in green the intermediate range, which usually means not pathogenic, or leads to a different phenotype, possibly with lower penetrance. The region in red is the pathogenic range. True expansions are red, intermediate expansions green and unexpanded blue. Controls are not shown, as they were not tested for expansions.


HD
SBMA
SCA1
SCA2
SCA3
SCA6
SCA7
SCA17
DRPLA
HDL2
FRAXA
FRAXE
−log10(p−value)
−log10(repeat length)
Control_size
50
100
150
200

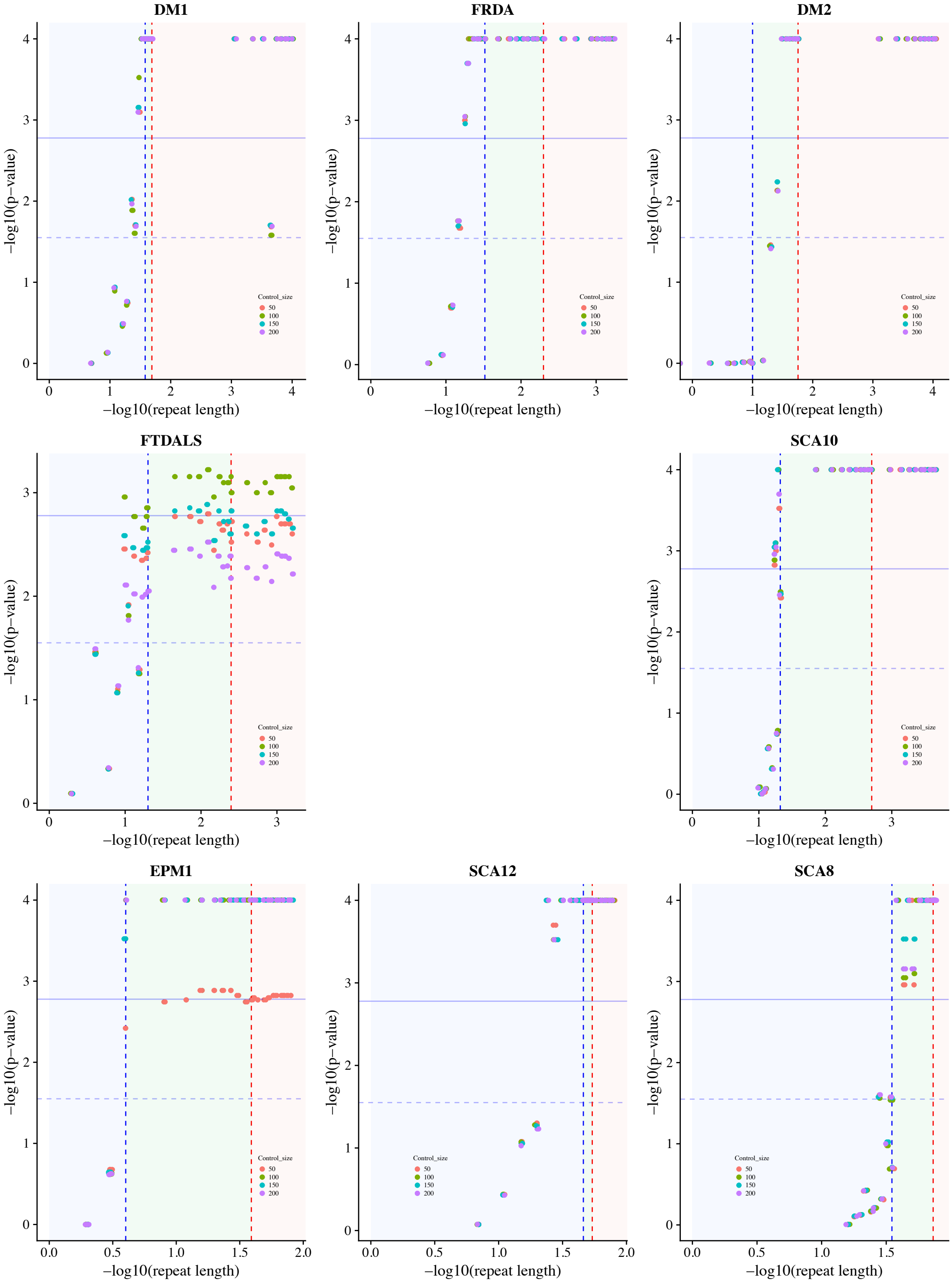


Figure S5 exSTRa p-value behavior with varying control cohort size.

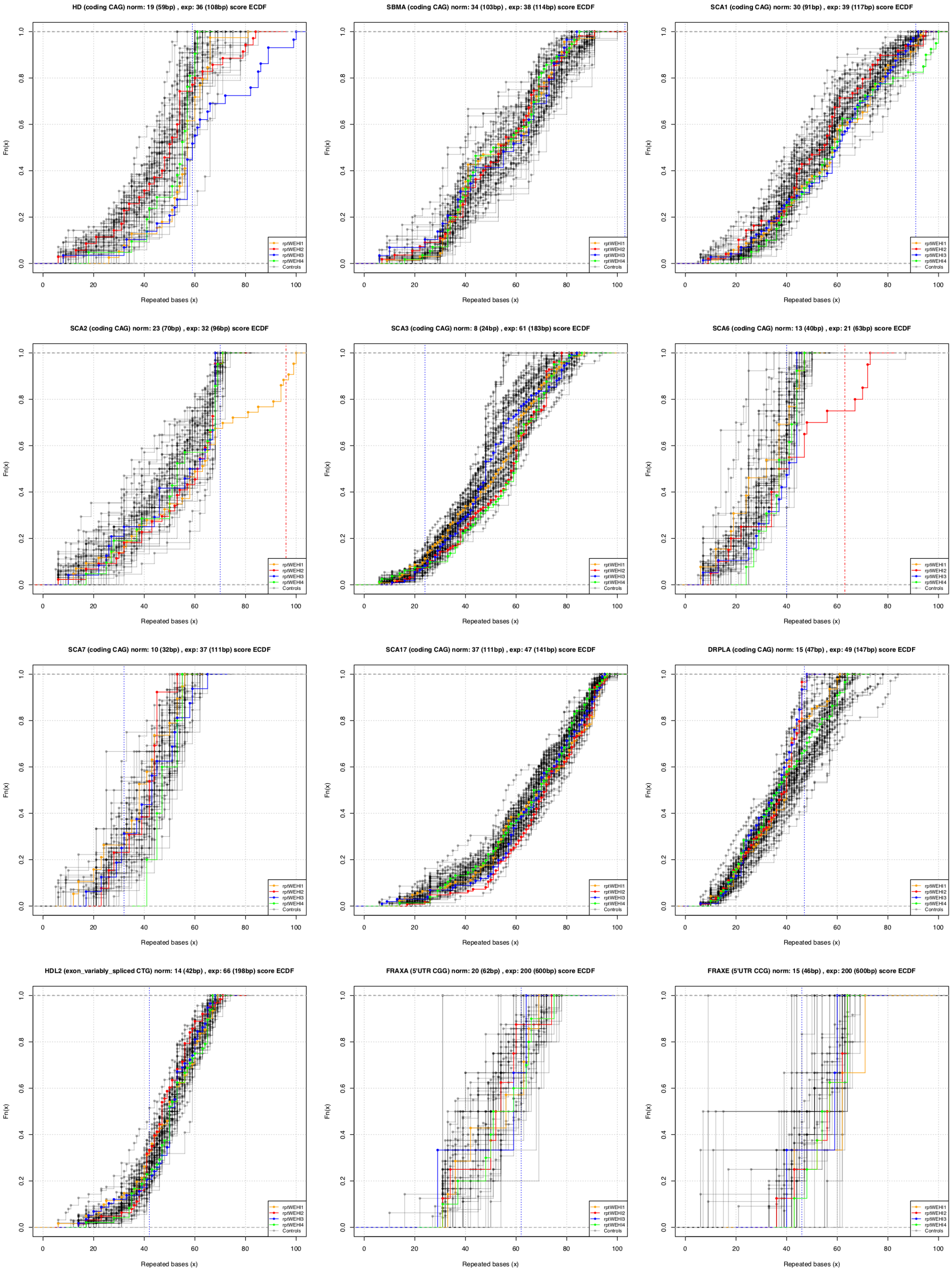
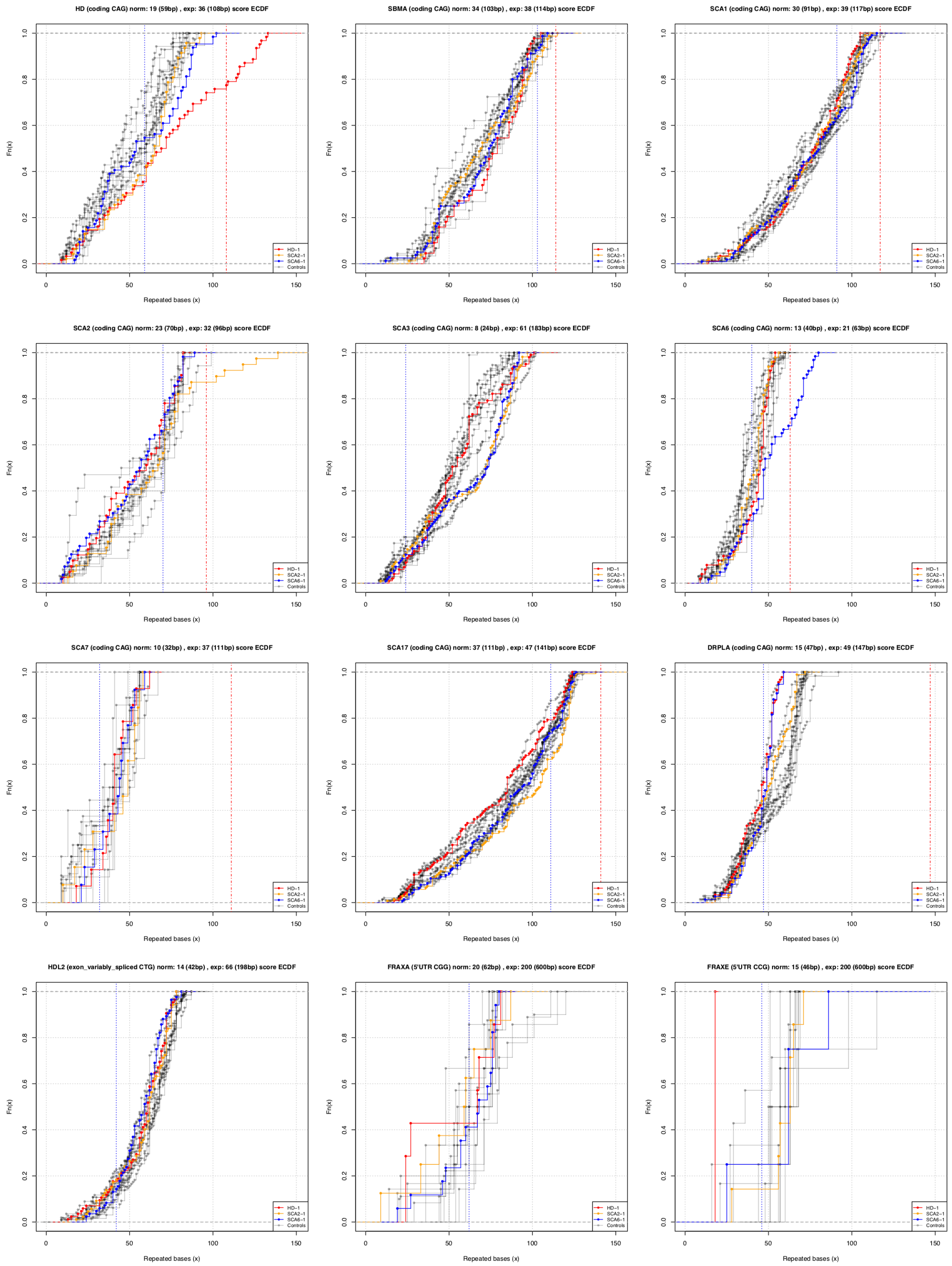


Figure S6. ECDFs for the 13 STR loci with coverage for the WES cohort (WES_PCR).



HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF

SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF

SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF

SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF

SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF

SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF

SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF

SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF

DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF

HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF

FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF

FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF

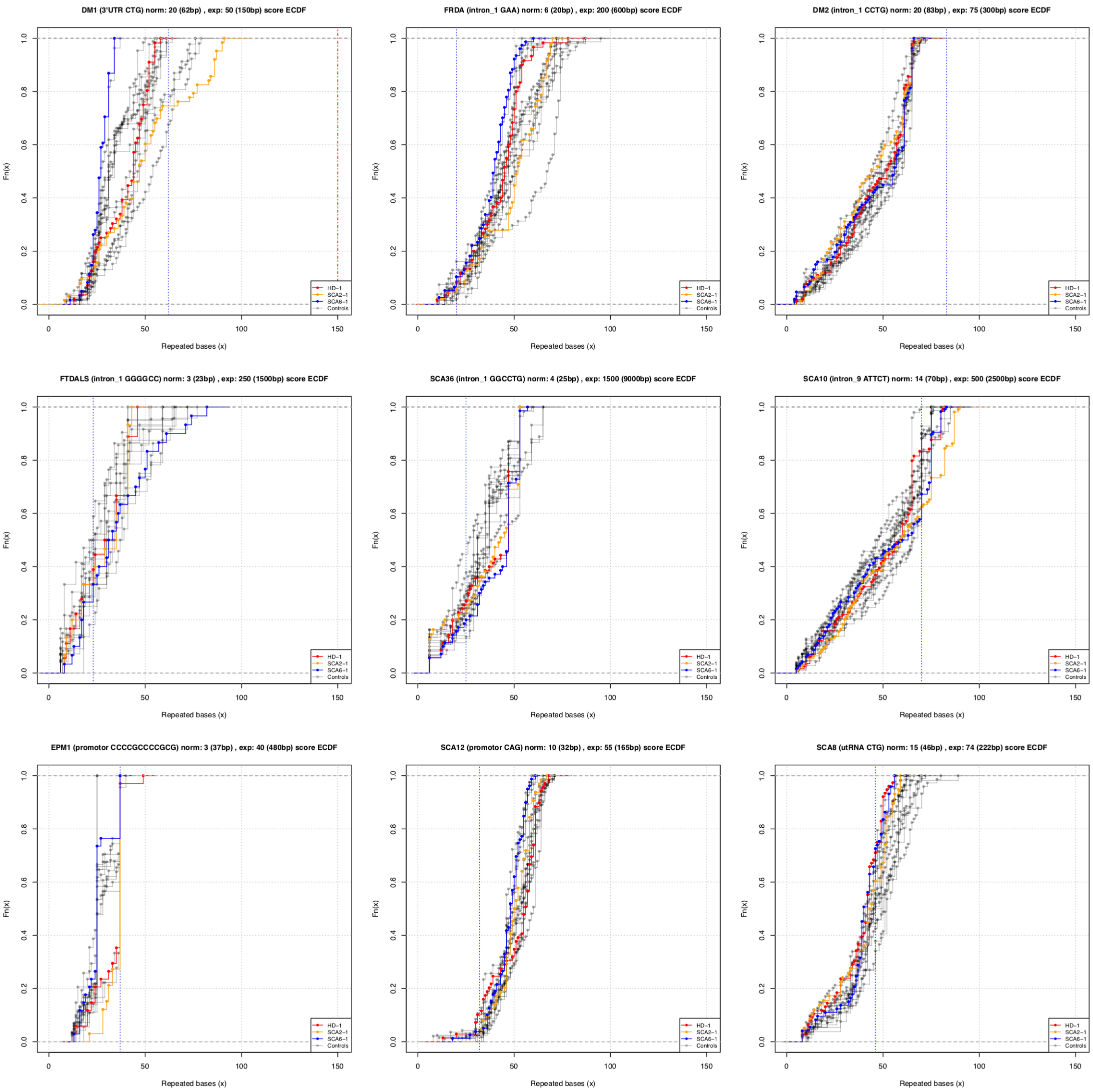
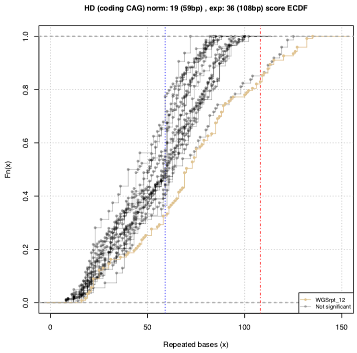
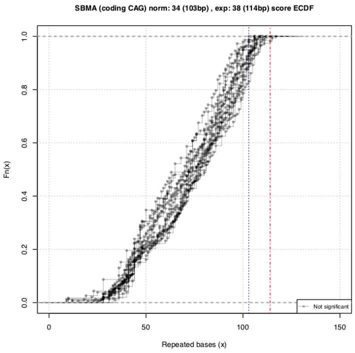
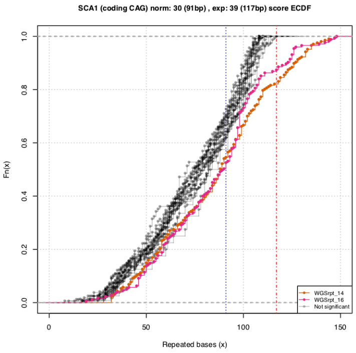
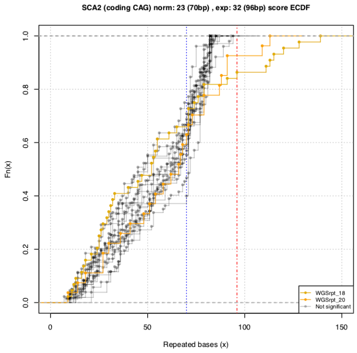


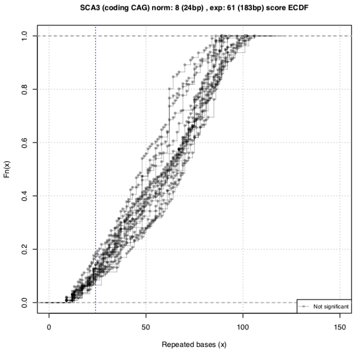
Figure S7. ECDFs for all 21 STR loci for the WGS with PCR cohort (WGS_PCR_1).

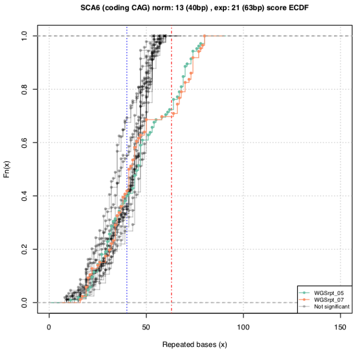

HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_12
Not significant

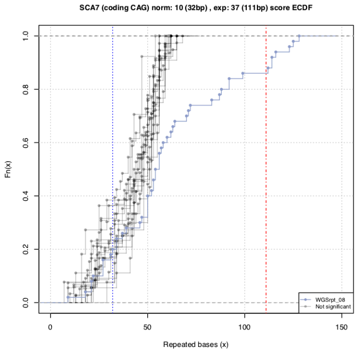

SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
Not significant

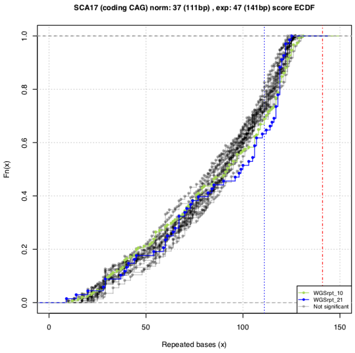

SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_14
WGSrpt_16
Not significant

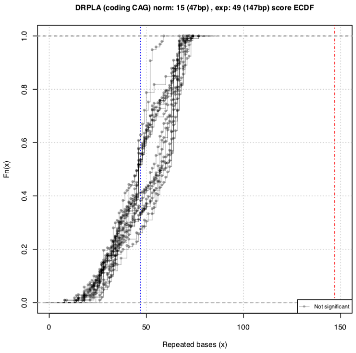

SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_18
WGSrpt_20
Not significant

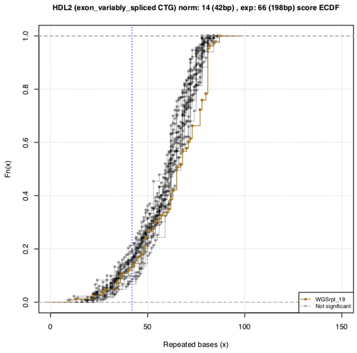

SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
Not significant

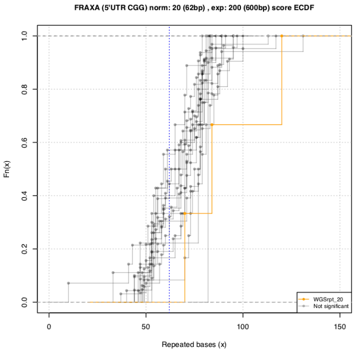

SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_05
WGSrpt_07
Not significant

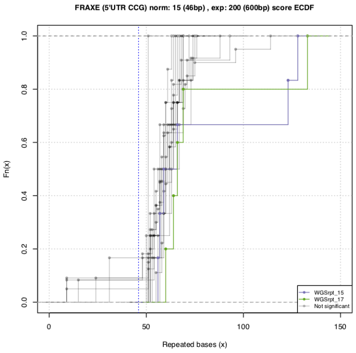

SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_08
Not significant


SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_10
WGSrpt_21
Not significant


DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
Not significant


HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_19
Not significant


FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_20
Not significant


FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
Repeated bases (x)
WGSrpt_15
WGSrpt_17
Not significant

