bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chromosome-wide characterization of Y-STR mutation rates using ultra-deep genealogies

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

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts, USA

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

[3]New York Genome Center, New York, NY, USA

[4]Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, USA.

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

[6]Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA

[7]Department of Genetics, Stanford University, Stanford, CA 94305, USA.

[8]The 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, USA

* To whom correspondence should be address (yaniv@cs.colubmia.edu)

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Although the utility of short tandem repeats on the Y-chromosome (Y-STRs) has long been recognized and leveraged in forensics, genealogy and paternity testing, the bulk of these applications have relied on only a few dozen loci identified as having remarkably high mutation rates. Recent efforts have expanded the set of Y-STRs with known mutation rates to two hundred markers, but the limited throughput of the capillary method for estimating mutation rates has left the mutability of most Y-STRs uncharacterized, particularly those with dinucleotide repeat units. To address this limitation, we developed a novel method capable of concurrently estimating the mutation rates of all Y-STRs by leveraging population-scale whole-genome sequencing data. Extensive simulations confirmed that our method robustly accounts for PCR stutter artifacts and obtains unbiased mutation rate estimates. Application of the method to orthogonal datasets from the 1000 Genomes Project and Simons Genome Diversity Project utilized evolutionary data from over 250,000 meioses to estimate the mutation rates of more than 700 Y-STRs with 2-6 base pair repeat units, yielding the largest such set to date. Comparison of these estimates with those from father-son studies indicated a high degree of concordance for loci that have been previously characterized. In addition, we identified nearly 100 previously uncharacterized Y-STRs with per-generation mutation rates greater than 1 in 3000. Altogether, our study provides a broadly applicable method for estimating Y-STR mutation rates from whole-genome sequencing cohorts, outlines a framework for imputing Y-STRs, vastly expands the number of identified loci with high discriminative power and provides the first chromosome-wide characterization of the mutation rates of dinucleotide short tandem repeats.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Over the past 20 years, a multitude of fields have increasingly leveraged Y-STRs due to their unique combination of high mutation rate and paternal inheritance pattern. Prior to the advent of genome-wide SNP genotype data, population genetics utilized these highly mutable markers to build phylogenies (Takezaki and Nei 1996; Forster et al. 2000) and to draw a host of demographic inferences (Pritchard et al. 1999). In forensics, Y-STRs are commonly used to resolve cases in which DNA samples contain multiple donors or are difficult to profile using traditional autosomal techniques (Kayser et al. 1997; Roewer 2009). Y-STRs are also widely used in genealogy to ascertain the relatedness of families (Kayser et al. 2007) and in paternity cases, even resolving historical debates such as the contentious paternal relationship between Thomas Jefferson and Sally Hemings' children (Foster et al. 1998). More recently, we employed these markers to demonstrate that one can infer the surname of an anonymous genome, a finding that stimulated important conversations related to genetic privacy (Gymrek et al. 2013).

Despite the immense utility of Y-STRs, the vast majority of their applications rely on only a few dozen markers. The small size of this panel is largely the result of the cumbersome and expensive method used to estimate Y-STR mutation rates. This process typically involves genotyping large pedigrees or thousands of father-son pairs using capillary electrophoresis, from which the frequency of discordant genotypes provides an estimate of the mutation rate (Heyer et al. 1997; Kayser et al. 2000; Dupuy et al. 2004; Gusmao et al. 2005). Recently, several large-scale studies expanded the set of loci with available mutation rates to include several hundred markers and identified a handful of rapidly mutating Y-STRs with mutation rates in excess of 10^{-2} mpg (Ballantyne et al. 2010; Burgarella and Navascues 2011). However, these studies characterized long Y-STRs with 3-6bp motifs that were identified in prior scans for polymorphic loci (Kayser et al. 2004), ignoring loci with fewer repeats or dinucleotide repeats. Since genome-wide studies of human populations have identified dinucleotide repeats as among the most abundant and heterozygous of the STR classes (Willems et al. 2014), characterizing these markers may identify new promising candidates for male lineage differentiation. In addition, characterizing the mutation rates of the full spectrum of Y-STRs will be instrumental towards understanding their mutational mechanisms and developing accurate sequence-based predictors of mutability for use across the genome.

Fortunately, the rapid advancement of next-generation sequencing technologies has provided a unique opportunity to address these issues. Coupled with vast improvements in the depth and quality of whole genome sequencing (WGS) datasets, the advent of STR genotyping tools has made it possible to genotype Y-STRs chromosome-wide (Gymrek et al. 2012; Highnam et al. 2013; Warshauer et al. 2013). As a result, mutation rate estimation procedures that leverage these datasets can perform unbiased scans for mutable loci instead of only considering previously ascertained sites. While it may seem appropriate to apply traditional STR mutation rate estimators based on microsatellite distance measures to these datasets (Goldstein et al. 1995; Slatkin 1995), these estimators assume simplistic STR mutation models and have been shown to be susceptible to haplogroup size fluctuations (Zhivotovsky et al. 2006). An alternative approach is to develop methods that also leverage the rich evolutionary information of recently generated high-resolution population-scale Y-chromosome (Francalacci et al. 2013; Poznik et al. 2013; Wei et al. 2013a). A recent study developed one such method, but it also required a simple mutation model and only partially utilized the phylogenetic

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

information (Ravid-Amir and Rosset 2010). Furthermore, all of the above methods assume error-free genotypes and are therefore poorly equipped to deal with the sources of error prevalent in WGS-based STR call sets.

In this study, we demonstrate how to effectively integrate Y-STR genotypes and Y-SNP phylogenies derived from whole-genome sequencing data to estimate Y-STR mutation rates. Using various simulations, we demonstrate that our approach results in unbiased mutation rate estimates for almost all considered mutation models, even in the presence of extensive PCR stutter. We then apply our approach to data from the Simons Genome Diversity Project (SGDP) and 1000 Genomes Project (1KGP) (Genomes Project et al. 2015) to estimate the mutation rates of over 700 Y-STRs, most of which have never been characterized. The resulting sets of estimates were remarkably concordant, uncovered a large number of unknown highly polymorphic markers and shed light on the sequence factors that govern Y-STR mutability.

Materials and Methods

Mutation Rate Method Overview

Our approach to estimating Y-STR mutation rates, which is outlined in Figure 1, is motivated by the notion that current Y-SNP phylogenies are sufficiently detailed and accurate to infer STR mutation models. Given a phylogeny and a set of STR genotypes, Felsenstein's pruning algorithm (Felsenstein 1981) and numerical optimization can be used to evaluate and improve the likelihood of a mutation model until convergence, providing an estimate of the mutation rate. However, due to the error-prone and low-coverage nature of WGS-based STR call sets, utilizing these genotypes will result in vastly inflated mutation rate estimates. To avoid these biases, we analyze the number of repeats observed in all individuals' reads to learn a locus-specific error model and use this error model to compute genotype posteriors. As these posteriors account for genotype uncertainty, we utilize them during the mutation model optimization process instead of fixed genotypes to obtain robust estimates. More detailed descriptions of each of the steps involved in this approach are contained in the sections below.

Y-SNP Phylogeny Construction

We downloaded Y-chromosome SNP calls for male SGDP samples from the project website and utilized VCFtools (Danecek et al. 2011) to remove loci where more than 10% of the calls were heterozygous. For the remaining polymorphic sites, we removed individual calls that were heterozygous, had fewer than 7 supporting reads or had more than 10% of reads supporting an uncalled allele. Lastly, we removed loci if fewer than 150 samples met these criteria or more than 10% of reads had zero mapping quality. These filters resulted in nearly 39,000 high quality polymorphic SNPs which were used to generate a maximum likelihood phylogeny using RAxML (Stamatakis 2014) and the options *–m ASC_GTRGAMMA –f d -asc-corr lewis.* We used Dendroscope (Huson and Scornavacca 2012) to root the resulting phylogeny along the branch marked by the M42 and M94 mutations, well-annotated markers associated with the split of the A haplogroup from all other haplogroups (Jobling and Tyler-Smith 2003). For the 1000 Genomes dataset, we utilized a RAxML-generated phylogeny generated by the 1000 Genomes Y-chromosome working group.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

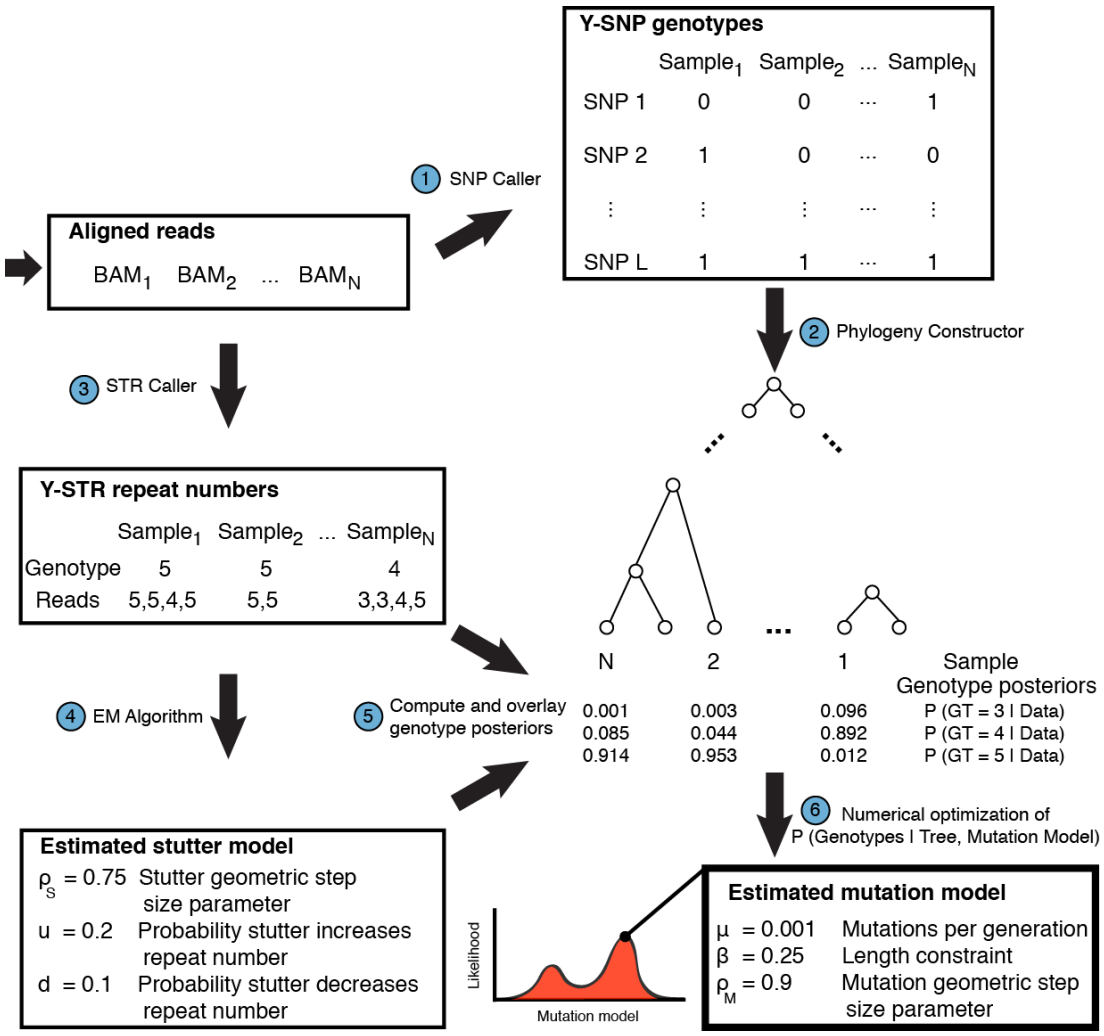

Figure 1: Y-STR mutation rate estimation method. Schematic of the steps required to estimate Y-STR mutation rates. The method first genotypes Y-SNPs (step 1) and utilizes these calls to build a single Y-SNP phylogeny for all Y-STRs (step 2). This phylogeny provides the evolutionary context required to infer the Y-STR mutational dynamics, with samples in the cohort lying on the leaves of the tree and all other nodes representing unobserved ancestors. Steps 3-6 are then run on each Y-STR individually. After utilizing an STR genotyping tool to determine 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 repeat counts, this model is used to compute each sample's genotype posteriors (step 5). After randomly initializing a mutation model, 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.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Modeling STR Genotyping Errors

PCR stutter artifacts are one of the primary causes of STR genotyping errors and typically involve the insertion or deletion of copies of the STR repeat unit in a subset of the reads at a locus. To mitigate the effects of these errors, we developed a method to learn locus-specific stutter models. Our stutter model Θ is parameterized by the allele frequencies for 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:

$$P(g_i = j|R,\Theta) \propto f_j \prod_{k=1}^{n_{reads,j}} \begin{cases} 1-u-d, & r_{k,i} = j \\ u\rho_s(1-\rho_s)^{r_{k,i}-j-1}, & r_{k,i} > j \\ d\rho_s(1-\rho_s)^{j-r_{k,i}-1}, & r_{k,i} < j \end{cases}$$

where g_i and $r_{k,i}$ denote the number of repeats in the locus and k[th] read for the i[th] individual, respectively. To learn these parameters, we employed an expectation-maximization framework in which the E-step computes the genotype posteriors for every sample under the observed read repeat counts and the current stutter model. The M-step then utilizes these posterior probabilities to update the stutter model parameters for *N* samples, *A* alleles and *Q* reads as follows:

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

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

In addition to PCR stutter, alignment errors may also cause reads to have a detected number of repeats that differ from their underlying genotype. As these errors are also incorporated when learning the stutter model, the stutter model accounts for the combined frequency of these errors and thereby generates robust posteriors.

