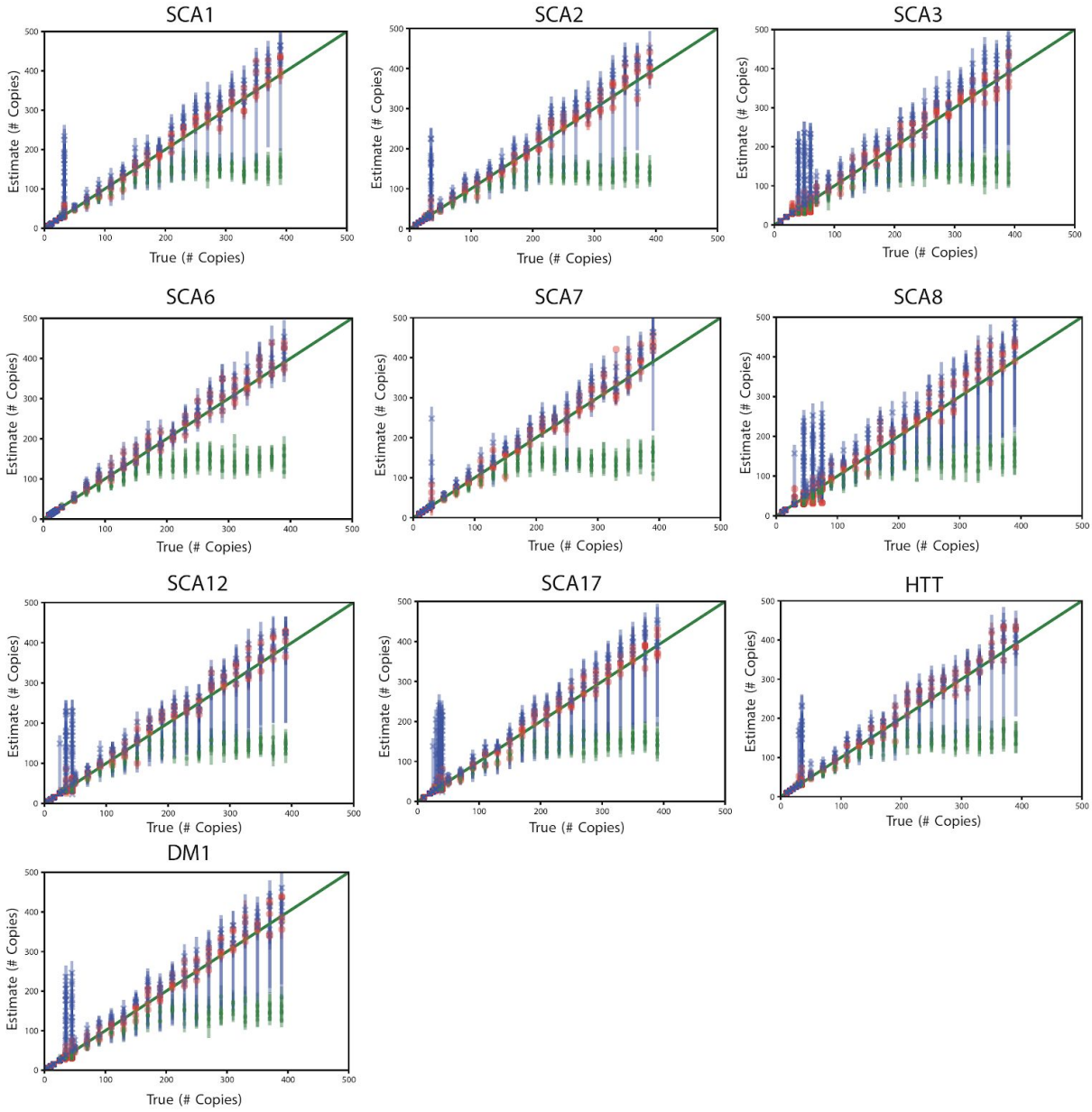
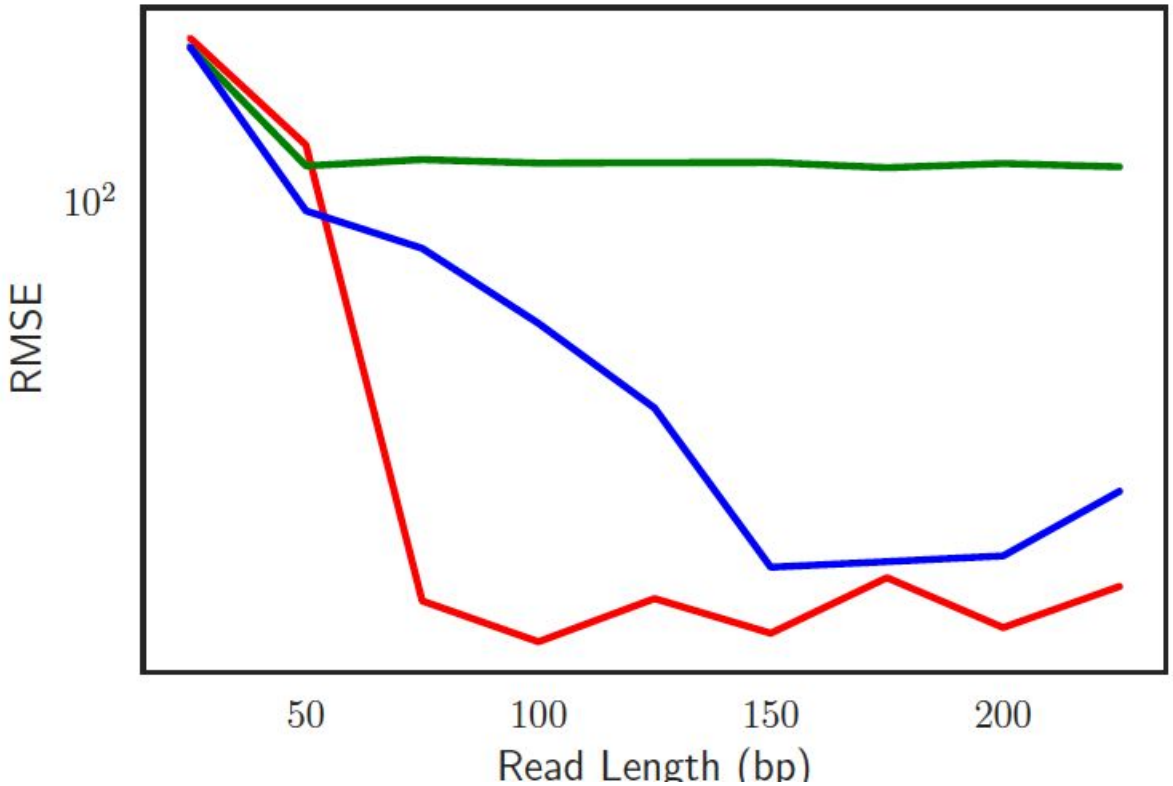
Supplementary Figures

