The American Journal of Human Genetics, Volume *101*

Supplemental Data

Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes

Haibao Tang, Ewen F. Kirkness, Christoph Lippert, William H. Biggs, Martin Fabani, Ernesto Guzman, Smriti Ramakrishnan, Victor Lavrenko, Boyko Kakaradov, Claire Hou, Barry Hicks, David Heckerman, Franz J. Och, C. Thomas Caskey, J. Craig Venter, and Amalio Telenti

Supplementary figures (9 total)

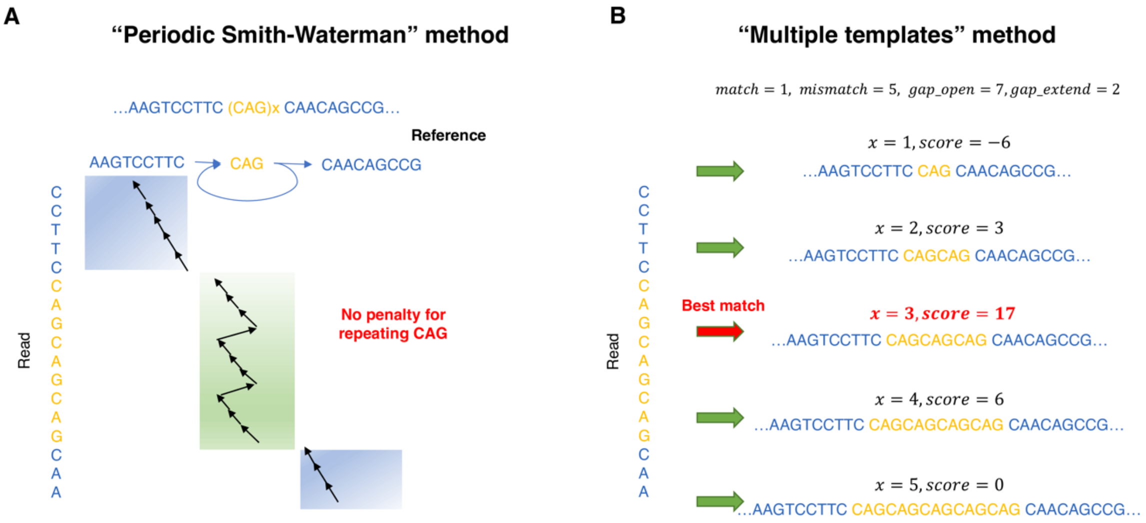
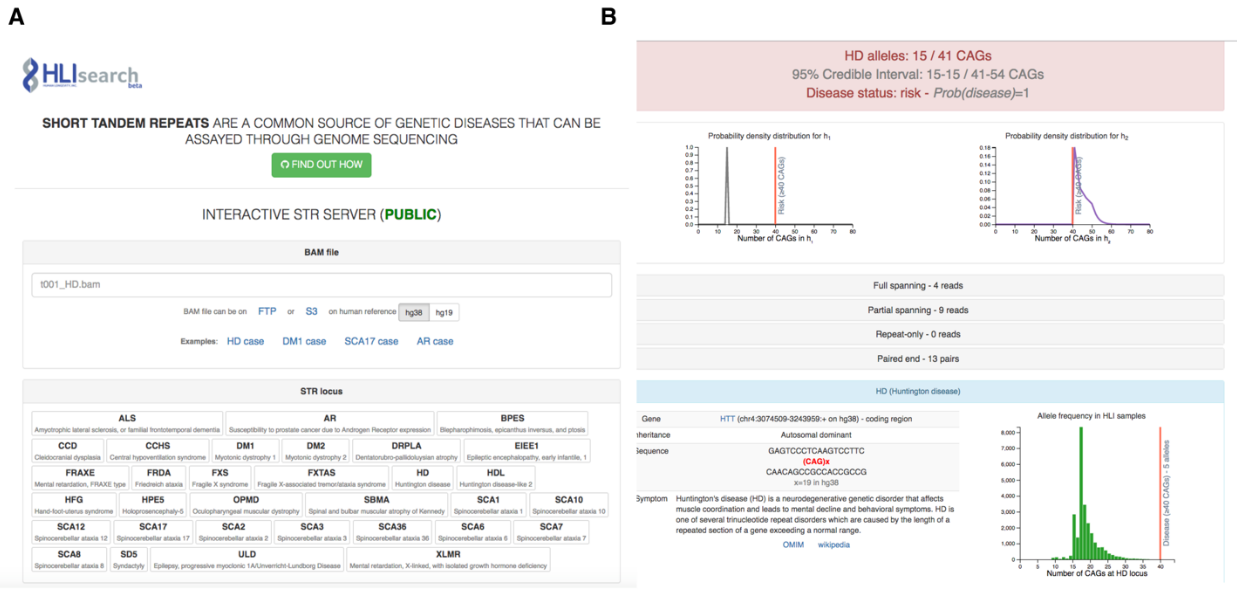