STR Mutation Model

We modeled STR mutations using a length-dependent variant of a generalized stepwise mutation model. The model is characterized by a per-generation mutation rate μ, a geometric step size distribution with parameter ρ_m and a spring-like length constraint β that causes alleles to mutate back towards a central allele denoted as having zero length. Alleles in this model can have negative values as an allele's value merely indicates the deviation, in repeats, from the central allele. Given a starting allele a_t, the probability of observing allele a_{t+1} the following generation is:

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

where the fraction of mutations increasing or decreasing the size of the STR are $f_i = \min\left(1, \max\left(0, \frac{1-\beta\rho_m a_t}{2}\right)\right)$ and $f_d = 1 - f_i$. 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 represent STRs with stronger tendencies to mutate back towards the central allele.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation Model Likelihood

We utilized Felsenstein's pruning algorithm (Felsenstein 1981) to evaluate the likelihood of an STR mutation model. Given a model M, dataset D comprised of observed STR genotypes and a SNP-based phylogeny T with root node R, the likelihood is

$$P(D|M,T) = \sum_r P(R=r, D|M,T) = \sum_r P(R=r)P(D|R=r,M,T)$$

Due to the structure of the phylogeny, the conditional probability of the data D_{N_i} below each interior node N_i given the node's genotype can be expressed in terms of transition probabilities to each child node C_j and the conditional probability of the data D_{C_j} in its subtree:

$$P\left(D_{N_i} \middle| N_i = p, M, T\right) = \prod_{C_j \in child(N_i)} \sum_{b \in alleles} P(C_j = b, D_{C_j} | N_i = p, M, T)$$

$$= \prod_{C_j \in child(N_i)} \sum_{b \in alleles} P(C_j = b | N_i = p, M, T)\, P(D_{C_j} | C_j = b, M, T)$$

While descending the phylogeny, this recursive relation applies until a node with no children is encountered. These nodes represent an observed sample and the conditional probability of the data in its subtree is merely given by its genotype likelihoods. Therefore, the likelihood of a mutation model can be calculated using a post-order tree traversal in which one computes the genotype likelihoods of each observed genotype and the conditional probability of the data in each interior node's subtree given the node's genotype. The total data likelihood is then readily computed using the root node's conditional probabilities and a uniform prior. Because normalizing the genotype likelihoods of each sample does not affect the relative model likelihoods, one can use genotype posteriors calculated using a uniform prior interchangeably. In addition, to avoid numerical underflow issues, we compute the total log-likelihood of the data instead of the raw likelihood.

STR Transition Probabilities

To accelerate the computation of parent-to-child transition probabilities along each branch in the phylogeny, we devised a means of rapidly computing the STR transition probabilities across hundreds of generations. Given a mutation model M and a vector of allele probabilities $\overrightarrow{p\,(a_t)}$ for generation *t,* the probability of observing allele v in the next generation is

$$p(a_{t+1} = v \mid M) = \sum_{a \in [a_{min},\ a_{max}]} p(a_{t+1} = v | a_t = a, M)\, p(a_t = a) = \gamma_{v,M}^T \overrightarrow{p(a_t)}$$

To calculate the probability of observing each allele in the next generation, we construct an *N*-by-*N* transition matrix Γ, where *N* is the number of STR alleles, rows 1, 2…N correspond to $\gamma_{a_{min,M}}^T, \gamma_{a_{min+1,M}}^T \ldots \gamma_{a_{max,M}}^T$ and each column represents the transition probabilities from one allele to all other alleles. We modify this matrix such that the first and last columns have one non-zero entry along the diagonal to prevent the boundary states from mutating and provide an assessment of how frequently they are encountered. We also modify the first and last rows of the matrix so that they represent transition inequalities that result in normalized transition probabilities. Recursive application of the transition matrix then readily results in the allele probabilities after *M* generations:

$$\overrightarrow{p(a_{t+M})} = \Gamma \overrightarrow{\mathrm{p}(a_{t+M-1})} = \Gamma\Gamma \overrightarrow{\mathrm{p}(a_{t+M-2})} = \Gamma^M \overrightarrow{\mathrm{p}(a_t)}$$

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We balance the tradeoff between computation time and boundary state collisions by utilizing the smallest allele range such that the minimum and maximum observed STR alleles have less than a 10^{-5} probability of drifting into the boundary states when progressing from the root node to the deepest leaf node.

Numerical Optimization of the Mutation Model

Given STR genotypes for a locus of interest, we developed 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 relative to this reference point. It then 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 (Nelder and Mead 1965) and the outlined method for computing each model's likelihood, the numerical optimization method iteratively updates the mutation model's parameters until the likelihood converges. After repeating this procedure using three different random initializations to increase the probability of discovering a global optimum, it selects the optimized set of parameters with the highest total likelihood.

Simulating Exact STR Genotypes

Values of μ, β, and ρ_m ranging from 10^{-5}-10^{-2}, 0–0.75, and 0.6–1.0 were used to simulate genotypes under a host of different mutation models. Using either the 1KGP phylogeny or the SGDP phylogeny, each simulation was performed 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:
    i. Calculate the child's allele probabilities using the branch length, the true mutation model and the parent node's genotype
    ii. Randomly select an STR allele based on 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) based on the required sample size

Simulating STR Sizes in Reads with PCR Stutter

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

1. Sample the number of observed reads $n_{reads,i}$ for each sample with genotype g_i from the read count distribution.
2. For each read from 1 through $n_{reads,i}$ sample a number $n \sim U(0,1)$
    I. If $n < d$, 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 $d \leq n < 1 - u$, 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 (see below) corresponding to loci in the 10th percentile by coverage. We also used a read count distribution representative of the median coverage in the SGDP dataset to assess performance at higher coverage.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Collection of previously published mutation rate estimates

We collated STR mutation rates from two previous large studies to enable comparison with our mutation rate estimates (Ballantyne et al. 2010; Burgarella and Navascues 2011). Utilizing only the estimates obtained from analyzing thousands of father-son pairs, we collected at least one mutation rate estimate for nearly 190 Y-STRs. To associate each marker with a locus in the reference genome, we utilized the published set of primer sequences and the isPCR tool (Hinrichs et al. 2006) to map the primers to hg19 coordinates. We then ran Tandem Repeats Finder (Benson 1999) (TRF) on each region and pinpointed the coordinates using the published repeat structure (Ballantyne et al. 2010) to generate a list of annotated STR regions. We also ran TRF on additional regions previously published as part of comprehensive Y-STR maps to obtain coordinates for a total of 261 annotated Y-STRs (Hanson and Ballantyne 2006).

STR Region Selection

We ran TRF on the hg19 assembly of the human reference genome and utilized previously described score thresholds to select only those regions significantly more repetitive than random genomic DNA (Willems et al. 2014). As this tool occasionally reports multiple overlapping repeats for a single genomic region, we merged overlapping entries in which the highest scoring entry contained 85% of the bases in the entries' union. Overlapping entries that failed this criterion but had the same period were further merged as they frequently represent loci comprised of two neighboring motifs (e.g. $[GATA]_{10}[TACA]_{8}$), while the remaining regions were omitted. We further removed regions that overlapped the annotated markers, failed to liftOver (Hinrichs et al. 2006) to the GrCh38 assembly or were lifted to the X chromosome. We then generated the complete STR reference using these regions and the annotated STRs described above.

Y-STR Call Set Generation

We downloaded BWA-MEM (Li 2013) alignments for 179 male and 108 female SGDP samples from the project website and extracted and merged the Y-chromosome alignments into a single BAM file using SAMtools (Li 2013). STR genotypes were then generated using HipSTR, a multi-sample haplotype-based STR caller that specifically accounts for the PCR stutter artifacts that drive most STR genotyping errors. HipSTR was run using the merged BAM, the hg19 STR regions described above, and the options *--min-reads 25 --haploid-chrs chrY --hide-allreads*. Similarly, we downloaded BWA-MEM alignments for 1320 male and 1371 female samples in the 1KGP phase 3 data release. As these alignments were relative to the *GrCh38* assembly, we ran HipSTR using the corresponding GrCh38 STR regions and the options *--min-reads 100 --haploid-chrs chrY --hide-allreads*.

Y-STR Call Set Filtration

To mitigate potential mutation estimate errors caused by pseudoautosomal and duplicated regions, we applied a series of stringent quality controls. We began by filtering the SGDP genotypes, as the 30X sequencing data and PCR-free protocol provided the highest quality dataset. To remove Y-STRs with putative homologous sites on the X chromosome, loci with more than 2 genotyped females were discarded. We further removed sites where more than 7.5% of reads had an indel in the STR flanks or 15% of reads had a stutter artifact, statistics that HipSTR reports based on the maximum likelihood alignment of each read relative to its sample's most probable haplotype. These loci likely represent instances in which duplicated copies of a polymorphic locus are mapping to a single reference genome locus, HipSTR failed to generate sufficient candidate alleles or the STR is flanked by an indel. For the remaining loci, we discarded

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

unreliable calls on a per-sample basis if more than 10% of an individual's reads had an indel in the flanks. Because the mutation model outlined above assumes that all alleles are integral copies of the repeat unit, we discarded loci where more than 5% of samples' genotypes violated this assumption. To avoid errors introduced by neighboring repeats, we omitted genotyped loci that overlapped one another or multiple STR regions, an issue that can arise when HipSTR expands an STR region to include proximal indels. Finally, we removed loci in which fewer than 100 samples had genotype posteriors above 66%, as 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 stutter frequency cutoff. These alterations account for the 1000 Genomes dataset's larger sample size and lack of a PCR-free protocol.

Estimating Y-STR Mutation Rates

For each locus in the SGDP and 1KGP call sets that passed the requisite quality control filters, we first used the EM algorithm to learn a PCR stutter model. The read STR sizes required to run this algorithm were obtained from the MALLREADS VCF field in which HipSTR reports the maximum likelihood STR size observed in each read that spans its sample's most probable haplotype. In conjunction with a uniform prior, this stutter model was then used 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. Finally, in conjunction with the optimization procedure and the appropriate scaled Y-SNP phylogeny, these genotype posteriors were used to obtain a point estimate of the mutation rate.

Confidence Interval Estimation

We utilized a delete-*d* jackknife approach to estimate mutation rate confidence intervals (Shao and Wu 1989). For each Y-STR, we sampled without replacement half of the STR genotypes above the genotype posterior threshold a total of 250 times and recalculated the log mutation rate using each of these subsets. Given these subsample estimates and the log estimate obtained using all samples, the standard error (SE) and confidence interval (CI) for the log mutation rate were calculated according to:

$$SE = \sqrt{\frac{1}{250}\sum_{i=1}^{250}\left(log\ \mu_i - \frac{1}{250}\sum_{j=1}^{250} log\ \mu_j\right)^2}, \qquad CI = log\ \mu_{tot} \pm 1.96 * SE$$

Effective Number of Meioses

For each phylogeny, we computed the sum of the branch lengths in generations after scaling (see results section). This resulted in estimates of ~177,600 and ~72,600 meioses in the 1KGP and SGDP phylogenies, respectively.

Estimating the Number of De Novo Mutations

To predict the number of de novo mutations on paternally inherited chromosomes:

1. We constructed a genome-wide reference of STRs using an approach identical to that for Y-STRs
2. We bootstrapped Y-STR loci for each repeat unit 2-4 base pairs in length 1000 times. For each bootstrapped dataset and each repeat unit length, we

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

i. Built a sequence-based mutation rate model using the sampled Y-STRs
ii. Utilized the fitted models to predict the mutation rate of each locus in the genome-wide reference with the same repeat unit based on its sequence properties
iii. Summed the resulting values to obtain an aggregate mutation estimate

3. We selected the 5th and 95th percentiles of aggregated estimates to obtain a 95% confidence interval for each repeat unit length

To build a sequence-based mutation rate model for each motif length, we assigned all fixed Y-STRs a log mutation rate of -5 (the minimum bound during optimization), all polymorphic Y-STRs the mean log estimated mutation rate between the two WGS datasets and utilized numerical optimization to fit a model of the form

$$\log \mu = \begin{cases} -5, & l < T \\ -5 + s(l - T), & l \geq T \end{cases}$$

where T is a threshold, s is the slope of the line and l is the length of the longest uninterrupted tract for each locus. These fitted models provide an estimate of the mean rate for a given tract length, but to account for uncertainty and to omit estimates for loci below the mutation rate optimization threshold, we predicted mutation rates as follows:

$$\widehat{\log \mu} = \begin{cases} -\infty, & l < T \\ -5 + s(l - T) + t_{N-2} * s_y \sqrt{\frac{1}{N} + \frac{N(l - \bar{L})^2}{N \sum L_i^2 - (\sum L_i)^2}}, & l \geq T \end{cases}$$

where t_{N-2} is sampled from a t-distribution with $N - 2$ degrees of freedom and N, L_i and $\bar{L}$ are the number, length and mean length of Y-STRs used to fit each model, respectively. To avoid potential biases, we did not generate predictions for loci whose tract lengths were above the maximum length used to train each model.