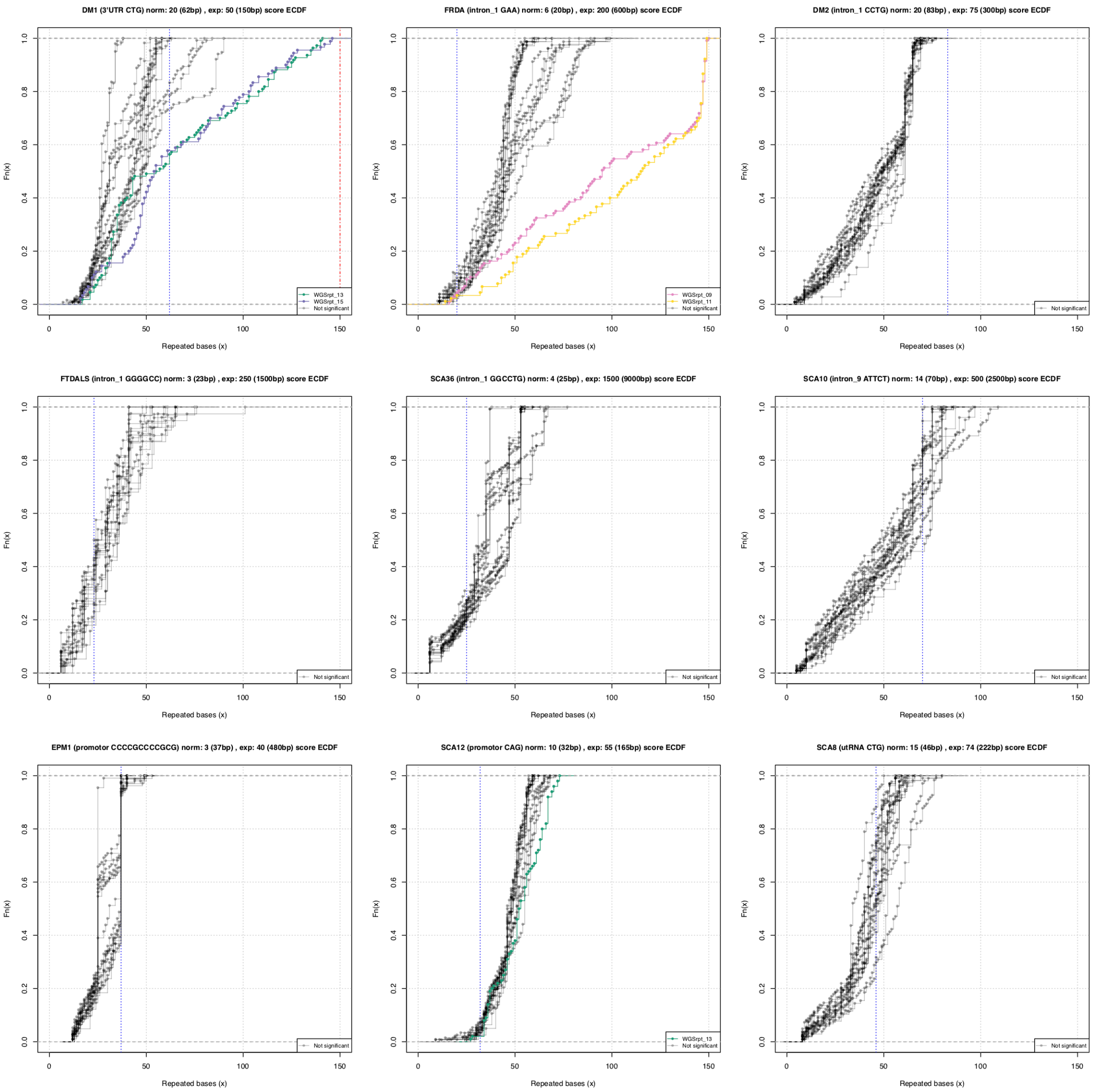
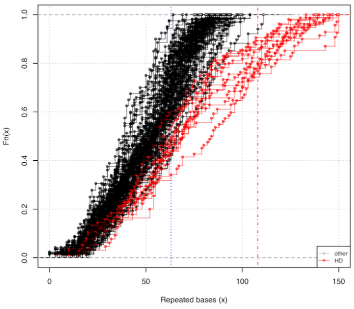
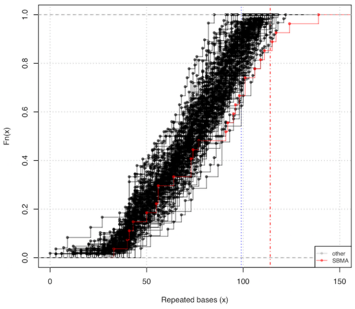


Figure S8. ECDFs for all 21 STR loci for the WGS with PCR cohort (WGS_PCR_2).

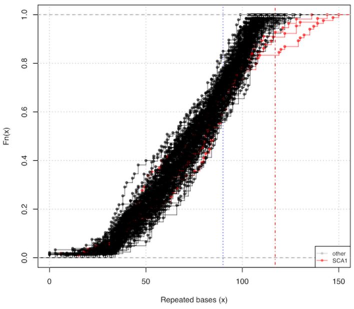
HD (coding CAG) norm: 21 (64bp) , exp: 36 (108bp) score ECDF



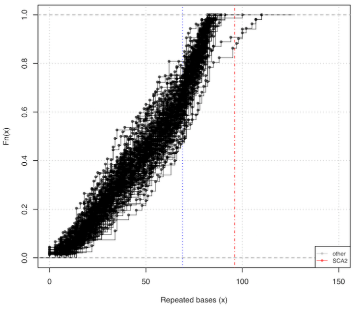
SBMA (coding CAG) norm: 33 (100bp) , exp: 38 (114bp) score ECDF



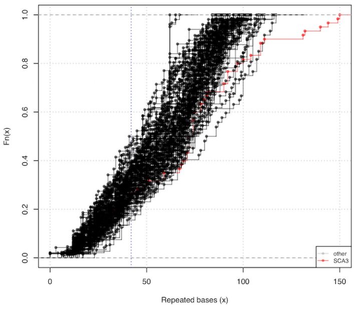
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF



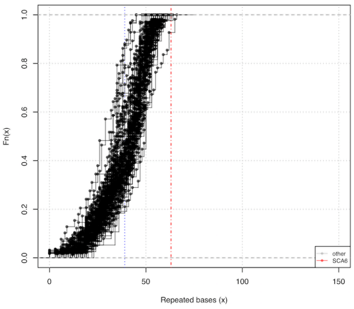
SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF



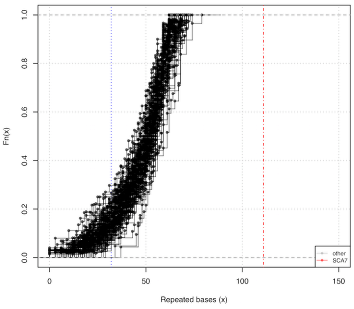
SCA3 (coding CAG) norm: 14 (42bp) , exp: 61 (183bp) score ECDF



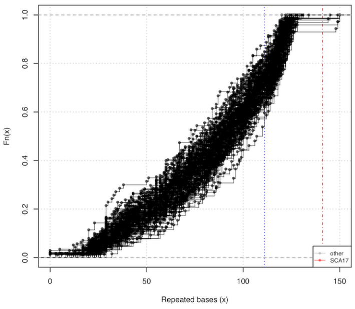
SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF



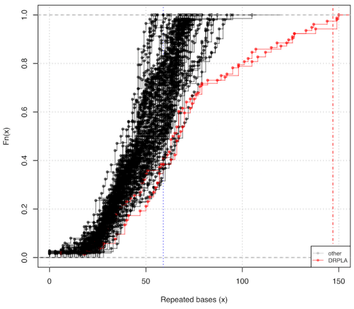
SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF



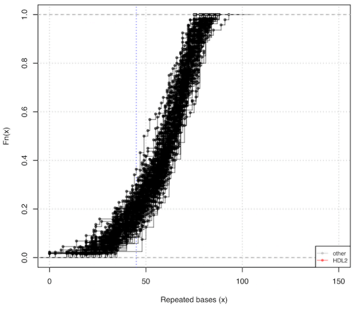
SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF



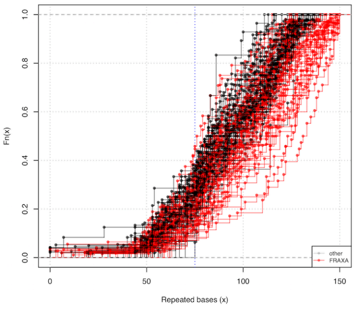
DRPLA (coding CAG) norm: 19 (59bp) , exp: 49 (147bp) score ECDF



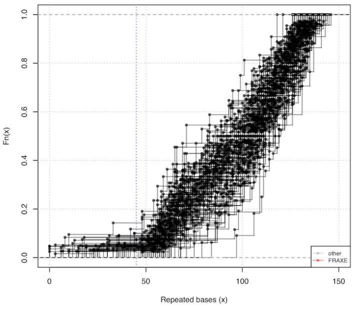
HDL2 (exon_variably_spliced CTG) norm: 15 (46bp) , exp: 66 (198bp) score ECDF



FRAXA (5'UTR CGG) norm: 25 (75bp) , exp: 200 (600bp) score ECDF



FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF



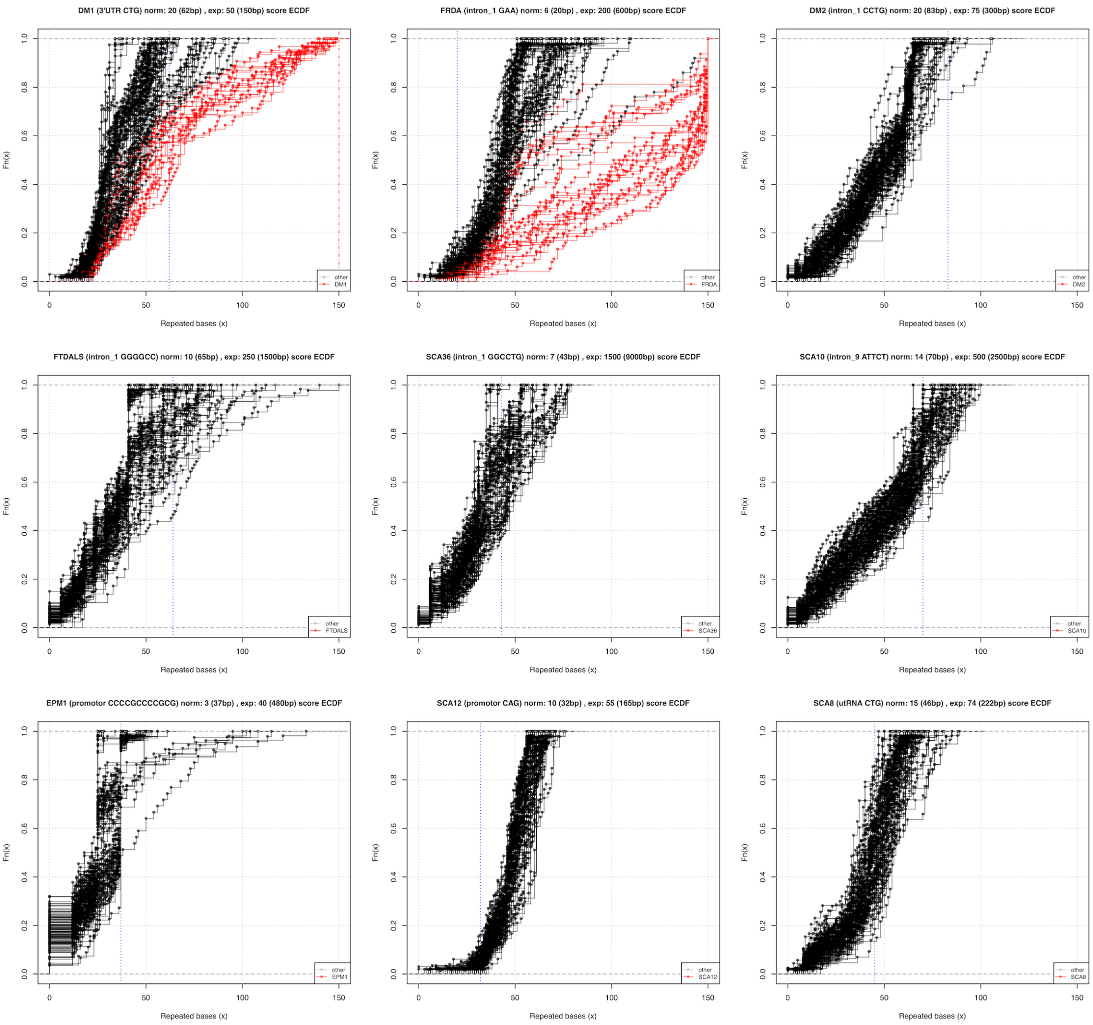
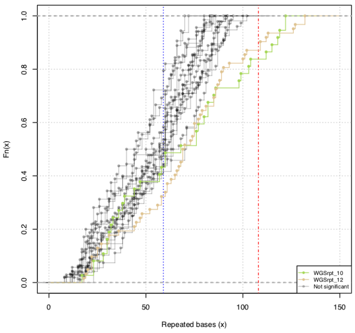
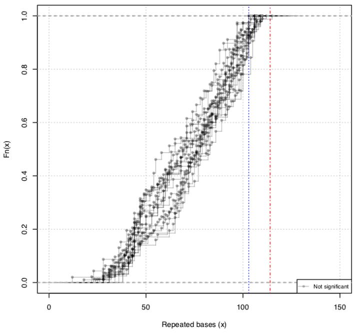


Figure S9. ECDFs for all 21 STR loci for the WGS without PCR cohort (WGS_PF).

