Interpreting short tandem repeat variations in humans using mutational constraint

SUPPLEMENTAL MATERIAL

Supplemental Notes

Supplemental Note 1: The traditional stepwise mutation model does not capture observed trends in STR variation

Our mutation rate estimation method relies on an STR mutation model that fits well to observed population-level data. While a variety of STR mutation models have been proposed[1], the most widely and traditionally used is the generalized stepwise mutation model (GSM), which allows STRs to add or delete one or more repeat units at each mutation with equal probabilities of expansion or contraction. Under this model, variance in allele size, and therefore squared allelic distance, should grow linearly with time according to random walk theory. However, multiple orthogonal lines of evidence suggest that a length-dependent bias in mutation direction is a key component of STR mutation. First, studies of *de novo* STR mutations consistently find a bias in mutation direction: the longest alleles are more likely to contract, whereas the shortest alleles are most likely to expand[2,3]. Second, looking across more distant time scales, variance in allele size nearly always grows substantially sublinearly with time to the most recent common ancestor (TMRCA) (**Figure 1B**). This saturation in the molecular clock over time is quite different than the linear trend predicted by the GSM (as depicted in Figure 3 of Sun *et al.*[2]). Taken together, these observations strongly suggest that mutational bias toward a central "optimal" allele length is a critical feature of STR evolution[1].

A length-biased version of the GSM more accurately describes observed STR mutation trends. As pointed out by Garza *et al*.[4], this is reminiscent of an Ornstein-Uhlenbeck (OU) stochastic process, which describes Brownian motion of an object with a spring-like force pushing that object back toward the central value. As the particle gets farther from the center, the force to go back toward the center increases. Here, we develop a discretized version of this process, which recapitulates the observed saturation in the STR molecular clock over time (**Supplemental Figure 1A,B**). Importantly, our model can be seen as an extension of GSM allowing for a length constraint: if the length constraint is set to 0, the OU model is equivalent to GSM. Specific details and limitations of our model are described in the main text and in **Supplemental Note 2.**

Supplemental Note 2: Modeling STR mutation as a discrete multi-step Ornstein-Uhlenbeck process

Introduction

The classical OU process describes the position of a continuous variable over time. However, STR mutations come in discrete step sizes. Miao[5] outlines a discrete analog of the OU, but the model only allows steps that increase or decrease by a single unit. This is insufficient for modeling STR mutations, as it is well known that STRs may mutate by more than one repeat unit in a single mutation.

Here we develop a generalized version of Miao's discrete OU process that allows steps of more than one unit. In **Supplemental Note 3**, we show that step size distributions from this model closely match those observed from *de novo* mutations. We use this to provide a realistic model of STR mutation, which serves as the basis for our maximum likelihood mutation rate estimation described in the main text.

Overview of the OU process

An OU process is described by the stochastic differential equation:

$$dx_t = \beta(\theta - x_t)dt + \sigma dB_t$$

where θ is the long run mean, x_t is the value at time t, β is a length constraint which pushes x back toward θ, σ is the standard deviation of the step size, and B is Brownian motion. For convenience, we assume θ is equal to 0, and for our purposes outlined below we do not need to know the value of θ.

This is a well-characterized process with properties[6]:

$$E[x_t \mid x_o = v] = ve^{\beta t}$$
$$Var[x_t \mid x_o = v] = \frac{\sigma^2(1-e^{-2\beta t})}{2\beta}$$

We are interested in the step size distribution and how that relates to STR mutations. Using the Markov Property and assuming $\theta = 0$:

$$E[x_{t+\Delta t} | x_t = v] = E[x_{\Delta t} \mid x_0 = v] = ve^{-\beta \Delta t}$$

and so:

$$E[x_{t+\Delta t} - x_t \mid x_t = v] = ve^{-\beta \Delta t} - v = v(e^{-\beta \Delta t} - 1) = -\beta v\, \Delta t + o(\Delta t)$$

For the variance:

$$Var[x_{t+\Delta t} - x_t \,|x_t = v] = \frac{\sigma^2(1-e^{-2\beta\Delta t})}{2\beta} = \sigma^2\Delta t + o(\Delta t)$$

So the step size has mean and variance of approximately $-\beta v$ and σ^2, respectively. For the continuous OU, a Gaussian process, the step sizes are drawn from $N(-\beta x,\ \sigma^2)$.

Discrete single-step OU

Miao derives a discrete version of the OU process allowing for steps of single units. He denotes the process as X_t^h with a tick size h, where each step increases or decreases the value of X_t^h by a single tick. For our purposes, an allele of size x_i increases by h with probability u_i, decreases by h with probability d_i, and stays the same with probability $1-(u_i+d_i)$. This is then matched to the continuous OU process by matching the values of the first two moments, an established procedure for discretizing continuous stochastic processes. Writing the first two moments of $E[X_{t+\Delta t}\,|X_t = x]$ in terms of u_i and d_i is straightforward (taken from Miao):

$$E[X^h_{t+\Delta t}|X^h_t = x_i] = u_i\Delta t(x_i+h) + (1-(u_i+d_i)\Delta t)x_i + d_i\Delta t(x_i-h) + o(\Delta t)$$
$$E[(X^h_{t+\Delta t})^2|X^h_t = x_i] = u_i\Delta t(x_i+h)^2 + (1-(u_i+d_i)\Delta t)x_i^2 + d_i\Delta t(x_i-h)^2 + o(\Delta t)$$

Setting these equal to the first two moments of the continuous OU and dropping the $o(\Delta t)$ term gives:

$$u_i = \frac{\sigma^2 + h\beta(\theta - x_i)}{2h^2}$$
$$d_i = \frac{\sigma^2 - h\beta(\theta - x_i)}{2h^2}$$

Since u_i and d_i are probabilities, they must be between 0 and 1. Imposing $u \geq 0$ and $d \geq 0$ gives a possible range of states. If X_t^h goes outside this range, we set u_i and d_i to 0 or 1 appropriately to force X_t^h back inside these boundaries.

Although not discussed by Miao, there are also limitations on the value of σ^2. This value describes the variance of the step size distribution. Here the steps only take values of -1, 0, or 1, and the variance of that distribution will always be at most 1. Therefore, we impose the additional restriction here that $\sigma^2 \leq 1$.

Discrete multi-step OU

The model described above only allows step sizes of a single unit. Here we extend this model to allow larger steps in some cases. The multistep discrete OU will be denoted as X_t^d and can be described as follows:

- Draw a step size k from a distribution D, where $k \in \{1, 2, \ldots \infty\}$ and D describes $P(k = i)$ with the requirements $\sum_{i=1}^{\infty} P(k = i) = 1$; $0 \leq P(k = i) \leq 1 \ \forall i$. Below $P(i)$ denotes the probability that we draw a step size i from distribution D.
- With probability u_i, X_t^d will change by $+kh$, with probability d_i it will change by $-kh$, and with probability $1 - u_i - d_i$ it will change by 0. Note that in the case where we define D such that $P(1) = 1$, this is the same as Miao's single step discrete model. (Recall that h is the tick size following Miao's notation. In cases of modeling STR mutation we have set h always equal to 1).

The first two moments of this process can be written as:

$$E[X_{t+\Delta t}^d | X_t^d = x_i] = \sum_{j=1}^{\infty} u_i \Delta t P(j)(x + jh) \ + \ \sum_{j=1}^{\infty} d_i \Delta t P(j)(x - jh) + (1 - (u_i + d_i)\Delta t) x_i + o(\Delta t)$$

$$E[(X_{t+\Delta t}^d)^2 | X_t^d = x_i] = \sum_{j=1}^{\infty} u_i \Delta t P(j)(x + jh)^2 \ + \ \sum_{j=1}^{\infty} d_i \Delta t P(j)(x - jh)^2 + (1 - (u_i + d_i)\Delta t) x_i^2 + o(\Delta t)$$

Following the example above, we set the first two moments equal to the first two moments of the continuous case. This gives:

$$u_i = \frac{\sigma^2 E[D] + \beta h(\theta - x_i) E[D^2]}{2h^2 E[D] E[D^2]}$$

$$d_i = \frac{\sigma^2 E[D] - \beta h(\theta - x_i) E[D^2]}{2h^2 E[D] E[D^2]}$$

where $E[D] = \sum_{i=1}^{\infty} iP(i)$ is the expected value of the step size drawn from distribution D and $E[D] = \sum_{i=1}^{\infty} i^2 P(i)$ is the expected squared step size. Note that for Miao's model, we have $E[D] = E[D^2] = 1$ and u_i and d_i reduce to the single step case.

