Supplemental Methods

April 24, 2017

1 Background

Short tandem repeats (STRs) are hyper-mutable sequences in the human genome that are often used in forensics and population genetics, and are also the underlying cause for many genetic diseases. There are challenges associated with accurately determining the length polymorphism of STR loci in the genome. In particular, accurate detection of pathological STR expansion is limited by the sequence read length during whole genome analysis.

We implemented TREDPARSE, a software that incorporates evidence from read alignment, paired-end distance distribution, as well as a sequence stutter model using a probabilistic framework to accurately infer the repeat sizes for 30 known disease loci. Using simulated data, we show that TREDPARSE consistently outperforms available software. We analyzed full genome sequences of 12,632 individuals at Human Longevity Inc. (HLI) [1] that were sampled to an average depth of $\sim 30 - 40\times$ with Illumina sequencing. We show that TREDPARSE has excellent precision on the whole genome sequencing datasets by confirming all the calls.

Simulation with synthetic data suggests that TREDPARSE out-performs many other callers of short tandem repeats. We compared TREDPARSE with commonly used general-purpose variant callers, including Manta, Isaac, and GATK. Not unexpectedly, they perform poorly on the simulated datasets [**Fig. S1**]. The general-purpose variant callers can detect small indels, but in most cases fail to recover the length of long alleles (i.e. large indels). Additionally, the indels could occur at different locations within the repeat tract, making direct calling of the repeat size difficult without further post-processing. Based on these comparisons, we found that most tools tested thus far were not effective at identifying the allele, especially when the size of the allele exceeds the typical read length [**Fig. S1**]. Since the longer allele is often the critical allele for determining the disease status since most STR-related diseases have dominant inheritance and repeat expansion is more often pathological, the lack of sensitivity at longer alleles suggests that most 'at risk' individuals could not be accurately typed using those tools.

2 Probabilistic model for calling STRs

2.1 Parameters in the model

To fully model the uncertainties of observing a set of reads that are generated by a certain repeat size, we built a probabilistic model for predicting the size of STRs, based on evidence from spanning reads, partial reads, repeat-only reads and spanning pairs. The spanning reads align to both flanks of the target repeat; the partial reads align to only one flank of the repeat; the repeat-only reads align entirely within the repeat tract and thus consist entirely of repeat units; the spanning pairs are read pairs that span the STR region, i.e. with one end on each side of the repeat. We will describe these components separately below, with the following notations:

- L: read length in base pairs (bp), e.g. $L = 150$ for 150 base pairs reads
- D: haplotype depth, average sequencing depth divided by ploidy. For diploid locus, it is equal to half of the sequencing depth
- F: number of base pairs required to call flanking sequence. By default, TREDPARSE requires at least 9 bp when matching flanking sequences so we have $F = 9$
- R: number of repeat units in the reference sequence
- K: repeat unit length, e.g. $K = 3$ for triplet 'CAG' repeats
- S: observed number of repeat units in a *spanning read*
- T: observed number of repeat units in a *partial read*
- U: number of *repeat-only reads* that consist of entirely repeats
- V: observed paired-end distance in bp for a *spanning pair*
- h_1, h_2: number of repeat units in two alleles, respectively. Without loss of generality, we assume $1 \leq h_1 \leq h_2 \leq h_{max}$. For a haploid locus (such as the X-linked locus in a male), we have $h_1 = h_2$

To avoid confusion, the *repeat length* is equal to the *repeat units* $\times$ *repeat unit length*. For example, the human reference genome (hg38) has $R = 19$ for the Huntington locus, which is a repeat of 'CAG' ($K = 3$), so $RK = 57$ is the total repeat length in base pairs.

Formally, our observations are a set of l spanning reads with repeat units $S_{1:l}$, m partial reads with repeat units $T_{1:m}$, U repeat-only reads, and n spanning pairs with paired end distance in base pairs $V_{1:n}$. Our goal is to estimate h_1 and h_2 that maximize the likelihood of the set of observations $\{S_{1:l}, T_{1:m}, U, V_{1:n}\}$.

