Genome-wide profiling of heritable and *de novo* STR variations

Thomas Willems^{1,2,*}, Dina Zielinski¹, Assaf Gordon¹, Melissa Gymrek^{3,4} and Yaniv Erlich^{1,5,6,*}

¹ New York Genome Center, New York, NY 10013, USA

² Computational and Systems Biology Program, MIT, Cambridge, MA 02139, USA

³ Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

⁴ Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

⁵ Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY 10027, USA

⁶ Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, USA

* To whom correspondence should be addressed: Y.E. (yaniv@cs.columbia.edu) or T.W. (tfwillems@gmail.com)

Abstract

Short tandem repeats (STRs) are highly variable elements that play a pivotal role in multiple genetic diseases, population genetics applications, and forensic casework. However, STRs have proven problematic to genotype from high-throughput sequencing data. Here, we describe HipSTR, a novel haplotype-based method for robustly genotyping, haplotyping, and phasing STRs from whole genome sequencing data and report a genome-wide analysis and validation of *de novo* STR mutations.

Main Text

The impact of genomics is contingent upon its ability to identify genetic variants. While tremendous progress has been made in identifying nearly every type of genetic variation, short tandem repeat (STR) variations remain largely understudied. Composed of repeating 1-6 base pair motifs, STRs are among the most polymorphic variants in the human genome and are present at over 1 million loci. STRs play a key etiological role in more than 30 Mendelian disorders¹ and recent evidence has underscored their profound regulatory role and potential involvement in complex traits²⁻⁴. Beyond medical genetics, STRs have proven useful for various applications in population genetics, forensics, and single cell lineage analysis. Despite their utility, STRs are typically omitted from large-scale genetic studies. While several STR callers exist⁵⁻⁷, they suffer from limited accuracy, difficulties in calling homopolymer runs, sensitivity to STR stutter noise, and limited functionality compared to mature SNP callers. As a result, large-scale projects⁸⁻¹⁰ are reluctant to report STR genotypes and are essentially blind to many of the most variable parts of the genome.

Here, we developed a novel algorithm called HipSTR (haplotype-inference and phasing for STRs) to create a mature tool for STR studies. HipSTR builds on our extensive experience with STR genotyping and addresses major limitations of existing STR tools^{5,11,12}. To improve accuracy, our tool uses a multitude of inference techniques that integrate additional information about the haplotype in which the STR resides (**Supplementary Fig. 1**). Briefly, HipSTR begins by learning a parametric model that captures the stutter noise profile specific to each locus. Using prior information about the genomic location of the repeat, it then harnesses this model to tune a hidden Markov model (HMM) that carefully realigns the STR-containing reads and mitigates the effects of stutter and sequencing errors (**Supplementary Figs. 1-2**). The realignment framework is highly flexible and can integrate population-scale data from other individuals and phased SNP scaffolds to find the most likely set of alleles, conferring robustness to the genotyping process (**Supplementary Figs. 3-4**). The output of HipSTR is a VCF file that consolidates all variants at an STR locus into a single line. The tool is written in C++ and is freely available at <https://hipstr-tool.github.io/HipSTR>.

We benchmarked HipSTR's accuracy by comparing its STR calls from whole genome sequencing (WGS) data to capillary electrophoresis data, the current gold standard for STR genotyping. To this end, we obtained 118 WGS datasets from the Simons Genome Diversity Project (SGDP)¹³ that were sequenced with an Illumina 100bp pair-end PCR-free protocol to at least 30x coverage. These samples also have capillary electrophoresis calls for 600 highly polymorphic STRs from the Marshfield panel¹⁴, providing a challenging test case for STR callers. The capillary calls showed an internal consistency of 98.8% when comparing STR genotypes of a few duplicated samples, setting an upper bound on the accuracy that could be achieved in our tests. In addition to using HipSTR, we also genotyped the same exact STRs with three general-purpose callers and two STR-specific tools: GATK HaplotypeCaller (GATK-HC)¹⁵, Platypus¹⁶, freebayes¹⁷, lobSTR⁵ and RepeatSeq⁶. We optimized the set of command line options to boost the accuracy of each tool (**Supplementary Table 1**) and developed a machine learning approach to find a combination of quality indicators for each tool to optimally filter STR calls (**Supplementary Table 2**).

Our results show that HipSTR outperformed all other tools (**Figure 1; Supplementary Table 3**). Without any quality-based filtering of the Marshfield panel calls, HipSTR achieved an accuracy of 95.2%, better than any other tested method. In contrast, the second best tool (GATK-HC) achieved an accuracy of 92.0% with default settings and an accuracy of 94.3% after considerable optimization, reporting 20% more erroneous STR genotypes than HipSTR. When filtering the 10% least confident genotypes, HipSTR's accuracy jumped to 98.8% and again substantially outperformed the next best tool.

Figure 1: STR genotype accuracy. We tested the accuracy of each variant caller to genotype STRs from the Marshfield panel. After generating calls for each tool, we ranked the confidence in the STR genotypes using a machine-learning classifier and explored the accuracy as a function of sensitivity (number of STR variants that crossed a certain confidence threshold). In all relevant circumstances, HipSTR's classifier resulted in the most accurate STR genotypes. Solid lines denote tools run using default settings, while dashed lines denote tools run using settings optimized for STR genotyping. The curve for freebayes run without optimization is not shown due to poor performance.

