HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2017 October 24.

Published in final edited form as:

Nat Methods. 2017 June ; 14(6): 590–592. doi:10.1038/nmeth.4267.

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

Thomas Willems[1,2,*], **Dina Zielinski**[1], **Jie Yuan**[1,3], **Assaf Gordon**[1], **Melissa Gymrek**[4,5], and **Yaniv Erlich**[1,3,6,*]

[1]New York Genome Center, New York, New York 10013, USA

[2]Computational and Systems Biology Program, MIT, Cambridge, Massachusetts 02139, USA

[3]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, New York 10027, USA

[4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA

[5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

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

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 and phasing STRs from Illumina 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,

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

Data availability

Sequencing datasets analyzed during the benchmarking and *de novo* variant analyses are publicly available using the accession codes listed in the Online Methods. The genotypes for STRs with a high confidence *de novo* mutation and the various BED files used to genotype STRs are available at https://github.com/HipSTR-Tool/HipSTR-paper. The long read MiSeq data is not publicly available as it could compromise the privacy of the sequenced individual.

Author Contributions

T.W., M.G and Y.E. designed the HipSTR algorithm and subsequent analyses. T.W. and A.G implemented the HipSTR software. T.W. and J.Y. performed the analyses. D.Z. experimentally validated the *de novo* mutations and analyzed the long MiSeq reads. T.W and Y.E wrote the manuscript.

Competing Financial Interests

Y.E. is a consultant for Arc Bio, a company interested in DNA forensics.

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 disorders[1] and recent evidence has underscored their profound regulatory role and potential involvement in complex traits[2–4]. Beyond medical genetics, STRs have applications in population genetics, forensics, and single cell lineage analysis. While several STR callers exist[5, 6], they suffer from limited accuracy, difficulties in calling homopolymer runs, sensitivity to PCR stutter noise, and limited functionality compared to SNP callers. As a result, large-scale projects[7–9] 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[5], addresses major limitations of existing STR tools, and is designed for Illumina reads (Supplementary Fig. 1). Briefly, HipSTR begins by learning a parametric model that captures each STR's stutter noise profile. Using the genomic location of the repeat, it then harnesses this profile and a hidden Markov model (HMM) to realign the STR-containing reads to candidate haplotypes and mitigate the effects of PCR stutter (Supplementary Fig. 2). The realignment framework is highly flexible and can integrate population-scale data from other individuals and phased SNP scaffolds to determine the most likely alleles, conferring robustness to the genotyping process (Supplementary Figs. 3–4). The output of HipSTR is a VCF file that consolidates all of an STR's variants into a single line.

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 263 WGS datasets from the Simons Genome Diversity Project (SGDP)[10] that were sequenced with an Illumina 100bp paired-end PCR-free protocol to at least 30× coverage. A subset of these samples also have capillary electrophoresis calls for 600 highly polymorphic STRs from the Marshfield panel[11], providing a challenging test case for STR callers. The capillary calls for a few duplicated samples showed an internal consistency of ~98.5%, setting an upper bound on the accuracy achievable in our tests. For comparison, we also genotyped the same STRs with two STR-specific tools: lobSTR[5] and RepeatSeq[6]. We optimized command line options to boost the accuracy of each tool and developed a machine learning approach to rank each tool's calls by quality (Supplementary Tables 1–2). Under all settings, HipSTR outperformed these tools. HipSTR achieved an overall accuracy of 95.2%, while lobSTR and RepeatSeq achieved overall accuracies of 88.2% and 57.8%, respectively (Figure 1; Supplementary Fig. 5; Supplementary Table 3). After filtering the 10% least confident genotypes, HipSTR again exhibited superior performance and its accuracy improved to 98.9%, saturating the capillary data's limit.

Using the same benchmarking framework, we compared HipSTR to five widely used variant callers: GATK HaplotypeCaller (GATK-HC)[12], Platypus[13], freebayes[14], SAMtools[15] and VarScan[16] (Figure 1; Supplementary Fig. 5; Supplementary Table 3). Variant callers that require reliable input alignments (SAMtools and VarScan) were the least accurate,

highlighting the challenges posed by repetitive regions. Local assembly-based approaches demonstrated improved accuracy, but HipSTR outperformed these tools under all relevant scenarios. GATK-HC was consistently the second best tool, but it required 4.4x more computation time than HipSTR and only achieved comparable accuracy after filtering a majority of calls. Analyses of downsampled data from the SGDP indicated that HipSTR also consistently outperformed GATK-HC across a wide range of sequencing coverage (Supplementary Fig. 6).