2.2 Spanning reads

The spanning reads are the reads that show both left and right flanking sequences of at least F bp. The spanning reads are quite straightforward, with the counted size matching or close to the true size. The spanning reads would show exactly the size of the underlying allele if there is no noise due to stuttering. The sharp peak becomes 'fuzzier' after incorporating the stuttering noise. We use the stuttering model trained by lobSTR [2] which considers the periodicity of the repeat as well as the GC content. The stuttering model allows a certain proportion of the spanning reads to show a different size than the true allele size.

To account for the stutter noise, we use the following model, similar to the stutter model used in lobSTR [2]. With probability $\pi(K)$, the read is a product of stutter noise, which is dependent on the repeat unit length K and also the GC content of the locus. If a read is a product of stutter, then with probability $Poisson(s;\lambda_K)$, the noisy read deviates by s units from the original allele, where $Poisson(s;\lambda_K)$ is a Poisson distribution with mean λ_K. Deviation can be either positive or negative with equal probability $\pi(K)/2$. Parameters $\pi(K)$ and λ_K were previously trained by lobSTR [2] for a range of values K.

Hence, the probability of generating a spanning read with S observed repeat units in the STR region from a hemizygous locus with an STR with h repeat units:

$$P_S(S|h) = \begin{cases} 1-\pi(K), & \text{if } S = h \\ \frac{\pi(K)}{2} Poisson(|S-h|-1, \lambda_K), & \text{otherwise} \end{cases} \tag{1}$$

For a diploid STR locus with h_1 and h_2 repeat units, we then have a mixed distribution with mixing rate π_S:

$$\pi_S = \frac{s(h_1)}{s(h_1)+s(h_2)} \tag{2}$$

where

$$s(h) = \begin{cases} L-2F-hK, & \text{if } hK < L-2F \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Note that there may not be any spanning reads expected when $s(h_1) = s(h_2) = 0$ if both allele lengths are longer than $L-2F$. In that case, we set $\pi_S = 0.5$. We then have the mixing distribution for observing a spanning read with size S, given that the true STR repeat units for each of the two alleles are h_1 and h_2:

$$P_S(S|h_1,h_2) = \pi_S P_S(S|h_1) + (1-\pi_S) P_S(S|h_2) \tag{4}$$

In the case of spanning reads, the longer allele typically has a *smaller* contribution to the number of observations [**Fig. S4D**].

2.3 Partial reads

The partial reads do not align all the way across the repeat region and shows only one flanking sequence. The partial reads have a probability mass function of discrete uniform distribution. Unlike the full spanning reads which show exactly the repeat units of the underlying allele, the partial reads only show a lower bound for the number of repeat units of the underlying allele. The inference is also analogous to the "German tank problem" but with replacement, under the condition that the allele cannot exceed $L - F$.

The probability of generating a partial read with T observed repeat units in the STR region from a hemizygous locus with an STR with h repeat units:

$$P_T(T|h) = Uniform(0, h) \tag{5}$$

For a diploid STR locus with h_1 and h_2 repeat units, we have a mixed distribution with mixing rate π_T:

$$\pi_T = \frac{t(h_1)}{t(h_1) + t(h_2)} \tag{6}$$

where

$$t(h) = \begin{cases} L - F, & \text{if } hK > L - F \\ hK, & \text{otherwise} \end{cases} \tag{7}$$

We then have the mixing distribution for observing a partial read with size T, given that the true STR repeat units for each of the two alleles are h_1 and h_2:

$$P_T(T|h_1, h_2) = \pi_T P_T(T|h_1) + (1 - \pi_T) P_T(T|h_2) \tag{8}$$

In the case of partial reads, the longer allele typically has a *larger* contribution to the number of observations [**Fig. S4E**].

2.4 Repeat-only reads

Reads that consist entirely of repeat units are called 'repeat-only' reads. When the repeat length hK is the same or longer than a read length L, repeat-only read are possible if they start in a region with size $hK - L$. The repeat-only reads allow the inference of repeats longer than the read length. Assuming each read can start anywhere in the genome, the expected number of repeat-only reads follows a Poisson distribution.

$$P_U(U|h_1, h_2) = Poisson(U; u(h_1) + u(h_2)) \tag{9}$$

where

$$u(h) = \begin{cases} D(hK - L)/L, & \text{if } hK > L \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