We also evaluated the ability of HipSTR to go beyond reporting length polymorphisms to full STR haplotypes. About half of the STRs in the genome display a repeat structure that includes short interruptions to the recurrent motif¹¹. Thus, two STR alleles that are identical by length can be discordant in sequence due to distinct evolutionary paths¹⁸ (**Supplementary Fig. 5**). Current STR callers and capillary electrophoresis methods only report an STR's length and cannot differentiate between these homoplasic alleles. This complicates population genetics analyses, reduces the ability to understand the effect of STR variations on regulatory or coding changes, and limits the information available for forensic fingerprinting. As HipSTR automatically reports the full sequence of STR variants, not just their length, we sought to test the accuracy of our strategy by running it on the well-characterized CEPH trio in the Illumina Platinum genomes dataset. We observed ~70,700 STR loci that passed our filters in which at least two alleles had the same length but different sequence. Only 304 loci (0.4%) did not follow Mendelian consistency, showing the robustness of our method to report STR sequence variations. Next, for the same trio, we measured the ability of HipSTR to physically phase the Marshfield STR genotypes onto SNP haplotypes. For 178 loci where the algorithm confidently phased the child, we were also able to determine the transmitted paternal and maternal STR alleles. In all 178 instances, HipSTR correctly phased the paternal STR allele onto the paternal SNP scaffold, demonstrating the accuracy of its physical phasing approach. Taken together, our results highlight that HipSTR not only accurately reports STR length polymorphisms but also adds valuable information about the repeat structure and the haplotype context of the variants.

We found that HipSTR is scalable and apt to the analysis of large-scale sequencing data. We ran HipSTR on 2,000 Illumina whole genome sequencing datasets with at least 30x coverage from a New York

To validate these mutations, we re-ran HipSTR with distinct Illumina datasets generated for these samples and compared the variants between the runs. Notably, we were able to replicate 358 (85%) of the *de novo* mutations as all samples' allele lengths matched perfectly between the two call sets. As expected, most *de novo* mutations predominantly occur at homopolymer repeats (293/358=81%) and to a much lesser extent at repeats with 2-6 base pair repeat units. To further validate our results, we genotyped high confidence *de novo* STR loci with capillary electrophoresis. As these STRs are highly mutable and prone to stutter noise, even Sanger sequencing does not deliver unequivocal genotypes for these loci. To gain confidence, we TOPO cloned the STR alleles from each member of the trio and Sanger sequenced at least eight independent clones per individual. Using this technique, we obtained high confidence Sanger calls for four STRs where HipSTR identified a *de novo* mutation (**Figure 2B, Supplementary Figs. 6-9**). In all cases, the Sanger calls for the parents confirmed the parental genotypes reported by HipSTR and the Sanger calls for the child confirmed the *de novo* allele reported by HipSTR, validating our method.

Next, we sought to distinguish *de novo* STR variants that arose in the germline of the trio's child (NA12878) from mutations that arose during the many cell-line passages of this sample. We used HipSTR to analyze the WGS data of her 11 offspring and inspected the genotypes for the 358 STRs where NA12878 had a replicable *de novo* mutation. For 31 loci, we found evidence that the *de novo* STR allele was transmitted to at least three offspring and was absent from the paternal genotype. For an additional 32 loci, the *de novo* allele of the mother was observed in at least three of her offspring, but the same (identical-by-state) allele also resided in her husband (NA12877) and the transmission could not be fully resolved. As our genome-wide scan identified at least 31 *de novo* STR mutations in the germline of NA12878 after examining ~35% of STRs in her genome, we predict a lower bound of over 100 *de novo* STR mutations per generation. Previous studies of *de novo* SNP rates quantified approximately 70 mutations per generation^{8,19}. Our analysis strongly suggests that the load of STR *de novo* variations is at least of the same order of magnitude, demonstrating their substantial source of genetic variability.

To summarize, our results show that HipSTR offers several advantages for STR calling. First, the technique is considerably more accurate than other general purpose and STR-specific variant callers and does so with exceptional computational tractability. Second, the tool has new capabilities such as phasing, haplotyping, and resolving STR homoplasy that are absent from other tools and can be important for population genetic analyses, forensic work, and STR association studies. Finally, our method enables highly specific detection of *de novo* STR variants. This new capability could open exciting avenues for understanding human genetic variation and the role of STRs in various diseases and complex traits.

Methods

The HipSTR algorithm

Modeling PCR stutter

PCR stutter artifacts add or remove copies of an STR's motif to sequencing reads, resulting in observed STRs that differ in size from the true underlying genotype. To mitigate these effects, we used a model that we developed and extensively validated in our previous work to discern between stutter noise and STR mutations on the Y chromosome¹². HipSTR constructs a stutter model θ_x for each STR locus x , which contains the probability that stutter adds (u) or removes (d) repeats from the true allele in an observed read, and a geometric distribution with parameter ρ_s that controls the size of the stutter-induced changes. In our framework, the probability of observing a stutter artifact δ repeat units in size is:

$$P(\text{stutter} = \delta | \theta_x) = \begin{cases} 1 - u - d, & \delta = 0 \\ u\rho_s(1 - \rho_s)^{\delta-1} & \delta > 0 \\ d\rho_s(1 - \rho_s)^{-\delta-1} & \delta < 0 \end{cases}$$