Limits on the state space and input parameters

The nature of this process enforces limits on the state space and input parameters. As in the single step case, we must have $u_i \geq 0$ and $d_i \geq 0$. Imposing this gives a state space limit of $[\theta - \frac{\sigma^2 E[D]}{h\beta E[D]^2}, \ \theta + \frac{\sigma^2 E[D]}{h\beta \ E[D]^2}]$. If X_t^d goes outside these boundaries, u_i or d_i will again be set to 0 or 1 appropriately to force it to a state within these bounds. For σ^2, the distribution will have the highest possible variance when $u_i = d_i = 0.5$. Since the distribution here is symmetric, its expectation is 0 and $\sigma^2 = E[D^2]$. We therefore enforce that $\sigma^2 \leq E[D^2]$. Note that in the case where $\sigma^2 = E[D^2]$, then $u_i + d_i = 1$ and there will

never be a step of size 0. If σ^2 is less than this, the 0 step size will receive non-zero probability and the measured mutation rate will have to be corrected to reflect this (see next section).

State holding time and its effect on mutation rate estimation

In the model description above, a step size of 0 can have a non-zero probability if $\sigma^2 < E[D^2]$. Therefore, even though we will generate mutations from this model at a rate μ, some of those "mutations" will result in no actual allele change. Therefore μ will be an overestimate of the rate of true mutation.

Note that at each mutation event, the probability of changing the allele is equal to $\lambda = u_i + d_i = \frac{\sigma^2}{h^2 E[D^2]}$. This is independent of the current state i. With $h = 1$, this is equal to $\lambda = \frac{\sigma^2}{E[D^2]}$. So $\lambda\mu$ will give the true per generation mutation rate. When estimating mutation rates, μ should be adjusted using this correction. For all future discussions we assume σ^2 is the maximum value, which avoids this correction.

Example discrete multi-step model

For a concrete example, assume we choose D to follow a geometric distribution with parameter p. Here p can be thought of as the probability that the step size is by a single unit. This distribution fits well to observed STR mutation sizes (**Supplemental Note 2**). Then we have $E[D] = \frac{1}{p}$ and $E[D^2] = \frac{2-p}{p^2}$. The up and down probabilities would be:

$$u_i = \frac{\sigma^2 p^2 + \beta h(2-p)(\theta - x_i)}{2h^2(2-p)}$$
$$d_i = \frac{\sigma^2 p^2 - \beta h(2-p)(\theta - x_i)}{2h^2(2-p)}$$

For the STR mutation model, we will assume the central allele $\theta = 0$ and $h = 1$. We will also set $\sigma^2 = E[D^2] = \frac{2-p}{p^2}$. This gives simplified up and down probabilities:

$$u_i = \frac{1 - \beta p\, x_i}{2}$$
$$d_i = \frac{1 + \beta p\, x_i}{2}$$

Note that when the current allele is 0, the step size distribution is symmetric. When the allele is less than 0, the step size distribution is weighted toward positive step sizes, and vice versa (**Supplemental Figure 1C**).

Limitations of the OU mutation model

We note that several well established features of STR mutation are not captured by our model, and represent potential future improvements:

- *Long STR expansions*: Most well known pathogenic STRs, such as those involved in Huntington's Disease or Fragile X Syndrome, are the result of large expansions of repeat tracks. Our method cannot currently be used to analyze repeat expansion loci for two major reasons. First, current tools for analyzing STRs are limited to loci that can be entirely spanned by a single read. For 100bp reads, this limits our detection to repeats of around 80bp or less in total length. Second, large pathogenic expansions clearly depart from the more stepwise mutation patterns observed at most shorter STRs, and thus likely occur under a different biological process. We hypothesize that the majority of STRs mutate under a stepwise model as described here, but that once a certain length threshold is crossed certain repeats become unstable and exhibit large mutation sizes that are more difficult to model.
- *Length-dependent mutation rate*: Our model assumes a single per-locus mutation rate. However, STR mutation rates have been shown to be dependent on allele length, with longer alleles more likely to mutate than shorter ones as modeled in Haasl and Payseur[7]. Our current model does not accommodate allele-specific mutation rates.
- *Interaction between alleles*: It has been hypothesized that interaction between the two STR alleles in an individual may shape mutation patterns[8]. Our model does not take this into account.

Supplemental Note 3: STR mutation properties observed from trio studies

Previous studies have examined properties of STR mutation by directly observing *de novo* mutation events. In this section, we summarize several of these studies and use the results to motivate our choice of model and parameters in the main text. **Supplemental Table 1** summarizes these studies spanning from Weber and Wong's early study of 24 mutations[9] to Sun *et al.*'s[2] recent examination of 2,058 mutations. Importantly, the vast majority of loci previously studied are di- and tetra- nucleotides that were ascertained specifically due to their high degree of polymorphism. Therefore, these

loci are unlikely to be representative of parameters of STR mutation genome-wide. However, they can be useful for gaining general insight into mutation patterns of STRs.

De novo *mutations are length-biased*

Nearly every study summarized in **Supplemental Table 1** observed a length-dependent bias in mutation direction. The three largest and most recent studies[2,3,10] showed that longer alleles are more likely to contract, and shorter alleles more likely to expand. Ellegren observed only a tendency of long alleles to contract, but he and Weber and Wong studied very few overall mutations and thus may have had limited power to detect this bias. Overall, these studies suggest a length bias is a key feature of STR mutation.

Our model, described in **Supplemental Note 2**, imposes a length bias denoted as β, which describes the pressure on alleles to mutate back to an "optimal" central allele. Mutation steps are drawn from a distribution of mean $-\beta x$, where x is the current allele length and assuming "0" is the optimal allele. β can be measured from mutation data by taking the negative slope of the best fit line for x_i vs. Δx_i, where x_i is the starting allele for mutation i and Δx_i is the mutation size.

Although no study to date has collected enough mutations to estimate this value per-locus, Sun *et al.* plot a similar relationship (population Z-score vs. proportion of mutations increasing) in aggregate across all loci analyzed. For tetranucleotides, assuming nearly all steps are of a single unit, the proportion of increasing mutations p_i can be converted to mean mutation size using the formula $E[\Delta x] = p_i - (1 - p_i) = 2p_i - 1$. Assuming differences in length Z-score are close to differences in repeat size, their Figure 2d suggests an estimate of $\beta = 0.3$ is reasonable. In reality this parameter is likely to vary between loci.

Observed step sizes follow a geometric distribution

Step-size patterns of *de novo* STR mutations varied markedly across repeat unit sizes: tetranucleotides almost always mutate by a single unit, whereas dinucleotides are much more likely to experience multi-step mutations. Little data exists for step sizes of homopolymers or tri-, penta-, and hexa-nucleotides, although Ballantyne *et al*'s results

suggest periods 3-6 tend to have single step sizes. Reported step size distributions fit well to a geometric distribution (**Supplemental Figure 2**), which we chose as our mutation model. **Supplemental Table 1** gives the estimated proportion of single-unit steps (p) and the corresponding value of σ^2. This parameter is also likely to vary significantly between loci. Because we could not obtain accurate per-locus estimates of σ^2, we also report "effective length constraint" $\beta_{eff} = \beta/\sigma^2$.

Supplemental Note 4: Calibrating standard errors on mutation rate estimates

Our mutation rate estimation method assumes that each pair of haplotypes is an independent observation. However, haplotype pairs often share some portion of evolutionary history, and thus this assumption is incorrect. This is especially true in our Y-STR mutation rate analysis, which considers all pairs of haplotypes, with each pair considered to be independent. Extensive tests of our method on simulated trees show that this lack of independence does not bias mutation rate results (**Figure 2, Supplemental Figure 4**). However, non-independence between data points is expected to artificially deflate standard errors, making estimates appear more precise than they are.

To correct for standard error deflation, we use an empirical method to scale standard errors such that the 95% interval, calculated as $log_{10}\hat{\mu} \pm 1.96SE$ indeed covers the true mutation rate with 95% probability. We found that standard error deflation grew approximately linearly with the absolute value of the log mutation rate (**Supplemental Figure 3A**), and thus scaled standard errors by a constant γ times the absolute value of the log maximum likelihood estimate, giving scaled confidence intervals of $log_{10}\hat{\mu} \pm 1.96\gamma|log_{10}\hat{\mu}|SE$.

We calibrated the constant γ by calculating a metric, denoted as "truth coverage" (C), which gives the probability that the true value of the mutation rate falls within the scaled 95% confidence intervals. In cases where the posterior probability distribution of the mutation rate is known, we calculate truth coverage for each locus as the total mass of

the PDF contained in the scaled confidence interval (method 1). The truth coverage of the dataset is the mean value across loci. Stated more formally:

$$C = \frac{1}{L} \sum_{i=1}^{L} \int_{log_{10}\hat{\mu}_i - 1.96SE_\gamma|log_{10}\hat{\mu}_i|}^{log_{10}\hat{\mu}_i + 1.96SE_\gamma|log_{10}\hat{\mu}_i|} \phi_i$$

Where $\hat{\mu}_i$ is the maximum likelihood mutation rate estimate at locus i, L is the total number of loci, and ϕ_i is the posterior probability distribution of the true value of the log mutation rate at locus i based on an orthogonal dataset. In the base case where the ground truth is known, the PDF simply consists of a point mass at the true value, and this metric gives the percent of loci for which the true value falls in the predicted confidence interval.