2.5 Paired-end reads

Additional information can be gathered from the group of paired end reads (or sometimes called "mates") that span the STR region. The observed distance between the two mate reads typically follow a distribution $p(V)$ for a specific sequencing library. This distribution can be inferred by compiling the distances between all (or a representative subset of) the paired-end reads across the genome. For alleles without indels in the STR region, the distribution of the observed distances should be distributed identically to $p(V)$ [3]. If there is a homozygous insertion or deletion in the STR region, the distribution of $p(V)$ would shift to $p(V + RK - hK)$. Expanded repeats (or longer h), when mapped onto the reference, show a *compression* of paired-end distances; conversely, shortened repeats (or shorter h) show an *expansion* of paired-end distances.

$$P_V(V|h) = \begin{cases} C \cdot p(V + RK - hK), & \text{if } hK < V + RK \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where C is a normalizing constant to ensure $P_V(V|h)$ sum to 1. Like repeat-only reads, the paired-end distance is also useful to extend the prediction of allele size beyond the length of the sequencing read.

For a diploid STR locus with h_1 and h_2 repeat units, we have a mixed distribution with mixing rate π_V:

$$\pi_V = \frac{v(h_1)}{v(h_1) + v(h_2)} \tag{12}$$

where

$$v(h) = 1 - \sum_{i=1}^{h} p(iK) \tag{13}$$

We then have the mixing distribution for the bp distance between the two ends for a spanning read pair, given that the true STR repeat units for each of the two alleles are h_1 and h_2:

$$P_V(V|h_1, h_2) = \pi_V P_V(V|h_1) + (1 - \pi_V) P_V(V|h_2) \tag{14}$$

The paired-end mode is only enabled when there are at least 5 spanning pairs across the STR locus. With too few observations, the variance of our maximum likelihood estimates based on spanning pairs alone can be substantial.

2.6 Integrated model

The integrated model combines evidence from spanning reads, partial reads, repeat-only reads and spanning pairs in a probabilistic model:

$$\begin{aligned} logP(\{S_{1:l}, T_{1:m}, U, V_{1:n}\}|h_1, h_2) = \\ \sum_{i=1:l} \log P_S(S_i|h_1, h_2) + \sum_{i=1:m} \log P_T(T_i|h_1, h_2) + \\ \log P_U(U|h_1, h_2) + \sum_{i=1:n} \log P_V(V_i|h_1, h_2) \end{aligned} \tag{15}$$

The maximum likelihood estimates are then obtained from the model through a grid search. In TREDPARSE, we set $h_{max} = 300$ which is close to the detection limit of all of our evidence, so the full grid search is at most 300 for haploid and 300×300 for diploid loci. In practice, a full search of the parameter space is not needed, thereby cutting down a substantial amount of compute time.

2.7 Confidence of calls

We combined the evidence in the integrated probabilistic model. We first compute the marginal distribution of $P(h_1|observations)$ and $P(h_2|observations)$, where:

$$P(h_1, h_2|observations) = \frac{P(observations|h_1, h_2)P(h_1, h_2)}{P(observations)} \tag{16}$$

Assuming uniform prior we have:

$$P(h_1, h_2|observations) = \frac{1}{Z} P(observations|h_1, h_2) \tag{17}$$

where Z is the normalizing constant. We then have:

$$\begin{aligned} P(h_1|observations) = \sum_{h_2=1}^{h_{max}} P(h_1, h_2|observations) = \frac{1}{Z} \sum_{h_2=1}^{h_{max}} P(observations|h_1, h_2) \\ P(h_2|observations) = \sum_{h_1=1}^{h_{max}} P(h_1, h_2|observations) = \frac{1}{Z} \sum_{h_1=1}^{h_{max}} P(observations|h_1, h_2) \end{aligned} \tag{18}$$

From these marginal distributions, we can compute the 95% credible intervals (CI) for $\hat{h_1}$ and $\hat{h_2}$. The $100(1-\alpha)\%$ CI of a distribution with parameter θ is defined as:

$$CI_{(100(1-\alpha)\%} = (l, u) : P(l \leq \theta \leq u) = \alpha \tag{19}$$

The 95% CI are not unique on a posterior distribution. In TREDPARSE, we use the 95% CI where there is equal $(1-\alpha)/2 = 2.5\%$ mass on each tail.