To estimate each locus' stutter model parameters, we extract the size of the STR observed in each read for all individuals in the population. We then use an Expectation-Maximization approach²⁰ to learn the parameters. The E-step computes each sample's genotype posteriors:

$$P(g_i = (j, k) | R, \theta_x^t) \propto f_j f_k \prod_{m=1}^{n_{reads,i}} \sum_{a \in j, k} \begin{cases} 1 - u - d, & r_{m,i} = r_a \\ u\rho_s(1 - \rho_s)^{r_{m,i}-r_a-1} & r_{m,i} > r_a \\ d\rho_s(1 - \rho_s)^{r_a-r_{m,i}-1} & r_{m,i} < r_a \end{cases}$$

Here, R denotes the set of all reads, g_i denotes the phased genotype for the i^{th} individual, $n_{reads,i}$ denotes the number of reads for the i^{th} individual, $r_{m,i}$ denotes the number of repeats in the m^{th} read for the i^{th} individual, r_a denotes the number of repeats in the a^{th} allele and f_j denotes the frequency of the j^{th} allele. For each possible phased genotype, the E-step also computes the conditional probability that a read originated from either allele:

$$P(s_{m,i} = j | g_i = (j, k), \theta_x^t) \propto \begin{cases} 1 - u - d, & r_{m,i} = r_j \\ u\rho_s(1 - \rho_s)^{r_{m,i}-r_j-1} & r_{m,i} > r_j \\ d\rho_s(1 - \rho_s)^{r_j-r_{m,i}-1} & r_{m,i} < r_j \end{cases}$$

Given N samples, A alleles and Q reads, the M-step then updates the stutter model parameters and allele frequencies using these probabilities:

$$\begin{aligned} u^{t+1} &= \frac{1}{Q} \sum_{i=1}^N \sum_{j=1}^A \sum_{k=1}^A P(g_i = (j, k) | R, \theta_x^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j, k} P(s_{m,i} = a | g_i = (j, k), \theta_x^t) I(r_{m,i} > r_a) \\ d^{t+1} &= \frac{1}{Q} \sum_{i=1}^N \sum_{j=1}^A \sum_{k=1}^A P(g_i = (j, k) | R, \theta_x^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j, k} P(s_{m,i} = a | g_i = (j, k), \theta_x^t) I(r_{m,i} < r_a) \\ \rho_s^{t+1} &= \frac{\sum_{i=1}^N \sum_{j=1}^A \sum_{k=1}^A P(g_i = (j, k) | R, \theta_x^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j, k} P(s_{m,i} = a | g_i = (j, k), \theta_x^t) I(r_{m,i} \neq r_a)}{\sum_{i=1}^N \sum_{j=1}^A \sum_{k=1}^A P(g_i = (j, k) | R, \theta_x^t) \sum_{m=1}^{n_{reads,i}} \sum_{a \in j, k} P(s_{m,i} = a | g_i = (j, k), \theta_x^t) |r_{m,i} - r_a|} \\ f_j^{t+1} &= \frac{1}{2N} \sum_{i=1}^N \sum_{k=1}^A P(g_i = (j, k) | R, \theta_x^t) + P(g_i = (k, j) | R, \theta_x^t) \end{aligned}$$

Intuitively, the update rules for the stutter probabilities u and d compute the fraction of times a read's STR allele is either larger or smaller than its underlying allele. The update rule for the step size parameter ρ_s is more involved, but it first restricts the computation to reads with non-zero stutter. It then computes the inverse of the mean weighted step size, consistent with a maximum likelihood estimator for a geometric distribution.

Generating candidate alleles

To identify an initial set of STR alleles, HipSTR selects reads that fully span the STR. It requires that both ends of a read match the reference genome for at least 10bp and that neither end of the read has a longer exact match with the reference genome 15bp upstream or downstream from its alignment. Based on this subset of reads, HipSTR includes a sequence as a candidate allele if it is present in two or more and at least 20% of a sample's reads.

HipSTR also uses an iterative approach to identify new candidate alleles. At the end of every round of genotyping, it computes the maximum-likelihood genotype for each sample and realigns every read relative to the most probable allele in its sample's genotype. Each of these alignments generates a sequence in the STR region. If the same sequence is observed in a sample in two or more alignments with stutter artifacts, HipSTR selects the sequence as a new candidate allele.

Computing genotype likelihoods

The genotype likelihood model integrates information about every read's phasing likelihood and alignment likelihood. For the m^{th} read for individual i , $P(p_{m,i}|h=1)$ and $P(p_{m,i}|h=2)$ denote the phasing likelihoods of the read originating from the first and second SNP haplotypes, while $P(s_{m,i}|a=j)$ denotes the alignment likelihood of the read to the j^{th} allele. We use a uniform prior for each unphased genotype, such that heterozygous phased genotypes have half the prior probability of their homozygous counterparts. The likelihoods for the phased genotypes for the i^{th} sample are:

$$P(g_i = (j, j)) \propto 2 \prod_{m=1}^{n_{\text{reads},i}} [P(p_{m,i}|h=1) + P(p_{m,i}|h=2)] P(s_{m,i}|a=j)$$

$$P(g_i = (j, k)) \propto \prod_{m=1}^{n_{\text{reads},i}} P(p_{m,i}|h=1) P(s_{m,i}|a=j) + P(p_{m,i}|h=2) P(s_{m,i}|a=k)$$

