Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates

Thomas Willems,[1,2,3] Melissa Gymrek,[1,3,4,5] G. David Poznik,[6,7] Chris Tyler-Smith,[8] The 1000 Genomes Project Chromosome Y Group, and Yaniv Erlich[1,3,9,10,*]

Short tandem repeats (STRs) are mutation-prone loci that span nearly 1% of the human genome. Previous studies have estimated the mutation rates of highly polymorphic STRs by using capillary electrophoresis and pedigree-based designs. Although this work has provided insights into the mutational dynamics of highly mutable STRs, the mutation rates of most others remain unknown. Here, we harnessed whole-genome sequencing data to estimate the mutation rates of Y chromosome STRs (Y-STRs) with 2–6 bp repeat units that are accessible to Illumina sequencing. We genotyped 4,500 Y-STRs by using data from the 1000 Genomes Project and the Simons Genome Diversity Project. Next, we developed MUTEA, an algorithm that infers STR mutation rates from population-scale data by using a high-resolution SNP-based phylogeny. After extensive intrinsic and extrinsic validations, we harnessed MUTEA to derive mutation-rate estimates for 702 polymorphic STRs by tracing each locus over 222,000 meioses, resulting in the largest collection of Y-STR mutation rates to date. Using our estimates, we identified determinants of STR mutation rates and built a model to predict rates for STRs across the genome. These predictions indicate that the load of de novo STR mutations is at least 75 mutations per generation, rivaling the load of all other known variant types. Finally, we identified Y-STRs with potential applications in forensics and genetic genealogy, assessed the ability to differentiate between the Y chromosomes of father-son pairs, and imputed Y-STR genotypes.

Introduction

Mutations provide the fuel for evolutionary processes. The rates at which new mutations arise play a central role in a range of genetic applications, including dating phylogenetic events,[1] informing disease studies,[2] and evaluating forensic evidence.[3] The advent of high-throughput sequencing has enabled genome-wide measurements of the number of de novo mutations via a broad range of strategies. A host of studies have evaluated the mutation rates of nearly every type of genetic variation, ranging from SNPs[4–7] and short indels[8] to large structural variations.[9] These sequencing studies have concluded that approximately 50–100 de novo mutations, most of which are point mutations, arise each generation. However, these studies have largely overlooked the contribution of short tandem repeats (STRs).

STRs are one of the most abundant types of repeats in the human genome. They consist of a repeating 2–6 bp motif and span a median of 25 bp. Approximately 700,000 STR loci exist in the human genome, and in aggregate, they occupy ~1% of its total length. STR variations have been implicated in more than 30 hereditary disorders,[10] and emerging lines of evidence have highlighted their involvement in complex traits in both humans[11–13] and model organisms.[14–16] The repetitive nature of STRs causes error-prone DNA-polymerase replication events that can insert or delete copies of the repeat motif in subsequent generations, leading to markedly elevated mutation rates.[17,18]

Previous studies estimated the rates and patterns of de novo STR mutations by using capillary electrophoresis genotyping of specialized sets of markers, such as the Marshfield panel, CODIS (Combined DNA Index System) markers, or specific Y chromosome STRs (Y-STRs). These studies have estimated that the average STR mutation rate per locus is 10^{-3} to 10^{-4} mutations per generation (mpg).[17,19–22] However, most STRs characterized in these studies were chosen for their relatively high levels of diversity in the population. As such, it is not clear whether their mutation rates and patterns reflect those of most STRs in the genome. Furthermore, given that most previously studied STRs have tri- and tetranucleotide motifs, the field lacks robust mutation-rate estimates for other motif lengths, specifically those of dinucleotides, the most prevalent type of STR. Finally, capillary electrophoresis has relatively low throughput, and most STRs were never genotyped in these studies, leaving the specific mutation rates of most STRs unknown.

The rapid advancement of next-generation sequencing technologies has provided the opportunity to genotype STRs beyond those on existing panels and to do so on a larger scale. Coupled with vast improvements in the depth, read length, and quality of whole-genome sequencing (WGS) datasets, algorithmic progress in STR genotyping tools has made it possible to robustly call these markers

[1]New York Genome Center, New York, NY 10013, USA; [2]Computational and Systems Biology Program, MIT, Cambridge, MA 02139, USA; [3]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02139, USA; [4]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA; [5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA; [6]Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA; [7]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [8]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; [9]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY 10027, USA; [10]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

*Correspondence: yaniv@cs.columbia.edu

http://dx.doi.org/10.1016/j.ajhg.2016.04.001.

from high-throughput data.[23–25] In our previous study, we found that 90% of the STRs in the genome are accessible to Illumina technology, and we showed that hemizygous STRs can be called with very high accuracy.[26]

Here, we leveraged population-scale high-throughput sequencing data to systematically estimate the mutation rates and analyze the mutational dynamics of STRs across the Y chromosome. To gain power, we used two independent datasets, the 1000 Genomes Project[27] and the Simons Genome Diversity Project (SGDP).[28] The Y chromosomes in these datasets confer rich genealogical information, enabling the analysis of complex STR mutation models without the need for familial information. To leverage this genealogical information, we developed MUTEA (Measuring Mutation Rates using Trees and Error Awareness), an algorithm that infers the mutational dynamics along the Y chromosome branches. After validating MUTEA via intrinsic and extrinsic tests, we scanned 4,500 Y-STRs and used the algorithm to infer the mutation rates of 702 polymorphic Y-STRs. To the best of our knowledge, this is the largest collection of Y-STR mutation rates to date. We show the value of this large collection of mutation rates by uncovering the sequence determinants of mutability, predicting the genetic load of de novo STR mutations across the genome, and exploring a series of forensic applications.

Material and Methods

Sequencing Datasets

We analyzed 179 male SGDP samples from widely dispersed populations across Africa, Asia, and the Americas. The SGDP sequenced these samples to over 30× coverage by using a PCR-free library-preparation protocol and 100 bp paired-end Illumina reads. Given that our previous results demonstrated that this protocol substantially reduces the rate of PCR stutter at STR loci,[29] the SGDP cohort provides a high-quality dataset for calling Y-STRs. We also analyzed 1,244 unrelated male samples from phase 3 of the 1000 Genomes Project. These samples are from 26 globally diverse populations and were sequenced to an average autosomal coverage of 7× with 75–100 bp paired-end Illumina reads.

Y-SNP Phylogeny

To construct the SGDP Y chromosome haplotype tree, we downloaded VCF files containing the Y-SNP calls generated by the SGDP analysis group. Because many of these SNPs lie in pseudoautosomal regions or regions with low mappability, we applied a series of filters to reduce the frequency of genotyping errors. Using VCFtools,[30] we first removed loci for which more than 10% of individuals were heterozygous. For the remaining SNPs, we removed individual SNP calls that were heterozygous, had fewer than seven supporting reads, or had more than 10% of reads supporting an uncalled allele. Lastly, we discarded SNP loci if fewer than 150 samples met these criteria or if more than 10% of reads had zero mapping quality. Overall, we obtained nearly 39,000 high-quality polymorphic SNPs.

We then used the high-quality SNPs to build the Y chromosome phylogenetic tree with RAxML[31] and the options -m ASC_GTRGAMMA -f d --asc-corr lewis. The SGDP samples included three representatives of haplogroup A1b1 and no members of the more basal clades (A00, A0, and A1a), so we used Dendroscope[32] to root the phylogeny along the branch marked by the M42 and M94 mutations, markers associated with the split between A1b1 and megahaplogroup BT. For the 1000 Genomes phase 3 dataset, we used a RAxML-generated phylogeny that was built by the 1000Y analysis group.[33]

Although the maximum-likelihood phylogeny generated for each dataset has numerical branch lengths, these lengths are not scaled in units of generations, as required by our method. We therefore tested two scaling approaches. First, we selected the factor that most closely equated the total number of generations in each phylogeny to the corresponding value on the basis of published Y-SNP mutation rates. To do so, we used a recently published Y-SNP mutation rate of 3×10^{-8} mutations per base per generation[34,35] and the numbers of called SNPs and called sites in each SNP dataset. As an alternative method, we scaled the trees by using mutation-rate estimates for 15 loci in the Y Chromosome Haplotype Reference Database (YHRD), a large compendium of individual Y-STR mutational studies (individually cited therein).[36] We chose to use these loci for calibration because their mutation-rate estimates are each based on more than 7,000 father-son pairs per locus and should therefore be relatively precise. For the 1000 Genomes data, we used the available PowerPlex capillary data for each locus, assumed error-free genotypes, scaled the phylogeny by using a range of factors, and used MUTEA (see below) to estimate the set of mutation rates for each scaling factor. The choice of scaling factor had essentially no effect on the correlation with the YHRD estimates, resulting in an R^2 of 0.89 across all tested factors (Figure S1). However, the total squared error between the estimates was minimized for a factor of ~2,800, which we therefore selected as the optimal scaling. For the SGDP data, we performed an analogous analysis by using HipSTR genotypes (see below) for 9 of these 15 loci, again resulting in a uniform R^2 of 0.91 and an optimal scaling factor of ~3,200 (Figure S1).

The resulting scaling factors were remarkably concordant between the methods, although the factors determined by the Y-SNP method were ~25% greater. However, to maximize the concordance with pedigree estimates, we used the second method. After scaling the branches, we found that the approximate total lengths of the SGDP and 1000 Genomes phylogenies were 60,000 and 160,000 meioses, respectively.

Defining and Identifying Y-STRs

To identify Y-STRs, we used a quantitative procedure developed in our previous work.[26] In brief, this procedure uses Tandem Repeats Finder (TRF) to score each genomic sequence according to its purity, length, and nucleotide composition.[37] It then uses extensive simulations of random nucleotide sequences to determine a scoring threshold that distinguishes random DNA from DNA that is truly repetitive and then selects regions with scores above this threshold as STRs. Our previous results suggested that this approach has less than a 1.4% probability of omitting a polymorphic STR and has a false-positive rate of approximately 1%.