We also compute the probability of sample being pathological (PP), given the inheritance model and the cutoff size c based on prior literature. We have:

$$PP = \begin{cases} \frac{1}{Z} \sum_{h_1=c}^{h_{max}} P(observations|h_1, h_2), & \text{if recessive inheritance} \\ \frac{1}{Z} \sum_{h_2=c}^{h_{max}} P(observations|h_1, h_2), & \text{if dominant inheritance} \end{cases} \tag{20}$$

The inheritance model and risk cutoff size c are both important in the calculation of PP. Recessive inheritance requires the shorter allele h_1 to be $\geq$ cutoff size c, while dominant inheritance requires the longer allele h_2 to be $\geq$ cutoff size. For X-linked cases, only one allele need be $\geq$ cutoff size in order to show the disease.

References

1. Telenti, A., Pierce, L. C., Biggs, W. H., di Iulio, J., Wong, E. H., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences pp. 201613365.

2. Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobstr: a short tandem repeat profiler for personal genomes. Genome research *22*, 1154–1162.

3. Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. (2009). Modil: detecting small indels from clone-end sequencing with mixtures of distributions. Nature methods *6*, 473–474.

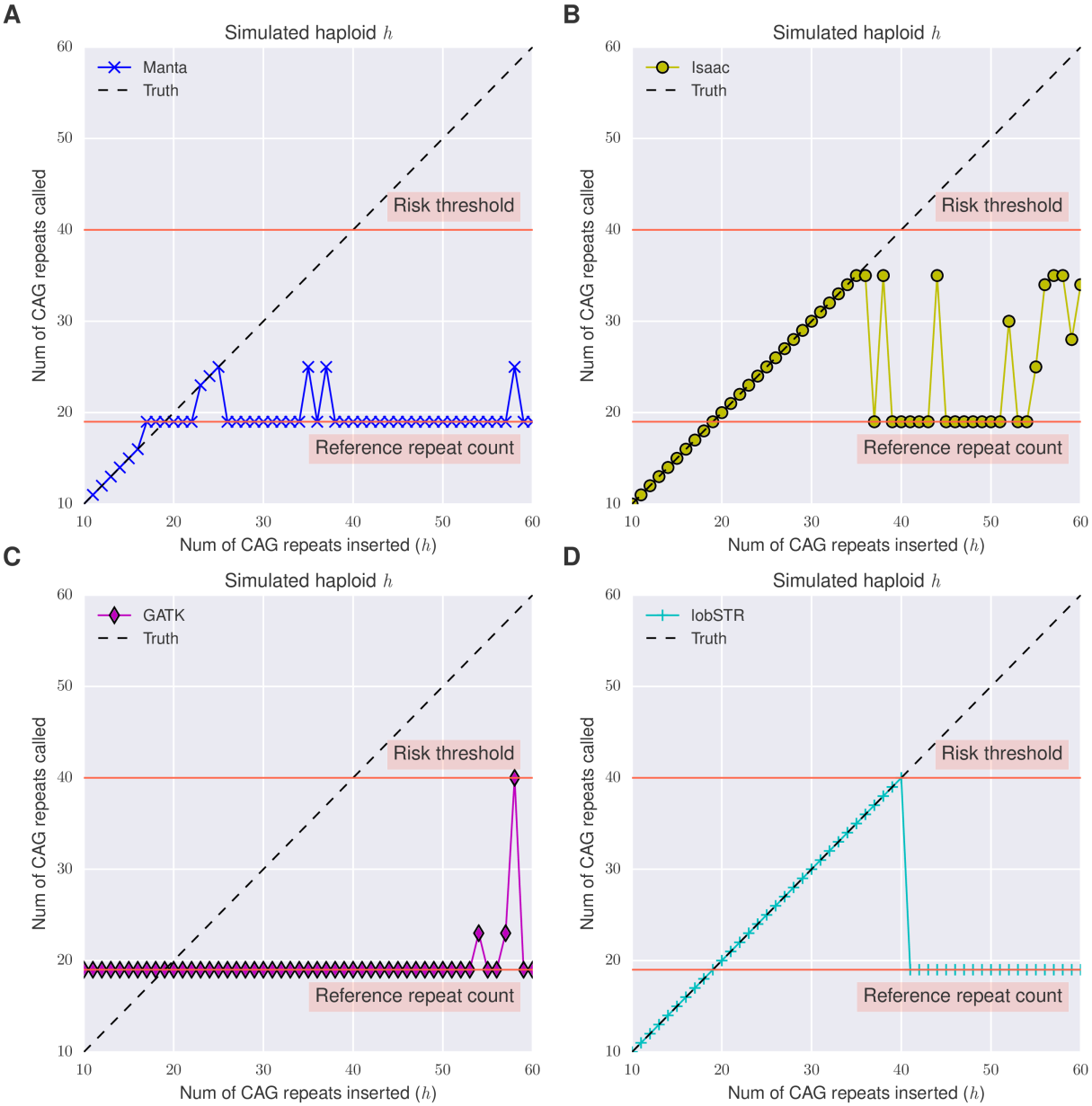

Figure S1: Simulations with synthetic datasets of implanted STR alleles at Huntington (HD) locus. We have tested performance of several variant callers, including (**A**) Manta (**B**) Isaac (**C**) GATK, (**D**) lobSTR.