To explore the performance of HipSTR with longer Illumina reads, we performed 2×300bp targeted MiSeq sequencing of a panel of long forensic STRs in a single individual from our lab collection. The resulting HipSTR calls perfectly matched the capillary results even for markers with alleles longer than 100bp (Supplementary Table 4), demonstrating that it has no intrinsic limitations in calling STRs other than Illumina read lengths.

Next, we evaluated the ability of HipSTR to report not only length polymorphisms but also full STR haplotypes. About half of the STRs in the genome display a repeat structure that includes short interruptions to the recurrent motif[17]. Thus, two STR alleles with identical lengths can differ in sequence due to distinct evolutionary paths[18] (Supplementary Fig. 7). Current STR callers and capillary electrophoresis methods only report an STR's length and cannot differentiate between homoplastic alleles. Similarly, general-purpose tools typically report multiple unphased variants per STR, limiting the utility of these calls (Supplementary Table 5). As HipSTR reports fully phased diploid STR sequences, we sought to test its accuracy using the CEPH trio in the Illumina Platinum genomes dataset. For ~70,700 STR loci that passed our filters, at least two alleles had identical lengths but different sequences. Only 304 of these loci (0.4%) were inconsistent with Mendelian inheritance, highlighting the robustness of the reported 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. Taken together, our results highlight that HipSTR not only accurately reports length polymorphisms but also adds valuable information about the sequence and haplotype context of STR 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 30× coverage available at the New York Genome Center. Using HipSTR to genotype 1.6 million STRs in the human genome only required an average of 10 CPU hours per sample. For each genome, HipSTR reported an average of ~360,000 STR loci that differed from the human reference.

Encouraged by the accuracy and scalability of HipSTR, we wondered about its ability to identity *de novo* STR mutations (Supplementary Fig. 8). After genotyping ~1.6 million STRs in the CEPH trio, HipSTR identified ~745,000 loci with at least one length variation. To enhance the specificity of our analysis, we applied stringent quality filters and restricted our analysis to ~265,000 STRs. Across these loci, HipSTR identified 423 *de novo* STR variants in which the child possessed an allele length not observed in the parental genotypes.

To validate these mutations, we re-ran HipSTR on distinct Illumina datasets generated for these samples and compared the variants between runs. Notably, 358 (85%) of the mutations replicated as all samples' allele lengths matched perfectly. These *de novo* mutations predominantly occurred at homopolymer repeats (293/358=81%). To further validate our results, we used Sanger sequencing to genotype a subset of these loci. We TOPO cloned the STR alleles from each member of the trio and sequenced at least eight independent clones per individual, yielding high confidence Sanger calls for four STRs (Figure 2, Supplementary Figs. 9–12). In all cases, the Sanger calls confirmed the parental genotypes and the *de novo* allele reported by HipSTR, validating our method.

Finally, we sought to distinguish *de novo* STR variants that arose in the germline of the child (NA12878) from mutations that arose during cell-line passages of this sample. We used HipSTR to analyze the WGS data of her 11 offspring at the 358 STRs with a replicable mutation. For 31 loci, the *de novo* allele was transmitted to at least three offspring and was absent from the paternal genotype. For an additional 32 loci, the *de novo* allele was observed in at least three of her offspring, but the husband carried the same allele and the transmission could not be fully resolved. As we identified these 31 transmitted mutations by examining only ~35% of NA12878's STRs, the load of *de novo* STR mutations may rival the approximately 70 *de novo* SNP mutations expected per generation[7, 19]. However, we cannot exclude the possibility that some of these detected mutations are actually due to cell-line mutations in NA12878's parents. Future studies using patient-derived DNA samples will therefore be invaluable towards assessing the true contribution of STRs to *de novo* variation.