Supplementary Figure 6



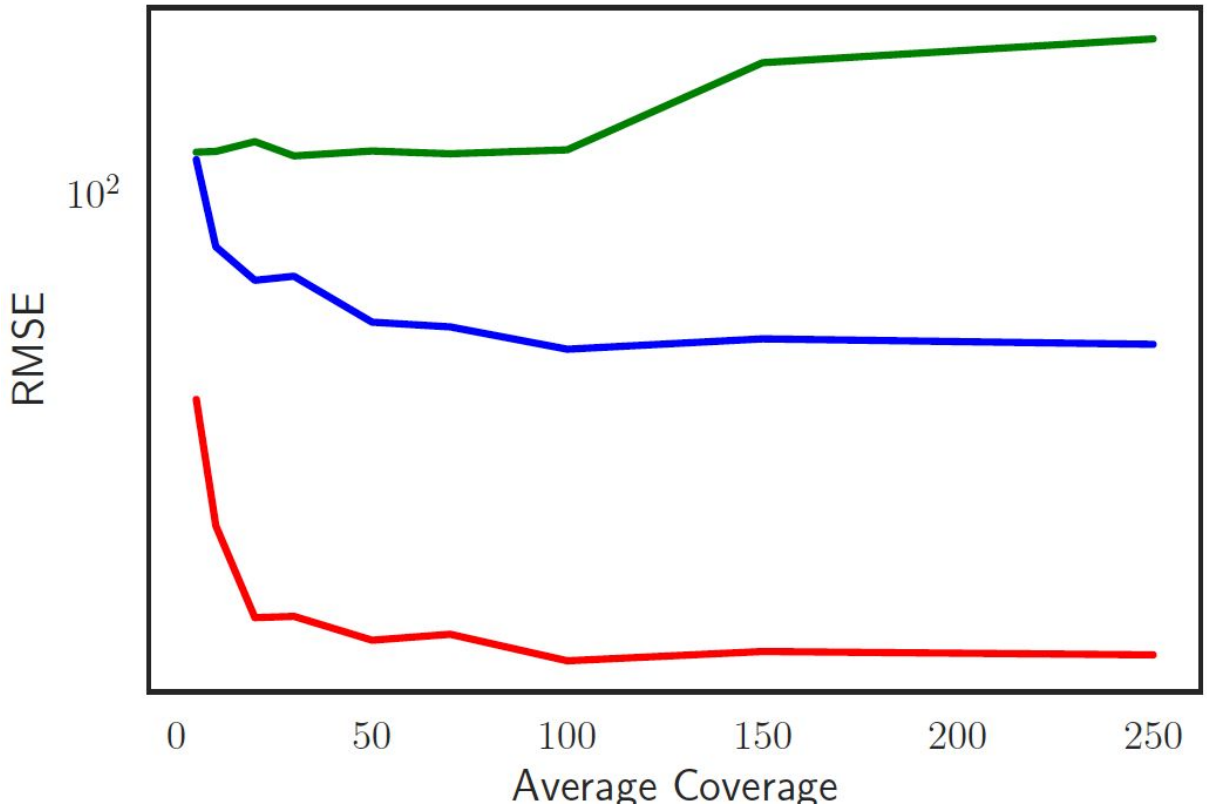
True vs. estimated repeat count for simulated data. Plots similar to **Figures 2B-D** are shown for each of the 10 loci analyzed in **Figure 2**. Red=GangSTR, blue=ExpansionHunter, green=TREDPARSE.