We applied this procedure to the Y chromosome sequence of the hg19 reference genome (UCSC Genome Browser). Because TRF occasionally identifies regions that overlap, we ensured that every locus had a unique STR annotation by using the following steps. (1) We merged two STR regions if the higher-scoring one contained 85% of the bases in the union of the regions. (2) We also merged overlapping entries that failed this criterion but had the same period. For example, adjacent $[GATA]_{10}$ and $[TACA]_8$ entries were merged into one STR. (3) Because we intended to use sequencing alignments relative to either hg19 or GRCh38 coordinates, we removed hg19 STR regions that failed to liftOver[38] to the GRCh38 assembly or were lifted from the Y chromosome to the X chromosome.

We also added coordinates for Y-STR loci whose mutation rates had been characterized in prior studies.[21,39] For these markers, we used the published set of primer sequences and the isPCR tool[38] to map the primers to hg19 coordinates. We then ran TRF on each region and pinpointed the coordinates by using the published repeat structure. Lastly, we applied TRF to additional regions previously published as part of comprehensive Y-STR maps to obtain coordinates for labeled markers whose mutation rates had not been previously characterized.[40] In total, we added 261 annotated Y-STRs, ~190 of which had mutation-rate estimates from prior studies. The complete Y-STR reference is available for download in both hg19 and GRCh38 coordinates (Web Resources).

Y-STR Call Set and Its Accuracy

We downloaded BWA-MEM[41] alignments for the SGDP samples from the project website and used SAMtools[42] to extract and merge the Y chromosome alignments into a single BAM file. STR genotypes were then generated with HipSTR, an improved version of lobSTR, an STR caller for Illumina data we developed in our previous studies.[23]

HipSTR provides additional capabilities over lobSTR because it uses a specialized hidden Markov model (HMM) to account for PCR stutter artifacts. In brief, to genotype an STR, HipSTR creates a list of candidate alleles from the alignments observed in the population. For each sample, it then realigns every read to each putative allele by using the HMM, selects the allele with the highest total likelihood as the genotype, and returns each read's alignment in relation to this genotype. This haplotype-based approach produces highly accurate STR genotypes and eliminates many read misalignments that can occur if reads are aligned individually or are only aligned to the reference genome. To genotype each STR region in the Y-STR reference described above, we ran HipSTR by using the merged BAMs and the following options: --min-reads 25 --haploid-chrs chrY --hide-allreads. Similarly, we downloaded BWA-MEM alignments from the 1000 Genomes phase 3 data release. Because these alignments were relative to the GRCh38 assembly, we ran HipSTR by using the corresponding GRCh38 STR regions and the options --min-reads 100 --haploid-chrs chrY --hide-allreads.

We employed several strategies to enhance the quality of the SGDP STR call set. (1) To avoid errors introduced by neighboring repeats, we omitted genotyped loci that overlapped one another or multiple STR regions. (2) We discarded loci if more than 5% of samples' genotypes had a non-integer number of repeats, such as a 3 bp expansion in an STR with a tetranucleotide motif. These types of events occur quite rarely and usually reflect genotyping errors rather than genuine STR polymorphisms.[23] (3) We removed Y-STRs that were called in at least two SGDP females because they are likely to have high X chromosome or autosome homology. (4) We omitted sites if more than 15% of reads had a stutter artifact or more than 7.5% of reads had an indel in the sequence flanking the STR. These HipSTR-reported statistics typically indicate that the locus is not well captured by HipSTR's genotyping model and can arise if duplicated sites map to the same location in the reference genome. (5) For the remaining loci, we discarded unreliable calls on a per-sample basis if more than 10% of an individual's reads had an indel in the flank sequence. (6) Finally, we removed loci in which fewer than 100 samples had genotype posteriors greater than 66%, because these loci had too few samples for accurate inference.

To filter the 1000 Genomes call set, we first removed loci that did not pass the SGDP dataset filters. We then applied a set of filters identical to those described above except that we only removed loci with more than 15 genotyped females and did not apply a cutoff for stutter frequency. These alterations account for the 1000 Genomes dataset's larger sample size and use of PCR amplification during library preparation.

Importantly, we found that both the SGDP and 1000 Genomes HipSTR call sets were of high quality. We compared our STR genotypes to capillary-electrophoresis datasets available for the same samples. For the SGDP samples, we observed a 99.7% concordance rate when we compared the HipSTR and capillary results for 3,300 calls at 48 Y-STRs.[43] For the 1000 Genomes samples, a comparison of 4,050 calls at 15 loci in the PowerPlex Y23 panel resulted in a 97.5% concordance rate.[44]

MUTEA: Theory

Previously developed methods estimate STR mutation rates from population data by comparing the mean squared difference in allele lengths between samples to the time to the most recent common ancestor (TMRCA).[45,46] However, these methods generally assume simple mutation models, can be sensitive to fluctuations in haplogroup size,[47] and require exact error-free genotypes. We therefore sought to develop an algorithm that can address these issues by leveraging detailed Y-SNP phylogenies.

Figure 1 outlines the steps underlying MUTEA. Under a naive setting without genotyping error, MUTEA uses Felsenstein's pruning algorithm[48] and numerical optimization to evaluate and improve the likelihood of a mutation model until convergence. However, because of the error-prone and low-coverage nature of WGS-based STR call sets, using these genotypes would result in vastly inflated mutation-rate estimates. To avoid these biases, MUTEA learns a locus-specific error model and uses this error model to compute genotype posteriors. It then uses these posteriors rather than fixed genotypes during the process of optimizing the mutation model to obtain robust estimates. In addition, for STR mutations, MUTEA uses a flexible computational framework that includes length constraints and allows for multi-step mutations. We describe each step below.

Likelihood of a Mutation Model

We used Felsenstein's pruning algorithm to evaluate the likelihood of an STR mutation model. Let M denote the STR mutation model, D denote the dataset containing STR genotype likelihoods, and T denote the Y chromosome phylogeny rooted at node R. The likelihood of the data is

$$\begin{aligned} P(D \mid M,T) &= \sum_r P(R=r, D \mid M,T) \\ &= \sum_r P(R=r \mid M,T) P(D \mid R=r, M,T). \end{aligned}$$

Let D_{N_i} denote the genotype likelihoods of all nodes that are in the subtree rooted at node N_i. If node N_i has genotype g, the conditional probability of the data in its subtree is given by

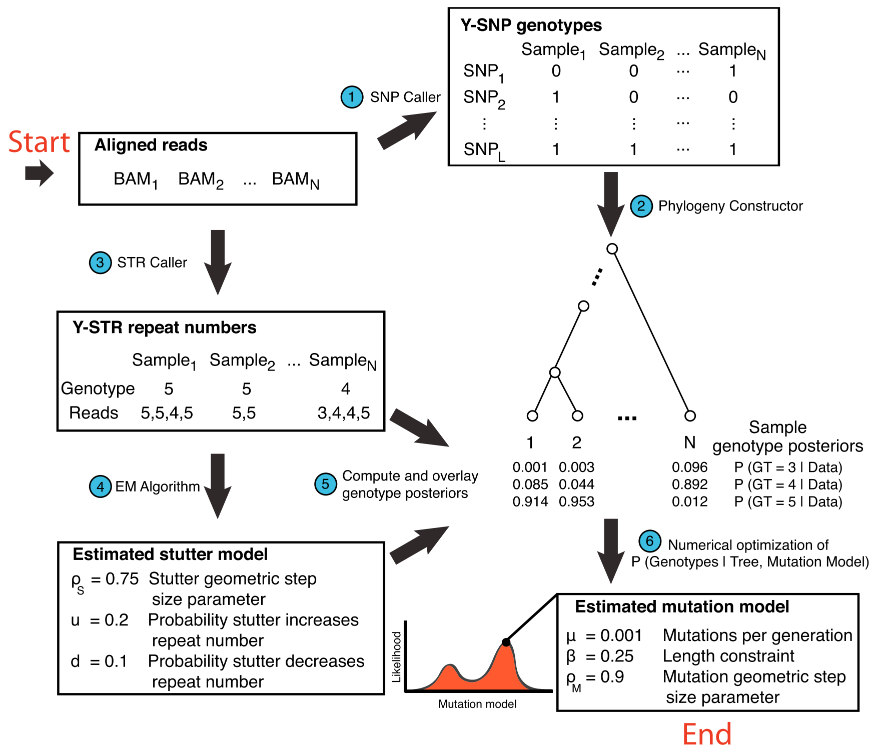

Figure 1. Method for Estimating Y-STR Mutation Rates

Schematic of our procedure for estimating Y-STR mutation rates. The method first genotypes Y-SNPs (step 1) and uses these calls to build a single Y-SNP phylogeny (step 2). This phylogeny provides the evolutionary context required for inferring Y-STR mutational dynamics; samples in the cohort occupy the leaves of the tree, and all other nodes represent unobserved ancestors. Steps 3–6 are then run on each Y-STR individually. After an STR genotyping tool is used for determining each sample's maximum-likelihood genotype and the number of repeats in each read (step 3), an EM algorithm analyzes all of these repeat counts to learn a stutter model (step 4). In combination with the read-level repeat counts, this model is used for computing each sample's genotype posteriors (step 5). After a mutation model is randomly initialized, Felsenstein's pruning algorithm and numerical optimization are used to repeatedly evaluate and improve the likelihood of the model until convergence. The mutation rate in the resulting model provides the maximum-likelihood estimate.

$$
\begin{aligned}
P(D_{N_i} \mid N_i = g, M, T) &= \prod\nolimits_{C_j \in \text{child}(N_i)} \sum\nolimits_{b \in \text{alleles}} P\left(C_j = b, D_{C_j} \mid N_i = g, M, T\right) \\
&= \prod\nolimits_{C_j \in \text{child}(N_i)} \sum\nolimits_{b \in \text{alleles}} P(C_j = b \mid N_i = g, M, T) \\
&\quad \times P\left(D_{C_j} \mid C_j = b, M, T\right).
\end{aligned}
$$

While descending the phylogeny, this recursive relation applies until a node with no children is encountered. These leaf nodes represent sequenced individuals, and the conditional probability of the data is given by the individuals' genotype likelihoods. Therefore, the likelihood of a mutation model can be calculated with a post-order tree traversal. First, the algorithm computes the genotype likelihoods at each leaf node. It then progresses to each internal node and calculates the conditional probability of the data for each potential genotype after computing its descendants' probabilities. Finally, upon reaching the root node, the total data likelihood is computed with the root node's conditional probabilities and a uniform prior for the root node's genotype.