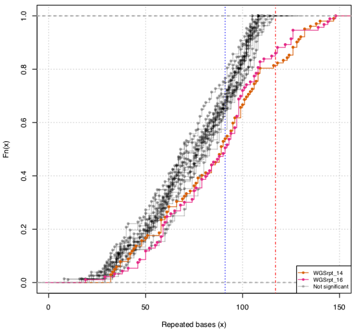
HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF



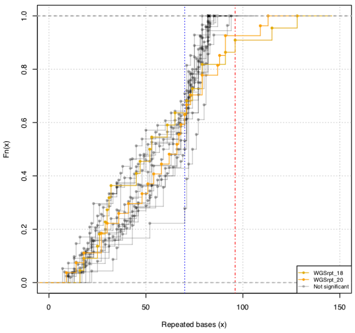
SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF



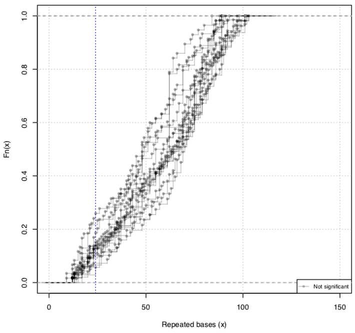
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF



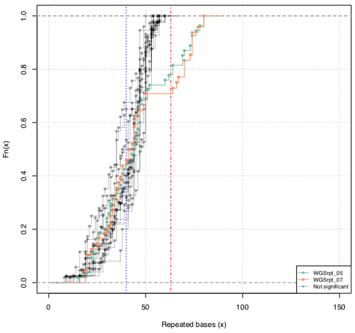
SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF



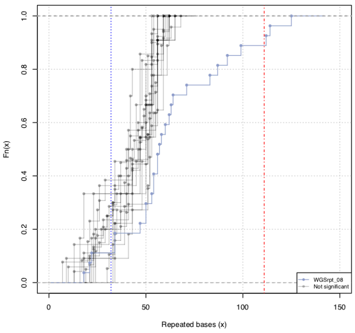
SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF



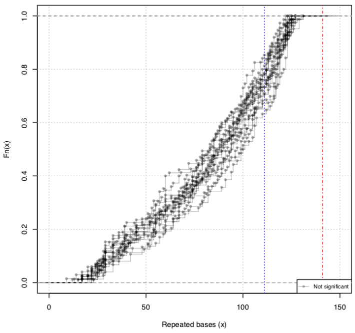
SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF



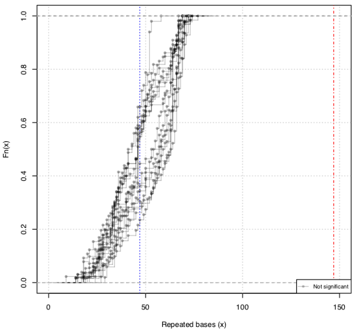
SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF



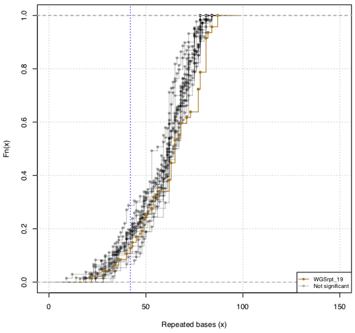
SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF



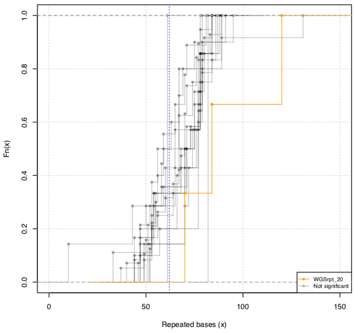
DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF



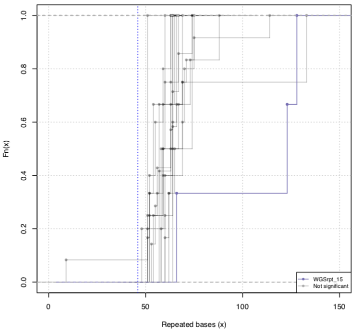
HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF



FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF



FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF



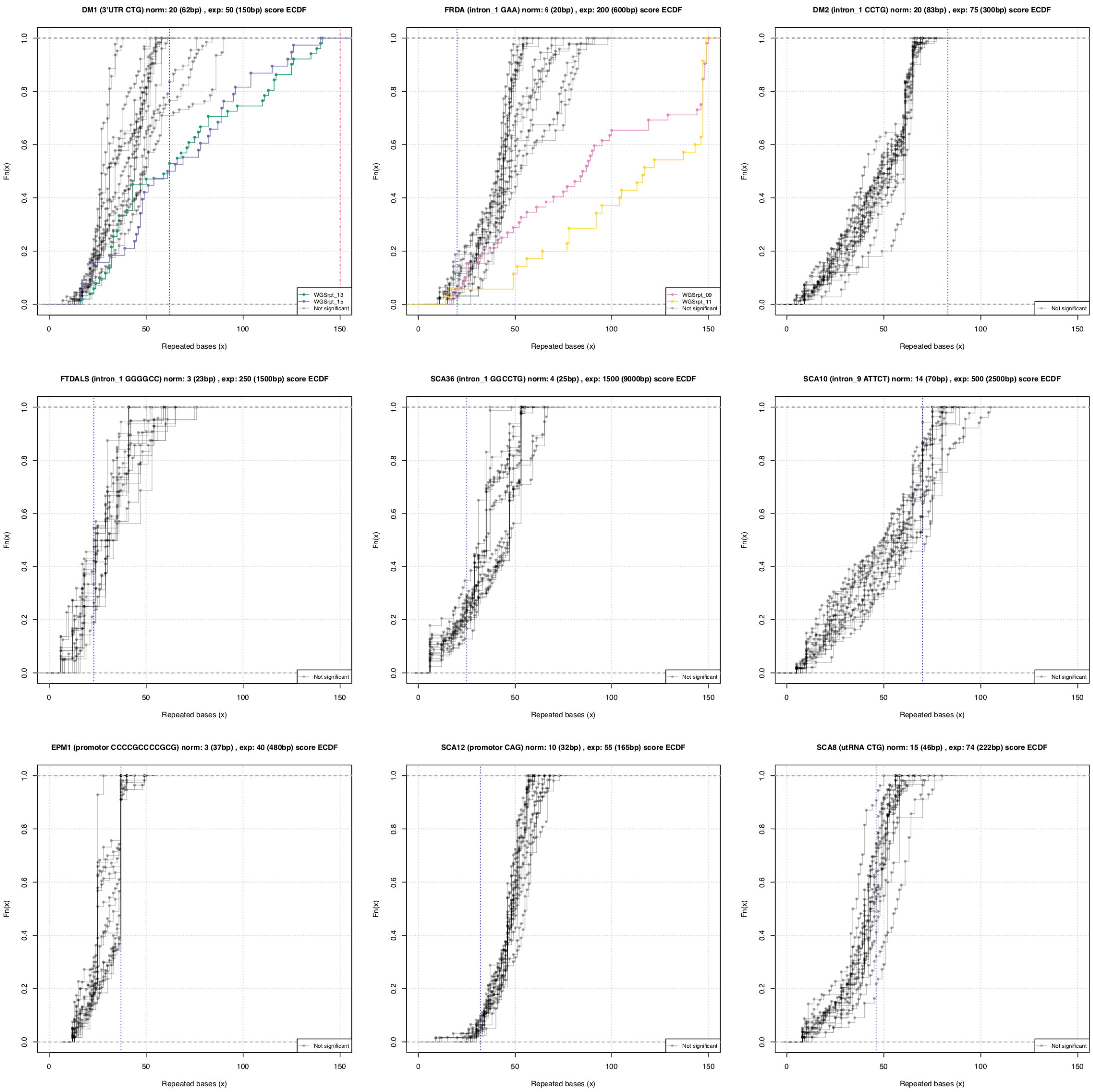
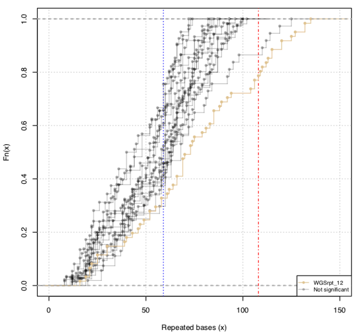
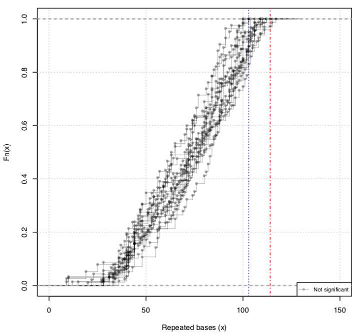


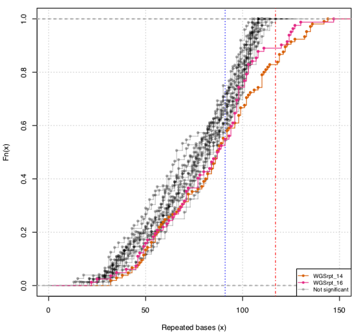
Figure S10. ECDFs for all 21 STR loci for the WGS with PCR 30X sub-cohort (WGS_PCR_2_30X_1).



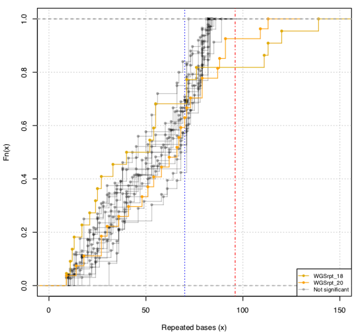
HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF



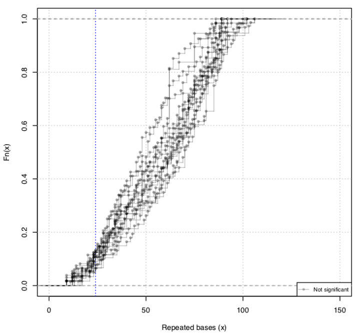
SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF



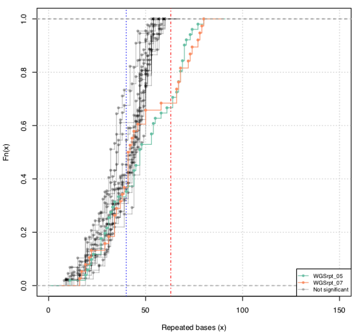
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF



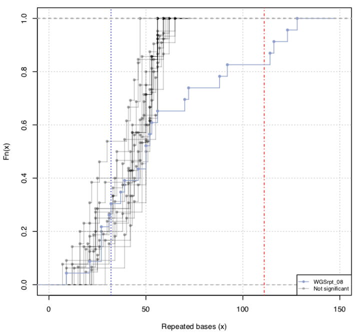
SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF



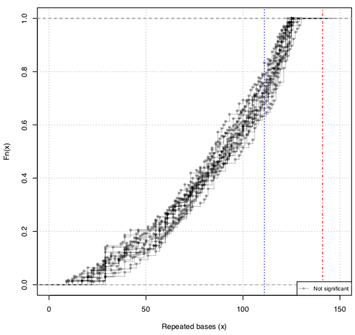
SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF



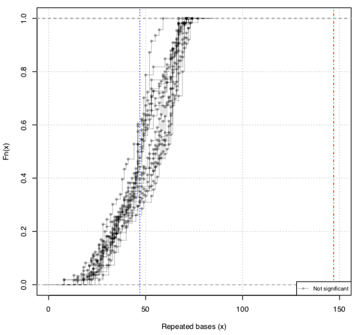
SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF



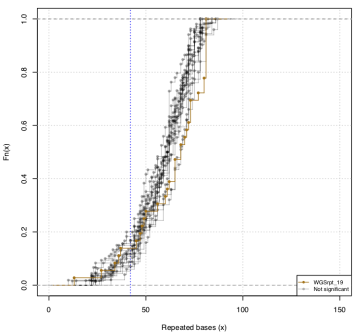
SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF



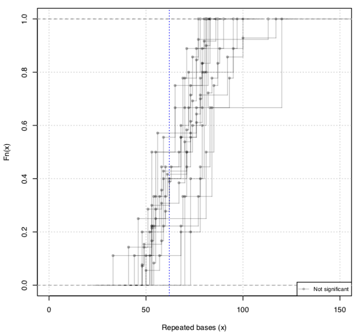
SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF



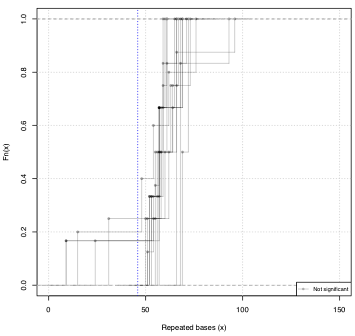
DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF



HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF



FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF



FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF

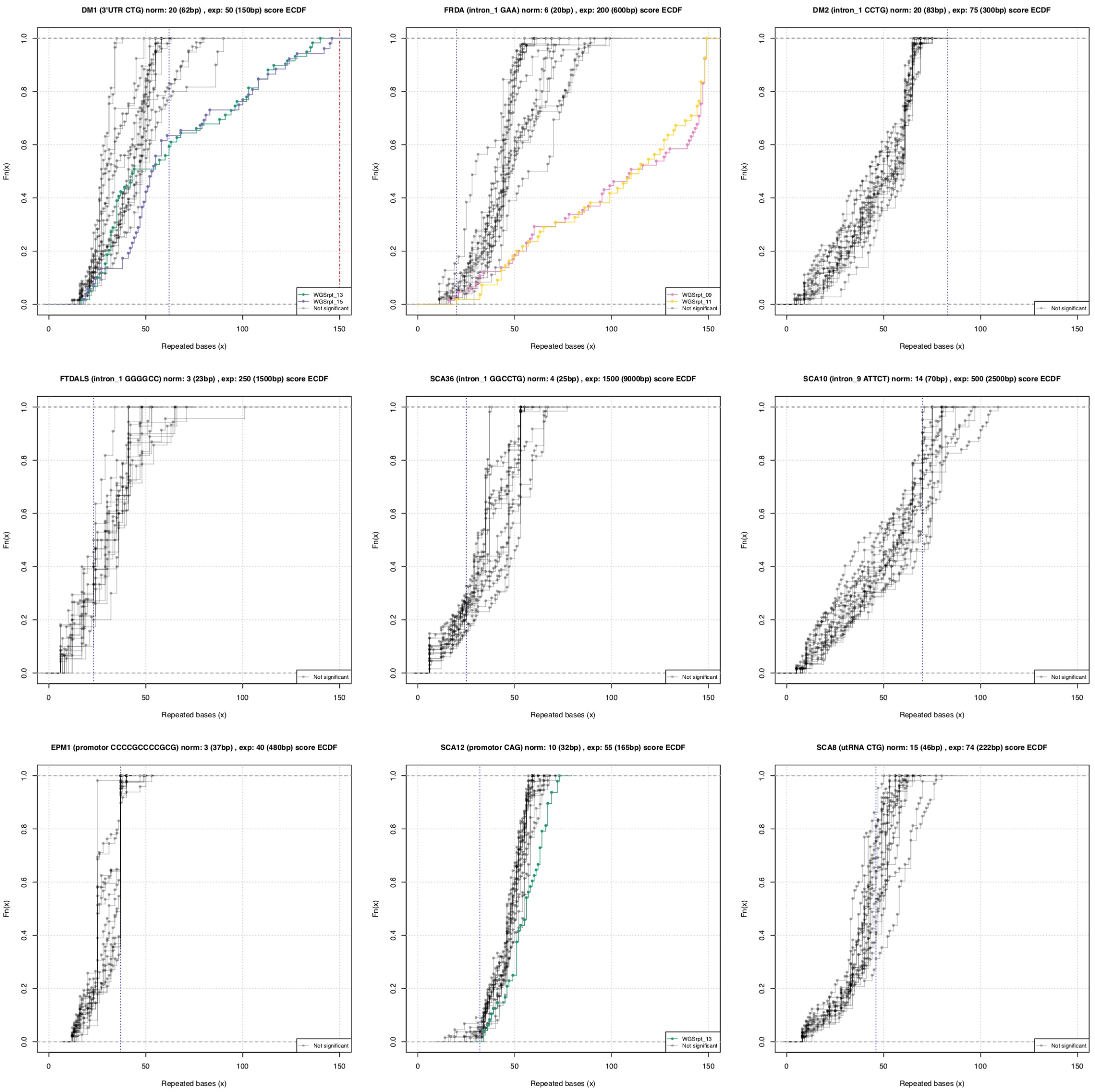


Figure S11. ECDFs for all 21 STR loci for the WGS with PCR 30X sub-cohort (WGS_PCR_2_30X_2).

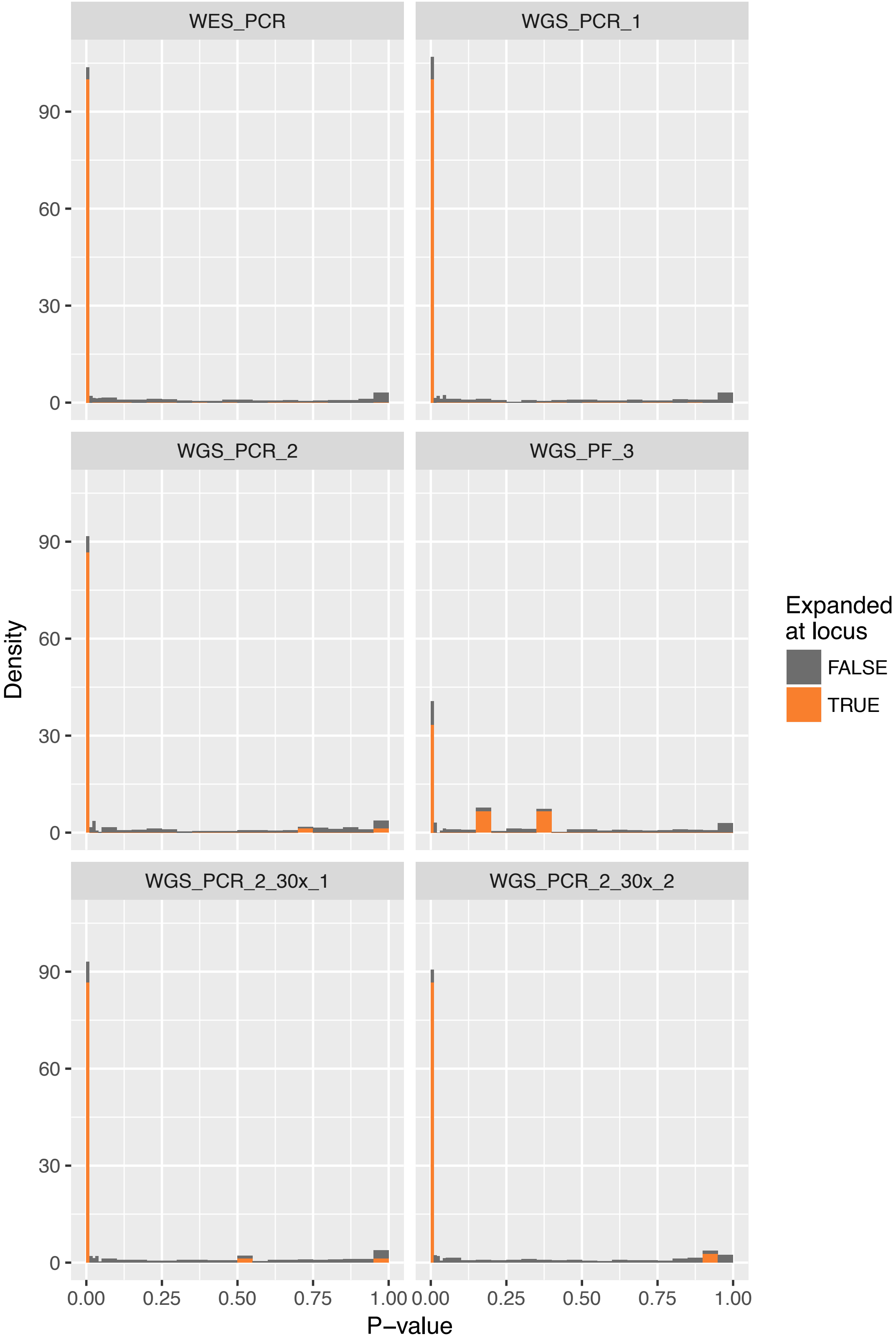


Figure S12. Histograms of the frequency density for the empirically derived p-values for all STR loci for all four cohorts, as well as the two 30X subsets for WGS_PCR_2 (top left panel = WES, top right panel = WGS_PCR_1, middle left = WGS_PCR_2, middle right = WGS_PF_3, bottom left = WGS_PCR_2_30X_1, bottom right = WGS_PCR_30X_2). The bins on the far left, where p<0.05, are plotted at smaller bin sizes of 0.01 whilst other bins were plotted with bin size 0.05 to show greater detail.