Supplementary Figure 7



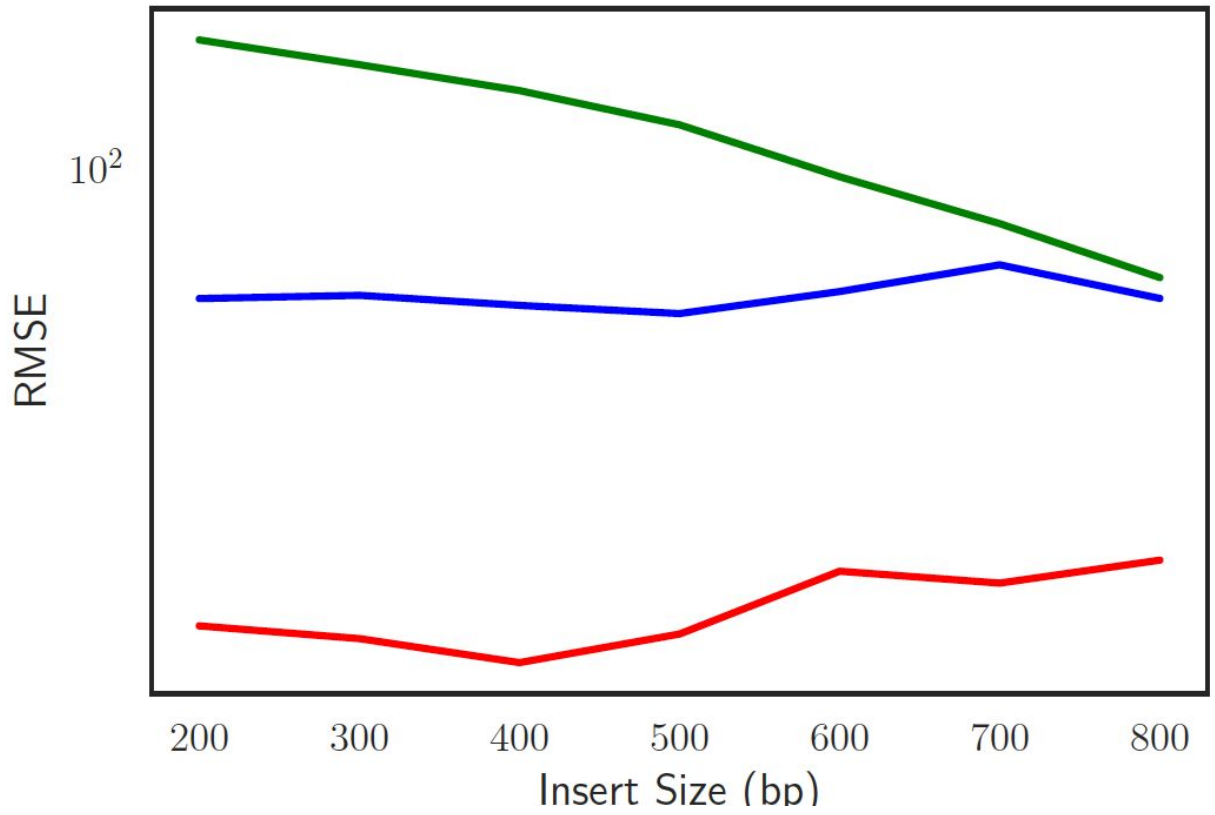
RMSE as a function of read length. Results shown for simulated data at the Huntington's Disease locus using simulation parameters described in the main text. Red=GangSTR, blue=ExpansionHunter, green=Tredparse.