In practice, we compute the total log-likelihood to avoid numerical underflow issues. Because normalizing the genotype likelihoods of each sample does not affect the relative model likelihoods, we calculated genotype posteriors by using a uniform prior and used them throughout our analysis.

STR Mutation Model

To model STR mutations, we used a generalized stepwise mutation model with a length constraint. Each mutation model M is characterized by three parameters: a per-generation mutation rate μ, a geometric step-size distribution with parameter ρ_M, and β, a spring-like length constraint that causes alleles to mutate back toward the central allele. In this framework, the central allele is assigned a value of 0, and nonzero allele values indicate the number of repeats from this reference point. Given a starting allele a_t observed at time t, the probability of observing a particular allele k in the following generation is

$$
p(a_{t+1} = k \mid a_t) = \begin{cases} 1 - \mu, k = a_t \\ \mu f_i \, \rho_M (1 - \rho_m)^{k - a_t - 1}, k > a_t \\ \mu f_d \, \rho_M (1 - \rho_m)^{a_t - k - 1}, k < a_t \end{cases},
$$

where the fraction of mutations increasing and decreasing the size of the STR is $f_i = 1 - \beta \, \rho_M a_t / 2$ and $f_d = 1 - f_i$, respectively; f_i values greater than 1 or less than 0 were clipped and set to 1 and 0, respectively. These two model features act as spring-like length constraints that attract alleles back toward the central allele. To avoid biologically implausible models, we constrained β to have non-negative values, where $\beta = 0$ reduces to a traditional generalized stepwise mutation model, and increasingly positive values of β model STRs with stronger tendencies to mutate back toward the central allele. Values of ρ_M close to 1 primarily restrict models to single-step mutations, whereas smaller values of this parameter enable frequent multi-step mutations.

Computing Likelihoods of STR Genotypes

To calculate the likelihood of the data D observed in the leaf nodes, we needed to account for STR genotyping errors. These errors are mainly caused by PCR stutter artifacts that insert or delete STR units in the observed sequencing reads. We therefore developed a method to learn each STR's distinctive stutter-noise profile.

Let Θ_x denote the stutter model for STR locus x. Θ_x is parameterized by the frequency of each STR allele (F_i), 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. Given a stutter model and a set of observed reads (R), the posterior probability of each individual's haploid genotype is

$$P(g_i = j \mid R, \Theta_x) \propto F_j \prod_{k=1}^{n_{\text{reads},i}} \begin{cases} 1 - u - d, r_{k,i} = s_j \\ u\rho_s(1-\rho_s)^{r_{k,i}-s_j-1}, r_{k,i} > s_j, \\ d\rho_s(1-\rho_s)^{s_j-r_{k,i}-1}, r_{k,i} < s_j \end{cases}$$

where g_i denotes the genotype of the i^{th} individual, $n_{\text{reads},i}$ denotes the number of reads for the i^{th} individual, $r_{k,\,i}$ denotes the number of repeats observed in the k^{th} read for the i^{th} individual, and s_j denotes the number of repeats in the j^{th} allele. Analogous to the step-size parameter in the mutation model, small values of ρ_s allow for frequent multi-step stutter artifacts, whereas values near 1 restrict artifacts to single-step changes.

We implemented an expectation-maximization (EM) framework to learn these model parameters.[49] The E step computes the genotype posteriors for every individual given the observed reads and the current stutter-model parameters. The M step then uses these posterior probabilities to update the stutter-model parameters as follows:

$$u^{t+1} = \frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A} P(g_i = j \mid R, \Theta^t) \sum_{k=1}^{n_{\text{reads},i}} I(r_{k,i} > s_j)$$

$$d^{t+1} = \frac{1}{Q}\sum_{i=1}^{N}\sum_{j=1}^{A} P(g_i = j \mid R, \Theta^t) \sum_{k=1}^{n_{\text{reads},i}} I(r_{k,i} < s_j)$$

$$\rho_s^{t+1} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{A} P(g_i = j \mid R, \Theta^t) \sum_{k=1}^{n_{\text{reads},i}} I(r_{k,i} \neq s_j)}{\sum_{i=1}^{N}\sum_{j=1}^{A} P(g_i = j \mid R, \Theta^t) \sum_{k=1}^{n_{\text{reads},i}} |r_{k,i} - s_j|}$$

$$F_j^{t+1} = \frac{1}{N}\sum_{i=1}^{N} P(g_i = j \mid R, \Theta^t).$$

Here, N denotes the number of samples, A denotes the number of putative alleles, Q denotes the number of sequencing reads, and I is the indicator function. Because ρ_S is the parameter of a geometric step-size distribution, the M step updates its value by using the inverse of the mean weighted step size for reads with nonzero stutter.

Locally misaligned reads can also introduce genotyping errors if they cause a miscalculation in a read's repeat length. However, these errors introduce artifacts that are relatively similar to those caused by PCR stutter. As a result, the EM procedure learns stutter models that correct for the combined frequencies of PCR stutter and misalignment, resulting in robust genotype posteriors for downstream analyses.

MUTEA Computation

Given genotype likelihoods for an STR of interest, we used a maximum-likelihood approach to estimate the underlying mutation model. Our approach first estimates the central allele of the mutation model by computing the median observed STR length and then normalizes all genotypes in relation to this reference point. Next, it randomly selects mutation-model parameters μ, β, and ρ_M, subject to the constraint that they lie within the ranges of 10^{-5}–0.05, 0–0.75, and 0.5–1.0, respectively. Using these bounds, the Nelder-Mead optimization algorithm,[50] and the outlined method for computing each model's likelihood, we iteratively update the mutation-model parameters until the likelihood converges. After repeating this procedure by using three different random initializations to increase the probability of discovering a global optimum, our algorithm selects the optimized set of parameters with the greatest total likelihood.

For each SGDP and 100 Genomes STR that passed the requisite quality-control filters, we first used the EM algorithm to learn a model of PCR stutter. To run this algorithm, we obtained the STR size observed in each read from the MALLREADS VCF field. HipSTR uses this field to report the maximum-likelihood STR size observed in each read that spans its sample's most probable haplotype. We then used the learned stutter model in conjunction with a uniform prior to compute the genotype posteriors for each sample with a HipSTR quality score greater than 0.66. Samples with quality scores below this threshold were omitted because the genotype uncertainty can result in erroneous reported read sizes. We used these genotype posteriors, together with the optimization procedure and the appropriate scaled Y-SNP phylogeny, to obtain a point estimate of the STR's mutation rate. Finally, using a delete-d jackknife procedure, we computed a 95% confidence interval (CI) for the estimated mutation rate (Appendix A).

Results

Verifying MUTEA by Using Simulations

We validated MUTEA's inferences by running the algorithm on simulated data from a wide range of Y-STR mutation models (Appendix B). We tested mutation rates (μ) from 10^{-5} to 10^{-2} mpg, a range that encompasses most known polymorphic Y-STRs. We also varied the distribution of step sizes for each STR mutation from a single step ($\rho_M = 1$) to a wide range of mutation steps ($\rho_M = 0.75$) and added various spring-like length constraints that ranged from no constraint ($\beta = 0$) to a strong attractor toward the central allele ($\beta = 0.5$).

MUTEA obtained unbiased estimates of the simulated mutation rate for nearly all scenarios (Figure S2). We observed a slight upward bias only for the estimates of the slowest simulated mutation rate ($\mu = 10^{-5}$) as a result of the lower bound imposed during numerical optimization. In contrast, mutation rates estimated with simpler mutation models limited to single-step mutations or no length constraints were far more biased in these scenarios (Figure S3). MUTEA's inferences were also robust to the presence of simulated PCR stutter noise. After forward simulating STRs, we simulated reads for each genotype and distorted their repeat numbers by using various models of PCR stutter (Appendix C). We then input these repeat counts into MUTEA instead of the STR genotypes. Although MUTEA was completely blind to the selected stutter parameters, it reported unbiased estimates of the Y-STR mutation rates, step sizes, and stutter models for nearly all scenarios (Figure 2; Figures S4–S6), although it had just a slight bias for the lowest simulated mutation rate, as was the case for the exact-genotype scenario described above. As a negative control, we again ran MUTEA on the stutter-affected reads but without employing the EM stutter-correction method. With this procedure, posteriors based on the fraction of reads supporting each genotype resulted in marked biases, particularly for low mutation rates, demonstrating the importance of correctly accounting for stutter artifacts in these settings (Figure 2; Figures S5 and S6).

