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

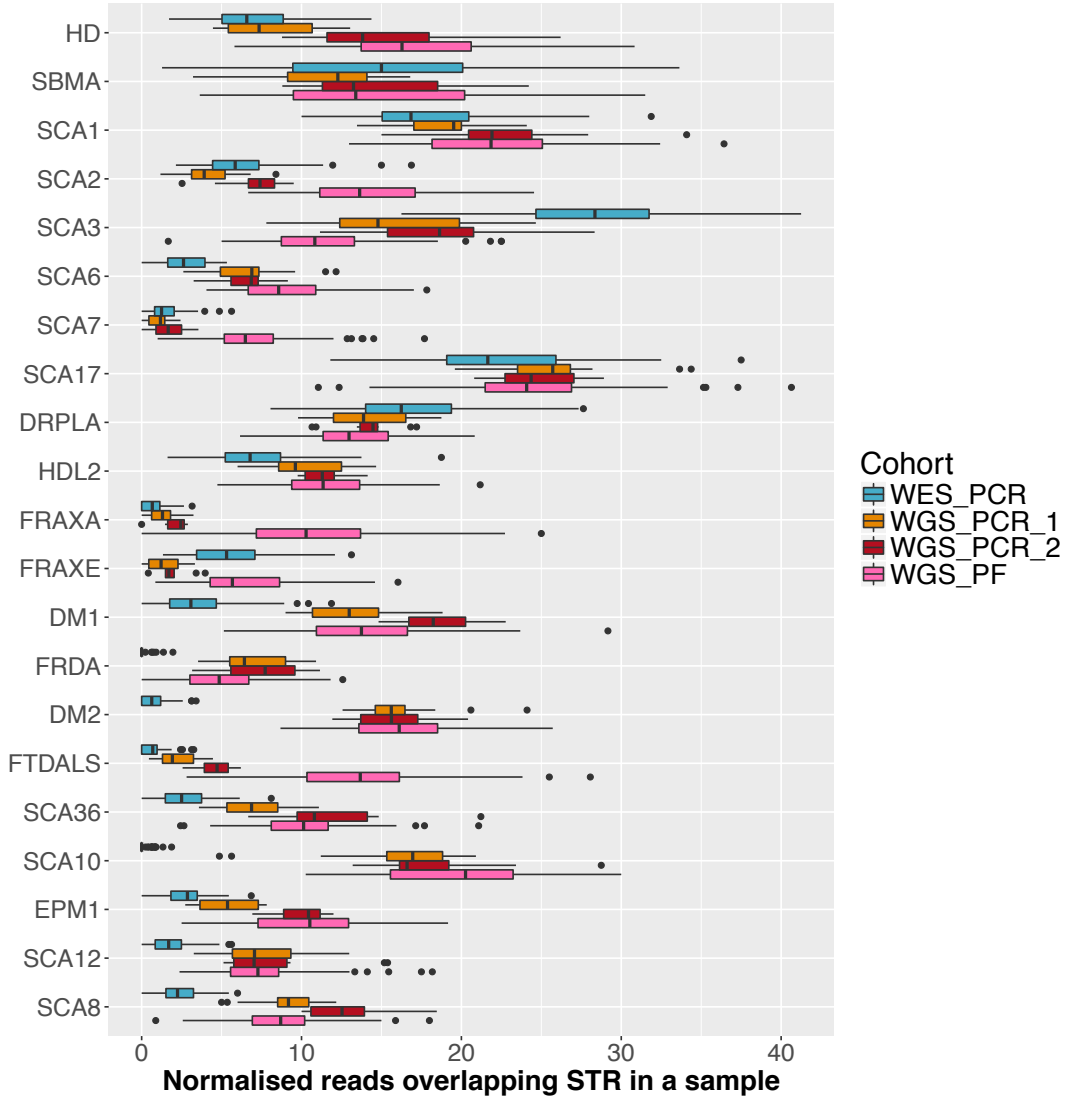


Figure S1 Normalised sequencing coverage comparison between the four sequencing cohorts, split by repeat expansion type.

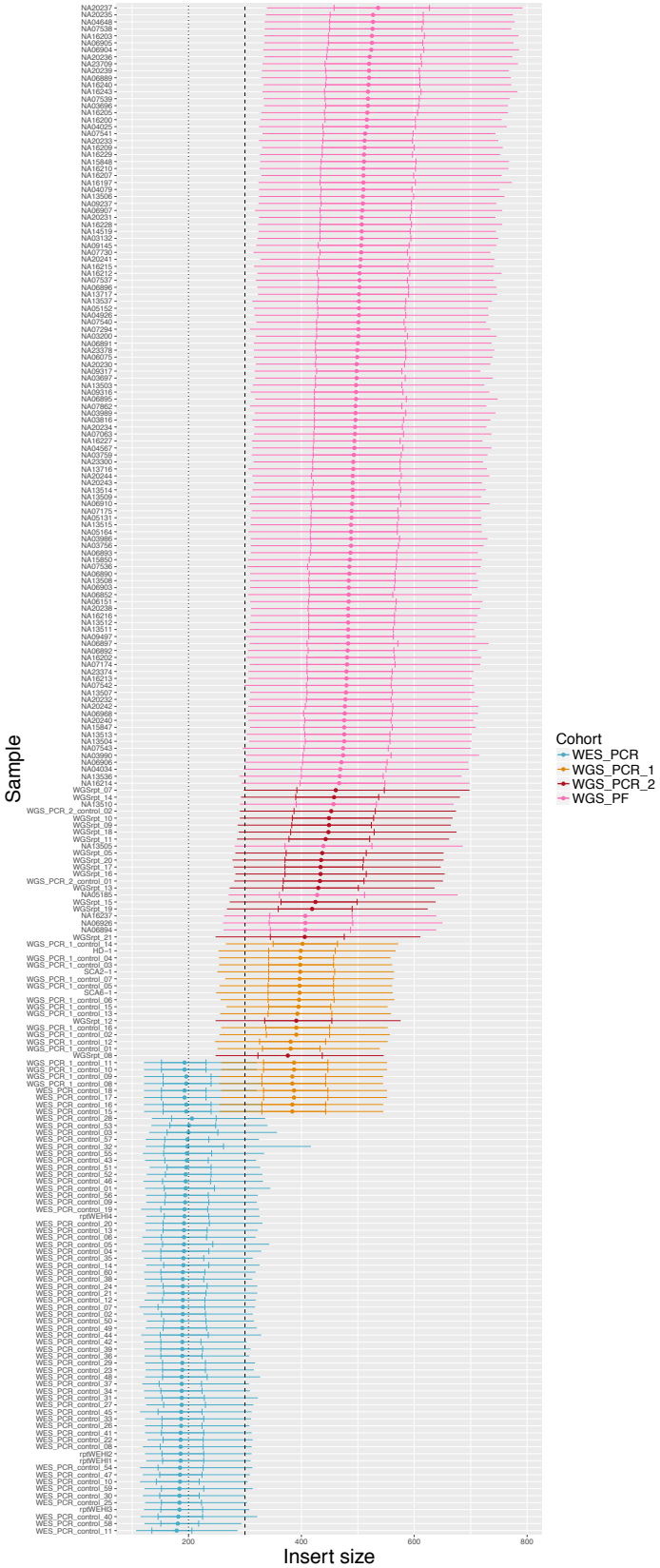
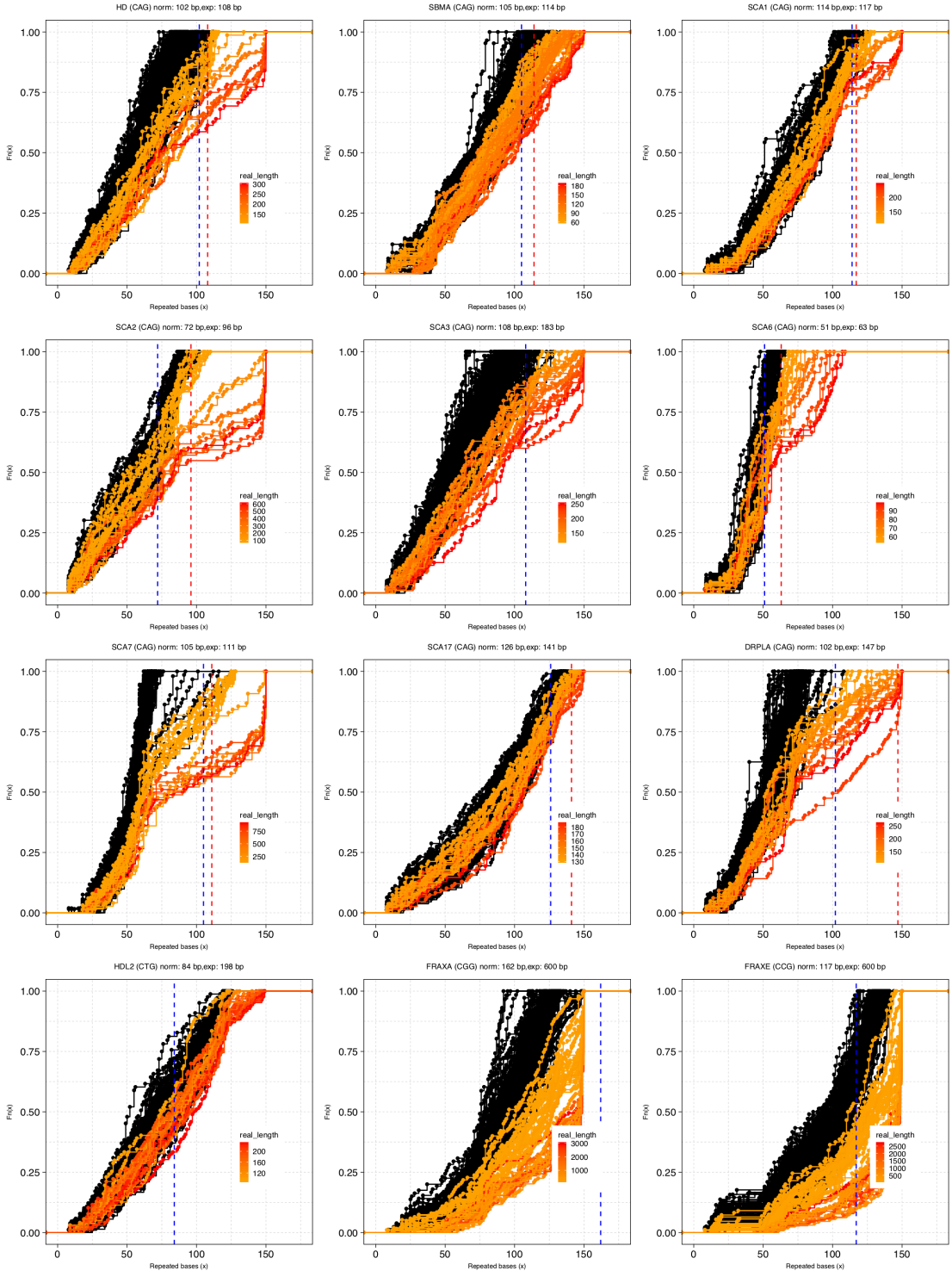