Supplementary Figure 8



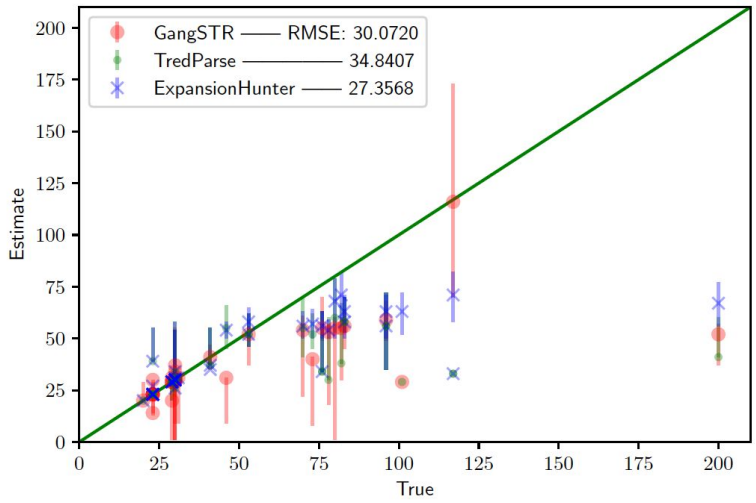
RMSE as a function of coverage. Results shown for simulated data at the Huntington's Disease locus using simulation parameters described in the main text. Red=GangSTR, blue=ExpansionHunter, green=Tredparse.

Supplementary Figure 9



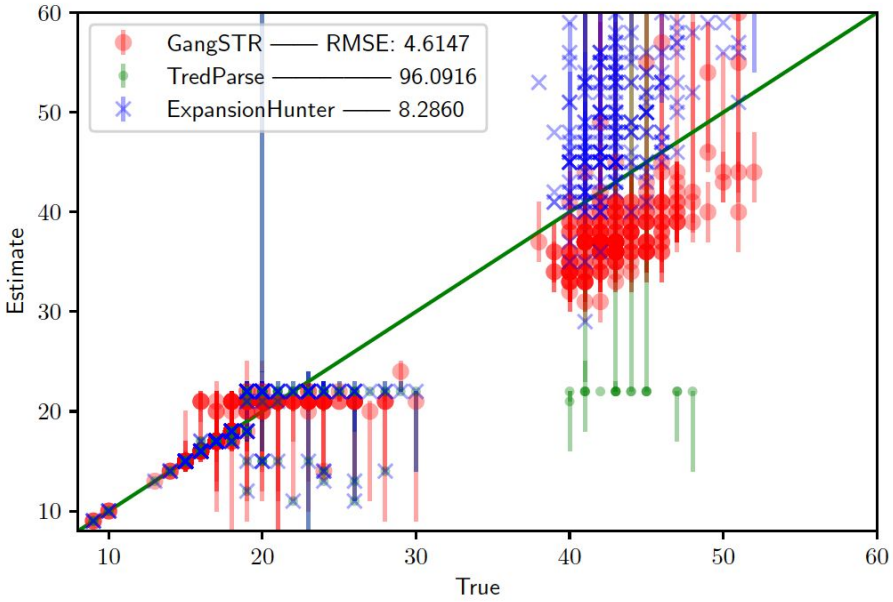
RMSE as a function of fragment length. Results shown for simulated data at the Huntington's Disease locus. Red=GangSTR, blue=ExpansionHunter, green=Tredparse.

Supplementary Figure 10



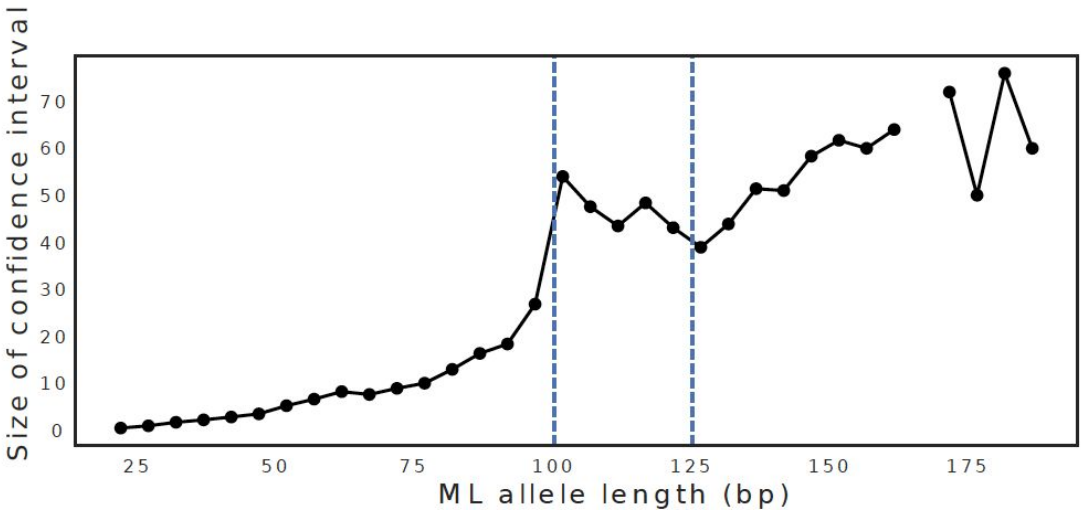
Comparison of repeat expansion tools at FMR1 using real WGS data. Dashed gray line gives the mean fragment length. Black solid line gives the diagonal. red=GangSTR; blue=ExpansionHunter; green=Tredparse.