Read phasing likelihoods

To compute the phasing likelihoods for each read, HipSTR examines bases in the read or its mate pair that are aligned to heterozygous SNPs. If the read originated from a haplotype, the likelihood of the base b_i matching the SNP base h_j is given by the base quality q_{b_i} while the likelihood of it not matching is one third of the residual probability. We express this as:

$$Q(b_i, h_j) = \begin{cases} q_{b_i}, & b_i = h_j \\ \frac{1 - q_{b_i}}{3}, & b_i \neq h_j \end{cases}$$

We compute each read's total phasing likelihood by multiplying $Q(b_i, h_j)$ for every base b_i aligned to a heterozygous haplotype SNP h_j in the read or its mate pair. In practice, SNP calls in and around STR regions are likely to be error-prone. We therefore exclude SNPs that are within 15 base pairs of the STR region when computing the phasing likelihoods.

Read alignment likelihoods

HipSTR assumes that each haplotype is composed of two distinct types of regions: flanking sequences and STR sequences. As the sources of error prevalent in these two types of regions differ dramatically, HipSTR uses a distinct model to align sequences to each type of region. It then combines the likelihoods from these two different models at the junctions of these regions. In particular, it requires that the read match the flanking sequence at the first base preceding the STR sequence and at the first base following an STR sequence.

Aligning reads to flanking sequences

Within flanking sequences, the alignment model accounts for Illumina sequencing errors using a previously described hidden Markov model²¹. To efficiently align reads in these regions, we use three matrices to recursively compute the maximum log-likelihood of aligning read bases $b_1 \dots b_i$ with haplotype bases $h_1 \dots h_j$. The matrices *Match* and *Ins* are used to track the likelihoods that read base b_i is aligned to haplotype base h_j or an insertion following base h_j , respectively. Matrix *Del* tracks the maximum log-likelihood that base b_i is followed by one or more deletions. In conjunction with values for $t_{X \rightarrow Y}$, the log-probability of transitioning from hidden state X to hidden state Y , we use the following recursions to fill in each matrix column-by-column:

$$Match(i, j) = Q(b_i, h_j) + \max \begin{cases} Match(i-1, j-1) & + t_{Match \rightarrow Match} \\ Del(i-1, j-1) & + t_{Match \rightarrow Del} \\ Ins(i-1, j-1) & + t_{Match \rightarrow Ins} \end{cases}$$

$$Ins(i, j) = Q(b_i, b_i) + \max \begin{cases} Ins(i-1, j) & + t_{Ins \rightarrow Ins} \\ Match(i-1, j) & + t_{Ins \rightarrow Match} \end{cases}$$

$$Del(i, j) = \max \begin{cases} Match(i, j-1) & + t_{Del \rightarrow Match} \\ Del(i, j-1) & + t_{Del \rightarrow Del} \end{cases}$$

Aligning reads to STR sequences

In STR regions, HipSTR utilizes an entirely different alignment model to account for STR-specific errors. As PCR stutter artifacts are prevalent in this domain, it assumes that a read's sequence differs from the underlying haplotype by at most one indel whose magnitude D is a multiple of the repeat unit length. If no stutter artifact has occurred, the likelihood of the observed sequence is governed by the agreement between each base in the read and its corresponding haplotype base. The probability of no stutter artifact and aligning base b_i and its preceding bases to an STR sequence $h_1 \dots h_L$ of length L is:

$$P(b_{\max(1, i-L+1)} \dots b_i, D=0 | h_1 \dots h_L) = (1-u-d) \prod_{k=0}^{\min(L-1, i-1)} Q(b_{i-k}, h_{L-k})$$

If a stutter deletion occurs, we assume that it can arise anywhere within the STR region. We iterate over these configurations, each of which has a likelihood given by the agreement between the sequenced bases and their corresponding haplotype bases:

$$P(b_{\max(1, i-L+D+1)} \dots b_i, D | h_1 \dots h_L) = \frac{d\rho_s(1-\rho_s)^{D-1}}{L-D+1} \sum_{d=0}^{L-D} \prod_{k=0}^{\min(d-1, i-1)} Q(b_{i-k}, h_{L-k}) \prod_{k=d}^{\min(L-D-1, i-1)} Q(b_{i-k}, h_{L-D-k})$$

Finally, if a stutter insertion occurs, we assume that it can precede any base in the STR region. As PCR stutter insertions typically contain sequences that copy the local repeat structure, we assume that inserted

sequences are periodic copies of the STR sequence directly preceding the insertion. We therefore measure the likelihood of inserted bases according to their agreement with this sequence. For an STR with a repeat motif of length M , iterating over each possible insertion position results in the likelihood:

$$P(b_{\max(1, i-L+D+1)} \dots b_i, D | h_1 \dots h_L) = \frac{u\rho_s(1-\rho_s)^{D-1}}{L+1} \sum_{d=0}^L \prod_{k=0}^{\min(d-1, i-1)} Q(b_{i-k}, h_{L-k}) \prod_{k=d+D}^{\min(L+D-1, i-1)} Q(b_{i-k}, h_{L+D-k})$$

$$\prod_{k=d}^{\min(d+D-1, i-1)} \begin{cases} Q(b_{i-k}, h_{L-d-((k-d) \bmod M)}), & L-d-((k-d) \bmod M) \geq 1 \\ Q(b_{i-k}, h_{M+L-d-((k-d) \bmod M)}), & \text{otherwise} \end{cases}$$

