# Supplementary Material for "A direct characterization of human mutation based on microsatellites"

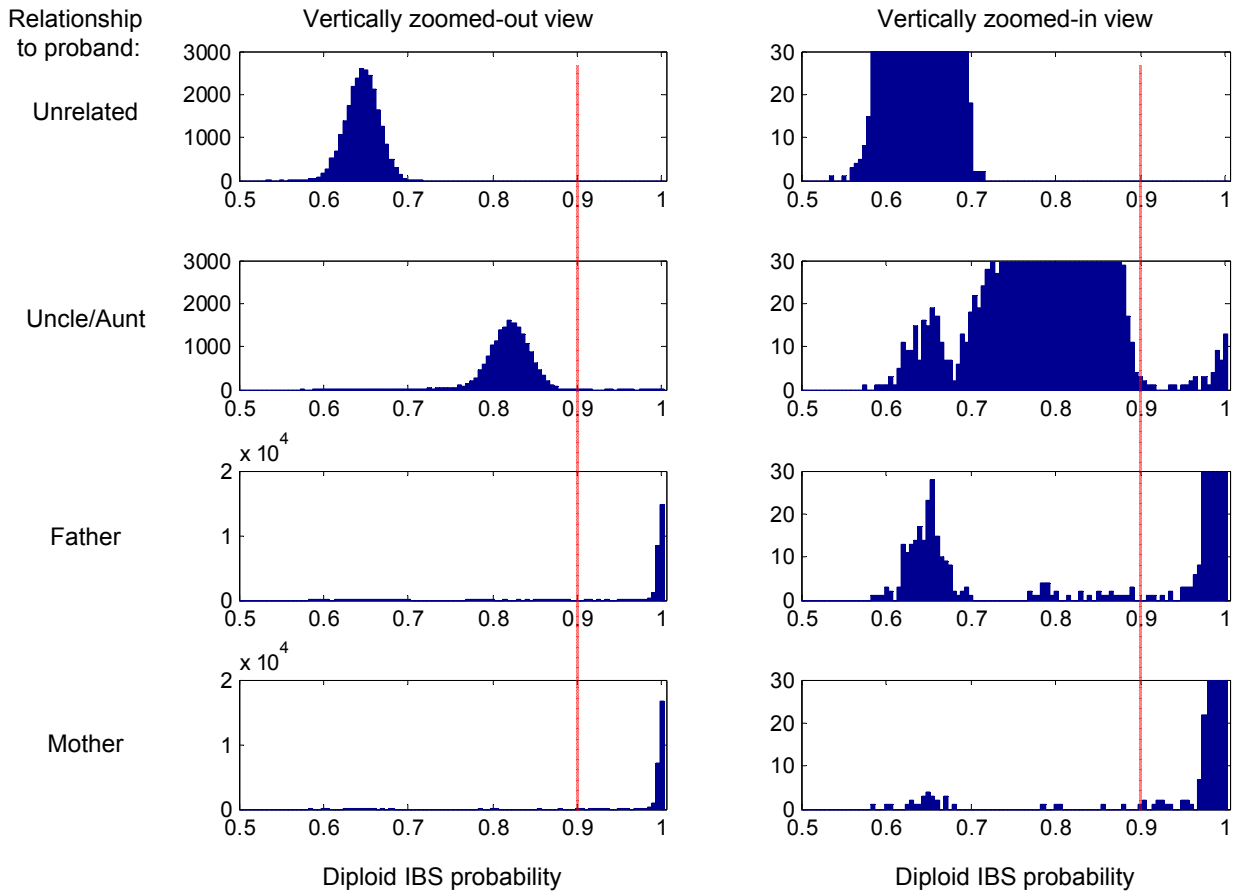
James X. Sun, Agnar Helgason, Gisli Masson, Sigríður Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich & Kari Stefansson





Nature Genetics: doi:10.1038/ng.2398

## Supplementary Figure 1. Removal of trios due to potential false-parenthood



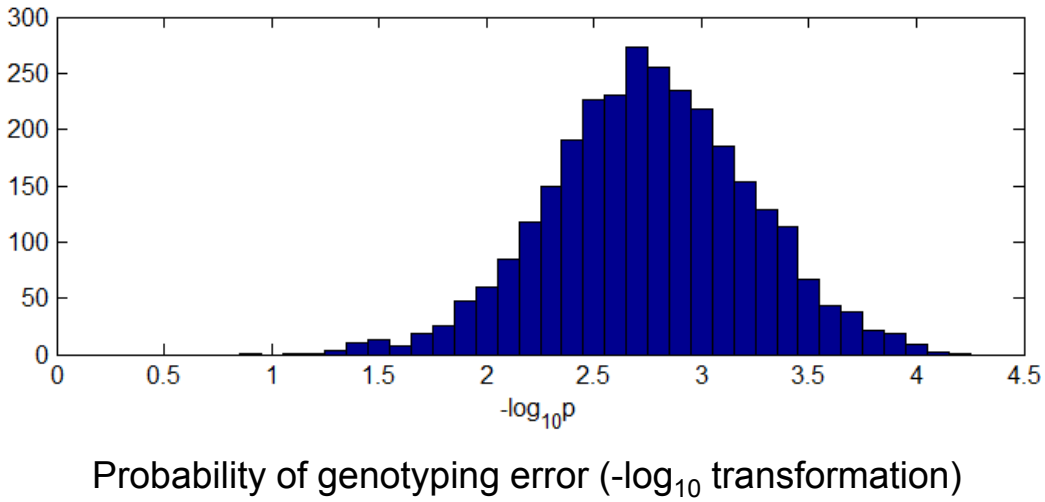

**Supplementary Figure 1. Removal of trios due to potential false-parenthood.** Trios were removed based on identity-by-state (IBS) probabilities between a parent and the proband, using all available microsatellite loci. In the figure, the first row is the empirically sampled IBS between pairs of unrelated individuals. The second row shows IBS between the proband and his/her uncle or aunt, allowing us to set a threshold that removes such trios as well. The 3rd and 4th rows are the IBS from the trios, assembled using the Icelandic genealogy. Based on the "null hypothesis" from the first two rows, the threshold for removal of trios was set at 0.9 (red line). A trio is removed if either the Father or the Mother falls below the threshold. Out of 25,067 trios, 235 were removed with this filter.

Definition of diploid IBS: Given individuals $A$ and $B$, assume that $n$ loci have been genotyped in both. At locus $i$, let the diploid genotype of $A$ be $A_i$, and that of $B$ be $B_i$. We call $A_i = B_i$ if any of the alleles match. For example, if $A_i$ = (4,6) and $B_i$ = (4,8), they are considered equal. Let $\mathbb{I}(A_i = B_i)$ be the indicator variable that is 1 if they are equal and 0 otherwise. Then, the IBS probability is defined as $pIBS(A,B) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(A_i = B_i)$.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 2. Estimated genotype error rate per locus**




**Supplementary Figure 2. Estimated genotype error rate per locus.** Distribution of genotype errors across loci is shown. The genotype error rate is defined as the probability that a single allele will be erroneous after genotyping. The horizontal axis shows the $-\log_{10}$ of the error rate. The median genotype error rate is $1.8 \times 10^{-3}$, with 95% of the density from $1.7 \times 10^{-4}$ to $1.4 \times 10^{-2}$.

Definition of genotype error rate at a given locus: Let $\hat{p}$ be the estimated probability of a genotype error when a single allele is observed, let $k$ be the number of times an allele is repeatedly genotyped, let $n_k$ be the total number of individuals who were each genotyped $k$-times, and let $y_k$ be the number of individuals with inconsistent genotypes. For example, if an individual is genotyped 10 times, 9 times yielding the genotype (4,6) and once yielding (5,6), this would be regarded as an inconsistent genotype. Then, the estimated probability of error is

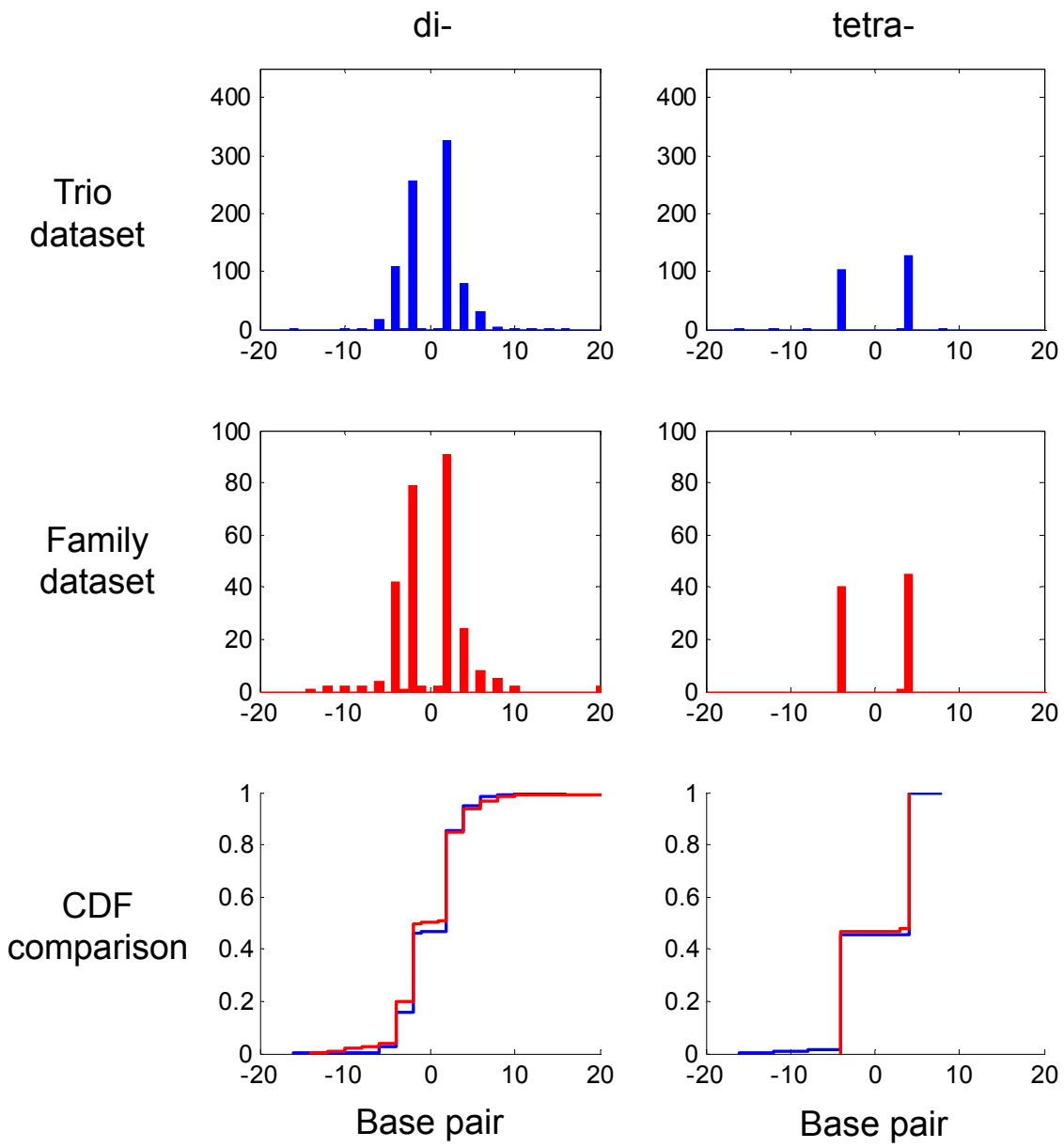
$$\hat{p} = \frac{\sum_k y_k}{\sum_k 2kn_k}$$

Supplementary Notes describe the derivation of this expression and its assumptions.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 3. Similarity between trio and family data in mutational length distribution**

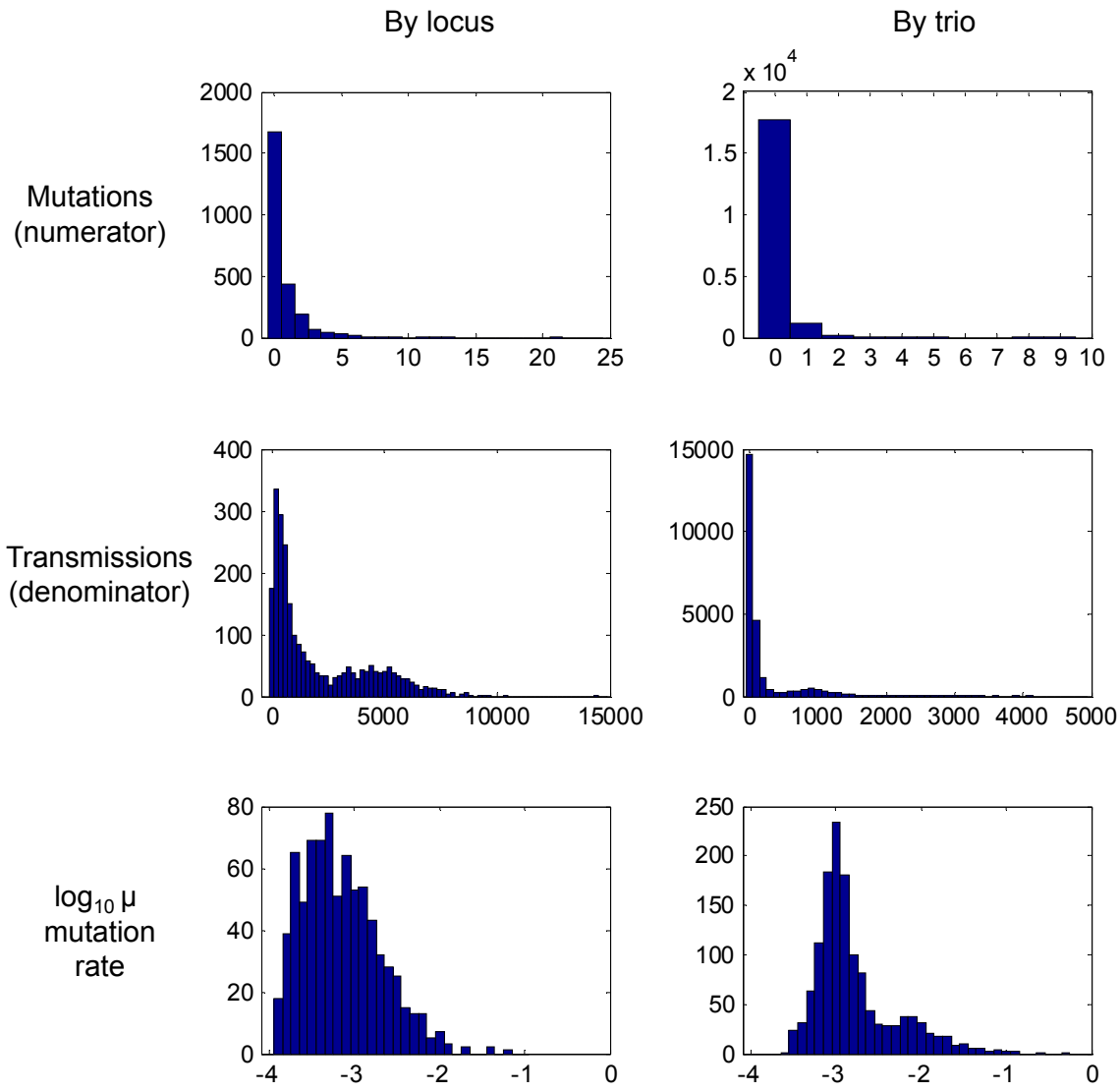



**Supplementary Figure 3. Similarity between trio and family data in mutational length distribution.** This figure separates the trio and family datasets from main text Fig 2B. Additionally, the bottom row compares the CDF between the datasets. The two-sample Kolmogorov-Smirnov test gives P-values of 0.807 and 1 for the di- and tetra- comparisons, respectively. Thus, in the mutational length distribution, there are no significant differences between the two datasets.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 4. Mutations by locus and by trio**

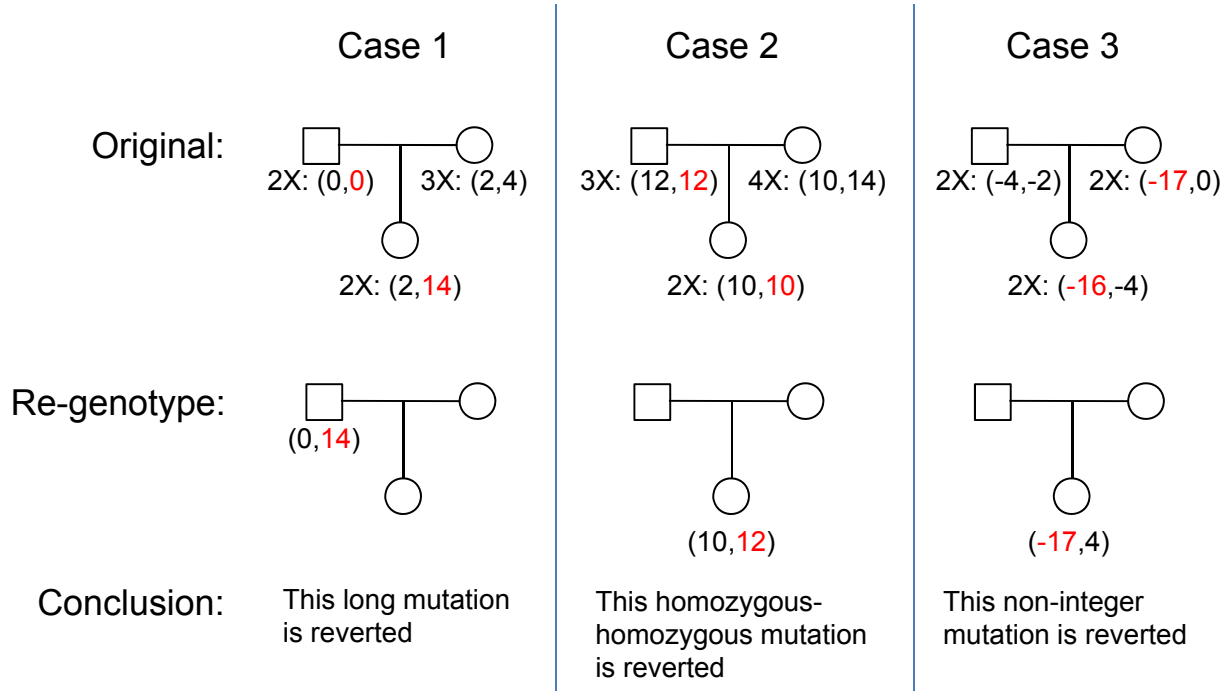



**Supplementary Figure 4. Mutations by locus and by trio.** The rows show histograms of mutations, transmissions, and the mutation rate per locus. Of the 2,477 loci, most loci do not contain any mutations. For the loci with at least 1 mutation, the histogram of $\log_{10}$ of the mutation rate resembles a truncated normal distribution, since our denominator is limited to at most about 10,000 per locus. The right column shows the corresponding plots by trio. Of the 24,832 trios, most do not contain a mutation. Due to the sparseness of mutations by locus and by trio, we combine locus and trio data as appropriate to perform our analyses.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 5. False-positive mutations from the trio approach**




**Supplementary Figure 5. False-positive mutations from re-genotyping in the trio approach.** From the set of trio mutations identified, we randomly chose 103 mutations and re-genotyped them. 3 false-positives were identified, which are shown here. All genotypes are in units of base pairs. The 1st case is an apparent mutation that is unusually long, with a mutational length of 14 bp. The 2nd case involves a homozygous parent transmitting to a homozygous child, which we believe is a more error-prone class as discussed in the text. The 3rd case is an apparent mutation of a single base pair, which is a non-integer multiple of the motif length (2 base pairs in this case).

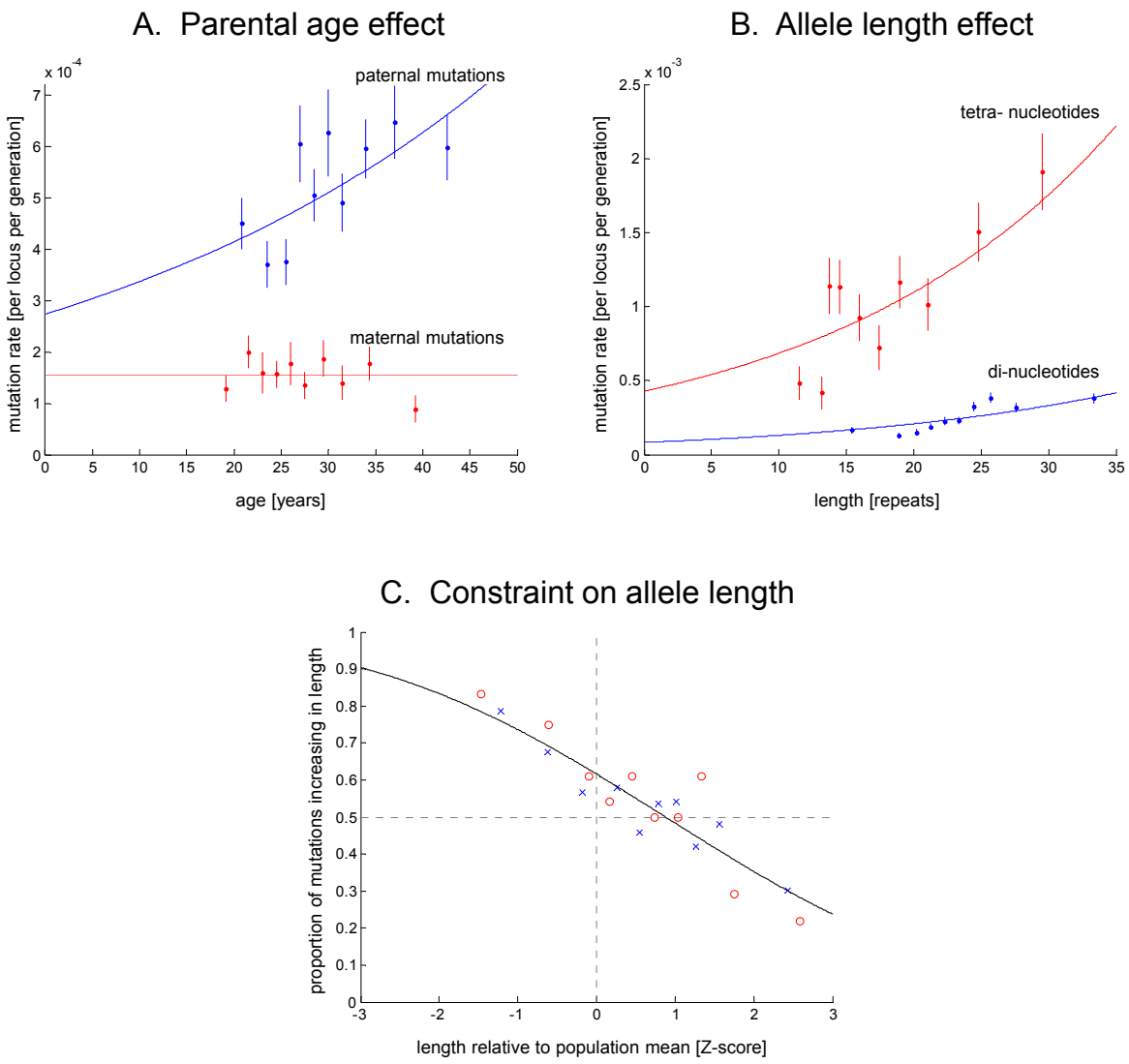
See Supplementary Notes and Supplementary Table 1 for a more elaborate analysis of false-positive rates when a mutation is either (1) excessively long, (2) a transmission from a homozygous parent to a homozygous child, or (3) a non-integer multiple of the motif length.