Supplementary Figure 11



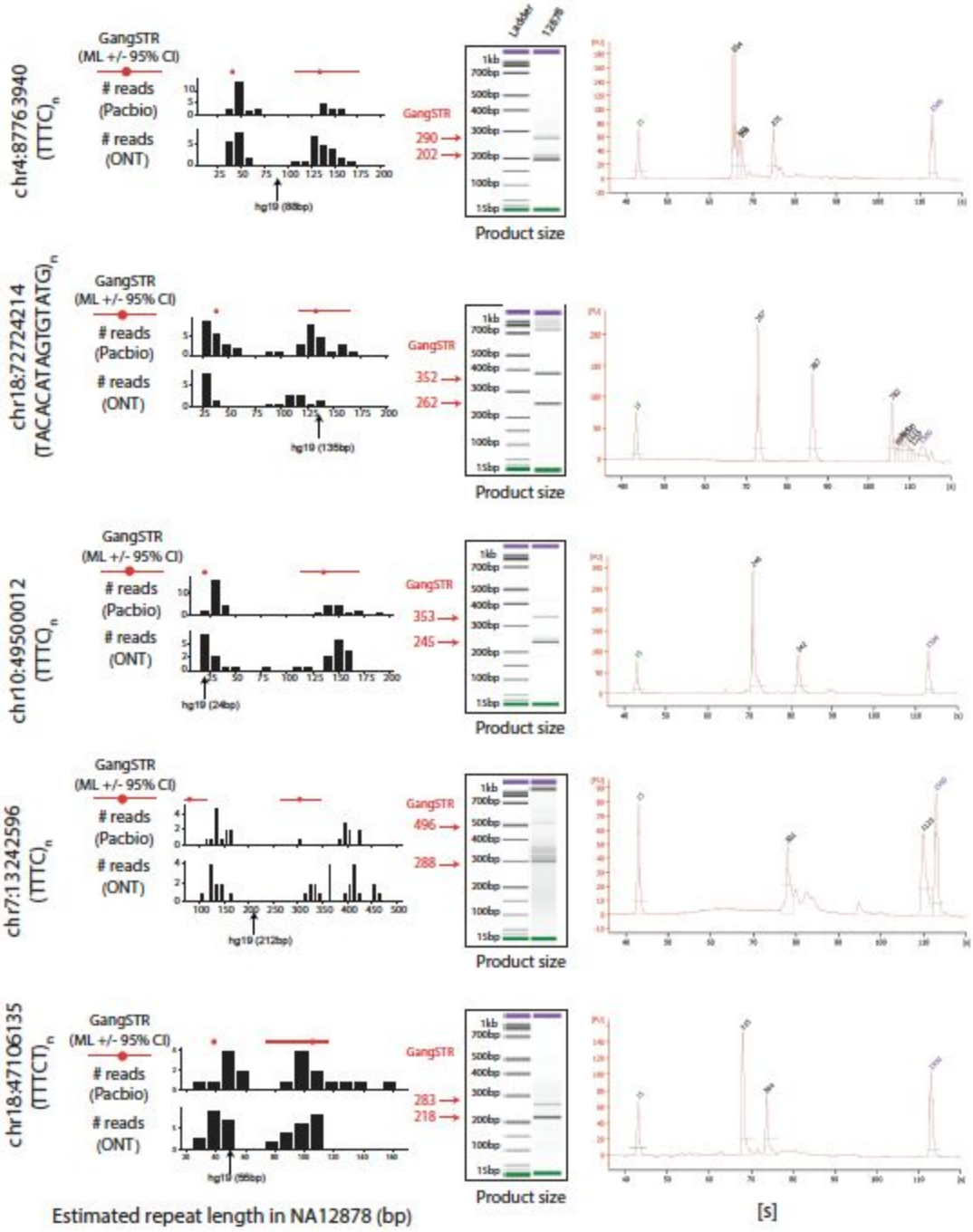
Comparison of repeat expansion tools at the Huntington's Disease locus using real whole exome sequencing data. Dashed gray line gives the mean fragment length. Black solid line gives the diagonal. red=GangSTR; blue=ExpansionHunter; green=Tredparse.

Supplementary Figure 12



Confidence interval size as a function of allele length in NA12878. The x-axis gives the allele length (binned at bp intervals) and the y-axis gives the mean size of the confidence intervals for all alleles with maximum likelihood lengths in that range. Blue dashed lines show the read length (101bp) and 126bp. TRs with alleles called in this range slightly above the read length give less precise allele length estimates.