In cases where the ground truth dataset consists of the number of observed mutations (m) out of a number of total observed meioses (n), we calculate truth coverage by first constructing an empirical confidence interval on the number of mutations for a given n, then determining how often the observed mutations falls in this interval (method 2). Specifically, we using the following steps:

1. Calculate the scaled standard error as $SE' = SE_\gamma log_{10}\hat{\mu}$
2. Draw a mutation rate μ' from $N(\hat{\mu}, SE')$.
3. Draw a number of mutations m' from a binomial distribution with n trials and probability of success μ'.
4. Repeat Steps 1-3 1,000 times to generate a distribution of m'.
5. Determine whether the observed number of mutations m falls within the 5th to 95th percentile of the determined distribution for m'.
6. Calculate C as the percent of loci for which step 5 passes.

We used these two approaches to calibrate standard errors for simulated autosomal data (**Supplemental Figure 3B**), Y-STR mutation rates (**Supplemental Figure 3C-D**), and autosomal mutation rates (**Supplemental Figure 3E**). For simulated data, ground truth values were known and thus method 1 was used. Autosomal standard errors were calibrated using method 2 against *de novo* mutation data from Sun *et al.*[2] across 1,634 loci with an average of 2,136 meioses observed each and from *de novo* mutation data

published for the CODIS markers (http://www.cstl.nist.gov/strbase/mutation.htm) across 11 loci with an average of 754,996 meioses each. For both simulated and observed autosomal mutation datasets, setting γ between 1-1.5 resulted in the desired truth coverage of approximately 95%. For downstream analyses we used $\gamma = 1.2$ to scale genome-wide autosomal standard errors.

Y-STR standard errors were calibrated using method 1 against posterior distributions returned by MUTEA[11] (**Supplemental Figure 3C**) across 702 loci and using method 2 against observed *de novo* mutation rates from Ballantyne *et al*.[10] across 52 loci with an average of 1,700 meioses each (**Supplemental Figure 3D**). For both analyses, we performed error calibration using both the 1000 Genomes and the SGDP datasets. Notably, standard error deflation was significantly stronger in the 1000 Genomes data, likely a result of the higher degree of shared history between individuals compared to the diverse genetic backgrounds present in the SGDP dataset. For both datasets, γ was significantly higher for Y-STRs than for autosomal loci. This trend is expected: whereas the data points for autosomal estimation essentially consist of randomly chosen haplotype pairs, the Y-STR analysis considers all haplotype pairs and thus has a much higher degree of non-independence across data points. For Y-STR analyses, we used $\gamma = 8$ to scale standard errors for the 1000 Genomes and $\gamma = 6$ to scale estimates from the SGDP data.

Supplemental Note 5: Gene-level analysis of STR constraint

Overall, we computed constraint scores for 1,424 STRs in coding regions across 1,180 individual genes. Most genes (83%) contained only a single STR. 13% contained two STRs, 3% contained 3 STRs, and 1% contained 4 or more. Pairs of STRs in the same gene had moderately more similar constraint scores than compared to all pairwise comparisons (median difference 1.00 vs. 2.31), although this difference was not quite statistically significant (Mann-Whitney U test; p=0.054; n_1=155, n_2=1,012,862), suggesting different STRs in the same gene contain independent information.

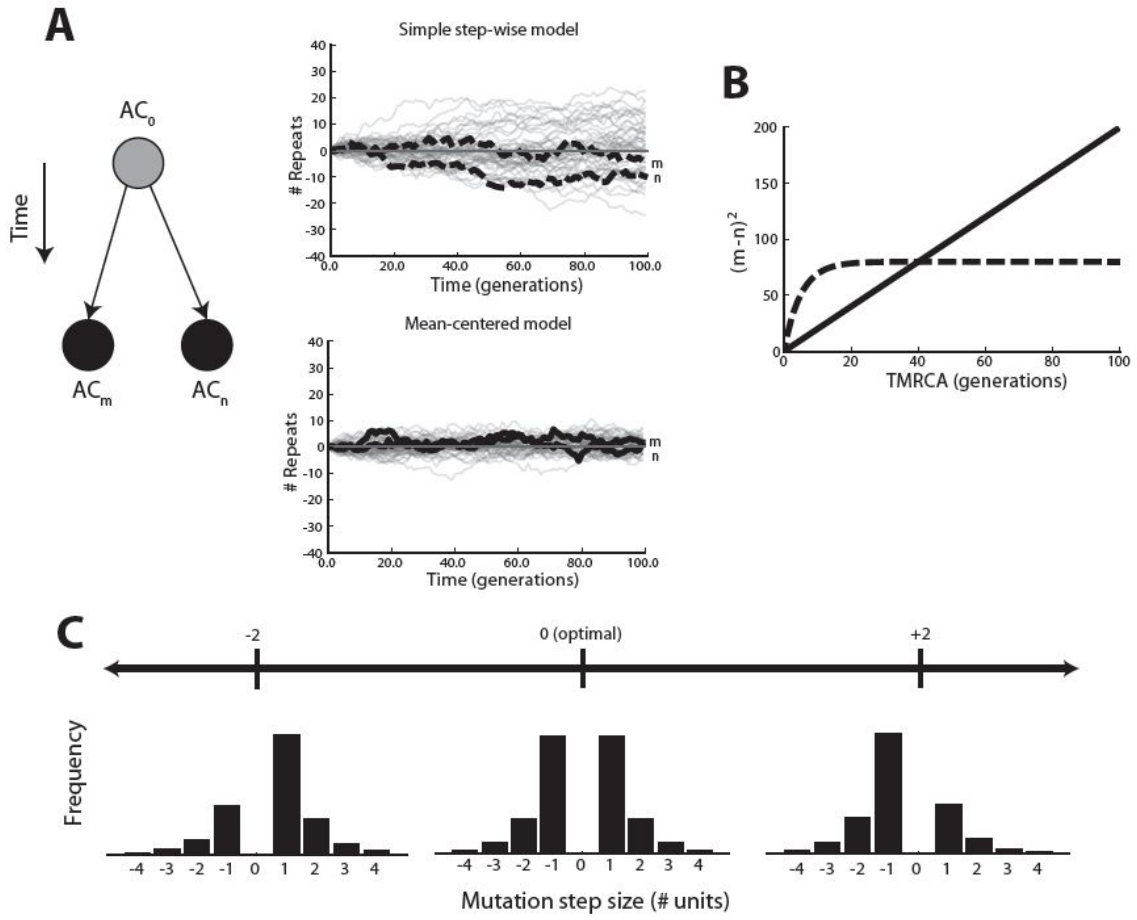
To determine whether our score implicates a role for STRs in genes with specific characteristics, we also examined the relationship between STR constraint and gene

expression levels across tissues as measured by GTeX[12]. Constraint scores were significantly stronger in the top 20% of expressed genes in nearly every tissue. STRs were most constrained in genes highly expressed in brain-related tissues (**Supplemental FIgure 14**). Intriguingly, this is consistent with the fact that most known pathogenic STRs results in neurological or psychiatric phenotypes[13].

We also compared STR constraint scores to gene-level scores computed by the Exome Aggregation Consortium (ExAC)[14] measuring tolerance to loss of function mutations (pLI scores) or missense mutations (missense Z score). Genes with high pLI scores (>0.9) had overall stronger STR constraint (Mann-Whitney U test; p=1.5e-97; n_1=463; n_2=7,564) (mean constraint -6.7 for high pLI, mean constraint -1.4 for low PLI). Similarly, genes with high missense Z scores (>3) had overall stronger STR constraint (Mann-Whitney U test; p=2.5e-51; n_1=272; n_2=7,755). On the other hand, for many genes the STRs and the SNPs tell a different story. 21 genes with pLI>0.9 had STRs with positive scores (not constrained). Interestingly, at least two of these STRs (in *ATN1* and *CACNA1A*) are involved in known Mendelian late-onset STR expansion disorders. Similarly, hundreds of highly constrained STRs are present in genes with low pLI and/or low missense Z scores. Thus, the two different variant types likely provide orthogonal sources of information in many cases.