Note that allele lengths illustrated above are relative lengths, which is an offset (in units of base pairs) based upon the absolute length of a reference individual's allele.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 6. Predictors of mutation rate and direction (logistic regression)**

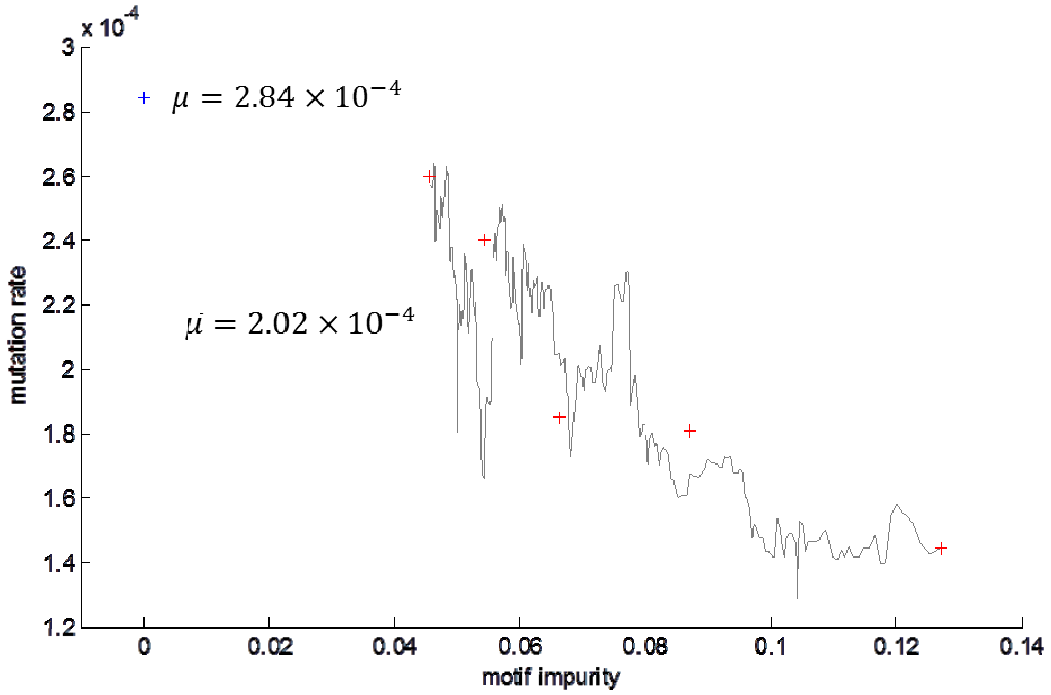



**Supplementary Figure 6. Predictors of mutation rate and direction (logistic regression).** Same as main text Fig 2, but with logistic regression curve fits. Note that while the data points shown here are from binning the data, as described in Fig 2, the logistic regressions are performed over the raw data, in which a binomial model of generating mutations (response variable) is assumed. Logistic regression over the raw data has more statistical power than linear regression over the binned data and is constrained to have non-negative mutation rates. The P-values in the main text are reported based on the logistic regression analysis.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 7. Imperfect repeats have a lower mutation rate**

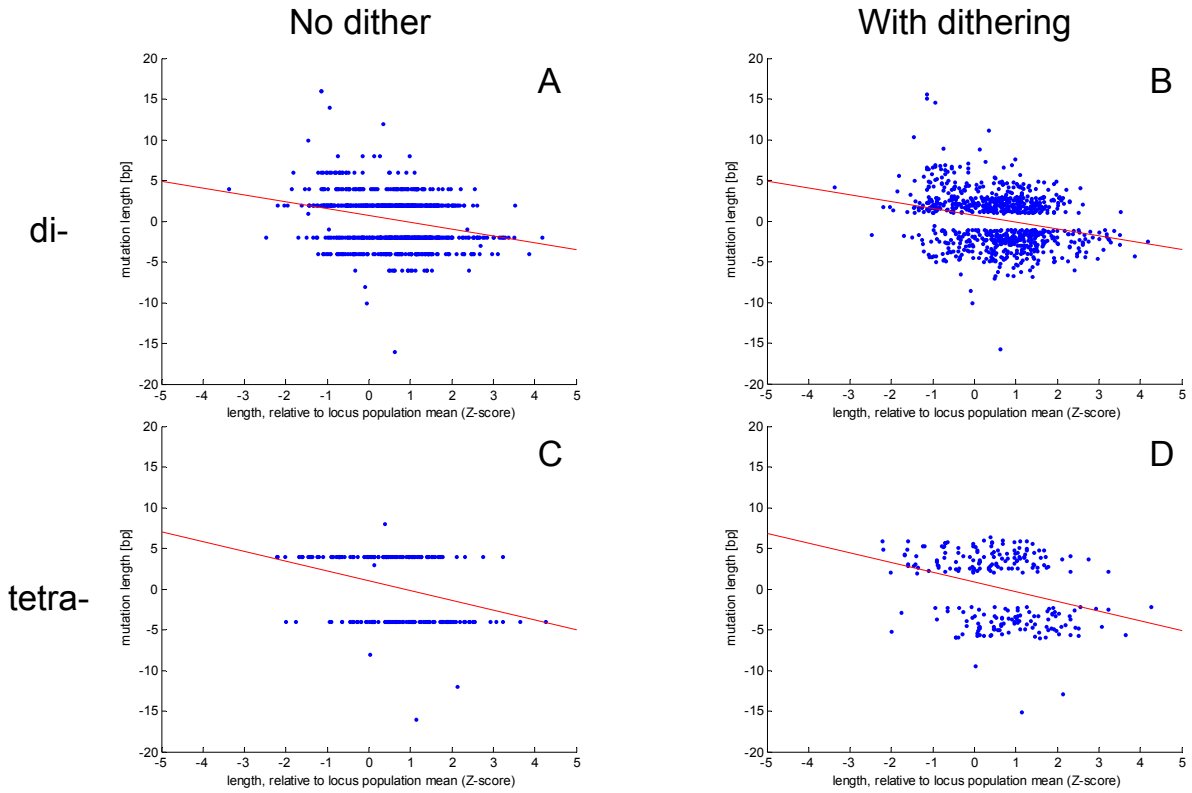



**Supplementary Figure 7. Imperfect repeats have a lower mutation rate.** The purity of a motif is computed using the human reference sequence *hg19* from the UCSC genome browser, and downloading data for "simple repeats", in which the "perMatch" column gives the percentage match of the human-genome reference microsatellite to the pure repeat. We define "motif impurity" as one minus this statistic. In blue is the aggregate of 1,036 di-nucleotide loci in which the repeats are perfect (e.g. CACACACACA), without any interrupting bases in the pattern. In red are the imperfect repeats (e.g. CACACA**T**CACA), binned according to the level of repeat impurity. In gray is the window-averaged mutation rate of the imperfect repeats. There are a total of 396 di-nucleotide loci with imperfect repeats. Logistic regression shows that the level of repeat impurity regresses significantly ($P = 3.1 \times 10^{-7}$) with mutation rate. The evidence here is compatible with the hypothesis that when a tandem repeat is interrupted, DNA polymerase slippage is less likely to occur, and hence the mutation rate becomes lower.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Figure 8. Length constraints in microsatellites (raw)



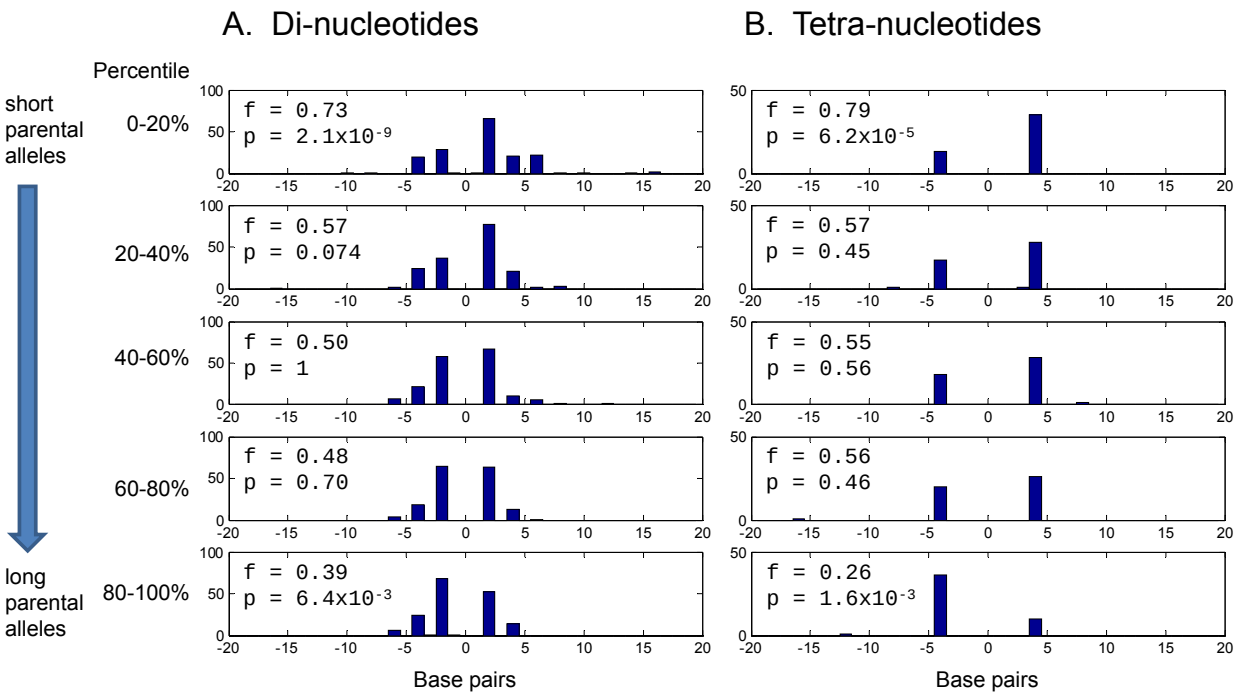

**Supplementary Figure 8. Length constraints in microsatellites (raw).** Relative length (x-axis) is in units of Z-scores, the number of standard deviations from the mean length at a given locus. The left panels plot relative length against the mutation length, in base pairs. The right panels provide dithering using a uniform distribution from -0.5 to 0.5 bp to reduce quantization on each mutation length. There is a significant negative correlation.

For di-nucleotides, panel A has: $r^2$=0.0739, slope=-0.838, P=1.48x$10^{-15}$.
For tetra-nucleotides, panel C has: $r^2$=0.106, slope= -1.202, P=3.33x$10^{-7}$.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Figure 9. Length constraints in microsatellites (binned)

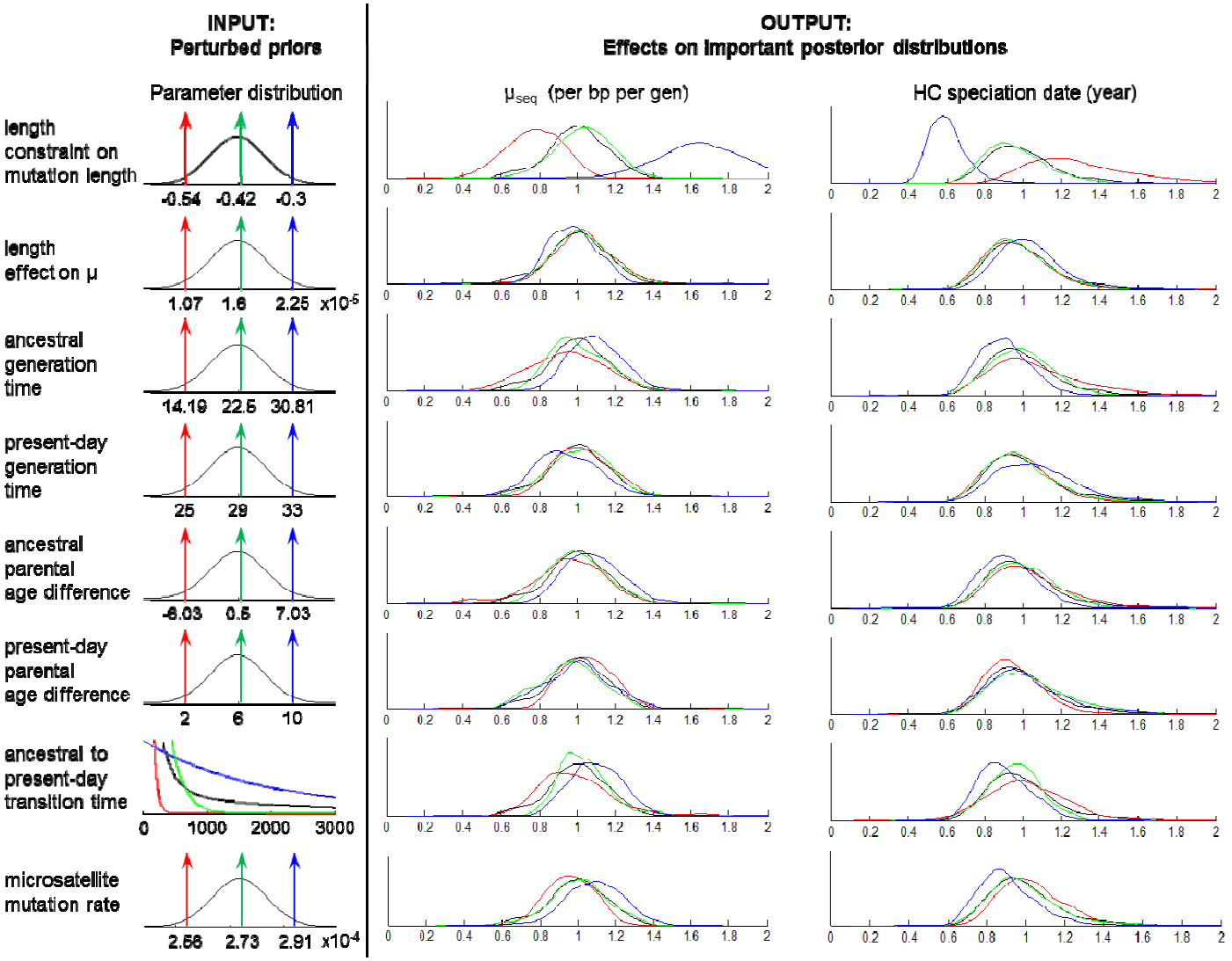



**Supplementary Figure 9. Length constraints in microsatellites (binned version).** This figure shows the mutation length distributions as a function of the length of the parental allele, relative to the mean length of a locus. When the parental allele is short (percentiles are displayed on the left), mutation length is biased towards the positive direction. When the parental allele is long, the mutation length is biased towards the negative direction. The fraction (f) of length expansions and the P-value (p) using a two-sided binomial test (the null hypothesis is that microsatellites have no directional bias), are shown in each histogram.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Figure 10. Sensitivity analysis of evolution model

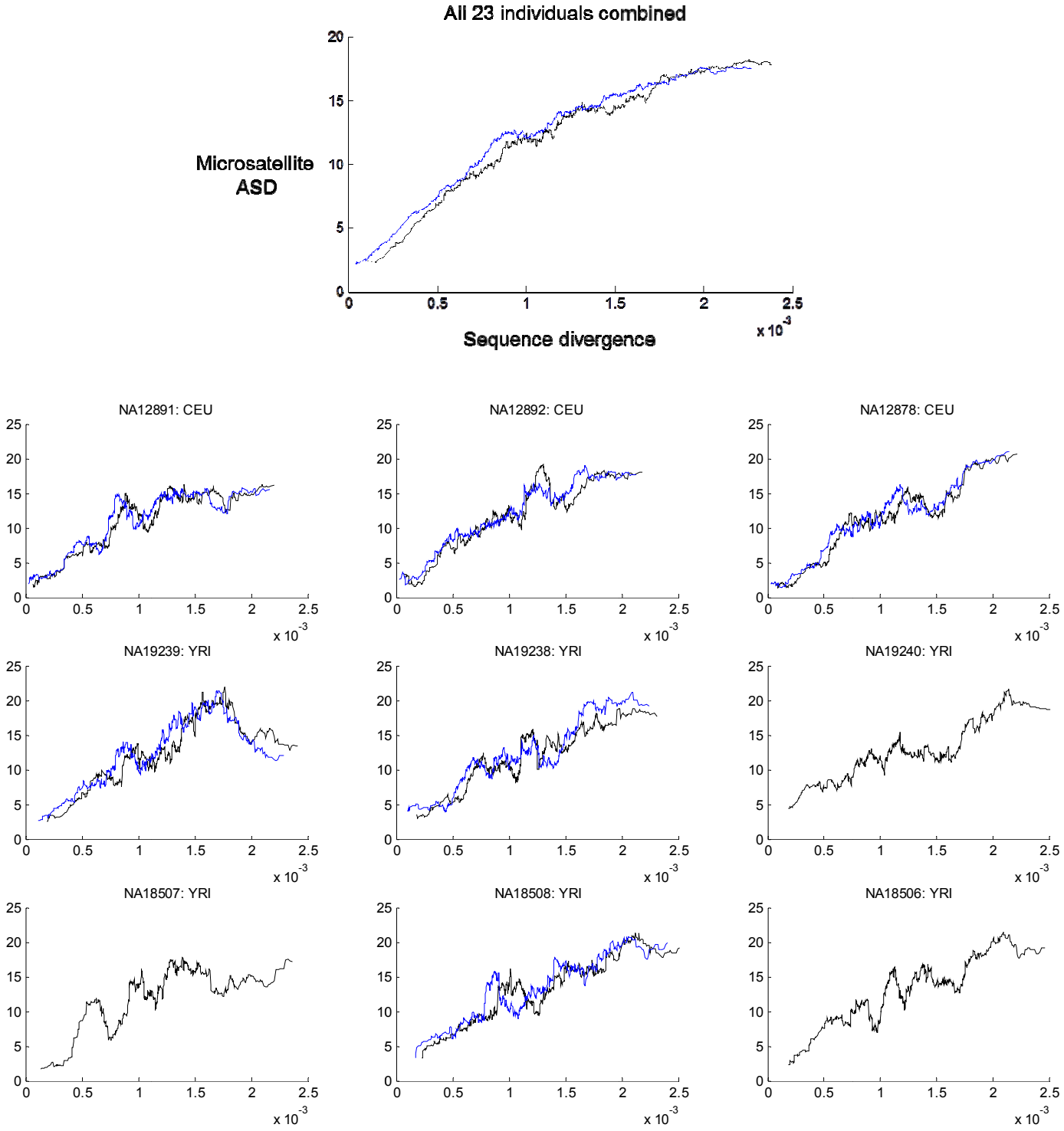



**Supplementary Figure 10. Sensitivity analysis of the evolution model.** Our model of evolution is robust to changes in the prior distributions. Eight parameters that we use as priors are in the left column, with the default distributions in black. We tested robustness by setting each prior to have different point values (the mean, 5th percentile, and 95th percentile of the default distribution in black), and exploring how this changes the posterior distributions (the coloring of the posteriors correspond to the respective priors, all scaled by the mean of the black posterior). In the case of the "ancestral to present-day transition" in the generation time ( in Supplementary Notes), the parameter distribution is a mixture of 3 exponentials (see Methods), and we test robustness by sampling from each separately. Our posterior estimates are not much affected by the input parameters as long as they fall within the range of the priors. The exception is the length constraint (top row) that governs the non-linear mapping between TMRCA and ASD (Fig 3), where we observe substantial differences. Note, however, that we obtain essentially the same posterior distribution when we use a point estimate corresponding to the mean of the prior distribution and the full prior distribution, which demonstrates the robustness of our inference procedure. Our evolutionary modeling updates its inference of the length constraint directly from comparing the microsatellite ASD to flanking sequence diversity; it is not solely based on our direct measurements. Thus, as long as we include the true value within the prior, we get robust results even for the length constraint parameter (Supplementary Notes).



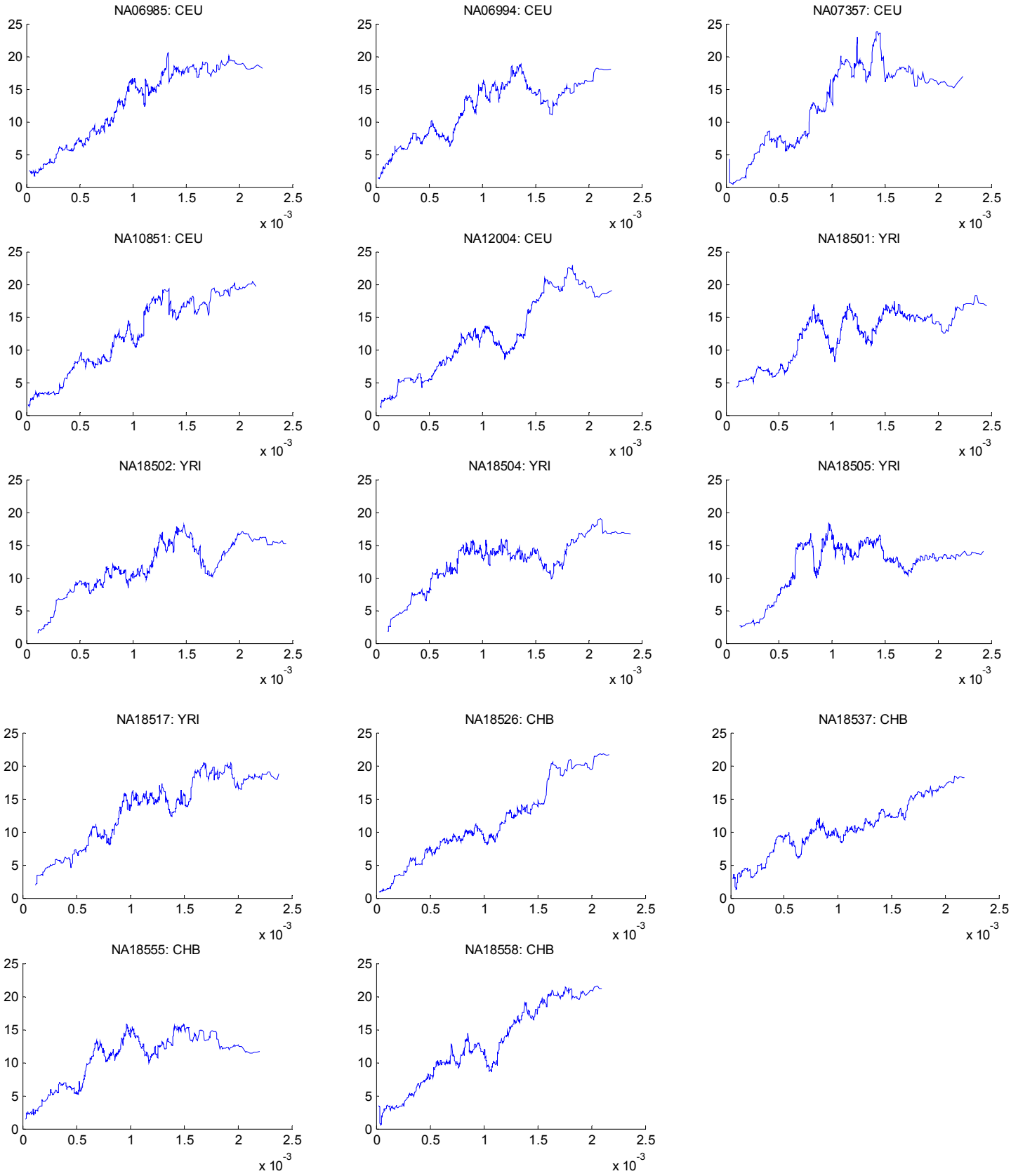
Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 11. Sequence divergence versus microsatellite ASD for 23 HapMap individuals**






Nature Genetics: doi:10.1038/ng.2398





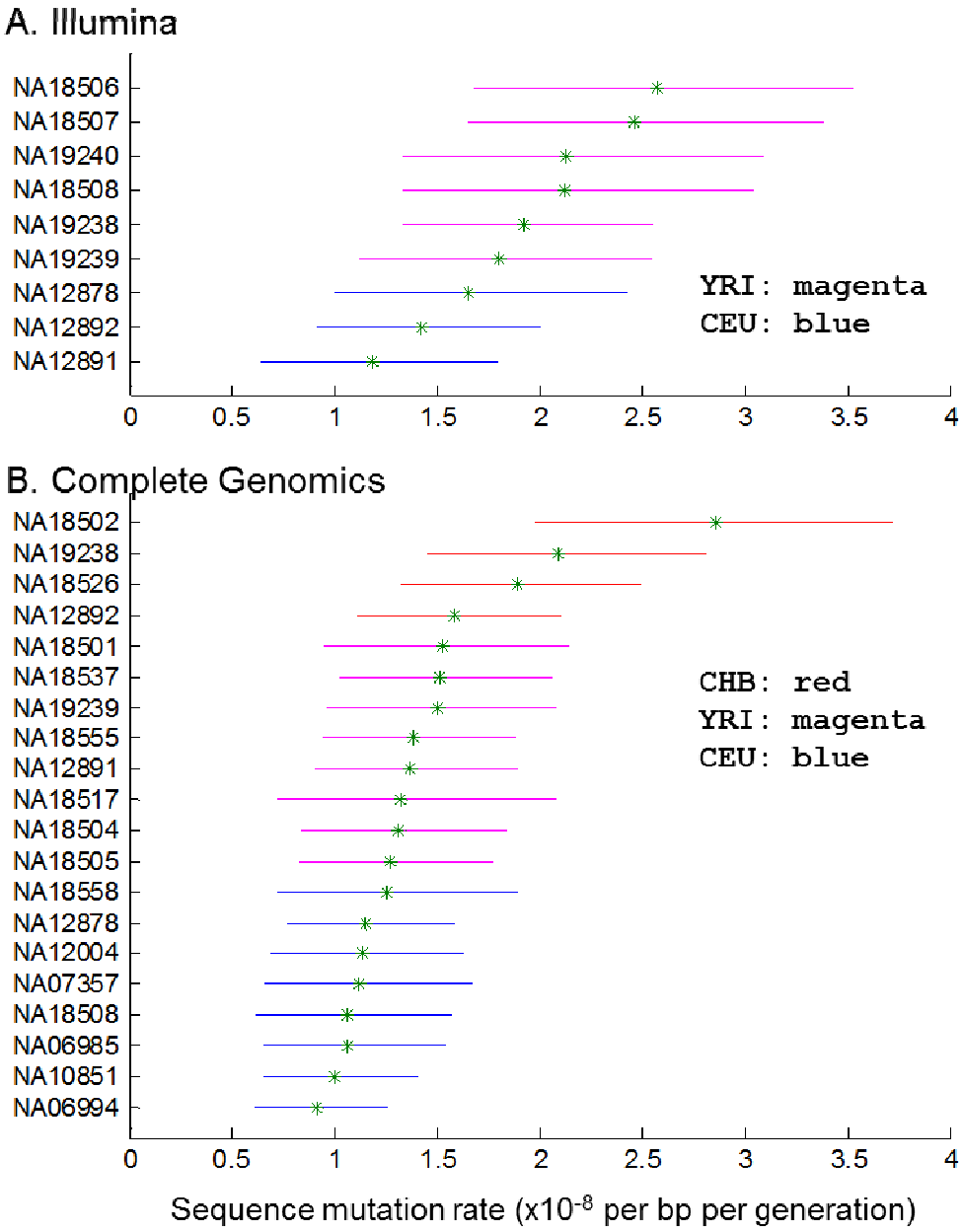

Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 11. Sequence divergence versus microsatellite ASD.** These plots are similar to that of Fig 3 but with the x-axis un-rescaled to TMRCA. The combined plot and separate plots for the 23 HapMap individuals are shown. We empirically validate the non-linear behavior predicted by our model by exploiting the fact that there exists considerable variability in sequence heterozygosity (hence TMRCA) across the genome. The x-axis shows the pairwise sequence heterozygosities from sequence data. The y-axis shows the ASD statistic from microsatellite data. In blue are sequence data from Complete Genomics (20 individuals), and in black are data generated using Illumina technology (9 individuals). Microsatellite ASD at each di-nucleotide locus and heterozygosity were computed for each individual and then combined and smoothed using a sliding-window average. We computed the local sequence heterozygosity based on the sequence flanking each microsatellite over a genetic distance window of 0.001 centimorgans in either direction and excluding a 1kb region where the microsatellite itself lies. The result shows a non-linear relationship between microsatellite ASD and sequence heterozygosity which is assumed to increase linearly with time, empirically demonstrating that our model of microsatellite evolution is more appropriate than the GSMM model.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 12. Inferred sequence mutation rate of 23 individuals**




