# Precision and Accuracy of Divergence Time Estimates from STR and SNPSTR Variation

*Uma Ramakrishnan and Joanna L. Mountain*

Department of Anthropological Sciences, Stanford University, Stanford, California

Inference of intraspecific population divergence patterns typically requires genetic data for molecular markers with relatively high mutation rates. Microsatellites, or short tandem repeat (STR) polymorphisms, have proven informative in many such investigations. These markers are characterized, however, by high levels of homoplasy and varying mutational properties, often leading to inaccurate inference of population divergence. A SNPSTR is a genetic system that consists of an STR polymorphism closely linked (typically $< 500$ bp) to one or more single-nucleotide polymorphisms (SNPs). SNPSTR systems are characterized by lower levels of homoplasy than are STR loci. Divergence time estimates based on STR variation (on the derived SNP allele background) should, therefore, be more accurate and precise. We use coalescent-based simulations in the context of several models of demographic history to compare divergence time estimates based on SNPSTR haplotype frequencies and STR allele frequencies. We demonstrate that estimates of divergence time based on STR variation on the background of a derived SNP allele are more accurate (3% to 7% bias for SNPSTR versus 11% to 20% bias for STR) and more precise than STR-based estimates, conditional on a recent SNP mutation. These results hold even for models involving complex demographic scenarios with gene flow, population expansion, and population bottlenecks. Varying the timing of the mutation event generating the SNP revealed that estimates of divergence time are sensitive to SNP age, with more recent SNPs giving more accurate and precise estimates of divergence time. However, varying both mutational properties of STR loci and SNP age demonstrated that multiple independent SNPSTR systems provide less biased estimates of divergence time. Furthermore, the combination of estimates based separately on STR and SNPSTR variation provides insight into the age of the derived SNP alleles. In light of our simulations, we interpret estimates from data for human populations.

## Introduction

The inference of within-species population history from molecular data is typically more challenging than the inference of species relationships; gene trees are inconsistent with population history more often than they are with species history (Arbogast et al. 2002). Both gene flow and stochasticity contribute to this inconsistency. An additional challenge is molecular resolution: for intraspecific inference, genetic regions with a relatively high mutation rate are most informative. An optimal strategy for intraspecific studies, therefore, is to examine a large number of independently evolving, rapidly mutating, molecular regions.

Single-nucleotide polymorphisms (SNPs) provide extensive data on population history, but few studies have attempted to estimate population parameters using these markers. The high mutation rate (e.g., Weber and Wong 1993; Ellegren 2000*a*; Huang et al. 2002) characterizing short tandem repeats (STRs, or microsatellites) (Di Rienzo et al. 1994; Tautz and Shlotterer 1994), coupled with their wide distribution in the human genome, make them effective markers for estimating divergence times between human populations. Various analytical approaches based on distance statistics (e.g., $(\delta\mu)^2$ by Goldstein et al. [1995] and $T_D$ by Zhivotovsky [2001]) and coalescent models in a Bayesian framework (BATWING [Wilson, Weale, and Balding 2003]) have been developed and applied to STR data (e.g., Jin et al. 2000; Zhivotovsky, Rosenberg, and Feldman 2003) in this context.

Although STR loci evolve rapidly and, therefore, provide needed molecular resolution (Grant and Kluge 2003), their mutation properties have disadvantages. Loci evolving according to a simple stepwise mutation model (Kimura and Ohata 1978) or a two-phase model (Di Rienzo et al. 1994) exhibit homoplasy, wherein two alleles identical by state are not identical by descent. If mutation rates are high, homoplasy obscures the gene history and leads to estimates of divergence time that are too recent. The extent and impact of such homoplasy has been evaluated using molecular characterization of STR flanking regions (e.g., Grimaldi and Crouau-Roy 1997; Makova et al. 1998; Van Oppen et al. 2000) and simulation (Estoup, Jarne, and Cornuet 2002). Simulation studies illustrated that the probability of two gene copies at the same locus being identical by state but not by descent (an index of homoplasy) was as high as 30%, especially with models including constraints on allele size. A certain fraction of this homoplasy can be estimated through further molecular analysis (molecularly accessible size homoplasy, or MASH), for example, by sequencing regions flanking the microsatellite locus (Estoup, Jarne, and Cornuet 2002).