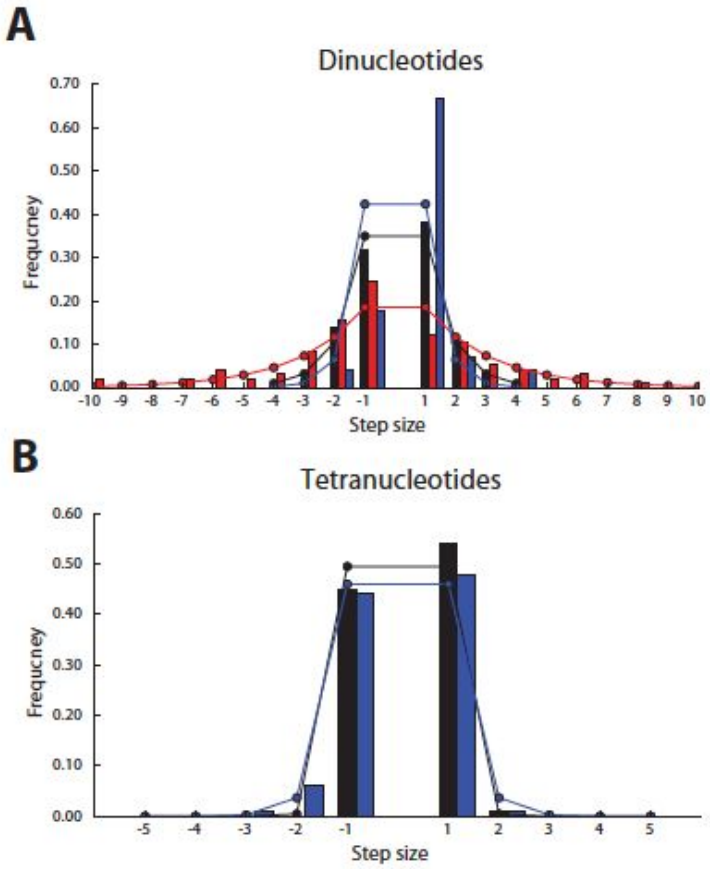
Supplemental Figures

Supplemental Figure 1



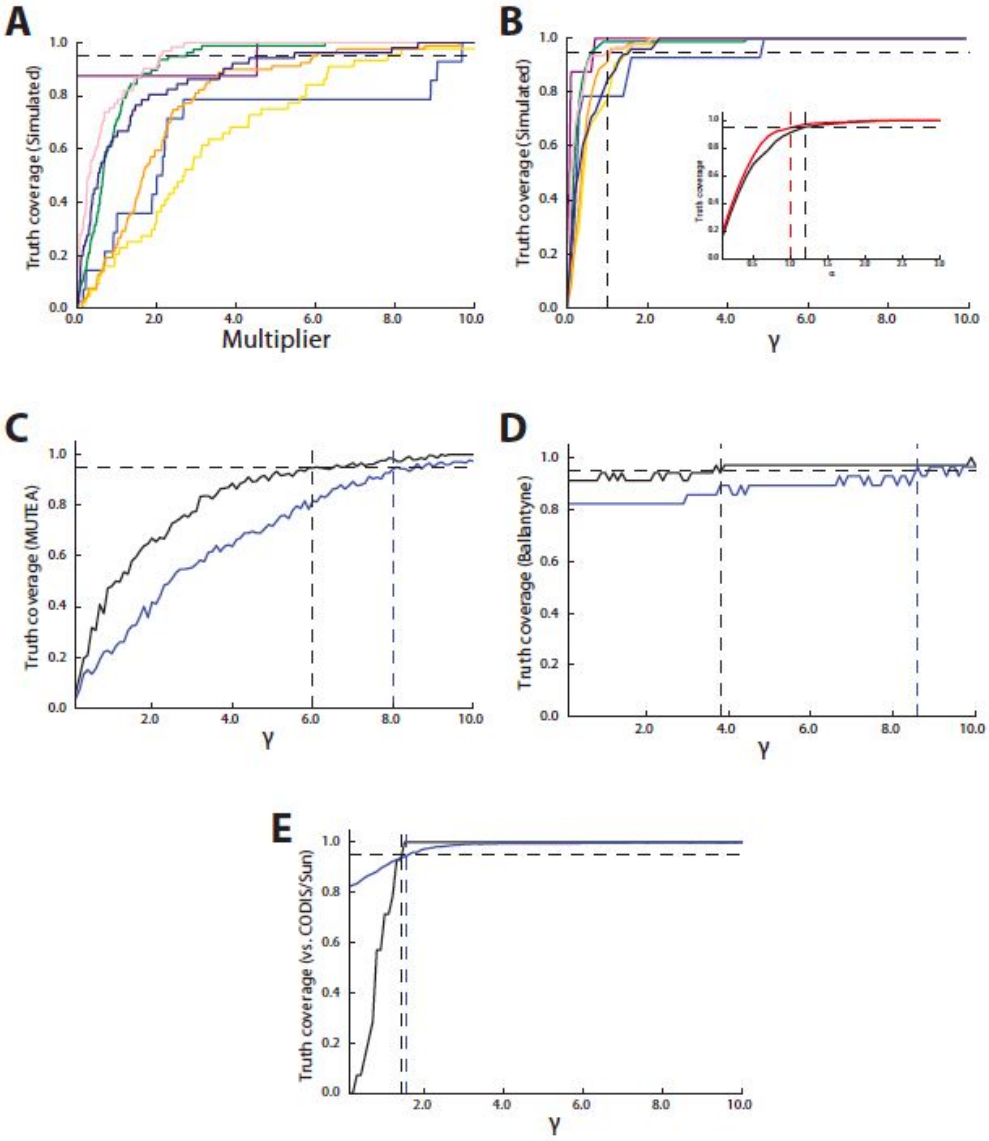
Modeling the STR mutation process. (A) A mean-centered random walk imposes a length constraint on allele size. The leftmost diagram represents two copies of an AC repeat descended from a common ancestor. The upper right plot shows the number of repeats at each leaf node vs. the number of generations passed for 100 simulations of a stepwise model (gray) and two example haplotypes (black, denoted as *m* and *n*). The lower right plot shows the same simulations using a mean-centered model. (**B**) **Allelic variance saturates at large TMRCAs.** The solid and dashed lines represent the relationship between squared allele difference and TMRCA under the stepwise model and mean-centered model, respectively. The mean-centered scenario recapitulates the saturation in the STR molecular clock observed in population data. (**C**) **Example step size distributions for a mean-centered model.** Histograms give probability distributions for step sizes assuming current alleles of -2 (left), 0 (center) or +2 (right) repeats from the optimal allele.

Supplemental Figure 2



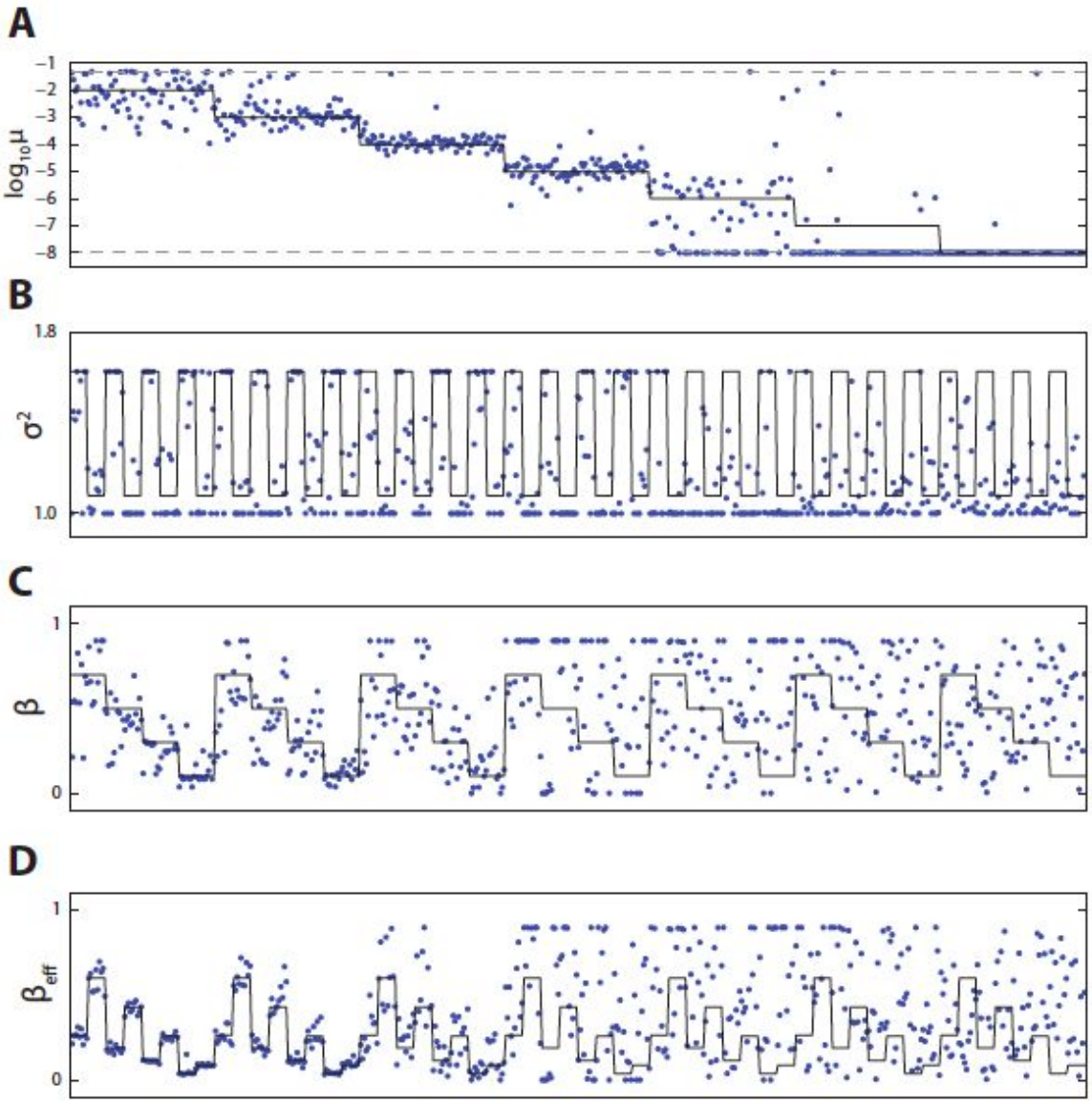
Previously reported step size distributions. (A) and **(B)** give step size distributions for dinucleotides and tetranucleotides, respectively. Black bars denote Sun *et al*.[2], red bars denote Huang *et al*.[3], and blue bars denote Ellegren[15]. Lines give the geometric distribution with parameter p, where p is the probability of a step of a single unit obtained from each study (**Supplemental Table 2**).