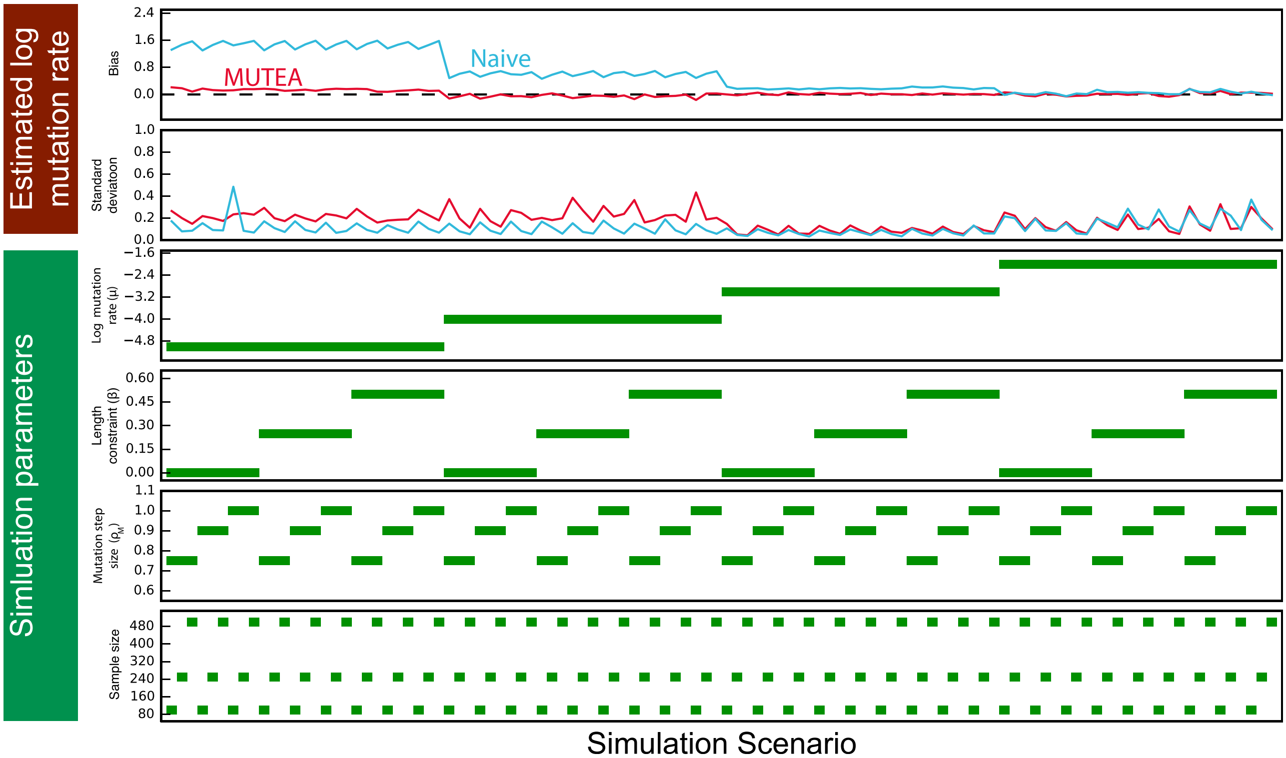

Figure 2. Validating MUTEA by Using Simulations
STR sequencing reads with PCR stutter noise were simulated for a variety of sample sizes and mutation models ("simulation parameters" panels). Applying MUTEA (red line) to these reads led to relatively unbiased mutation-rate estimates (upper panel) with small SDs (second panel). As a negative control, we also applied a naive approach to correct for stutter noise (blue line). This approach computed genotype posteriors by using the fraction of supporting reads, resulting in markedly biased mutation-rate estimates.

MUTEA Estimates Are Internally and Externally Consistent

Encouraged by the robustness of our approach, we turned to analyze real Y-STR data from the SGDP and the 1000 Genomes Y-STR call sets. In total, we examined ~4,500 STR loci, 702 of which displayed length polymorphisms in both datasets, and the rest were nearly fixed. We ran MUTEA on each of these polymorphic STRs to estimate its mutation rate (μ), expected step size (ρ_M), and stutter parameters (u, d, and ρ_s) in both datasets (Table S1).

The MUTEA mutation-rate estimates were largely consistent between the datasets (Figure 3). We obtained an R^2 of 0.92 when we compared the log mutation-rate estimates from the 1000 Genomes and SGDP datasets for the 702 polymorphic markers. Importantly, this high concordance was achieved despite substantial differences between the analyzed populations, sample sizes, and quality of the sequencing data. The 1000 Genomes data should have higher rates of stutter than the SGDP data because of the PCR amplification used in the preparation of the sequencing library. Consistent with this expectation, MUTEA learned higher stutter probabilities in the 1000 Genomes data than in the SGDP data for most loci (Figure S7, left panels). Nonetheless, the mutation-rate estimates were highly concordant. In addition, we found that despite differences in the overall probability of stutter, the downward and upward stutter rates were highly correlated between the two datasets ($R^2 = 0.88$ and $R^2 = 0.68$, respectively, on the log scale), reflecting the algorithm's ability to capture each locus's distinctive error profile (Figure S7, right panels).

Genotyping technology played only a small role in explaining the estimated concordance between the two datasets. We re-ran MUTEA on the 1000 Genomes Y-tree by using capillary genotypes for 15 Y-STR loci that were available for the same samples (Figure 3). Comparing the resulting log mutation-rate estimates to those obtained with sequencing-generated genotypes, we obtained an R^2 of 0.98. These comparisons demonstrate that our method obtains robust locus-specific mutation-rate estimates while accounting for varying degrees of PCR stutter artifacts and alignment and genotyping errors. Furthermore, the inter-dataset concordance suggests either that there are very few errors in the phylogenies or that these errors have little impact on the resulting mutation-rate estimates.

We also validated our mutation-rate estimates by comparing them to results from previous studies that used pedigree-based designs and capillary electrophoresis for genotyping. In these studies, Burgarella et al.[39] and Ballantyne et al.[21] estimated Y-STR mutation rates for specialized panels of Y-STRs by examining approximately 500 and 2,000 father-son duos, respectively, per Y-STR. We observed only a moderate replicability between the reported mutation rates from these two prior studies (R^2 of 0.34; Figure 3).

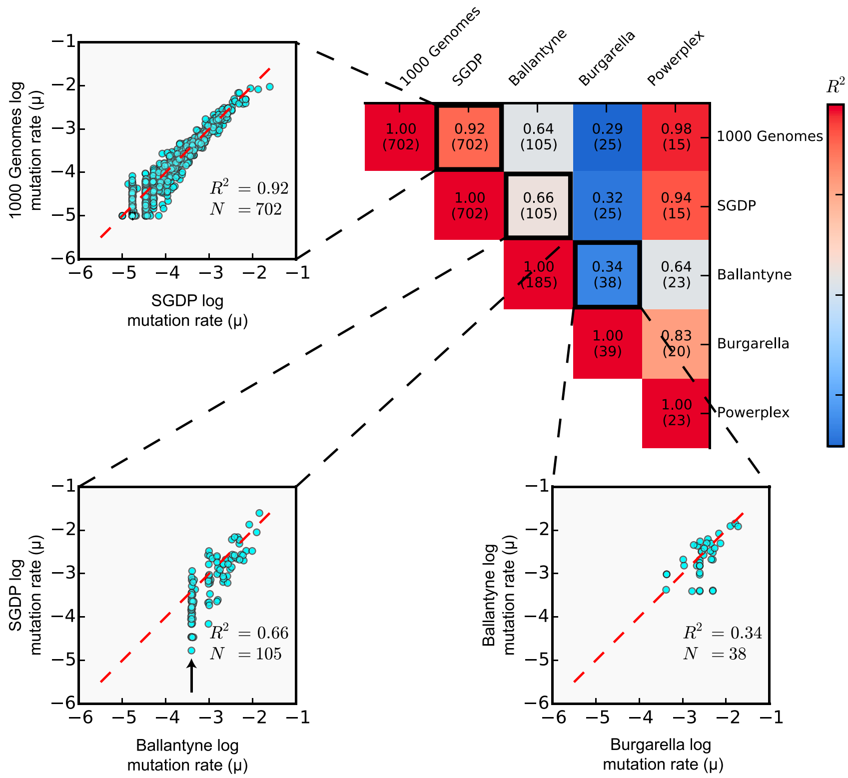

Figure 3. Concordance of Mutation-Rate Estimates across Datasets
The heatmap in the upper right corner presents the correlation between log mutation rates obtained from two father-son capillary-based studies ("Ballantyne"[21] and "Burgarella"[39]) and those we obtained by using the 1000 Genomes WGS data ("1000 Genomes"), the Simons Genome WGS data ("SGDP"), and the capillary data available for samples in 1000 Genomes ("Powerplex"). Each cell indicates the number of markers involved in the comparison and the resulting R^2. Representative scatterplots for three of these comparisons depict the pair of estimates for each marker (cyan) and the x = y line (red). The black arrow in the comparison of SGDP and Ballantyne shows the effective lower limit of the Ballantyne et al. mutation-rate estimates.

This low correlation presumably stems from the very small number of transmissions used by Burgarella et al. On the other hand, we observed an R^2 of ~0.65 when we compared either the SGDP or the 1000 Genomes estimates to those from Ballantyne et al., despite considerably different methodological approaches (Figure 3). One limitation of this comparison is that Ballantyne et al. could not report precise mutation rates for slowly mutating Y-STRs because of the number of meioses examined in their study. As a result, their estimates were effectively restricted to a lower bound of $\mu = 10^{-3.5}$ mpg (Figure 3, inset). In contrast, our deep phylogeny enabled us to accurately estimate much lower rates, highlighting the advantage of analyzing population data rather than father-son pairs for slowly mutating STRs. Comparing our estimates to those from Burgarella et al. resulted in an R^2 of ~0.3, but restricting this evaluation to the subset of loci they characterized by using more than 5,000 father-son duos resulted in a substantially higher R^2 of 0.87 (Figure S8). These results demonstrate that our estimates are concordant with prior father-son based results, provided that the latter were generated with sufficiently many pairs.