A
"Periodic Smith-Waterman" method
...AAGTCCTTC (CAG)x CAACAGCCG...
Reference
AAGTCCTTC
CAG
CAACAGCCG
Read
C C T T C C A G C A G C A G C A A
No penalty for repeating CAG
B
"Multiple templates" method
$match = 1,\ mismatch = 5,\ gap\_open = 7, gap\_extend = 2$
$x = 1, score = -6$
...AAGTCCTTC CAG CAACAGCCG...
$x = 2, score = 3$
...AAGTCCTTC CAGCAG CAACAGCCG...
Best match
$x = 3, score = 17$
...AAGTCCTTC CAGCAGCAG CAACAGCCG...
$x = 4, score = 6$
...AAGTCCTTC CAGCAGCAGCAG CAACAGCCG...
$x = 5, score = 0$
...AAGTCCTTC CAGCAGCAGCAGCAG CAACAGCCG...
Read
C C T T C C A G C A G C A G C A A

Figure S1. Comparison of two sequence alignment methods exploiting the periodicity of the STR sequences. (**A**) "Periodic Smith-Waterman" method modifies the recurrence table when performing the dynamic programming step so that repeat units are not penalized during matching; (**B**) "Multiple templates" method aligns read to a series of templates embedded with varying number of repeats, using standard SW alignment with a fixed scoring scheme. The alignment yields a series of alignments with different scores which are then compared to determine the repeat size that corresponds to the highest score. "Multiple templates" method is the selected alignment method in TREDPARSE.



SHORT TANDEM REPEATS ARE A COMMON SOURCE OF GENETIC DISEASES THAT CAN BE ASSAYED THROUGH GENOME SEQUENCING

INTERACTIVE STR SERVER (**PUBLIC**)

BAM file

t001_HD.bam

BAM file can be on FTP or S3 on human reference hg38 hg19

Examples: HD case DM1 case SCA17 case AR case

STR locus

ALS Amyotrophic lateral sclerosis, or familial frontotemporal dementia
AR Susceptibility to prostate cancer due to Androgen Receptor expression
BPES Blepharophimosis, epicanthus inversus, and ptosis
CCD Cleidocranial dysplasia
CCHS Central hypoventilation syndrome
DM1 Myotonic dystrophy 1
DM2 Myotonic dystrophy 2
DRPLA Dentatorubro-pallidoluysian atrophy
EIEE1 Epileptic encephalopathy, early infantile, 1
FRAXE Mental retardation, FRAXE type
FRDA Friedreich ataxia
FXS Fragile X syndrome
FXTAS Fragile X-associated tremor/ataxia syndrome
HD Huntington disease
HDL Huntington disease-like 2
HFG Hand-foot-uterus syndrome
HPE5 Holoprosencephaly-5
OPMD Oculopharyngeal muscular dystrophy
SBMA Spinal and bulbar muscular atrophy of Kennedy
SCA1 Spinocerebellar ataxia 1
SCA10 Spinocerebellar ataxia 10
SCA12 Spinocerebellar ataxia 12
SCA17 Spinocerebellar ataxia 17
SCA2 Spinocerebellar ataxia 2
SCA3 Spinocerebellar ataxia 3
SCA36 Spinocerebellar ataxia 36
SCA6 Spinocerebellar ataxia 6
SCA7 Spinocerebellar ataxia 7
SCA8 Spinocerebellar ataxia 8
SD5 Syndactyly
ULD Epilepsy, progressive myoclonic 1A/Unverricht-Lundborg Disease
XLMR Mental retardation, X-linked, with isolated growth hormone deficiency

HD alleles: 15 / 41 CAGs
95% Credible Interval: 15-15 / 41-54 CAGs
Disease status: risk - *Prob(disease)*=1

Full spanning - 4 reads

Partial spanning - 9 reads

Repeat-only - 0 reads

Paired end - 13 pairs

HD (Huntington disease)

Gene	HTT (chr4:3074509-3243959:+ on hg38) - coding region
nheritance	Autosomal dominant
Sequence	GAGTCCCTCAAGTCCTTC (CAG)x CAACAGCCGCCACCGCCG x=19 in hg38
Symptom	Huntington's disease (HD) is a neurodegenerative genetic disorder that affects muscle coordination and leads to mental decline and behavioral symptoms. HD is one of several trinucleotide repeat disorders which are caused by the length of a repeated section of a gene exceeding a normal range. OMIM wikipedia

Figure S2. An interactive server for computing the STR calls. (**A**) Input includes the BAM location, reference genome version and the STR locus of interest. (**B**) Output includes detailed information about the STR – including call results, posterior probability density of the size of risk alleles, and various types of reads affecting the final calls.