Supplemental Figure 3



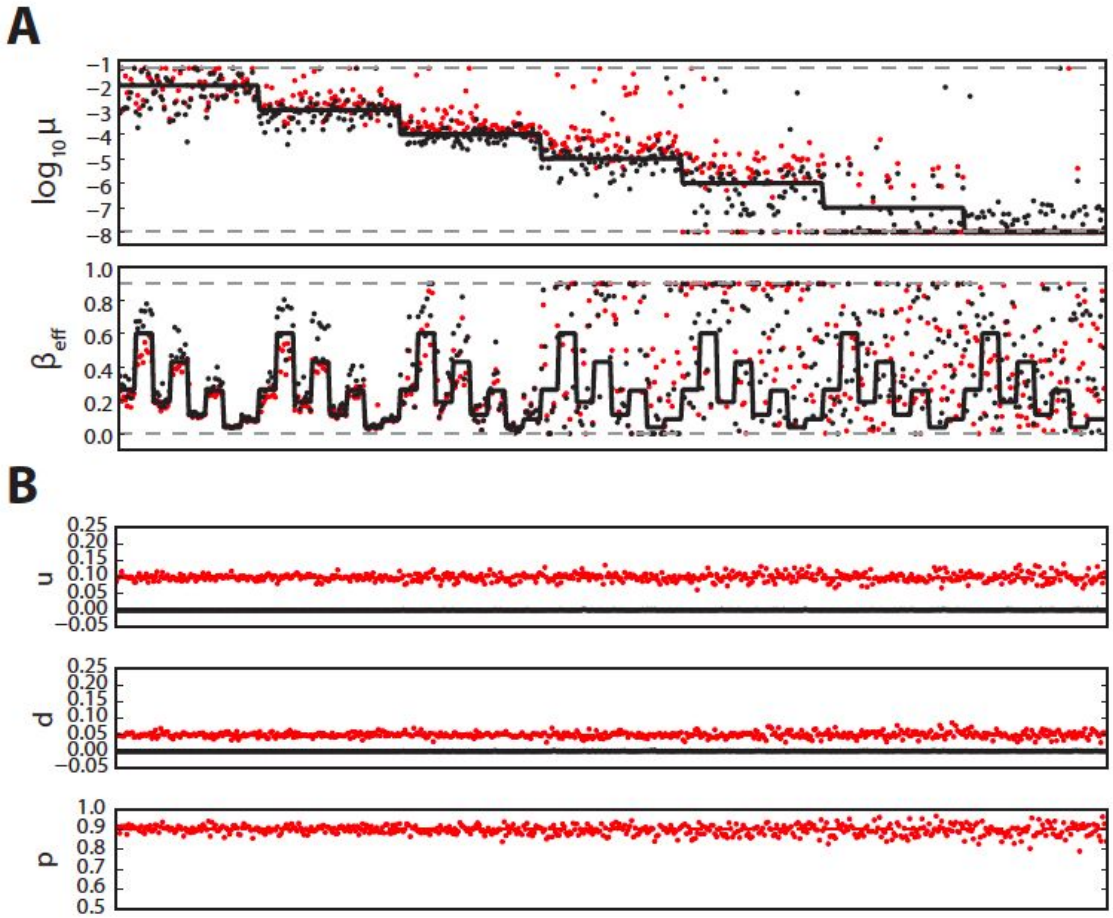
Calibrating standard errors (A) Multiplier vs. truth coverage on simulated data. For each mutation rate estimate, the standard error was multiplied by a constant (x-axis). Truth coverage is calculated as the percentage of loci for which the true mutation rate falls within the maximum likelihood estimate +/- 1.96*standard error*multiplier. Colors represent a range of simulated mutation rates: purple=10^{-8}, blue=10^{-7}, gold=10^{-6}, orange=10^{-5}, green=10^{-4}, pink=10^{-3}, darkblue=10^{-2}. **(B) Scaled multiplier (γ) vs. truth coverage.** The x-axis represents the multiplier from **(A)** scaled by the absolute value of the log of the maximum likelihood mutation rate estimate. Inset shows data aggregate across mutation rates for simulations with (red) and without (black) stutter noise. **(C) Calibrating γ against MUTEA Y-STR estimates.** Blue=1000 Genomes, Black=SGDP. **(D) Calibrating γ against Ballantyne Y-STR estimates.** Blue=1000 Genomes, Black=SGDP. **(E) Calibrating α for autosomal STRs.** Blue=autosomal STRs from Sun *et al.*[2] Black=published CODIS mutation rates http://www.cstl.nist.gov/strbase/mutation.htm.

Supplemental Figure 4



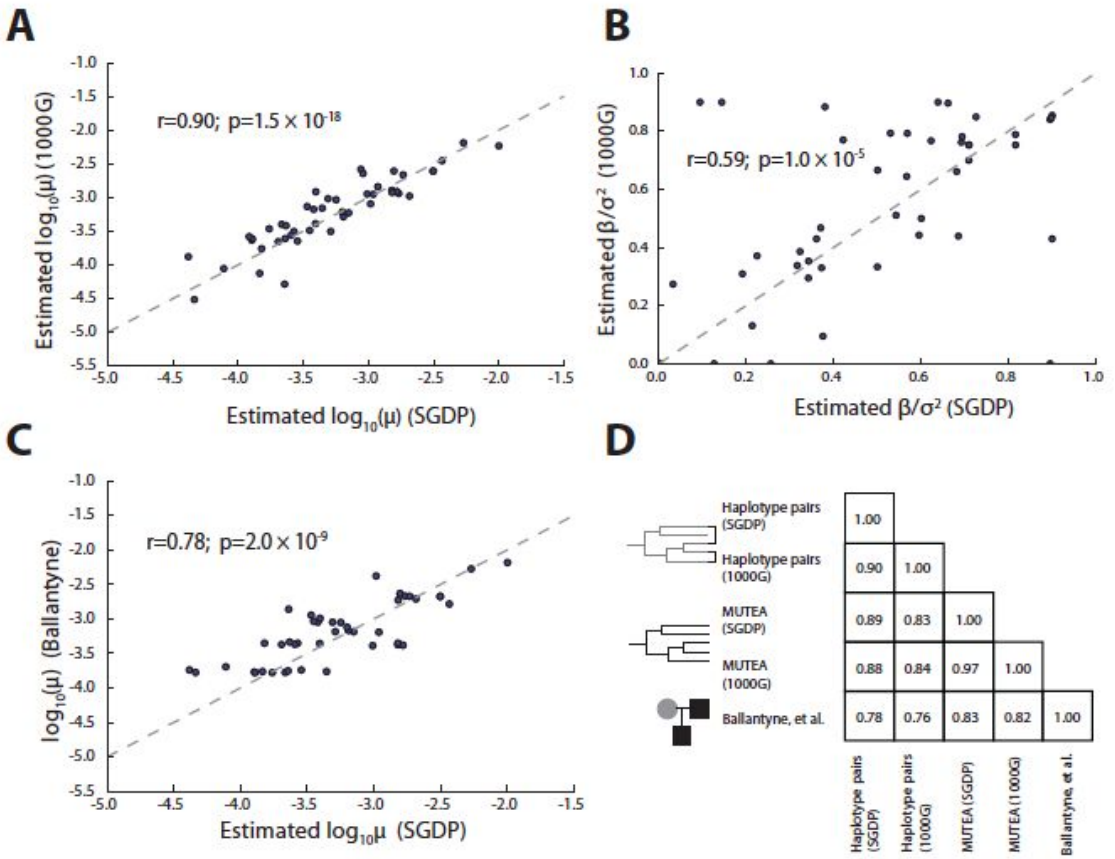
Per-locus simulation results. Plots are shown for the $\log_{10}$ mutation rate (**A**), step size parameter (**B**), length constraint (**C**), and effective length constraint, defined as β/σ^2. Black lines give simulated values. Blue dots give estimated values for each simulation.