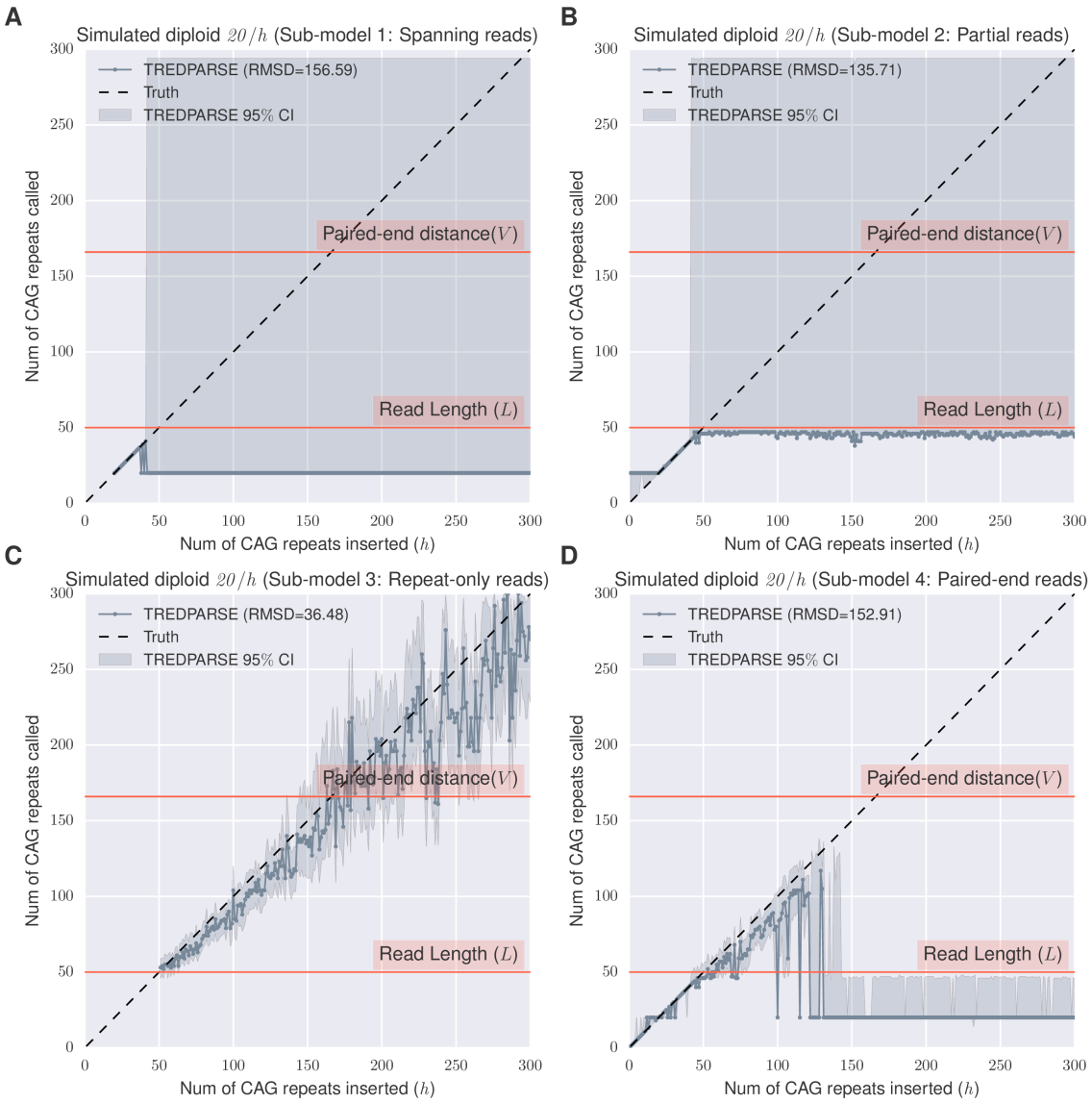

Figure S2: Predictive power of each of the four types of evidence. Each type of evidence has their own specific predictive range. In this simulation, TREDPARSE is run (**A**) using only spanning reads; (**B**) using only partial reads; (**C**) using only repeat reads; (**D**) using only paired-end distance. Shaded region represent 95% credible interval for TREDPARSE estimates of h. $RMSD$ represents root-mean-square deviation, calculated as $RMSD = \frac{1}{N}\sqrt{\sum_{i=1:N}(h_i - \hat{h_i})^2}$.