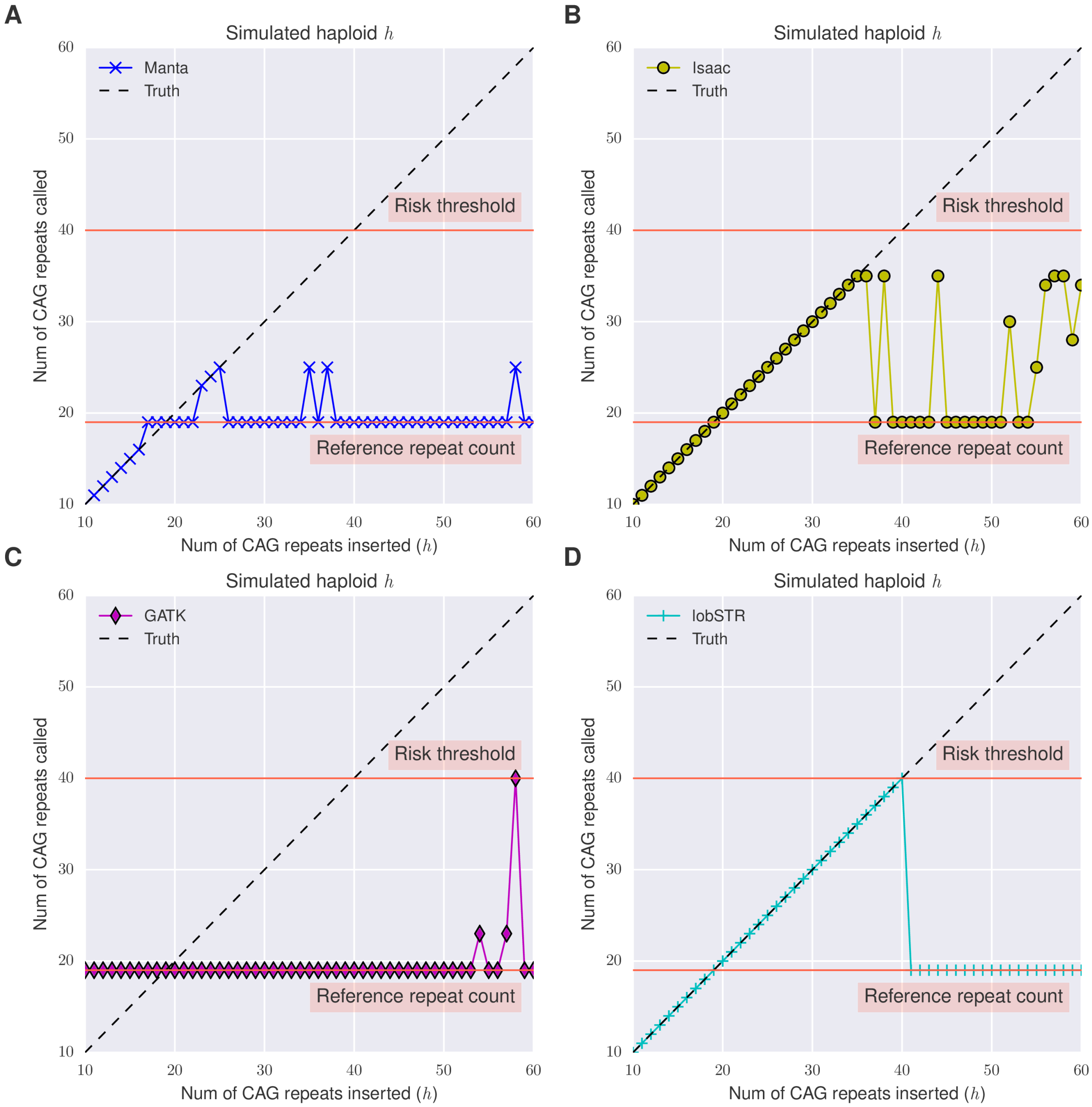


Figure S3. Simulations with synthetic datasets of implanted STR alleles at Huntington (HD) locus. We have tested performance of several variant callers, including (**A**) Manta (**B**) Isaac (**C**) GATK, (**D**) lobSTR.