Supplemental Figure 5



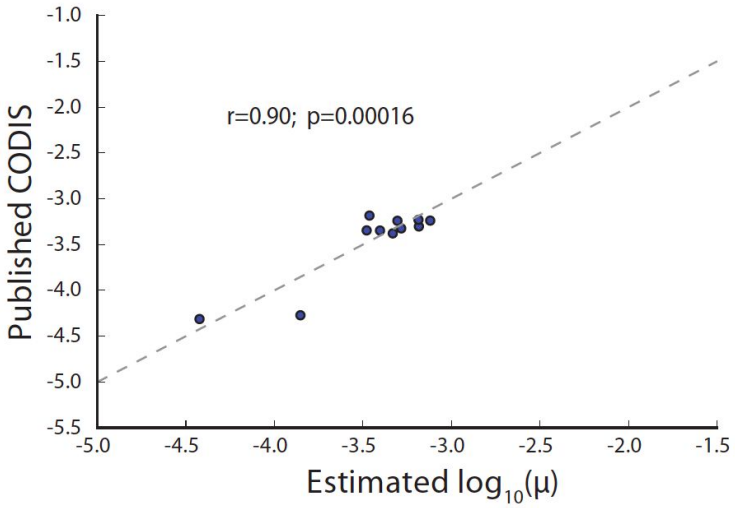
Modeling genotyping errors (A) Adjusting genotypes for stutter errors reduces bias. Solid black lines show simulated values of mutation rate and effective length constraint. Red dots give values estimated from genotypes with simulated stutter errors. Black dots give estimates after inferring stutter parameters and adjusting genotype likelihoods. Dashed gray lines give boundaries enforced during numerical likelihood maximization. **(B) Stutter parameters are accurately recovered from simulated data.** Black points represent data with no simulated stutter errors (d=0, u=0). Red points represent reads simulated with 5x coverage with stutter parameters u=0.1, d=0.05, and p=0.9.

Supplemental Figure 6



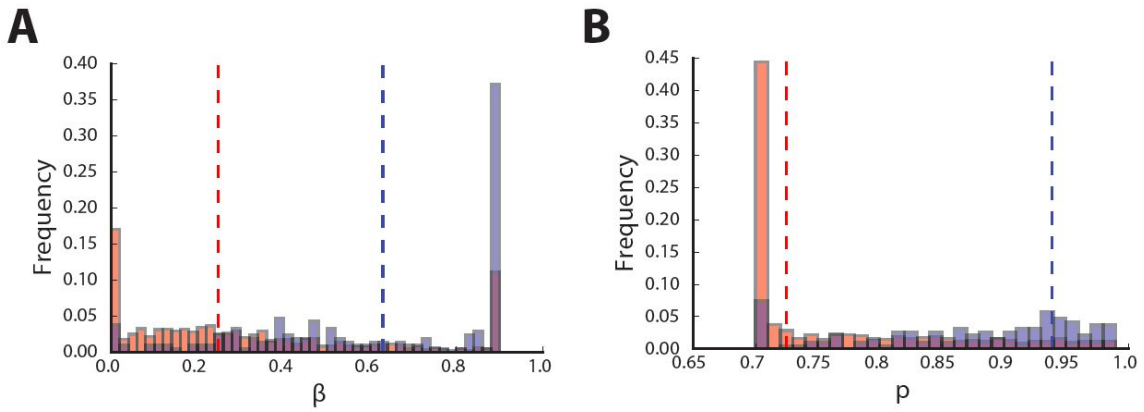
Validating mutation parameters at Y-STRs. Mutation parameter estimates for mutation rate **(A)** and effective length constraint **(B)** are highly concordant across datasets. **(C) Y-STR mutation rate estimates are concordant with *de novo* studies.** Each point represents a single Y-STR. Gray dashed lines denote the diagonal. (**D**) Pairwise Pearson correlation between Y-STR studies.

Supplemental Figure 7



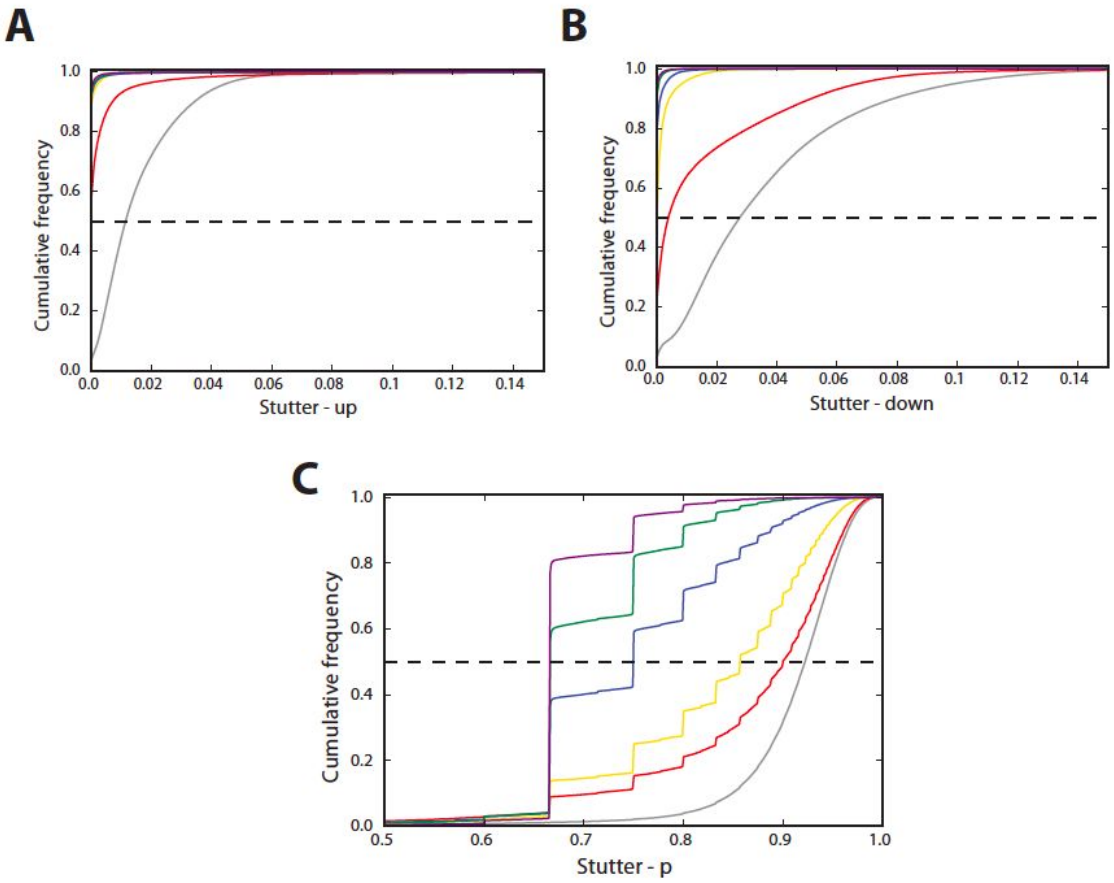
CODIS mutation rate estimates are concordant with *de novo* studies. Gray dashed line denotes the diagonal.

Supplemental Figure 8



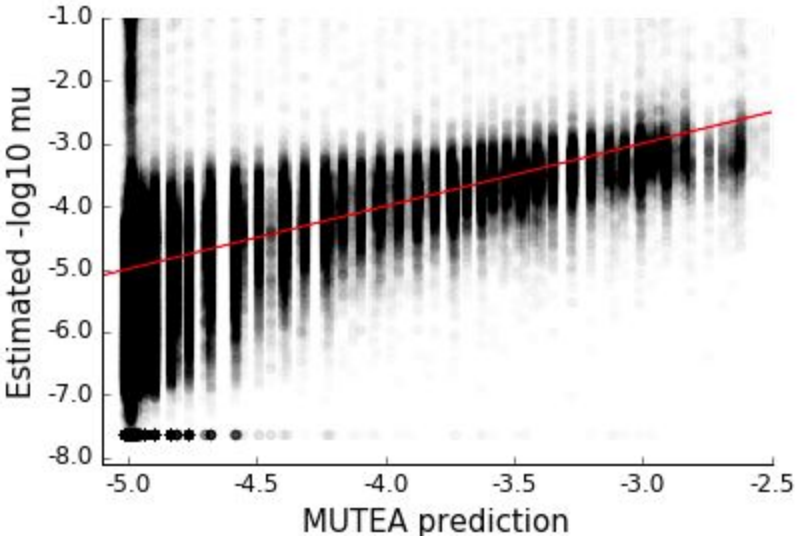
Comparison of autosomal mutation parameters with *de novo* studies. Shown are estimated length constraint (**A**) and step size parameter (**B**) for 1,634 STRs also analyzed by Sun *et al*.[2] Dashed lines give median estimate across loci. Solid lines give empirical mutation rate from trio data analyzed by Sun *et al.* Red=dinucleotides; blue=tetranucleotides. A comparison of mutation rates is shown in **Figure 2D**.