Figure S2 Insert sizes by sample. Samples are in decreasing order of the median insert size that is indicated by a circle. Bars extend to cover 90% of insert sizes, at the 5th and 95th percentile. The interquartile range (IQR), covering 50% of the data, is indicated by small vertical bars. The dotted 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
real_length
300
250
200
150
SBMA (CAG) norm: 105 bp,exp: 114 bp
real_length
180
150
120
90
60
SCA1 (CAG) norm: 114 bp,exp: 117 bp
real_length
200
150
SCA2 (CAG) norm: 72 bp,exp: 96 bp
real_length
600
500
400
300
200
100
SCA3 (CAG) norm: 108 bp,exp: 183 bp
real_length
250
200
150
SCA6 (CAG) norm: 51 bp,exp: 63 bp
real_length
90
80
70
60
SCA7 (CAG) norm: 105 bp,exp: 111 bp
real_length
750
500
250
SCA17 (CAG) norm: 126 bp,exp: 141 bp
real_length
180
170
160
150
140
130
DRPLA (CAG) norm: 102 bp,exp: 147 bp
real_length
250
200
150
HDL2 (CTG) norm: 84 bp,exp: 198 bp
real_length
200
160
120
FRAXA (CGG) norm: 162 bp,exp: 600 bp
real_length
3000
2000
1000
FRAXE (CCG) norm: 117 bp,exp: 600 bp
real_length
2500
2000
1500
1000
500
Fn(x)
Repeated bases (x)
0.00
0.25
0.50
0.75
1.00
0
50
100
150

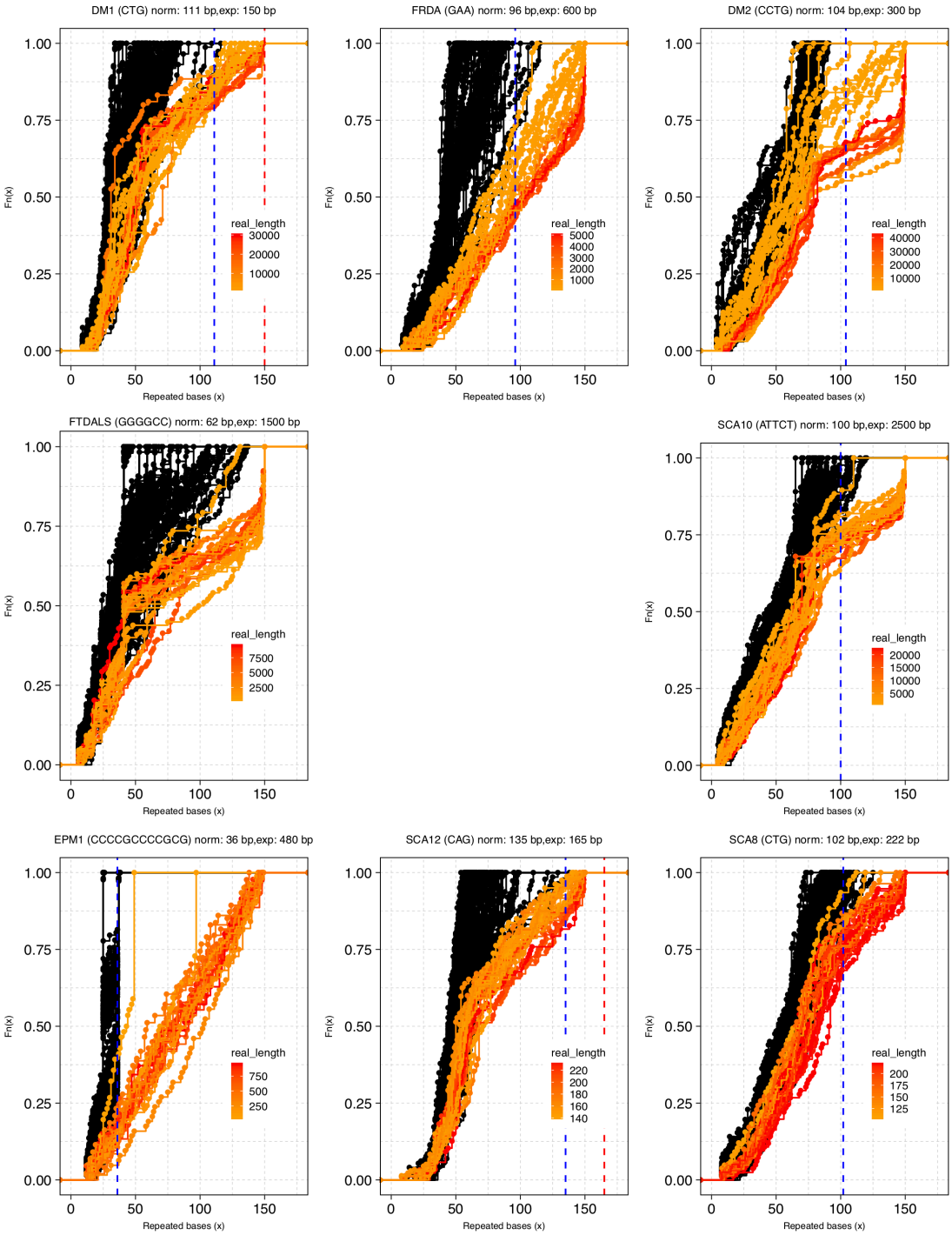
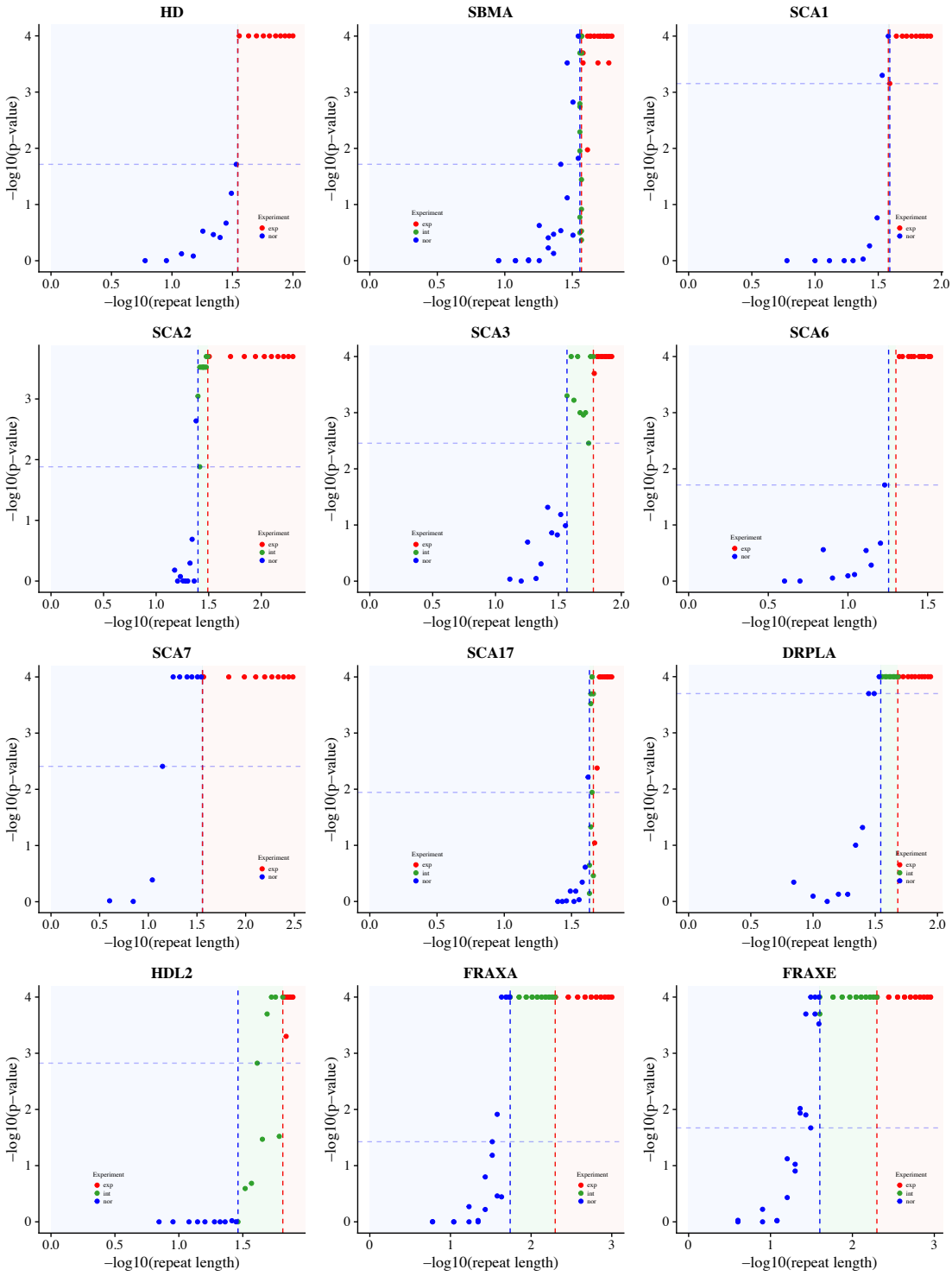


Figure S3 exSTRa eCDF plots for simulated data labeled by size of repeat expansion. Each panel depicts one STR with 210 controls (black) and 20 intermediate size tandem repeat alleles and 20 expanded repeat alleles, with intermediates and expansions coloured in red, with smallest repeat alleles in yellow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