Characteristics and Determinants of Y-STR Mutations

Next, we analyzed the STR mutation patterns. To obtain a single mutation-rate estimate for each Y-STR, we averaged the estimates from the SGDP and 1000 Genomes datasets. We found that the distribution of Y-STR mutation rates had a substantial right tail, such that most STRs mutated at very slow rates and only a few loci mutated at high rates (Figure 4). On average, a polymorphic Y-STR mutates at a rate of 3.8×10^{-4} mpg and has a median mutation rate of 8.7×10^{-5} mpg. The average Y-STR mutation rate is an order of magnitude lower than previous estimates from panel-based studies. This difference cannot be explained by our phylogenetic measurement procedure given that inspection of the same markers yielded relatively concordant numbers. Instead, it most likely stems from the ascertainment strategy of STR panels, which select highly diverse loci that do not reflect the mutation rates of most STRs. One caveat in this analysis is that very long Y-STR markers were not accessible to Illumina reads. These loci might affect the calculated average mutation rate and, to a smaller extent, the median mutation rate. Consistent with these explanations, our mutation-rate estimates for previously characterized loci were upwardly enriched in relation to our estimates for all markers (Figure 4).

Leveraging our catalog of Y-STR mutation rates, we searched for loci with relatively high mutation rates. These loci help to distinguish Y chromosomes of highly related individuals and can help to precisely date patrilineal relatedness among individuals, which is important for forensics and genetic genealogy. Most of the markers with the greatest estimated mutation rates have been characterized in prior studies (Table 1), but we identified six loci whose mutation rates were estimated to be greater than ~2×10^{-3} mpg and are yet to be reported (Tables 2 and 3). Two of these markers, DYS548 and DYS467, have been used in previous genealogical panels, but to the best of our knowledge, their mutation rates were never reported. In addition, we identified more than 65 loci with dinucleotide motifs and mutation rates greater than ~3.33×10^{-4} mpg (Table 3; Table S1).

We observed wide variability in the mutation rates and patterns between motif length classes. STRs with

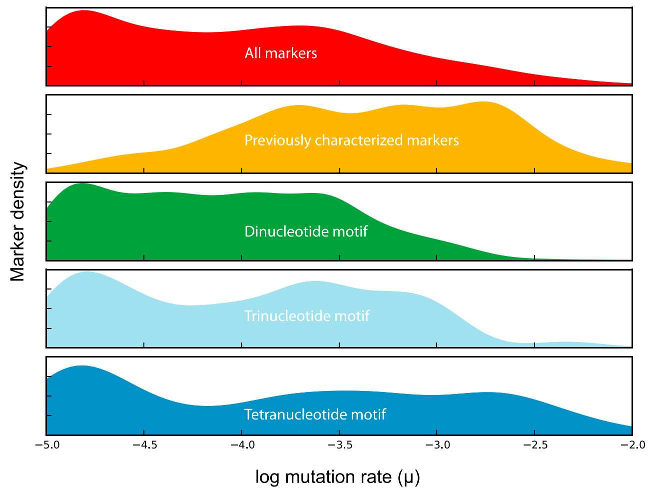

Figure 4. Distribution of Y-STR Mutation Rates
In red, we show the distribution of mutation rates across all STRs in this study. The set of loci with previously characterized mutation rates (orange) is substantially enriched with more-mutable loci. When stratified by motif length, loci with tetranucleotide motifs (dark blue) are the most mutable and are followed by loci with trinucleotide (light blue) and dinucleotide (green) motifs.

tetranucleotide motifs had the greatest median mutation rate ($\mu = 1.76 \times 10^{-4}$ mpg) and were followed by those with trinucleotide ($\mu = 1.22 \times 10^{-4}$ mpg), pentanucleotide ($\mu = 1.19 \times 10^{-4}$ mpg), dinucleotide ($\mu = 7.7 \times 10^{-5}$ mpg), and hexanucleotide motifs ($\mu = 3.28 \times 10^{-5}$ mpg) (Figure 4). However, within each motif class, mutation rates varied by two or more orders of magnitude, indicating that other factors contribute to STR variability and highlighting that aggregate mutation-rate statistics depend on the set of loci under consideration. We also found marked differences in the mutation patterns between motif classes. Loci with dinucleotide motifs and mutation rates greater than 10^{-4} mpg had a median step-size parameter of $\rho_M = 0.8$, implying that many of the de novo mutations are expected to be greater than one repeat unit. On the other hand, the median step-size parameter for longer motif classes within this mutation-rate range was closer to 1, implying that nearly all de novo events involve single-step mutations.

Next, we harnessed the large number of Y-STR mutation-rate estimates to identify the sequence determinants of mutation rates. For STRs without repeat-structure interruptions, the length of the major allele explains a substantial fraction of the variance in log mutation rates for loci with di-, tri-, and tetranucleotide motifs ($R^2 = 0.83$, $R^2 = 0.67$, and $R^2 = 0.82$, respectively; pentanucleotide motifs were not assessed because of a small number of data points). However, when we analyzed all STRs, including those with interruptions, the length of the major allele was a poor predictor and explained only a modest amount of the variance ($R^2 = 0.16$, $R^2 = 0.25$, and $R^2 = 0.42$; Figure 5, left panels). To construct an improved model, we analyzed the relationship between the log mutation rate and the length of the longest uninterrupted repeat tract, regardless of the number of interruptions (Figure 5, right panels). This model explained more than 75% of the variance in mutability for each of the three motif length classes. To assess the impact of the repeat motif on the mutation rate, we stratified loci with dinucleotide motifs by repeat sequence (AC, AG, or AT) and once again regressed the log mutation rate on the length of either the major allele or the longest uninterrupted tract (Figure S9). Major-allele length was again a relatively poor predictor of the log mutation rate, but uninterrupted-tract length explained more than 80% of the variance for each motif. Although these motif-specific models improved the R^2, the increase was quite limited, suggesting that conditioned on the uninterrupted-tract length, the repeat motif itself plays a minor role in the mutation rate. Taken together, our results show that a simple model of motif size and the length of the longest uninterrupted tract largely explains STR mutation rates.

Predicting Genome-wide STR Mutation Rates

Using the determinants found above, we estimated the number of de novo mutations across the entire genome. For each repeat-motif length, we trained a non-linear mutation-rate predictor by using the uninterrupted-tract lengths and mutation rates of the polymorphic Y-STRs. To account for the fixed STRs in our dataset and to better fit the model at shorter tract lengths, we assigned each fixed locus a mutation rate of 10^{-5} mpg, the lower mutation-rate boundary used by MUTEA (Figure S10), and we jointly trained the predictors across all STRs. To validate these predictors, we used them to estimate the mutation rates of paternally transmitted autosomal CODIS markers, which the National Institute of Standards and Technology (NIST) had previously estimated via conventional means. Our predictors explained about 75% of the variance in the log mutation rates for these markers. In addition, the median mutation rate reported by NIST ($\mu = 1.3 \times 10^{-3}$ mpg) closely matched the result reported by our predictors ($\mu = 1.0 \times 10^{-3}$ mpg), suggesting that they generate reliable predictions.

Next, we ran our predictors on each STR in the human genome with 2–4 bp motifs, resulting in mutation-rate estimates for each of the ~590,000 loci (Table S2). Because our model was trained with Y-STR mutation rates, these estimates refer only to the paternally inherited half of the genome. We discarded estimated rates below 1.25×10^{-5} mpg, because these are too close to the MUTEA lower boundary and might therefore be upwardly biased. After filtering, our model predicted that there are ~70,000 STRs with mutation rates greater than 10^{-4} mpg and ~44,000 loci with mutation rates greater than 1 in 3,000 mpg and that an STR should mutate at an average rate of 4.4×10^{-4} mpg. Stratifying our results by motif length, we predict 29, 3, and 33 de novo STR mutations for loci with di-, tri-, and tetranucleotide motifs, respectively, on the paternally inherited set of chromosomes.

Table 1. The Most Mutable Y-STRs with Previously Characterized Mutation Rates

Chr	hg19 Start	hg19 End	Motif	Mutation Rate (mpg)	Homogeneous-Tract Length (bp)	Name
Y	7,053,359	7,053,426	AAAG	1.37×10^{-2}	68	DYS576
Y	7,867,880	7,867,943	AAAG	9.20×10^{-3}	64	DYS458
Y	6,861,231	6,861,298	AAAG	7.80×10^{-3}	72	DYS570
Y	14,515,312	14,515,363	AGAT	5.08×10^{-3}	48	DYS439
Y	8,426,378	8,426,443	AAG	4.67×10^{-3}	69	DYS481
Y	21,520,224	21,520,275	AGAT	4.50×10^{-3}	48	DYS549
Y	18,718,889	18,718,940	AGAT	4.20×10^{-3}	52	Y-GATA-A10
Y	4,270,960	4,271,019	AGAT	3.77×10^{-3}	60	DYS456
Y	19,372,273	19,372,328	AGAT	2.88×10^{-3}	48	DYS543
Y	14,761,101	14,761,160	AGAT	2.65×10^{-3}	46	DYS442

The following abbreviation is used: Chr, chromosome.