Y-STR Imputation Method

Given a set of samples with Y-SNP genotypes and a reference panel with Y-SNP and Y-STR genotypes, we extended the mutation rate estimation method 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. We then use this learned mutation model to pass two sets of messages along the tree and compute exact posteriors for each node. Samples with observed genotypes correspond to leaves in the tree and their posteriors represent imputation probabilities. In particular, for a node N_i in a binary phylogeny with parent P_i, sibling S_i and children C_{1i} and C_{2i}, its probability conditioned on the observed genotypes is given by

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

where D_{N_i} and D_{-N_i} denote the data in and not in node N_i's subtree, respectively. The first term in this expression is computed using a bottom-up traversal of the tree from the leaves to the root node. Each node in the tree combines the probabilities of its two children using the recurrence

$$P(D_{C_{i1}} | N_i) = \sum_{a \in alleles} P(D_{C_{1i}}, C_{1i} = a | N_i) = \sum_a P(D_{GC_{1i}}, D_{GC_{2i}}, C_{i1} = a | N_i)$$

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$= \sum_{a} P(C_{1i} = a \mid N_i)\, P\left(D_{GC_{1i}} \middle| C_{1i} = a\right) P(D_{GC_{2i}} \mid C_{1i} = a)$$

where 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 second term in the node posterior expression is computed using a top-down traversal of the tree from the root to the leaves. After assigning the root node a uniform probability, each node combines information from its parent and sibling:

$$P(N_i \mid D_{-N_i}) = \sum_{a \in alleles} P(N_i, P_i = a \mid D_{S_i}, D_{-P_i}) = \sum_{a} P\left(N_i, P_i = a, D_{S_i} \middle| D_{-P_i}\right) / P\left(D_{S_i} \middle| D_{-P_i}\right)$$

$$= \sum_{a} P(P_i = a \mid D_{-P_i})\, P\left(D_{S_i} \middle| P_i = a, D_{-P_i}\right) P(N_i | P_i = a, D_{S_i}\, D_{-P_i}) / P\left(D_{S_i} \middle| D_{-P_i}\right)$$

$$\propto \sum_{a} P(P_i = a \mid D_{-P_i})\, P\left(D_{S_i} \middle| P_i = a\right) P(N_i | P_i = a)$$

Results

Mutation Rates Estimates with Perfect Genotypes

We validated our mutation rate estimation algorithm by simulating STR genotypes under various mutation models and assessing how accurately we could recover the true mutation rate when given error-free observations. For the majority of models, we obtained an unbiased estimate of the log mutation rate using both phylogenies (**Figure S1**). Slight upward biases were observed for the smallest simulated mutation rate (10^{-5} mpg), but these stem from the lower bound imposed during numerical optimization. As previous studies have developed estimators involving simplified mutation models, we sought to assess how these simplifications might affect estimates. Restricting the optimized models to single step mutations resulted in stronger upward biases for low mutation rate scenarios, strong downward biases for high mutation rate scenarios and higher variance in the estimates (**Figure S2**). The effect of disabling the length constraint for optimized models was much less pronounced, but also resulted in large downward biases for many rapidly mutating scenarios. Altogether, these results illustrate that if given error-free genotypes, our method is well powered to obtain accurate estimates across a host of mutation scenarios but that making overly simplistic assumptions about STR mutation models can result in marked biases.

Mutation Rates Estimates with PCR Stutter

We extended the above simulations to introduce the effects of PCR stutter, a primary driver of STR genotyping errors. After simulating STR genotypes under various mutation models, we generated observed reads using various stutter models and distributions of reads per sample. Application of the EM algorithm to this data resulted in relatively unbiased estimates of each stutter model parameter for nearly all scenarios (**Figure S3**). Slight downward biases were observed for the geometric step size parameter when both stutter frequencies were 1%, but this is likely caused by the scarcity of informative instances of stutter in this setting. The variance of the stutter parameter estimates decreased substantially with increases in sample size and mean number of reads per sample, as these led to more stutter-informative reads.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We next sought to assess whether the stutter parameter estimates were sufficiently precise for mutation rate inference. To this end, we estimated mutation rates after computing genotype posteriors using the learned stutter models. For comparison, we also generated estimates when posteriors were computed with exact knowledge of the stutter model or using a naive approach based on the fraction of reads supporting each allele. In scenarios with low average coverage, the fraction-based posteriors resulted in marked biases, particularly for low mutation rates, demonstrating the importance of correctly accounting for stutter artifacts in these settings (**Figure 2, Figures S4-S5**). In contrast, posteriors generated using the estimated and exact stutter models obtained relatively unbiased mutation rate estimates across all scenarios and yielded estimates with similar variance. The primary exception to these trends was the slight upward bias observed for rates of 10^{-5} mpg, but this bias was also observed in the simulations with exact genotypes. Collectively, these results indicate that the combination of the EM and mutation rate algorithms obtain robust estimates suitable for downstream analyses.

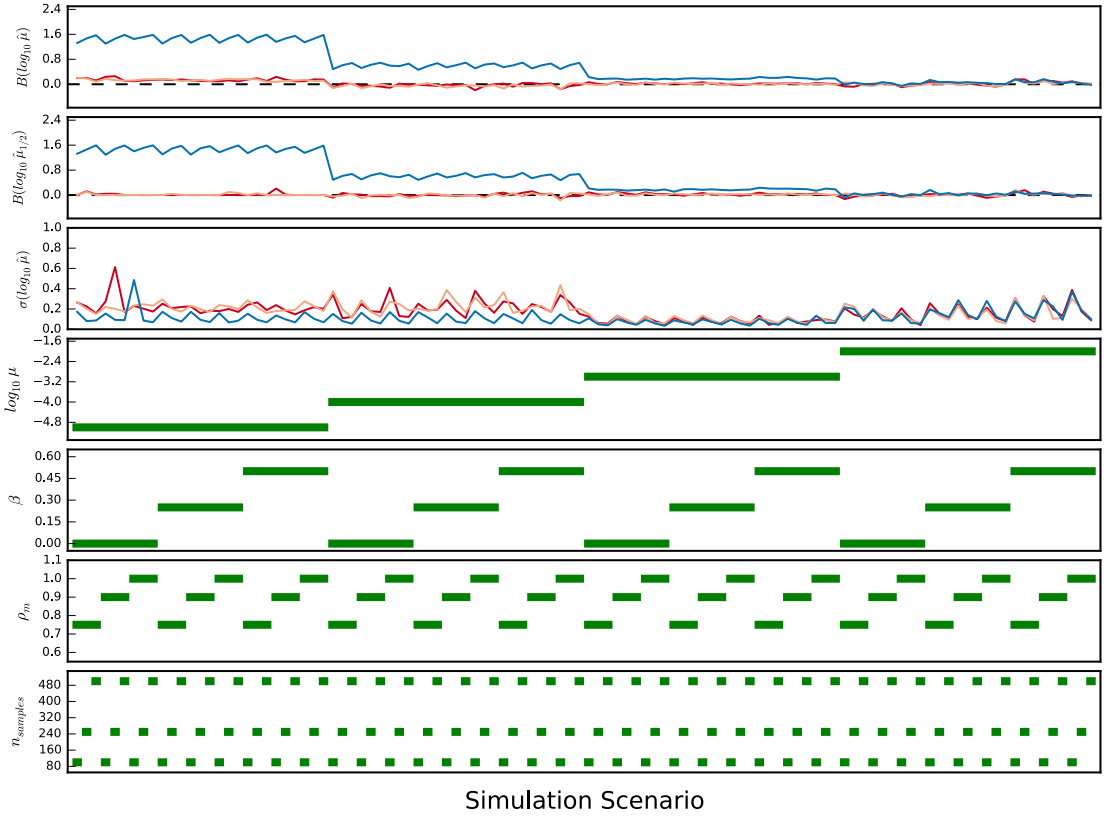

Figure 2: Accuracy of mutation rate estimates in the presence of PCR stutter. STR genotypes were simulated for a variety of sample sizes and mutation models (bottom four panels). Reads for each sample's genotype were then simulated using a PCR stutter model with $d = 0.15, u = 0.01$ and $\rho_s = 0.8$ and using 1, 2 and 3 reads for 65%, 25% and 10% of samples, respectively. Across 25 iterations for each simulation scenario, genotyper posteriors computed using the fraction of supporting reads (blue) resulted in markedly biased mutation rate estimates (top two panels), while posteriors computing using the exact stutter model (red) and EM stutter model (orange) resulted in relatively unbiased estimates with similar standard deviations (third panel).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Call Set Validation

To assess the level of genotyping errors present in each call set, we stringently filtered each call set and compared them to capillary electrophoresis datasets involving a subset of the same male samples. For 565 samples in the 1000 Genomes Project, the concordance for 3500 calls at 13 loci in the PowerPlex Y23 panel was 97.5%, indicating that the low coverage data was not prohibitive for obtaining accurate Y-STR genotypes. An analogous comparison of 3300 calls at 48 loci for 76 SGDP samples resulted in an even higher concordance of 99.7%. These comparisons were restricted to loci with 3-5 base pair motifs and therefore may not reflect the quality for loci with shorter motifs due to their increased propensity for stutter. Nonetheless, they are indicative of the high quality of the data for larger repeat motifs.

Scaling Phylogeny Branch Lengths

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 sought to determine an appropriate scaling factor using mutation rate estimates for 15 loci in the Y-chromosome Haplotype Reference Database (YHRD) (Willuweit et al. 2007). We chose these loci as a calibration point because their estimates are based on more than 7000 father-son pairs per locus and should therefore be relatively precise. For the 1000 Genomes data, we used the PowerPlex capillary data for each locus, assumed error-free genotypes, scaled the phylogeny using a range of factors and estimated the set of mutation rates using each scaling factor. The choice of scaling factor had essentially no affect on the correlation with the YHRD estimates, resulting in an R^2 of 0.89 (**Figure S6**). However, the total squared error between the estimates was minimized for a factor of 2800, which we therefore selected as the optimal scaling. For the SGDP data, we performed an analogous analysis using HipSTR genotypes for 9 of these 15 loci, again resulting in a uniform R^2 of 0.91 and an optimal factor of roughly 3200 (**Figure S6**).

An alternative approach to scaling each phylogeny is to select the factor that best matches the total number of generations in the tree to the value based on published Y-SNP mutation rates. To explore how this approach might impact the scaling, we calculated factors using a recently published Y-SNP mutation rate of 3E-8 mutations per generation (Xue et al. 2009; Helgason et al. 2015) and the total numbers of called SNPs and called sites in each SNP dataset. The resulting estimates for the 1KGP data and SGDP data were remarkably concordant with those above, as they were only 14% and 34% greater. However, to maximize our concordance with pedigree estimates, we chose to utilize the first set of scaling factors outlined above.

Y-STR Stutter Models

We applied the EM algorithm to each of the filtered call sets to learn per-locus stutter models. Across both datasets, the learned parameters demonstrated a strong bias in favor of stutter-induced contractions versus expansion for nearly all loci (**Figure S7)**. The stutter parameter estimates were highly correlated between the two datasets, reflecting the algorithm's ability to capture each locus' distinctive error profile, but as expected the PCR-free protocol resulted in significantly lower stutter rates for the SGDP dataset relative to the 1KGP dataset (**Figure S7**). Within each dataset, the rates of stutter exhibited an inverse correlation with repeat unit size for a given allele length and a positive correlation with allele length for a given repeat unit size **(Figure S8)**.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Y-STR Mutation Rate Estimates

After utilizing the learned stutter models to compute genotype posteriors, we applied the mutation rate estimation algorithm to each polymorphic locus, resulting in estimates for 702 loci with 2-6 base pair motifs. Stratifying these estimates by motif length indicated a wide degree of variability both within and between classes (**Figure 3**). Within each class, mutation rates varied by two or more orders of magnitude, indicating that Y-STR mutation rates are highly dependent on the loci under consideration. Relative to other Y-STR classes, loci with previously characterized rates had substantially higher estimates, illustrating that they've been selected for their high mutability (**Figure 3**, **Tables 1-2**). While the bulk of uncharacterized loci with tri- and tetranucleotide motifs were substantially less polymorphic, we identified 29 of these loci with mutation rates greater than $10^{-3.5}$ mpg, of which the five most mutable loci had rates ranging from $10^{-2.29}$ - $10^{-2.44}$ mpg **(Table 2**). Dinucleotide repeats were also highly polymorphic and 70 of these loci had mutation rates above $10^{-3.5}$ mpg.

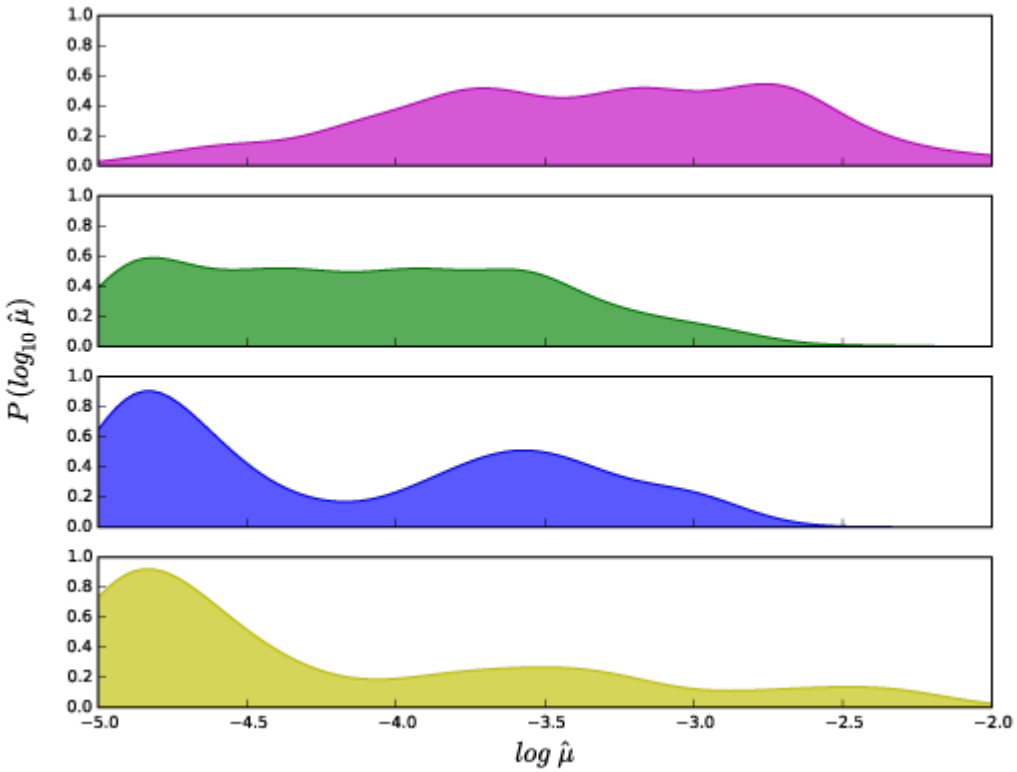

Figure 3: Distribution of Y-STR mutation rates. Loci with previously characterized mutation rates (purple) are substantially more mutable than uncharacterized loci with dinucleotide (green), trinucleotide (blue) and tetranucleotide (yellow) motifs. Nonetheless, a substantial number of these uncharacterized loci are highly polymorphic with mutation rates greater than $10^{-3.5}$ mutations per generation.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation Rate Concordance

In order to assess the reliability of our mutation rate estimates, we measured the concordance between the two sets of WGS-based estimates obtained in this study. Despite substantial differences in the quality of the sequencing data, the analyzed populations and the study sizes, we obtained an R^2 of 0.92 between the 1KGP and SGDP log mutation rate estimates (**Figure 4**). This high concordance extended to slowly mutating markers, as estimates for loci with a mean estimated mutation rate below $10^{-3.5}$ mpg had an R^2 of 0.80. To assess the potential impact of genotyping errors on these estimates, we regenerated them using the 1000 Genomes capillary genotypes for 23 loci, resulting in R^2 of 0.98 and 0.94 with the SGDP and 1KGP estimates for the same loci. These comparisons illustrate that our method obtains robust locus-specific values while accounting for varying degrees of PCR stutter artifacts and genotyping errors. Furthermore, the inter-dataset concordance suggests that there are either very few errors in the phylogenies or that these errors have little impact on the resulting mutation rate estimates.