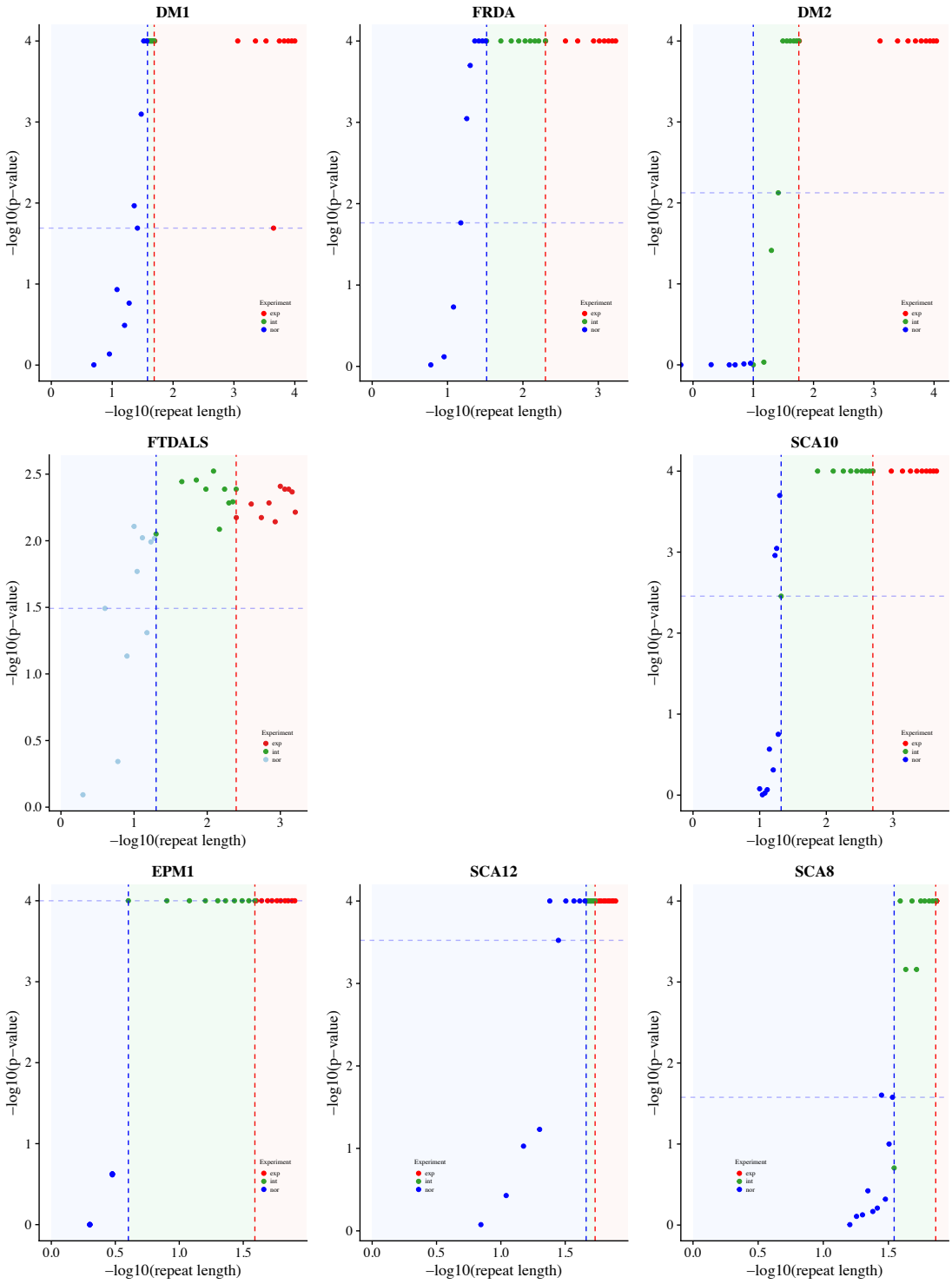
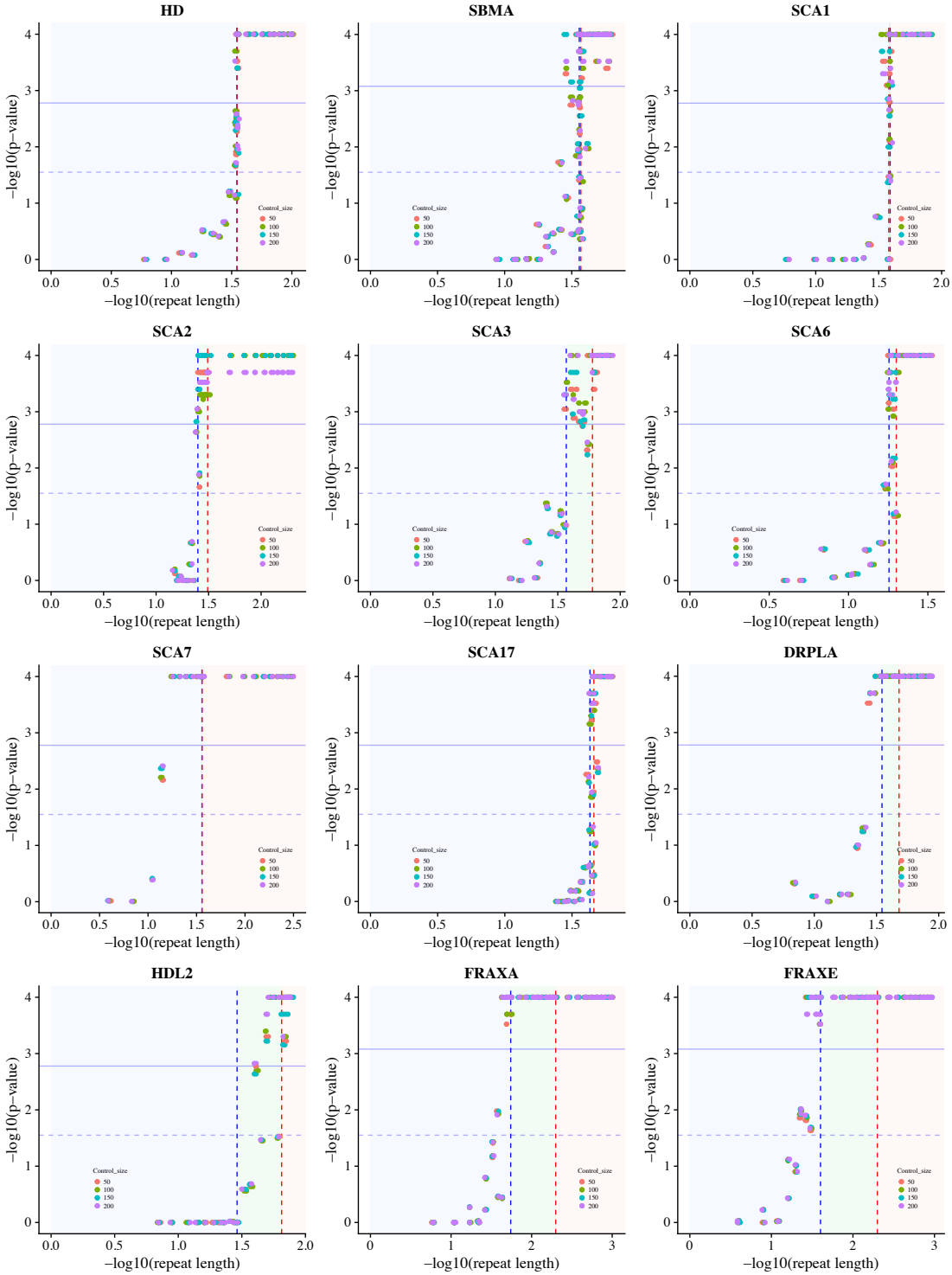


Figure S4 exSTRa p-value behavior with varying repeat length. Regions in light blue are normal ranges, regions in green the intermediate range, which usually means not pathogenic, or leads to a different phenotype, possibly with lower penetrance. The region in red is the pathogenic range. True expansions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

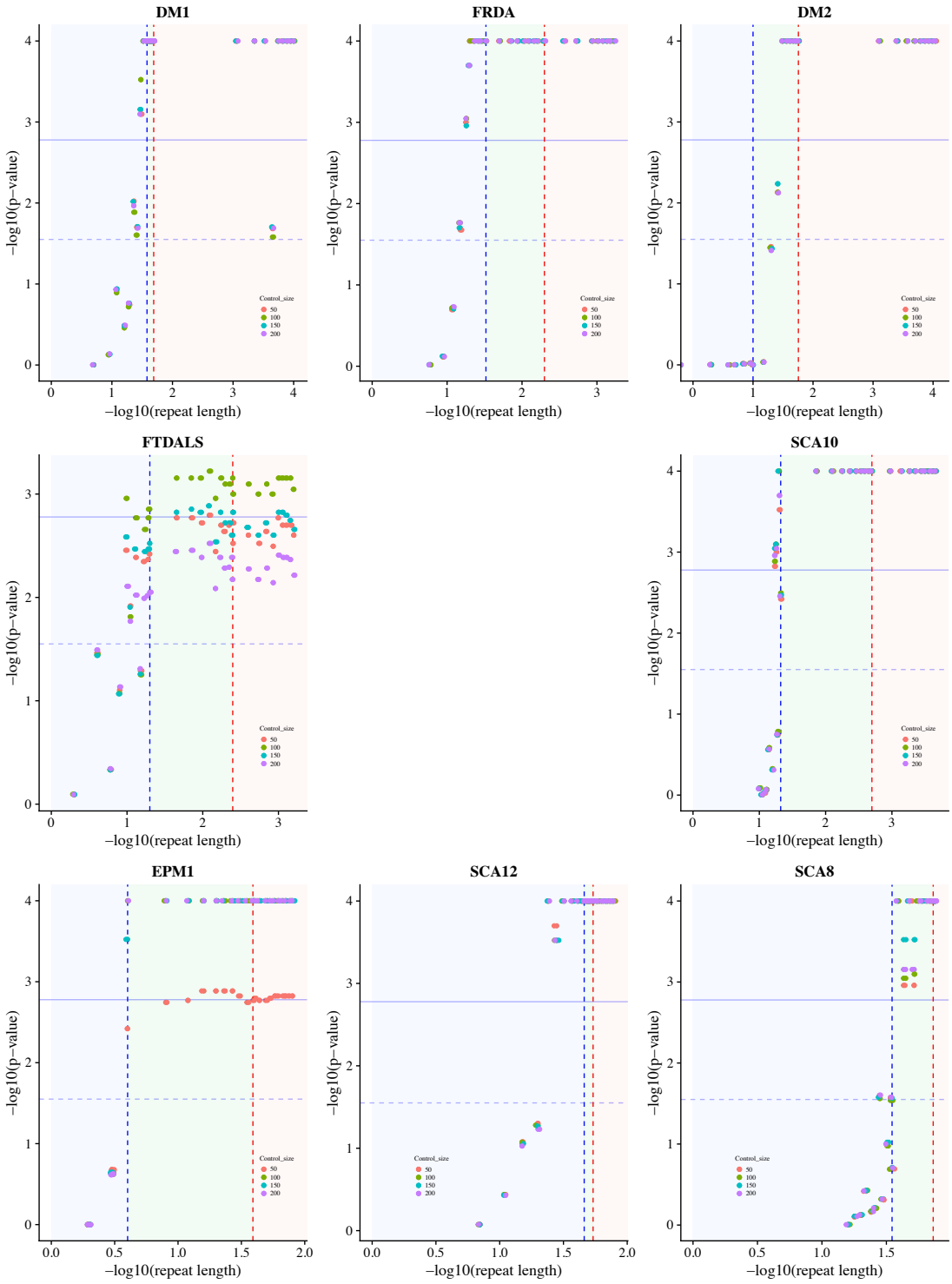


Figure S5 exSTRa p-value behavior with varying control cohort size.