Supplemental Figure 9



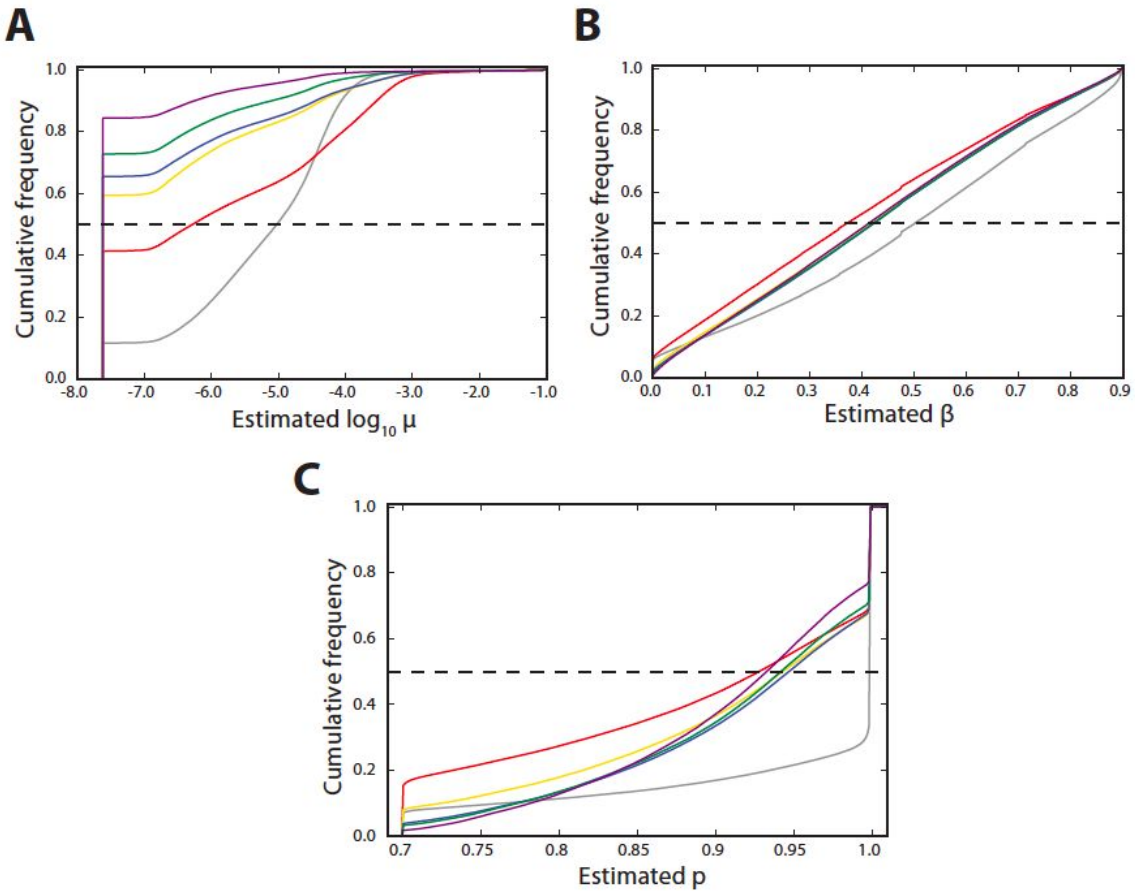
Per-locus stutter parameter estimates by repeat motif length. A. Probability of stutter to delete repeat units. B. Probability of stutter to insert repeat units. C. Parameter describing the geometric distribution of step sizes. Each plot shows the cumulative distribution across all autosomal loci. Gray=homopolymers, red=dinucleotides, gold=trinucleotides, blue=tetranucleotides, green=pentanucleotides, purple=hexanucleotides.

Supplemental Figure 10



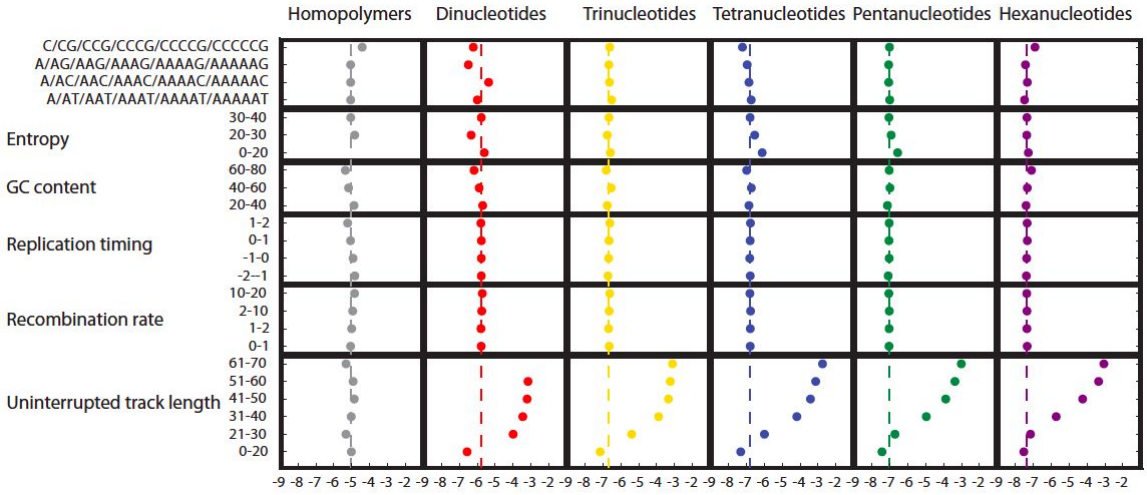
Comparison of per-locus mutation rate estimates vs. rates predicted by Willems, *et al.* [11] Each point represents a locus. The red line represents the diagonal.

Supplemental Figure 11



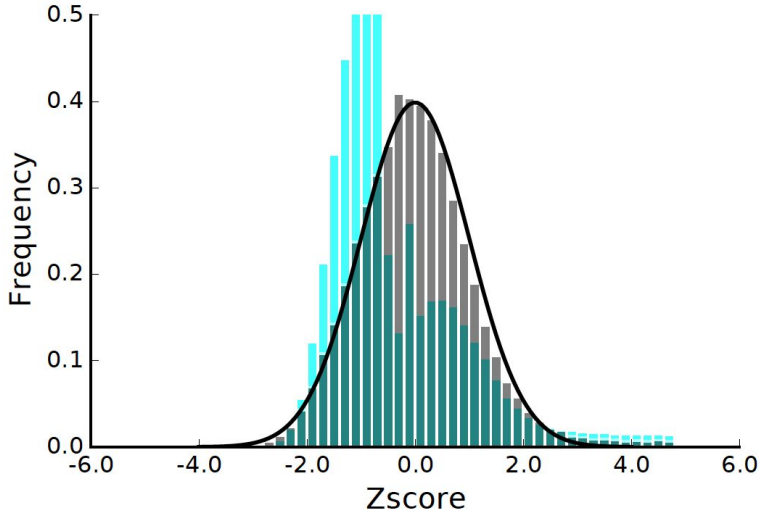
Per-locus estimates of mutation parameters. Plots give the cumulative distributions of the per-locus estimates of mutation rate (**A**), length constraint (**B**), and step size parameter (**C**). Dashed lines represent the 50th percentile. Gray=homopolymers, red=dinucleotides, gold=trinucleotides, blue=tetranucleotides, green=pentanucleotides, purple=hexanucleotides.