Next, we assessed the accuracy of our mutation rate estimates by comparing them to results from prior studies based on roughly 1500 and 500 father-son transmissions per Y-STR **(Figure 4**) (Ballantyne et al. 2010; Burgarella and Navascues 2011). The R^2 between these two studies was only 0.34, a low concordance that likely stems from the small sample size and large uncertainty in the Burgarella et al. estimates. By comparison, the SGDP and Ballantyne et al. estimates had an R^2 of 0.66. Although markedly higher, this concordance was substantially reduced by the plateau in the Ballantyne et al. estimates at $10^{-3.5}$ mpg, a threshold that stems from loci without any detected mutations. While accurately characterizing these loci using the father-son approach would require tens of thousands of additional pairs, our method easily obtains replicable estimates below this threshold by leveraging over 222,000 meioses in the phylogenies. An analogous comparison of the SGDP and Burgarella et al. estimates resulted in an R^2 of 0.32. However, restricting this comparison to a subset of loci characterized using more than 5000 father-son pairs resulted in a substantially higher R^2 of 0.87 (**Figure S9**). Collectively, these comparisons demonstrate that our method accurately replicates father-son based estimates based on sufficient pairs, but that the father-son approach is poorly suited to quantifying the point estimates for mutation rates of slowly mutating markers.

Discriminative Power

Given the large number of markers with novel mutation rates, we sought to assess the potential gains in discriminative power they might provide. We therefore computed the probability of observing at least one mutation over one generation for various groups of loci. Utilizing the full panel of 190 Y-STRs characterized by Ballantyne et al. resulted in a discrimination probability of 42%. Extending this set of markers to incorporate those with novel rates in this study increased this probability to 50%. However, because of the constraints imposed by the Illumina sequencing reads, we were unable to genotype many of the long markers in the Ballantyne et al. study, particularly most of the 13 rapidly mutating markers with mutation rates greater than 0.01 mpg. The subset of their markers we were able to genotype resulted in a discrimination probability of 12% and incorporating our novel marker estimates improved this probability to 24%.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

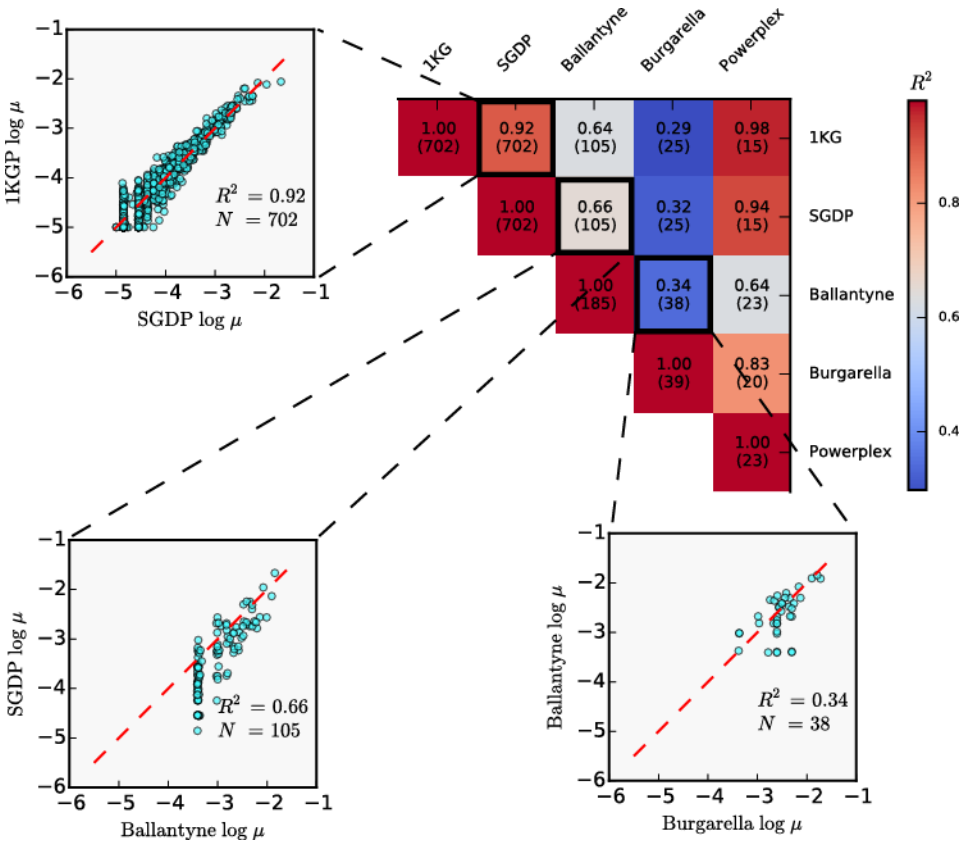

Figure 4: Concordance of mutation rate estimates. A comparison of the log mutation rates obtained from two father-son based studies (Ballantyne and Burgarella) with those obtained in this study using the 1000 Genomes WGS data (1KGP), the Simons Genome WGS data (SGDP) and the 1000 Genomes capillary data (Powerplex). Each square of the heatmap 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 diagonal (red line).

Sequence Determinants of Y-STR Mutability

To assess the extent to which sequence characteristics drive STR mutation rates, we analyzed how allele length, repeat-motif length, and interruptions to the repeat structure affect mutability. For STRs with and without interruptions, major allele length only explained a modest amount of the variance in log mutation rate for loci with di-, tri-, and tetra-nucleotide motifs (R^2 = 0.16, R^2 = 0.25, and R^2 = 0.42) (**Figure 5**). Restricting this analysis to STRs without interruptions substantially improved the variance explained (R^2 = 0.83, R^2 = 0.67, and R^2 = 0.82), suggesting that interruptions to the repeat structure disrupt the correlation between allele length and mutability. A subsequent analysis of the relationship between the log mutation rate and the length of the longest uninterrupted repeat tract indicated that this was a more general predictor of mutability (**Figure 5**), as it explained over 75% of the variance for each of the three motif lengths regardless of the number of interruptions. Stratifying loci with dinucleotide repeat units by motif indicated that these trends also apply at a much finer scale (**Figure S10**). Major allele length was once again a relatively poor predictor of the log mutation rate for loci with AC, AG and

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

AT repeat motifs, but uninterrupted tract length explained over 80% of the variance for each motif.

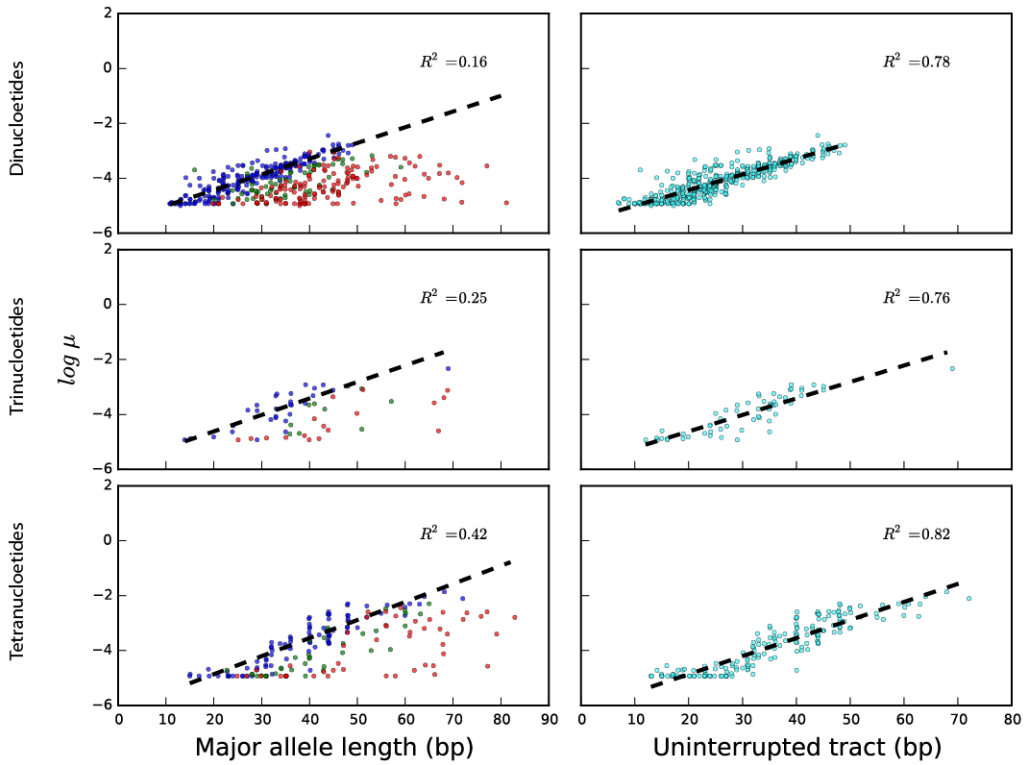

Figure 5: Sequence determinants of Y-STR mutability. Stratified by repeat motif length (rows) and major allele length, YSTRs with no interruptions to the repeat structure (blue) are generally more mutable than those with one interruption (green) or more than one interruption (red). While major allele length is a poor predictor of mutability, the length of the longest interrupted tract is a very strong predictor of the log mutation rate for each motif length (cyan).

De novo mutations

We sought to use the mutation rates and sequence properties of each Y-STR to predict the expected number of genome-wide de novo mutations. Because the Y-STR mutation rates are only applicable to the male germ line due to differing numbers of meioses, we restricted this analysis to the number of de novo mutations on paternally inherited chromosomes. To generate a prediction, we built sequence-based models of Y-STR mutability, obtained per-locus estimates by applying these models to each locus in a genome-wide reference of STRs and aggregated the results (**Methods**). After utilizing bootstrapping and sampling techniques to account for uncertainty in both the fitted models and the predicted values, we obtained 95% confidence intervals of 27-34, 2-11 and 37-102 mutations for loci with di-, tri- and tetranucleotide motifs. The 95% confidence interval for the total number of expected de novo mutations on paternally inherited chromosomes was 72-140 mutations. These estimates are likely conservative as we omitted loci with 5-6 base pair motifs due to a relatively small numbers of Y-STRs

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and omitted genome-wide loci that were longer than the Y-STRs used to train each model.

Imputing Y-STRs

We extended the mutation rate estimation procedure to develop a Y-STR imputation approach. Briefly, after building a SNP phylogeny relating all samples and learning a mutation model as outlined in **Figure 1**, this approach 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. To assess the accuracy of this technique, we once again turned to the capillary PowerPlex Y23 genotypes for the 1KGP dataset, as this panel is one of the most commonly used in forensic and genealogical settings. Over 100 iterations, we randomly constructed reference and imputation panels of 500 and 70 samples, utilized the reference panel's Y-STR genotypes to infer a mutation model and compute node posteriors, and compared the imputed genotypes for the imputation panel to their true underlying values. The resulting imputed probabilities roughly matched their true accuracy, indicating that the posteriors computed using this technique are well calibrated (**Figure S11**). When using all imputed genotypes, even those with probabilities below 50%, this approach resulted in an overall accuracy of 66% across markers (**Table 3**). However, discarding imputed genotypes with probabilities less than 70% resulted in an overall accuracy of 88% and retained more than 40% of the calls. On a marker-by-marker basis, accuracy was generally inversely proportional to the estimated mutation rates, with the most slowly mutating markers having accuracies on the order of 95%. This trend stems from the fact that as the mutation rate increases, shorter branch lengths are required to obtain an estimate with similar confidence.

Discussion

Over the past two decades, tremendous advances in sequencing technology have fundamentally transformed the applications of Y-STRs. The initial scarcity of available SNP genotypes resulted in the development of methods capable of inferring coalescent models from Y-STR genotypes alone. Methods designed to also learn STR mutational dynamics either marginalized over these coalescent models (Nielsen 1997) or aimed to simultaneously infer the coalescent and mutational models (Wilson and Balding 1998; Wilson et al. 2003). With the advent of population-scale WGS datasets, many of these STR-centric approaches have instead utilized SNPs, resulting in substantially more detailed phylogenies. On the Y-chromosome, these detailed phylogenies now provide the evolutionary context required to interpret Y-STR mutations, obviating the need for expensive tree enumeration or marginalization approaches. However, the errors prevalent in WGS-based Y-STR genotypes require methods capable of accounting for genotype uncertainty, preventing the application of many traditional microsatellite distance measures designed for capillary data (Goldstein et al. 1995; Slatkin 1995).