Experiments

Constructing a gold standard STR dataset

We downloaded capillary genotypes for the 628 STRs of the Marshfield panel from <https://web.stanford.edu/group/rosenberglab/data/rosenbergEtAl2005/combinedmicrosats-1048.stru> and other information from https://web.stanford.edu/group/rosenberglab/data/pembertonEtAl2009/Pemberton_AdditionalFile1_1124_2009.txt. These datasets were generated in a prior study¹⁴. Using the isPCR²² tool, we mapped each STR's primers to the hg19 reference genome. Together with available repeat structure information, we used Tandem Repeats Finder²³ to scan between the primer sites and pinpoint the exact genomic coordinates of each STR.

Capillary STR genotypes report the lengths of amplified DNA fragments containing the STR region. These lengths correctly capture STR variations but also capture indels outside of STRs if they fall within the amplified regions. To mitigate the effects of these indels, we downloaded assembly-based FermiKit²⁴ calls for each of the 263 SGDP samples from a publicly available repository (<https://github.com/lh3/sgdp-fermi>). For each STR, we identified samples with indels located in between the two primer sites. We then masked the genotype for a sample if any of its indels occurred more than 15bp upstream or downstream from the STR region, as these are unlikely to originate from the STR itself.

Assessing variant caller performance

We downloaded BWA-MEM alignments²⁵ available for 263 publicly available SGDP samples (accession ERP010710). Using a BED file containing each STR's genomic coordinates, we ran every tool with default options. We then explored the effects of alternate command line options by rerunning each tool using every combination of the settings listed in **Supplementary Table 1**. The combination of settings that led to the highest level of agreement with the capillary genotypes was then selected as optimal. Despite multiple attempts, we were unable to run STR-FM⁷ outside of the Galaxy environment and therefore could not include this tool in our benchmarking experiments.

Most general-purpose variant callers produce multiple unphased variants per STR region. Without phase information, it is frequently not possible to determine the lengths of an individual's two STR alleles present at a given locus. To overcome this issue, we summed the sizes of all indels each caller produced within an STR region for a given sample. The frequency with which these sums exactly matched those predicted by the capillary electrophoresis data was then used as the variant caller's accuracy.

Phased trio SNP scaffolds

We used the HaplotypeCaller module in GATK v3.5.0-g36282e4 to jointly genotype all members of the

trio using the aligned and sorted BAMs for runs ERR194147, ERR194160 and ERR194161 generated as described above. In accordance with guidelines for hard filtering SNP calls, we used GATK's SelectVariants and VariantFiltration modules to select only those SNPs with a passing FILTER, $QD > 2$, $FS < 60$, $MQ > 40$, $MQRankSum > -12.5$ and $ReadPosRankSum > -8$. Next, we downloaded v5a of Beagle's²⁶ reference panels for Phase 3 of the 1000 Genomes Project from the tool's website and removed three samples that are part of the CEPH pedigree (NA12878, NA12889 and NA12890). Using v4.0-r1399 of Beagle and these filtered reference panels, we phased the trio's SNP calls using the *ped* option and *phase-its=10*, *burnin-its=10*, *impute-its=0* and *impute=false*.

Generating CEPH trio STR genotypes for the Marshfield markers

We downloaded FASTQ files from the Illumina Platinum Genomes Project and an additional study of the effects of PCR amplification on sequencing errors²⁷, resulting in data from twelve different sequencing runs (NA12878: ERR194147, SRR826463, SRR826467, SRR826469; NA12891: ERR194160, SRR826427, SRR826448, SRR826465; NA12892: ERR194161, SRR826428, SRR826473, SRR826471). We individually aligned each run to the hg19 reference genome using the BWA-MEM algorithm from bwa v0.7.12-r1039. Using the resulting BAMs for all samples concurrently, we ran HipSTRv0.2 with the options *--def-stutter-model --use-all-reads --min-reads 25* and *--read-qual-trim #*. To generate STR genotypes that are phased onto SNP scaffolds, we reran HipSTR using the same arguments but also specified the *--snp-vcf* option and used the VCF containing phased SNP scaffolds as input.

Evaluating HipSTR's physical phasing accuracy

We began by restricting our analysis to Marshfield markers in which the STR alleles transmitted from each parent to the child could be determined from the unphased genotypes alone. We then required that the child have a confidently phased STR genotype as indicated by a FORMAT field with PQ > 0.9 (minimum phased genotype posterior of 90%). Using the SNPs 50 kilobases upstream and 50 kilobases downstream of the STR region, we determined the SNP haplotype each parent transmitted to the child by requiring that each of the child's SNP haplotypes exactly match one parental haplotype and that these matches involve both the maternal and paternal haplotypes. After enforcing each of these requirements, 178 markers were available for downstream analyses. For each of these markers, HipSTR's phased genotypes correctly placed the paternally transmitted STR allele onto the paternally transmitted SNP haplotype, resulting in perfect phasing accuracy.