Compound genetic marker systems that include one or more SNPs tightly linked to an STR also partially reveal STR homoplasy. The haplotypes of such systems can be divided into two categories: STR variation on the ancestral SNP allele background and STR variation on the derived SNP allele background. The haplotypes with the derived SNP share a more recent common ancestor than do the haplotypes with the ancestral SNP. The STR variation on the derived SNP background is, therefore, likely to be characterized by far less homoplasy. The younger the SNP, the less homoplasy we expect to see in the STR variation on the background of the derived allele. The nonrecombining region of the human Y chromosome (de Kniff 2000) constitutes a highly informative example and has been used to estimate divergence times between groups (e.g., Hurles et al. 1999; Kayser et al. 2003;

Key words: SNP, STR, divergence time estimation, accuracy and precision, linked markers.

E-mail: uramakri@stanford.edu.

*Mol. Biol. Evol.* 21(10):1960–1971. 2004
doi:10.1093/molbev/msh212
Advance Access publication July 14, 2004

Knight et al. 2003). The Y chromosome, however, like the mitochondrial genome, reflects only a fraction of human history and has been subject to the influence of variation in effective population size (reflecting variation in the level of polygyny in human populations as revealed in Kayser et al. [2003], among other factors) and possible selective sweeps.

In response to the need for independently and rapidly evolving genetic systems, Mountain et al. (2002) developed protocols for autosomal SNPSTR systems, consisting of one or more SNPs tightly linked to an STR marker. SNPSTR systems satisfy three requirements: (1) close physical linkage of two or more polymorphisms (typically < 500 bp), (2) significant differences in mutation rate between polymorphisms, and (3) potential for a large number of independently evolving, compound haplotypic systems.

Using two SNPSTR systems, Mountain et al. (2002) demonstrated support for the hypothesis that anatomically modern humans first migrated out of Africa relatively recently. Tishkoff and et al (1996, 2000) reached similar conclusions using analogous systems with STRs linked to ALU polymorphisms rather than SNPs. Although these initial studies provided qualitative evidence for the potential contributions of SNPSTR haplotype data to the study of human evolution, they provided no quantitative evaluation of the approach in the context of parameter estimation.

The SNP(s) of compound genetic systems such as SNPSTRs reveal homoplasy at the linked STR locus. STR variation on the background of a relatively recent SNP allelle with respect to the time of a population divergence may, therefore, provide a particularly good estimate of population divergence time. In this paper, we simulate coalescent processes to evaluate how informative these novel marker systems are for estimating the time of divergence between two populations. Three genetic distance measures are used to estimate population divergence time. We compare divergence time estimates calculated from total STR variation and from STR variation on the derived SNP allele, evaluating these estimates in terms of accuracy and precision. Additionally, we explore the impact of population divergence coupled with population expansion, bottlenecks, and gene flow (processes not incorporated in models on which estimators are based) on estimates of divergence time from SNPSTR variation and from STR variation. We then evaluate the sensitivity of these measures to the timing of the mutation event that generated the SNP as well as to the variability in mutational properties of STR loci. Finally, we calculate divergence time estimates between African and non-African populations for published SNPSTR data and interpret the results in light of the simulations. Although this study was motivated by questions regarding human evolution, results are also broadly relevant to intraspecific studies of other species (Makova et al. 1998; Hey et al. 2004).

## Materials and Methods

### Simulations

We evaluated estimates of divergence time by comparing the accuracy and precision of those estimates. Accuracy (measured as 1–the magnitude of the bias) corresponds to the proximity between the true value and the estimate. It is measured as the difference between the median of estimated divergence time (over 1,000 simulation runs for each locus) and the true divergence time. Precision is represented here by the range between the 95th and the 5th percentiles of the same distribution and is highly correlated with the variance. Error, reflecting both accuracy and precision, is calculated as follows:

$$Error = \sqrt{\frac{\sum_{n}(T_{est} - T_{true})^2}{n}} \qquad (1)$$

where $T_{true}$ is the true divergence time, $T_{est}$ is the estimated divergence time, and $n$ is the number of simulations. A high value of error implies low accuracy and/or low precision.

These statistical properties of estimates of divergence time (see *Estimating Divergence Time*) were calculated for simulated STR and simulated SNPSTR data (see *Coalescent Simulations, Ascertainment*), a range of population histories (see *Demographic Scenarios*), and a range of SNP and STR mutation models (see *Sensitivity Analyses*).