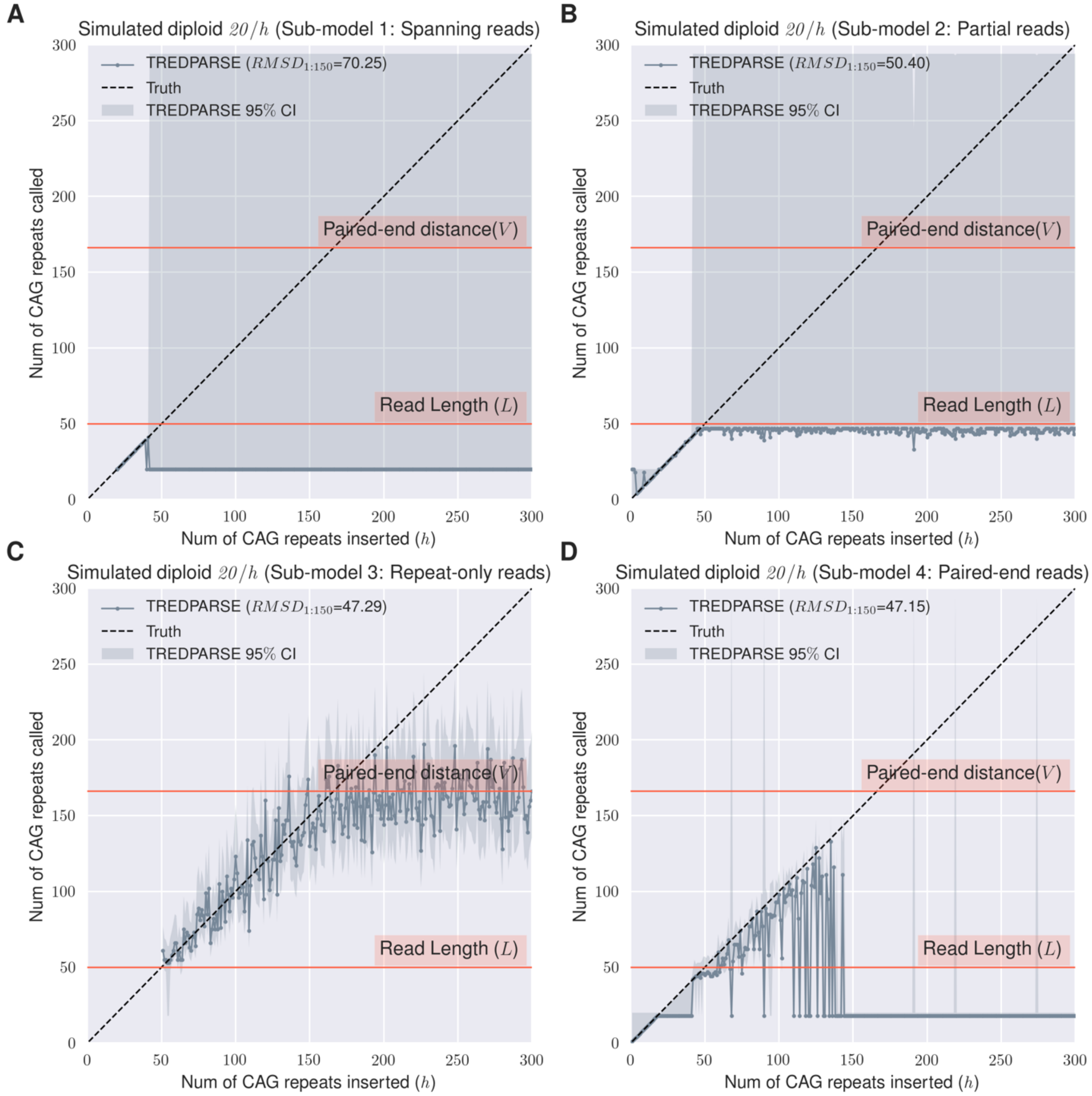


Figure S4. Predictive power of each of the four types of evidence. Each type of evidence has their own specific predictive range. In this simulation, TREDPARSE is run (**A**) using only spanning reads; (**B**) using only partial reads; (**C**) using only repeat reads; (**D**) using only paired-end distance. Shaded region represents 95% credible interval for TREDPARSE estimates of h. $RMSD$ represents root-mean-square deviation, calculated as $RMSD = \frac{1}{N}\sqrt{\sum_{i=1:N}\left(h_i - \widehat{h_i}\right)^2}$, where $N = 150$.

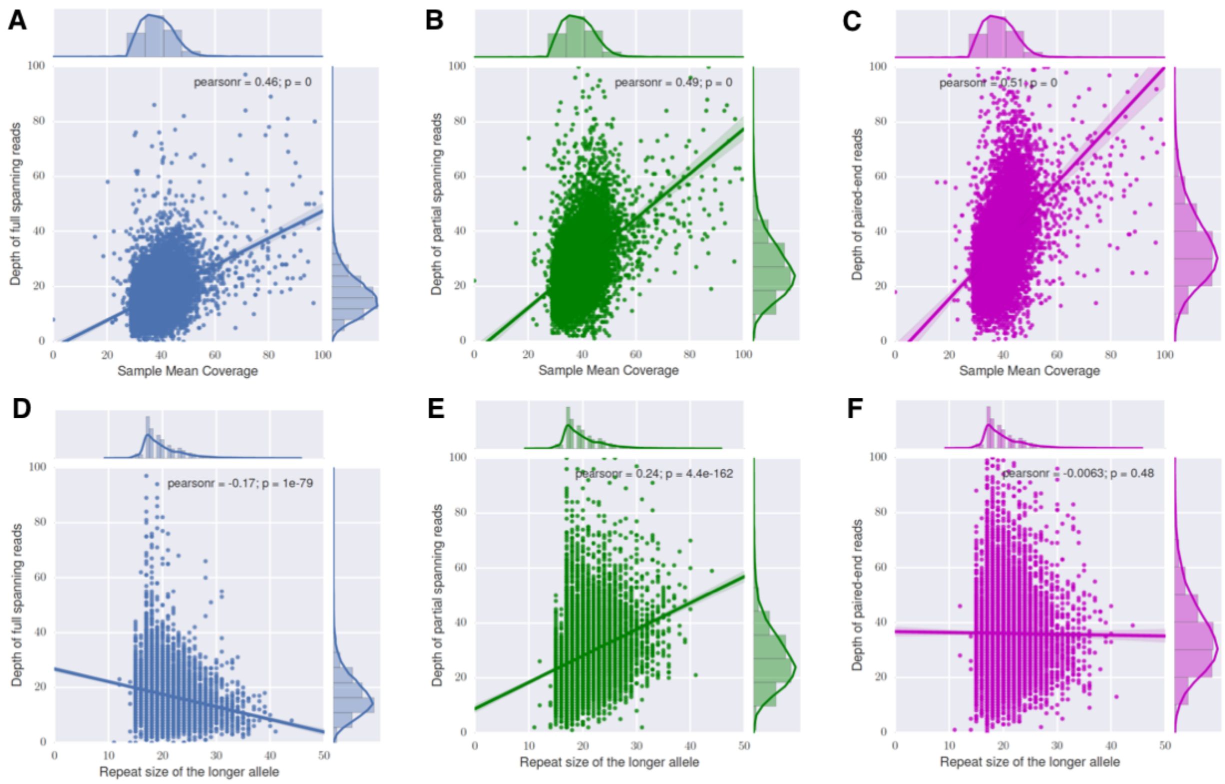
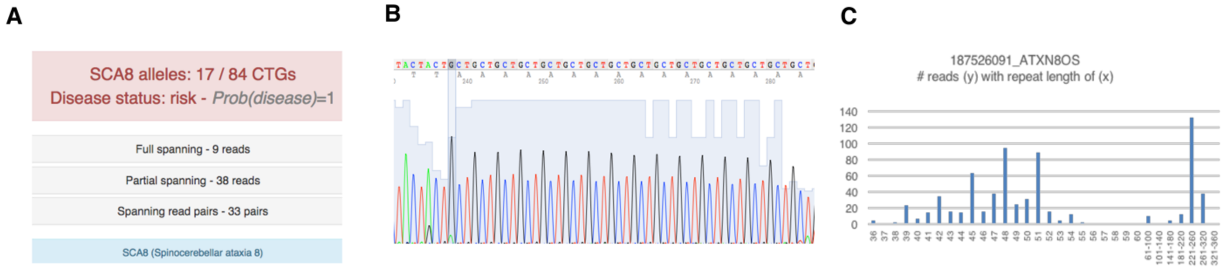