Overall, we predict that 76–85 de novo STR mutations occur each generation for the full set of chromosomes. To account for the maternal chromosomes, we extrapolated our paternal results by using prior estimates of the male-to-female STR mutation-rate ratio (3.3:1 to 5.5:1[19,51]). We posit that our estimates for STR de novo mutational load are likely to be conservative. First, we omitted loci with 5–6 bp motifs for which we did not have sufficient data to build a mutation-rate model. Second, for autosomal STRs whose uninterrupted-tract lengths exceeded the maximal length observed in our study, we estimated their mutation rates by using the maximal Y-STR length. Given the strong positive correlation between tract length and mutation rate observed in our study, these loci are probably far more mutable. Despite our conservative approach, the estimated number of genome-wide de novo STR mutations rivals that of any known class of genetic variation, including SNPs (~70 events per generation), indels (one to three events), and SV and interspersed repeats (less than one event per generation).[6,7,9,52] As such, our results highlight the putative contribution of STRs to de novo genetic variation.

Y-STRs in Forensics and Genetic Genealogy

We assessed the applicability of our Y-STR results to the genetic genealogy and forensic DNA communities. First, we considered whether it would be possible to distinguish between closely patrilineally related individuals from high-throughput sequencing data. On the basis of the entire Y-STR set reported by our study, we expect roughly one de novo mutation to occur every four generations. In addition, from WGS data, one also expects to identify approximately one de novo SNP every 2.85 generations,[35] resulting in a 60% theoretical probability of differentiating between a father and son's Y chromosome haplotype by high-throughput sequencing. Previous studies have suggested that capillary genotyping of 13 rapidly mutating Y-STRs can discriminate between father-son pairs in 20%–27% of cases.[21,53] However, these particular markers are largely inaccessible to WGS data because of their long lengths and highly repetitive flanking regions, which preclude unique mapping. With increased interest in high-throughput sequencing among genetic genealogy services (e.g., FullGenomes and Big Y by FamilyTreeDNA) and the forensics community, our results suggest that WGS can achieve better patrilineal discrimination than common panel-based methods. Of course, the main caveat is that WGS technology is at least an order of magnitude more expensive than a panel-based approach. However, if the current trajectory of declining sequencing costs continues, shotgun sequencing to discriminate between closely patrilineally related individuals might soon become economically viable.

We also assessed the accuracy of imputing Y-STR profiles from Y-SNP data. This capability could be useful in forensic cases involving a highly degraded male sample, from which it would be difficult to obtain complete Y-STR profiles. In such cases, because there are many more SNPs than STRs on the Y chromosome, it might be possible to salvage some of those markers with a high-throughput method and impute Y-STRs profiles for compatibility with common forensic or genealogical databases.

For imputation, we created a framework called MUTEA-IMPUTE. In brief, after building a SNP phylogeny relating all samples and learning a mutation model as outlined in Figure 1, MUTEA-IMPUTE passes two sets of messages along the phylogeny to compute the exact marginal posteriors for each node, resulting in imputation probabilities for samples without observed Y-STR genotypes (Appendix D). We assessed the accuracy of our algorithm by imputing the 1000 Genomes individuals for the PowerPlex Y23 panel, a set of markers regularly used in forensic cases involving sex crimes. Over 100 iterations, we randomly constructed reference panels of 500 samples and used MUTEA-IMPUTE to calculate the maximum a posteriori genotypes for a distinct set of 70 samples.

Table 2. The Most Mutable Y-STRs with Tetranucleotide Motifs and Previously Uncharacterized Mutation Rates

Chr	hg19 Start	hg19 End	Motif	Mutation Rate (mpg)	Homogeneous-Tract Length (bp)	Name
Y	14,612,456	14,612,520	AGAT	5.07×10^{-3}	59	DYS467
Y	5,409,729	5,409,801	AAAG	5.06×10^{-3}	61	NA
Y	19,500,594	19,500,656	AAAG	4.89×10^{-3}	63	NA
Y	14,200,743	14,200,802	AGAT	4.54×10^{-3}	56	NA
Y	21,665,702	21,665,764	AAAT	3.66×10^{-3}	50	DYS548

Abbreviations are as follows: Chr, chromosome; and NA, not available.

Despite the small size of the reference panel, we were able to correctly impute an average of 66% of the genotypes without any quality filtration (Table S3). Importantly, the resulting imputed probabilities roughly matched the average accuracy, indicating that the posteriors computed by this technique are well calibrated (Figure S11). Discarding imputed genotypes with posteriors below 70% resulted in an overall accuracy of 88% and retained about 40% of the calls. On a marker-by-marker basis, accuracy was generally inversely proportional to the estimated mutation rates, such that the most slowly mutating markers had accuracies on the order of 95%. This trend stems from the fact that as the mutation rate increases, obtaining an estimate with similar confidence requires shorter branch lengths. We envision that a larger panel will substantially increase the ability to correctly impute Y-STRs and might facilitate work with highly degraded samples, a common issue in forensic casework.

Discussion

Advances in sequencing technology have fundamentally altered Y-STR analyses. The initial scarcity of SNP genotypes led to the development of methods for inferring coalescent models from Y-STR genotypes alone. Methods designed to also learn STR mutational dynamics either marginalized over these coalescent models[54] or aimed to simultaneously infer the coalescent and mutational models.[55,56] With the advent of population-scale WGS datasets, many of these STR-centric approaches have instead used SNPs, resulting in substantially more detailed phylogenies. For the Y chromosome, these detailed phylogenies now provide the evolutionary context required for interpreting Y-STR mutations, obviating the need for computationally expensive tree enumeration or marginalization approaches. However, the errors prevalent in WGS-based Y-STR genotypes require methods capable of accounting for genotype uncertainty, precluding the application of many traditional microsatellite distance measures designed for capillary data.[45,46]

In this study, we developed MUTEA, a method that leverages population-scale sequencing data to estimate Y-STR mutation rates. One inherent advantage of our approach is its ability to model and learn many of the salient features of microsatellite mutations. By incorporating a geometric step-size distribution, we allow both single-step mutations that predominate at tetranucleotide loci[19,57] and multistep mutations that frequently occur at dinucleotide loci.[19,58] In addition, the model's length-constraint parameter captures the intra-locus phenomenon of shorter STR alleles preferentially expanding and longer alleles preferentially contracting.[58,59] Because these parameters are learned from observed STR genotypes, our method avoids many biases that stem from imposing single-step mutations or assuming parameters a priori.

In addition to having a flexible mutation-model framework, our approach has both high throughput and a high dynamic range. With WGS data, we were able to assess every Y-STR that is accessible to Illumina sequencing, dramatically increasing the catalog of polymorphic loci with estimated mutation rates. In addition, by leveraging deep Y chromosome phylogenies, we were able to obtain mutation-rate estimates for very slowly mutating loci. Our estimates were highly replicable and consistent, as demonstrated by the strong concordance between the estimates from the two WGS datasets.

Our approach has several inherent limitations. Because Illumina datasets are currently composed of 75–100 bp reads, we were unable to genotype and characterize the mutation rates of both long Y-STRs and Y-STRs that reside in heterochromatic regions. Because of the strong relationship between tract length and mutation rate, we anticipate that more rapidly mutating loci reside on the Y chromosome. In addition, we were unable to characterize the mutation rates of homopolymers because base quality scores degraded rapidly as allele length increased. As a result, future studies might benefit from reapplying our analyses as sequencing technologies, particularly those enabling longer reads, continue to mature. Another limitation is that our mutation model does not capture the full complexity of STR mutational dynamics, given that it ignores intra-locus mutation-rate variation.[60] Incorporating these and other mutational characteristics might be of interest to future studies.

One longstanding question regarding Y-STR mutation rates has been the apparent discrepancy between

Table 3. The Most Mutable Y-STRs with Dinucleotide Motifs and Previously Uncharacterized Mutation Rates

Chr	hg19 Start	hg19 End	Motif	Mutation Rate (mpg)	Homogeneous-Tract Length (bp)	Name
Y	2,807,025	2,807,064	AT	3.62×10^{-3}	44	NA
Y	2,708,412	2,708,457	AG	1.75×10^{-3}	46	NA
Y	3,832,234	3,832,278	AC	1.66×10^{-3}	45	NA
Y	6,398,638	6,398,684	AC	1.62×10^{-3}	49	NA
Y	17,109,092	17,109,141	AC	1.57×10^{-3}	48	NA

Abbreviations are as follows: Chr, chromosome; and NA, not available.

evolutionary and pedigree-based mutation rates. Several studies have suggested that evolutionary rates are three to four times lower, resulting in substantial inconsistencies in Y-STR-based lineage dating and large discrepancies from Y-SNP-based TMRCA estimates.[20,47,61] Because our study harnessed evolutionary data, we sought to avoid any potential issues by scaling each phylogeny such that our estimates best matched those from pedigree-based studies. Nonetheless, our investigations into an alternative scaling based on a SNP molecular clock resulted in similar scaling factors that only differed by ~25%. Coupled with the strong concordance we observed with pedigree-based estimates, our study provides little evidence for a substantial difference between mutation rates estimated from these two types of data. Future work might benefit from assessing whether these previously reported discrepancies were due to the simplified Y-STR mutation models that the approaches used to obtain evolutionary-based Y-STR mutation rates.

Our large corpus of mutation-rate estimates has enabled us to dissect the sequence factors governing STR mutability. We determined that the length of the longest uninterrupted tract is a strong predictor of the log mutation rate. This observation matches the exponential relationship between mutation rate and tract length previously reported in several pedigree-based studies.[21,51,57,59] We also found that the total length of the major allele was a poor predictor. Coupled with the fact that Y-STRs without interruptions were much more mutable than interrupted ones with the same major-allele length, our study provides strong evidence that interruptions to the repeat structure decrease mutation rates. This finding supports what has long been posited in STR evolutionary models[62,63] and has been shown in a handful of small-scale experimental studies of STR mutability.[64,65] However, it contradicts the recent findings of Ballantyne et al., who observed no effect.[21]

Another open question is why STRs with dinucleotide motifs have lower mutation rates, given their higher levels of polymorphisms in the population. A previous large-scale panel-based study reported that loci with dinucleotide motifs have lower mutation rates than do loci with tetranucleotide motifs.[19] Our survey confirmed this observation without STR ascertainment directly based on their polymorphism rates. However, genome-wide analyses of STRs have shown that dinucleotides have more diverse allelic spectra than do tetranucleotides.[23,26] These results are unlikely to be due to genotyping errors given that a study of an individual sequenced to a depth of 120× also showed that dinucleotide repeats are more polymorphic than other types of STRs.[23] One potential explanation is that STRs with dinucleotide motifs have larger step sizes but lower mutation rates. However, we cannot exclude other explanations, such as a difference in length constraint.