In this study, we developed a novel method to leverage these datasets. One inherent advantage of our approach is its ability to model and learn many of the salient features of microsatellite mutations. Through the incorporation of a geometric step size distribution, we allow both single step mutations that predominate at tetranucleotide loci (Kayser et al. 2000; Sun et al. 2012) as well as multistep mutations that frequently occur at dinucleotide loci (Huang et al. 2002; Sun et al. 2012). In addition, the model's length constraint parameter replicates the intra-locus phenomenon of shorter STR alleles preferentially expanding and longer alleles preferentially contracting (Xu et al. 2000;

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Huang et al. 2002). As these parameters are learned from the observed STR genotypes, our method avoids many biases that stem from imposing single-step mutations or assuming parameters a priori. Nonetheless, our mutation model does not capture the fully complexity of STR mutational dynamics as it ignore intra-locus mutation rate variation (Ellegren 2000). Incorporating these and other mutational characteristics may be of interest in future studies.

In addition to its mutational model flexibility, our approach is advantageous because of its ability to leverage evolutionary data. From the large number of meioses in each phylogeny, it can obtain extremely replicable and accurate estimates, as demonstrated by the strong concordance between our WGS-based estimates and their strong concordance with father-son estimates based on sufficient pairs. In contrast, the estimates of Ballantyne et al. and Burgarella et al. showed poor concordance, likely due to the small number of pairs used in one of these studies. This underscores the fact that without vast numbers of samples, pedigree-based approaches cannot obtain precise point estimates for more slowly mutating markers. We believe that this limitation, coupled with our method's ability to analyze any WGS dataset and hundreds of STRs in parallel, make it a simple and scalable alternative to pedigree-based estimation approaches.

One longstanding concern regarding Y-STR mutation rates has been the apparent discrepancy between evolutionary and pedigree-based mutation rates. A host of studies have suggested that evolutionary rates are 3-4 times lower, resulting in substantial inconsistencies in Y-STR based lineage dating and large discrepancies from Y-SNP based TMRCA estimates. (Zhivotovsky et al. 2004; Zhivotovsky et al. 2006; Wei et al. 2013b). Because this 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 15-30%. Coupled with the strong concordance we observed with pedigree estimates, our study provides little evidence for a substantial difference between mutation rates estimated from these two types of data.

Empowered by the accuracy and parallelizability of our method, we were able to obtain Y-STR mutation rates on an unprecedented scale. The set of estimates for over 700 polymorphic STRs is, to our knowledge, the largest Y-STR set to date, substantially expanding upon those previously obtained for 190 loci (Ballantyne et al. 2010; Burgarella and Navascues 2011). Two of the largest prior studies of autosomal STR mutability characterized 350 and 2500 markers using traditional family-based approaches (Huang et al. 2002; Sun et al. 2012), but these studies only observed mutations for 50 and 800 of these loci. As a result, the scope of our study also parallels those of the largest autosomal studies.

Despite the large-scale nature of our study, it has several inherent limitations. Because we analyzed sequencing datasets comprised of 80-100bp Illumina reads, we were unable to genotype and characterize the mutation rates of many long Y-STRs. Given the strong positive correlation between tract length and mutation rate observed here and in previous studies, we anticipate that many long dinucleotide loci will be extremely mutable and will add significant discriminative power to Y-STR panels. We were also unable to include homopolymers in this study despite their shorter lengths due to a rapid degradation in base quality scores, but we anticipate that many of these markers are

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

also highly polymorphic. As a result, future studies may benefit from reapplying our analysis to both of these sets of markers as sequencing technologies, especially those enabling long reads, continue to mature.

Nonetheless, given the extent of our set of estimates, we were able to shed new light on the sequence factors governing STR mutability. While prior Y-STR studies have primarily focused on loci with longer alleles and 3-6 base pair motifs, the results here extend these analyses to shorter loci and repeats with dinucleotide motifs. In particular, we found that for all examined repeat unit lengths, the longest uninterrupted tract length is an extremely strong predictor of the log mutation rate, replicating the exponential trend between mutation rate and tract length previously observed in a host of pedigree-based studies (Brinkmann et al. 1998; Kayser et al. 2000; Xu et al. 2000; Ballantyne et al. 2010). In contrast, allele length alone 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 (Kruglyak et al. 1998; Sainudiin et al. 2004) and has been shown in a handful of small-scale experimental studies of STR mutability (Petes et al. 1997; Bacon et al. 2000) but contradicts the recent findings of Ballantyne et al in which no effect was observed. This discrepancy may stem from the fact that they primarily considered longer repeats with uninterrupted tract lengths at least 8 repeat units long.

In addition to estimating Y mutation rates, we've outlined a Y-STR imputation method that is, to the best of our knowledge, the first of its kind. A preliminary assessment of this method's accuracy indicated that imputation accuracies of up to 95% can be achieved for some of the most slowly mutating markers in the PowerPlex Y23 but that the performance is much poorer for more rapidly mutating markers. However, the accuracy of our approach is essentially linear in the shortest time to the most recent common ancestor. As a result, as population-scale sequencing datasets for the Y-chromosome continue to expand in scope to tens of thousands of individuals, we expect its accuracy to increase substantially. We also anticipate that our Y-STR mutation rate method and its relevant extensions may be applied to autosomal STRs. Although recombination complicates the generation of sufficiently detailed phylogenies, tools capable of inferring ancestral recombination graphs and the associated phylogenies continue to improve (Minichiello and Durbin 2006; Rasmussen et al. 2014). As a result, it may be possible to apply these approaches to ensembles of trees and aggregate the results.

The large corpus of mutation rate estimates has also enabled novel predictions about genome-wide STR variation. Prior studies have estimated a rate of approximately 75 de novo SNP mutations per generation (Conrad et al. 2011; Francioli et al. 2015) but have largely ignored STRs despite their elevated mutation rates. Based on our projections for de novo mutations on paternally inherited chromosomes, the number of de novo STR mutations is likely to exceed that of SNPs. An increasingly large number of candidate gene studies and genome-wide analyses have highlighted instances in which STR variations modulate gene expression (Gebhardt et al. 1999; Shimajiri et al. 1999; Contente et al. 2002; Gymrek et al. 2015). We therefore hope that as others aim to dissect the genetic basis of complex diseases and traits, this study will motivate them to consider STRs as the causal genetic elements.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

M.G. was supported by the National Defense Science and Engineering Graduate Fellowship. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was partially supported by NIJ grant 2014-DN-BX-K089 (Y.E. and T.W.)

Web Resources

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

Dendroscope, http://dendroscope.org/

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

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

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

1000 Genomes PowerPlex Y23 genotypes, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20140107_chrY_str_haplotypes/s_PowerPLexY23_1000Y_QA_20130107.txt

1000 Genomes Project Y-chromosome phylogeny, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/1000Y.Supplementary.Data.File.2016.01.04.tar.gz

Y-Chromosome STR Haplotype Reference Database mutation rates, https://yhrd.org/pages/resources/mutation_rates

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

References

Bacon AL, Farrington SM, Dunlop MG. 2000. Sequence interruptions confer differential stability at microsatellite alleles in mismatch repair-deficient cells. *Human molecular genetics* **9**(18): 2707-2713.

Ballantyne KN, 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. *American journal of human genetics* **87**(3): 341-353.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**(2): 573-580.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American journal of human genetics* **62**(6): 1408-1415.

Burgarella C, Navascues M. 2011. Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data. *European journal of human genetics : EJHG* **19**(1): 70-75.

Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella KV et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nature genetics* **43**(7): 712-714.

Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. 2002. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature genetics* **30**(3): 315-320.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**(15): 2156-2158.

Dupuy BM, Stenersen M, Egeland T, Olaisen B. 2004. Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Human mutation* **23**(2): 117-124.

Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature genetics* **24**(4): 400-402.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**(6): 368-376.

Forster P, Rohl A, Lunnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, Brinkmann B. 2000. A short tandem repeat-based phylogeny for the human Y chromosome. *American journal of human genetics* **67**(1): 182-196.

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C. 1998. Jefferson fathered slave's last child. *Nature* **396**(6706): 27-28.

Francalacci P, Morelli L, Angius A, Berutti R, Reinier F, Atzeni R, Pilu R, Busonero F, Maschio A, Zara I et al. 2013. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341**(6145): 565-569.

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands C, van Duijn CM, Swertz M, Wijmenga C et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics* **47**(7): 822-826.

Gebhardt F, Zanker KS, Brandt B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *The Journal of biological chemistry* **274**(19): 13176-13180.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**(7571): 68-74.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**(1): 463-471.

Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML et al. 2005. Mutation rates at Y chromosome specific microsatellites. *Human mutation* **26**(6): 520-528.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome research* **22**(6): 1154-1162.

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* **339**(6117): 321-324.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ et al. 2015. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*.

Hanson EK, Ballantyne J. 2006. Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Legal medicine* **8**(2): 110-120.

Helgason A, Einarsson AW, Guethmundsdottir VB, Sigurethsson A, Gunnarsdottir ED, Jagadeesan A, Ebenesersdottir SS, Kong A, Stefansson K. 2015. The Y-chromosome point mutation rate in humans. *Nature genetics* **47**(5): 453-457.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human molecular genetics* **6**(5): 799-803.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic acids research* **41**(1): e32.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. 2006. The UCSC Genome Browser Database: update 2006. *Nucleic acids research* **34**(Database issue): D590-598.

Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, Liu YZ, Li JL, Recker RR, Deng HW. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. *American journal of human genetics* **70**(3): 625-634.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology* **61**(6): 1061-1067.

Jobling MA, Tyler-Smith C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature reviews Genetics* **4**(8): 598-612.

Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M et al. 1997. Evaluation of Y-chromosomal STRs: a multicenter study. *International journal of legal medicine* **110**(3): 125-133, 141-129.

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, Mehdi SQ, Rosser Z, Stoneking M, Jobling MA et al. 2004. A comprehensive survey of human Y-chromosomal microsatellites. *American journal of human genetics* **74**(6): 1183-1197.

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger 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. *American journal of human genetics* **66**(5): 1580-1588.

Kayser M, Vermeulen M, Knoblauch H, Schuster H, Krawczak M, Roewer L. 2007. Relating two deep-rooted pedigrees from Central Germany by high-resolution Y-STR haplotyping. *Forensic science international Genetics* **1**(2): 125-128.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences of the United States of America* **95**(18): 10774-10778.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997*.

Minichiello MJ, Durbin R. 2006. Mapping trait loci by use of inferred ancestral recombination graphs. *American journal of human genetics* **79**(5): 910-922.

Nelder JA, Mead R. 1965. A Simplex Method for Function Minimization. *The Computer Journal* **7**(4): 308-313.

Nielsen R. 1997. A likelihood approach to populations samples of microsatellite alleles. *Genetics* **146**(2): 711-716.

Petes TD, Greenwell PW, Dominska M. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* **146**(2): 491-498.

Poznik GD, Henn BM, Yee MC, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**(6145): 562-565.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* **16**(12): 1791-1798.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. *PLoS genetics* **10**(5): e1004342.

Ravid-Amir O, Rosset S. 2010. Maximum likelihood estimation of locus-specific mutation rates in Y-chromosome short tandem repeats. *Bioinformatics* **26**(18): i440-445.

Roewer L. 2009. Y chromosome STR typing in crime casework. *Forensic science, medicine, and pathology* **5**(2): 77-84.

Sainudiin R, Durrett RT, Aquadro CF, Nielsen R. 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* **168**(1): 383-395.

Shao J, Wu CFJ. 1989. A General Theory for Jackknife Variance Estimation. *The Annals of Statistics* **17**(3): 1176-1197.

Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y. 1999. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS letters* **455**(1-2): 70-74.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**(1): 457-462.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9): 1312-1313.

Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D et al. 2012. A direct characterization of human mutation based on microsatellites. *Nature genetics* **44**(10): 1161-1165.

Takezaki N, Nei M. 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**(1): 389-399.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, Budowle B. 2013. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic science international Genetics* **7**(4): 409-417.

Wei W, Ayub Q, Chen Y, McCarthy S, Hou Y, Carbone I, Xue Y, Tyler-Smith C. 2013a. A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome research* **23**(2): 388-395.

Wei W, Ayub Q, Xue Y, Tyler-Smith C. 2013b. A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. *Forensic science international Genetics* **7**(6): 568-572.

Willems T, Gymrek M, Highnam G, Genomes Project C, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome research* **24**(11): 1894-1904.

Willuweit S, Roewer L, International Forensic YCUG. 2007. Y chromosome haplotype reference database (YHRD): update. *Forensic science international Genetics* **1**(2): 83-87.

Wilson IJ, Balding DJ. 1998. Genealogical inference from microsatellite data. *Genetics* **150**(1): 499-510.

Wilson IJ, Weale ME, Balding DJ. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166**(2): 155-188 %@ 1467-1985X.

Xu X, Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nature genetics* **24**(4): 396-399.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current biology : CB* **19**(17): 1453-1457.

Zhivotovsky LA, Underhill PA, Cinnioglu 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. *American journal of human genetics* **74**(1): 50-61.