Figure S5. Amount of evidence as a function of sample read depth coverage and length of the repeat allele at Huntington (HD) locus. Based on our HLI samples, relationships are shown between sample mean coverage and depth of (**A**) spanning reads, (**B**) partial reads, (**C**) spanning pairs, as well as the repeat units of the longer allele and read depth of (**D**) spanning reads, (**E**) partial reads, (**F**) spanning pairs.



SCA8 alleles: 17 / 84 CTGs
Disease status: risk - *Prob(disease)*=1

Full spanning - 9 reads

Partial spanning - 38 reads

Spanning read pairs - 33 pairs

SCA8 (Spinocerebellar ataxia 8)

TREDPARSE at SCA8
Call 17/84

Sanger identifies only one allele
Call 17/17

ONP confirms the longer allele
predicted by TREDPARSE

Figure S6. Example of validation of TREDPARSE calls using Sanger and Oxford Nanopore sequencing. In this example, there is disagreement between (**A**) TREDPARSE call (two alleles 17 and 84) and (**B**) Sanger sequencing which identified only the allele with size 17. (**C**) Oxford Nanopore sequencing confirms the longer allele, showing two peaks of allele sizes that both match the prediction of TREDPARSE. Sample mean coverage of the input BAM is 33×.

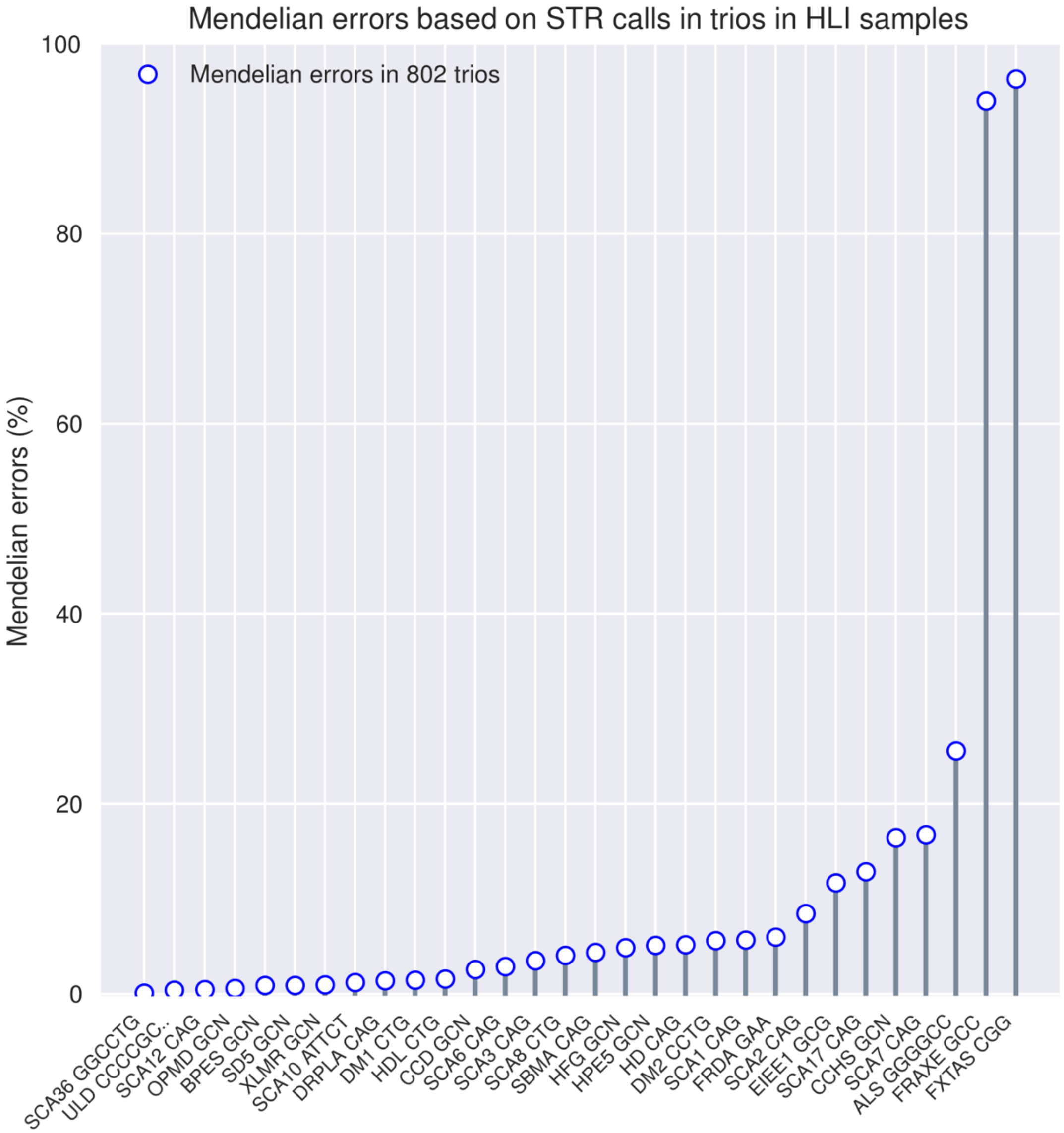


Figure S7. Mendelian errors based on trios and duos in HLI samples. For each of the 802 trio families in HLI samples, a Mendelian error was called if the child alleles are improbable given the parent alleles. The error rate was calculated as the number of families containing a Mendelian error divided by the total number of families that have all three individuals called by TREDPARSE.