**Supplementary Figure 12. Inferred sequence mutation rate of 23 individuals.** This is a graphical representation of Supplementary Table 8. The asterisk is the mean mutation rate, and the bars are the 90% Bayesian credible intervals. Populations are coded by color. Note that while the individual mutation rates are not significantly different from each other, the populations do exhibit some clustering, where CEU Europeans have a lower mutation rate than either YRI



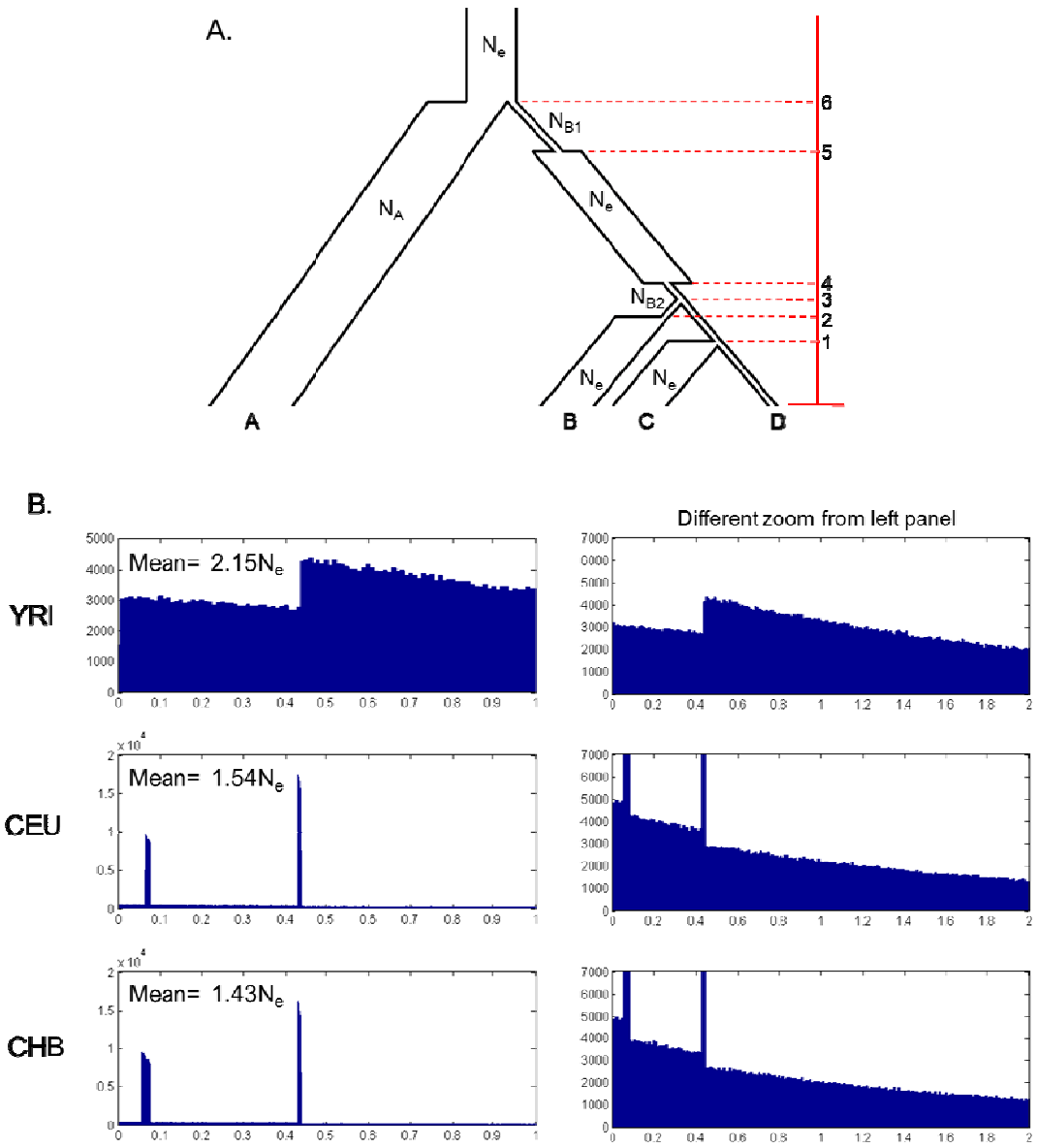
Nature Genetics: doi:10.1038/ng.2398

Africans or CHB Han Chinese. We see two possible explanations for non-random clustering within populations. (1) One possibility is random fluctuation: the differences are not statistically significant, and the clustering within populations could thus simply reflect correlated histories within populations. (2) A second possibility is ascertainment bias for microsatellites with high heterozygosity in Europeans (to make them more useful for disease gene mapping). To understand how this bias could cause underestimation of the mutation rate especially in Europeans, we note that ascertaining for highly polymorphic microsatellites is expected to inflate the measured ASD compared with the expectation based on the true mutation rate, thus overestimating the TMRCA. This in turn results in an underestimate of the sequence heterozygosity, since if we infer that more time elapsed in the process of generating the observed mutations, we will estimate a lower mutation rate. Such an ascertainment bias would be expected to be strongest in people of European ancestry as we observe (since they are most closely related to Icelanders), while it would be more mild in more distant populations (CHB and YRI).



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 13. Demographic model for coalescent simulation**

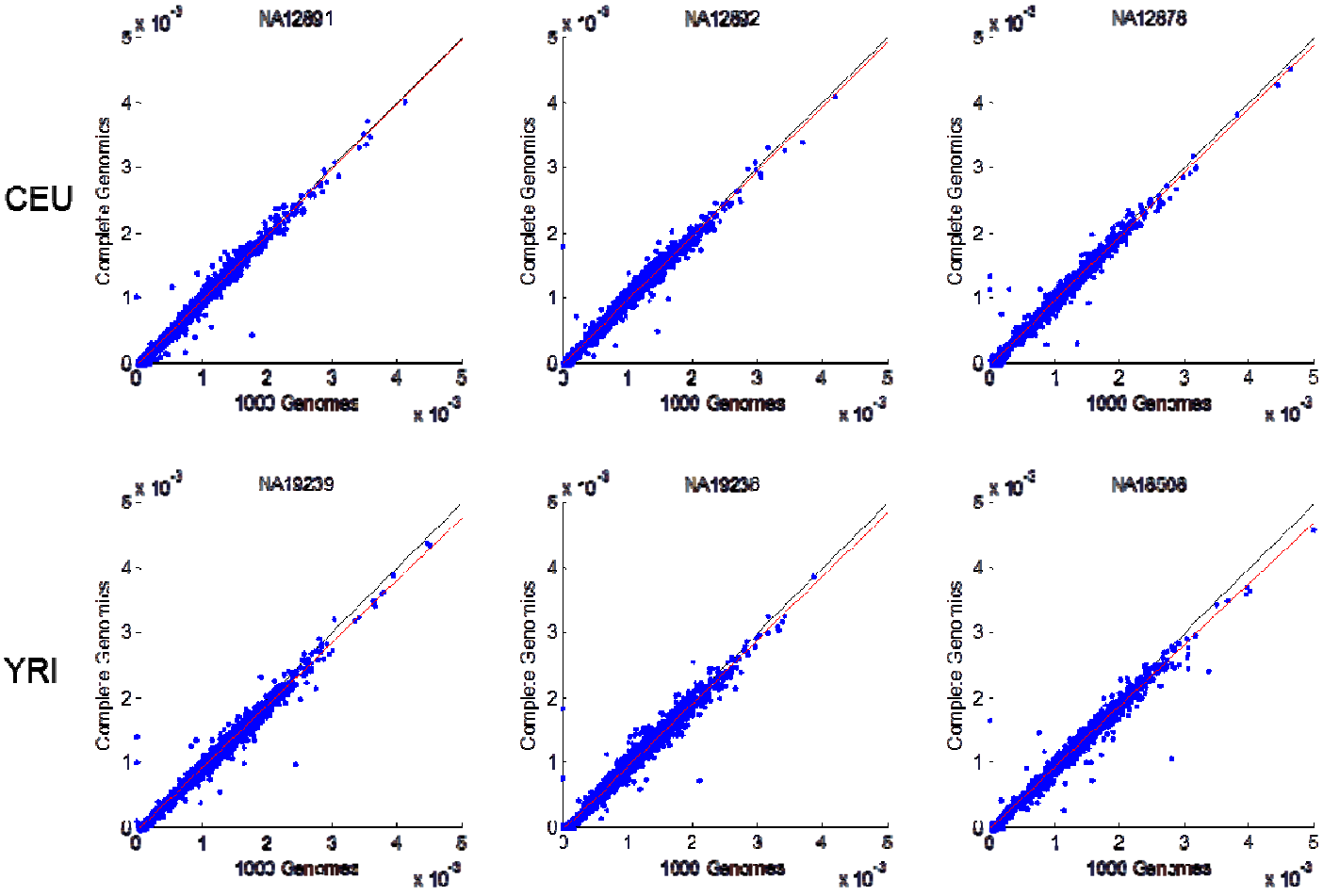



**Supplementary Figure 13. Demographic model for coalescent simulation.** In panel A is a model of 4 populations, with A=West Africans (YRI), B=Western Europeans (CEU), C=Han Chinese (CHB), D=Native Americans. This is a 2-bottleneck model suggested by Keinan et al.[1], with $N_e$=10,000, $N_A$=1.1$N_e$, $N_{B1}$=0.02$N_e$, $N_{B2}$=0.05$N_e$, $t_1$=0.0147*4$N_e$, $t_2$=0.016*4$N_e$, $t_3$=0.018*4$N_e$, $t_4$=0.019*4$N_e$, $t_5$=0.107*4$N_e$, $t_6$=0.109*4$N_e$. Panel B shows the distribution of within-population 2-sample coalescent times, scaled by $N_e$. There are more coalescent events within bottlenecks, as shown by the peaks in the distribution for CEU and CHB. We use this model to verify that our inference of mutation rates is robust to differences in demographic histories across populations.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 14. Heterozygosity: CGI versus Illumina**

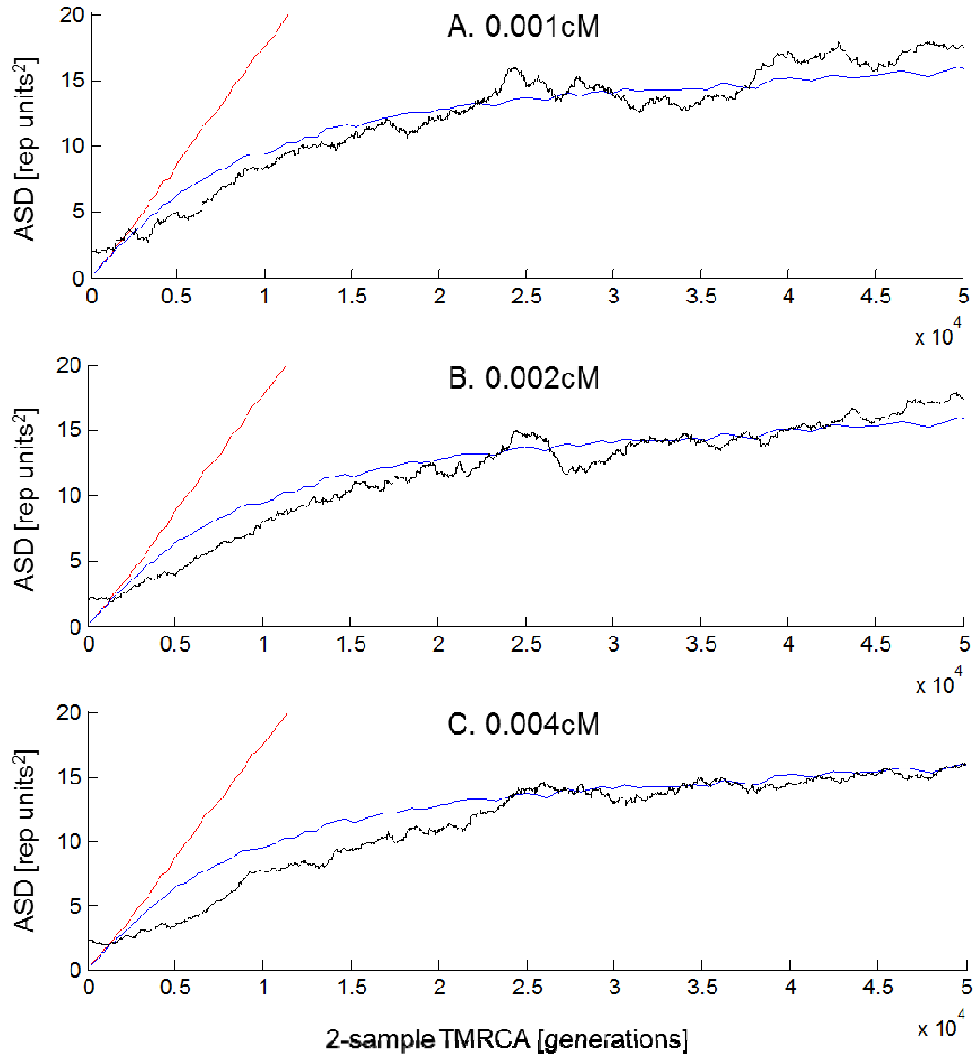



**Supplementary Figure 14. Heterozygosity: CGI versus Illumina.** Six individuals have sequence data from both CGI and Illumina. Here we compare heterozygosities. The Illumina heterozygosity is slightly higher than that of CGI.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Figure 15. Genetic windows for sequence heterozygosity




**Supplementary Figure 15. Varying genetic windows for sequence heterozygosity.** To extract sequence heterozygosity around each microsatellite, a suitable window length is required. If this window size is too short, sequence heterozygosity becomes imprecise. If the window is too large, crossing multiple recombination events, then the sequence heterozygosity approaches the genome-wide average, rather than local. We tried 3-different window sizes with thresholds at 0.001, 0.002, and 0.004 cM. Shown in black is the empirical curve of microsatellite ASD versus sequence-based 2-sample TMRCA, averaged across the 23 HapMap individuals. The TMRCA is estimated from sequence heterozygosity using a sequence mutation rate of $1.82 \times 10^{-8}$, which is the value we inferred (main manuscript Table 2). The red and blue curves are simulations: in red is the standard random walk (GSMM) model, and in blue is our evolution model. As shown in the figure, all 3 window sizes clearly show a saturation of the ASD curve, closely matching our model. The threshold with 0.001cM is noisier due to less sequence data, however, the fit seems slightly better. Thus, this is the threshold we use, and panel A is the one used for Figure 3 of the main manuscript.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 16. UCSC web query for obtaining microsatellite information**

A.


B.

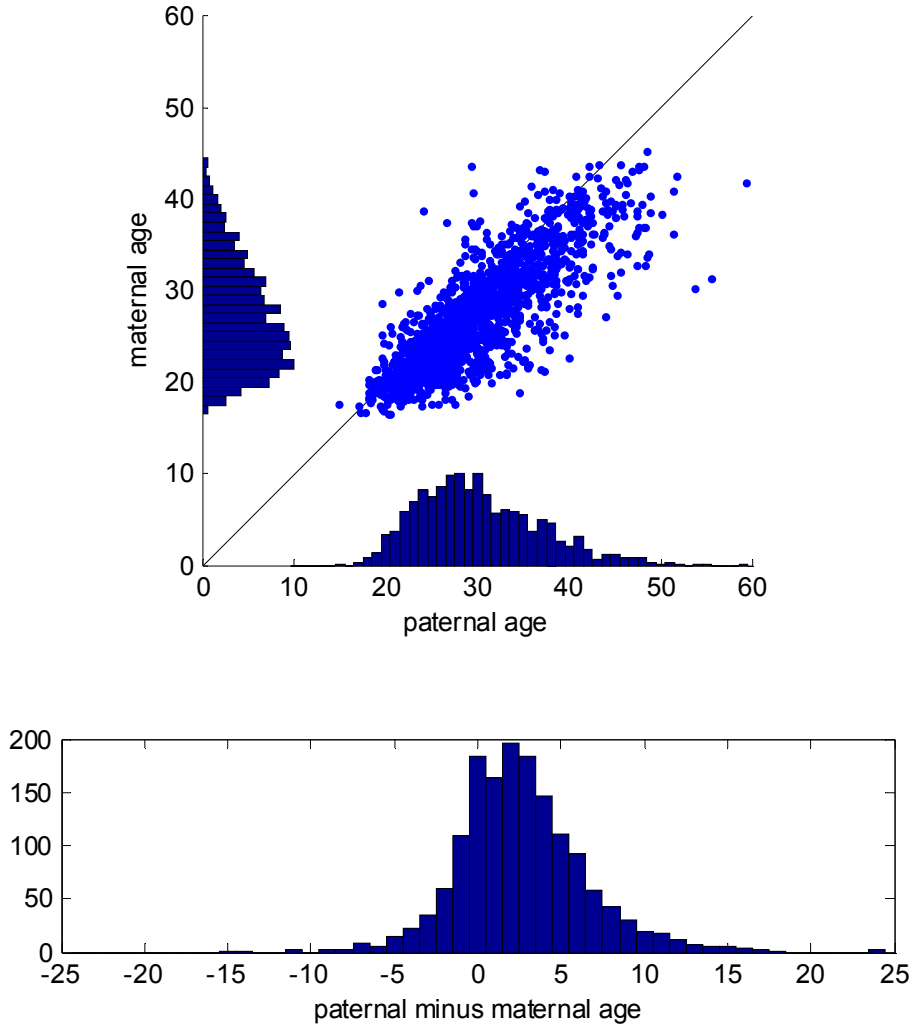
| #chrom | chromStart | chromEnd | name | period | copyNum | perMatch | sequence |
|---|---|---|---|---|---|---|---|
| chr1 | 0 | 468 | trf | 6 | 77.2 | 95 | TAACCC |
| chr1 | 20725 | 20822 | trf | 2 | 47.5 | 75 | TC |
| chr1 | 34698 | 34739 | trf | 4 | 10 | 94 | AAAT |
| chr1 | 40344 | 40376 | trf | 2 | 16 | 100 | GT |
| chr1 | 44575 | 44680 | trf | 4 | 25.8 | 87 | TTTC |
| chr1 | 56023 | 56493 | trf | 2 | 262 | 71 | TA |
| chr1 | 56067 | 56495 | trf | 5 | 87.4 | 73 | ATATA |
| chr1 | 61991 | 62026 | trf | 4 | 8.8 | 87 | ATAC |
| chr1 | 73654 | 73904 | trf | 4 | 64.8 | 86 | AAAG |
| chr1 | 73726 | 73844 | trf | 6 | 18.2 | 69 | AAAGAA |
| chr1 | 88862 | 88905 | trf | 4 | 10.8 | 100 | TTTA |
| chr1 | 88909 | 88979 | trf | 1 | 70 | 76 | T |

**Supplementary Figure 16. UCSC web query for obtaining microsatellite information.** To obtain information for repeat motif (column: "sequence"), repeat length (column: "copyNum"), motif purity (column: "perMatch"), we obtained the output of Tandem Repeat Finder from the UCSC genome browser, with settings shown in panel A, and an excerpt of the output in panel B.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Figure 17. Distribution of parental age at child-birth**




**Supplementary Figure 17. Distribution of parental age at child birth.** These are the parental age of trios used in our mutation rate analyses. The paternal age has a mean and standard deviation of 30.1 and 6.5 years, while the maternal age has a mean and standard deviation of 27.4 and 5.9 years. Combining parents, the generation-time has a mean and standard deviation of 28.8 and 6.4 years.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 1. Experimental validation of mutations

| Mutations from family data set | Mutation Counts | Targeted re-genotyping | | | Electropherogram review | | | Intersection of sites | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TP | FP | FP/(TP+FP) | TP | FP | FP/(TP+FP) | TP | FP | FP/(TP+FP) |
| **Class 1 mutations** | 326 | 74 | 2 | 0.026 | 262 | 8 | 0.030 | 57 | 2 | 0.034 |
| **Class 2 mutations** | | | | | | | | | | |
| Homozygous parent and offspring | 21 | 10 | 2 | 0.167 | 20 | 0 | 0.000 | 9 | 2 | 0.182 |
| Non-integer multiple of motif length | 10 | 0 | 2 | 1.000 | 6 | 3 | 0.333 | 0 | 2 | 1.000 |
| Excessively long (>6bp) | 18 | 7 | 2 | 0.222 | 13 | 3 | 0.188 | 6 | 3 | 0.333 |
| More than 1 of the above | 1 | 0 | 0 | N/A | 1 | 0 | 0.000 | 0 | 0 | N/A |
| **Total** | 376 | | | 0.058 | | | 0.043 | | | 0.072 |

Experimental validation of mutations from the family data are shown here. See Supplementary Figure 5 for validation of the trio data.

TP = True Positives, i.e. candidate mutations that are verified to be true.
FP = False Positives, i.e. candidate mutations that are rejected by the verification.

Class 1 mutations are the ones that do not belong to Class 2, which are likely to have a higher false identification rate. Class 2 mutations include: (1) both parent and offspring were homozygous, (2) the mutation length was a non-integer multiple of the motif size, or (3) the mutation length was longer than 6 nucleotides.

In our re-genotyping efforts, to maximize our discovery of false-positives, we targeted our re-genotyping efforts toward Class 2. No such sampling bias was used in the electropherogram review. In combining the results of re-genotyping and electropherogram review, we examined only overlap data, calling a candidate mutation as a false-positive if either method rejects the mutation.