Zhivotovsky LA, Underhill PA, Feldman MW. 2006. Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Molecular biology and evolution* **23**(12): 2268-2270.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1. Most mutable markers with previously characterized mutation rates

Chrom	Hg19 start	Hg19 end	Motif	Mean log μ	Homogeneous tract (bp)	Annotated Name
Y	7053359	7053426	AAAG	-1.86	68	DYS576
Y	7867880	7867943	AAAG	-2.04	64	DYS458
Y	6861231	6861298	AAAG	-2.11	72	DYS570
Y	14515312	14515363	AGAT	-2.29	48	DYS439
Y	8426378	8426443	AAG	-2.33	69	DYS481
Y	21520224	21520275	AGAT	-2.34	48	DYS549
Y	18718889	18718940	AGAT	-2.38	52	Y-GATA-A10
Y	4270960	4271019	AGAT	-2.42	60	DYS456
Y	19372273	19372328	AGAT	-2.54	48	DYS543
Y	14761101	14761160	AGAT	-2.58	46	DYS442

Table 2. Five most mutable tetranucleotide and dinucleotide markers with previously uncharacterized mutation rates

Chrom	Hg19 start	Hg19 end	Motif	Mean log μ	Homogeneous tract (bp)	Annotated Name
Y	14612456	14612520	AGAT	-2.29	59	DYS467
Y	5409729	5409801	AAAG	-2.29	61	N/A
Y	19500594	19500656	AAAG	-2.31	63	N/A
Y	14200743	14200802	AGAT	-2.34	56	N/A
Y	21665702	21665764	AAAT	-2.44	50	DYS548
Y	2807025	2807064	AT	-2.44	44	N/A
Y	2708412	2708457	AG	-2.76	46	N/A
Y	3832234	3832278	AC	-2.78	45	N/A
Y	6398638	6398684	AC	-2.79	49	N/A
Y	17109092	17109141	AC	-2.80	48	N/A

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3. Imputation accuracy for each locus in the PowerPlex Y23 Panel

		Posterior > 0%		Posterior > 70%	
Marker	$\hat{\mu}$ **(mpg)**	**% Calls**	**% Correct**	**% Calls**	**% Correct**
DYS392	0.0006	100	92.7	96.1	94.1
DYS438	0.0007	100	93.4	95.1	95.5
DYS437	0.0008	100	92.3	94.8	94.1
DYS393	0.0015	100	83.1	75.7	88.7
DYS448	0.0018	100	81.4	80.3	89.0
DYS533	0.0018	100	78.6	74.4	86.2
DYS643	0.0020	100	80.2	77.4	86.3
DYS391	0.0023	100	73.8	53.3	79.9
Y-GATA-H4	0.0024	100	72.1	46.2	86.8
DYS390	0.0026	100	76.6	50.9	84.3
DYS385a	0.0027	100	74.4	63.9	87.5
DYS389I	0.0029	100	71.7	28.7	88.0
DYS19	0.0029	100	70.2	34.2	88.7
DYS635	0.0034	100	67.6	54.8	81.4
DYS456	0.0039	100	61.9	20.5	90.2
DYS549	0.0046	100	56.2	5.0	86.1
DYS439	0.0054	100	52.2	3.0	88.3
DYS481	0.0054	100	56.2	24.2	82.4
DYS385b	0.0055	100	55.2	16.8	87.3
DYS389II	0.0060	100	49.4	6.3	87.1
DYS458	0.0084	100	38.5	0.6	46.7
DYS570	0.0101	100	41.6	0.6	100.0
DYS576	0.0102	100	34.5	0.5	59.0
All		**100**	**67.6**	**43.7**	**88.6**

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplemental Figures

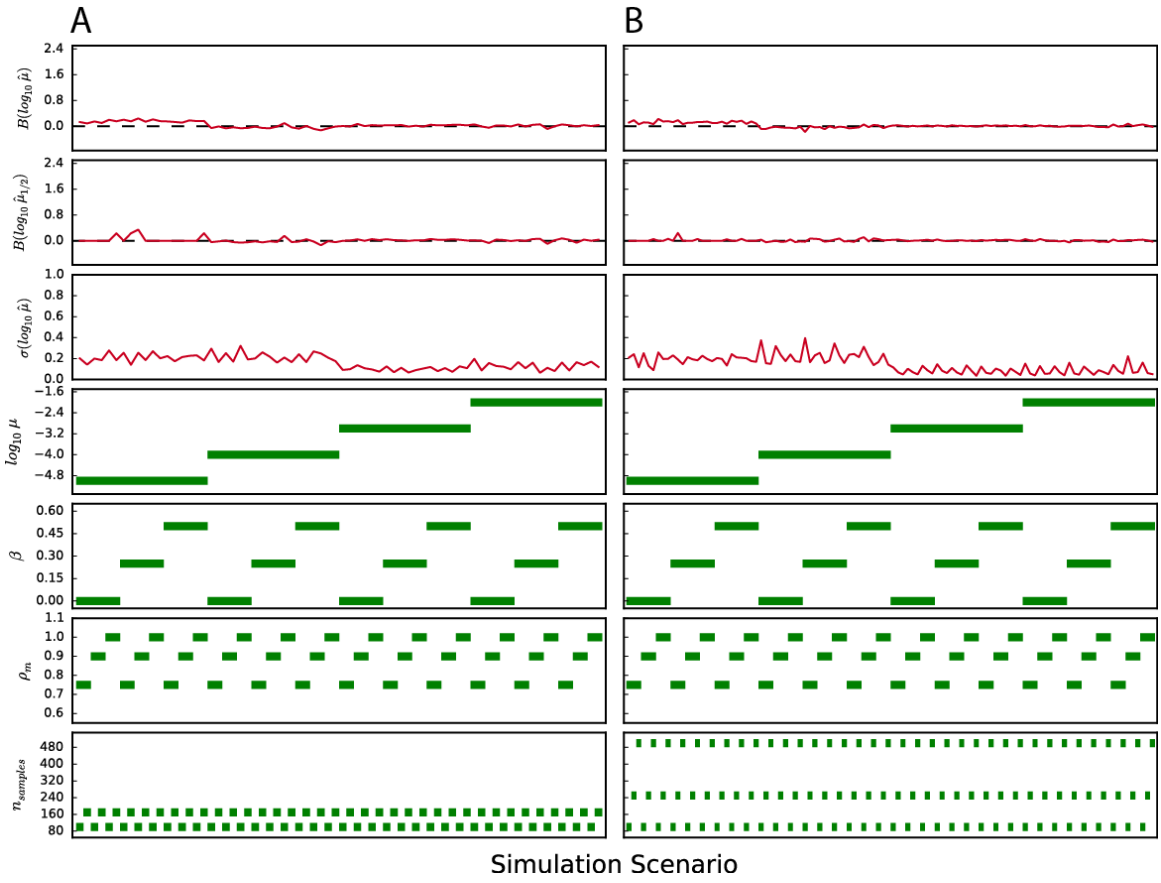

Figure S1: Accuracy of mutation rate estimates based on exact genotypes
STR genotypes were simulated for a variety of sample sizes and mutation models (bottom four panels) for both the Simons Genomes phylogeny (A) and 1000 Genomes phylogeny (B). Across 25 iterations for each simulation scenario, mutation rates computed after assigning each sample's genotype unity posterior probability are unbiased (top two panels) and have reasonably low standard deviations (third panel).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

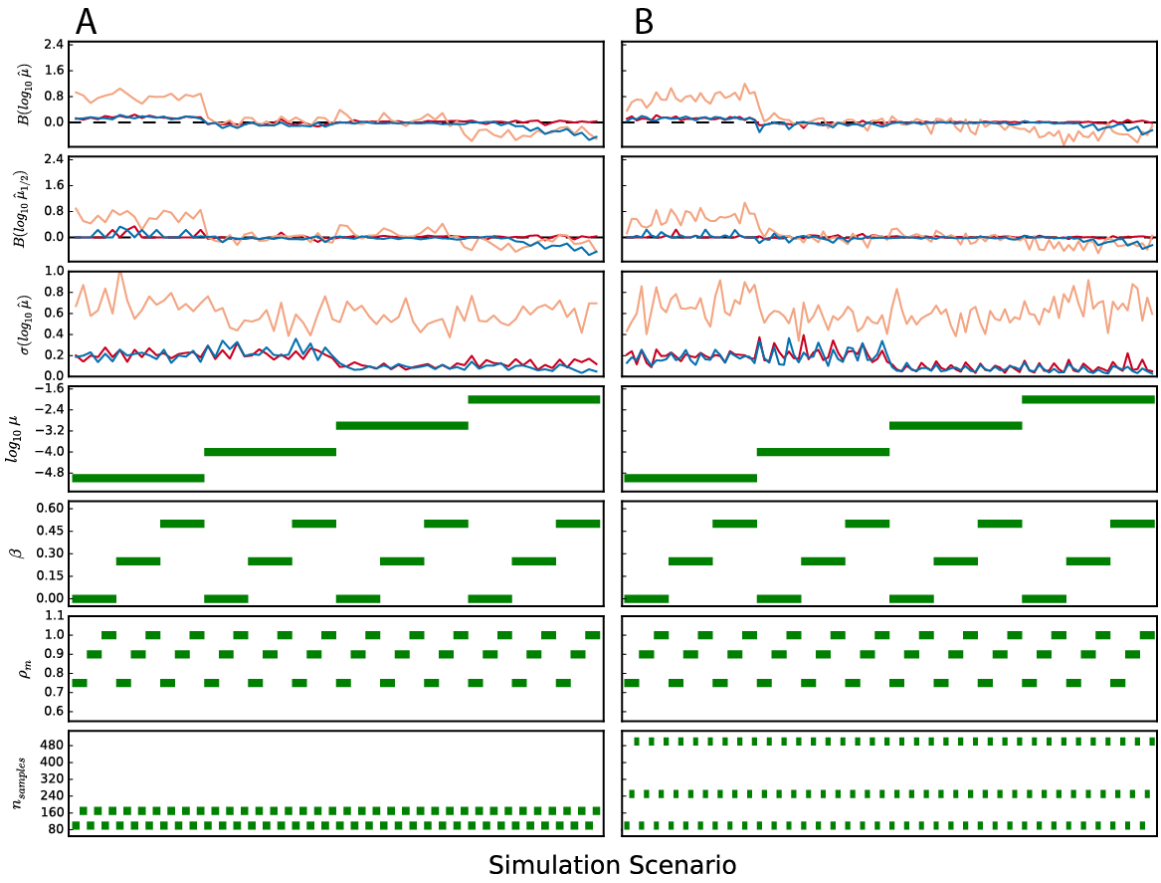

Figure S2: Simplifying mutation models results in biased mutation rate estimates STR genotypes were simulated for a variety of sample sizes and mutation models (bottom four rows) for both the Simons Genomes phylogeny (A) and 1000 Genomes phylogeny (B). Across 25 iterations for each simulation scenario, mutation rates computed after assigning each sample's genotype unity posterior probability are biased (top two rows) if the estimated model is restricted to singe-step mutations (orange) or no length constraint (blue) but not if the estimated model is unconstrained (red)

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

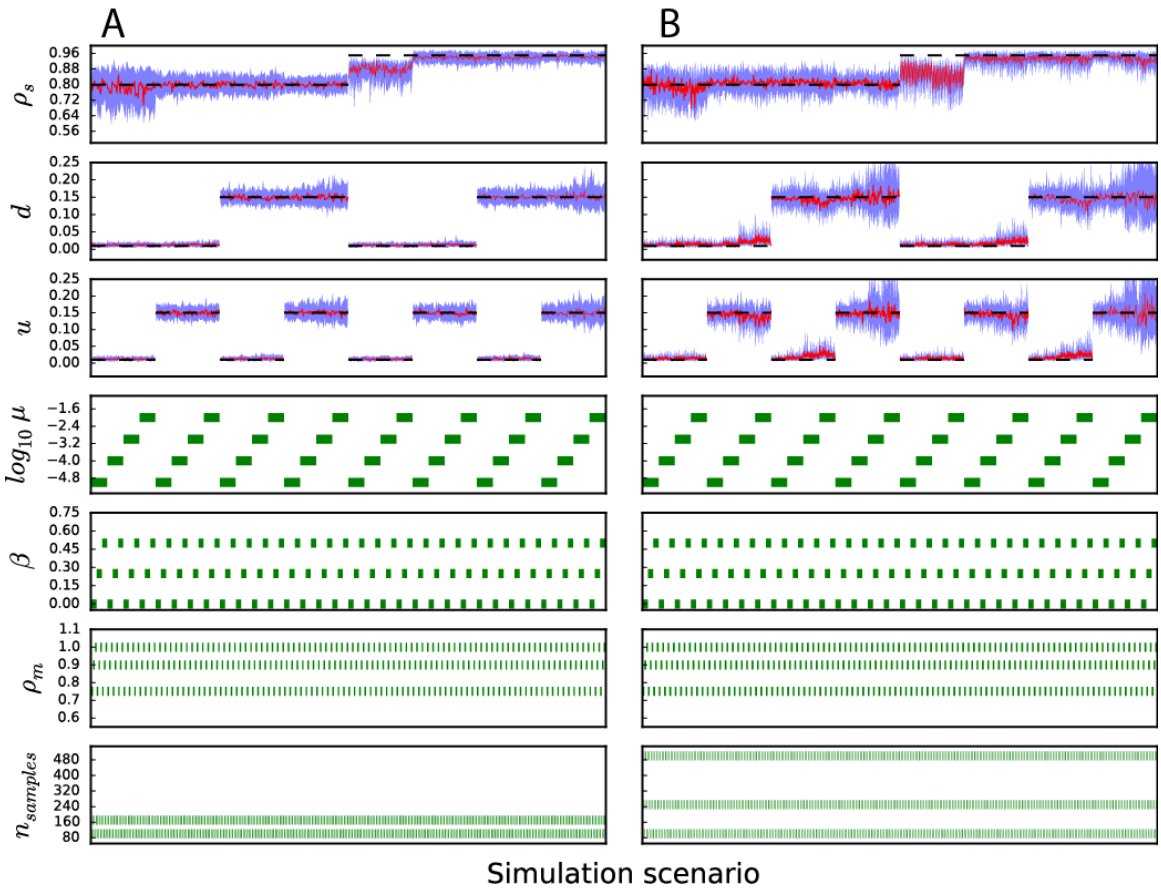

Figure S3: The EM-based stutter estimation method accurately recovers the underlying stutter model