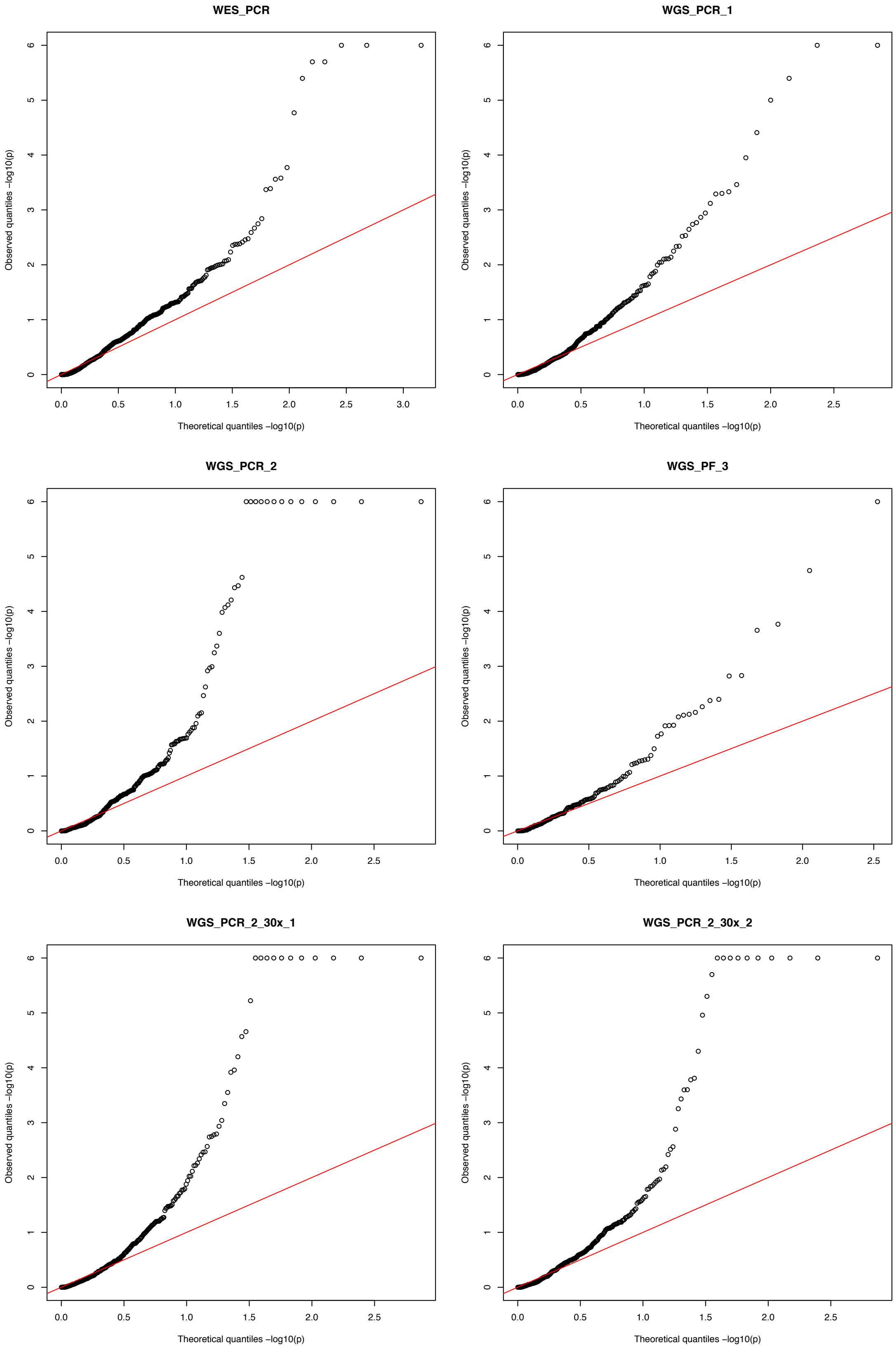


Figure S13. Q-Q plots for the empirically derived p-values for all STR loci for all four cohorts, as well as the two 30X subsets for WGS_PCR_2 (top left panel = WES, top right panel = WGS_PCR_1, middle left = WGS_PCR_2, middle right = WGS_PF_3, bottom left = WGS_PCR_2_30X_1, bottom right = WGS_PCR_30X_2). X-axis has –log10 transformed uniform distribution quantiles, which are plotted against the empirically derived –log10 transformed p-value.

Software	**Publication**	**Computational Burden#** **Known Loci/Genome-wide**	**Statistical Test**	**Reported WGS/WES Analysis capability**	**Software ease of use**	**Ability to search genome wide**	**Graphical Output**	**Length of STR expansion detection bias**
Expansion Hunter	Dolzhenko et al, Genome Research 2017	Low/Low	None – estimates allele sizes. Significance determined based on thresholds.[P]	WGS	High	Possible	No	Repeats with long motifs, e.g. c9orf72^ gain extra evidence for expansion with usage of IRR* reads
TREDPARSE	Tang et al, AJHG, 2017	Low/Unknown	Likelihood of pathogenicity, genetic model, estimates allele sizes[P]	WGS	High	Possible	Yes	Does not detect expansions that exceed its detection threshold (300 repeats)
STRetch	Dashnow et al, Genome Biol, 2018	High/Medium+	Likelihood Ratio Test with reads mapping to decoy. Estimates allele sizes.	WGS	Low	Easy	No	Short expansions may not map to the decoy chromosomes and remain undetected, e.g. SCA6[&]
exSTRa	Tankard et al, this manuscript	Low/Medium	Permutation based outlier detection test	WGS & WES	Medium	Possible	Yes	No known bias

Table S1. Summary of computational methods, evaluation framework and limitations for ExpansionHunter, exSTRa, STRetch and TREDPARSE. #=Computational Burden has been split into two components: known loci (a small subset of all STR loci) and genome-wide, representing thousands of STR loci. **P**=requires prior information for STR in terms of allele size to aid statistical test**.** ^=The C9orf72 repeat expansion is a hexamer repeat. [&]SCA6 is the smallest repeat expansion currently known, *IRR = in read repeat. Updated and adapted from Bahlo et al, F1000Research, 2018,+STRetch is inherently different to the other three methods in runtime since it

requires a realignment of all reads to its augmented reference, hence the “High” computational cost for the known loci. Computational costs for the statistical tests should rise linearly for additional STRs tested, with STRetch and exSTRa more computationally expensive than ExpansionHunter and TREDPARSE because they perform permutation tests to estimate p-values.

Cohort	Sample	Total Reads	Mean	Median	Duplication
WES_PCR	WES_PCR_control_01	139,513,764	96.01	80	6.9%
WES_PCR	WES_PCR_control_02	62,353,356	43.43	36	4.5%
WES_PCR	WES_PCR_control_03	238,172,010	153.26	128	11.5%
WES_PCR	WES_PCR_control_04	58,129,456	40.57	34	4.0%
WES_PCR	WES_PCR_control_05	145,193,758	99.37	83	6.9%
WES_PCR	WES_PCR_control_06	62,134,938	43.95	37	4.1%
WES_PCR	WES_PCR_control_07	50,181,708	35.87	30	5.5%
WES_PCR	WES_PCR_control_08	66,955,438	47.6	40	4.0%
WES_PCR	WES_PCR_control_09	64,382,836	44.42	37	2.5%
WES_PCR	WES_PCR_control_10	30,678,508	18.38	16	1.2%
WES_PCR	WES_PCR_control_11	31,469,068	18.98	16	1.2%
WES_PCR	WES_PCR_control_12	72,726,312	53.08	45	4.0%
WES_PCR	WES_PCR_control_13	72,612,894	53.45	45	4.0%
WES_PCR	WES_PCR_control_14	80,590,976	57.52	49	3.7%
WES_PCR	WES_PCR_control_15	59,659,362	42.4	35	3.5%
WES_PCR	WES_PCR_control_16	59,659,362	42.4	35	3.5%
WES_PCR	WES_PCR_control_17	64,947,428	45.97	38	3.5%
WES_PCR	WES_PCR_control_18	64,947,428	45.97	38	3.5%
WES_PCR	WES_PCR_control_19	61,810,190	43.55	37	4.9%
WES_PCR	WES_PCR_control_20	72,176,900	51.84	44	5.3%
WES_PCR	WES_PCR_control_21	61,188,452	53.02	45	4.4%
WES_PCR	WES_PCR_control_22	78,890,270	55.18	47	4.4%
WES_PCR	WES_PCR_control_23	77,933,824	56.04	48	4.4%
WES_PCR	WES_PCR_control_24	75,209,662	55.68	47	4.0%
WES_PCR	WES_PCR_control_25	106,336,552	79.2	67	9.2%
WES_PCR	WES_PCR_control_26	76,593,848	55.12	47	3.6%
WES_PCR	WES_PCR_control_27	77,592,098	56.92	48	3.9%
WES_PCR	WES_PCR_control_28	114,297,146	76.46	66	13.7%
WES_PCR	WES_PCR_control_29	74,242,926	53.39	45	3.9%
WES_PCR	WES_PCR_control_30	101,662,468	78.86	67	6.8%
WES_PCR	WES_PCR_control_31	113,194,258	82.04	70	11.7%
WES_PCR	WES_PCR_control_32	109,980,714	79.12	68	8.8%
WES_PCR	WES_PCR_control_33	104,260,718	79.85	68	7.2%
WES_PCR	WES_PCR_control_34	58,997,374	44.4	38	2.7%
WES_PCR	WES_PCR_control_35	60,072,178	44.58	38	4.4%
WES_PCR	WES_PCR_control_36	61,760,484	46.91	40	3.5%
WES_PCR	WES_PCR_control_37	56,915,474	42.91	37	3.5%
WES_PCR	WES_PCR_control_38	60,614,514	45.51	39	3.0%
WES_PCR	WES_PCR_control_39	55,326,730	41.05	35	3.4%
WES_PCR	WES_PCR_control_40	62,545,440	45.27	38	4.0%
WES_PCR	WES_PCR_control_41	58,499,634	42.61	36	4.1%
WES_PCR	WES_PCR_control_42	58,035,986	42.02	35	3.9%
WES_PCR	WES_PCR_control_43	65,329,052	45.2	38	3.7%
WES_PCR	WES_PCR_control_44	62,781,160	43.44	37	6.2%
WES_PCR	WES_PCR_control_45	58,649,916	42.6	35	4.5%
WES_PCR	WES_PCR_control_46	63,582,040	44.41	37	4.3%
WES_PCR	WES_PCR_control_47	89,591,992	52.39	44	2.6%
WES_PCR	WES_PCR_control_48	87,561,816	52.46	44	2.9%
WES_PCR	WES_PCR_control_49	97,835,338	58.1	48	2.9%
WES_PCR	WES_PCR_control_50	89,557,392	53.19	45	2.6%
WES_PCR	WES_PCR_control_51	101,165,530	70.66	60	12.5%
WES_PCR	WES_PCR_control_52	114,720,190	83.33	71	10.2%
WES_PCR	WES_PCR_control_53	108,440,198	77.59	66	9.8%
WES_PCR	WES_PCR_control_54	59,788,846	43.79	37	3.9%
WES_PCR	WES_PCR_control_55	56,578,500	39.12	33	3.8%
WES_PCR	WES_PCR_control_56	63,339,278	44.18	37	3.7%
WES_PCR	WES_PCR_control_57	60,093,432	41.9	36	2.9%
WES_PCR	WES_PCR_control_58	106,707,804	79.01	67	9.8%
WES_PCR	WES_PCR_control_59	106,570,138	80.02	68	9.4%
WES_PCR	WES_PCR_control_60	67,782,869	72.1	61	6.7%
WES_PCR	rptWEHI1	93,689,702	57.49	48	3.2%
WES_PCR	rptWEHI2	96,342,624	58.09	48	3.1%
WES_PCR	rptWEHI3	85,887,382	54.97	46	3.2%
WES_PCR	rptWEHI4	80,398,670	56.56	48	3.9%
WGS_PCR_1	HD-1	1,490,961,246	66.1	69	37.4%
WGS_PCR_1	SCA2-1	1,452,983,981	64.44	66	37.8%
WGS_PCR_1	SCA6-1	1,585,248,814	70.73	73	35.3%
WGS_PCR_1	WGS_PCR_1_control_01	996,511,742	46.02	48	24.5%
WGS_PCR_1	WGS_PCR_1_control_02	770,818,821	35.47	37	42.6%
WGS_PCR_1	WGS_PCR_1_control_03	1,061,318,492	48.06	50	31.3%
WGS_PCR_1	WGS_PCR_1_control_04	1,116,929,170	48.87	50	28.0%
WGS_PCR_1	WGS_PCR_1_control_05	963,162,036	43.55	45	35.3%
WGS_PCR_1	WGS_PCR_1_control_06	1,083,837,380	47.95	49	29.9%
WGS_PCR_1	WGS_PCR_1_control_07	1,034,524,662	44.63	46	32.6%
WGS_PCR_1	WGS_PCR_1_control_08	1,600,013,709	72.37	74	37.4%
WGS_PCR_1	WGS_PCR_1_control_09	1,600,013,709	72.37	74	37.4%
WGS_PCR_1	WGS_PCR_1_control_10	1,437,787,592	65.49	67	44.4%
WGS_PCR_1	WGS_PCR_1_control_11	1,437,787,592	65.49	67	44.4%
WGS_PCR_1	WGS_PCR_1_control_12	1,450,901,977	64.07	66	42.3%
WGS_PCR_1	WGS_PCR_1_control_13	1,751,030,705	81.14	83	32.6%
WGS_PCR_1	WGS_PCR_1_control_14	1,646,345,811	75.79	78	29.1%
WGS_PCR_1	WGS_PCR_1_control_15	1,155,693,820	53.47	56	31.9%
WGS_PCR_1	WGS_PCR_1_control_16	1,067,537,829	48.86	51	31.3%
WGS_PCR_2	WGS_PCR_2_control_01	1,690,757,788	77.21	79	12.7%
WGS_PCR_2	WGS_PCR_2_control_02	1,670,045,093	77.36	79	14.4%
WGS_PCR_2	WGSrpt_05	1,763,448,305	83.02	85	14.1%
WGS_PCR_2	WGSrpt_07	1,743,429,928	84.94	87	13.7%
WGS_PCR_2	WGSrpt_08	1,714,347,858	83.09	84	15.1%
WGS_PCR_2	WGSrpt_09	1,758,081,790	81.38	84	12.1%
WGS_PCR_2	WGSrpt_10	1,764,184,511	81.53	83	12.7%
WGS_PCR_2	WGSrpt_11	1,711,175,531	79.09	82	12.8%
WGS_PCR_2	WGSrpt_12	1,519,487,865	72.69	75	14.3%
WGS_PCR_2	WGSrpt_13	1,626,730,877	76.37	78	11.9%
WGS_PCR_2	WGSrpt_14	1,759,223,150	86.55	88	8.0%
WGS_PCR_2	WGSrpt_15	1,582,360,421	73.38	76	14.6%
WGS_PCR_2	WGSrpt_16	1,747,015,570	84.7	87	13.2%
WGS_PCR_2	WGSrpt_17	1,672,344,799	79.15	82	10.4%
WGS_PCR_2	WGSrpt_18	1,705,757,541	81.24	83	14.8%
WGS_PCR_2	WGSrpt_19	1,550,742,464	70.15	72	13.4%
WGS_PCR_2	WGSrpt_20	791,723,778	35.85	37	13.2%
WGS_PCR_2	WGSrpt_21	654,118,132	30.35	31	30.2%