In obtaining the total false identification rate, due to sampling bias towards the Class 2 mutations, we calculated an overall rate that weights the number of Class 1 and Class 2 candidate mutations, i.e. to obtain the final value of 0.072, we have:

$$\frac{50}{376}\cdot\frac{7}{22}+\frac{326}{376}\cdot\frac{2}{59}=0.072$$



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 2. Validation of 14 microsatellite mutations with next generation sequence data

| Mutation Information | | | | | PCR genotype | | | NGS genotype (total length of motifs) | | | NGS genotype (actual alleles observed) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Locus | Repeat motif | Type | Parent (F/M) | Allele change | Father | Mother | Proband | Father | Mother | Proband | Father | Mother | Proband | Confirmed? |
| D11S4191 | $AC_n$ | Trio | F | 16→14 | 0/16 | 0/0 | 16/0 | 36/52 | 36/36 | 36/50 | 17xAC:3, 18xAC:8, 25xAC:1, 26xAC:3 | 14xAC:1, 16xAC:1, 17xAC:3, 18xAC:20, 19xAC:1 | 17xAC:3, 18xAC:9, 24xAC:3, 25xAC:7, 26xAC:1 | ✓ |
| D12S1297 | $TCTA_n$ | Trio | ? | 0→4 | -4/0 | 0/0 | 0/4 | 36/40 | 40/40 | 40/44 | 9xTCTA:16, 10xTCTA:15 | 9xTCTA:1, 10xTCTA:40 | 10xTCTA:9, 11xTCTA:16 | ✓ |
| D12S372 | $CTAT_n$ $CTAC_m$ | Trio | F | 4→8 | -4/4 | 0/0 | 0/8 | 48/56 | 52/52 | 52/60 | 9xCTAT+3xCTAC:4, 12xCTAT+2xCTAC:7 | 9xCTAT+3xCTAC:1, 11xCTAT+2xCTAC:11, 10xCTAT+3xCTAC:4 | 10xCTAT+2xCTAC:1, 11xCTAT+2xCTAC:8, 13xCTAT+2xCTAC:14 | ✓ |
| D17S794 | $GT_n$ $GT_m$ | Trio | F | 0→2 | 0/0 | 0/6 | 0/2 | 40/40 | 40/46 | 40/42 | 12xGT+6xGT:1, 13xGT+6xGT:1, 14xGT+6xGT:16, 17xGT+7xGT:1 | 14xGT+6xGT:7, 15xGT+6xGT:1, 15xGT+7xGT:2, 16xGT+7xGT:7 | 13xGT+6xGT:3, 14xGT+6xGT:17, 15xGT+6xGT:9 | ✓ |
| D20S852 | $GT_n$ | Trio | F | 0→-2 | 0/8 | -10/4 | -2/4 | 30/38 | ? | 28/34 | 14xGT:2, 15xGT:8, 16xGT:1, 19xGT:5 | No data | 12xGT:2, 13xGT:4, 14xGT:8, 15xGT:1, 16xGT:1, 17xGT:7 | ✓ |
| D20S902 | $CA_n$ $CA_m$ | Trio | M | 2→-2 | -2/-2 | 2/4 | -2/-2 | 50/54 | 54/56 | 50/54 | 11xCA+14xCA:9, 11xCA+16xCA:6, 11xCA+17xCA:1 | 11xCA+16xCA:2, 10xCA+18xCA:6 | 11xCA+14xCA:4, 11xCA+16xCA:5 | NO |
| D21S1908 | $CA_n$ | Trio | F | 2→0 | 2/2 | 2/6 | 0/6 | 32/32 | 32/36 | 30/36 | 15xCA:2, 16xCA:31, 18xCA:1 | 16xCA:10, 17xCA:2, 18xCA:8 | 15xCA:9, 18xCA:11 | ✓ |
| D2S2254 | $GT_n$ | Trio | F | 20→18 | 0/20 | -2/16 | -2/18 | 32/54 | 30/48 | 30/50 | 13xGT:1, 16xGT:10, 27xGT:4 | 14xGT:2, 15xGT:12, 16xGT:1, 23xGT:1, 24xGT:6 | 14xGT:3, 15xGT:18, 16xGT:1, 24xGT:1, 25xGT:9 | ✓ |
| D3S3620 | $TG_n$ | Trio | F | 2→0 | -4/2 | -4/-4 | -4/0 | 36/42 | 36/36 | 36/40 | 18xTG:4, 20xTG:3, 21xTG:5 | 17xTG:3, 18xTG:12 | 18xTG:10, 19xTG:1, 20xTG:5 | ✓ |
| D5S1397 | $CTTTT_n$ $CTTT_m$ | Trio | M | 4→8 or 12→8 | 12/17 | 4/12 | 8/17 | 52/57 | 44/57 | 48/57 | 0xCTTTT+13xCTTT:5, 1xCTTTT+13xCTTT:6 | 0xCTTTT+9xCTTT:1, 0xCTTTT+11xCTTT:5, 0xCTTTT+13xCTTT:11 | 0xCTTTT+12xCTTT:5, 1xCTTTT+13xCTTT:8 | ✓ |
| D5S1503 | $TAGA_n$ | Trio | F | 0→4 or 8→4 | 0/8 | 4/8 | 4/4 | 48/56 | 52/56 | 52/52 | 12xTAGA:16, 13xTAGA:2, 14xTAGA:5 | 13xTAGA:8, 14xTAGA:11 | 13xTAGA:40 | ✓ |
| D8S1763 | $TG_n$ | Trio | F | 4→6 | 2/4 | 0/0 | 0/6 | 30/32 | 28/28 | 28/34 | 14xGT:1, 15xGT:19, 16xGT:22 17xGT:2 | 13xGT:2, 14xGT:18, 15xGT:1 | 12xGT:1, 13xGT:1, 14xGT:13, 17xGT:10 | ✓ |
| D12S372 | $CTAT_n$ $CTAC_m$ | Fam. | F | 20→16 | 4/20 | 0/4 | 0/16 | 56/72 | ? | 52/68 | 10xCTAT+3xCTAC:1, 11xCTAT+3xCTAC:5, 15xCTAT+3xCTAC:5 | No data | 11xCTAT+2xCTAC:7, 13xCTAT+3xCTAC:2, 14xCTAT+3xCTAC:9 | ✓ |
| D13S796 | $CTGT_n$ $CTAT_m$ | Fam. | F | 12→16 | 12/20 | 4/12 | 4/16 | 72/80 | 64/72 | 64/76 | 3xCTGT+15xCTAT:7, 3xCTGT+16xCTAT:1, 3xCTGT+17xCTAT:3 | 2xCTGT+14xCTAT:7, 3xCTGT+15xCTAT:9 | 2xCTGT+13xCTAT:1, 2xCTGT+14xCTAT:3, 3xCTGT+15xCTAT:1, 3xCTGT+16xCTAT:6 | ✓ |

Note: We used next generation sequence (NGS) data from Illumina GAllx and HiSeq2000 instruments to validate a subset of the mutations that we inferred based on PCR and electrophoresis with fluorescently labeled primers. These data were produced as a part of a large scale project in Iceland, where individuals have been sequenced to a depth of ~10-30X. Sequencing reads were aligned to the hg18 reference genome with BWA27 and duplicates were marked with Picard [http://picard.sourceforge.net/]. An inspection of the overlap between trios and families with candidate mutations and those with NGS data revealed 12 trios and 2 families that could be used for the purpose of verification, in the sense that there were at least 7 informative sequence reads for each relevant individual (minimally, the proband and the parent carrying the wild-type allele). In each case, sequence reads spanning the variable part of the microsatellite (i.e. with flanking sequence on both sides) were identified and carefully aligned by hand. This strategy was adopted because the available alignment algorithms did not seem to provide convincing results – particularly for the more complex microsatellites, composed of multiple different repeat motifs. Genotypes were called on the basis of these alignments in the following manner. First the modal allele was identified and called as allele 1 in the genotype. If this allele was present in ≥80% of the reads then the individual was deemed to be a homozygote. If not, then 5% was subtracted from the frequency of alleles that differed by one mutational step from allele 1 (in order to account for the



Nature Genetics: doi:10.1038/ng.2398

presence of apparent somatic mutational variation) and the second most frequent allele was identified. If this allele was found in ≥20% of the reads, then the genotype was called as allele 1 / allele 2. If no other allele was found in ≥20% of the reads, then the genotype was defined as an allele 1 homozygote. The table shows the results of this genotyping, which was performed blindly in relation to the electrophoretic genotypes. The allele lengths for the electrophoretic and NGS genotypes are not reported in the same scale. The former are lengths relative to the shorter allele observed in a particular reference individual (used for this purpose in all microsatellite genotyping at deCODE Genetics). The latter are absolute combined lengths of the variable repeat motifs based on sequence data (in the next three columns we show the distribution of allele lengths that were the raw data used to call the genotype). Results are consistent between both data sets in all cases but one (locus D20S902), where the electrophoretic genotype indicates that the father is a homozygote, but the NGS data reveals that the father is a heterozygote carrying an allele with a length consistent with the candidate mutation.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Table 3. Di-nucleotide microsatellite mutations by motif type**

| Repeat-type, by motif | mutations | transmissions | rate | std error |
|---|---|---|---|---|
| AC/CA/GT/TG | 1102 | 4063534 | 2.71 | 0.08 |
| AG/GA/CT/TC | 27 | 93352 | 2.89 | 0.56 |
| AT/TA | 12 | 8760 | 13.70 | 3.95 |
| CG/GC | 0 | 0 | N/A | N/A |



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 4. Differences in $\alpha$

| Mutation class | Trio data | | | | Family data | | | |
|---|---|---|---|---|---|---|---|---|
| | Paternal | Maternal | α | [95% CI] | Paternal | Maternal | α | [95% CI] |
| homozygous to homozygous | 123 | 81 | 1.52 | [1.15 2.04] | 13 | 8 | 1.63 | [0.62 4.25] |
| homozygous to heterozygous | 146 | 43 | 3.40 | [2.50 4.91] | 57 | 21 | 2.71 | [1.69 4.57] |
| heterozygous to homozygous | 104 | 42 | 2.48 | [1.75 3.56] | 25 | 14 | 1.79 | [0.95 3.88] |
| heterozygous to heterozygous | 471 | 82 | 5.74 | [4.59 7.38] | 184 | 41 | 4.49 | [3.25 6.50] |
| Total | 844 | 248 | 3.40 | [2.97 3.94] | 279 | 84 | 3.32 | [2.63 4.26] |

$\alpha$ is the ratio of the paternal mutation rate to the maternal mutation rate. Since we are only examining full trios and families (i.e. probands that have both parents genotyped), the paternal and maternal transmissions are the same, hence $\alpha$ is just the ratio of the mutations.

We split our mutations by trio/family data and by mutation class. A "homozygous to homozygous" mutation is when a parent with homozygous alleles transmits a mutation to a child with homozygous alleles, e.g. parent = (6,6) and child = (8,8).

To construct the 95% confidence interval for $\alpha$, we assume that the partition of paternal and maternal events is generated via a binomial distribution. For example, in the total mutations for trio data, assume that the paternal counts are generated with $Binomial(n, p)$, where $n = 844 + 248 = 1092$ and $p = \frac{844}{1092} = 0.773$. $\alpha$ is simulated enough times to suppress Monte Carlo noise, and then the 95% CI is obtained. Note that although we have 1,695 mutations from the trio data, only 1,092 are used here, because the parent transmitting the mutation is ambiguous for the rest (Supplementary Notes).

Comparing the trio data to the family data, $\alpha$ is not significantly different, as the 95% CI significantly overlap for each mutation class.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 5. Predictors of the mutation process

| | p-values for assessing significance in the tested variable | | |
|---|---|---|---|
| **Tested variable†** | **mutation rate** | **magnitude in step size*** | **directionality*** |
| motif length (di- vs. tetra-) | $<10^{-12}$ | $1.78 \times 10^{-9}$ | 0.58 |
| absolute length‡ | $<10^{-12}$ | 0.19 | 0.16 |
| variance in allele length distribution in Icelanders | $<10^{-12}$ | 0.70 | 0.11 |
| repeat impurity | $3.1 \times 10^{-7}$ | 0.12 | 0.26 |
| distance from exons (measured by B-statistic††) | $2.2 \times 10^{-6}$ | 0.71 | 0.74 |
| DNA replication timing | 0.005 | 0.07 | 0.69 |
| recombination rate | 0.02 | 0.49 | 0.59 |
| sequence divergence, human-chimp (10Kb window) | 0.24 | 0.61 | 0.67 |
| recombination hotspot | 0.42 | 0.83 | 0.79 |
| physical distance from telomeres | 0.86 | 0.24 | 0.40 |
| Heterozygosity | $<10^{-12}$ | 0.28 | 0.46 |
| | | | |
| parental gender | $<10^{-12}$ | 0.04 | 0.01 |
| paternal age | $9.3 \times 10^{-5}$ | 0.67 | 0.18 |
| maternal age | 0.47 | 0.33 | 0.66 |
| | | | |
| relative length*** | N/T** | $1.41 \times 10^{-7}$ | $<10^{-12}$ |

† Because our data are mostly di-nucleotides, and di and tetra-nucleotides show major differences in their characteristics, all tested variables excluding motif length, are tested only using di-nucleotides.

†† The B-statistic predicts the intensity of background selection, according to McVicker et al.[2]

‡ When regressing to mutation rate, absolute length is the mean absolute length of each locus. When regressing to step-size variance and directionality, absolute length is defined as that of the parental allele.

* For each mutation, if the mutational length is X, then the magnitude in step size is defined as the absolute value of X, and the directionality is defined as the sign of X.

** Not testable.

*** Relative length is the Z-score of the allele length, relative to the allelic distribution at the microsatellite locus. See the Methods of the main manuscript for a formal definition.



Nature Genetics: doi:10.1038/ng.2398

# Supplementary Table 6. Interactions between covariates

| Covariate $x_1$ | Covariate $x_2$ | $r^2$ | P-value $x_1$ | P-value $x_2$ | P-value $x_1 \cdot x_2$ |
|---|---|---|---|---|---|
| Genotype error rate | absolute length | 0.004 | 2.37E-01 | 2.01E-04 | 9.20E-03 |
| human-chimp divergence | absolute length | 0.000 | 8.51E-01 | 6.04E-01 | 9.75E-01 |
| human-chimp divergence | Genotype error rate | 0.002 | 2.12E-01 | 5.50E-02 | 1.67E-01 |
| recombination rate | absolute length | 0.001 | 3.30E-01 | 1.48E-13 | 5.62E-01 |
| recombination rate | Genotype error rate | 0.002 | 2.97E-01 | 5.02E-10 | 4.88E-01 |
| recombination rate | human-chimp divergence | 0.053 | 2.17E-03 | 2.56E-01 | 1.03E-03 |
| DNA replication time | absolute length | 0.000 | 1.56E-03 | 7.05E-14 | 7.47E-03 |
| DNA replication time | Genotype error rate | 0.004 | 1.98E-01 | 3.34E-12 | 1.10E-01 |
| DNA replication time | human-chimp divergence | 0.006 | 5.48E-02 | 4.31E-01 | 3.73E-02 |
| DNA replication time | recombination rate | 0.005 | 4.11E-01 | 3.77E-03 | 7.69E-03 |
| ASD | absolute length | 0.045 | 1.07E-04 | 1.41E-04 | 1.75E-01 |
| ASD | Genotype error rate | 0.019 | 5.80E-01 | 7.83E-06 | 4.45E-02 |
| ASD | human-chimp divergence | 0.000 | 4.80E-01 | 9.04E-01 | 8.55E-01 |
| ASD | recombination rate | 0.000 | 3.35E-33 | 5.94E-04 | 3.15E-03 |
| ASD | DNA replication time | 0.001 | 5.90E-33 | 3.91E-01 | 9.21E-01 |
| B-stat | absolute length | 0.000 | 1.60E-01 | 2.14E-05 | 6.46E-01 |
| B-stat | Genotype error rate | 0.000 | 4.71E-02 | 8.96E-03 | 1.49E-02 |
| B-stat | human-chimp divergence | 0.188 | 1.03E-01 | 4.20E-01 | 4.98E-02 |
| B-stat | recombination rate | 0.155 | 1.33E-01 | 5.69E-02 | 7.69E-02 |
| B-stat | DNA replication time | 0.103 | 1.65E-03 | 2.14E-03 | 3.83E-03 |
| B-stat | ASD | 0.000 | 8.35E-01 | 2.98E-15 | 2.64E-01 |
| recombination hotspot | absolute length | 0.002 | 1.08E-02 | 3.36E-14 | 9.32E-03 |
| recombination hotspot | Genotype error rate | 0.000 | 8.70E-01 | 1.31E-13 | 9.93E-01 |
| recombination hotspot | human-chimp divergence | 0.005 | 2.20E-01 | 3.14E-01 | 1.87E-01 |
| recombination hotspot | recombination rate | 0.220 | 2.94E-01 | 1.84E-02 | 2.25E-01 |
| recombination hotspot | DNA replication time | 0.002 | 1.66E-01 | 1.45E-02 | 6.33E-01 |
| recombination hotspot | ASD | 0.001 | 1.16E-01 | 8.76E-31 | 1.75E-01 |
| recombination hotspot | B-stat | 0.015 | 1.65E-01 | 2.98E-04 | 2.17E-01 |
| physical position | absolute length | 0.000 | 1.28E-01 | 9.51E-11 | 1.24E-01 |
| physical position | Genotype error rate | 0.000 | 7.45E-01 | 1.38E-06 | 8.24E-01 |
| physical position | human-chimp divergence | 0.007 | 4.98E-01 | 2.95E-01 | 4.69E-01 |
| physical position | recombination rate | 0.001 | 8.39E-01 | 1.88E-02 | 2.88E-01 |
| physical position | DNA replication time | 0.005 | 6.33E-01 | 9.53E-02 | 7.88E-01 |
| physical position | ASD | 0.001 | 4.46E-03 | 3.38E-07 | 3.49E-03 |
| physical position | B-stat | 0.004 | 3.40E-01 | 7.38E-03 | 5.01E-01 |
| physical position | recombination hotspot | 0.002 | 3.15E-01 | 8.80E-02 | 2.23E-01 |
| repeat impurity | absolute length | 0.180 | 5.82E-01 | 5.68E-31 | 9.37E-03 |
| repeat impurity | Genotype error rate | 0.001 | 8.29E-04 | 1.53E-06 | 4.12E-04 |
| repeat impurity | human-chimp divergence | 0.000 | 4.14E-01 | 2.37E-01 | 3.40E-01 |
| repeat impurity | recombination rate | 0.000 | 1.12E-01 | 1.43E-02 | 6.32E-01 |
| repeat impurity | DNA replication time | 0.002 | 1.20E-02 | 9.11E-03 | 1.31E-01 |



Nature Genetics: doi:10.1038/ng.2398

| | | | | | |
|---|---|---|---|---|---|
| repeat impurity | ASD | 0.014 | 9.70E-01 | 1.27E-28 | 6.92E-01 |
| repeat impurity | B-stat | 0.003 | 3.60E-06 | 1.19E-06 | 7.12E-06 |
| repeat impurity | recombination hotspot | 0.001 | 5.09E-02 | 3.25E-01 | 2.63E-01 |
| repeat impurity | physical position | 0.000 | 3.31E-01 | 7.44E-01 | 7.45E-01 |
| Heterozygosity | absolute length | 0.099 | 3.89E-03 | 2.11E-01 | 8.00E-01 |
| Heterozygosity | Genotype error rate | 0.014 | 6.49E-01 | 6.87E-06 | 1.68E-02 |
| Heterozygosity | human-chimp divergence | 0.001 | 3.75E-01 | 7.18E-01 | 7.71E-01 |
| Heterozygosity | recombination rate | 0.000 | 8.50E-48 | 5.55E-02 | 3.00E-01 |
| Heterozygosity | DNA replication time | 0.005 | 1.48E-53 | 2.47E-02 | 6.83E-02 |
| Heterozygosity | ASD | 0.416 | 2.31E-02 | 3.95E-04 | 9.65E-13 |
| Heterozygosity | B-stat | 0.002 | 3.13E-19 | 2.20E-01 | 7.60E-01 |
| Heterozygosity | recombination hotspot | 0.000 | 1.44E-52 | 1.13E-01 | 2.89E-01 |
| Heterozygosity | physical position | 0.000 | 3.14E-16 | 2.95E-02 | 3.22E-02 |
| Heterozygosity | repeat impurity | 0.019 | 1.79E-48 | 5.31E-01 | 4.03E-01 |



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 7. Bayesian parameters for evolution modeling

| Class | Description | Sampling distribution | Mean (SD) | Units |
|---|---|---|---|---|
| Generation interval | $g_{anc}$ Generation time in the human-chimp ancestor | Normal | 22.5 (4.24) | years |
| | $g_{now}$ Present-day human generation time | Normal | 29.0 (2.04) | years |
| | $t_0$ Inflection point of the logistic curve | Mixture of 3 exponentials of equal probability | 50<br>200<br>2000 | thousand years |
| Parental age difference (paternal minus maternal) | $\Delta_{anc}$ Age difference in the human-chimp ancestor | Normal | 0.50 (3.33) | years |
| | $\Delta_{now}$ Present-day human parental age difference | Normal | 6.00 (2.04) | years |
| Mutation rate as a function of generation interval | $\beta_{0,pat}$ Paternal mutation rate, baseline (at age 0) | multivariate t (sampled from Fig 2A) | see Fig 2A | $\mu$ |
| | $\beta_{0,mat}$ Maternal mutation rate, baseline (at age 0) | | | $\mu$ |
| | $\beta_{1,pat}$ Slope of paternal mutation rate with age | | | $\mu$ per year |
| | $\beta_{1,mat}$ Slope of maternal mutation rate with age | | | $\mu$ per year |
| Mutation rate with length | $m_\mu$ Slope of mutation rate vs. absolute allele length | Normal | 1.66 (0.30) x$10^{-5}$ | $\mu$ per repeat unit |
| Length constraint | Slope of mutational direction vs. relative allele length | Normal | -0.419 (0.060) | repeat units per SD |
| For human-chimp divergence time | $\pi_{HC}/\pi_E$ Ratio of human-chimp to Western European sequence divergence | Normal | 15.4 (0.356) | dimensionless |
| For human-chimp speciation time | $\tau_{HC}/t_{HC}$ Ratio of human-chimp speciation time to genetic divergence time | Normal | 0.663 (0.041) | dimensionless |
| For human-orangutan divergence time | $\pi_{HO}/\pi_{HC}$ Ratio of human-orangutan to human-chimp sequence divergence | Normal | 2.65 (0.075) | dimensionless |

Note: This table gives the prior distributions used in our Bayesian modeling analysis, obtained from surveys of the literature and discussions with experts in relevant fields (our approach to obtain these priors is also discussed in the Methods section). The experts we consulted were John Hawks and David Pilbeam regarding the ape fossil record; Kevin Langergraber and Linda Vigilant regarding primate generation intervals and plausible generation intervals in the ancestral population; and Jack Fenner regarding the recent human generation interval. We thank all these colleagues for useful discussions and advice.

The parameters above the thick black line are "global parameters" used for microsatellite evolution modeling, in which the same set of parameter values apply to all loci, per simulation. The parameters below the line are used after the posterior TMRCA of Western Europeans has been obtained.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 8. Mutation rate estimates and sequence heterozygosities in 23 individuals

| **Illumina dataset** | | **Sequence heterozygosity** | | **Mutation rate estimates (x $10^{-8}$)** | | | |
|---|---|---|---|---|---|---|---|
| Population | ID | mean | std error | mean | std error | 5th percentile | 95th percentile |
| **CEU** | **NA12891** | 0.000860 | 0.000026 | 1.65 | 0.44 | 1.00 | 2.43 |
| **CEU** | **NA12892** | 0.000838 | 0.000026 | 1.92 | 0.37 | 1.33 | 2.56 |
| **CEU** | **NA12878** | 0.000838 | 0.000026 | 1.42 | 0.34 | 0.91 | 2.01 |
| **YRI** | **NA19239** | 0.001112 | 0.000027 | 1.80 | 0.44 | 1.12 | 2.55 |
| **YRI** | **NA19238** | 0.001048 | 0.000027 | 2.46 | 0.53 | 1.65 | 3.38 |
| **YRI** | **NA18508** | 0.001174 | 0.000028 | 1.18 | 0.35 | 0.64 | 1.79 |
| YRI | NA19240 | 0.001168 | 0.000028 | 2.57 | 0.56 | 1.68 | 3.53 |
| YRI | NA18507 | 0.001077 | 0.000031 | 2.12 | 0.53 | 1.33 | 3.04 |
| YRI | NA18506 | 0.001141 | 0.000030 | 2.13 | 0.54 | 1.33 | 3.09 |

| **Complete Genomics dataset** | | **Sequence heterozygosity** | | **Mutation rate estimates (x $10^{-8}$)** | | | |
|---|---|---|---|---|---|---|---|
| Population | ID | mean | std error | mean | std error | 5th percentile | 95th percentile |
| **CEU** | **NA12891** | 0.000804 | 0.000025 | 1.36 | 0.31 | 0.90 | 1.90 |
| **CEU** | **NA12892** | 0.000804 | 0.000025 | 1.58 | 0.30 | 1.11 | 2.10 |
| **CEU** | **NA12878** | 0.000780 | 0.000026 | 1.15 | 0.25 | 0.77 | 1.58 |
| CEU | NA06985 | 0.000800 | 0.000027 | 1.06 | 0.28 | 0.65 | 1.54 |
| CEU | NA06994 | 0.000850 | 0.000029 | 0.91 | 0.20 | 0.61 | 1.25 |
| CEU | NA07357 | 0.000794 | 0.000027 | 1.12 | 0.31 | 0.66 | 1.67 |
| CEU | NA10851 | 0.000848 | 0.000029 | 1.00 | 0.23 | 0.66 | 1.40 |
| CEU | NA12004 | 0.000841 | 0.000028 | 1.13 | 0.29 | 0.69 | 1.63 |
| **YRI** | **NA19239** | 0.001035 | 0.000026 | 1.50 | 0.35 | 0.96 | 2.08 |
| **YRI** | **NA19238** | 0.000980 | 0.000026 | 2.09 | 0.42 | 1.44 | 2.81 |
| **YRI** | **NA18508** | 0.001089 | 0.000027 | 1.06 | 0.29 | 0.62 | 1.57 |
| YRI | NA18501 | 0.001062 | 0.000026 | 1.52 | 0.37 | 0.95 | 2.14 |
| YRI | NA18502 | 0.001062 | 0.000027 | 2.86 | 0.53 | 1.98 | 3.72 |
| YRI | NA18504 | 0.001059 | 0.000026 | 1.31 | 0.31 | 0.84 | 1.84 |
| YRI | NA18505 | 0.001076 | 0.000027 | 1.27 | 0.29 | 0.82 | 1.77 |
| YRI | NA18517 | 0.001083 | 0.000027 | 1.32 | 0.41 | 0.72 | 2.08 |
| CHB | NA18526 | 0.000798 | 0.000027 | 1.89 | 0.36 | 1.32 | 2.50 |
| CHB | NA18537 | 0.000766 | 0.000026 | 1.51 | 0.32 | 1.02 | 2.06 |
| CHB | NA18555 | 0.000779 | 0.000026 | 1.38 | 0.28 | 0.94 | 1.88 |
| CHB | NA18558 | 0.000770 | 0.000027 | 1.25 | 0.36 | 0.72 | 1.90 |

Mutation rates (in units of X*1e-8 /bp/generation) and Bayesian posterior intervals for each individual are shown here. In bold are individuals that overlap between the two datasets. See Supplementary Figure 12 for a graphical representation.