STR genotypes were simulated for a variety of sample sizes and mutation models (green lines in bottom four rows). Using various PCR stutter model parameters (dashed black lines in top three rows), observed reads were generated for each of the samples. In conjunction with the EM stutter estimation algorithm, we utilized these reads to estimate the simulated stutter model. The concordance between the median parameter estimates (red lines) across 25 iterations of each scenario and the true parameters reflects the algorithm's ability to obtain robust estimates. Blue lines indicate the lower and upper quartiles of the estimates for each scenario. A, 1, 2, 3, 4, 5 or 6 observed reads were generated for 19%, 27%, 21%, 15%, 8% and 10% of the samples using the Simons Genome phylogeny, respectively. B, 1,2 or 3 observed reads were generated for 65%, 25% and 10% of the samples using the 1000 Genomes phylogeny, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

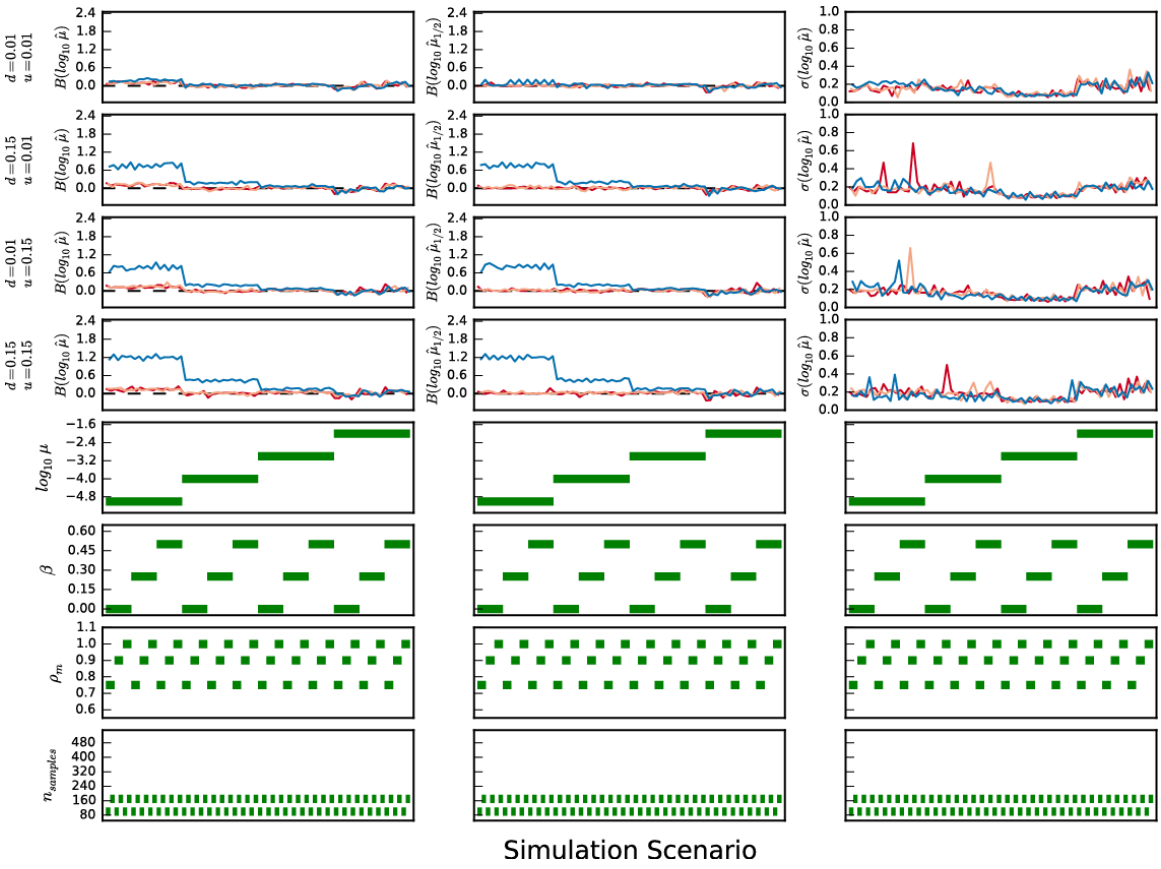

Figure S4: Estimating mutation rates from stutter-affected reads using the Simons Genomes phylogeny

STR genotypes were simulated for a variety of sample sizes and mutation models (bottom four rows) using the Simons Genome phylogeny. Each sample's genotype and four different stutter models (d and u in top four rows) were then used to generate 1, 2, 3, 4, 5 or 6 observed reads for 19%, 27%, 21%, 15%, 8% and 10% of samples. Across 25 iterations for each simulation scenario, genotyper posteriors computed using the fraction of supporting reads (blue) resulted in markedly biased mutation rate estimates (first two columns), while posteriors computing using the exact stutter model (red) and EM stutter model (orange) resulted in relatively unbiased estimates with similar standard deviations (third column).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

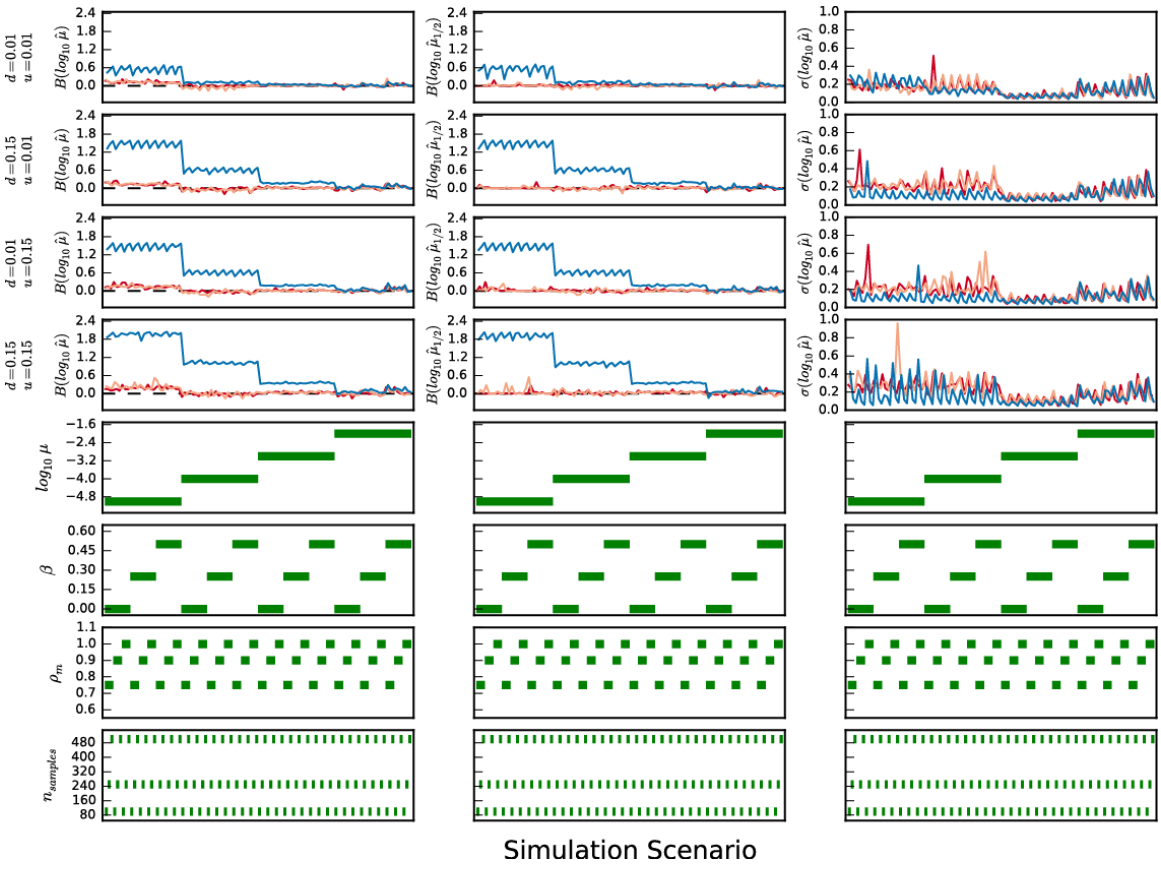

Figure S5: Estimating mutation rates from stutter-affected reads using the 1000 Genomes phylogeny

STR genotypes were simulated for a variety of sample sizes and mutation models (bottom four rows) using the 1000 Genomes phylogeny. Each sample's genotype and four different stutter models (d and u in top four rows) were then used to generate 1, 2, or 3 observed reads for 65%, 25% and 10% of samples. Across 25 iterations for each simulation scenario, genotyper posteriors computed using the fraction of supporting reads (blue) resulted in markedly biased mutation rate estimates (first two columns), while posteriors computing using the exact stutter model (red) and EM stutter model (orange) resulted in relatively unbiased estimates with similar standard deviations (third column).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

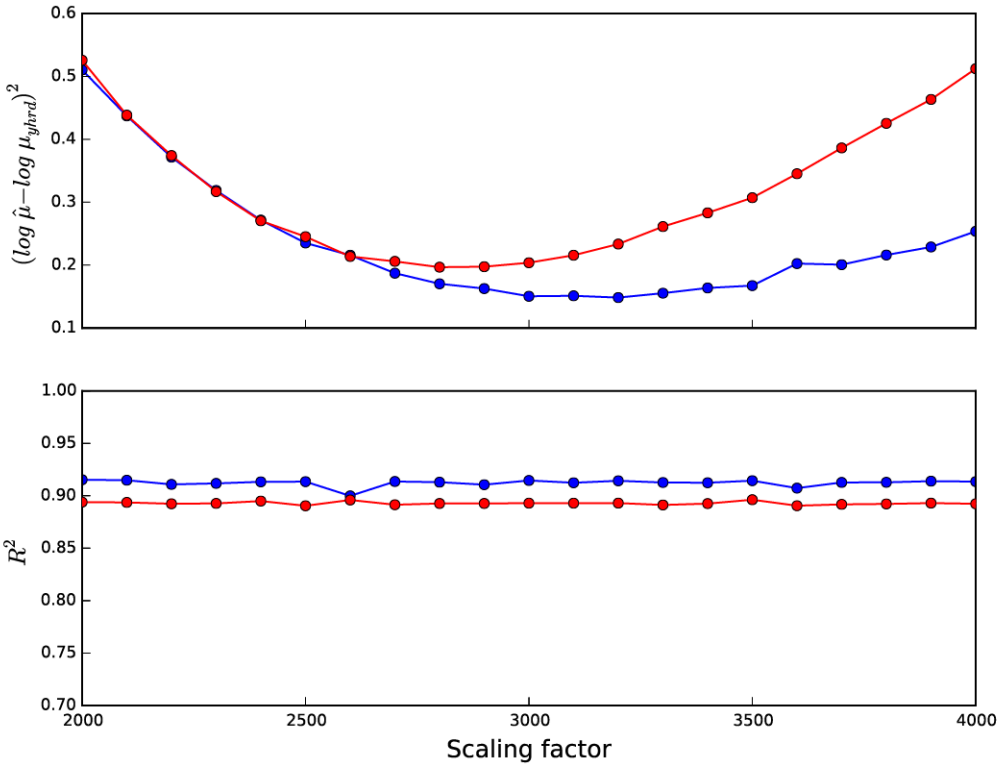

Figure S6: Scaling the Y-SNP phylogenies
Mutation rate estimates for loci in the Y-Chromosome Haplotype Reference Database were compared to estimates for the same loci obtained using the Simons Genome Project (blue), the 1000 Genomes Project (red) and a range of scaling factors. While the scaling factor had little effect on the R^2, it substantially impacted the total squared error in the log estimates. The values minimizing this squared error were chosen as the optimal factor for each phylogeny.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

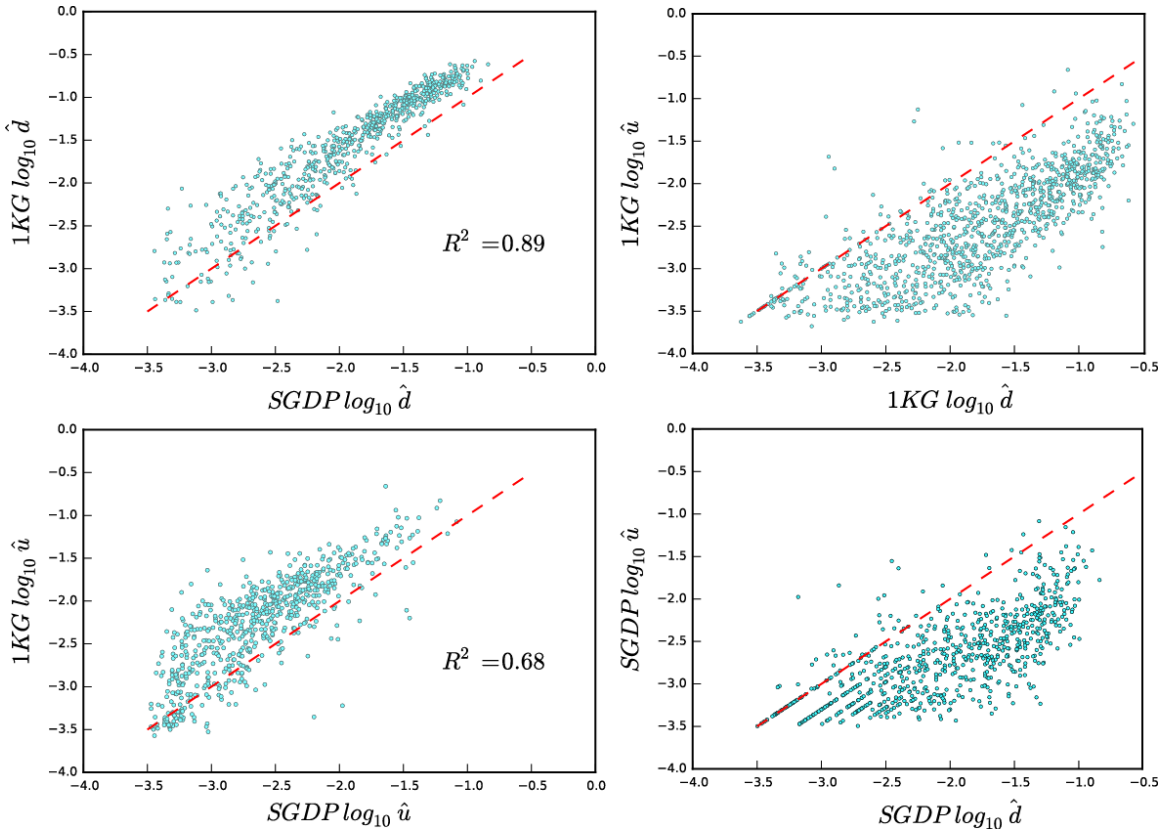

Figure S7: Relationship between stutter probabilities within and across datasets For a given Y-STR locus, the probabilities of stutter increasing (u) or decreasing (d) the size of the STR in each read were highly correlated (first column). However, the 1000 Genomes stutter rates largely fell above the diagonal (red line), indicating the higher rates of stutter in this dataset. Within each dataset, nearly all loci had a higher rate of downward stutter than upward stutter (second column).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