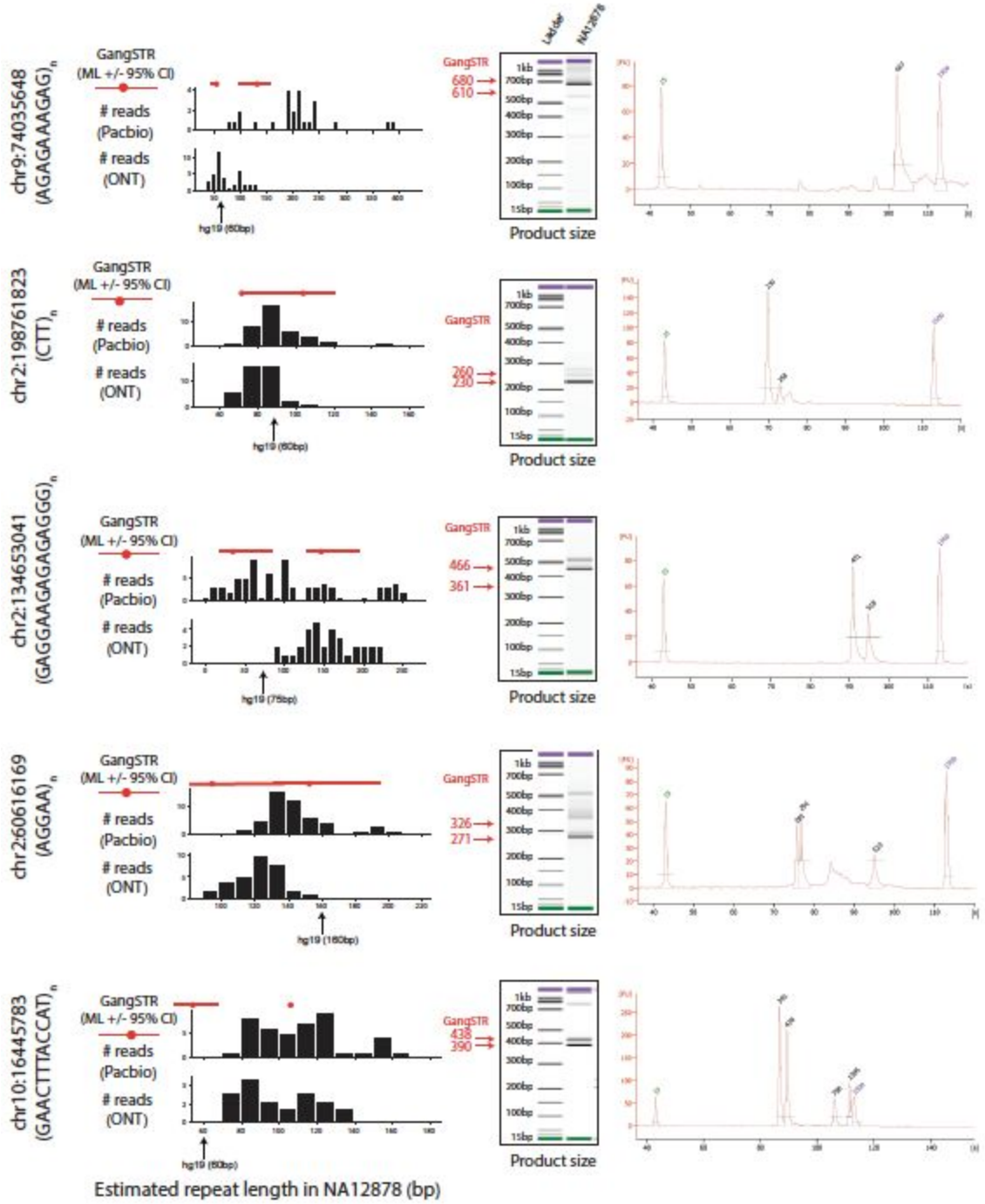
Supplementary Figure 13



Validation of long TR genotypes in NA12878. For each of the five loci shown, left plots compare GangSTR genotypes to those predicted by long reads. Red dots give the maximum likelihood repeat lengths predicted by GangSTR and red lines give the 95% confidence intervals for each allele. Black histograms give the distribution of repeat lengths supported by PacBio (top) and ONT (bottom) reads. The black arrow denotes the length in hg19. The middle plots show PCR product sizes for each locus as estimated using capillary electrophoresis. Left bands

show the ladder and right bands show product sizes in NA12878. Green and purple bands show the lower and upper limits of the ladder, respectively. Red arrows and numbers give product sizes expected for the two alleles called by GangSTR. Right plots give capillary electrophoresis traces. The x-axis shows seconds and y-axis shows arbitrary fluorescent units. Peaks are annotated with estimated product sizes.