Nature Genetics: doi:10.1038/ng.2398

## Supplementary Table 9. Ascertainment bias around microsatellite loci

| Population | HapMap ID | msat region | random region | ratio |
|---|---|---|---|---|
| CEU | NA12891 | 0.088 | 0.085 | 1.037 |
| CEU | NA12892 | 0.087 | 0.082 | 1.067 |
| CEU | NA12878 | 0.090 | 0.085 | 1.057 |
| YRI | NA19239 | 0.118 | 0.113 | 1.041 |
| YRI | NA19238 | 0.110 | 0.105 | 1.046 |
| YRI | NA18508 | 0.119 | 0.116 | 1.025 |
| YRI | NA19240 | 0.121 | 0.114 | 1.059 |
| YRI | NA18507 | 0.112 | 0.107 | 1.043 |
| YRI | NA18506 | 0.118 | 0.114 | 1.036 |
| | | | | |
| human-chimp | | 2.347 | 2.248 | 1.044 |
| human-macaque | | 7.978 | 7.884 | 1.012 |

We compared sequence heterozygosity (in units of $X*10^{-2}$) of regions surrounding our set of microsatellites to that of a random region. On average, the sequence heterozygosity was about 4% higher, suggesting that we have a slight bias towards the deeper trees in the human genome. Our modeling of evolutionary parameters explicitly corrects for such biases in two ways. First, we correct for unusual mutation rates around microsatellites by normalizing inferences by the ratio of local human-macaque sequence divergence to genome-wide average human-macaque sequence divergence. Second, we correct for unusual gene tree depths around microsatellites by making all inferences based on the comparison of local microsatellite ASD to heterozygosity in the flanking sequence data.



Nature Genetics: doi:10.1038/ng.2398

**Supplementary Notes**

# Chapter 1: Estimating the Genotyping Rate

Based on the discordance rate of multiple-genotyped alleles, we estimated the per-allele genotype error rate for each locus. Formally, at a particular microsatellite locus, a single allele is observed after genotyping. There is a non-zero probability that the genotyping yielded an erroneous allele length. What is this probability of error?

Let $\hat{p} =$ Our goal. $0 \leq \hat{p} \leq 1$.
$k =$ Number of times an allele is repeatedly genotyped.
$n_k =$ Total number of individuals who were each genotyped k-times.
$y_k =$ Number of individuals that resulted in inconsistent genotypes.

For a given individual at a given locus, suppose the true bi-allelic genotype is **a**, and after genotyping, $\mathbf{b}_i$ is observed.

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \longrightarrow \mathbf{b}_i = \begin{pmatrix} a_0 + \epsilon_{i0} \\ a_1 + \epsilon_{i1} \end{pmatrix}$$

To further simplify, suppose that after repeatedly genotyping *k* times (*k* is a known quantity), with $\varepsilon_{ij}$ IID (independent and identically distributed) with probability *p* of being nonzero, we only observe the indicator random variable *X*:

$$X = 1 - \mathbb{I}(\boldsymbol{b_1} = \boldsymbol{b_2} = \cdots = \boldsymbol{b_k})$$

Assuming that the probability of making *k* identical errors is negligibly small, then

$$X \sim \text{Bernoulli}[\theta] = \text{Bernoulli}[1 - (1-p)^{2k}]$$

Suppose for *n* individuals genotyped *k* times at this particular locus, *p* is unknown but constant. Our goal is to find the optimal estimate for parameter *p*.

Thus, our data is modeled as IID $X_1 \ldots X_n \sim \text{Bernoulli}[\theta]$.

By using the maximum likelihood estimate (MLE) for the Bernoulli family, and applying the invariance property of MLE, the MLE for *p* is

$$\begin{aligned} \widehat{p} &= 1 - (1 - \bar{X})^{\frac{1}{2k}} \\ &\approx \frac{\bar{X}}{2k} \end{aligned}$$

The approximation is a 1st-order Taylor expansion around $\bar{X} = 0$, and hence is good only for sufficiently small genotype error probabilities, which we expect in this case. With this approximation, $\theta \approx 2kp$. We use this approximation for all subsequent analyses.



Nature Genetics: doi:10.1038/ng.2398

Above we gave the derivation of a single *k*. For multiple *k*, what is the best estimate of *p*, assuming *p* is constant for all *k*? To derive the correct MLE, let $Y_k = n_k \bar{X}_k$, where the subscript *k* emphasizes the dependence on *k*. It can be shown that $Y_k$ is a sufficient statistic for *p*, and

$$Y_k \sim \text{Binomial}[n_k, \theta_k] \approx \text{Binomial}[n_k, 2kp]$$

Importantly, $Y_k$ are independent for different *k*, but clearly not identically distributed.

$$\begin{aligned} l(p|\mathbf{Y}) &= \ln \prod_k \binom{n_k}{y_k} (2kp)^{y_k} (1-2kp)^{n_k - y_k} \\ &= \sum_k \ln \binom{n_k}{y_k} + y_k \ln(2kp) + (n_k - y_k) \ln(1-2kp) \end{aligned}$$

Differentiating and setting equal to 0 yields:

$$\frac{1}{\widehat{p}} \sum_k y_k = \sum_k \frac{2k(n_k - y_k)}{1 - 2k\widehat{p}}$$

Unfortunately, *p* cannot be expressed explicitly. A numerical algorithm such as Newton's Method is needed to find *p*. However, if we use the Poisson approximation to the binomial, i.e. $n_k$ is large and *2kp* is small, then an analytical solution can be found:

$$Y_k \sim \text{Poisson}[\lambda_k] \approx \text{Poisson}[n_k 2kp]$$

$$\begin{aligned} l(p|\mathbf{Y}) &= \ln \prod_k e^{-n_k 2kp} (n_k 2kp)^{y_k} / y_k! \\ &= \sum_k y_k \ln(n_k 2kp) - n_k 2kp - \ln y_k! \end{aligned}$$

Differentiating and setting equal to 0 yields:

$$\widehat{p} = \frac{\sum_k y_k}{\sum_k 2kn_k}$$

We use this formula to estimate the per allele genotype error rate at each microsatellite locus. Supplementary Figure 2 shows the distribution of error rates across the 2,477 loci. The median rate is $1.8 \times 10^{-3}$, with a 95% central range of $1.7 \times 10^{-4}$ to $1.4 \times 10^{-2}$. Since this number is comparable to the expected microsatellite mutation rate, a simple search for mutations using trios genotyped at 1× coverage will lead to many erroneous mutations. Thus, we developed the "trio approach" and "family approach" to obtain mutations that are highly likely to be genuine.



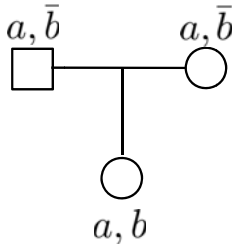
Nature Genetics: doi:10.1038/ng.2398

# Chapter 2: Details of the trio approach in mutation detection

## Mutations with ambiguous parental origin

In the trio approach, since we do not phase the alleles using neighboring microsatellites, there are cases in which the parental origin is ambiguous. Below we describe how this scenario occurs.

Let *a* and *b* be distinct alleles. Let  be any allele that is not *b*. If there are multiple instances of , they are not required to be equal. Then, the following mutant case has ambiguous parental origin:

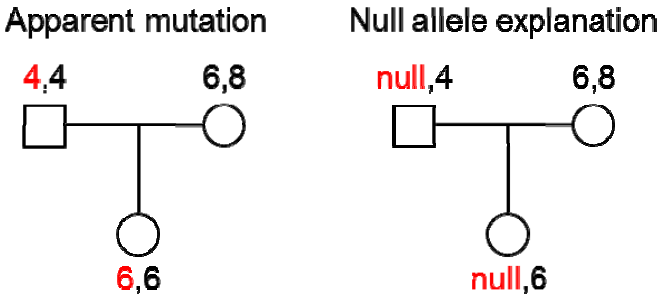



In this pattern, allele *a* is the allele that is also present in the parents, and allele *b* is the mutant. However, since we cannot identify the parental origin of *a*, that of *b* is also ambiguous. Note that we do not attempt to assign *b* to the parent who has a smaller delta in the mutational length, if such a parent exists.

## Excessive mutations from homozygous-parent to homozygous-child

After identifying mutations, we discovered that certain loci exhibited many more *de novo* mutations from homozygous parents to homozygous children than would be expected based on Hardy-Weinberg equilibrium. We suspected that these loci might be generating false mutations due to polymorphisms under PCR primer sites, leading to allele-specific PCR mis-amplification.

An example is shown below (left panel), in which there is an apparent mutation from father's allele 4 to child's allele 6. Alternatively, this can be explained by a null allele (right panel). This could be due to (1) a polymorphism in the PCR primer site, resulting in mis-amplification, or (2) a deleted allele, both of which would mean that there is no real mutation.






Nature Genetics: doi:10.1038/ng.2398

We removed loci that have an excess rate of homozygous-to-homozygous mutations, compared with the expectation from Hardy-Weinberg equilibrium. To do this, for each locus we compare the observed homozygosity of all alleles to the observed homozygosity of the mutations. We perform a one-sided binomial test and remove any locus with a p-value < 0.05 (plus a Bonferroni correction by a factor of 2477, the number of loci examined). Formally, for each locus let

$p =$ Observed homozygosity of all alleles genotyped. $0 \leq p \leq 1$.
$n =$ Number of mutations observed.
$k =$ Number of mutations that are from a homozygous-parent to a homozygous-child

$$\text{P-value} = \sum_{i=k}^{n} \binom{n}{i} p^{2i}(1-p^2)^{n-i}$$

Note that we have $p^2$ instead of $p$ because we are observing two homozygous genotypes simultaneously. In this manner, 49 loci were removed from the trio approach.



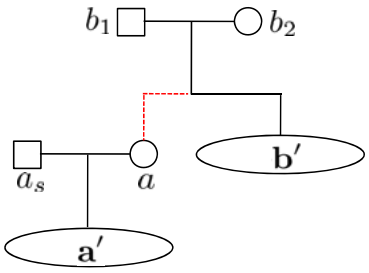
Nature Genetics: doi:10.1038/ng.2398

# Chapter 3: Details of the family approach in mutation detection

## Assigning alleles to haplotypes: a constraint satisfaction problem

Since Allegro cannot determine haplotypes in the presence of a mutation (a Mendelian inheritance error), we initially mask out any locus that generates inheritance errors. Based on neighboring loci, Allegro imputes haplotypes into the masked loci. To optimally assign haplotypes to alleles, this problem can now be posed as a constraint satisfaction problem (CSP) and solved.

**Goal:** Given the family structure below, a set of haplotypes, and a set of alleles at a particular locus, assign haplotypes to alleles in a way that is consistent with the family structure.




**Solution:**

We formulate this problem in terms of a constraint satisfaction problem (CSP). Suppose we have individuals $I_1, I_2, \ldots, I_m$ and haplotypes $H_1, H_2, \ldots, H_n$, where $n$ is even. Then, we can write the alleles in a sparse matrix format, as shown below. Each row is an individual, each column is a haplotype, and each matrix entry is the pair of alleles of the corresponding individual. Since each individual has 2 haplotypes, we have 2 matrix entries per row. The CSP problem is then to find the suitable unique number for each matrix entry.

Formally, the set of variables is the non-empty entries of the matrix, denoted as $X_{ij}$. In the example below, there are 6 variables. Each variable has a domain of values. Since loci are diploid, we have 2 values per domain. There are two constraints for this CSP: (1) The non-empty entries of each column must be equal. (2) The non-empty entries of each row must be different, unless the domain is a homozygote, such as "7, 7". The desired outcome of the CSP is shown below.

**CSP in the presence of mutation.** Without mutations, we simply run the algorithm over the entire family in one batch. However, suppose that there is a candidate mutant in the proband, then a single batch CSP would yield an empty solution. To resolve this, we instead use the following steps: (1) Run CSP over $b_1$, $b_2$, and *b'*. This group should carry the ancestral allele. (2) Run CSP over $a$, $a_s$, and *a'*. This group should carry the mutant allele. At this point, we should have the 6 six haplotypes assigned to the alleles, with 1 haplotype assigned inconsistently between the two groups. Thus, in combining the results, we have successfully identified the haplotype carrying the mutant, the mutant allele, and the ancestral allele.

**Example.** In this family, we have 2 members of *a'* and 2 members of *b'*. We first run CSP over the ancestral group, yielding:



Nature Genetics: doi:10.1038/ng.2398

| | $H_1$ | $H_2$ | $H_3$ | $H_4$ | | | $H_1$ | $H_2$ | $H_3$ | $H_4$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $b_1$ | $4,8$ | $4,8$ | | | | $b_1$ | $8$ | $4$ | | |
| $b_2$ | | | $2,8$ | $2,8$ | $\longrightarrow$ | $b_2$ | | | $2$ | $8$ |
| $b'$ | $2,8$ | | $2,8$ | | | $b'$ | $8$ | | $2$ | |
| $b'$ | | $4,8$ | | $4,8$ | | $b'$ | | $4$ | | $8$ |

This yields a haplotype assignment of

$$\{H_1 = 8, H_2 = 4, H_3 = 2, H_4 = 8\}$$

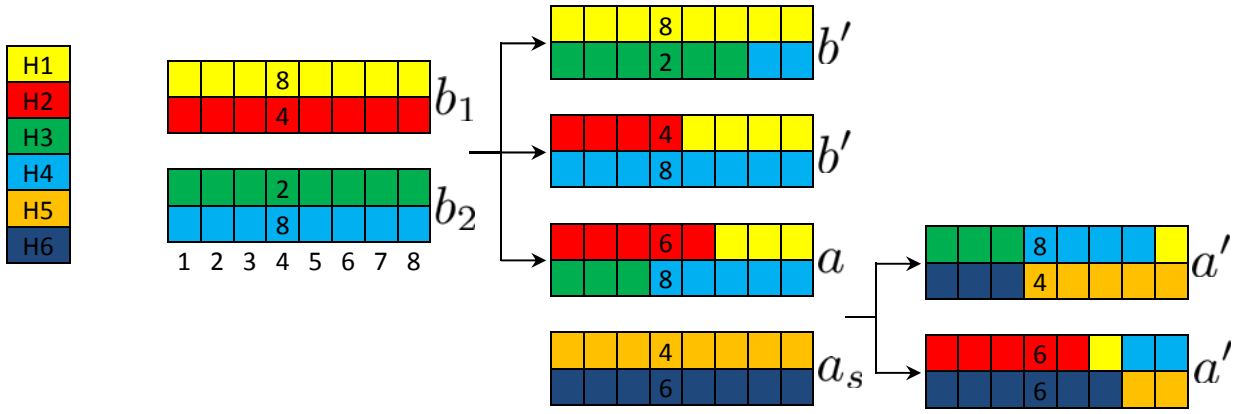
Next, we run CSP of the mutant group, yielding:

| | $H_2$ | $H_4$ | $H_5$ | $H_6$ | | | $H_2$ | $H_4$ | $H_5$ | $H_6$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $a$ | $6,8$ | $6,8$ | | | | $a$ | $6$ | $8$ | | |
| $a_s$ | | | $4,6$ | $4,6$ | $\longrightarrow$ | $a_s$ | | | $4$ | $6$ |
| $a'$ | | $4,8$ | $4,8$ | | | $a'$ | | $8$ | $4$ | |
| $a'$ | $6,6$ | | | $6,6$ | | $a'$ | $6$ | | | $6$ |

This yields a haplotype assignment of

$$\{H_2 = 6, H_4 = 8, H_5 = 4, H_6 = 6\}$$

We see that haplotype 2 is inconsistent between the two sets of assignments. Therefore, haplotype 2 is the one of interest, carrying ancestral allele 4 and mutant allele 6. Below is the full haplotype of the entire region and the 4th microsatellite locus as the mutating one:




Note that in this example, if we instead used the trio approach, i.e. we are limited to the data of $b_1 = (4,8)$, $b_2 = (2,8)$, $a = (6,8)$. The mutant allele of 6 would be detected, but we would not be able to find the parental origin of the mutation. Thus, by using additional family members and neighboring loci, the family approach allows parental assignment of the mutation.



Nature Genetics: doi:10.1038/ng.2398

# Chapter 4: Testing the Heterozygote Instability Hypothesis

Amos et al.[3] suggested that if the parental allele is heterozygous, the mutation rate will be elevated compared to homozygous parental alleles. This would have significant implications as population size (N) is related to heterozygosity, and thus $\mu = f(N)$ would significantly undermine the population genetics assumption that $N$ and $\mu$ are independent.

We tested the Heterozygote Instability hypothesis as follows:

The Heterozygote Hypothesis: If the parent is more heterozygous (i.e. length differences of alleles are large), then the mutation rate is higher.

Prediction of the hypothesis: For each microsatellite mutation, the magnitude of length difference in the parent who transmitted the mutation is expected to be larger than that of an individual randomly sampled at the same microsatellite locus.

Definitions:

| | |
|---|---|
| $\Omega$ | The entire sample space of individuals genotyped. |
| $S'$ | The subspace of parents who transmitted mutations. |
| $S$ | The subspace of individuals who do not belong to $S'$ (complement of $S'$). |
| $A_j\ B_j$ | A random sample of a pair of alleles from $S$ at locus $j$. |
| $A'_j\ B'_j$ | Likewise, but sampled from $S'$. |
| $L_j$ | The length difference of the alleles, i.e. $L_j = \lvert A_j - B_j \rvert$ |
| $L'_j$ | Likewise, but sampled from $S'$. |

Formalized hypothesis: Given the definitions, and assuming the hypothesis is true, then $L' - L > 0$ is true over the set of loci $J$.

Testing the hypothesis:

Dataset: 363 mutations from the family approach. We do not use trio mutations for this analysis, because in trio mutations we have directly filtered based on the excessive homozygosity of certain mutant loci. Since the filter directly influences the parameter we are trying to estimate, we cannot use the larger trio dataset.

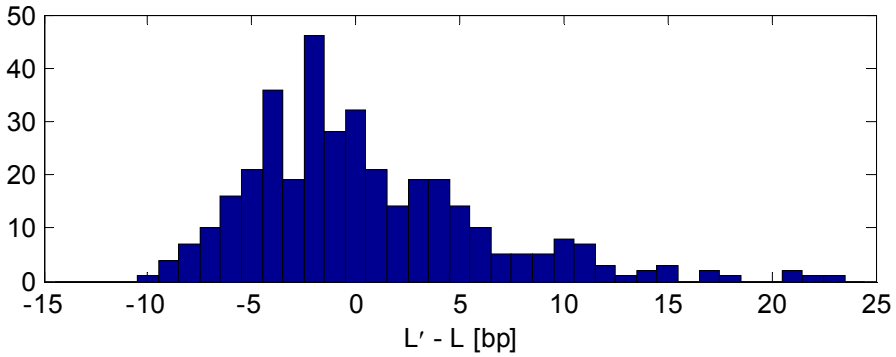
Sampling $L'$: We use the parents who transmitted the mutations. Thus, $l'_j =$ parental allele difference for case $j$.



Nature Genetics: doi:10.1038/ng.2398

Sampling $L$: For each mutation case, we take that locus' allelic distribution, and independently sample $n$ length differences and take the average. More precisely, at case $j$, we sample and compute $l_j = \frac{1}{n}\sum_{i=1}^{n} l_{j,i}$..

Results: Below is the histogram for the 363 data points of $l'_j - l_j$, with $n = 1000$. To test whether the mean is significantly different from 0, we perform a one-sample two-sided t-test, as was done by Amos et al., and obtain $t_{362} = 1.48,\ p = 0.14$. Therefore, our data provide no significant support for the Heterozygote Instability hypothesis.






Nature Genetics: doi:10.1038/ng.2398

# Chapter 5: Microsatellite evolution modeling to infer TMRCA

## I. Overview

Using the mutational characteristics that we observed, we can build a model of microsatellite evolution through time. Given additional parameters summarizing evolutionary history, such as the coalescent time ($t_{MRCA}$) of modern-day Western Europeans, we can simulate allelic distributions of microsatellites at any genotyped locus. By optimally matching statistics (such as ASD) of the simulated allelic distribution to that of the empirically observed data, we can infer parameters of interest such as $t_{MRCA}$.

Given any local region of the genome, $t_{MRCA}$ between individuals in that region (assuming no recombinations occurred in the region) must be constant, regardless of whether the genomic features examined are microsatellites or nucleotide substitutions. Therefore, once we have determined $t_{MRCA}$ at each microsatellite locus, we can use that value in conjunction with neighboring sequence divergence to infer parameters such as the sequence mutation rate. Furthermore, given a ratio of human-chimpanzee $t_{MRCA}$ to Western-European divergence, we can use our Western-European $t_{MRCA}$ to estimate the genetic divergence of present-day humans to chimpanzees. A key point is that all inferences here are performed without a calibration to the fossil record.

## II. Model design

At a particular microsatellite locus, a single run consists of simulating a coalescent tree, adding mutations onto the branches of the tree, and finally collecting simulated data at the leaf nodes. By default, the coalescent tree has time in units of generations. When conducting inferences that require time in years, we rescale the branch lengths into years following a generation-time function, as described below.

### 1. Demography: Generating the coalescent tree

We use the 2-bottleneck model from Keinan et al.[1] (Fig S13). Coalescent trees are sampled using this demography.

### 2. Variation of generation-time in history

In modern-day human populations, the average time per generation is about 29 years[4]. However, this number is likely to have been different in the past. To simulate variation in generation-time, we use the logistic curve

$$g(t) = g_{anc} + \frac{g_{now} - g_{anc}}{1 + \exp\left(\frac{t - t_0}{t_0/4}\right)}$$



Nature Genetics: doi:10.1038/ng.2398

Where we define

| | |
|---|---|
| $g_{anc}$ | Generation time of the common ancestor of humans and chimpanzees |
| $g_{now}$ | Generation time of present-day humans |
| $t_0$ | Inflection point of an assumed rapid change between $g_{anc}$ and $g_{now}$ |

These 3 parameters are stochastic. The shapes of the distribution, means, and variances are given in Supplementary Table 7. To determine $g(t)$, we first sample these 3 parameters from their distributions.

## 3. Scale coalescent tree into units of years

The $g(t)$ logistic function is the transformation factor from generations to years. When it is necessary to make inferences in years, we use $g(t)$ to rescale branch lengths as follows: The mean generation-time between a node and its parent is analytically calculated as

$$\bar{g}(t_1, t_2) = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} g(t)dt \; = g_{now} + \frac{g_{now} - g_{anc}}{r_2 - r_1}\log\frac{1 + \exp(r_1 - 4)}{1 + \exp(r_2 - 4)}$$

Where we define

| | |
|---|---|
| $t_1$ | Time of current node, in units of generations |
| $t_2$ | Time of parental node, in units of generations |
| $r_1$ | $4t_1/t_0$ |
| $r_2$ | $4t_2/t_0$ |

Once $\bar{g}(t_1, t_2)$ is calculated, that particular branch length is trivially scaled into time in units of years.

## 4. Mutation generation

Mutations are added onto the coalescent tree, sequentially from the root to the leaves. We first generate the baseline mutation rate, which is governed by the mean number of repeats of the microsatellite locus (Fig 2C). Furthermore, using our empirical observations, we build into our model that the mutation rate changes dynamically as generation-time and allele length change (Fig 2A,C) as we propagate from the root to the leaves of the tree. Finally, as mutations are generated, there is a constraint on allele length (Fig 2D). The details are given below.

(a) The locus-specific baseline mutation rate: For a given locus, we first establish the mutation rate $\mu_0$, which is constant throughout the coalescent tree. This baseline mutation rate is determined using the mean absolute length.



Nature Genetics: doi:10.1038/ng.2398

(b) Generation-time effect: In Fig 2A we observed that parental age affects mutation rate. Since generation-time $g(t)$ is modeled as varying as we travel down the coalescent tree, $g(t)$ causes a dynamic change in the mutation rate. In Fig 2A we demonstrated a difference in the paternal and maternal behavior, and we therefore first split generation-time into paternal time $g_{pat}(t)$ and maternal time $g_{mat}(t)$:

$$g_{pat}(t) = g(t) + 0.5 \cdot \Delta(t)$$
$$g_{mat}(t) = g(t) - 0.5 \cdot \Delta(t)$$

$\Delta(t)$ is the mean difference between paternal and maternal age, at time *t*. Note that this is a time-varying quantity too, as Δ of present-day humans could be different from that of the human-chimp common ancestor. In particular, we model $\Delta(t)$ as entirely analogous to the logistic function of $g(t)$.

$$\Delta(t) = \Delta_{anc} + \frac{\Delta_{now} - \Delta_{anc}}{1 + \exp\left(\frac{t - t_0}{t_0/4}\right)}$$

$\Delta_{now}$ and $\Delta_{anc}$ are sampled values. (See Supplementary Table 7 for the distributions, means, and variances used.) $t_0$ uses the same value sampled from $g(t)$ and hence is not a new sample. Once $g_{pat}(t)$ and $g_{mat}(t)$ are determined, we can obtain the gender-specific mutation rates and the gender-averaged mutation rate:

$$\mu_{pat}(t) = \beta_{0,pat} + \beta_{1,pat} \cdot g_{pat}(t)$$
$$\mu_{mat}(t) = \beta_{0,mat} + \beta_{1,mat} \cdot g_{mat}(t)$$
$$\mu_g(t) = \left(\mu_{pat}(t) + \mu_{mat}(t)\right)/2$$

Where we define

| | |
|---|---|
| $\beta_{0,pat}, \beta_{0,mat}$ | The intercepts of regressions in Fig 2A |
| $\beta_{1,pat}, \beta_{1,mat}$ | The slopes of regressions in Fig 2A |

To take into account the stochasticity of the slopes and intercepts, these quantities are sampled from the data, using a Bayesian analysis of simple linear regression (or equivalently, a draw from the multivariate student-*t* distribution).