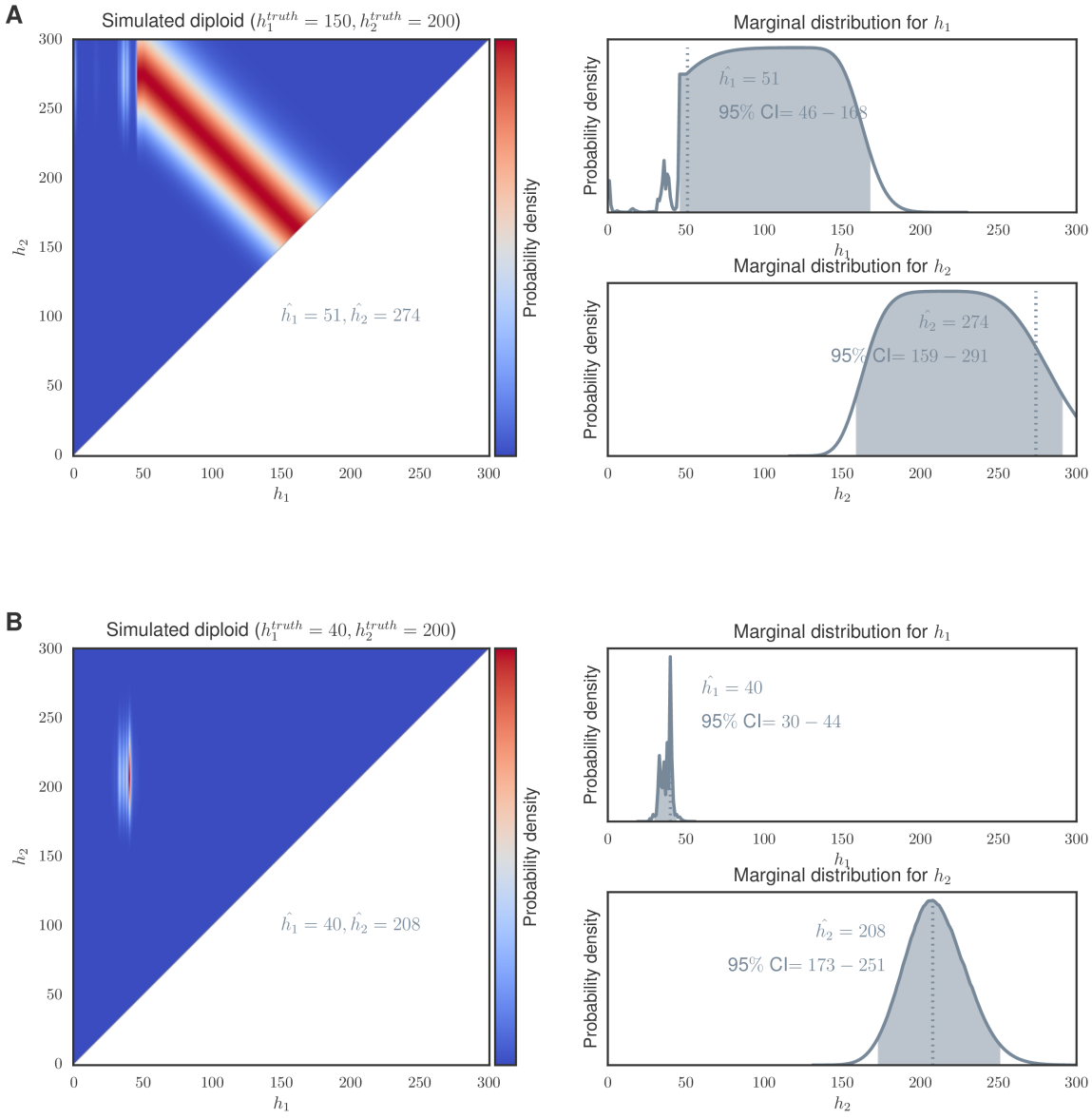

Figure S3: Examples of posterior probability density function based on the integrated model to call STRs. (**A**) Simulated diploid with $h_1 = 150, h_2 = 200$ showing a situation that has both alleles exceeding the paired-end distance with strong negative dependence; (**B**) Simulated diploid with $h_1 = 40, h_2 = 200$ showing a non-smooth marginal distribution showing some uncertainties in the classification of spanning reads, partial reads and repeat-only reads around $h_1 = 40$.