Supplementary Figure 14



Example repeats with discordant lengths across long read technologies in NA12878. For each of the five loci shown, left plots compare GangSTR genotypes to those predicted by long reads. Red dots give the maximum likelihood repeat lengths predicted by GangSTR and red lines give the 95% confidence intervals for each allele. Black histograms give the distribution of repeat lengths supported by PacBio (top) and ONT (bottom) reads. The black arrow denotes the length in hg19. The middle plots show PCR product sizes for each locus as estimated using

capillary electrophoresis. Left bands show the ladder and right bands show product sizes in NA12878. Green and purple bands show the lower and upper limits of the ladder, respectively. Red arrows and numbers give product sizes expected for the two alleles called by GangSTR. Right plots give capillary electrophoresis traces. The x-axis shows seconds and y-axis shows arbitrary fluorescent units. Peaks are annotated with estimated product sizes.

Supplementary Tables

Supplementary Table 1: Target pathogenic repeats used in benchmarking experiments.

Abbreviation	Disease	Gene	Motif	Repeat location	Pathogenic cutoff
SCA1	Spinocerebellar ataxia 1	*ATXN1*	CAG	chr6:16327636–16327722 (hg38) chr6:16327867-16327953 (hg19)	39
SCA2	Spinocerebellar ataxia 2	*ATXN2*	CAG	chr12:111598951–111599019 (hg38) chr12:112036755-112036823 (hg19)	33
SCA3	Spinocerebellar ataxia 3	*ATXN3*	CAG	chr14:92071011–92071034 (hg38) chr14:92537355-92537378 (hg19)	60
SCA6	Spinocerebellar ataxia 6	*CACNA1A*	CAG	chr19:13207859–13207897 (hg38) chr19:13318673-13318711 (hg19)	20
SCA7	Spinocerebellar ataxia 7	*ATXN7*	CAG	chr3:63912686–63912715 (hg38) chr3:63898362-63898391 (hg19)	34
SCA8	Spinocerebellar ataxia 8	*ATXN8OS*	CTG	chr13:70139384–70139428 (hg38) chr13:70713516-70713560 (hg19)	80
SCA12	Spinocerebellar ataxia 12	*PPP2R2B*	CAG	chr5:146878729–146878758 (hg38) chr5:146258292-146258321 (hg19)	51
SCA17	Spinocerebellar ataxia 17	*TBP*	CAG	chr6:170561908–170562021 (hg38) chr6:170870996-170871109 (hg19)	43
HTT	Huntington's Disease	*HTT*	CAG	chr4:3074877–3074933 (hg38) chr4:3076604-3076660 (hg19)	40
DM1	Myotonic Dystrophy 1	*DMPK*	CTG	chr19:45770205–45770264 (hg38) chr19:46273463-46273522 (hg19)	50

Table modified from Tang *et al.* Table 1. Repeat locations are given for both hg19 and hg38 genomic coordinates.

Supplementary Table 2: Computational performance of repeat expansion tools.

Tool	Average CPU time
GangSTR - 100 bootstraps	13.2s
GangSTR - no bootstrap	10.2s
Tredparse	170.4s
ExpansionHunter (no cov)	12.8s
ExpnasionHunter	652.0s

All timing experiments were run on a 64 bit machine running CentOS 7.4.1708 using a single Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz processor. All tools were run on the 10 loci shown in **Supplementary Table 1**.

Supplementary Table 3: Summary of long repeat alleles identified in NA12878
See GangSTR_SuppTable3.xlsx

Supplementary Table 4: Motif enrichment for expanded repeats in NA12878

Motif	# Expansions in NA12878	# in hg19	P-value
AAAG/CTTT	87	11,669	1.67e-94
AAAGG/CCTTT	26	499	2.16e-48
AAAAG/CTTTT	8	4,168	8.11e-05
AAAAT/ATTTT	7	8,939	2.89e-02
AAG/CTT	7	11,272	7.92e-02

P-values were computed using one-sided Fisher's exact test.