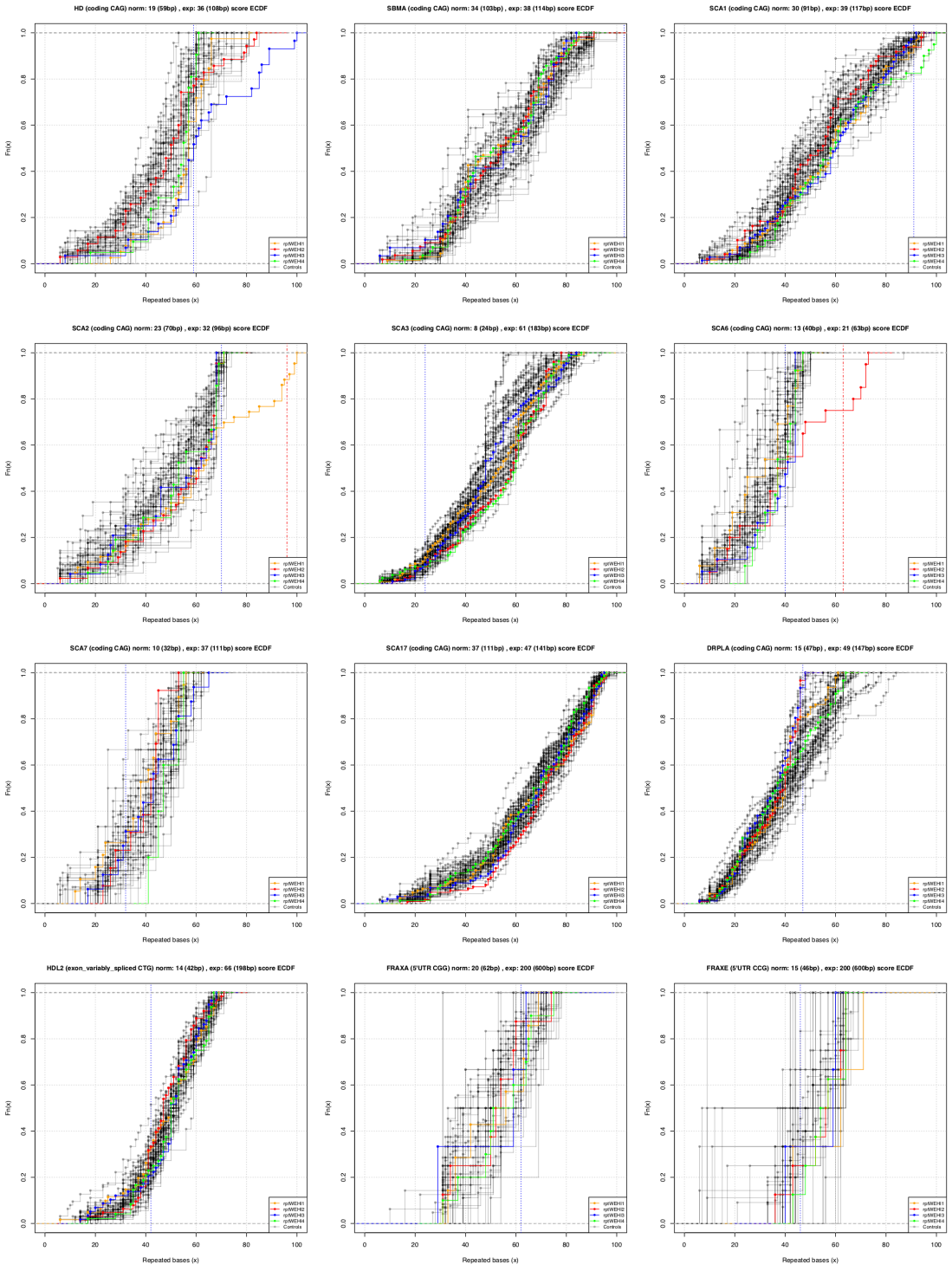
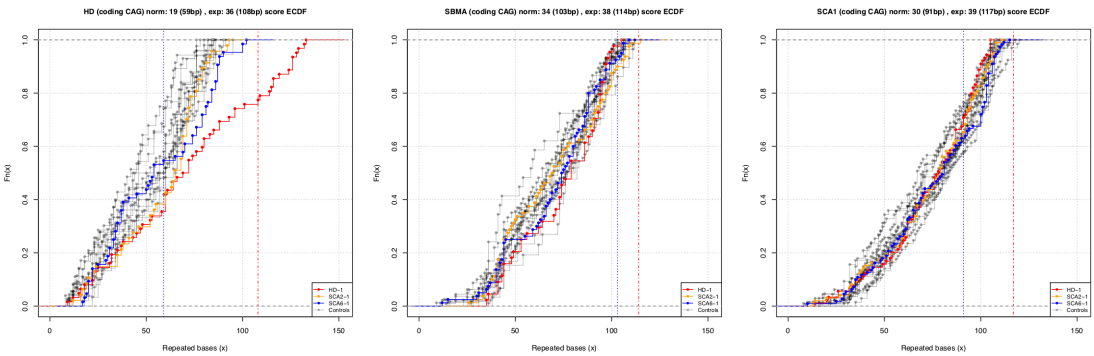
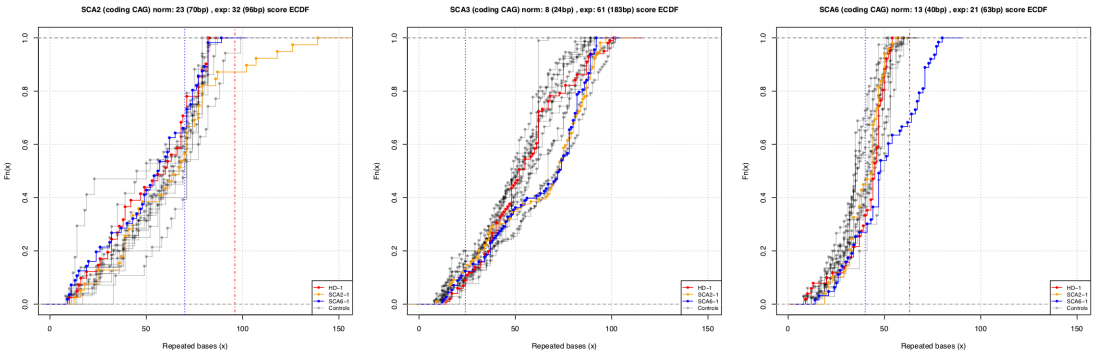
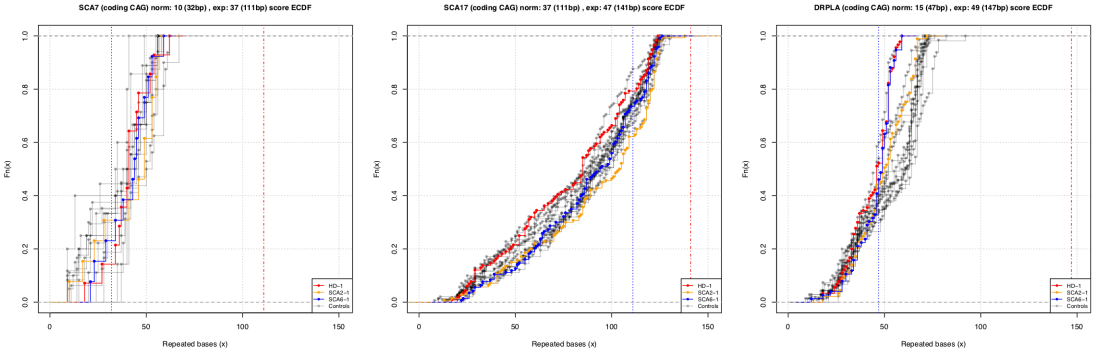
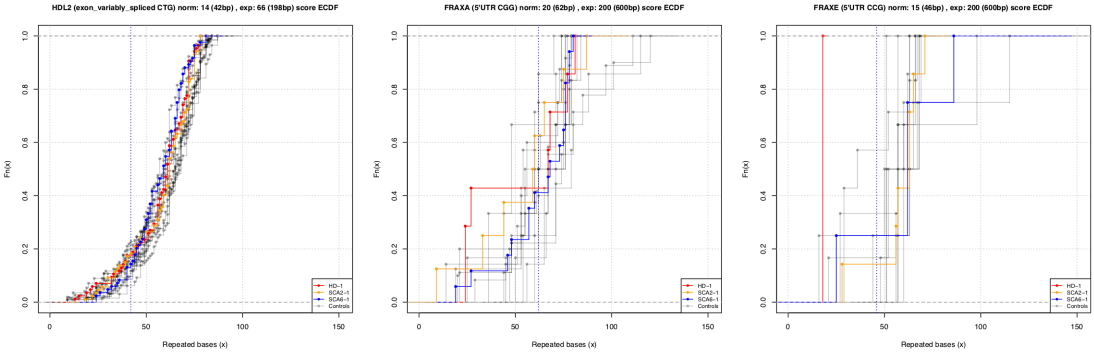


Figure S6. ECDFs for the 13 STR loci with coverage for the WES cohort (WES_PCR).


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
Fn(x)
Repeated bases (x)
HD−1
SCA2−1
SCA6−1
Controls


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
Fn(x)
Repeated bases (x)
HD−1
SCA2−1
SCA3−1
SCA6−1
Controls


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
Fn(x)
Repeated bases (x)
HD−1
SCA2−1
SCA6−1
Controls


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
Fn(x)
Repeated bases (x)
HD−1
SCA2−1
SCA6−1
Controls