Running HipSTR on a population-scale dataset

WGS data for 2000 individuals sequenced using 150bp paired-end Illumina reads and more than 30x coverage was internally available at the New York Genome Center. We applied HipSTR to genotype STRs in this dataset. Using the BWA-MEM aligned BAMs for each sample as input, we jointly genotyped 200 samples at a time with HipSTR and all default options. We aggregated the timing statistics across each of the individual runs to determine the total run time of ~20,000 CPU hours. NYGC's IRB committee approved all human subject experiments prior to this analysis.

Using HipSTR to identify *de novo* mutations

We downloaded FASTQs from the Illumina Platinum Genomes project containing 200x sequencing data for NA12877 and NA12878 (runs ERR174310-ERR174341). As before, we used BWA-MEM to align the reads in each of these runs individually. Using all of these BAMs and two previously generated BAMs for NA12891 and NA12892 (ERR194160 and ERR194161), we ran HipSTR with the options *--def-stutter-model --require-pairs, --min-reads 25* and a BED file containing 1.6 million STR regions.

We applied a series of stringent filters to the call set to reduce the likelihood that genotyping errors introduce false positive *de novo* calls. Using the FORMAT fields available in the HipSTR VCF, we required that all three individuals in the trio have a minimum genotype posterior (Q) of 0.9, no more than 10% of reads containing either a stutter artifact or a flanking sequence indel (DSTUTTER/DP and

DFLANKINDEL/DP) and at least 10 reads spanning the STR region (MALLREADS). Lastly, we required that the ratio of spanning reads supporting the alleles for each individual be at least 20% (MALLREADS).

To identify an initial set of candidate mutations, we examined STRs that satisfied all of these requirements. For 423 loci where the child (NA12878) carried an allele length not observed in either of the parents (NA12891 and NA12892), we identified a potential *de novo* mutation.

Validating *de novo* mutations using orthogonal datasets

We downloaded FASTQs containing 300x Illumina sequencing data for NA12878 from the Genome in a Bottle consortium and FASTQs for NA12878, NA12891 and NA12892 from the 1000 Genomes Project (SRR622457, SRR622458 and SRR622459, respectively). After generating BAM files using BWA-MEM, we collated these alignments with others generated in previous analyses (NA12878: SRR826463, SRR826467, SRR826469; NA12891: SRR826427, SRR826448, SRR826465; NA12892: SRR826428, SRR826471, SRR826473). We then reran HipSTR using these BAMs and the options *--def-stutter-model --require-pairs --min-reads 25* and a BED file containing the 423 STR regions with previously detected *de novo* mutations.

Without performing any filtering, we compared the HipSTR calls from these datasets to the calls generated during the discovery phase. For 358 of the markers, each member of the trio had allele lengths that matched perfectly between the two call sets, resulting in a set of sites with high confidence *de novo* mutations. The genotypes for these STRs with high confidence *de novo* mutation are available as a VCF at <https://github.com/HipSTR-Tool/HipSTR-paper>.

Sanger sequencing validation

Primers were designed around the STR coordinates to generate 300-600 bp PCR products (**Supplementary Table 4**). All coordinates are according to GRCh37/hg19. Primers were tested using *in silico* PCR for unique products. Genomic DNA for NA12878, NA12891, and NA12892 was obtained from the Coriell Institute (Camden, NJ, USA). DNA was amplified for 30 cycles in 25 ul reactions according to the manufacturer's recommended cycling conditions using Q5 High Fidelity Polymerase (NEB catalog #M0494) to reduce stutter, generating blunt end products. Amplicons were purified on magnetic beads (Thermo Fisher Scientific ChargeSwitch PCR Clean-Up Kit, catalog #CS12000) and cloned into linearized pMiniT (NEB catalog #E1202). Plasmids were transformed into 50ul of chemically competent *E. coli* (Lucigen *E. coli* Chemically Competent Cells catalog #60108). Outgrowth cultures (50ul) were incubated overnight on ampicillin plates. Individual colonies were selected and cultured overnight in 2mL LB + ampicillin (100ug/mL). DNA was extracted and column purified (Thermo Fisher Scientific PureLink Quick Plasmid Miniprep Kit, catalog #K210010). Sanger sequencing of at least 8 clones per individual per locus was performed by Eton Bioscience (Newark, NJ, USA) using the supplied primers for the pMiniT plasmid. Only results with flanking sequences upstream and downstream of the STR of sufficient quality were included in the final counts.

Assessing *de novo* transmission to children

We downloaded FASTQs containing 50x Illumina sequencing data for each of the 11 children of NA12878 (ERR194148, ERR218433, ERR324432, ERR324433, ERR194152, ERR324434, ERR194154, ERR194155, ERR324435, ERR194157 and ERR194162). Using these BAMs as input, we ran HipSTR with the options *--def-stutter-model --require-pairs --min-reads 25*.

Coordinates

All reported coordinates are based on the hg19 genome build.

Acknowledgments

Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by NIH grant 2014-DN-BX-K089 (T.W., D.Z, A.G., M.G., Y.E.) and a generous gift by Andria and Paul Heafy.