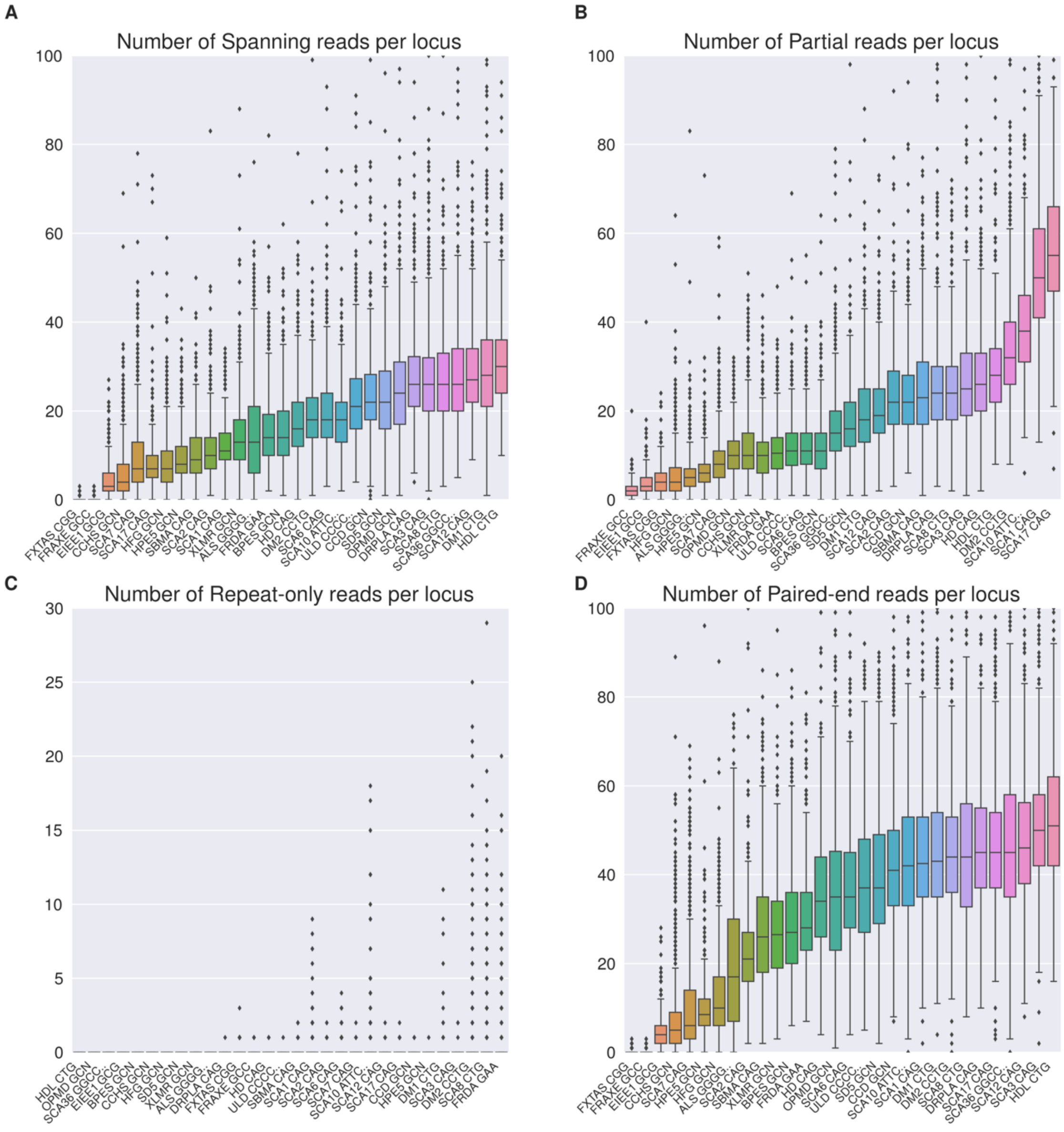


Figure S8. Box plot of amount of read evidence across 30 TRED loci in HLI samples. We plot the distribution of number of reads at each locus, for all read types – including (**A**) spanning reads; (**B**) partial reads; (**C**) repeat-only reads; (**D**) paired-end reads. Box shows the inter-quartile (IQR) range, with bar in the middle indicating median values. Outliers are shown in dots outside the IQR range. In the repeat-only reads panel (C), the