Supplementary Table 5: Experimental validation of long TRs in NA12878

Locus (hg19)	F primer	R primer	Hg19 size (bp)	Predicted sizes (bp)	Observed sizes (bp)	Long reads
chr4:87763940-87764027	AGCTGTCCTGAGTTGCATCA	GACTGAGGCAGGAGAAATGC	242	202/290	194/275	Y
chr18:72724214-72724348	GGGCACCTGTGCTGAAAT	ATGAGTCGTTGGCAAAGTGT	352	262/352	257/387	Y
chr10:49500012-49500035	CCCCTCACCTCTTGTCTTTG	GCTACTTGGGAGCTGAGGTG	241	245/353	240/342	Y
chr7:13242596-13242807	GCATTTTCCTGATGGCTAAA	TTAGCCGGGTGTGGTAGC	400	288/496	302/1,133	Y
chr18:47106135-47106189	CCCTGATGCTCAGTCTTTCC	CCTGGGGAACAAGAGTGAAA	228	218/283	215/264	Y
chr9:74035648-74035707	GGCAAGGAGAAACAGATACCA	TTGCTGCAAAGGACGTGA	600	610/680	667	N
chr2:198761823-198761912	CCATAATGATACCTTTGGGGATA	TCCTCTATTTCAGCACAACTAGATACA	245	230/260	230/258	N
chr2:134653041-134653115	TTCCCTAGGGGAAGAGGAAG	CATGGTCACCGATAAGACCTTT	391	361/466	451/518	N
chr2:60616169-60616328	CTGGGCCACAGAATGAGACT	TTTACAGGTTGGCCACACAA	331	271/326	281/291	N
chr10:16445783-16445842	TGCCCAATAAGTATGAGAAGAACA	AAGTTCAAAAGGCCAGACCA	390	390/438	391/428	N
chr17:59583055-59583146	AAGACGGCAGTAAGCCAGAA	GGAGTGAACACGAGACAGCA	250	218/266	205/235	N

Predicted sizes give the maximum likelihood estimate from GangSTR for each allele. Observed sizes give the top two peaks observed in capillary electrophoresis. Note, traces were often messy and showed evidence for more than two alleles. See **Supplementary Figures 13** and **14** for raw capillary data. Long reads column is “Y” if GangSTR calls were concordant with both Pacbio and ONT. “N” indicates that long reads were discordant either with each other or with GangSTR calls. We chose five loci concordant with long read data for validation and six loci with inconclusive results using long reads.

Supplementary Table 6: Summary of GangSTR genotypes in 150 genomes

Cohort	# Loci genotyped	# loci heterozygous for allele >100bp	# loci homozygous for allele >100bp	# loci heterozygous for allele >150bp	# loci homozygous for allele >150bp
African (n=50)	513,998	44.4	7.9	5.5	0.76
East Asian (n=50)	515,494	44.2	10.5	5.5	0.72
European (n=50)	516,662	39.6	9.3	4.7	0.72

Each value gives the mean across all individuals analyzed.

Supplementary Table 7: Long repeats with discordant allele frequency spectra across populations

STR Locus (hg19)	Motif	ANOVA p-value	Annotation	Mean length (EUR)	Mean length (AFR)	Mean length (EAS)
chr21:36720944-36721033	AATAG	1.8e-13	Intron (*RUNX1*)	22.07	16.21	24.10
chr2:158410717-158410780	AAAG	2.0e-12	Intron (*ACVR1C*)	18.07	21.69	24.48
chr5:17883973-17884024	AAAG	4.3e-9	Intron (*BC028204*)	24.10	15.77	21.50
chr2:119670373-119670420	AAAG	4.3e-7	Intergenic	19.30	14.28	14.25
chr14:56113179-56113263	AAAGG	2.9e-7	Intron (*KTN1*)	18.22	14.72	18.34
chr7:26995285-26995349	AAAGG	4.2e-6	Intron (*SKAP2*)	18.60	15.47	18.73
chr3:163246863-163246934	AAAG	7.4e-7	Intergenic	20.86	19.97	27.06
chr2:88703915-88704010	AAAG	1.2e-6	Intergenic	24.69	21.85	26.52
chr17:32835083-32835187	AAAGG	1.8e-6	Intergenic	19.67	16.93	18.31
chr4:54751299-54751364	AAG	2.8e-6	Intron (*PDGFRA*)	28.56	18.61	27.75
chr12:80408618-80408697	AAAGG	7.2e-6	Intergenic	15.81	17.72	16.76
chr7:113785487-113785551	AAAAG	1.1e-5	Intron (*FOXP2*)	13.83	14.61	10.25
chr18:70874677-70874735	AAAACATATATATGT	1.8e-5	Intron (*LOC400655*)	3.61	6.44	3.42
chr1:11491766-11491837	AAAG	2.2e-5	Intergenic	20.74	16.67	20.11
chr3:102856807-102856876	AAAAT	2.8e-5	Intergenic	17.29	13.45	18.95
chr22:23333498-23333585	AGAGAGGG	3.1e-5	Intergenic	10.04	9.63	11.43
chr2:127327339-127327466	AAAG	4.3e-5	Intergenic	24.82	26.95	29.03

chr6:21965536-21965640	AAGAAGGA GGAGGGG	5.7e-5	Intron (*LINC00340*)	3.57	4.65	5.00
chr10:118332630-118332725	AAAG	6.6e-5	Intergenic	23.32	22.86	29.56
chr13:32333513-32333580	AAAG	6.7e-5	Intron (*FXFP2*)	23.33	23.82	18.50

Lengths are given in multiples of the repeat unit.