Table S2: Coverage and alignment statistics for samples from cohorts WES, WGS_PCR_1 and WGS_PCR_2.

STR Locus	Reference
Huntington Disease (HD)	Rubinsztein, David C., et al. "Phenotypic characterization of individuals with 30–40 CAG repeats in the Huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36–39 repeats." *American journal of human genetics* 59.1 (1996): 16.
Kennedy Disease (SBMA)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.
Spinocerebellar ataxia 1 (SCA1)	Ranum, Laura PW, et al. "Molecular and clinical correlations in spinocerebellar ataxia type I: evidence for familial effects on the age at onset." *American journal of human genetics* 55.2 (1994): 244.
Spinocerebellar ataxia 2 (SCA2)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.
Machado- Joseph disease (SCA3)	Limprasert, Pornprot, et al. "Analysis of CAG repeat of the Machado-Joseph gene in human, chimpanzee and monkey populations: a variant nucleotide is associated with the number of CAG repeats." *Human molecular genetics* 5.2 (1996): 207-213.
Spinocerebellar ataxia 2 (SCA6)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.
Spinocerebellar ataxia 2 (SCA7)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.

Spinocerebellar ataxia 2 (SCA17)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.
Dentatorubral-pallidoluysian atrophy (DRPLA/ATN1)	Butland, Stefanie L., et al. "CAG-encoded polyglutamine length polymorphism in the human genome." *BMC genomics* 8.1 (2007): 126.
Huntington disease-like 2 (HDL2)	Seixas, Ana I., et al. "Loss of junctophilin-3 contributes to huntington disease-like 2 pathogenesis." *Annals of neurology* 71.2 (2012): 245-257.
Fragile-X site A (FRAXA)	Fu, Ying-Hui, et al. "Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox." *Cell* 67.6 (1991): 1047-1058.
Fragile-X site E (FRAXE)	Knight, S. J., et al. "Triplet repeat expansion at the FRAXE locus and X-linked mild mental handicap." *American journal of human genetics* 55.1 (1994): 81.
Myotonic dystrophy 1 (DM1)	Magaña, J. J., et al. "Distribution of CTG repeats at the DMPK gene in myotonic distrophy patients and healthy individuals from the Mexican population." *Molecular biology reports* 38.2 (2011): 1341-1346.
Friedreich ataxia (FRDA)	Montermini, Laura, et al. "The Friedreich ataxia GAA triplet repeat: premutation and normal alleles." *Human molecular genetics* 6.8 (1997): 1261-1266.
Myotonic dystrophy 2 (DM2)	Liquori, Christina L., et al. "Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9." *Science* 293.5531 (2001): 864-867.

Amyotrophic lateral sclerosis-frontotemporal dementia (FTDALS)	DeJesus-Hernandez, Mariely, et al. "Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS." *Neuron* 72.2 (2011): 245-256.
Spinocerebellar ataxia 36 (SCA36)	García-Murias, María, et al. "'Costa da Morte'ataxia is spinocerebellar ataxia 36: clinical and genetic characterization." *Brain* 135.5 (2012): 1423-1435.
Spinocerebellar ataxia 10 (SCA10)	Matsuura, Tohru, et al. "Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10." *Nature genetics* 26.2 (2000): 191-194.
Spinocerebellar ataxia 12 (SCA12)	Holmes, Susan E., et al. "Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12." *Nature genetics* 23.4 (1999): 391-392.

Table S3: Literature sources for expansion distributions for all 21 STR loci

OMIM	Model	Gene	Capture	Location	strcat_all	chrom	start	end	strand
309550	X	FMR1	Yes	Xq27.3	http://strcat.teamerlich.org/chart/chrX/146993555/146993629	chrX	146,993,554	146,993,629	+
309548	X	FMR2	Yes	Xq28	http://strcat.teamerlich.org/chart/chrX/147582125/147582273	chrX	147,582,158	147,582,204	+
229300	AR	FXN	No	9q13	http://strcat.teamerlich.org/chart/chr9/71652201/71652220	chr9	71,652,200	71,652,220	+
160900	AD	DMPK	No	19q13	http://strcat.teamerlich.org/chart/chr19/46273463/46273524	chr19	46273462	46273524	-
602668	AD	ZNF9/CNBP	No	3q21.3	http://strcat.teamerlich.org/chart/chr3/128891420/128891502	chr3	128891419	128891502	-
603516	AD	ATXN10	No	22q13.31	http://strcat.teamerlich.org/chart/chr22/46191235/46191304	chr22	46191234	46191304	+
254800	AR	CSTB	Single	21q22.3	http://strcat.teamerlich.org/chart/chr21/45196324/45196360	chr21	45196323	45196360	-
143100	AD	HTT	Yes	4p16.3	http://strcat.teamerlich.org/chart/chr4/3076604/3076667	chr4	3076603	3076667	+
313200	X	AR	Yes	Xq12	http://strcat.teamerlich.org/chart/chrX/66765159/66765261	chrX	66765158	66765261	+
164400	AD	ATXN1	Yes	6p23	http://strcat.teamerlich.org/chart/chr6/16327865/16327955	chr6	16327864	16327955	-
183090	AD	ATXN2	Yes	12q24	http://strcat.teamerlich.org/chart/chr12/112036754/112036823	chr12	112,036,753	112,036,823	-
109150	AD	ATXN3	Yes	14q32.1	http://strcat.teamerlich.org/chart/chr14/92537355/92537396	chr14	92537354	92537396	-
183086	AD	CACNA1A	Yes	19p13	http://strcat.teamerlich.org/chart/chr19/13318673/13318712	chr19	13318672	13318712	-
164500	AD	ATXN7	Yes	3p14.1	http://strcat.teamerlich.org/chart/chr3/63898361/63898392	chr3	63898360	63898392	+
607136	AD	TBP	Yes	6q27	http://strcat.teamerlich.org/chart/chr6/170870995/170871105	chr6	170870994	170871105	+
125370	AD	DRPLA/ATN1	Yes	12p13.31	http://strcat.teamerlich.org/chart/chr12/7045880/7045938	chr12	7045879	7045938	+
608768	AD	ATXN8OS	No	13q21	http://strcat.teamerlich.org/chart/chr13/70713516/70713561	chr13	70713515	70713561	+
604326	AD	PPP2R2B	No	5q32	http://strcat.teamerlich.org/chart/chr5/146258291/146258322	chr5	146258290	146258322	-
606438	AD	JPH3	Single	16q24.3	http://strcat.teamerlich.org/chart/chr16/87637889/87637935	chr16	87637888	87637935	+
105550	AD	C9orf72	No	9p21	http://strcat.teamerlich.org/chart/chr9/27573483/27573544	chr9	27,573,482	27,573,544	-
614153	AD	NOP56	Yes	20p13	http://strcat.teamerlich.org/chart/chr20/2633379/2633421	chr20	2,633,378	2,633,421	+

Table S4: Bait Capture information for WES data, generated using the Agilent V5+UTR capture platform. Model refers to the genetic model, with AD = autosomal dominant, X = X-linked, AR = autosomal recessive. Bait information is given in the Agilent SS V5+UTR column with "Yes" indicating presence of a pair of baits, with on each side of the STR locus, "No" no baits, and "Single" indicating a single bait, only on one side. The ability to capture sequence is determined by whether sequencing 'baits' are in the vicinity (within ~50 bps) of the STR. Strcat gives the location to the STR catalogue generated by Willems et al. Chrom, start and end refer to physical map co-ordinates according to hg19.

Table S5: Individual level expansion call results for cohorts WES, WGS_PCR_1, WGS_PCR_2, and split WGS_PCR_2 cohorts for exSTRa, ExpansionHunter. BF, Bonferroni correction, performed correcting for 21 STR loci tested; mismatch calls are shown in bold; NC, Not Called, meaning no expanded STR was detected; TREDPARSE –L, TREDPARSE expansion calls based on likelihood; TREDPARSE-T, TREDPARSE expansion calls based on threshold. Available as an Excel spreadsheet (SupplementaryTable_S4.xlsx).

Table S6: WGS_Pf_3 analysis results comparing exSTRa, ExpansionHunter, STRetch, TREDPARSE. Per sample expansion calls for 118 WGS samples. Available as an Excel spreadsheet (SupplementaryTable_S5.xlsx).

Alignment

Alignment of each pair of FASTQ files was performed with Bowtie2[1] to the hg19 human genome reference build in very sensitive local mode, with maximum insert sizes of 800 bp for WES samples and 1000 bp for WGS samples. BAM files were sorted and merged with the Novosort tool. Duplicate marking was performed with Picard. Local realignment and base score recalibration was performed with the GATK IndelAligner tool and the Base Quality Score Recalibration tool[2] to produce input ready BAM files.

Software

The first step of the analysis is performed with a Perl module, called Bio::STR::exSTRa, which carries out a heuristic procedure to extract repeat content. In summary, this procedure uses the data from the reference database for the 21 loci presented in Table 1 to identify all reads that map to each of the STR loci, for each individual to be examined. The number of repeat motifs contained by each read are determined by the heuristic procedure, which examines each read for the repeat units that that STR is known to contain. This allows for some mismatches due to impure repeats and sequencing errors. Additionally, this is more computationally efficient than determining the exact repeat start and end and is more robust as determining the edge of the repeat can be difficult near the end of a read in the presence of mismatches.