We can summarize $\mu_g(t)$ using the matrix notation below:

$$\mu_g(t) = \frac{1}{2}[1 \quad 1] \cdot \left(\begin{bmatrix} \beta_{1,pat} & 0 \\ 0 & \beta_{1,mat} \end{bmatrix} \begin{bmatrix} 1 & 1/2 \\ 1 & -1/2 \end{bmatrix} \begin{bmatrix} g_{anc} & g_{now} \\ \Delta_{anc} & \Delta_{now} \end{bmatrix} \begin{bmatrix} 1 - f(t) \\ f(t) \end{bmatrix} + \begin{bmatrix} \beta_{0,pat} \\ \beta_{0,mat} \end{bmatrix}\right)$$

Where $f(t) = \frac{1}{1+\exp\left(\frac{t-t_0}{t_0/4}\right)}$

We highlight two special cases:



Nature Genetics: doi:10.1038/ng.2398

i. If mutations are entirely generation-like, i.e. $\beta_1$ for both parents are 0, then the expression simplifies to $\mu_g(t) = (\beta_{0,pat} + \beta_{0,mat})/2$. Thus, as expected in this case, the mutation rate does not vary as a function of generation interval.

ii. If mutations are entirely year-like, i.e. $\beta_0$ for both parents are 0 and $\beta_{1,pat} = \beta_{1,mat}$, then the expression simplifies to $\mu_g(t) = \beta_1 \cdot g(t)$. Hence the mutation rate per generation perfectly correlates with generation-time. However, the mutation rate per year, $\mu_g(t)/g(t)$, becomes a constant.

(c) Generating the instantaneous mutation rate: At any point along the coalescent tree, the instantaneous mutation rate is a function of the baseline rate, generation-time, and allele length. We combine these three factors to generate the mutation rate $\mu(t)$:

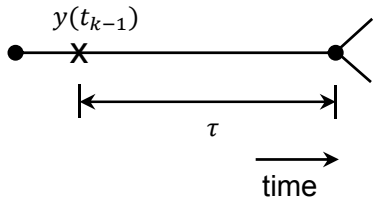
$$\mu(t) = (m \cdot y(t) + \mu_0) \cdot \frac{\mu_g(t)}{\mu_g(0)}$$

Where we define

| | |
|---|---|
| $y(t)$ | The allelic length of the branch at time *t* |
| $m$ | The slope in Fig 2C that relates allelic length to mutation rate |
| $\mu_0$ | The baseline mutation rate described in part (a) |
| $\mu_g(t)$ | The mutation rate as a function of generation time, as described in (b) |
| $\mu_g(0)$ | The present-day mutation rate, as determined by the $\mu_g(t)$ |

Note that this mutation rate model simplifies to that of the generalized stepwise mutation model (GSMM) if $m = 0$ and $\mu_g(t) = \mu_g(0)$.

(d) Generating mutation events: Suppose we are on a branch (shown below) where the *(k-1)*-th mutation occurred at $t_{k-1}$, which is marked by the "X". The allele length immediately following that event is $y(t_{k-1})$ and the generation-time is $g(t_{k-1})$. Mutation events are simulated forward in time, from the root of the tree, using an exponential distribution with mean $\mu(t_{k-1})$, which is determined from the equation in part (c). After a random sample $T \sim Exp(\mu(t_{k-1}))$ is drawn, if $T < \tau$, generate a mutation with length $Y(t_k)$ and update $\tau$ to be $\tau - T$. Otherwise, there are no more mutations in the branch and move on to the next branch. Details for generating $y(t_k)$ are described in the next section.




The process for generating mutation events for a coalescent tree re-scaled into units of years is very similar, except that the mutation rate at any point in time is divided by the generation-time, e.g. we set the mutation rate per year to be $\mu(t_{k-1})/g(t_{k-1})$.



Nature Genetics: doi:10.1038/ng.2398

(e) Generating microsatellite lengths for each mutation event: In the GSMM, the microsatellite length $y(t_k)$ is the parental length plus the mutational length, which is an independent random sample from the mutation length distribution, defined as $x$ for the *k*-th mutation event. However, using our empirical observations (Fig 2D, Fig S7, Fig S8), we model the fact that longer microsatellites tend to mutate to a shorter length, and vice versa, as a linear function:

$$y(t_k) = y(t_{k-1}) + x(t_k) + \frac{y(t_{k-1})}{\sigma} m$$
$$= \left(1 + \frac{m}{\sigma}\right) y(t_{k-1}) + x(t_k)$$

Where we define

| | |
|---|---|
| $x(t_k)$ | The mutation length, drawn randomly from the mutation length distribution in Fig 2B |
| $y(t_{k-1})$ | The microsatellite allele length, just prior to the mutation |
| $y(t_k)$ | The microsatellite allele length, just after the mutation |
| $m$ | The slope in Fig S8A. This quantity is negative, generating the length constraint. |
| $\sigma$ | The standard deviation of the allelic distribution of the locus, based on empirical data |

Observations:

- Note that while $\sigma$ is locus specific, $m$ was obtained from the combined mutational data of all loci.
- If $m = 0$, this equation reduces to the GSMM.
- At the root of the coalescent tree, we begin with allele length of $x_0$, which is determined from the empirical allele length distribution. However, we set $y(t_{root}) = 0$ when propagating mutations. When collecting allele lengths at the leaf nodes, $x_0$ is added back in.
- $y(t_{k-1})/\sigma$ produces a Z-score (horizontal axis of Fig S8A) showing the degree of deviation from the mean length, and through multiplication with slope $m$, gives the strength of the return-to-mean length constraint.

## III. Model simulation

For an individual whose genome sequence is available, diploid microsatellites genotypes are simulated as follows:

1. **Generate 1 set of genome-wide parameters** (Supplementary Table 7), which are common across loci, sampling from the prior distributions obtained from the literature and our direct measurements in this study. This includes the genome-wide sequence mutation rate and microsatellite mutation rate.



Nature Genetics: doi:10.1038/ng.2398

2. **At locus $i = 1$, generate locus-specific mutation rate $\mu_{msat,i}$.** The local microsatellite mutation rate is the genome-wide rate multiplied by $l_i/l_{genome}$, where $l_{genome}$ is the genome-wide mean microsatellite length, and $l_i$ is the locus-specific length (averaged across individuals). The local variation in microsatellite mutation rate is modeled to be purely due to allele length variation, which strongly influences mutation rate (Figure 2C).

3. **At the locus, generate locus-specific mutation rate $\mu_{seq,i}$.** Analogous to step 2, the local sequence mutation rate is the genome-wide rate multiplied by $D_i/D_{genome}$, where $D_i$ is the local human-macaque divergence, and $D_{genome}$ is the genome-wide human-macaque divergence. The local variation in sequence mutation rate is modeled to be purely due to human-macaque divergence variation, which is known to strongly influence mutation rate.

4. **At the locus, generate coalescent time $t_i$,** using local sequence heterozygosity if available. The key is that the coalescent tree is shared between microsatellites and sequence, and if the local sequence heterozygosity is highly precise, it puts a strong constraint on the local TMRCA. The coalescent time is drawn from a gamma distribution with mean: $\frac{N_i+1}{\lambda_i+1/\tau_{genome,i}}$, where $\lambda_i = 2\mu_{seq,i}D_i$, $N_i/D_i$ is the local heterozygosity, and $\tau_{genome,i} = \theta_{genome}/2\mu_{seq,i}$ is the genome-wide average TMRCA. Note that if $D_i$ is small, we revert to the genome-wide TMRCA, but if $D_i$ is large, the locus-specific heterozygosity overwhelms the genome-wide estimate. The gamma distribution is demography-free: If $D_i$ is small, the distribution converges to an exponential with mean $\tau_{genome,i}$. To test our inference's robustness to demographic differences across populations, we use a 2-bottleneck demographic model (Supplementary Figure 13) and sample the coalescent time using rejection sampling with the following steps: (1) Sample $\tau_{genome,i}$ with demography (distributions for each population shown in Supplementary Figure 13B); (2) calculate the importance ratio of $r = \exp\left[(N_i - \lambda_i t) \cdot \ln(\lambda_i t) - \sum_{i=1}^{N} \ln i + \sum_{i=1}^{\lfloor \lambda_i t \rfloor} \ln i\right]$; (3) accept $t$ with probability $r$; (4) If rejected, go to step (1).

5. **Simulate mutations.** Mutations are sequentially generated from the root of the coalescent tree, using our model of microsatellite evolution which has length constraints and time-varying mutation rate as follows: At time $t$ on the coalescent tree, the mutation rate is determined using parental length $y(t)$, mutation rate $\mu_i$, and the mutation rate relative to the present, taking into account variation in generation-time: $\mu_g(t)/\mu_g(0)$. We model this as: $\mu_i(t) = \left(m_\mu \cdot y(t) + \mu_i\right) \cdot \mu_g(t)/\mu_g(0)$. The slope parameter $m_\mu$ is empirically determined from Figure 2C. The waiting time until a mutation is sampled from an exponential distribution with mean of $1/\mu(t)$ generations. Once a mutation event occurs, its length is $l_{child} = (1 + m/\sigma)\, l_{parent} + X$, where $m$ is the negative slope reflecting the length constraint in Supplementary Figure 8, $\sigma$ is the standard deviation of the allelic distribution at a locus, $l_{parent}$ is the parent allele length, and $X$ is the mutational length, sampled from the histogram in Figure 2B. At the root of the tree, without-loss-of-generality the absolute length is set to be 0. Using this scheme of generating mutation events and mutation lengths, we begin at the root of the tree and iterate until the leaves are reached. The leaves are the sets of sampled microsatellite alleles, which are used to compute ASD. To obtain time in units of years, we



Nature Genetics: doi:10.1038/ng.2398

rescale branch lengths of the coalescent tree and mutation rates by $g(t)$, which is the generation-interval logistic function described above.

6. **Record ASD between the two microsatellites**, and go to Step 2, with $i$ incremented by 1.

We use a Markov Chain Montel Carlo (MCMC) approach to obtain the posterior distribution for present-day sequence mutation rate in a single diploid individual. This algorithm is a variation of "algorithm F" of Marjoram et al[5], and is as follows:

1. Sample a set of global parameters $\lambda$ from their prior distribution (Supplementary Table 7).
2. Propose a move of the sequence mutation rate from $\mu_{seq}$ to $\mu'_{seq}$. We use $\mu'_{seq}$ as a random walk, sampled from a normal distribution with mean $\mu_{seq}$, and standard deviation $0.5 \times 10^{-8}$.
3. At locus $i$:
    a. Generate 1000 pairs of microsatellite alleles using our evolution model with parameters $\mu'_{seq}$ and $\lambda$.
    b. Calculate ASD. Thus, we now have 1000 samples of simulated ASD.
    c. Compute the error distance $d_i = (mean(ASD_{sim}) - ASD_{real})^2$ between the simulated ASD and the real ASD of the individual.
4. Sum the error distance across all loci: $d_{total} = \sqrt{\sum_i d_i}$. If $d_{total} < \epsilon$, accept and set $\mu_{seq}$ to be $\mu'_{seq}$ and go to step 2. Otherwise, reject $\mu'_{seq}$. We choose $\epsilon$ such that the overall acceptance rate of the MCMC is between 10% and 50%. (Note that since the proposal function is symmetric, and we choose a flat prior on $\mu_{seq}$, we do not need to calculate the ratio as described in Step F4 of Marjoram et al., because the ratio is always 1.)

The result of MCMC is a correlated $\mu_{seq}|\lambda$ chain. To collect independent samples, the autocorrelation function of the chain is calculated and the correlogram is plotted. The first lag in which the correlation coefficient drops below 0.1 is recorded. Call this $n_{lag}$. Then, we thin the chain and collect at every $n_{lag}$-th sample. Finally, we run 1000 independently sampled $\mu_{seq}|\lambda$ and combine the thinned samples to produce the overall posterior distribution for $\mu_{seq}$.



Nature Genetics: doi:10.1038/ng.2398

# Chapter 6: Testing the microsatellite evolution model

## Overview

To test our procedure for using the microsatellite mutation model to estimate evolutionary parameters, we use two approaches. First, we show that our inferences based on the model produce unbiased sequence mutation rate estimates. To do this, we simulate microsatellite alleles and sequence heterozygosity using a 2-bottleneck demographic model (Fig S13), with a known sequence mutation rate and effective population size. Then, with the simulated sequence and microsatellite data, we infer the sequence mutation rate and compare it to the truth.

Second, we show that the model is robust to each parameter's prior probability distribution: we use different parameter values for our prior and show that our inferences of the sequence mutation rate and human-chimpanzee speciation time are not greatly affected (Fig S10).

## I. Simulated data shows that the model is unbiased

Procedure:

1. Choose a sequence mutation rate to use in simulation: [1.0, 1.5, 2.0, 2.5, 3.0] $x10^{-8}$ per bp per generation. Use $N_e$ of 12,500 for the 2-bottleneck demography model (Fig S13). Generate a set of global parameters (Supplementary Table 7).

2. Based on the demographic model and mutation rate chosen for the simulation, generate the local TMRCA for each individual at each locus, followed by the local sequence heterozygosity and microsatellite ASD. Generate the local sequence heterozygosity using a Poisson process, and the local microsatellite ASD using our model of evolution.

3. Run the Markov Chain Monte Carlo inference to obtain a posterior sequence mutation rate estimate for each individual, without any knowledge of the values from Step 1 used in generating the data (we also do not use knowledge about the values of the global parameters used in the simulations).

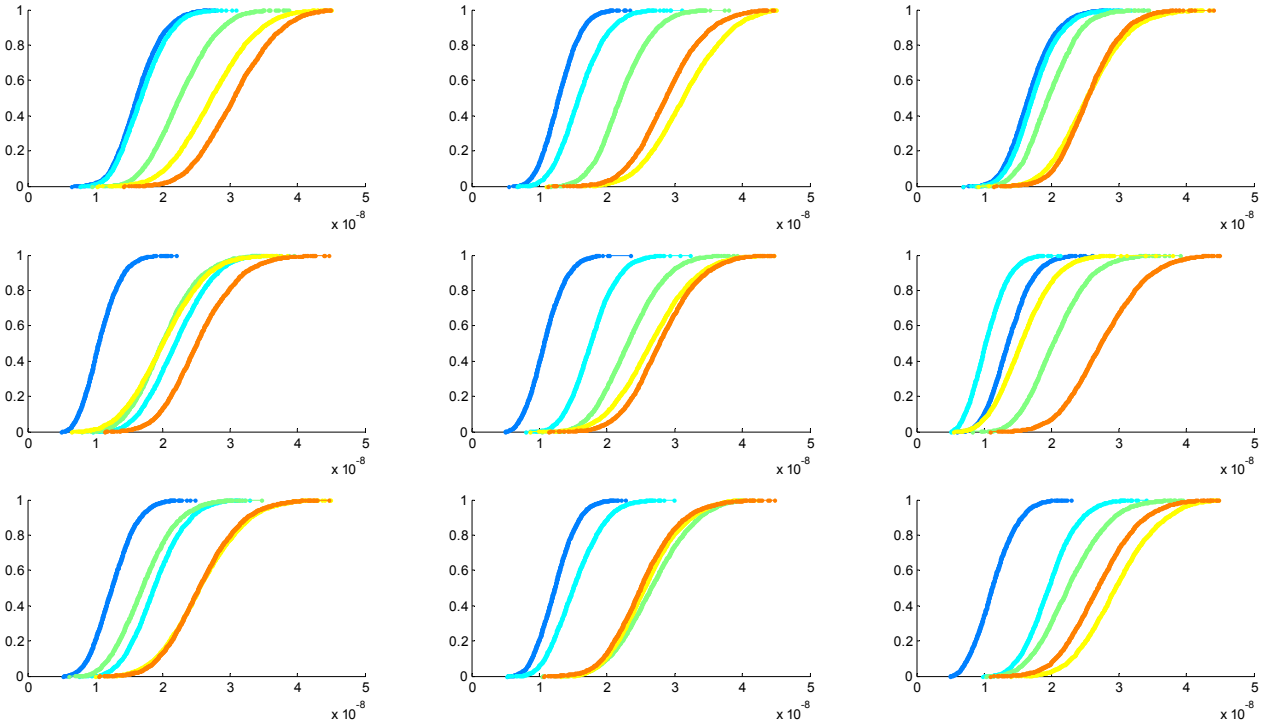
4. Obtain inferences for 9 individuals, for each of 5 mutation rates, resulting in 45 posterior distributions for sequence mutation rate. With these results, we can report the fraction of simulations in which the true TMRCA falls in the 90% Bayesian credible interval.

Results:

The CDFs (cumulative distribution function) of posterior sequence mutation rate are shown below, one panel per individual. There are 5 curves for each individual, each corresponding to a different true mutation rate: [Blue=1.0, Cyan=1.5, Green=2.0,

Yellow=2.5, Red=3.0] x$10^{-8}$. The table summarizes the results by the percentile (of the posterior distribution) in which the true mutation rate lies. Only in 3 of 45 cases (6.7%) does the true mutation rates fall outside the 90% Bayesian credible interval.

| | True sequence mutation rate | | | | |
|---|---|---|---|---|---|
| | **1.0E-08** | **1.5E-08** | **2.0E-08** | **2.5E-08** | **3.0E-08** |
| **Person 1** | 0.018 | 0.317 | 0.297 | 0.349 | 0.462 |
| **Person 2** | 0.137 | 0.412 | 0.302 | 0.123 | 0.607 |
| **Person 3** | 0.011 | 0.247 | 0.553 | 0.485 | 0.846 |
| **Person 4** | 0.427 | 0.055 | 0.514 | 0.826 | 0.815 |
| **Person 5** | 0.399 | 0.214 | 0.253 | 0.398 | 0.670 |
| **Person 6** | 0.107 | 0.944 | 0.485 | 0.983 | 0.675 |
| **Person 7** | 0.208 | 0.166 | 0.759 | 0.470 | 0.802 |
| **Person 8** | 0.211 | 0.502 | 0.101 | 0.461 | 0.838 |
| **Person 9** | 0.347 | 0.102 | 0.312 | 0.199 | 0.727 |



## II. The model is robust to changes in the parameter prior distributions

Each parameter in our evolution model has a prior distribution governing its uncertainty. We therefore explored how changing the value of the parameter—within the plausible range given by the prior—influences our inferences about the sequence mutation rate and human-chimpanzee speciation time.



Nature Genetics: doi:10.1038/ng.2398

To test for robustness of our priors, for each of 8 parameters (Fig S10), instead of using the default prior distribution, we set them to point values at three different points: the lower 95% CI, the mean, and the upper 95% CI. Then, this altered set of parameters was fed through our inference process. The primary purpose of this exercise was to see whether an extreme value of the prior, if used, would cause our inferences to change greatly. Reasonable extreme values are at the boundary of our prior distribution specifications. The second purpose is to see whether shrinkage in the variance (to zero) of any prior would cause a significant shrinkage in the variance of the posterior estimates. Note that we only perturb one parameter at a time.

As shown in Fig S10, using our model of evolution, our inference of sequence mutation rate and human-chimpanzee speciation date is reasonably robust to changes in the prior, both in the mean and in the standard error of the inferred distributions. We observe the following:

- Aside from the length constraint parameter, when we use extreme values, the inference on the sequence mutation rate does not change significantly. This suggests that (1) our priors are reasonably tight such that no significant changes are observed, or (2) the model is not heavily dependent on that parameter. For example, case (1) holds for the microsatellite mutation rate parameter: although the microsatellite mutation rate can in principle affect our inferences greatly since it has a linear effect on ASD, it is determined with high precision by our direct observations of mutations, with a 95% CI of 2.56-2.91 $x10^{-4}$; thus, the extreme values of this prior do not affect our inferences substantially.

- The length constraint governs the non-linear mapping between TMRCA and ASD (Fig 3), and changes to it (Fig S11) can cause large changes to our inferences on the sequence mutation rate. Our prior distribution for this parameter was determined entirely based on the direct observation of mutations (Fig S8, Supplementary Table 7), and not on comparisons between microsatellite ASD and sequence heterozygosity (Fig 3, Fig S11). As a result, the length constraint prior was not determined to a high level of precision. This is in fact desirable, because in the inference machinery, we use the empirical data of Fig S11 (comparison to flanking sequence data) to further infer the length constraint parameter, rather than being extremely precise about the prior. The result from Fig S10 shows the power of using this information: If we give the length constraint parameter the default prior, the resulting sequence mutation rate distribution is not different from the green spike prior, and this is because the data of Fig S11 strongly constrains the true value of this parameter, which falls within the prior distribution. On the other hand, if we actually forced an unreasonable prior, such as the red or blue spikes, the data of Fig S11 could not influence the length constraint in any way, and since this is such an important parameter in our model, the resulting inferences are inaccurate.



Nature Genetics: doi:10.1038/ng.2398

# Chapter 7: Constraints on sequence mutation rate from calibration to the fossil record

## (i) Overview

We were interested in obtaining constraints on the sequence substitution rate based on calibration to the fossil record, to which we could compare our absolute estimate based on direct measurement of the mutation rate at microsatellites.

## (ii) Assumptions

For the analyses in this note, we make a number of simplifying assumptions:

• $d_{HC}$, the divergence per base pair between human and chimpanzee, is 0.0130. This number is derived from the Enredo-Pecan-Ortheus (EPO) 6-way primate whole genome alignments[6].

• $d_{HO}/d_{HC}$ the divergence per base pair between human and orangutan divided by that between human and chimpanzee at aligned bases is 2.65, as argued in the main text.

• $\tau_{HC}$, human-chimpanzee speciation time, is >4.2 Mya, based on the date of the *Australopithecus amanensis* fossil which is believed to be on the hominin lineage since the split from chimpanzee[7].

• $\tau_{HC}/t_{HC}$, the ratio of human-chimpanzee time of last gene flow to human-chimpanzee average autosomal divergence time, is <0.73. This bound (also discussed in the text) is based on human-chimpanzee genetic divergence near genes on chromosome X, close to sites where humans and chimpanzees share an allele not seen in gorilla, orangutan and macaque. Here, the ratio $\tau_{HC}/t_{HC}$ is 0.73. Thus, the time of most recent gene flow between humans and chimpanzees is <0.73.

• $t_{HO}$, human-orangutan genetic divergence time is <23 Mya. This is based on a view that the *Proconsul* fossil places an upper bound on human-orangutan speciation time of $\tau_{HO} < 18\ Mya$[8,9]. We assume that $t_{HO} - \tau_{HO} < 5\ Mya$, that is, the human-orangutan average autosomal genetic divergence time is at most 5 Mya older than human-orangutan speciation time.

• The mutation rate per year has been constant since human-orangutan genetic divergence. (For the upper bound on the mutation rate, we only require the assumption that it has been constant since human-chimpanzee genetic divergence).

• The present-day human generation time has a lower bound 25.6 years per generation and an upper bound of 32.4 years per generation. This range is derived from our prior distribution of present-day generation time of 29 ± 2.04 from Supplementary Table 7, and using the 90% confidence interval.

## (iii) Upper bound on mutation rate: $<3.7\times10^{-8}$ /bp/gen. from *Australopithecus anamensis*



Nature Genetics: doi:10.1038/ng.2398

$\tau_{HC}$>4.2 Mya (since *Australopithecus anamensis* is a hominin)

$\Rightarrow$ $t_{HC}$ > 5.8 Mya (since $\tau_{HC}/t_{HC}$ <0.73)

$\Rightarrow$ $\mu_{year}^{seq} < 1.1 \times 10^{-9}$ (since $\mu_{year}^{seq} = d_{HC}/2t_{HC}$ = 0.0130/(2×5.8×$10^6$)

$\Rightarrow$ $\mu_{generation}^{seq} < 3.7 \times 10^{-8}$ (since $\mu_{generation}^{seq} < 32.3\mu_{year}^{seq}$ )

**(iv) Lower bound on mutation rate: >1.9×$10^{-8}$ /bp/generation from *Proconsul***

$\tau_{HO}$<18 Mya (from *Proconsul*)

$\Rightarrow$ $t_{HO}$ < 23 Mya (since we assume that $t_{HC} < \tau_{HC}$ + 5 Mya)

$\Rightarrow$ $\mu_{year}^{seq} > 7.5 \times 10^{-10}$ (since $\mu_{year}^{seq} = d_{HC}(\frac{d_{HO}}{d_{HC}})/2t_{HO}$ = 0.0130(2.65)/(2×23×$10^6$))

$\Rightarrow$ $\mu_{generation}^{seq} > 1.9 \times 10^{-8}$ (since $\mu_{generation}^{seq} > 25.6\mu_{year}^{seq}$ )

The most likely way that this lower bound could be in error would be if the mutation rate were not constant over time since human-orangutan genetic divergence. For example, if the mutation rate slowed down on the African great ape lineage (and perhaps also on the orangutan lineage) since the two diverged—perhaps associated with the increase in their body size as documented in the fossil record—the lower bound would be substantially less.