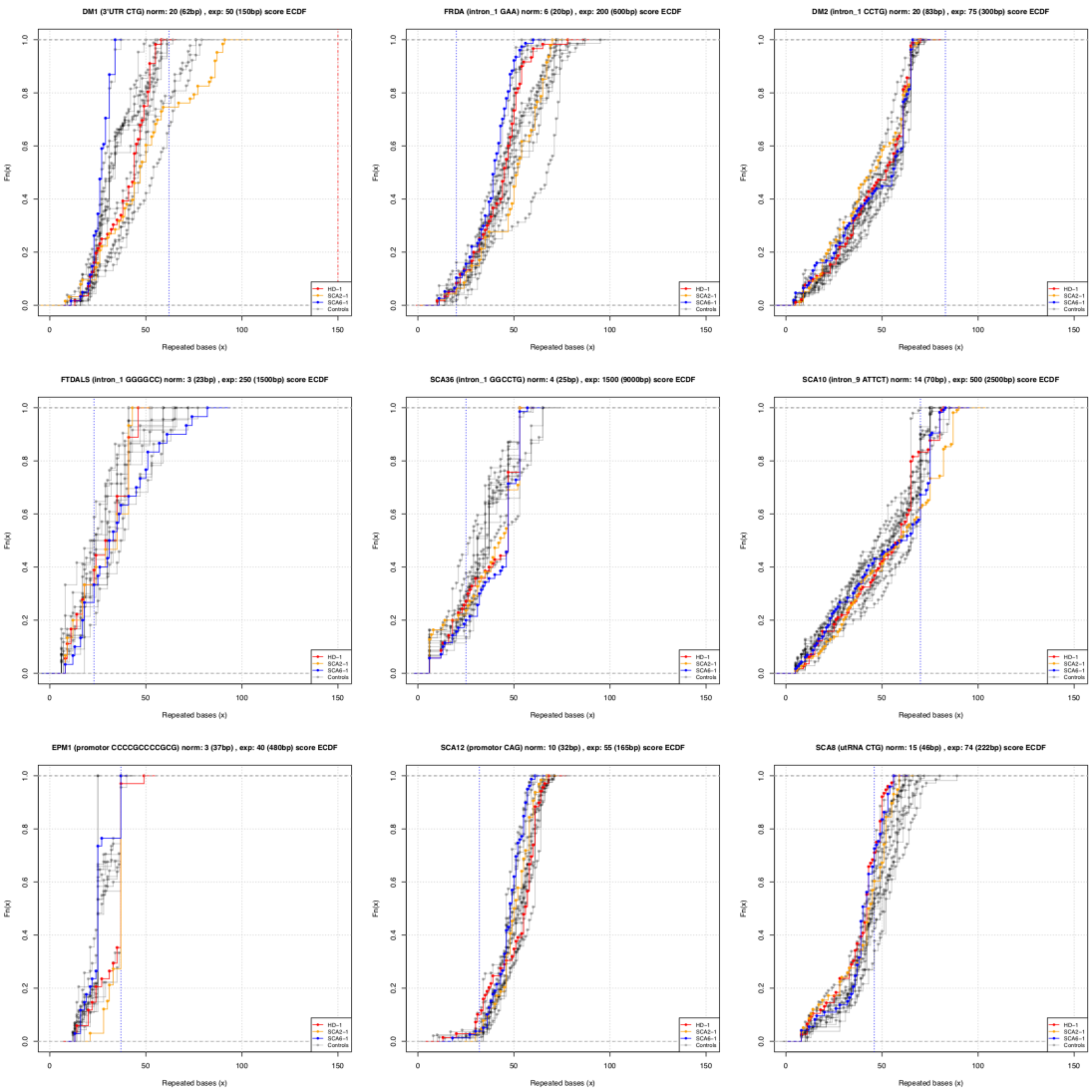
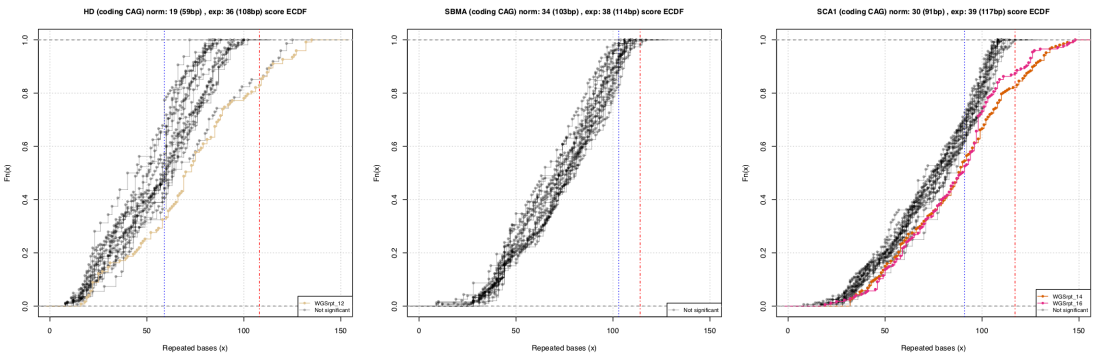
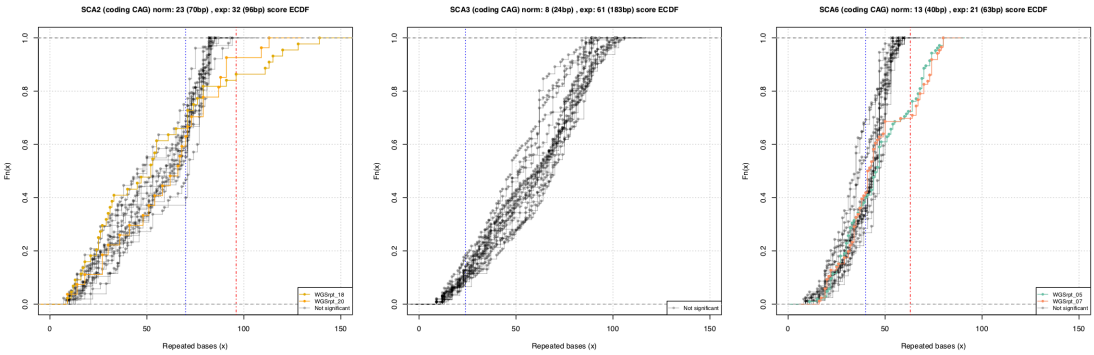
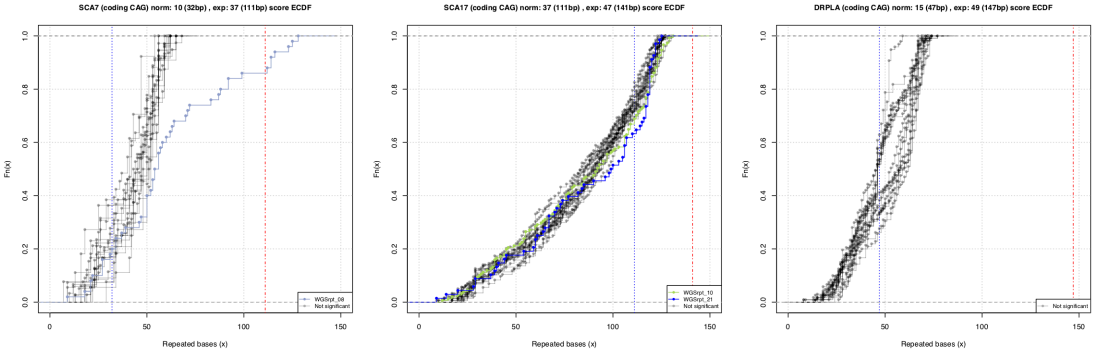
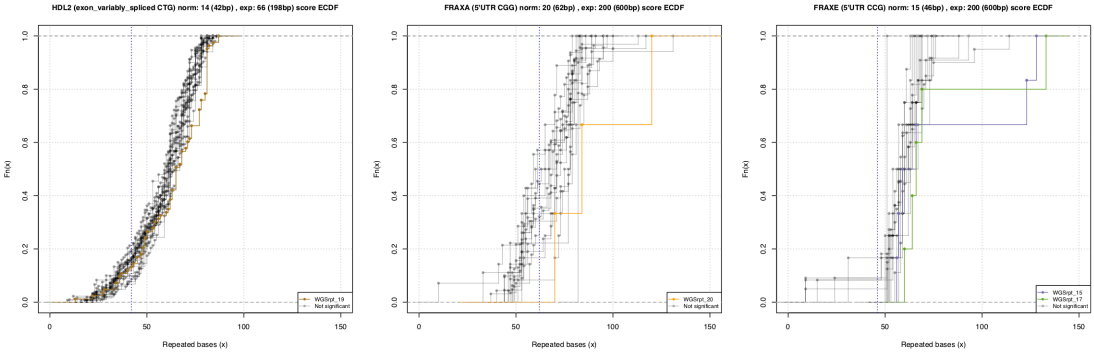


Figure S7. ECDFs for all 21 STR loci for the WGS with PCR cohort (WGS_PCR_1).


HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF
SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_12
Not significant
Not significant
WGSrpt_14
WGSrpt_16
Not significant


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
Fn(x)
Repeated bases (x)
WGSrpt_18
WGSrpt_20
Not significant
Not significant
WGSrpt_05
WGSrpt_07
Not significant


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
Fn(x)
Repeated bases (x)
WGSrpt_08
Not significant
WGSrpt_10
WGSrpt_21
Not significant
Not significant


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
Fn(x)
Repeated bases (x)
WGSrpt_19
Not significant
WGSrpt_20
Not significant
WGSrpt_15
WGSrpt_17
Not significant

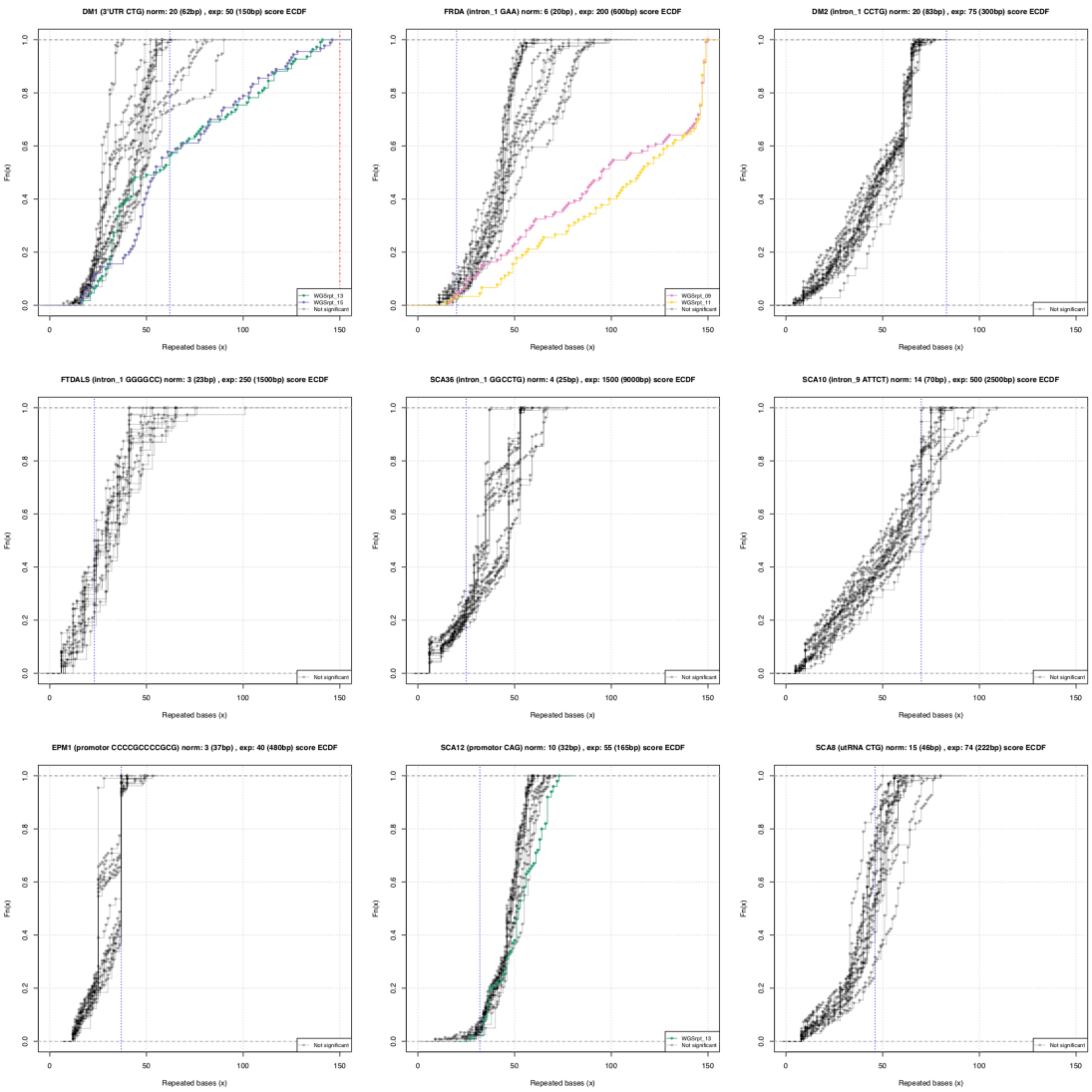
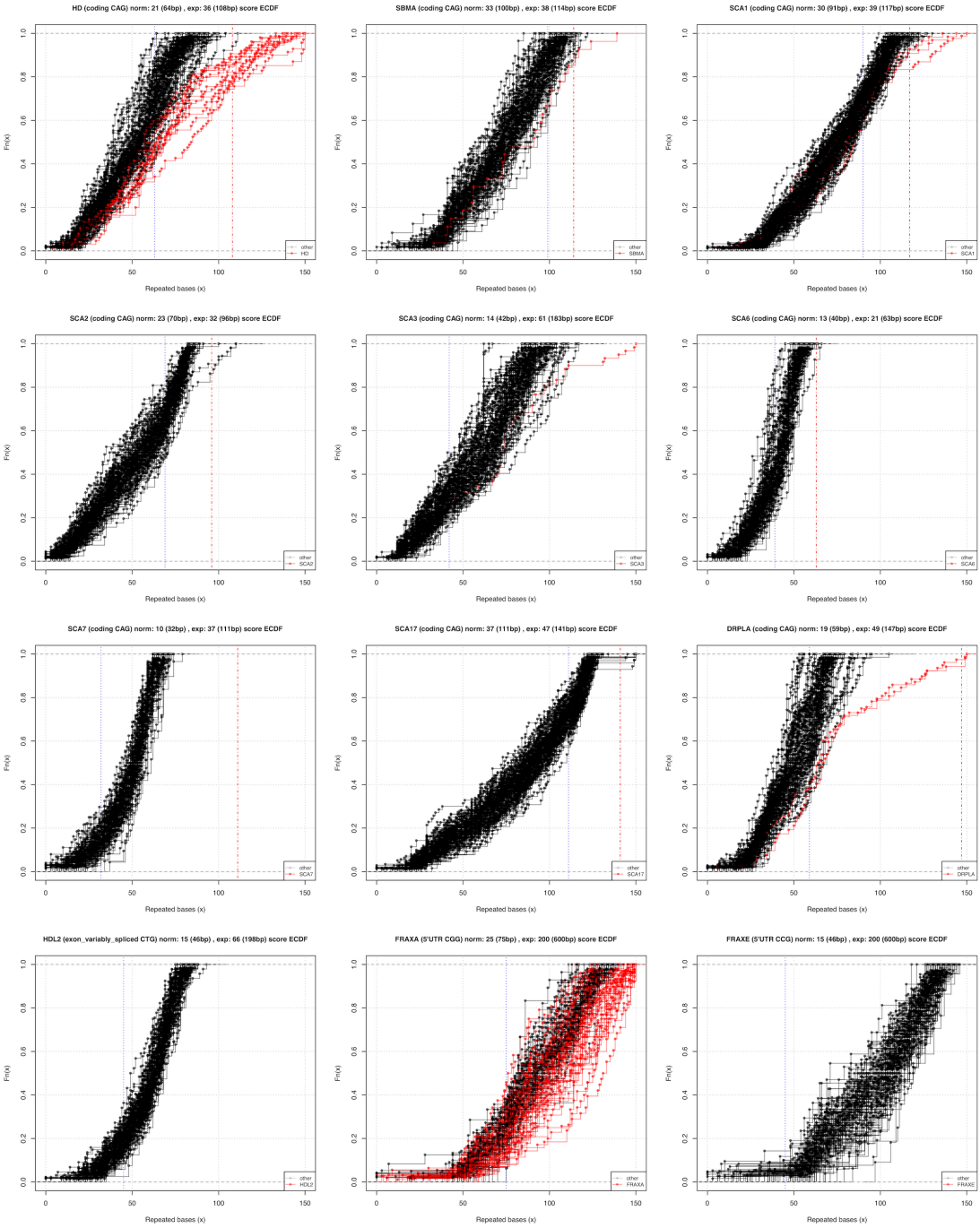


Figure S8. ECDFs for all 21 STR loci for the WGS with PCR cohort (WGS_PCR_2).