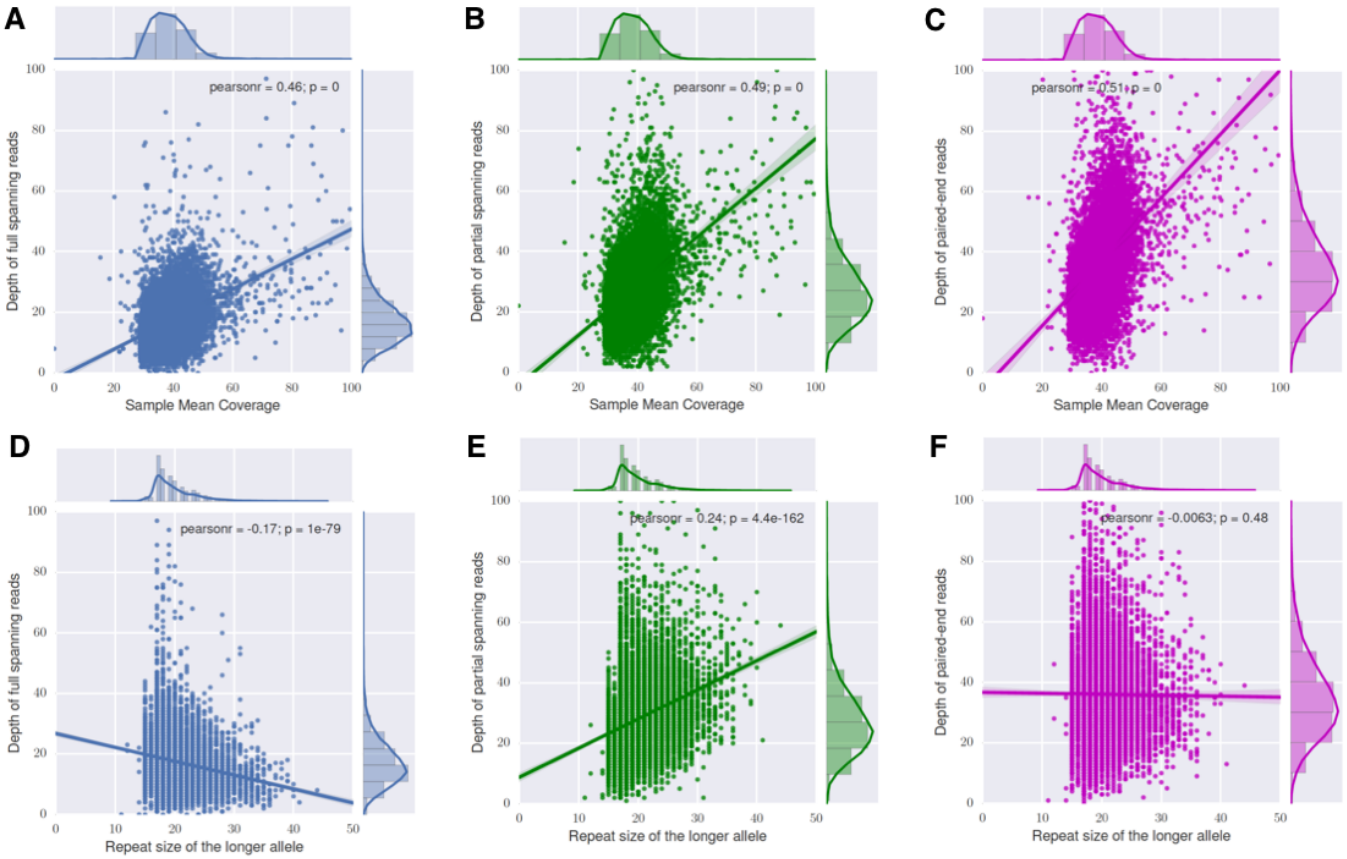

Figure S4: Amount of evidence as a function of sample read depth coverage and length of the repeat allele at Huntington (HD) locus. Based on our HLI samples, relationships are shown between sample mean coverage and depth of spanning reads (**A**), partial reads (**B**), spanning pairs (**C**), as well as the repeat units of the longer allele and read depth of spanning reads (**D**), partial reads (**E**), spanning pairs (**F**).