References

1. Mirkin, S.M. Expandable DNA repeats and human disease. *Nature* **447**, 932-940 (2007).
2. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics* (2015).
3. Contente, A., Dittmer, A., Koch, M.C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature genetics* **30**, 315-320 (2002).
4. Hefferon, T.W., Groman, J.D., Yurk, C.E. & Cutting, G.R. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3504-3509 (2004).
5. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome research* **22**, 1154-1162 (2012).
6. Highnam, G. et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic acids research* **41**, e32 (2013).
7. Functamman, A. et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome research* **25**, 736-749 (2015).
8. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).
9. Zook, J.M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* **32**, 246-251 (2014).
10. Genomes Project, C. et al. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
11. Willems, T. et al. The landscape of human STR variation. *Genome research* **24**, 1894-1904 (2014).
12. Willems, T. et al. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *American journal of human genetics* **98**, 919-933 (2016).
13. Mallick, S. et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature advance online publication* (2016).
14. Rosenberg, N.A. et al. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS genetics* **1**, e70 (2005).
15. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).
16. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46**, 912-918 (2014).
17. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* (2012).
18. Estoup, A., Jarne, P. & Cornuet, J.M. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular ecology* **11**, 1591-1604 (2002).
19. Francioli, L.C. et al. Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics* **47**, 822-826 (2015).
20. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38 (1977).

21. Albers, C.A. et al. Dindel: accurate indel calls from short-read data. *Genome research* **21**, 961-973 (2011).
22. Hinrichs, A.S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**, D590-598 (2006).
23. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573-580 (1999).
24. Li, H. FermiKit: assembly-based variant calling for Illumina resequencing data. *Bioinformatics* **31**, 3694-3696 (2015).
25. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
26. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics* **81**, 1084-1097 (2007).
27. Weisenfeld, N.I. et al. Comprehensive variation discovery in single human genomes. *Nat Genet* **46**, 1350-1355 (2014).

Supplementary Figure 1

Supplementary Figure 1: Overview of the HipSTR algorithm. In step 1, an expectation-maximization algorithm learns the PCR stutter model for the locus of interest in order to account for the frequency of these errors during genotyping. Step 2 utilizes well-anchored alignments that span the STR to identify candidate alleles in the STR region and builds haplotypes using these alleles and the sequence upstream and downstream of the STR. In step 3, the PCR stutter model and an HMM are used to align every read to every candidate haplotype. Step 4 analyzes reads that overlap heterozygous SNPs to determine the likelihood that each read came from either strand of the phased SNP haplotypes. Finally, step 5 combines these two sets of likelihoods to determine each sample's maximum likelihood genotype. Every read is realigned relative to its sample's optimal genotype and if new reads span the STR, HipSTR returns to step 2 to identify new candidate alleles and repeat the process. This iterative procedure continues until no new candidate alleles are identified, at which point the maximum likelihood genotypes are output to a VCF file.

Supplementary Figure 2

Flanking region

Method = Illumina error model + Pairwise HMM

STR region

Method = Locus stutter model + Single indel of size $D \in \dots -2L, -L, 0, L, 2L, \dots$

$D = 0$ (No indel in read)

$P(\text{no indel})$

$D > 0$ (Insertion in read)

$D < 0$ (Deletion in read)

Supplementary Figure 2: HipSTR alignment models. HipSTR uses two distinct types of alignment models to obtain robust alignments likelihoods. In regions flanking the STR (top), a pairwise hidden Markov model is used to accounts for Illumina sequencing errors. In contrast, within STR regions (bottom), HipSTR assumes that the main source of artifacts are stutter errors. If no stutter error occurs, the likelihood of observing a sequence of characters is given by the agreement between the bases in the read (blue) and the corresponding bases on the haplotype (green). Otherwise, HipSTR assumes that a single stutter indel occurs (red) and that it arose at each position with equal probability. The alignment likelihood for these scenarios are then obtained by marginalizing over all configurations.

Supplementary Figure 3

Supplementary Figure 3: Physically phasing STRs onto SNP scaffolds. When provided with phased SNP haplotypes, HipSTR uses reads that overlap heterozygous SNPs to phase the STR genotypes (blue) onto the SNP haplotypes (red). The schematic provides a conceptual outline of how this works for an AC repeat flanked by two heterozygous SNPs with known phase.

Supplementary Figure 4

Reference
genome

Input alignments

After HipSTR

Supplementary Figure 4: HipSTR in action. Example alignments for NA12878 to a particular Marshfield STR before (top) and after (bottom) processing by HipSTR. While the input alignments are riddled with local alignment errors, HipSTR's algorithm resolves the alignments into two parsimonious STR insertions of 4 and 16 base pairs.

Supplementary Figure 5

Supplementary Figure 6: High-confidence de novo mutation detected by HipSTR.

The figure depicts an STR located at chr20:45121567 in the hg19 reference holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Each subsequent row depicts a read's maximum likelihood alignment after processing with HipSTR for the child (NA12878), father (NA12891) and mother (NA12892) in the trio. While the alignments strongly support that both the father and mother match the reference allele, the alignments for NA12878 support the reference allele and a 1bp deletion. Only a subset of the alignments for NA12878 (12.5%), NA12891 (50%) and NA12892 (50%) are displayed to facilitate visualization.

Supplementary Figure 7: High-confidence de novo mutation detected by HipSTR.