Bio::STR::exSTRa : A heuristic procedure to extract repeat units per read

For simplicity, the following description of the data and analysis methods is only for a single locus. The algorithm is repeated independently at each locus.

Read information is extracted from a database of STR locations, such as 2–6bp repeat unit features generated using the Tandem Repeats Finder[3], which is also available as the Simple Repeats track of UCSC Genome Browser. Information is extracted for one STR at a time, with the following algorithm repeated for each STR:

1. The method identifies ‘anchor’ reads that facilitates identifying reads within or overlapping the STR. To qualify as an anchor, the reads are required to map within 800 bp of the STR, with the anchor orientated towards the STR. An anchor may overlap the STR.

2. The anchor-mate mapping is checked. If the anchor-mate is mapped near the STR and is not overlapping or adjacent, then the read is discarded, while those reads overlapping the STR are taken forward to the next analysis step. Sometimes the read is unmapped, or mapped to another locus, which is then recovered for further interrogation in the next step.

3. Remaining anchor-mates have their sequence content matched for the presence of the repeat unit in the correct direction, allowing for the repeat to start at any base, or phase, of the repeat unit. For example, if the repeat unit is CAG, the method can also match AGC and GCA. The number of bases found to be part of the repeat unit is counted to derive a repeat-score for that read, that is designated at a given locus as x_{ij} for sample i and read j (note that the maximum defined j depends on the sample). If both ends of a read-pair overlap within an STR, both reads undergo this procedure and each end is given a score that can be resolved during the statistical analysis of the data

(the implementation in this paper did not investigate resolving these further, with both ends left in the analysis if any). An example of matching (lower case) a CAG on the opposite strand, thus matching CTG at any starting base, or phase, of the motif, i.e. CTG, TGC and GCT:

CGTTCAC**ctg**GATGTGAACT**ctg**TC**ctg**ATAGGTCCCC**ctgctgctgctgctgctgctg**Tt
gctgcTTTt**gctgc**TGT**ctg**AAA

This 87 bp sequence has 48 bp marked (bold and lower case) as part of the repeat.

4. The method filters out reads where the score is lower than expected in random nucleotide sequences. While not precisely true, the assumption applied is that the four nucleotides are uniformly distributed and independent with respect to other positions. Short motifs are more likely to appear by chance. The method filters out scores where $x_{ij}<lk/4^k$, where l is the read length and k is the motif length. 800 bp has been chosen to avoid discarding reads overlapping the STR, with the insert size of read pairs having median ~360 bp. Some protocols may need to analyse reads further than 800 bp. This can be adjusted when calling the Perl module.

The output of this Perl module consists of a tab-delimited file consisting of a table where each row in the table is the repeat content of any read from a particular individual that has been identified as mapping to an STR locus that was to be investigated.

Note that these data do not represent the true size of the allele that the read has captured but where the method predicts an individual with repeat expansion allele at a particular STR locus to show an excess of reads and read content mapping to that STR.

R package exSTRa : detecting outlier distributions of repeat content in reads

Analysis methods for the second part of the analysis method are embedded in an R package, called exSTRa (expanded STR algorithm). The output data from step 1 can be loaded and the data visualized. In particular visualizations of the data are performed with empirical cumulative distribution functions, or ECDFs.

The analysis of the samples is treated as an outlier detection problem. For the N individuals in the cohort the method compares each individual in turn to all others, including itself for robustness, for all STR loci that will be tested for repeat expansions. Since more reads with greater numbers of the repeat motif will be visible in an individual with a repeat expansion at a particular locus, the data at the repeat locus being interrogated is used in a statistical test of a difference of distribution in number of repeats that are observed for a particular individual in comparison to the set of controls. Individuals with an expanded repeat demonstrate a shift in the distribution in comparison to individuals with normal size alleles comprising their genotype for the STR locus being examined. To visualize the results, the output is plotted as empirical cumulative distribution functions (ECDFs) in R.

Statistical Test

We developed a statistical test to detect outlier samples in comparison to a background set of samples. These outlier samples are likely to be individuals harbouring repeat expansions. To apply this test the method utilizes an empirical quantile imputation procedure, implemented in the R function quantile(). This function calculates empirical quantiles for any desired probability, for example probability = 0.5 generates the median observation in a dataset, but it is also capable of generating quantiles at probability points that have not been observed, by interpolating the probability distribution function based on the empirical observations. We make use of this function to firstly generate the same number of 'observations' for all samples to be tested, defined as M. In general, n is defined so that it is the largest number of observations for all of the samples, but other values could also be chosen, such as the median number of observations. The R function quantile() is applied to generate this dataset which consists of N samples, with M observations/quantiles, leading to a dataset with N by M datapoints, or quantiles. This dataset is defined as $Y=(y_{ij})$, where y_{ij} is the repeat content of the j^{th} quantile from the i^{th} individual.

The test statistic, which we call T_i, is defined as the average of multiple t-statistics generated at each quantile j, above a preset threshold $0 \leq h < 1$, which we usually define $h = 0.5$.

$$T_i = \frac{1}{D} \sum_{j:Pr(y_{ij}) \geq h}^{M} t_{ij}$$

$$D = |\{j : Pr(y_{ij}) \geq h\}|$$

Sixteen of the 21 STR repeat expansion loci to be examined have a dominant mode of inheritance, with only one copy of the expanded allele. This can be observed with the ECDF plots for the autosomal dominant STR loci, where deviations in the repeat composition of reads are only noticeable after the median quantile, when the y-axis (which is the probability) exceeds 0.5. Observations below this threshold are likely to carry no signal, and are thus would not contribute to any test statistic attempting to discriminate between expansions and normal sized alleles.

Each quantile test statistic, $t_{ij,}$ is calculated similarly to a two-sample T-test like test statistic, but using a trimmed mean and variance, to robustly allow for the occurrence of more than one expansion in the background distribution, which is the case in the cohorts we tested but which will also likely be the case in other cohorts. The trimming percentage, or percentage of samples that are used is a parameter that can be set by the user in exSTRa, but the default is set at 0.15. Trimming is performed bilaterally, for both the lower and upper tails of the distributions, resulting in at least 30% of the samples being trimmed.

$$\begin{aligned}
t_{ij} &= \frac{y_{ij} - m_j}{S_j} \\
m_j &= \frac{1}{n_j} \sum_{j:l_j \leq y_{ij} \leq u_j} y_{ij} \\
n_j &= |\{j : l_j \leq y_{ij} \leq u_j\}| \\
S_j &= s_j \sqrt{1 + \frac{1}{n_j}}
\end{aligned}$$

where l_i is the first observation included from the lower tail of the distribution after the trimmed observations and u_i the last observation included from the upper tail of the distribution, with all observations beyond this trimmed. s_j is the sample standard deviation of the trimmed samples.

We derive p-values for these test statistics using a simulation procedure.

Since the number of individuals in our simulations is not large and only test a single individual, standard permutation tests will not result in sufficient sampling of the empirical distribution thus resulting in a very coarse-grained empirical distribution. Instead we take advantage of the well-described empirical distributions of the samples by directly simulating from the background distribution, which represents the distribution of normal, or non-expanded alleles. We perform this using robust methods to ensure that samples with expanded alleles do not influence the simulation in the simulation study.

For simulation s we simulate M quantiles for N samples, by assuming that the distributions at each quantile follow large sample theory and are thus approximately normally distributed with mean m_j and standard deviation d_j, where j denotes the quantile. The method then tests this assumption by performing visual inspections of the distribution of quantiles after standardization with the R function qqnorm() and the approximation was reasonable.

The method then uses the median as our estimator for the mean, and the median absolute deviation (MAD) as our robust estimator for the standard deviation. Thus,

$$\begin{aligned}
\hat{m}_j &= median\{\mathbf{y}_{.\mathbf{j}}\} \\
\hat{d}_j &= \frac{1}{(\Phi^{-1}(3/4))} MAD\{\mathbf{y}_{.\mathbf{j}}\} \\
MAD\{\mathbf{y}_{.\mathbf{j}}\} &= median\{|y_{ij} - median\{y_{.j}\}|\}
\end{aligned}$$

Where $y_{.j} = \{y_{1j}, \ldots, y_{Nj}\}$, and $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal distribution. The R function mad() incorporates the scaling factor that ensures consistency with the standard deviation when observations are normally distributed.

The method then uses the rnorm() function in R to randomly generate the N new observations for each quantile, using the STR locus and quantile specific estimators for the mean and standard deviation. The data is then sorted for each sample, as some of the new observations are no longer monotonically increasing as per definition of quantiles.

Finally, the test statistic T_s is calculated as defined above, but using the new data set generated from the simulation, where the first sample in the simulated data set is arbitrarily chosen to be the sample to be tested as an outlier. The method then repeat this for a desired number of simulations, say B, and then calculates the empirical p-value for our test statistic pT_i using standard methods, where:

$$p_{T_i} = \frac{\sum_{s=1}^{B} I([T_i > T_1^s]) + 1}{B + 1}$$

Here $I(.)$ is the indicator function. T_1^s is the test statistic for the dataset. The method calls individuals as expanded or not for each STR locus examined based on a Bonferroni corrected threshold at the 0.05 significance level, based on the number of STR tested for each sample.

Standard deviations for the empirical p-value estimator were also calculated as follows.

$$\begin{aligned} SD(\hat{p}) &= \sqrt{\frac{\frac{1+\sum_{i=1}^{B} x_i}{B+1}(1 - \frac{\sum_{i=1}^{B} x_i}{B+1})}{B}} \\ x_i &= I([T_i > T_1^S]) \end{aligned}$$

Calling expansions with ExpansionHunter, STRetch and TREDPARSE

We performed analysis with ExpansionHunter (version 2.5.3)[4], STRetch (GitHub commit 94d0516)[5] and TREDPARSE (GitHub commit 83881b4)[6], on the cohorts at the 21 repeat expansion loci listed in Table 1. The input data was the same BAM files generated as described above. Only specification files (in JSON format) for the DM1, DRPLA, FRAXA, FRDA, FTDALS1, HD, SBMA, SCA1 and SCA3 loci were provided with ExpansionHunter. The JSON files for the remaining loci were obtained by personal communication with Egor Dolzhenko (Illumina, Inc. San Diego, CA, USA). For data aligned with bowtie2, the --min-anchor-mapq parameter was set to 44, while for the original alignments of the Coriell samples this parameter was set to 60. The --read-depth parameter was set the median coverage for each sample in the WES_PCR cohort, otherwise this was computed by ExpansionHunter for the WGS samples. The list of STR loci provided with STRetch does not include FRDA, which was added manually. The EPM1 repeat motif is 12 bp and is not assessed using STRetch, which aligns to an augmented reference genome containing a decoy chromosome for each STR repeat motif up to 6 bp in size.

ExpansionHunter and TREDPARSE-T call allele lengths and genotypes. To call individuals as having expansions requires the user to define thresholds on allele sizes as to what constitutes an appropriate threshold. For FRAXA, we additionally tested using the premutation threshold (labelled FRAXA_pre), in addition to testing for full expansions. To call an expansion, we used the same thresholds as Dolzhenko et al[4] (based on McMurray[7]) or the largest reported normal allele size at other loci. Other thresholds will change the sensitivity and specificity. TREDPARSE-L expansions calls were recorded for all samples labelled as "risk". exSTRa p-values were Bonferroni corrected over the number of STRs tested. STRetch reports p-values adjusted for multiple testing over all STRs genome wide, however unadjusted p-values were extracted and Bonferroni corrected over just the number of STRs tested. A threshold of $p < 0.05$ was used for significance.

References

1. 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.
2. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303.
3. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-580.
4. 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., Johnson, N.H., et al. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res 27, 1895-1903.
5. 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.
6. 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.
7. McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11, 786-799.