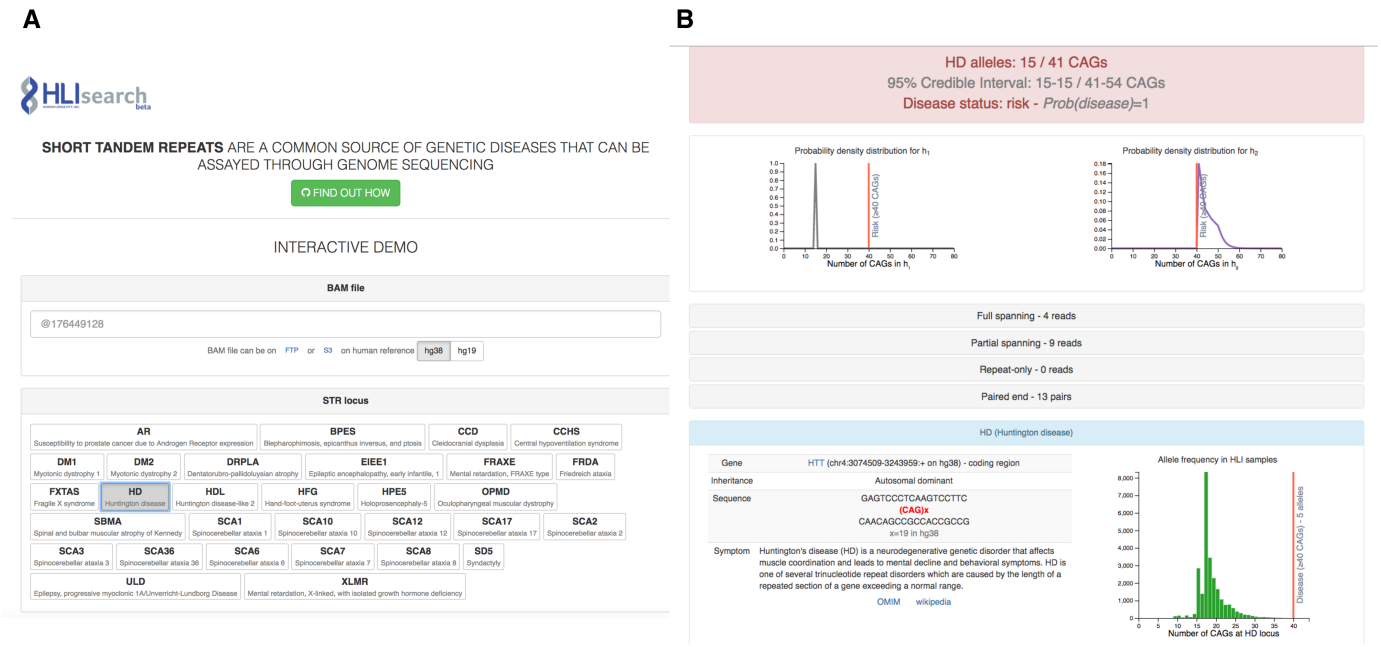

Figure S5: An interactive server for computing the STR calls. (**A**) Input includes the BAM location, reference genome version and the STR locus of interest. (**B**) Output includes detailed information about the STR – including call results, posterior probability density of the size of risk alleles, and various types of reads affecting the final calls.