**(v) Upper bound on human-chimpanzee speciation date from fossil record <6.3 Mya**

For comparison to the upper bound on human-speciation obtained by direct calibration to the microsatellite-based molecular clock, we also use the fossil record of human-orangutan divergence to produce a complementary bound based on the fossil record. As in (iv), we write:

$\tau_{HO}$<18 Mya (from *Proconsul*)

$\Rightarrow$ $t_{HO}$ < 23 Mya (since we assume that $t_{HO} < \tau_{HO}$+ 5 Mya)

$\Rightarrow$ $t_{HC}$ < 8.7 Mya (since $t_{HC} = t_{HO}/(\frac{d_{HO}}{d_{HC}})$ = (23 Mya)/2.65)

$\tau_{HC}$ < 6.3 Mya (since $\tau_{HC} = t_{HC}(\tau_{HC}/t_{HC})$, and $\tau_{HC}/t_{HC} << 0.73$, see Supplementary Note Chapter 8)

As in (iv), the most plausible way that this lower bound could be in error would be if the mutation rate were not constant over time since human-orangutan genetic divergence.



Nature Genetics: doi:10.1038/ng.2398

# Chapter 8: Constraints on human-chimpanzee speciation date

### (i) Motivation for estimating the ratio of human-chimpanzee speciation to divergence

Our calibration of the molecular clock allows us to estimate the genetic divergence time of humans and chimpanzees $\bar{t}_{HC}$, averaged across the autosomes. However, the speciation date $\tau_{HC}$—defined in this study as the date of last gene flow between the ancestors of humans and chimpanzees—is also of biological interest. To infer $\tau_{HC}$, we require a Bayesian prior distribution on the ratio of these two quantities: $\tau_{HC}/\bar{t}_{HC}$. This is the most difficult of our prior distributions to formulate, and the following note describes how we construct our distribution based on obtaining a number of point estimates of the ratio, as well as conservative upper bounds.

### (ii) A point estimate of $\tau_{HC}/\overline{t}_{HC}$ = 0.61 from modeling of a simple demographic history

Burgess and Yang 2008

For a best estimate of the ratio $\tau_{HC}/\overline{t}_{HC}$, we use the results from Burgess and Yang 2008, who analyzed a data set of 7.4 Mb of aligned sequence from human, chimpanzee, gorilla, orangutan and macaque across "neutral" autosomal loci using the MCMCcoal software[10]. This software analyzes the 5-species alignment data under the simplifying assumptions that:

(i) The phylogeny is ((((human, chimpanzee),gorilla),orangutan),macaque)
(ii) The speciation events were instantaneous.
(iii) The populations in the intervening periods were constant in size and panmictic.
(iv) All the analyzed loci are unlinked, neutral and free of recombination

Under these assumptions, MCMCcoal estimates the ancestral population sizes and speciation times, conditional on the observed divergent site pattern. On page 7 of Burgess and Yang 2008, the authors estimate that $1 - \tau_{HC}/\overline{t}_{HC}$ = 0.39 (thus, $\tau_{HC}/\overline{t}_{HC}$ = 0.61) under a model of no gene flow after initial speciation.

Dutheil et al. 2009

Dutheil et al. 2009 made inferences under the same demographic assumptions, but using a different approach based on a coalescent Hidden Markov Model (CoalHMM) that also exploits information from recombination between adjacent loci[11]. We inferred $\tau_{HC}/\overline{t}_{HC}$ for the four autosomal loci ("targets") that Dutheil et al. analyzed, using their "bias-corrected" estimates of demographic parameters in their Table 2. After translating the quantities to estimates of $\tau_{HC}/\overline{t}_{HC}$ , we obtained results in the range of Burgess and Yang 2008: 0.67 (Target 1), 0.57 (Target 106), 0.60 (Target 121) and 0.66 (Target 122). We use the Burgess and Yang 2008 estimate of $\tau_{HC}/\overline{t}_{HC}$ = 0.61 for our primary calculations because it is based on more data and because it falls within the range of the Dutheil et al. estimates.

### (iii) Conservative upper bound on the ratio: $\tau_{HC}/\overline{t}_{HC}$ < 0.73

Analyzing subsets of the genome to obtain a conservative upper bound on $\tau_{HC}/\overline{t}_{HC}$

The published studies infer demographic parameters for human-chimpanzee speciation under a simplified model that assumes constant population size, sudden speciation, and no impact of natural selection on the genome. However, the truth likely differs from this model, as Yang



Nature Genetics: doi:10.1038/ng.2398

found in 2010 when he carried out a formal test of the fit of the data from Burgess and Yang 2008 to the model assumed in that study[12]. Thus, while the simplified models provide a useful initial estimate, deviations from the assumptions might mean that the time of last gene flow between humans and chimpanzee was more ancient or more recent.

To obtain a conservative upper bound on the ratio $\tau_{HC}/\overline{t}_{HC}$, we take advantage of an idea of Patterson et al. 2006[8]. The idea is to compute human-chimpanzee genetic divergence (dividing by human-macaque divergence to correct for variation in the local mutation rate across the genome) in subsets of the genome where the genetic divergence is expected to be less than the genome-wide average for population genetic reasons. Human-chimpanzee genetic divergence at all loci in the genome must be older than the speciation time (by definition, if we define speciation as the time of last gene flow). Thus, the ratio of the local divergence at any subset of the genome to the genome-wide average provides an upper bound on the speciation date $\tau_{HC}$.

A new 5-way alignment of human-chimpanzee-gorilla-orangutan-macaque (HCGOM)

*Overview of a 100x larger dataset generated for studying human-chimpanzee-gorilla speciation*
Patterson et al. 2006 analyzed datasets consisting of about 9 Mb of aligned DNA rom human, chimpanzee, gorilla, orangutan and macaque[8]. Here we describe how we generated a similar dataset with about 100x more data. In brief, we restricted to data generated using traditional Sanger long-read sequencing data from five genomes, and used an alignment and filtering procedure described in Mallick et al. 2009[13] (the detailed filters we applied are given below). In comparison to other multi-species alignments methodologies (e.g. EPO[6]), which have as a goal the maximization of the number of covered nucleotides, our alignment procedure filters out a larger fraction of the data, since for the purpose of making inferences about population history, we do not mind losing data as long as what is left is of high reliability. These filters resulted in 849.6 Mb of 5-species genomic alignment on the autosomes (48.58 million bi-allelic divergent sites passing filters), and 32.6 Mb on chromosome X (1.62 million bi-allelic divergent sites passing filters). These datasets are available on request from the authors.

*Genome assemblies used as input*
The raw data consisted of 5 whole genome assemblies based on Sanger long-read sequencing data. These consisted of the human genome reference sequence (*hg18*), and four assisted assemblies that we built ourselves so as to have full control over the data: chimpanzee (7.3× coverage), orangutan (6.2× coverage), macaque (6.3× coverage) and gorilla (1.8× coverage). Since we assembled the genomes ourselves, we had a sequence quality score at each nucleotide that did not automatically assign low quality to bases overlapping at within-species single nucleotide polymorphisms (SNPs), which is a feature of some genome assemblies that makes it difficult to carry out population genetic analyses.

*Generating local alignments*
We applied a stringent local alignment procedure that took advantage of the long range synteny information available from the genome assemblies[13], and then applied the following filters:

- Restrict to loci that have alignments of all 5 species over at least 100 bp
- Restrict to loci for which a unique consensus sequence is available from all 5 species

*Identifying divergent sites for analysis*
We identified sites that were divergent across the species after applying the following filters:



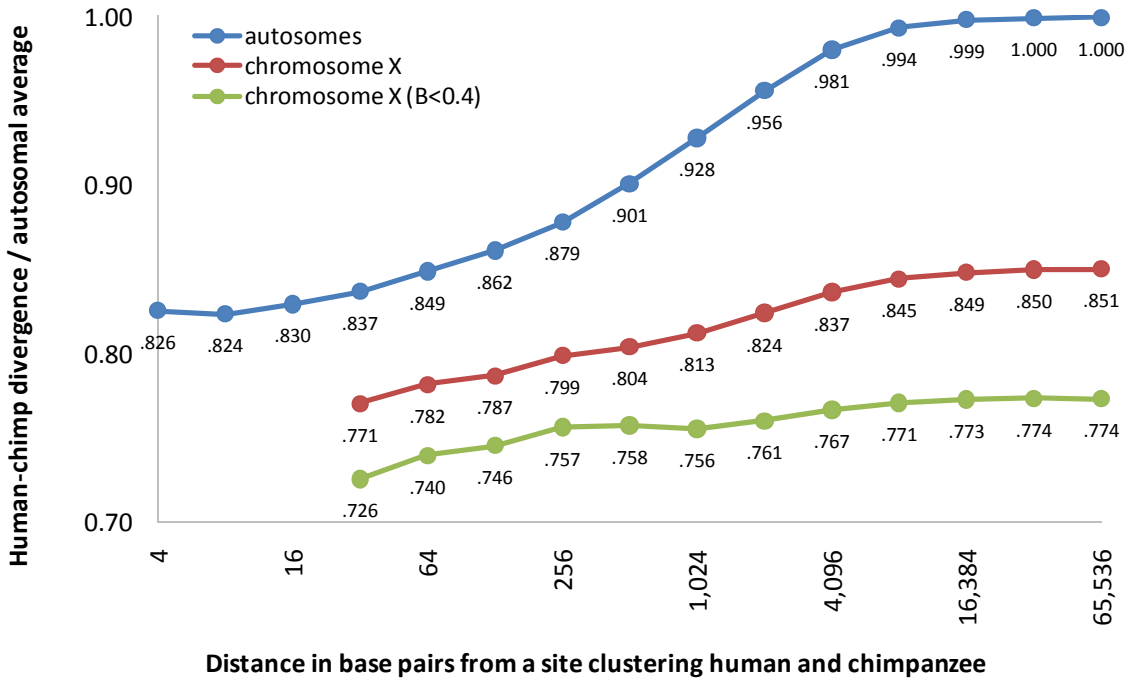
Nature Genetics: doi:10.1038/ng.2398

- Filter out sites with 3 or more alleles across species
- Filter out sites where any species has a Phred sequence quality score of <30
- Filter out sites where any species has a Phred score of <15 within 5 bp on either side.
- Filter out sites within 1 bp of an insertion/deletion in any of the species.
- Filter out sites within 5 bp of the end of an alignment
- Filter out sites within 1 bp of any other divergent site, as these sites have consistently different properties indicating that they are determined less reliably
- Filter out divergent sites that could potentially reflect a C→T mutation in the first base of a hyper-mutable CpG dinucleotide on either DNA strand (these are subject to high rates of recurrent mutation, which could complicate tests of relative divergence time).

*Post-processing to remove potential misalignments*

We filtered out entire alignments where the pattern of divergent sites showed evidence of an extreme excess on a single lineage compared with genome-wide pattern, which could reflect erroneous alignment due to low copy number repeats (paralogs). For 7 species pairs—Human-chimpanzee, Human-gorilla, Chimp-gorilla, Human-orang, Chimp-orang, Orang-macaque—we counted the number of divergent sites reflecting changes on one lineage or the other, using the other species to polarize. We compared the ratio of sites on the tested lineage to the average genome-wide (performing the analysis separately for chromosome X and the autosomes), and removed alignments with P <0.001 by a chi-square test for any of the seven comparisons

**Figure S8.1: Bounds on human-chimp speciation based on proximity to sites clustering humans and chimps.** (Blue curve) We stratify the autosomal data based on the distance to the closest site clustering humans and chimps to the exclusion of gorilla. Within 4bp, the divergence is 0.826 of the autosomal average. (Red curve) Repeating the same computation on chromosome X, the average divergence as a fraction of the autosomes is 0.851, and within 32 bp of a human-chimp clustering site is 0.771. (Green curve) We again present data for the X chromosome, but now restrict to the quarter of the data with B-statistic <0.4 reflecting an expectation of further reduced divergence due to directional selection in the ancestral population. The average X chromosome divergence in this subset of the data is 0.774, and within 32 bp of human-chimp clustering sites, it is 0.726.






Nature Genetics: doi:10.1038/ng.2398

Bound B: Genetic divergence on chromosome X divided by the autosomes ($\tau_{HC}/\overline{t}_{HC} < 0.851$)
The second upper bound on ratio of human-chimpanzee speciation time also exploits a strategy first described in Patterson et al. 2006, and is based on dividing the human-chimpanzee genetic divergence as a fraction of human-macaque on chromosome X by that on the autosomes. The motivation is that there is an *a priori* reason to expect that genetic divergence on chromosome X will be lower than on the autosomes. In a constant-sized, freely mixing population, there are 3 copies of chromosome X for every 4 copies of the autosomes, leading to a lower predicted coalescence time at X chromosome loci in the common ancestral population of humans and chimpanzees. In addition, selection operates differently on chromosome X and the autosomes (because of the exposure of recessive alleles in males), further motivating a search to explore whether the genetic divergence is unusually low.

In our new dataset, we computed the ratio of human-chimpanzee to human-macaque divergence on chromosome X divided by that on the autosomes, filtering out the pseudo-autosomal regions of chromosome X (<2.710 Mb and >154.585 Mb). After applying the correction for recurrent mutation (nearly identical results are obtained without the correction), we obtained an upper bound of $\tau_{HC}/t_{HC} < 0.851$. This is one standard error from the estimate of $\tau_{HC}/t_{HC} < 0.835 \pm 0.016$ from Patterson et al. 2006, and so the two inferences are statistically consistent.

Bound C: Chromosome X loci close to sites clustering humans and chimps ($\tau_{HC}/\overline{t}_{HC} < 0.771$)
We combined the two ideas from Patterson et al. 2006 (bounds A and B) to obtain an even more stringent upper bound. Using our 32.6 Mb of X chromosome alignment, we computed the ratio of human-chimpanzee to human-macaque divergence close to sites that cluster humans and chimpanzees to the exclusion of gorilla. Figure S8.1 (blue curve) shows that just as on the autosomes, the closer one is to a human-chimpanzee clustering site, the lower the normalized human-chimpanzee divergence. We compute the human-chimpanzee divergence divided by human-macaque divergence in the vicinity of these sites, and divide by the autosomal average after correction for recurrent mutation, resulting in a bound of $\tau_{HC}/t_{HC} < 0.771$ based on data from <32 bp away from informative sites. (We focus on the <32 bp distance because of noisy estimates in lower bin sizes, although the estimates are qualitatively consistent for smaller bin sizes as well: 0.773 (<16 bp), 0.752 (<8 bp) and 0.725 (<4 bp).)

Bound D: Chromosome X loci subject to directional selection close to HC sites ($\tau_{HC}/\overline{t}_{HC} < 0.726$)
We next studied genetic divergence between humans and chimpanzees at a subset of the genome that was not exploited in Patterson et al. 2006: loci that are at increased likelihood of having been subject to directional selection in the ancestral population of humans and chimpanzees (due to hitchhiking and selection at linked sites), thus reducing the average genetic divergence between the two species. McVicker et al. 2009 showed that loci that are close to exons or conserved non-coding sequences have a reduced genetic divergence between humans and chimpanzees compared with the average in the genome, which is likely to reflect directional selection in the ancestral population (either positive selective sweeps or negative background selection)[2]. For each nucleotide, they also computed a quantity, B, which predicts the genetic divergence without using any information from genetic variation and comparative genomics at all, and only using its proximity to functional elements. We confirmed that the B statistic is strongly predictive of divergence in our data by stratifying human-chimpanzee genetic divergence along chromosome X by the B-statistic (Figure S8.2). Figure S8.2 shows long regions of low divergence on



Nature Genetics: doi:10.1038/ng.2398

chromosome X where B is low (and which further bound the human-chimpanzee speciation time), interspersed with regions of high divergence where B is high. The pattern in this plot can only be explained by strong directional natural selection in the ancestral population of humans and chimpanzees prior to human-chimpanzee speciation. The cause remains a mystery. Possibilities include an increased rate of background selection in the ancestral population of humans and chimpanzee, an increased rate of positive selection, or selection to remove Dobzhansky-Muller incompatibilities following hybridization[8]. Determining which factors are responsible is outside the scope of this note.

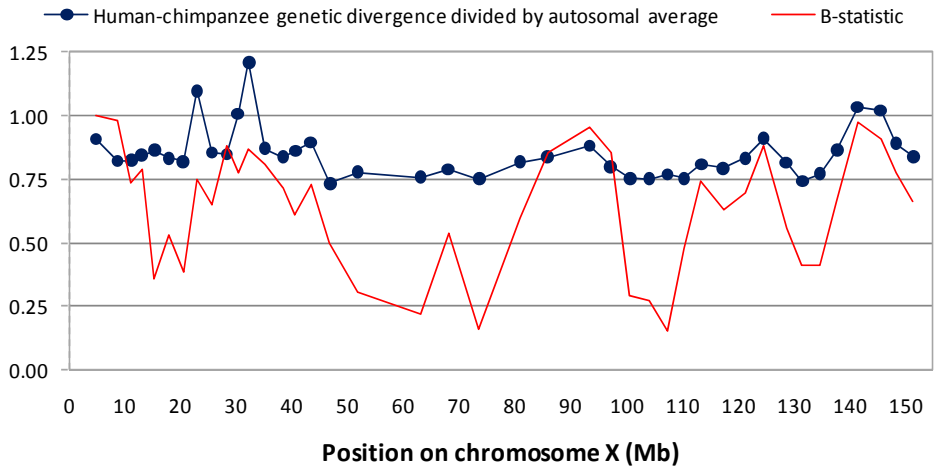



**Figure S8.2: B-statistic predicts chromosome X divergence.** We analyzed 41 equally sized bins of 40,000 sites excluding pseudoautosomal regions, and plotted human-chimp divergence as a fraction of human-macaque genetic divergence. This strongly correlates to the B-statistic, and there are large regions (e.g. 46.6-86.7 Mb, and 95.6-136.1 Mb) with low average B that also have low average divergence.

To take advantage of the correlation of divergence with selection to set a new constraint on the date of human-chimpanzee speciation, we stratified human-chimpanzee genetic divergence along chromosome X into ten approximately equal-sized bins based on the B-statistic, performing the analysis separately for chromosome X and the autosomes. Figure S8.3 shows that the bin with the smallest B-statistic on the X chromosome gives a new upper bound on $\tau_{HC}/\overline{t}_{HC}$<0.82, even without using the additional information from proximity to human-chimpanzee clustering sites.

**Table S8.2: Summary of the bounds on human-chimpanzee genetic divergence**

| Bound | Description | $\tau_{HC}/\overline{t}_{HC}$ |
|---|---|---|
| A | Genetic divergence near sites clustering humans and chimpanzees | < 0.826 |
| B | Genetic divergence on chromosome X divided by the autosomes | < 0.851 |
| C | Chromosome X loci close to HC sites (A+B) | < 0.771 |
| D | X loci close to HC sites and B<0.4 | < 0.726 |

Motivated by the power of the B-statistics to predict human-chimpanzee genetic divergence, we combined all three ideas for finding segments of the genome with reduced divergence to produce an even more stringent (but still conservative) upper bound on human chimpanzee speciation compared with any of the approaches by themselves: (i) Restriction to chromosome X, (ii) Restriction to loci strongly affected by directional selection (B<0.4, where the genetic divergence in Figure S8.3B appears to asymptote), and (iii) Restriction to sites that are within 32 bp of a divergent site that clusters human and chimpanzee to the exclusion of gorilla. From this subset of the data, we obtain a new upper bound of $\tau_{HC}/\overline{t}_{HC}$ < 0.726 (green curve in Figure S8.1). For completeness the numbers for the even lower bin sizes are: 0.742 (<16 bp), 0.730 (<8 bp) and



Nature Genetics: doi:10.1038/ng.2398

0.671 (<4 bp).) Table S8.2 lists the various bounds. In what follows and the main text, we use the strongest (D), conservatively rounding it off to $\tau_{HC}/\overline{t}_{HC} < 0.73$.

The upper bound of $\tau_{HC}/\overline{t}_{HC} < 0.73$ is conservative and robust

We conclude this section by noting that the true value of the ratio is likely to be less than 0.73.

*(a) Upper bounds using X chromosome data are conservative:* Our upper bound on human-chimpanzee speciation based on data from the X chromosome is conservative. The reason is that we are dividing by human-macaque divergence to normalize for differences in the mutation rate across loci in the genome, assuming that the average time since the most recent common ancestor (TMRCA) between humans and macaques is identical across the genome. In fact, the TMRCA varies, and is expected to be less on chromosome X than on the autosomes, since in the ancestral population of humans and macaques, the ancestral effective population size is expected to have been less on chromosome X than the autosomes (3/4). As discussed in Patterson et al. 2006, the true TMRCA could plausibly be 0-5% lower on average on chromosome X due to this effect, which will result in an overestimate of our upper bound by the same amount[8].

*(b) Upper bounds using X data are not strongly affected by changes in male-to-female mutation rate.* In 2009, Presgraves and Yi suggested that the finding of Patterson et al. 2006 of a greatly reduced genetic divergence time on chromosome X relative to the autosomes might be an artifact of changing male-to-female mutation rates among great apes, for example, due to an acceleration of the male mutation rate on the chimpanzee lineage due to more male competition for mates leading to larger numbers of sperm cell divisions and a higher male mutation rate[14]. To evaluate whether there is evidence that this might affect our inferences, we computed the human-chimpanzee genetic divergence as a fraction of human-macaque divergence across the X chromosome, after separating the data by mutations on the human lineage and chimpanzee lineage since divergence. The inference on the human-specific lineage is $\tau_{HC}/\overline{t}_{HC} < 0.850$, and on the chimpanzee-specific lineage is $\tau_{HC}/\overline{t}_{HC} < 0.852$, suggesting that this is not a major effect.

*(c) Although $\tau_{HC}/\overline{t}_{HC} < 0.73$ is a hard bound we conservatively treat it as a soft bound.* While $\tau_{HC}/\overline{t}_{HC} < 0.73$ is in principle a hard upper bound—in the sense that we have found loci where the genetic divergence is 72.6% of the autosomal average making this a maximum on human-chimpanzee speciation time—in fact we conservatively treat it as a soft bound in the main text, where we use it as the upper 5% bound of a 90% Bayesian prior probability distribution on the ratio $\tau_{HC}/\overline{t}_{HC}$. Thus, with 5% probability, we allow for the possibility that the true ratio is larger, which means that our quoted upper bound on human-chimpanzee speciation reported in the main text is actually somewhat less stringent than it should be.

### (iv) Point estimates of $\tau_{HC}/\overline{t}_{HC}$ = 0.61-0.68 from modeling of background selection

In this section, we obtain new point estimates of the ratio $\tau_{HC}/\overline{t}_{HC}$ that take advantage of the modeling analyses in McVicker et al. 2009[2], which account for the impact of directional selection on human-chimpanzee genetic divergence to obtain not just an upper bound, but also a best estimate of the ratio. This kind of modeling analysis is important, since as shown in Figure S8.2-S8.3, directional selection is clearly having an important impact on our data.



Nature Genetics: doi:10.1038/ng.2398

We first used the modeling of autosomal data directly reported in the McVicker et al. 2009 paper[2]. In Table 1 of their paper (page 7), they give parameter estimates under their model taking into account a fitted model of background selection on the autosomes, which translate to an estimate of $\tau_{HC}/\overline{t}_{HC}$ = 0.61, matching the estimate from Burgess and Yang.

As an additional estimate using >100 times more data than was analyzed by McVicker et al. 2009, we examined the correlation of B-statistic with genetic divergence in our own data. If the model underlying the B-statistic is correct, then the value of B (on its scale of 0-1) predicts the reduction in genetic diversity in the human-chimpanzee ancestral population at a locus, compared with the expectation if there were no selection at all. Assuming that the B-statistics are measured with perfect accuracy and the model is correct, if we measure human-chimpanzee genetic divergence as a fraction of the autosomal average in ten bins of B-statistic, and fit a line, then the y-intercept gives the expected human-chimpanzee genetic divergence at loci in the genome where the time to the common ancestor in the ancestral population was zero; that is, they give the date of human-chimpanzee speciation.

**Figure S8.3: Human-chimpanzee divergence divided by the autosome average, stratified by B.** We divided (A) the autosomal and (B) chromosome X data into 10 equally sized bins, based on McVicker B-statistics. Blue lines show least squares fits to all ten data points, and red lines leave out three points that contribute to non-linearity and may reflect model failure (the two points with the lowest B and the one point with the highest B). The y-intercepts provide an estimate of human-chimp speciation as a fraction of the autosomal divergence; that is, the expected genetic divergence assuming no genetic variation in the ancestors.