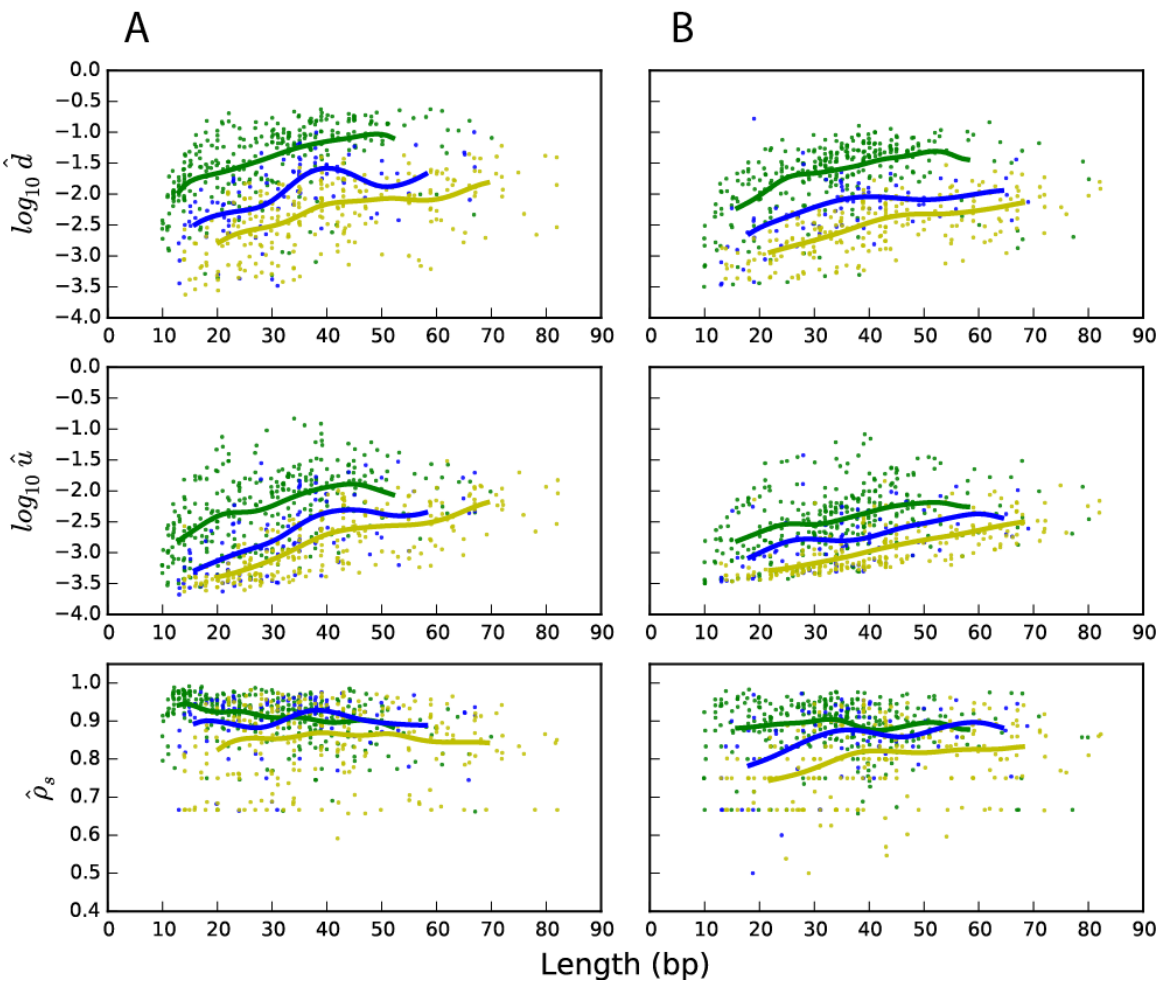

Figure S8: Sequence-based determinants of stutter probabilities
The learned stutter models for the 1000 Genomes (A) and Simons Genomes (B) datasets suggest that the probability of stutter increasing (u) or decreasing (d) the size of the STR rose with allele length for loci with di- (green), tri- (blue) and tetranucleotide (yellow) motifs. Each point denotes the estimate and reference allele length for a single Y-STR in each dataset, while the solid lines indicate lines fit using a second-degree kernel.

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

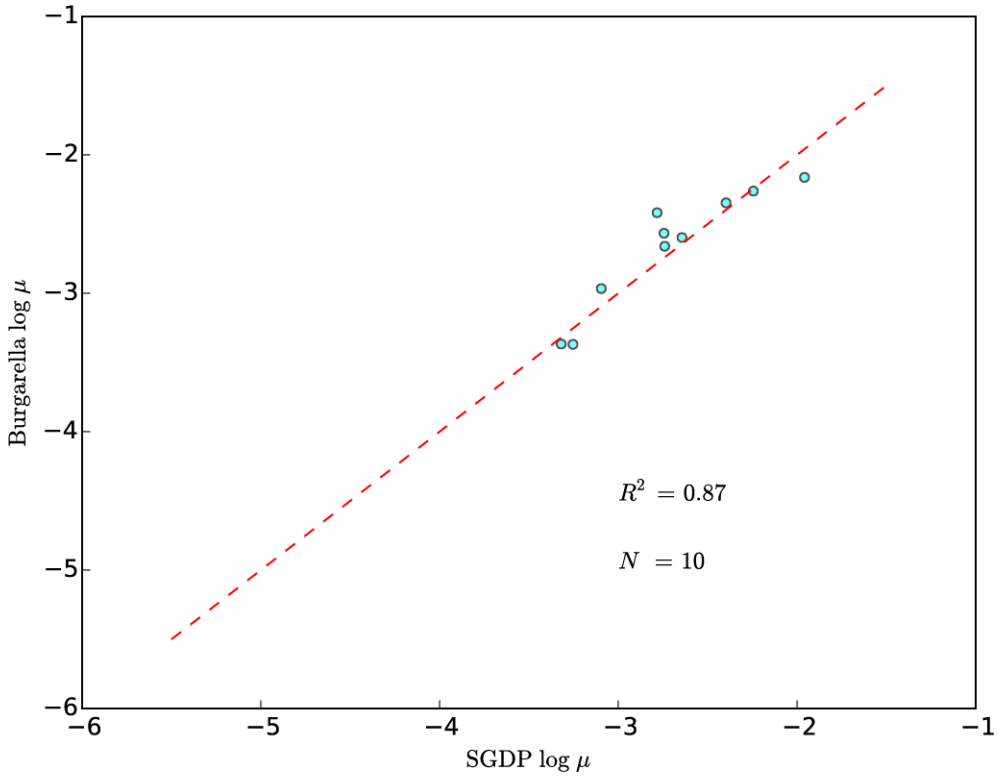

Figure S9: Concordance between SGDP estimates and Burgarella estimates based on large numbers of father-son pairs

Ten mutation rate estimates generated by Burgarella et al. using more than 5000 father-son pairs are highly concordant with estimates from the SGDP data and largely fall along the diagonal (red line).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

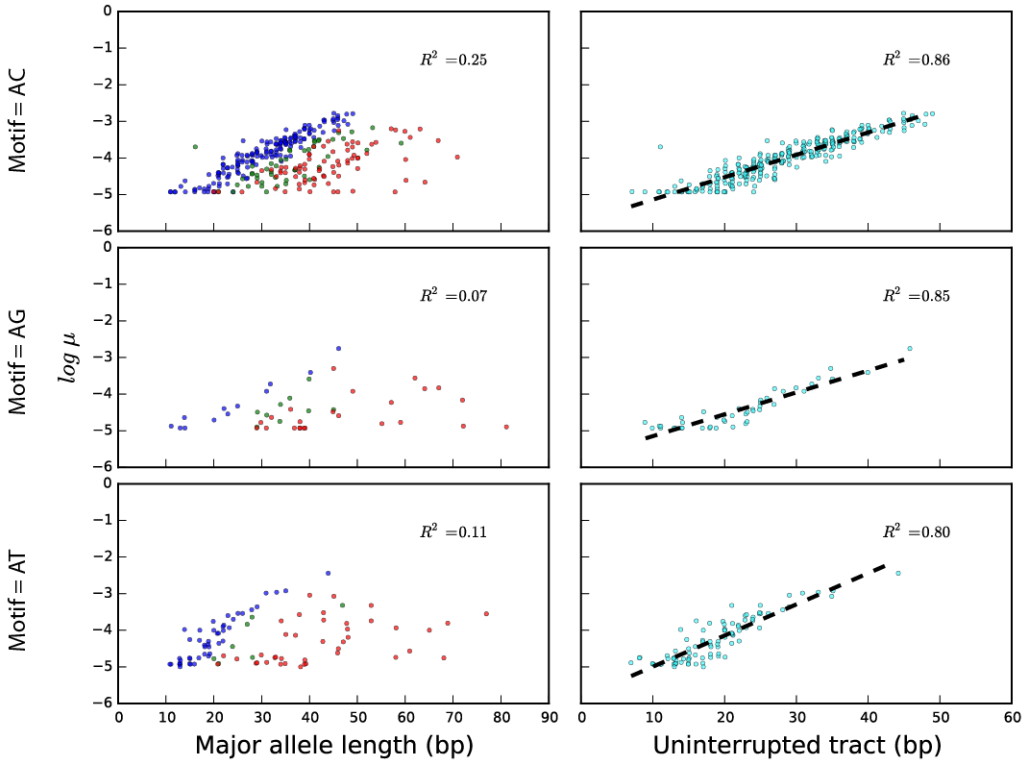

Figure S10: Sequence determinants of Y-STR mutability for loci with dinucleotide repeat units

Stratified by repeat motif (rows) and major allele length, loci with no interruptions to the repeat structure (blue) are generally more mutable than those with one interruption (green) or more than one interruption (red). While major allele length is a poor predictor of mutability, the length of the longest interrupted tract is a very strong predictor of the log mutation rate for each motif length (cyan).

bioRxiv preprint doi: https://doi.org/10.1101/036590; this version posted January 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

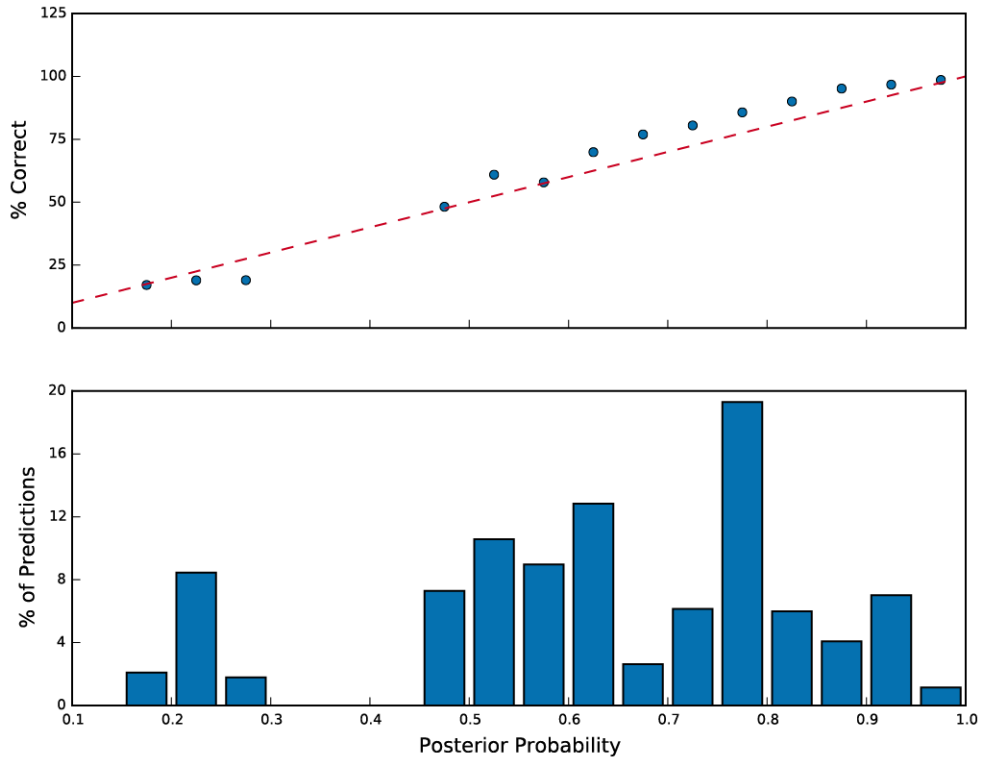

Figure S11: Y-STR imputation results in well-calibrated posteriors
Y-STR genotypes for each locus in the PowerPlex Y23 panel were imputed across 1000 iterations using a reference panel of 500 samples and 70 imputed samples. The accuracy for each posterior probability bin (top panel) largely followed the diagonal (red line), reflecting that the imputation probabilities reflect the true probability of correct imputation.