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
Fn(x)
Repeated bases (x)
other
HD
SBMA
SCA1
SCA2
SCA3
SCA6
SCA7
SCA17
DRPLA
HDL2
FRAXA
FRAXE

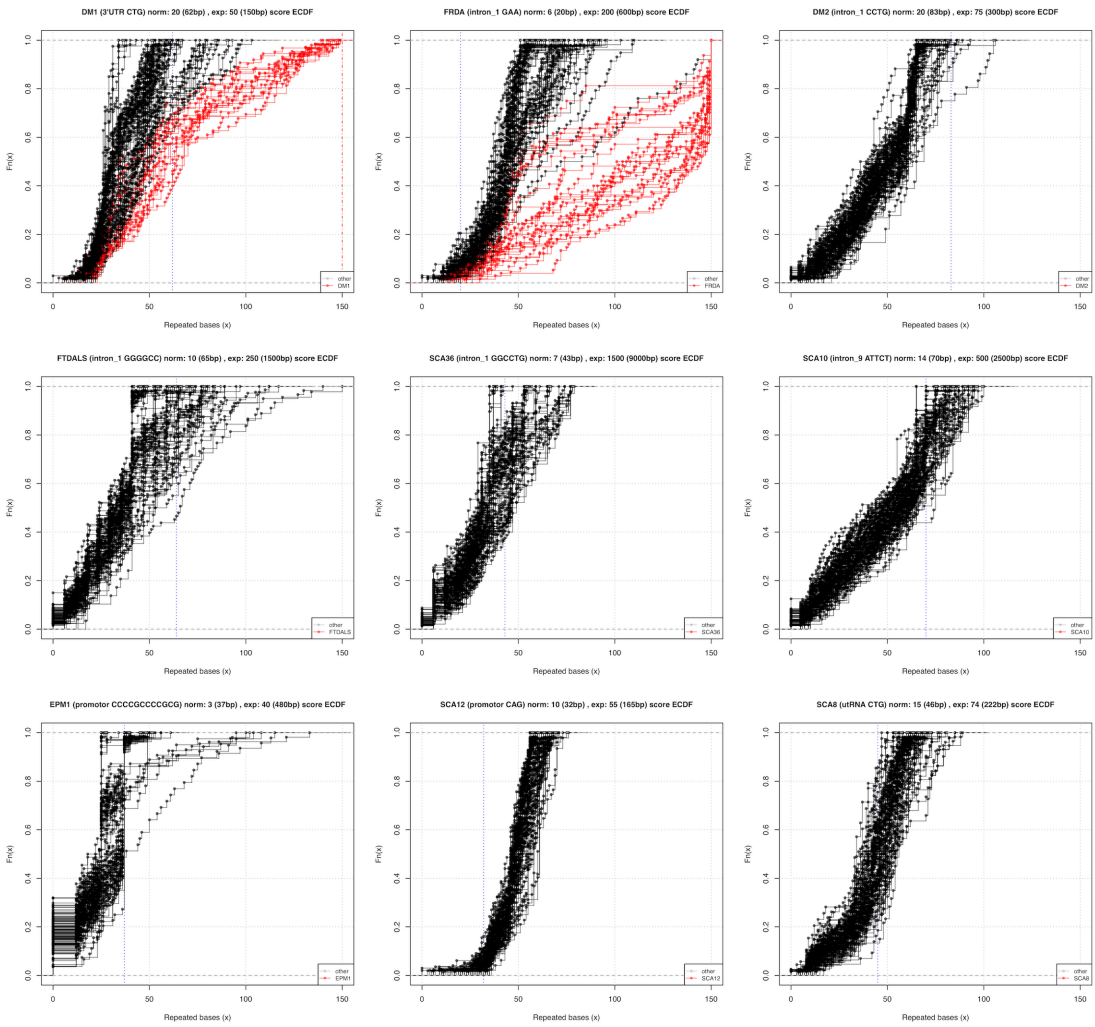
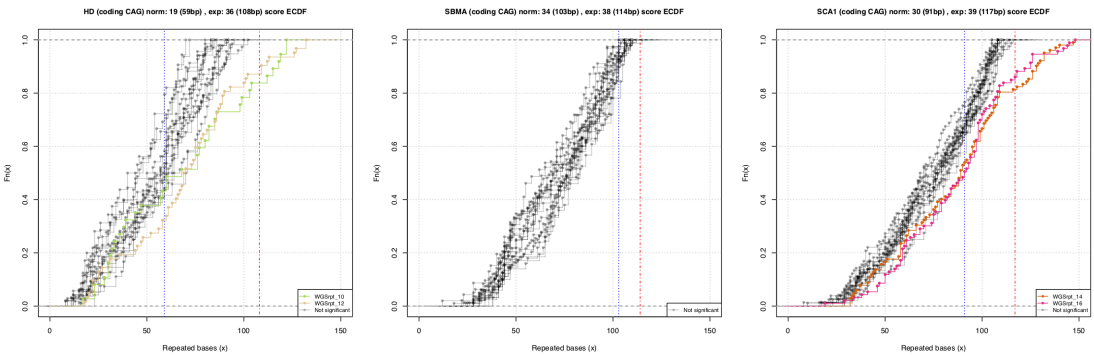
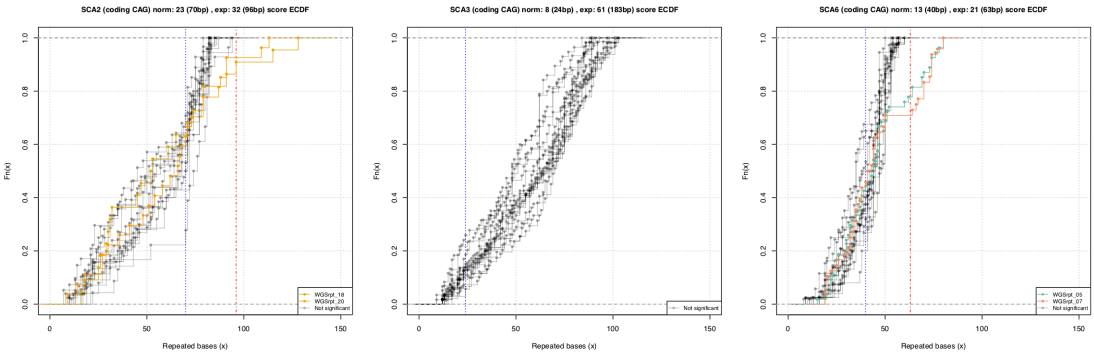
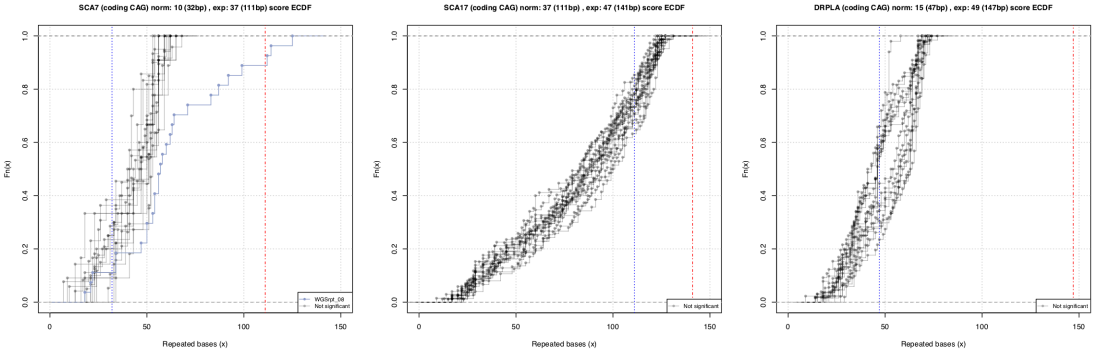
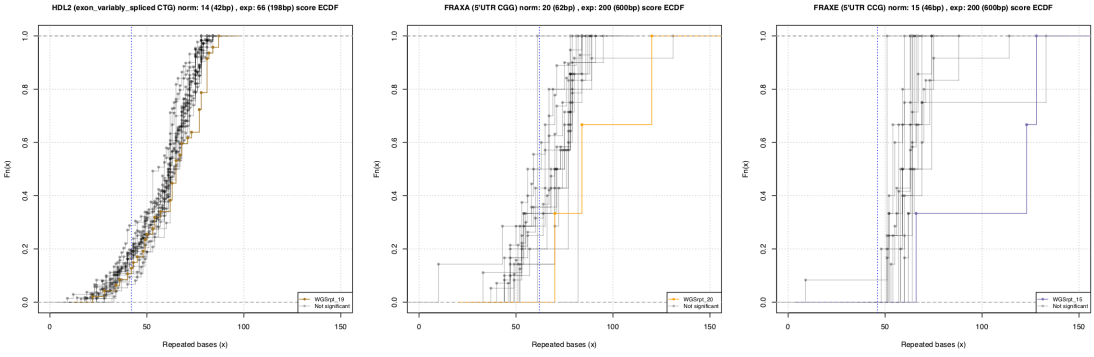


Figure S9. ECDFs for all 21 STR loci for the WGS without PCR cohort (WGS_PF).


HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF
WGSrpt_10
WGSrpt_12
Not significant
SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF
Not significant
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF
WGSrpt_14
WGSrpt_16
Not significant
Fn(x)
Repeated bases (x)


SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF
WGSrpt_18
WGSrpt_20
Not significant
SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF
Not significant
SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF
WGSrpt_05
WGSrpt_07
Not significant
Fn(x)
Repeated bases (x)


SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF
WGSrpt_08
Not significant
SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF
Not significant
DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF
Not significant
Fn(x)
Repeated bases (x)


HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF
WGSrpt_19
Not significant
FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF
WGSrpt_20
Not significant
FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF
WGSrpt_15
Not significant
Fn(x)
Repeated bases (x)

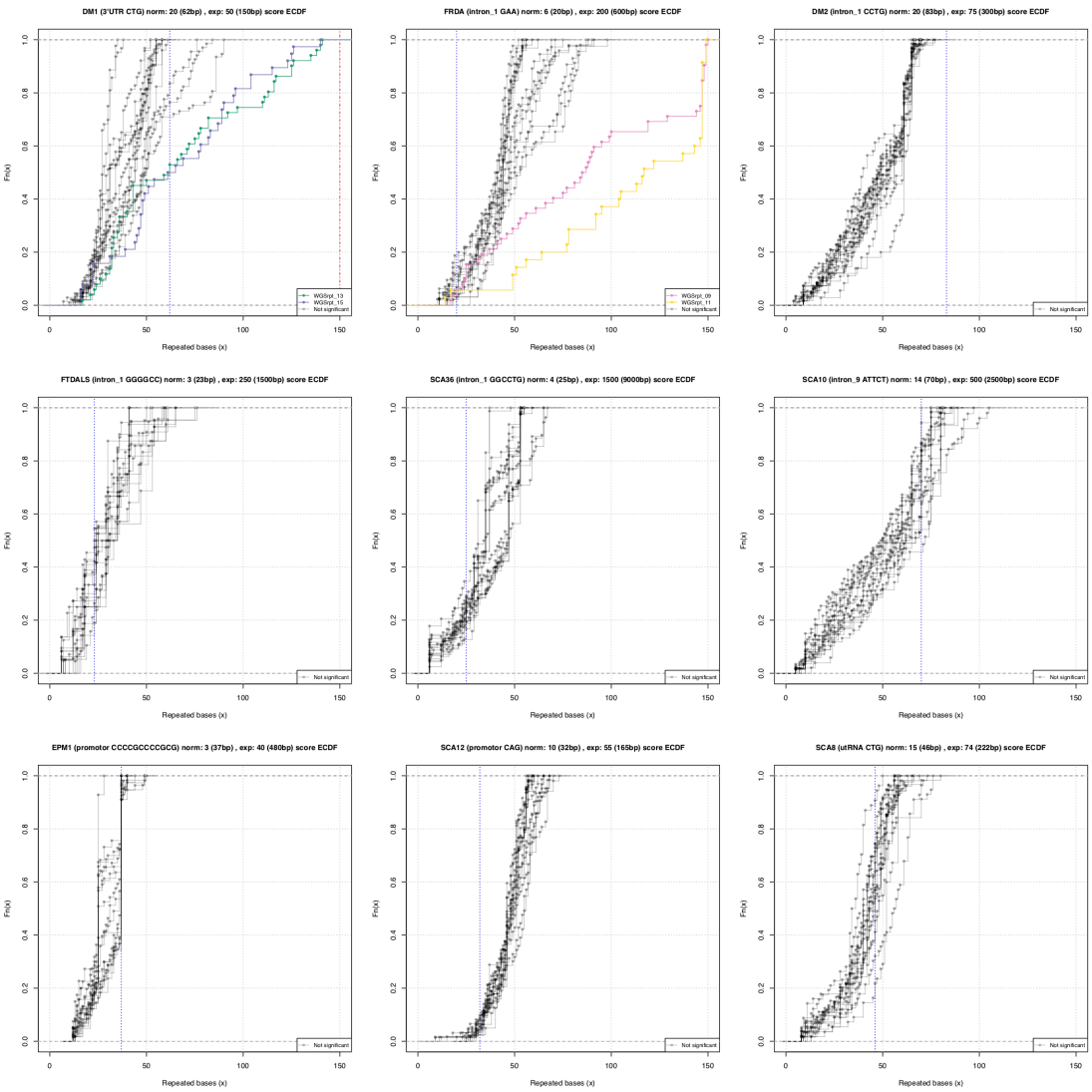
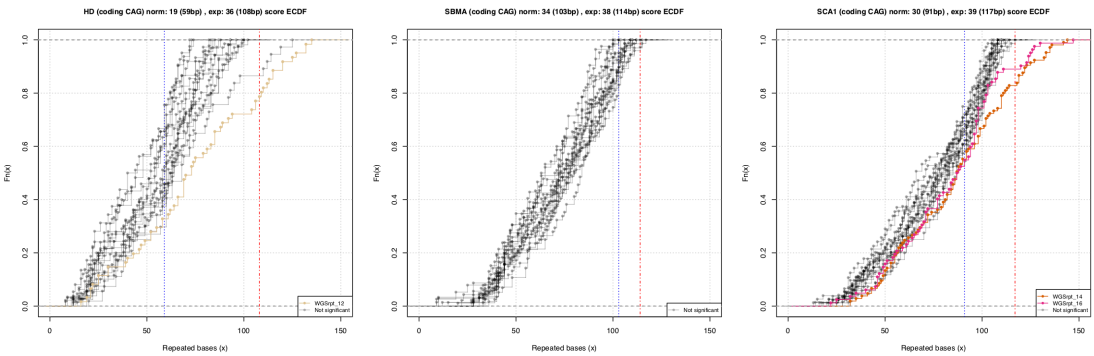
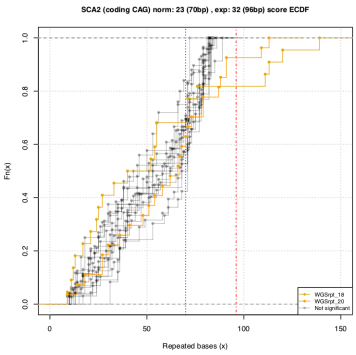


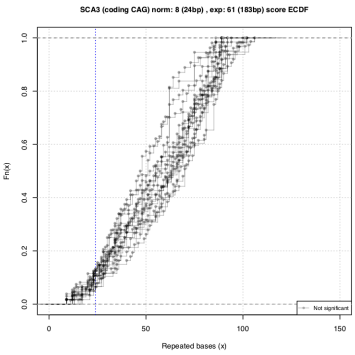
Figure S10. ECDFs for all 21 STR loci for the WGS with PCR 30X sub-cohort (WGS_PCR_2_30X_1).

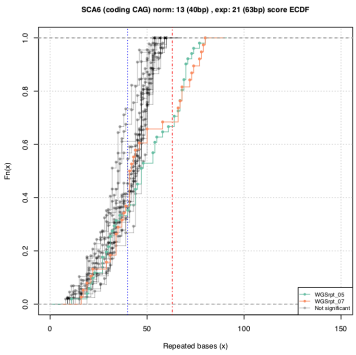

HD (coding CAG) norm: 19 (59bp) , exp: 36 (108bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_12
Not significant

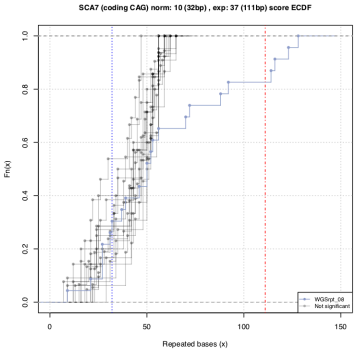
SBMA (coding CAG) norm: 34 (103bp) , exp: 38 (114bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant

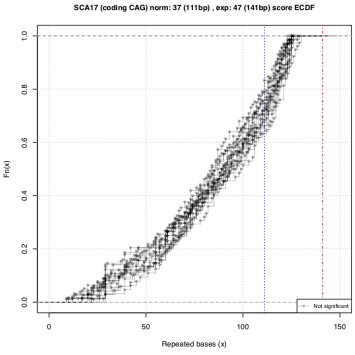
SCA1 (coding CAG) norm: 30 (91bp) , exp: 39 (117bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_14
WGSrpt_16
Not significant

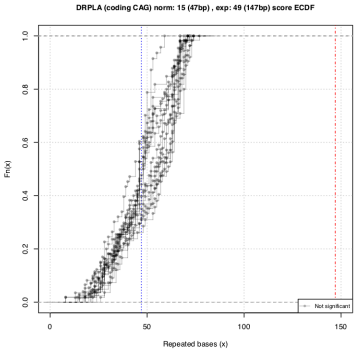

SCA2 (coding CAG) norm: 23 (70bp) , exp: 32 (96bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_18
WGSrpt_20
Not significant

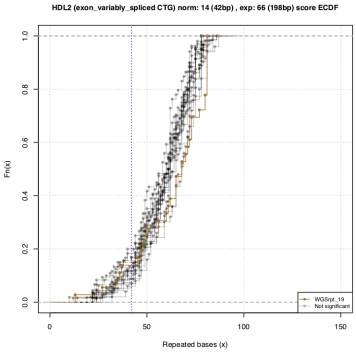

SCA3 (coding CAG) norm: 8 (24bp) , exp: 61 (183bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant

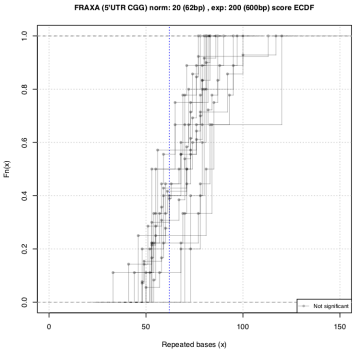

SCA6 (coding CAG) norm: 13 (40bp) , exp: 21 (63bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_05
WGSrpt_07
Not significant

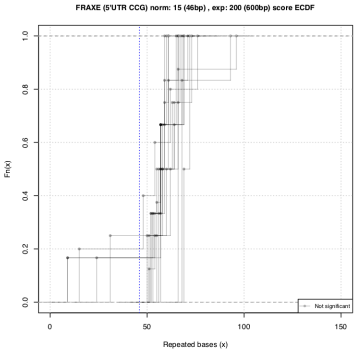

SCA7 (coding CAG) norm: 10 (32bp) , exp: 37 (111bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_08
Not significant


SCA17 (coding CAG) norm: 37 (111bp) , exp: 47 (141bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant


DRPLA (coding CAG) norm: 15 (47bp) , exp: 49 (147bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant


HDL2 (exon_variably_spliced CTG) norm: 14 (42bp) , exp: 66 (198bp) score ECDF
Fn(x)
Repeated bases (x)
WGSrpt_19
Not significant


FRAXA (5'UTR CGG) norm: 20 (62bp) , exp: 200 (600bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant


FRAXE (5'UTR CCG) norm: 15 (46bp) , exp: 200 (600bp) score ECDF
Fn(x)
Repeated bases (x)
Not significant