box plot is not visible due to zero counts of the repeat-only reads for most loci in most individuals.

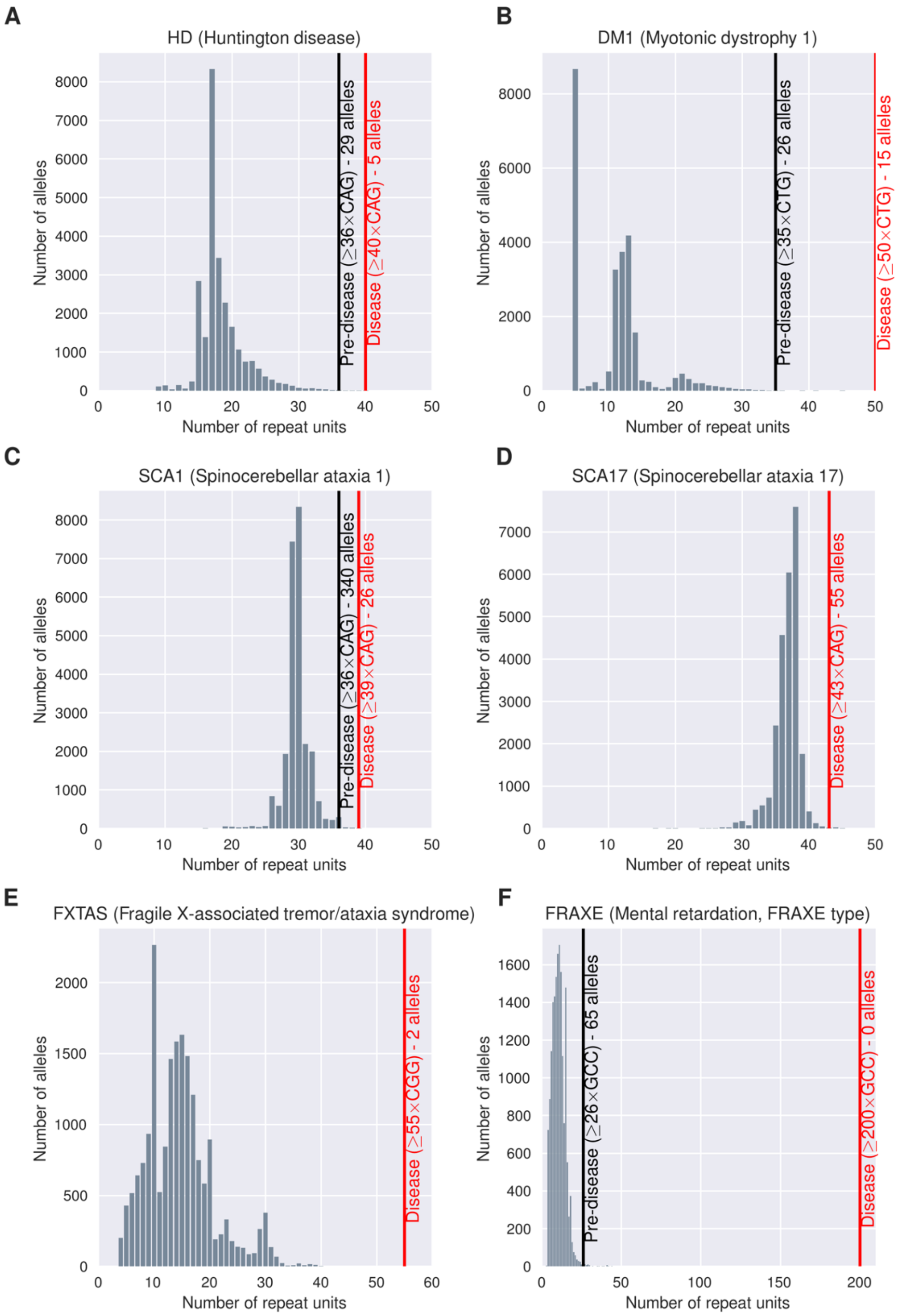


Figure S9. Allele frequencies within the whole genome samples at selected STR loci. Selected STR loci (**A-F**) include Huntington's disease (HD), Myotonic dystrophy (DM1), Spinocerebellar Ataxia Type 1 (SCA1), Spinocerebellar Ataxia Type 17 (SCA17), Fragile X-associated tremor/ataxia syndrome (FXTAS) and Mental retardation, FRAXE type (FRAXE).