Supplemental Figure 12



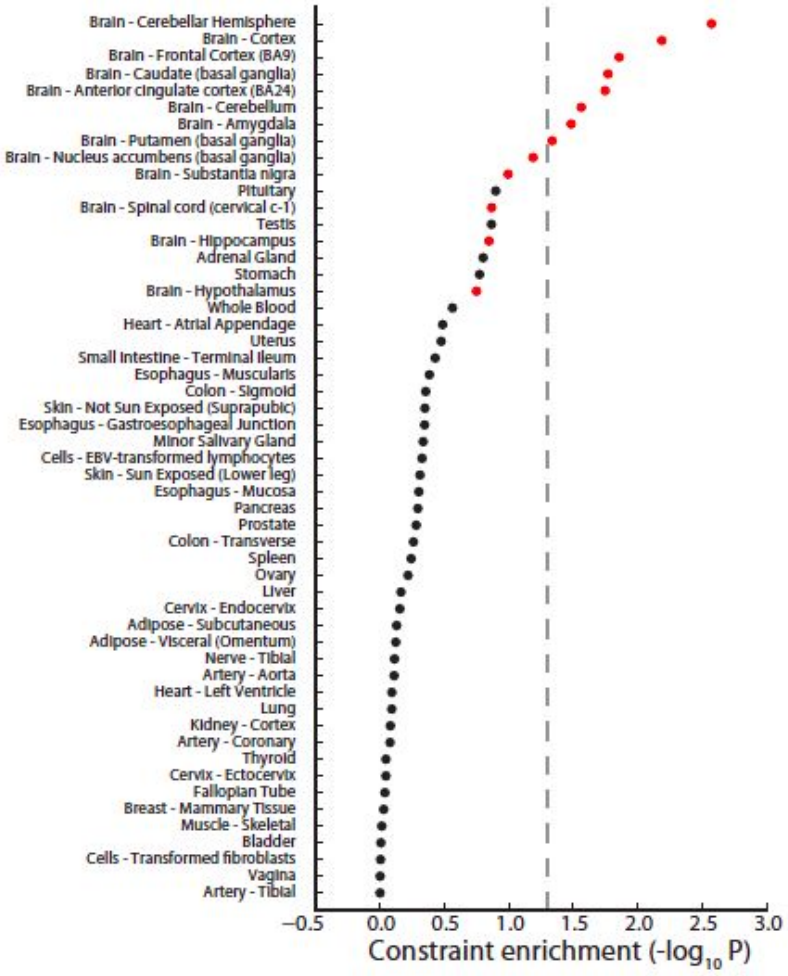
Relationship between mutation rate and local sequence features. Each dot represents the mean mutation rate for each category. The dashed line gives the mean mutation rate across all loci for each motif length. (gray=homopolymers, red=dinucleotides, gold=trinucleotides, blue=tetranucleotides, green=pentanucleotides, purple=hexanucleotides).

Supplemental Figure 13



Constraint score distribution. Distribution of constraint scores for loci with mutation rates not at the lower optimization boundary (gray) and for loci with high or undefined standard errors with mutation rates likely below our mutation threshold (cyan). Black line gives a standard normal distribution.

Supplemental Figure 14



Enrichment of constrained STRs in highly expressed genes. Red denotes brain tissues. Gray line gives p=0.05. Constraint score distributions were compared in the top 20% vs. the bottom 80% of expressed genes in each tissue.

Supplemental Tables

Supplemental Table 1

Study	# STRs	# Mutations	Mutation rate	Step size[1]
Weber and Wong[9]	28	24*	1.2×10^{-3}	$p_{di,tetra} = 0.87$; $\sigma^2_{di,tetra} = 1.49$
Ellegren[15]	52	102	-	$p_{di} = 0.85$; $\sigma^2_{di} = 1.59$ $p_{tetra} = 0.92$; $\sigma^2_{tetra} = 1.28$
Huang *et al*. 2002[3]	362	97	1.9×10^{-4}	$p_{di} = 0.37$; $\sigma^2_{di} = 11.91$
Ballantyne *et al*.[10] (Y-STRs)	186	924	3.78×10^{-4} to 7.44×10^{-2}	$p_{tri-hexa} = 0.96$; $\sigma^2_{tri-hexa} = 1.13$
Sun *et al*.[2]	2,477	2,058	10.01×10^{-4} (tetra) 2.73×10^{-4} (di)	$p_{di} = 0.68$; $\sigma^2_{di} = 2.85$ $p_{tetra} = 0.99$; $\sigma^2_{tetra} = 1.03$

Previously reported parameters of STR mutation

[1] p denotes the probability that a mutation is a single unit. σ^2 is calculated as $(2-p)/p^2$. Subscripts denote the length of the repeat motif.

*Verified and likely events reported in Table 1. Includes both di and tetranucleotides combined.

Supplemental Table 2

Motif length	p	u	d	u+d.
1	0.92	0.012	0.027	0.040
2	0.90	0.00070	0.0047	0.0056
3	0.86	0.00024	0.0011	0.0013
4	0.75	0.00017	0.00038	0.00058
5	0.67	0.00018	0.00028	0.00047
6	0.67	0.00018	0.00020	0.00041

Per-locus stutter parameter estimates. The median value of each parameter estimated across all autosomal loci is shown. “u+d” gives the median total probability of stutter error.

Supplemental Table 3

Motif length	$\log_{10}\mu$	$\text{Log}_{10}\mu$ (SE≥0)	β	p
1	-5.0	-4.8	0.60	1.00
2	-6.3	-4.7	0.40	0.93
3	-7.6	-5.9	0.44	0.94
4	-7.6	-5.9	0.45	0.95
5	-7.6	-6.2	0.44	0.94
6	-7.6	-6.7	0.42	0.93

Per-locus mutation median parameter estimates. The third column gives median mutation rates for STRs that had defined or non-zero standard errors, meaning they were not on the optimization boundary of our method.

Supplemental Table 4

Feature	Pearson r	P-value
Uninterrupted length	0.38	<10e-200
Length	0.15	<10e-200
GC content	-0.044	<10e-200
Entropy	0.030	7.4e-160
Period	-0.27	<10e-200
Replication timing	-0.030	2.3e-162
Purity score	0.22	<10e-200

Impact of local sequence features on STR mutation rates. P-values of <10e-200 denotes p-values under Python's numerical threshold using the scipy.stats.pearsonr function.

Supplemental Table 5

Chrom	Start	Motif	Gene	Zscore	OMIM annotation	Category
18	19752073	ACC	GATA6	-12.88	Atrioventricular septal defect 5; Atrial septal defect 9; Pancreatic agenesis and congenital heart defects; Persistent truncus arteriosus; Tetralogy of Fallot	Most constrained
2	5833526	ACG	SOX11	-12.60	Mental retardation, autosomal dominant	Most constrained
14	99641544	AGG	BCL11B	-12.31	Immunodeficiency	Most constrained
17	40345560	AGC	GHDC	-12.02	NA	Most constrained
17	71205859	AGC	FAM104A	-12.02	NA	Most constrained
8	28209226	AGC	ZNF395	-12.02	NA	Most constrained
3	129546680	AGC	TMCC1	-11.88	NA	Most constrained
19	51961617	AGC	SIGLEC8	-11.88	NA	Most constrained
2	237076427	CCG	GBX2	-11.88	NA	Most constrained
2	105472776	ACC	POU3F3	-11.88	NA	Most constrained
9	35561913	ACCC	FAM166B	2.78		Least constrained
10	21805467	AGG	SKIDA1	2.97		Least constrained
5	112824025	AGC	MCC	4.09	Colorectal cancer	Least constrained
1	17026022	CCG	ESPNP	4.21		Least constrained
13	25671799	AGC	PABPC3	4.99		Least constrained
3	18391133	AGC	SATB1	5.18		Least constrained
19	39019599	AGG	RYR1	8.35	Central core disease; Minicore myopathy with external ophthalmoplegia; Neuromuscular disease, congenital, with uniform type 1 fiber; King-Denborough syndrome	Least constrained
10	29821971	AAG	SVIL	9.36		Least constrained
7	73462847	AGC	ELN	10.01	Supravalvar aortic stenosis, cutis laxa	Least constrained
11	73020369	AGG	ARHGEF17	14.04		Least constrained

Most and least constrained protein coding STRs.

References

1. Sainudiin, R., Durrett, R. T., Aquadro, C. F. & Nielsen, R. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* **168,** 383–395 (2004).
2. Sun, J. X. *et al.* A direct characterization of human mutation based on microsatellites. *Nat. Genet.* **44,** 1161–1165 (2012).
3. Huang, Q.-Y. *et al.* Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* **70,** 625–634 (2002).
4. Garza, J. C., Slatkin, M. & Freimer, N. B. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* **12,** 594–603 (1995).
5. Miao, D. W.-C. Analysis of the discrete Ornstein-Uhlenbeck process caused by the tick size effect. *J. Appl. Probab.* **50,** 1102–1116 (2013).
6. Karlin, S. & Taylor, H. M. *An Introduction to Stochastic Modeling, Third Edition*. (Academic Press, 1998).
7. Haasl, R. J. & Payseur, B. A. Microsatellites as targets of natural selection. *Mol. Biol. Evol.* **30,** 285–298 (2013).
8. Amos, W., Kosanović, D. & Eriksson, A. Inter-allelic interactions play a major role in microsatellite evolution. *Proc. R. Soc. B* **282,** 20152125 (2015).
9. Weber, J. L. & Wong, C. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2,** 1123–1128 (1993).
10. Ballantyne, K. N. *et al.* Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* **87,**

341–353 (2010).

11. Willems, T. *et al.* Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *Am. J. Hum. Genet.* **98,** 919–933 (2016).
12. The GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
13. La Spada, A. R. & Taylor, J. P. Repeat expansion disease: progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* **11,** 247–258 (2010).
14. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
15. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* **24,** 400–402 (2000).