To summarize, our results show that HipSTR offers several advantages for STR calling. First, the technique is considerably more accurate than other variant callers and has exceptional computational tractability. Second, HipSTR offers new capabilities such as phasing, haplotyping, and reporting full STR sequences, important features for population genetic analyses, forensic work, and STR association studies. Finally, our method enables highly specific detection of *de novo* STR mutations. As HipSTR is limited to Illumina sequencing data, future efforts may benefit from adapting it to linked reads or longer reads from platforms such as PacBio and Oxford Nanopore. It is our hope that these efforts, in addition to applying HipSTR to increasingly rich Illumina datasets, will help unravel the role of STRs in human diseases and complex traits.

Online 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 STR sizes that differ 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 chromosome[20]. 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 of δ repeat units is:

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

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[21] approach 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 each read originated from either allele:

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

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

$$u^{t+1}=\frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A}P(g_i=(j,k)|R,\theta_x^t)\sum_{m=1}^{n_{reads,i}}\sum_{a\in j,k}P\left(s_{m,i}=a|g_i=(j,k),\theta_x^t\right)I(r_{m,i}>r_a)$$

$$d^{t+1}=\frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A}P(g_i=(j,k)|R,\theta_x^t)\sum_{m=1}^{n_{reads,i}}\sum_{a\in j,k}P\left(s_{m,i}=a|g_i=(j,k),\theta_x^t\right)I(r_{m,i}<r_a)$$

$$\rho_s^{t+1}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A}P(g_i=(j,k)|R,\theta_x^t)\sum_{m=1}^{n_{reads,i}}\sum_{a\in j,k}P\left(s_{m,i}=a|g_i=(j,k),\theta_x^t\right)I(r_{m,i}\neq r_a)}{\sum_{i=1}^{N}\sum_{j=1}^{A}\sum_{k=1}^{A}P(g_i=(j,k)|R,\theta_x^t)\sum_{m=1}^{n_{reads,i}}\sum_{a\in j,k}P(s_{m,i}=a|g_i=(j,k),\theta_x^t)|r_{m,i}-r_a|}$$

$$f_j^{t+1}=\frac{1}{2N}\sum_{i=1}^{N}\sum_{k=1}^{A}P(g_i=(j,k)|R,\theta_x^t)+P(g_i=(k,j)|R,\theta_x^t)$$

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 i^{th} sample's phased genotypes are:

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

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

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 with known phase. 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:

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

We compute each read's total phasing likelihood by multiplying $Q(b_i, h_j)$ for every base b_i in the read or its mate pair that is aligned to a heterozygous haplotype SNP h_j. 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. If no phased SNP information is available, HipSTR assigns equal phasing likelihoods to both of the strands.

Read alignment likelihoods—HipSTR assumes that each haplotype is composed of two distinct types of regions: flanking sequences and STR sequences (Supplementary Figure 1). As the sources of error prevalent in these two types of regions differ dramatically, HipSTR uses distinct models to align sequences to each type of region (Supplementary Figure 2). It then combines the likelihoods from these two different models at the junctions of these

regions by requiring 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—In flanking sequences, the alignment model accounts for Illumina sequencing errors using a previously described hidden Markov model[22]. To efficiently align reads in these regions, we use three matrices to recursively compute the maximum log-likelihood of aligning read bases $b_1 \ldots b_i$ with haplotype bases $h_1 \ldots h_j$. The matrices *Match* and *Ins* are used to track the log-likelihoods that read base b_i is aligned to haplotype base h_j or an insertion following haplotype 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:

$$Match(i,j) = log\left(Q\left(b_i, h_j\right)\right) + max \begin{cases} \mathrm{Match(i-1,j-1)} & +\mathrm{t}_{\mathrm{Match}\rightarrow\mathrm{Match}} \\ \mathrm{Del(i-1,j-1)} & +\mathrm{t}_{\mathrm{Match}\rightarrow\mathrm{Del}} \\ \mathrm{Ins(i-1,j-1)} & +\mathrm{t}_{\mathrm{Match}\rightarrow\mathrm{Ins}} \end{cases}$$

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

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

Aligning reads to STR sequences—In STR regions, HipSTR utilizes an alignment model that accounts 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 M. 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 $h_1 \ldots h_L$ of length L is:

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

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:

$$P\left(b_{\max(1,i-L+D+1)} \ldots b_i, D | h_1 \ldots h_L\right) = \frac{d\rho_s(1-\rho_s)^{\frac{D}{M}-1}}{L-D+1}$$

$$\sum_{d=0}^{L-D} \prod_{k=0}^{\min(d-1,i-1)} Q(b_{i-k}, h_{L-k}) \prod_{k=d}^{\min(L-D-1,i-1)} Q(b_{i-k}, h_{L-D-k})$$

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. Iterating over each possible insertion position results in the likelihood:

$$P\left(b_{\max(1,i-L-D+1)}\ldots b_i, D|h_1\ldots h_L\right)$$
$$=\frac{u\rho_s(1-\rho_s)^{\frac{D}{M}-1}}{L+1}\sum_{d=0}^{L}\prod_{k=0}^{\min(d-1,i-1)}Q\left(b_{i-k},h_{L-k}\right)\prod_{k=d+D}^{\min(L+D-1,i-1)}Q(b_{i-k},h_{L+D-k})$$
$$\prod_{k=d}^{\min(d+D-1,i-1)}\begin{cases}Q\left(b_{i-k},\ \mathrm{h}_{\mathrm{L-d-((k-d)mod\ M)}}\right), & \mathrm{L-d-((k-d)modM)}\geq 1\\ Q\left(b_{i-k},\ \mathrm{h}_{\mathrm{M+L-d-((k-d)modM)}}\right), & otherwise\end{cases}$$

Experiments

Constructing a gold standard STR dataset—We downloaded capillary genotypes for 628 STRs in 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_11242009.txt. Using the is PCR[23] tool, we mapped each STR's primers to the hg19 reference genome. We then used Tandem Repeats Finder[24] to scan between the primer sites and identified the genomic coordinates of each STR based on the published repeat structure.

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 FermiKit[25] 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 capillary 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.

Assessing variant caller performance—We downloaded BWA-MEM alignments[26] for 263 publicly available Simons Genome Diversity Project samples (accession ERP010710). Using a BED file containing each STR's genomic coordinates, we ran every tool with default options. For general-purpose tools that have no prior knowledge about repetitive regions, we padded the STR regions by 15bp prior to genotyping to improve the chance that it captured indels near the STR boundaries. 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-FM[27] outside of the Galaxy environment and we therefore could not include this tool in our benchmarking experiments.

Most general-purpose variant callers report multiple unphased variants per STR region. Without phase information, it is frequently impossible to determine the lengths of an individual's two STR alleles present at a locus. To overcome this issue, we summed the sizes of all indels each caller reported 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.

To assess how variant filtration affects performance, we selected FORMAT and INFO fields from each tool's VCF file that might be indicative of genotype robustness (Supplementary Table 2). We then used gradient boosting to train ensembles of regression trees that convert these fields into a confidence level. Using five-fold cross-validation, we trained these classifiers on 20% of the calls and ranked the remaining 80% of calls by confidence level. Figure 1 and Supplementary Figure 5 depict the relationship between mean accuracy and minimum confidence level observed across the five iterations.

To gauge the effect of sequencing coverage on variant caller performance, we used the SAMtools *view* command and the *–s* option to downsample the SGDP reads to 5×, 10× or 20× median coverage. We then reran GATK-HC and HipSTR using the command line parameters optimized for STR genotyping (Supplementary Table 1) but also specified the *--use-all-reads* HipSTR option.

Calling STRs from longer reads—Targeted 300bp paired end sequencing was performed on an Illumina MiSeq (supplied by Kailos Genetics). After aligning FASTQ files with BWA-MEM, we used HipSTR to call the GlobalFiler markers with the options *--haploid-chrs chrY --no-rmdup --min-reads 10 --read-qual-trim 5*. To compare with capillary data, we used information from NIST (http://www.cstl.nist.gov/strbase/str_fact.htm) to determine the number of repeats present in the hg19 reference genome (Supplementary Table 4). We then computed the HipSTR predicted capillary sizes by adding HipSTR's estimated base pair differences (GB FORMAT field) to the number of repeats in hg19.