The figure depicts an STR located at chr19:37201975 in the hg19 reference genome with a TG repeat (top row). Each subsequent row depicts a read's maximum likelihood alignment after processing with HipSTR for the child (NA12878), father (NA12891) and mother (NA12892) in the trio. While the alignments strongly support that both the father and mother match the reference allele, the alignments for NA12878 support both the reference allele and a 2bp insertion. Only a subset of the alignments for NA12878 (12.5%) and NA12892 (50%) are displayed to facilitate visualization.

The figure depicts an STR located at chr18:27043336 in the hg19 reference genome with an ATC repeat (top row). Each subsequent row depicts a read's maximum likelihood alignment after processing with HipSTR for the child (NA12878), father (NA12891) and mother (NA12892) in the trio. While the alignments strongly support a 0 bp/3bp heterozygous insertion for the father and a 3bp homozygous insertion for the mother, the alignments for NA12878 support a heterozygous 3bp/6bp insertion. Only a subset of the alignments for NA12878 (12.5%), NA12891 (50%) and NA12892 (50%) are displayed to facilitate visualization.

Supplementary Figure 9

Supplementary Figure 9: High-confidence de novo mutation detected by HipSTR. The figure depicts an STR located at chr1:221172558 in the hg19 reference genome with an ATCT repeat (top row). Each subsequent row depicts a read's maximum likelihood alignment after processing with HipSTR for the child (NA12878), father (NA12891) and mother (NA12892) in the trio. While the alignments strongly support an 8bp homozygous deletion for both parents, the alignments for NA12878 support both a 4bp and 8bp deletion. Only 25% of the alignments for NA12878 are displayed to facilitate visualization.

Supplementary Tables

Supplementary Table 1: Optimal settings for STR genotyping

Method	Version	Explored Options	Optimal Options
freebayes	1.0.2-6 (g3ce827d)	genotype-qualities no-partial-observations report-genotype-likelihood-max	genotype-qualities no-partial-observations
GATK-HC	3.5-0 (g36282e)	max_alternate_alleles pcr_indel_model kmerSize	max_alternate_alleles 25
HipSTR	0.2	None	None
lobSTR	4.0.0	filter-mapq0 filter-clipped max-repeats-in-ends min-read-end-match	None
Platypus	0.8.1	maxVariants minVarFreq	maxVariants 0.25 minVarFreq 0.0
RepeatSeq	0.8.2	None	None

Supplementary Table 2: Features used to build STR genotype classifiers

Method	Feature Name	Source ¹	Description
freebayes	GQ	F	Genotype quality
	DP	F	Genotype read depth
	GLDIFF	F	Difference between genotype likelihoods of the called and next best genotypes
	READRATIO ²	F	Minimum ratio of read depths for called alleles
	UNCALLEDFRAC ²	F	Fraction of reads observed for an uncalled allele
GATK-HC	QD	I	Ratio of variant quality to read depth
	FS	I	P-value for strand bias
	ReadPosRankSum	I	Z-score for test of alt vs. ref read position bias
	GQ	F	Genotype quality
	DP	F	Genotype read depth
HipSTR	Q	F	Genotyper posterior probability
	SPANDP	F	Minimum number of reads spanning either allele
	DSTUTTER/DP	F	Fraction of reads with a stutter artifact
	DFLANKINDEL/DP	F	Fraction of reads with an indel flanking the STR
lobSTR	SB	F	P-value for strand bias
	Q	F	Genotype posterior probability
	DP	F	Genotype read depth
	DISTENDS	F	Average difference between distance of STR to read ends
Platypus	QD	I	Ratio of variant quality to read depth
	SbPval	I	P-value for strand bias
	GQ	F	Genotype quality
	GOD	F	Genotype goodness of fit value
RepeatSeq	GQ	F	Genotype quality
	DP	F	Genotype read depth

¹ Describes whether the field is calculated from FORMAT fields (F) or INFO fields (I) in the variant caller's VCF

² Computed using the RO and DPR FORMAT fields

Supplementary Table 3: Capillary-based benchmarking of STR genotyping tools

Method	Correct Calls ¹	Total Calls ¹	Accuracy (%)	Run Time (hrs)	Variants per STR ²
HipSTR	58937	61942	95.2	1.3	1
GATK-HC*	59364	62966	94.3	5.8	5.5
GATK-HC	57947	62966	92	6.3	5.5
lobSTR	54627	61971	88.2	0.1	1
Platypus*	50196	62937	79.8	15.7	4.3
freebayes*	39197	57349	63.5	4.8	2.8
Platypus	39220	62347	62.9	1.9	2.9
RepeatSeq	30289	52393	57.8	0.3	1
freebayes	9011	61741	15.7	3.6	3

* Tool was run with settings optimized for accurate STR calling

1 The reported number of calls refers to the 118 samples with capillary genotypes

2 Average number of VCF records per STR region

Supplementary Table 4: Primers for STR Sanger experiments

Locus (hg19)	Repeat Unit	Primer 1	Primer 2
chr20:45121567-45121577	A	CAAGGCAGGTGGTTCATGAG	TGCTTACCACAGGAAATGCC
chr19:37201975-37201994	TG	CCTCGGTGATCTCTCCAACA	GGAGTTCAAAACCAGCCCAG
chr18:27043336-27043370	ATC	ATTAGTCAGCCCGAGGTACA	TGTCTCTATGAGGAAAGGCTTGT
chr1:221172558-221172630	ATCT	TCCAGGGTTCAAGCGATTCT	ATGTCCTCCCATGGCCTTAC