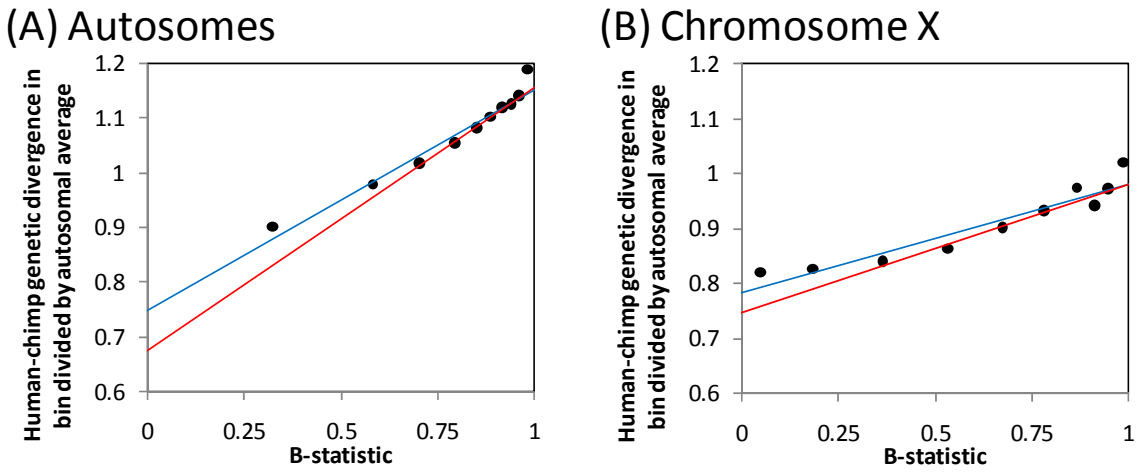



Figure S8.3 shows the empirical relationship of genetic divergence between human and chimpanzee to B-statistics on the autosomes and chromosome X separately. There is evident non-linearity, mostly in the two bins with the lowest B-statistics. A potential explanation (even if the model is correct) is "regression to the mean". The assignment of B-statistics to individual nucleotides is noisy and thus the bin of nucleotides with the lowest B-statistics is likely to contain a substantial fraction of nucleotides that are not in fact so constrained by selection as indicated by their assigned B-statistic. Thus, the observed human-chimpanzee divergence in these bins is not as reduced as predicted. We therefore fit lines not just to all ten bins, but also to a subset of seven bins that exclude the two with the lowest B-statistics, and the highest bin (which appears to be an outlier perhaps due to structural variation). In the middle seven bins, the points appear linear. The extrapolated y-intercept from the fitted (red) regression line is $\tau_{HC}/\overline{t}_{HC}$ = 0.68 on the autosomes, giving a new point estimate. (On chromosome X, it is $\tau_{HC}/\overline{t}_{HC}$ = 0.75



Nature Genetics: doi:10.1038/ng.2398

(Figure S8.3), but we focus here on the autosomes since McVicker et al. 2009 had much better autosomal data to use in their modeling analysis and obtained a much better fit of their B-statistic model to the data on the autosomes. Moreover, the best estimate of the ration on chromosome X is clearly too high, as it exceeds the upper bound of section (iii).)

### (v) Prior distribution on $\tau_{HC}/\overline{t}_{HC}$

Above, we described several inferences about the ratio of human-chimpanzee speciation to average human-chimpanzee genetic divergence:

(a) We described a point estimate of $\tau_{HC}/\overline{t}_{HC}$ (0.61) based on the modeling analyses under neutral evolution from Burgess and Yang, which is consistent with Dutheil and colleagues.
(b) We described a conservative upper bound of <0.73.
(c) We described point estimates of $\tau_{HC}/\overline{t}_{HC}$ (0.61-0.68) from modeling analyses that take into account background selection using insights from McVicker et al. 2009.

Taking these various inferences into account, we propose a prior distribution on $\tau_{HC}/\overline{t}_{HC}$ that is normally distributed, and that allows 5% of its density above 0.73 and 10% of its density below 0.61. Thus, its mean is 0.663, and its standard deviation is 0.041 (Figure S8.4). This distribution captures the observation that none of the point estimates are substantially below 0.61, and that we have a strong upper bound at 0.73 (which conservatively, we treat as a soft upper bound, although in fact it would be very surprising if the true value was higher).

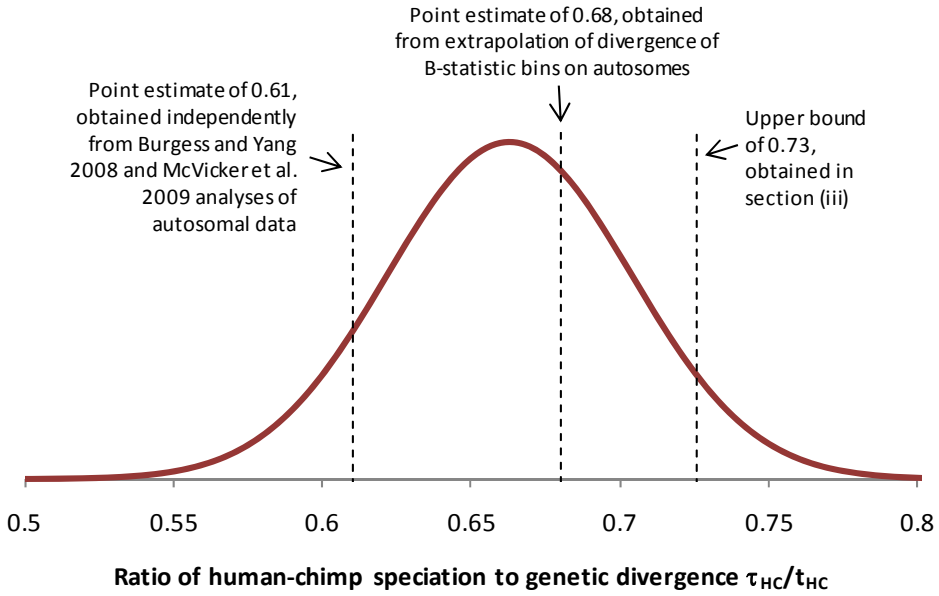



**Figure S8.4: Prior distribution on the ratio of human-chimp speciation to genetic divergence, $\tau_{HC}/\overline{t}_{HC}$.** This distribution has a mean of 0.663 and a standard deviation of 0.041, set so that 10% of the density is below 0.61 and 5% of the density is above 0.73. The inferences that we use to inform this prior are indicated by dashed lines.

We conclude by discussing what the effect on our inferences would be if the true value of the ratio was below 0.61, which is especially relevant since two of the point estimates were at this value. Lower values would reduce the posterior estimate of the human-chimpanzee speciation date, which is already lower in our paper than would be consistent with some interpretations of the fossil record. Figure 4 of the paper allows readers to ignore our prior, and instead infer the speciation date that would be obtained for any choice of $\tau_{HC}/\overline{t}_{HC}$. This analysis shows that



Nature Genetics: doi:10.1038/ng.2398

speciation dates above 6.8 Mya (the current minimum date of the *Sahelanthropus* fossil) require a ratio of $\tau_{HC}/\overline{t}_{HC}$ >0.70.



Nature Genetics: doi:10.1038/ng.2398

# Chapter 9: Hierarchical Bayes Model

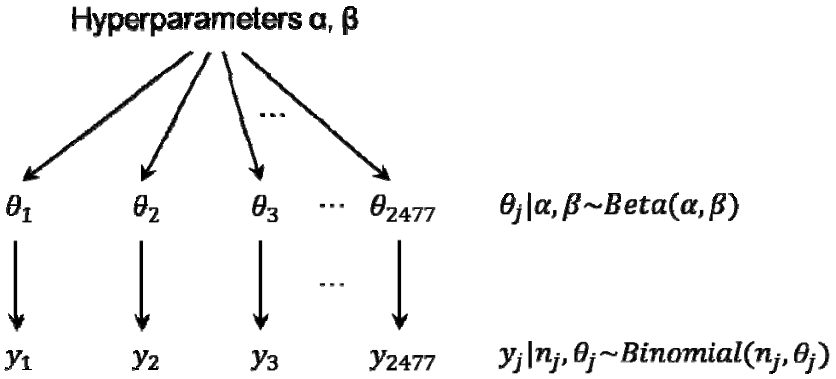
Because of inter-locus variation in mutation rate, quantities such as the standard error of the mutation rate, pooled across loci, become non-trivial to estimate. To infer this quantity, we model the data using a Hierarchical Bayes Model (HBM).

The framework of the HBM is as follows: (1) Describe the data generating process using a set of equations, that is, the method to generate data (mutation events) given the parameters. (2) Derive the posterior distribution, which is conditioned upon the data. (3) Using the set of posterior equations with the empirical data as input, sample the posterior distribution using direct-sampling or MCMC techniques. (4) Perform extensive model-checking to ensure that the HBM performs appropriately.

## Hierarchical model of the mutation process

1. Data generative process

   For loci , the numbers of mutations are modeled as independent binomial samples: , where is the number of observations and assumed to be known. is the mutation rate. We use a conjugate distribution with hyperparameters that are the same for all .




2. The joint posterior density is as follows:

$$p(\theta, \alpha, \beta|y) = \frac{p(\theta, \alpha, \beta)p(y|\theta, \alpha, \beta)}{p(y)} \quad (1)$$

$$\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta) \quad (2)$$

$$= p(\alpha, \beta)\prod_j p(\theta_j|\alpha, \beta)p(y_j|\theta_j) \quad (3)$$

$$= p(\alpha, \beta)\prod_j Beta(\alpha, \beta)Bin(n_j, \theta_j) \quad (4)$$

$$\propto p(\alpha, \beta)\prod_j \frac{1}{B(\alpha, \beta)}\theta_j^{\alpha-1+y_j}(1-\theta_j)^{\beta-1+n_j-y_j} \quad (5)$$

Line 1 is by Bayes rule.



Nature Genetics: doi:10.1038/ng.2398

Line 2 is the product of the hyper-prior distribution, the parameter distribution, and the likelihood.
Line 3 follows by conditional independence of the parameter and data.
Lines 4 and 5 follow from our data generative model. $B(\alpha, \beta)$ is the beta function.

3. In order to sample from the posterior, we first find $p(\alpha, \beta|y)$ by integrating over each $\theta_j$ from 0 to 1, obtaining:

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_j \frac{B(\alpha + y_j, \beta + n_j - y_j)}{B(\alpha, \beta)}$$

4. A suitable hyper-prior distribution $p(\alpha, \beta)$: We would like to choose a diffuse prior. However, an improper prior such as $p(\alpha, \beta) = 1$ doesn't work because $p(\alpha, \beta|y)$ cannot integrate to 1. This is because

$$\lim_{\alpha,\beta\to\infty} \frac{B(\alpha + y_j, \beta + n_j - y_j)}{B(\alpha, \beta)} = 1$$

Instead, we choose a diffuse (uniform) density on $(\frac{\alpha}{\alpha+\beta}, (\alpha + \beta)^{-\frac{1}{2}})$, which are the mean and approximately proportional to the standard deviation of $\theta_j|\alpha, \beta \sim Beta(\alpha, \beta)$. From equation 5.9 of Gelman et al[15], this leads to $p(\alpha, \beta) \propto (\alpha + \beta)^{-\frac{5}{2}}$. Hence,

$$p(\alpha, \beta|y) \propto (\alpha + \beta)^{-5/2} \prod_j \frac{B(\alpha + y_j, \beta + n_j - y_j)}{B(\alpha, \beta)}$$

**Drawing simulations from the posterior distributions**

1. The first step is to crudely estimate the parameters $\theta, \alpha, \beta$. From the data, we find $mean\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-4}$ and $var\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-6}$, obtaining estimates of $(\theta, \alpha, \beta) = (5 \times 10^{-4}, 0.05, 99)$.

2. Next, we look for the posterior mode of $p(\alpha, \beta|y)$. When calculating values of the posterior, to avoid numerical issues, we compute the log posterior, then exponentiate at the end. We can use the EM algorithm to find the mode, using our crude estimates as a starting point. Alternatively, for this 2 dimensional problem, we can simply use a grid of $(\alpha, \beta)$ to look for $\max_{\alpha,\beta} p(\alpha, \beta|y)$ in the vicinity of the crude estimates. We find that the posterior mode is located at $(\alpha, \beta) = (0.68, 1480)$. At the mode, this would correspond to $E[\theta|\alpha, \beta] = 4.6 \times 10^{-4}$ and $var[\theta|\alpha, \beta] = 3 \times 10^{-7}$. Our variance here is about 10 times smaller than that of our crude estimates. This is because $var\left(\frac{y_j}{n_j}\right) = 5 \times 10^{-6}$ was estimating $var(\theta)$, taking into account variability in $(\alpha, \beta)$.

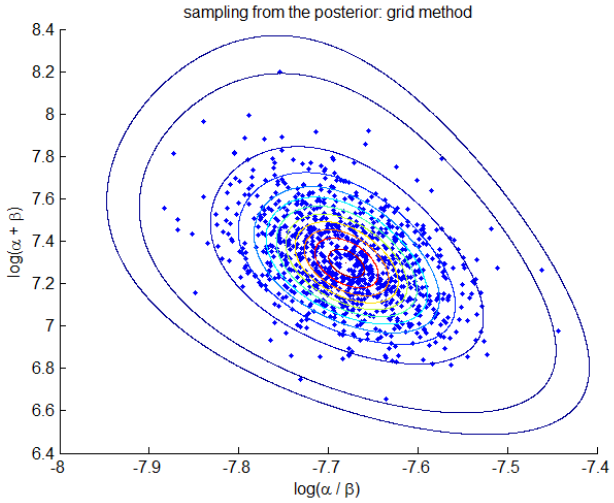
   Below is a contour plot of $p(\alpha, \beta|y)$, re-parameterized in terms of $\left(\log\frac{\alpha}{\beta}, \log \alpha + \beta\right)$, with contours at 0.0001, 0.001, and at 0.05, 0.15, 0.25, …, 0.95 of the modal value.

3. Given our sense of how $p(\alpha, \beta|y)$ behaves, we now sample from the posterior. We directly sample via grids. This method is feasible because we are sampling only in 2 dimensions. Using the contour plot above, we compute the grid of points where most of the density lies. Then, we numerically sum one dimension to obtain the marginal distribution, say $p(\alpha|y)$. $\alpha$ is then



Nature Genetics: doi:10.1038/ng.2398

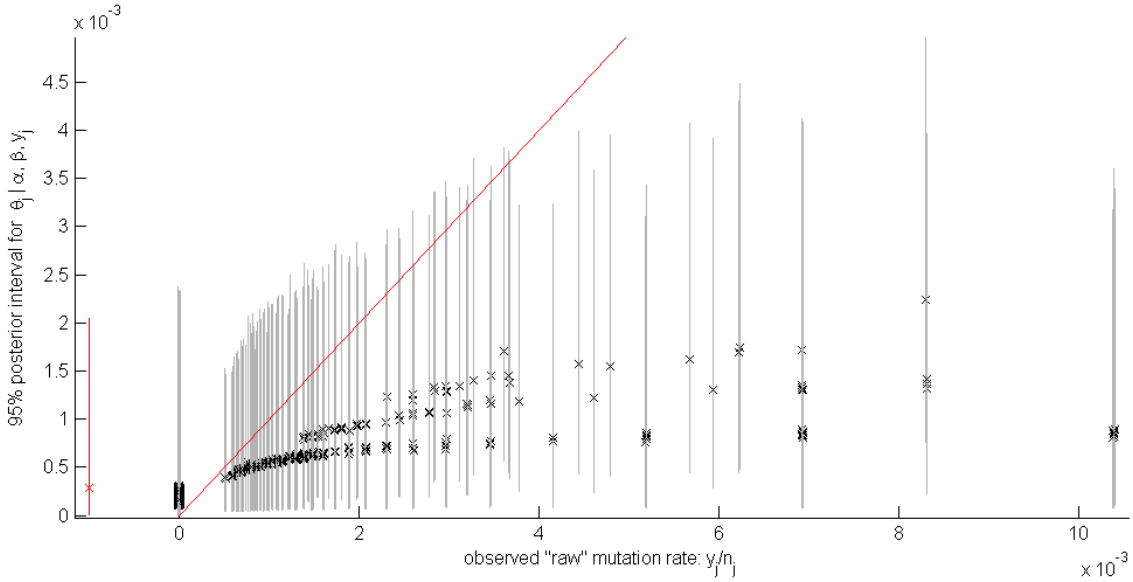
sampled using the inverse-CDF method. Then we sample $\beta$ using the inverse-CDF method again, this time on $p(\beta|\alpha, y)$. 1000 samples of $(\alpha, \beta)$ are shown below.




4. After sampling from $p(\alpha, \beta|y)$, we sample $\theta$ using $p(\theta|\alpha, \beta, y)$. Note that the posterior for $\theta$ is beta distributed, and has parameters that combine the data and the hyper-parameters:

$$p(\theta_j|\alpha, \beta, y_j) = Beta(\alpha + y_j, \beta + n_j - y_j)$$

With the hierarchical framework, for each sample of $(\alpha, \beta)$, we sample the entire set of 2,477 $\theta_j$. This is one experiment. Since we have 1,000 samples of $(\alpha, \beta)$, we run 1,000 experiments and obtain a confidence bound for each $\theta_j$. The plot below shows our posterior for $\theta_j$. The horizontal axis gives the 2,477 mutation rates, taken as the raw ratio of mutant to observed events. The vertical axis gives the posterior. Crosses "x" are the median. Gray vertical bars show the 95% posterior confidence interval. The y=x line is in red. The red vertical line on the left shows the median and confidence interval of a locus that has $n_j = 0$, an uninformative locus. Note that the slope of a regression line through the crosses would be substantially less than 1. This is the effect of "smoothing" the raw mutation rates, using the combined information from all loci.






Nature Genetics: doi:10.1038/ng.2398

# Chapter 10: Inferences based on direct estimates of the sequence mutation rate (this section is added as a note in proof)

## Introduction

After this manuscript was accepted, a paper by Kong et al. reported a direct estimate of the sequence substitution rate based on whole-genome sequencing of 79 trios[16].

The two studies are concordant in inferring that the male mutation rate is 3-4 times higher than the female mutation rate, and that male mutation rate increases rapidly with age while female mutation rate does not. However, there are a couple of differences that affect inferences of dates in evolutionary history.

The first difference is that the dependence of mutation rate on paternal age in Kong et al. 2012 is stronger than we estimate for microsatellites. Mutation rate is estimated to double every 16.5 years for sequence data, compared with every 38 years for microsatellites. We hypothesize that this reflects the different mutation processes for sequence substitutions and microsatellites.

The second difference is that the direct estimate of the sequence substitution rate of $1.20\times10^{-8}$/bp/generation in Kong et al. 2012 is outside the 90% credible interval of $1.40\text{-}2.28\times10^{-8}$/bp/generation inferred here based on modeling of the microsatellite mutation process and extrapolation of the sequence mutation rate. Part of the discrepancy is due to different assumptions about present-day generation intervals: recalibrating the mutation rate estimates from Figure 2 of Kong et al. 2012 to the male and female generation intervals assumed for this study, we obtain a slightly higher estimate of $1.26\times10^{-8}$/bp/generation. However, even with this correction the Kong et al. 2012 estimates are less than ours.

## Comparison of dates inferred from the two independent estimates of mutation rate

To explore the effect of the direct estimates of the sequence mutation process on our inferences of evolutionary parameters, we used the fitted dependence on age in Kong et al.[16] (blue dashed curve in their Fig. 2). Using the notation of Supplementary Note Chapter 5, the per-generation mutation rate is:

$$\mu_{maternal}(t) = 14.2/2.63 \times 10^{9}$$
$$\mu_{paternal}(t) = \exp(2.61 + 0.042 \cdot t)/2.63 \times 10^{9}$$

We assume here that there is no error in these fitted parameters.

To understand the implication for dates in human evolution, we used a Bayesian procedure similar to that of Supplementary Note, Chapter 5 to integrate these sequence mutation rate estimates with ten prior distributions on evolutionary parameters that we developed for the microsatellite modeling and which are summarized in Supplementary Table 7. These correspond to: (1) ancestral generation time; (2) present-day generation time; (3) ancestral male-female parental age difference; (4) present-day male-female parental age difference; (5) ancestral-to-present-day transition time; (6) Western European heterozygosity per base pair; (7) West African heterozygosity per base pair; (8) Ratio of human-chimpanzee to Western European sequence divergence; (9) Ratio of human-chimpanzee speciation time to genetic divergence time; and (10) Ratio of human-orangutan to human-chimpanzee sequence divergence

## Results

Table S10.1 shows that based on the mutation rates inferred from Kong et al. 2012, the average time since the most recent common ancestor of two Western Europeans is 880-1,100 thousand years ago, the inferred time since human-chimpanzee divergence is13.0-17.2 million years ago (Mya), the inferred time



Nature Genetics: doi:10.1038/ng.2398

since human-orangutan divergence is 34.0-46.2 Mya, and the inferred date of human-chimpanzee speciation is 8.32-11.8 Mya. The dates implied by Kong et al. 2012 are in some cases more than twice those inferred from the microsatellite data, even though the present-day generation mutation rate is only ~1.5-fold higher. This is due to the stronger dependence of mutation rate on generation interval for the sequence- than for the microsatellite-based rate estimates.

We discuss two implications of these results.

Implication for the human-chimpanzee speciation date: A first implication is that the inferred human-chimpanzee speciation date of 8.32-11.8 Mya is greater than the 6.8-7.2 Mya estimate for *Sahelanthropus tschadensis*, a fossil that has been interpreted as being on the hominin lineage since the split from chimpanzees, and is excluded by the dates that emerge from the microsatellite analysis. If we accept the sequence-based estimates of mutation rate, *S. tschadensis* is no longer in tension with the genetic data.

Implication for human-orangutan genetic divergence: While using the sequence-based estimates of mutation rates makes it possible to reconcile *S. tschadensis* with being on the hominin lineage, the new dates are in tension with the fossils relevant to human-orangutan genetic divergence. The inferred human-orangutan genetic divergence date of 34.0-46.2 Mya is so much older than the upper bound from the fossil record of <18 Mya on human-orangutan speciation that the date is implausible (we discuss these constraints further in Supplementary Note, Chapter 7). A possible reconciliation to this conundrum was suggested by Scally et al. 2012[17] who hypothesized that there might have been a slowdown of the mutation rate on the African great ape lineage and on the orangutan lineage simultaneously since their ancestors separated, perhaps associated with the known increase in body size on both lineages. This slowdown would result in an overestimate of the date of human-orangutan genetic divergence using models like those in this paper that assume a molecular clock whose rate has been constant over time. However, this scenario also requires us to hypothesize a combination of unlikely events: (a) the slowdown would need to have been coincidental in both lineages to explain the observations, and (b) the slowdown would also have to have been extraordinarily dramatic: about 3-fold in both lineages in the period ancestral to human-chimpanzee divergence to produce as extreme an effect as is observed.

**Discussion**

The differences in the dates implied by the microsatellite- and sequence-based mutation rate estimates are striking. If the rate from the microsatellite data is too high, this might be due to a higher rate of false-positives than we measured empirically or inaccuracies in the model we fit to the data. If the direct measurement of the sequence-based mutation rates is too low, this might be due to the stringent filtering that Kong et al. 2012 applied to remove false-positive sites, which could have resulted in a substantial false-negative rate. Accurate estimates of the human mutation rate are important for evolutionary studies, and an important area for future research is to determine which rates are most appropriate.

**Table S10.1 Comparison of inferred evolutionary parameters from sequence & microsatellite data**

| | Duplicated from Table 2 | | Using Kong et al.'s mutation rates[16] | |
|---|---|---|---|---|
| | Mean | $5^{th} - 95^{th}$ percentile | mean | $5^{th} - 95^{th}$ percentile |
| **Genetic divergence times (millions of years)** | | | | |
| $t_{CEU}$: Western Europeans | 0.546 | 0.426 – 0.709 | 1.01 | 0.88 – 1.10 |
| $t_{YRI}$: Yoruba (African) | 0.720 | 0.562 – 0.933 | 1.33 | 1.17 – 1.44 |
| $t_{HC}$: human-chimpanzee | 7.49 | 5.80 – 9.77 | 15.3 | 13.0 – 17.2 |
| $t_{HO}$: human-orangutan | 19.8 | 15.2 – 25.9 | 40.5 | 34.0 – 46.2 |
| $\tau_{HC}$: human-chimpanzee speciation time | 4.97 | 3.75 – 6.57 | 10.1 | 8.32 – 11.8 |

Note: 90% Bayesian credible intervals are obtained from the Bayesian posterior distribution.



Nature Genetics: doi:10.1038/ng.2398

## References


1. Keinan, A., Mullikin, J.C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**, 1251-5 (2007).
2. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471 (2009).
3. Amos, W., Flint, J. & Xu, X. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genet* **9**, 72 (2008).
4. Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R. & Stefansson, K. A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am J Hum Genet* **72**, 1370-88 (2003).
5. Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* **100**, 15324-8 (2003).
6. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-28 (2008).
7. Leakey, M.G., Feibel, C.S., McDougall, I. & Walker, A. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565-71 (1995).
8. Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103-8 (2006).
9. MacLatchy, L., Gebo, D., Kityo, R. & Pilbeam, D. Postcranial functional morphology of Morotopithecus bishopi, with implications for the evolution of modern ape locomotion. *J Hum Evol* **39**, 159-83 (2000).
10. Burgess, R. & Yang, Z. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol Biol Evol* **25**, 1979-94 (2008).
11. Dutheil, J.Y. *et al.* Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* **183**, 259-74 (2009).
12. Yang, Z. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome biology and evolution* **2**, 200-11 (2010).
13. Mallick, S., Gnerre, S., Muller, P. & Reich, D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome research* **19**, 922-33 (2009).
14. Presgraves, D.C. & Yi, S.V. Doubts about complex speciation between humans and chimpanzees. *Trends in ecology & evolution* **24**, 533-40 (2009).
15. Gelman, A., Carlin, J., Stern, H. & Rubin, D. *Bayesian Data Analysis*, (2004).
16. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).
17. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-75 (2012).




Nature Genetics: doi:10.1038/ng.2398