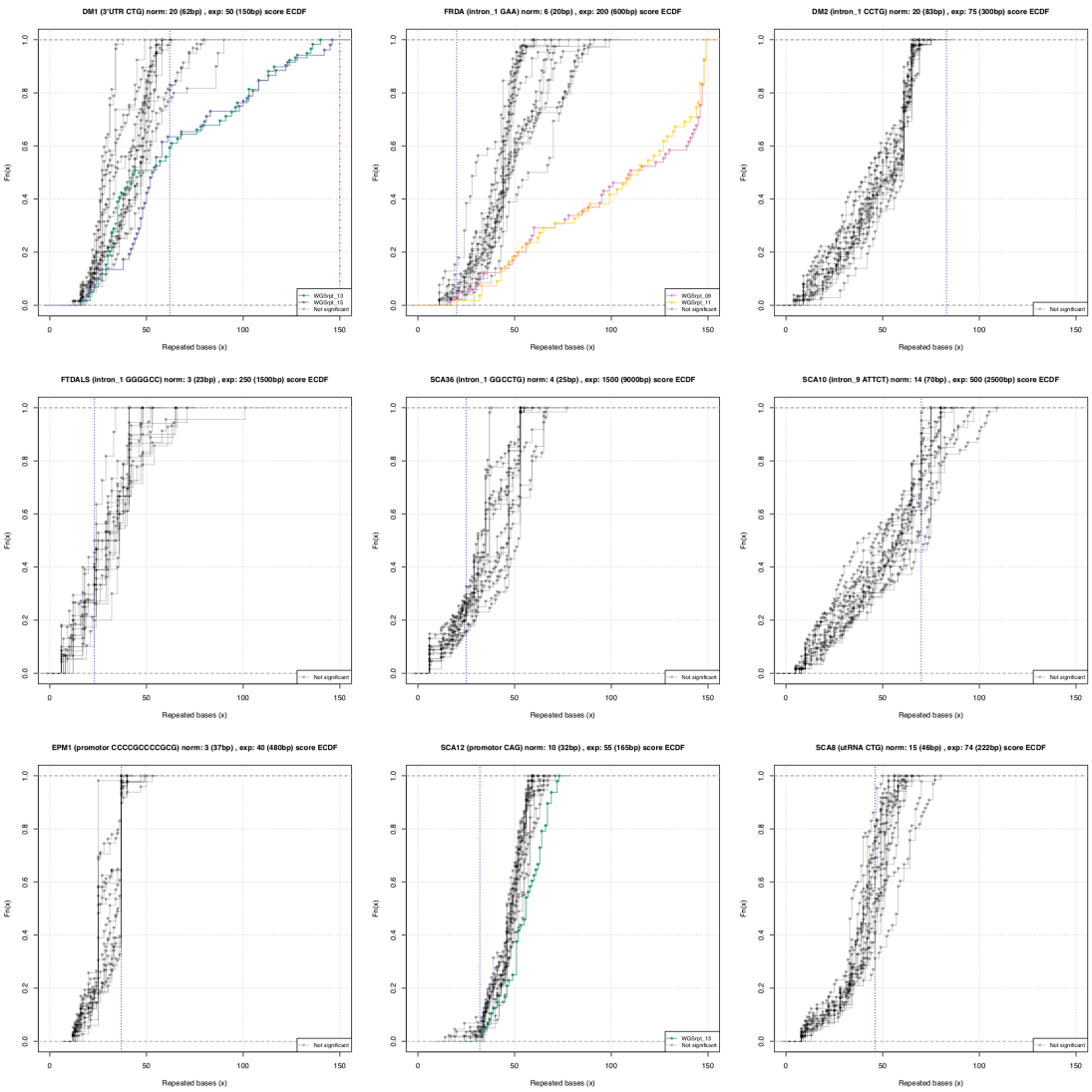


Figure S11. ECDFs for all 21 STR loci for the WGS with PCR 30X sub-cohort (WGS_PCR_2_30X_2).

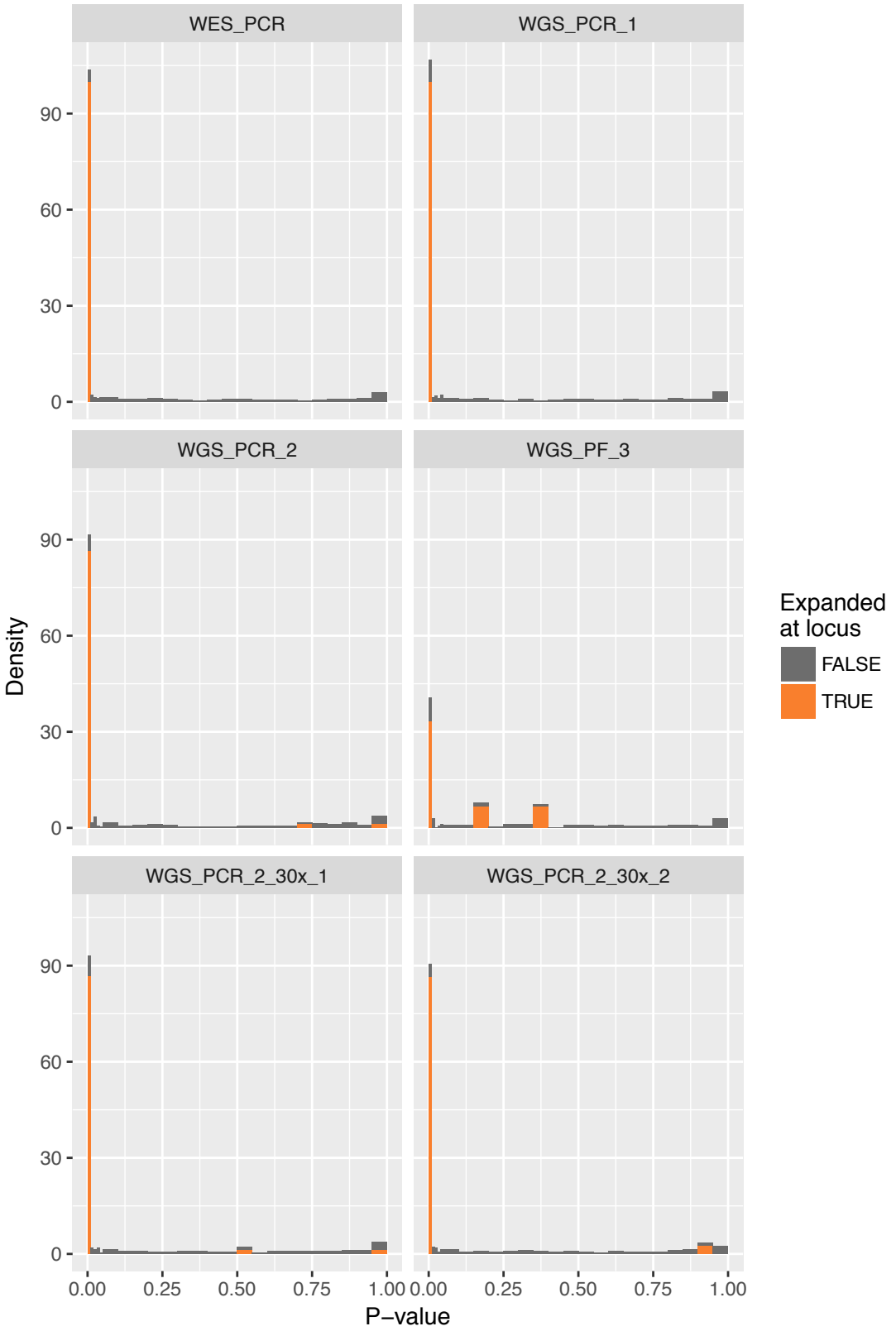


Figure S12. Histograms of the frequency density for the empirically derived p-values for all STR loci for all four cohorts, as well as the two 30X subsets for WGS_PCR_2 (top left panel = WES, top right panel = WGS_PCR_1, middle left = WGS_PCR_2, middle right = WGS_PF_3, bottom left = WGS_PCR_2_30X_1, 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.

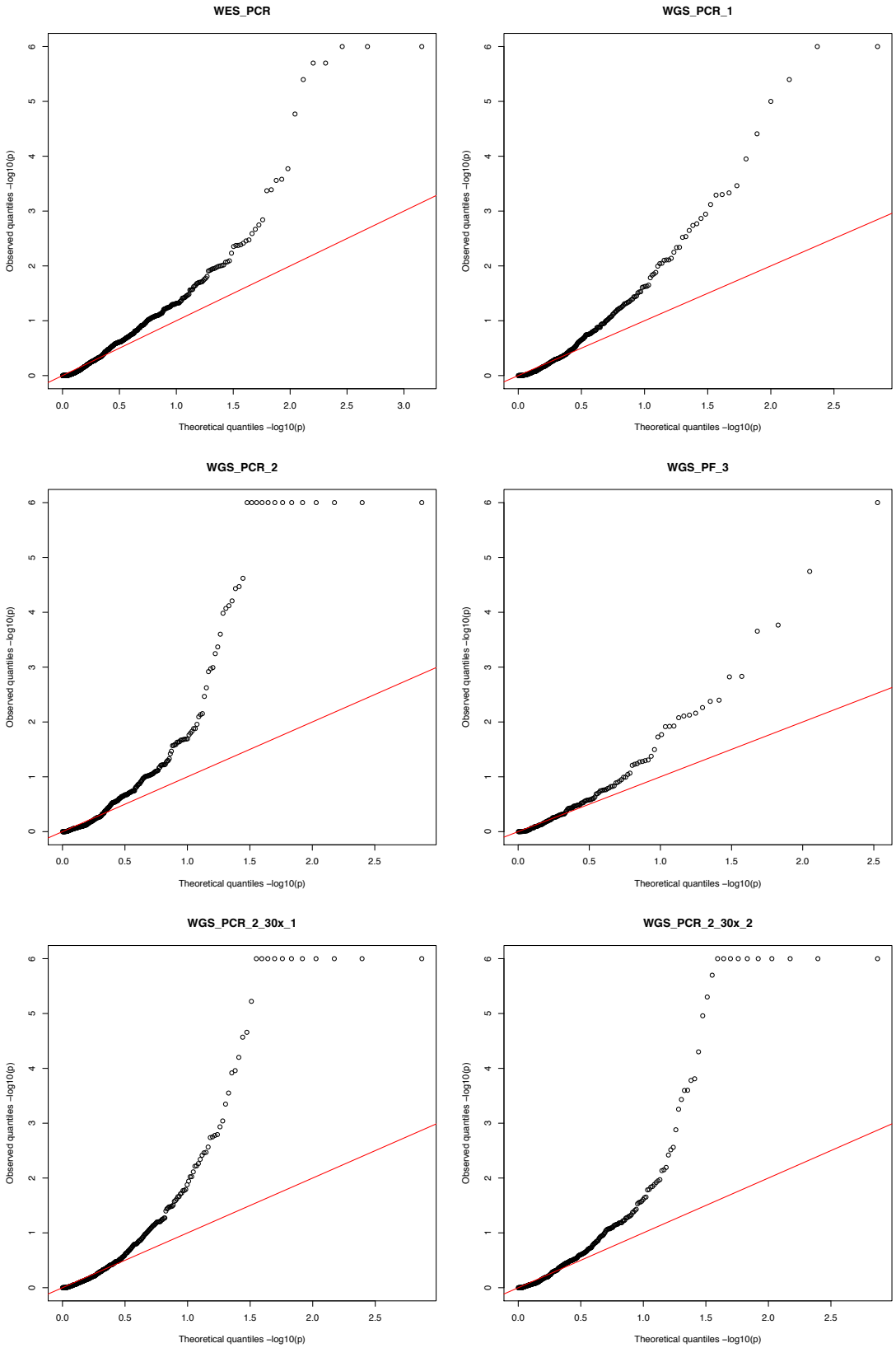


Figure S13. Q-Q plots for the empirically derived p-values for all STR loci for all four cohorts, as well as the two 30X subsets for WGS_PCR_2 (top left panel = WES, top right panel = WGS_PCR_1, middle left = WGS_PCR_2, middle right = WGS_PF_3, bottom left = WGS_PCR_2_30X_1, bottom right = WGS_PCR_30X_2). 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**ctgctgctgctgctgctgctgctg**Tt**gctgc**TTTt**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.

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

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,

$$\hat{m}_j = median\{\mathbf{y}_{.j}\}$$

$$\hat{d}_j = \frac{1}{(\Phi^{-1}(3/4))} MAD\{\mathbf{y}_{.j}\}$$

$$MAD\{\mathbf{y}_{.j}\} = median\{|y_{ij} - median\{y_{.j}\}|\}$$

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_i^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.

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

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.