Phased trio SNP scaffolds—We used the HaplotypeCaller module in GATK v3.5-0-g36282e4 to jointly genotype all members of the CEPH trio using aligned and sorted BAMs for runs ERR194147, ERR194160 and ERR194161. 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's[28] 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, *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 errors[29], 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 BWA-MEM and analyzed all of the resulting BAMs concurrently using HipSTRv0.2 and 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 with the phased SNP scaffolds VCF described above 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 HipSTR FORMAT field with PQ > 0.9 (minimum phased genotype posterior of 90%). Using the SNPs 50kb upstream and 50kb 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 a maternal and a paternal haplotype. After enforcing all 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 30× coverage was internally available at the New York Genome Center. 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 200× 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% (computed from MALLREADS).

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

Validating *de novo* mutations using orthogonal datasets—We downloaded FASTQs containing 300× Illumina sequencing data for NA12878 from the Genome in a Bottle consortium and FASTQs for NA12878, NA12891 and NA12891 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.

Sanger sequencing validation—Primers were designed around the STR coordinates to generate 300–600bp PCR products (Supplementary Table 6). Primers were tested using isPCR 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. cloni 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 50× Illumina sequencing data for each of the 11 children of NA12878 (ERR194148, ERR218433, ERR324432, ERR324433, ERR194152, ERR324434, ERR194154, ERR194155, ERR324435, ERR194157 and ERR194162) and aligned them using BWA-ME. 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.

Code availability—The latest version of HipSTR and detailed usage information are freely available under the GNU General Public License v2.0 at https://hipstr-tool.github.io/HipSTR

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by NIJ grant 2014-DN-BX-K089 (T.W., D.Z, A.G., M.G., Y.E.) and a generous gift by Andria and Paul Heafy. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number UM1HG008901. We thank Kailos Genetics for providing the 300bp Illumina sequencing data. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007; 447:932–940. [PubMed: 17581576]
2. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet. 2002; 30:315–320. [PubMed: 11919562]
3. Gymrek M, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016; 48:22–29. [PubMed: 26642241]
4. Hefferon TW, Groman JD, Yurk CE, Cutting GR. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc Natl Acad Sci USA. 2004; 101:3504–3509. [PubMed: 14993601]
5. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012; 22:1154–1162. [PubMed: 22522390]
6. Highnam G, et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res. 2013; 41:e32. [PubMed: 23090981]
7. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–475. [PubMed: 22914163]
8. Zook JM, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014; 32:246–251. [PubMed: 24531798]
9. The Genomes Project C. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
10. Mallick S, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538:201–206. [PubMed: 27654912]
11. Rosenberg NA, et al. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 2005; 1:e70. [PubMed: 16355252]
12. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
13. Rimmer A, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014; 46:912–918. [PubMed: 25017105]
14. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012 arXiv preprint arXiv:1207.3907.
15. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

16. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22:568–576. [PubMed: 22300766]
17. Willems T, et al. The landscape of human STR variation. Genome Res. 2014; 24:1894–1904. [PubMed: 25135957]
18. Estoup A, Jarne P, Cornuet JM. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol. 2002; 11:1591–1604. [PubMed: 12207711]
19. Francioli LC, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015; 47:822–826. [PubMed: 25985141]
20. Willems T, et al. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am J Hum Genet. 2016; 98:919–933. [PubMed: 27126583]
21. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977; 39:1–38.
22. Albers CA, et al. Dindel: accurate indel calls from short-read data. Genome Res. 2011; 21:961–973. [PubMed: 20980555]
23. Hinrichs AS, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006; 34:D590–598. [PubMed: 16381938]
24. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27:573–580. [PubMed: 9862982]
25. Li H. FermiKit: assembly-based variant calling for Illumina resequencing data. Bioinformatics. 2015; 31:3694–3696. [PubMed: 26220959]
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 arXiv preprint arXiv:1303.3997.
27. Fungtammasan A, et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 2015; 25:736–749. [PubMed: 25823460]
28. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81:1084–1097. [PubMed: 17924348]
29. Weisenfeld NI, et al. Comprehensive variation discovery in single human genomes. Nat Genet. 2014; 46:1350–1355. [PubMed: 25326702]