Our large compendium of mutation-rate estimates has also enabled predictions about genome-wide STR variation. Prior studies have estimated a rate of approximately 75 de novo mutations per generation[4,8] but have largely ignored STRs, despite their elevated mutation rates. On the basis of our projections for paternally inherited chromosomes, the number of de novo STR mutations is likely to rival the combined contribution of all other types of genetic variants. Given that several lines of evidence have highlighted the involvement of STR variations in complex traits,[11–13,66] it will be important to assess the biological impact of these de novo STR variations on human phenotypes.

Appendix A: Estimating CIs

We used a delete-d jackknife approach to estimate mutation-rate CIs.[67] For each Y-STR, we sampled without replacement half of the STR genotypes a total of 100 times and estimated the log mutation rate by using each of these subsets. Given these subsample estimates and the log estimate obtained from all samples, the SE and CI for the log mutation rate were calculated as follows:

$$\mathrm{SE} = \sqrt{\frac{1}{100}\sum_{i=1}^{100}\left(\log \mu_i - \frac{1}{100}\sum_{j=1}^{100}\log \mu_j\right)^2},$$
$$\mathrm{CI} = \log \mu_{\mathrm{tot}} \pm 1.96 * \mathrm{SE},$$

where μ_{tot} is the estimate based on the full dataset.

Appendix B: Simulating Exact STR Genotypes

We used values of μ, β, and ρ_M ranging from 10^{-5} to 10^{-2}, 0 to 0.5, and 0.75 to 1.0, respectively, to simulate genotypes under a wide range of mutation models. Using either

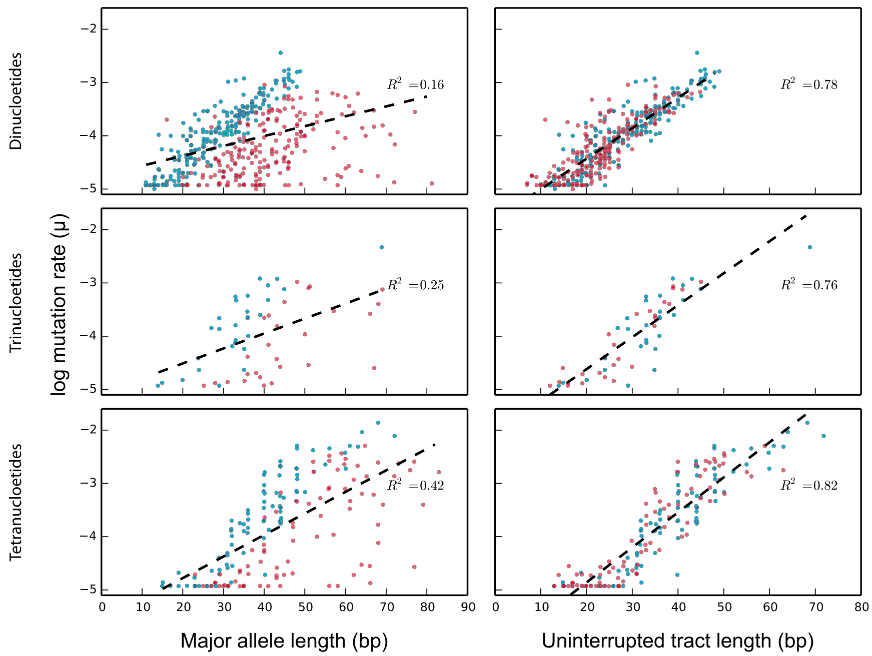

Figure 5. Sequence Determinants of Y-STR Mutability
Each panel plots the estimated log mutation rates (y axis) of STRs against either the major-allele length (x axis, left panels) or the length of the longest uninterrupted tract (x axis, right panels) for various sizes of repeat motifs (rows). The black lines represent the mutation rate predicted by a simple linear model. For a given allele length (left panels), Y-STRs with no interruptions to the repeat structure (blue) are generally more mutable than those with one or more interruptions (red). Whereas major-allele length alone is poorly correlated with mutation rate (left panels), the length of the longest uninterrupted tract (right panels) is strongly correlated regardless of the number of interruptions.

the 1000 Genomes phylogeny or the SGDP phylogeny, we performed each simulation as follows:

1. Randomly assign the root node an STR allele between -4 and 4, and mark it as active.
2. Remove an active node, and mark it as inactive. For each of this node's children, do the following:
   - i. Calculate the child's allele probabilities by using the branch length, the true mutation model, and the parent node's genotype.
   - ii. Randomly select an STR allele on the basis of these probabilities.
   - iii. Mark the descendant node as active.
3. While active nodes remain, go to step 2.
4. Report the exact STR alleles for a random subset of the samples (leaf nodes) on the basis of the required sample size.

Appendix C: Simulating STR Sizes in Reads with PCR Stutter

We first used the procedure above to simulate STR genotypes down the phylogeny. We then used the true genotype for a particular sample g_i and a given stutter model to simulate the STR sizes observed in each read as follows:

1. Sample the number of observed reads $n_{\text{reads},\,i}$ for each sample with genotype g_i from the read-count distribution.
2. For each read from 1 through $n_{\text{reads},\,i}$, sample a number c ~U (0,1).
   - i. If $c < d$, randomly sample an artifact size a_j from a geometric distribution with parameter ρ_s. Report the read's STR size as $g_i - a_j$.
   - ii. If $d \le c < 1 - u$, report the read's STR size as g_i.
   - iii. Otherwise, randomly sample an artifact size a_j from a geometric distribution with parameter ρ_s. Report the read's STR size as $g_i + a_j$.

To assess whether estimates would be accurate for even the most sparsely sequenced loci, we used read-count distributions obtained from both Y-STR call sets corresponding to loci in the tenth coverage percentile. For Figure 2, we used a stutter model with $d = 0.15$, $u = 0.01$, and $\rho_s = 0.8$, and we used one, two, and three reads for 65%, 25%, and 10% of samples, respectively.

Appendix D: Y-STR Imputation

We extended MUTEA to impute missing STR genotypes. Using the approach outlined in Figure 1, we first construct a phylogeny relating all samples and learn a mutation model. Then, we use this learned mutation model to pass two sets of messages along the tree and compute exact posteriors for each node, resulting in imputation probabilities for samples with missing genotypes. For node N_i with parent P_i, sibling S_i, and children C_{1i} and C_{2i}, its conditional genotype probability given the observed data D is

$$\begin{aligned} P(N_i \mid D) &= P\left(N_i \mid D_{C_{1i}}, D_{C_{2i}}, D_{-N_i}\right) \\ &= P\left(N_i, D_{C_{1i}}, D_{C_{2i}} \mid D_{-N_i}\right) / P\left(D_{C_{1i}}, D_{C_{2i}} \mid D_{-N_i}\right) \\ &= P\left(N_i \mid D_{-N_i}\right) P\left(D_{C_{1i}}, D_{C_{2i}} \mid N_i, D_{-N_i}\right) / P\left(D_{C_{1i}}, D_{C_{2i}} \mid D_{-N_i}\right) \\ &\propto P\left(N_i \mid D_{-N_i}\right) P\left(D_{C_{1i}} \mid N_i\right) P\left(D_{C_{2i}} \mid N_i\right). \end{aligned}$$

Here, D_{N_i} and D_{-N_i} denote the genotype likelihoods in and not in node N_i's subtree, respectively. We note that each of these terms is conditioned on the STR mutational model M and the Y chromosome phylogeny T, but we have omitted these terms here and below for brevity.

The second and third terms in the node posterior expression are computed with a bottom-up traversal of the tree from the leaves to the root node. Each node in the tree combines information from its two children by using the recurrence

$$
\begin{aligned}
P\left(D_{C_{i1}} \mid N_i\right) &= \sum\nolimits_{a \in \text{ alleles}} P\left(D_{C_{1i}}, C_{1i} = a \mid N_i\right) \\
&= \sum\nolimits_{a} P\left(D_{GC_{1i}}, D_{GC_{2i}}, C_{i1} = a \mid N_i\right) \\
&= \sum\nolimits_{a} P(C_{1i} = a \mid N_i) P\left(D_{GC_{1i}} \mid C_{1i} = a\right) \\
&\quad \times P\left(D_{GC_{2i}} \mid C_{1i} = a\right).
\end{aligned}
$$

Here, GC_{1i} and GC_{2i} denote the two children of node C_{1i}. This recurrence applies to all nodes except the leaves, where genotype posteriors or a uniform prior are used for samples with and without genotype information, respectively.