Supplementary tables (5 total)

Table S1. STR disease prevalence in HLI samples when compared to the known prevalence estimates based on literature review.

[See EXCEL file **Table-S1.xlsx**]

Table S2. Number of spanning, partial, repeat-only and paired-end reads identified by TREDPARSE for each of the 138 individuals with risk alleles.

[See EXCEL file **Table-S2.xlsx**]

Table S3. Validation of TREDPARSE using 6 cell line samples from the Genetic Testing Reference Materials Coordination Program (GeT-RM). lobSTR was run with parameters `` `--max-diff-ref 150 --realign --noise_model illumina_v3.pcrfree` ``. TREDPARSE was run with parameter `` `--useclippedreads` ``.

Sample ID	Disorder	Truth	lobSTR	Expansion-Hunter	TREDPARSE	TREDPARSE status
NA05164	[DM1] - Myotonic dystrophy 1	21/340	11/21	56/63	21/74	short allele matches, long allele -266
NA06075	[DM1] - Myotonic dystrophy 1	12/66	12/12	12/38	12/64	short allele matches, long allele -2
NA20236	[FXTAS/FXS] - Fragile X-associated tremor/ataxia syndrome / Fragile X syndrome	31/53	NO CALL	16/18	18/145	no matches
NA20239	[FXTAS/FXS] - Fragile X-associated tremor/ataxia syndrome / Fragile X syndrome	20/183 -193	NO CALL	20/22	20/163	short allele matches, long allele -20
NA20250	[HD] - Huntington disease	15/40	15/18	15/45	15/40	both alleles match
NA20252	[HD] - Huntington disease	22/66	22/22	22/55	22/59	short allele matches, long allele -7

Table S4. Validation of samples with risk alleles by CLIA Sanger sequencing and Oxford Nanopore. Oxford Nanopore (ONP) repeat sizes follow a wide distribution and we use the median of a distinctive peak to represent the allele size. lobSTR was run with parameters `` `--max-diff-ref 150 --realign --noise_model illumina_v3.pcrfree` ``. One SBMA tested sample was determined to be a *carrier* but had been included in the validation to test the existence of risk allele. ND: no data.

Sample ID	Disorder	Sanger	ONP	lobSTR	Expansion-Hunter	TREDPARSE	TREDPARSE status
1	[HD] - Huntington disease	15/41	14/34	15/19	15/41	15/41	Validated by Sanger
2	[HD] - Huntington disease	21/41	18/33	19/21	21/54	21/40	Validated by Sanger (long allele -1)
3	[HD] - Huntington disease	16/43	15/38	16/16	16/55	16/44	Validated by Sanger (long allele +1)
4	[HD] - Huntington disease	25/40	19/37	19/25	25/40	25/40	Validated by Sanger
5	[SBMA] - Spinal and bulbar muscular atrophy of Kennedy	19/21	34/34	23/23	40/40	40/40	Validated by ONP
6	[SBMA] - Spinal and bulbar muscular atrophy of Kennedy	24/25	24/34	25/25	25/36 (risk *carrier*)	25/36 (risk *carrier*)	Validated by ONP
7	[SCA1] - Spinocerebellar ataxia 1	29/39	ND	29/29	30/40	29/39	Validated by Sanger
8	[SCA1] - Spinocerebellar ataxia 1	30/39	ND	30/39	31/40	30/39	Validated by Sanger

9	[SCA1] - Spinocerebellar ataxia 1	29/42	ND	29/29	53/56	29/45	Validated by Sanger (long allele +3)
10	[SCA2] - Spinocerebellar ataxia 2	22/39	ND	22/22	ND	22/39	Validated by Sanger
11	[SCA8] - Spinocerebellar ataxia 8	ND	14/80	15/15	14/62	15/84	Validated by ONP
12	[SCA8] - Spinocerebellar ataxia 8	17/17	16/84	13/16	16/87	17/84	Validated by ONP
13	[SCA17] - Spinocerebellar ataxia 17	38/46	ND	38/38	38/45	38/46	Validated by Sanger
14	[SCA17] - Spinocerebellar ataxia 17	37/43	ND	37/43	37/43	37/43	Validated by Sanger
15	[SCA17] - Spinocerebellar ataxia 17	38/43	35/40	26/38	38/43	38/43	Validated by both
16	[SCA17] - Spinocerebellar ataxia 17	37/44	ND	37/44	37/44	37/44	Validated by Sanger
17	[SCA17] - Spinocerebellar ataxia 17	37/45	33/42	37/37	37/52	37/44	Validated by both
18	[SCA17] - Spinocerebellar ataxia 17	36/43	ND	36/43	36/43	36/43	Validated by Sanger
19	[DM1] - Myotonic dystrophy 1	5/44+	ND	5/5	58/67	5/53	Validated by Sanger

Table S5. Number of spanning, partial, repeat-only, paired-end reads and STR calls identified by TREDPARSE for each of the trio families used in the Mendelian error estimates. Each worksheet contains a single STR locus. STR calls are in the form of “X I Y” for autosomal loci, and “X I .” for male individuals called at X-linked loci to indicate hemizygosity. Rows highlighted in green for “Correct”, red for “Error” and yellow for “Missing”.

[See EXCEL file Table-S5.xlsx]