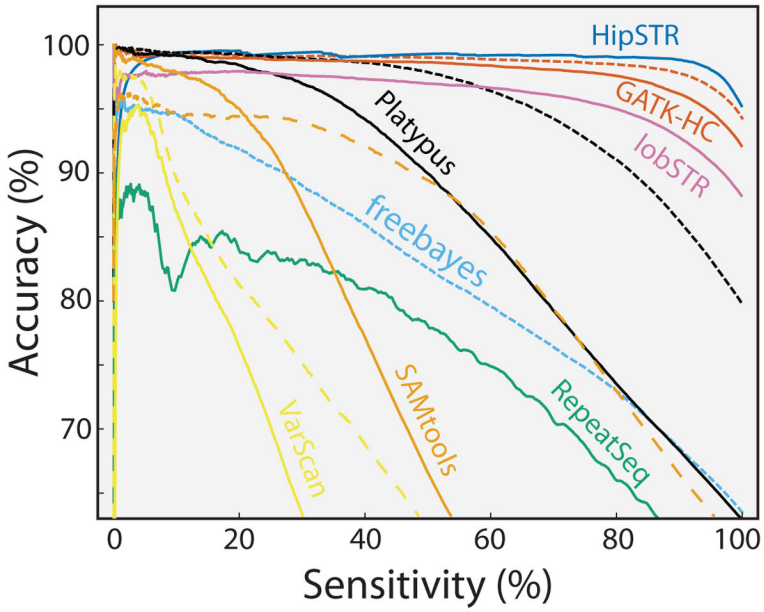

Figure 1. Performance of variant callers in STR regions

The accuracy of each tool's calls is shown as a function of sensitivity for the Marshfield STR panel. Solid and dashed lines denote tools run using default settings and settings optimized for STR genotyping, respectively.

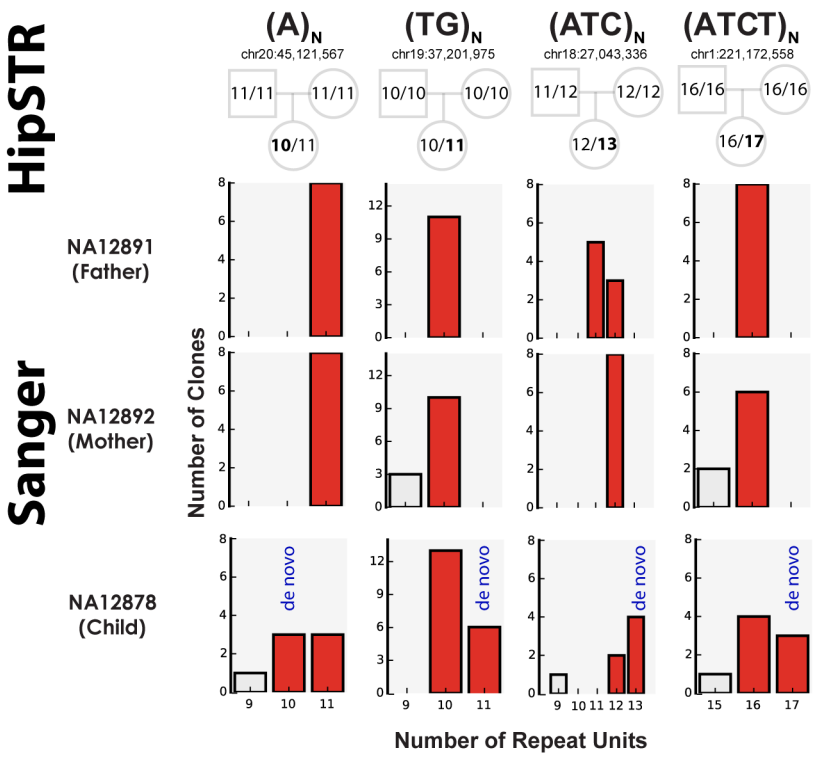

Figure 2. Experimental validation of *de novo* STR mutations

The top panel depicts the number of repeats HipSTR identified in each family member for four STRs with a predicted *de novo* mutation (novel allele in bold). The bottom three panels illustrate the number of clones with repeat sizes predicted (red) or not predicted (gray) by HipSTR during Sanger sequencing of these same individuals.