Similarly, the first term in the node posterior expression is computed with a top-down traversal of the tree from the root to the leaves. After the root node is assigned a uniform prior probability, each node combines information from its parent and sibling:

$$
\begin{aligned}
P\left(N_i \mid D_{-N_i}\right) &= \sum\nolimits_{a \in \text{ alleles}} P\left(N_i, P_i = a \mid D_{S_i}, D_{-P_i}\right) \\
&= \sum\nolimits_{a} P\left(N_i, P_i = a, D_{S_i} \mid D_{-P_i}\right) / P\left(D_{S_i} \mid D_{-P_i}\right) \\
&= \sum\nolimits_{a} P\left(P_i = a \mid D_{-P_i}\right) P\left(D_{S_i} \mid P_i = a, D_{-P_i}\right) \\
&\quad \times P\left(N_i \mid P_i = a, D_{S_i} D_{-P_i}\right) / P\left(D_{S_i} \mid D_{-P_i}\right) \\
&\propto \sum\nolimits_{a} P\left(P_i = a \mid D_{-P_i}\right) P\left(D_{S_i} \mid P_i = a\right) P\left(N_i \mid P_i = a\right).
\end{aligned}
$$

Supplemental Data

Supplemental Data include 11 figures and 3 tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.04.001.

Acknowledgments

M.G. was supported by the National Defense Science and Engineering Graduate Fellowship. G.D.P. was supported by the National Science Foundation Graduate Research Fellowship under grant DGE-1147470. C.T.-S. was supported by Wellcome Trust grant 098051. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by National Institute of Justice grant 2014-DN-BX-K089 (to Y.E. and T.W.). Y.E. is a scientific advisory board member of Identity Genomics, BigDataBio, and Solve Inc. G.D.P is an employee of 23andMe. None of these entities played a role in the design, execution, interpretation, or presentation of this study.

Received: February 13, 2016
Accepted: April 1, 2016
Published: April 25, 2016

Web Resources

1000 Genomes Project BAM alignments, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/

1000 Genomes Project capillary genotypes, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140107_chrY_str_haplotypes/YSTRs_PowerPLexY23_1000Y_QA_20130107.txt

Dendroscope, http://dendroscope.org/

HipSTR, https://github.com/tfwillems/HipSTR

MUTEA, https://github.com/tfwillems/MUTEA

RAxML, http://sco.h-its.org/exelixis/web/software/raxml/index.html

Simons Genome Diversity Project, https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/

Simons Genome Diversity Project capillary genotypes, ftp://ftp.cephb.fr/hgdp_supp9/genotype-supp9.txt

Y-STR references, HipSTR call sets, and Y-SNP phylogenies, https://github.com/tfwillems/ystr-mut-rates

References

1. Scally, A., and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. *13*, 745–753.
2. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
3. Kayser, M., and de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. Nat. Rev. Genet. *12*, 179–192.
4. Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al.; 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.
5. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636–639.
6. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.
7. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B., et al.; UK10K Consortium (2016). Timing, rates and spectra of human germline mutation. Nat. Genet. *48*, 126–133.
8. Francioli, L.C., Polak, P.P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C.M., Swertz, M., Wijmenga, C., van Ommen, G., et al.; Genome of the Netherlands Consortium (2015). Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. *47*, 822–826.
9. Itsara, A., Wu, H., Smith, J.D., Nickerson, D.A., Romieu, I., London, S.J., and Eichler, E.E. (2010). De novo rates and selection of large copy number variation. Genome Res. *20*, 1469–1481.
10. Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.
11. Contente, A., Dittmer, A., Koch, M.C., Roth, J., and Dobbelstein, M. (2002). A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat. Genet. *30*, 315–320.

12. Gebhardt, F., Zänker, K.S., and Brandt, B. (1999). Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. J. Biol. Chem. *274*, 13176–13180.
13. Shimajiri, S., Arima, N., Tanimoto, A., Murata, Y., Hamada, T., Wang, K.Y., and Sasaguri, Y. (1999). Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett. *455*, 70–74.
14. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K.J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. Science *324*, 1213–1216.
15. Sureshkumar, S., Todesco, M., Schneeberger, K., Harilal, R., Balasubramanian, S., and Weigel, D. (2009). A genetic defect caused by a triplet repeat expansion in Arabidopsis thaliana. Science *323*, 1060–1063.
16. Weiser, J.N., Love, J.M., and Moxon, E.R. (1989). The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. Cell *59*, 657–665.
17. Weber, J.L., and Wong, C. (1993). Mutation of human short tandem repeats. Hum. Mol. Genet. *2*, 1123–1128.
18. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.
19. Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., and Stefansson, K. (2012). A direct characterization of human mutation based on microsatellites. Nat. Genet. *44*, 1161–1165.
20. Zhivotovsky, L.A., Underhill, P.A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., et al. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am. J. Hum. Genet. *74*, 50–61.
21. Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., et al. (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am. J. Hum. Genet. *87*, 341–353.
22. Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., and de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. *6*, 799–803.
23. Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. *22*, 1154–1162.
24. Highnam, G., Franck, C., Martin, A., Stephens, C., Puthige, A., and Mittelman, D. (2013). Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res. *41*, e32.
25. Warshauer, D.H., Lin, D., Hari, K., Jain, R., Davis, C., Larue, B., King, J.L., and Budowle, B. (2013). STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. Forensic Sci. Int. Genet. *7*, 409–417.
26. Willems, T., Gymrek, M., Highnam, G., Mittelman, D., and Erlich, Y.; 1000 Genomes Project Consortium (2014). The landscape of human STR variation. Genome Res. *24*, 1894–1904.
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
28. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.
29. Gymrek, M. (2016). PCR-free library preparation greatly reduces stutter noise at short tandem repeats. bioRxiv, http://dx.doi.org/10.1101/043448.
30. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
31. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
32. Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. *61*, 1061–1067.
33. Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., Wilson Sayres, M.A., Ayub, Q., McCarthy, S.A., Narechania, A., Kashin, S., et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat. Genet. http://dx.doi.org/10.1038/ng.3559.
34. Helgason, A., Einarsson, A.W., Guðmundsdóttir, V.B., Sigurðsson, Á., Gunnarsdóttir, E.D., Jagadeesan, A., Ebenesersdóttir, S.S., Kong, A., and Stefánsson, K. (2015). The Y-chromosome point mutation rate in humans. Nat. Genet. *47*, 453–457.
35. Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., et al.; Asan (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. *19*, 1453–1457.
36. Willuweit, S., and Roewer, L.; International Forensic Y Chromosome User Group (2007). Y chromosome haplotype reference database (YHRD): update. Forensic Sci. Int. Genet. *1*, 83–87.
37. Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.
38. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.
39. Burgarella, C., and Navascués, M. (2011). Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. Eur. J. Hum. Genet. *19*, 70–75.
40. Hanson, E.K., and Ballantyne, J. (2006). Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. Leg Med (Tokyo) *8*, 110–120.
41. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:13033997.
42. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
43. Vermeulen, M., Wollstein, A., van der Gaag, K., Lao, O., Xue, Y., Wang, Q., Roewer, L., Knoblauch, H., Tyler-Smith, C., de Knijff, P., and Kayser, M. (2009). Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. Forensic Sci. Int. Genet. *3*, 205–213.
44. Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S.M., Santos, L.H., Anslinger, K., Bayer, B., et al.

(2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Sci. Int. Genet. *12*, 12–23.
45. Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics *139*, 457–462.
46. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., and Feldman, M.W. (1995). An evaluation of genetic distances for use with microsatellite loci. Genetics *139*, 463–471.
47. Zhivotovsky, L.A., Underhill, P.A., and Feldman, M.W. (2006). Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. Mol. Biol. Evol. *23*, 2268–2270.
48. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. *17*, 368–376.
49. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Series B Stat. Methodol. *39*, 1–38.
50. Nelder, J.A., and Mead, R. (1965). A Simplex Method for Function Minimization. Comput. J. *7*, 308–313.
51. Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. *62*, 1408–1415.
52. Kloosterman, W.P., Francioli, L.C., Hormozdiari, F., Marschall, T., Hehir-Kwa, J.Y., Abdellaoui, A., Lameijer, E.W., Moed, M.H., Koval, V., Renkens, I., et al.; Genome of Netherlands Consortium (2015). Characteristics of de novo structural changes in the human genome. Genome Res. *25*, 792–801.
53. Ballantyne, K.N., Ralf, A., Aboukhalid, R., Achakzai, N.M., Anjos, M.J., Ayub, Q., Balažic, J., Ballantyne, J., Ballard, D.J., Berger, B., et al. (2014). Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. Hum. Mutat. *35*, 1021–1032.
54. Nielsen, R. (1997). A likelihood approach to populations samples of microsatellite alleles. Genetics *146*, 711–716.
55. Wilson, I.J., and Balding, D.J. (1998). Genealogical inference from microsatellite data. Genetics *150*, 499–510.
56. Wilson, I.J., Weale, M.E., and Balding, D.J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A Stat. Soc. *166*, 155–188.
57. Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Krüger, C., Krawczak, M., Nagy, M., Dobosz, T., et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am. J. Hum. Genet. *66*, 1580–1588.
58. Huang, Q.Y., Xu, F.H., Shen, H., Deng, H.Y., Liu, Y.J., Liu, Y.Z., Li, J.L., Recker, R.R., and Deng, H.W. (2002). Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. *70*, 625–634.
59. Xu, X., Peng, M., and Fang, Z. (2000). The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. *24*, 396–399.
60. Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. Nat. Genet. *24*, 400–402.
61. Wei, W., Ayub, Q., Xue, Y., and Tyler-Smith, C. (2013). A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci. Int. Genet. *7*, 568–572.
62. Kruglyak, S., Durrett, R.T., Schug, M.D., and Aquadro, C.F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA *95*, 10774–10778.
63. Sainudiin, R., Durrett, R.T., Aquadro, C.F., and Nielsen, R. (2004). Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics *168*, 383–395.
64. Petes, T.D., Greenwell, P.W., and Dominska, M. (1997). Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. Genetics *146*, 491–498.
65. Bacon, A.L., Farrington, S.M., and Dunlop, M.G. (2000). Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. Hum. Mol. Genet. *9*, 2707–2713.
66. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.
67. Shao, J., and Wu, C.F.J. (1989). A General Theory for Jackknife Variance Estimation. Ann. Stat. *17*, 1176–1197.