#### Coalescent Simulations

The coalescent (Hudson 1990) was used to simulate genetic variation for both STR and SNPSTR systems resulting from different demographic scenarios. We modified the publicly available coalescent simulation program, SIMCOAL (Excoffier, Novembre, and Schneider 2000), which currently models unlinked systems, to accommodate linked marker systems. The modified version is available at www.stanford.edu/group/mountainlab. Coalescent simulations assume neutral evolution, randomly mating populations, and nonoverlapping generations. For all demographic scenarios described below, populations with effective size of 10,000 diploid individuals were modeled, with 45 individuals (90 chromosomes) sampled per population. Given the close physical linkage (< 500 bp) separating the SNP and STR of existing SNPSTR systems for humans (Mountain et al. 2002), we assumed complete linkage between the SNP and the STR; the SNP and the STR mutations were modeled on the same genealogy. STR loci evolved according to a simple stepwise mutation model (mutation rate = 0.0005). All modeled STR loci (apart from the sensitivity analysis) had the same mutation rate. A survey of 377 microsatellite loci typed for more than 1,000 humans by the Marshfield Genotyping Institute revealed that the average number of alleles is approximately 10.8, with a standard deviation of 3.6 alleles. Only nine of 377 loci had a size range of more than 20 alleles. We, therefore, modeled a range constraint (the number of possible allelic states) of 20. The SNP was modeled as a biallelic system (ancestral or derived allele for each chromosome) and as a unique event in history of the two populations. Once we had generated the genealogy (via the standard SIMCOAL routine), we identified the branches of the genealogy where the derived SNP mutation might have arisen, given the specified time of mutation. One of these branches was picked at random, and all

**Table 1**
**Two-Population Divergence Models**

| Model # | Divergence Time[a] | SNP Mutation Time[a] | Gene Flow Begins[a] | Gene Flow Ends[a] | r | m | Bottleneck |
|---|---|---|---|---|---|---|---|
| 1a | 5000 | 6000 | – | – | 0 | 0 | – |
| 1b | 2000 | 2400 | – | – | 0 | 0 | – |
| 2a | 5000 | 6000 | 5000 | 0 | 0 | 0.00005 | – |
| 2b | 2000 | 2400 | 2000 | 0 | 0 | 0.00005 | – |
| 3a | 5000 | 6000 | 5000 | 4000 | 0 | 0.00005 | – |
| 3b | 2000 | 2400 | 2000 | 1600 | 0 | 0.00005 | – |
| 4a | 5000 | 6000 | 1000 | 0 | 0 | 0.00005 | – |
| 4b | 2000 | 2400 | 400 | 0 | 0 | 0.00005 | – |
| 5a | 5000 | 6000 | – | – | 0.000461[b] | 0 | – |
| 5b | 2000 | 2400 | – | – | 0.001160[b] | 0 | – |
| 6a | 5000 | 6000 | – | – | 0.000925[c] | 0 | – |
| 6b | 2000 | 2400 | – | – | 0.002320[c] | 0 | – |
| 7a | 5000 | 6000 | – | – | 0.001520[d] | 0 | 5% |
| 7b | 2000 | 2400 | – | – | 0.003800[d] | 0 | 5% |
| 8a | 2000 | 2700 | – | – | 0 | 0 | – |
| 8b | 2000 | 3000 | – | – | 0 | 0 | – |
| 8c | 2000 | 8000 | – | – | 0 | 0 | – |

[a] All times in terms of number of generations in the past; r is exponential growth rate/generation, m is migration rate/generation.

[b] After population divergence, effective population size increases to 100,000.

[c] After population divergence, effective size increases to 1,000,000.

[d] After bottleneck, effective size increases to 100,000.

descendents of this lineage inherited the derived SNP allele. The timing of the mutation that led to the derived allele (for the SNP) was fixed before divergence. For most simulations, this time was set to 120% of the divergence time (see *Demographic Scenarios*).

## *Demographic Scenarios*

We investigated 17 two-population demographic histories. Parameters differing between models included population divergence time, population growth rate, migration, and the presence/absence of a population bottleneck. Migration was assumed to be symmetric. See table 1 for detailed descriptions of these parameters for all models.

Both ancient (5,000 unscaled generations before present) and recent (2,000 unscaled generations before present) population divergence were modeled. In both cases, the divergence time is not scaled by effective size. These divergence scenarios were then investigated in combination with migration, population growth, and population bottlenecks.

We investigated the impact of gene flow on estimates of divergence time. A migration rate of 0.00005 per generation ($N_em = 1$ [i.e., 1 migrant per generation]) was modeled after population divergence in three situations: (1) continuous gene flow after divergence (models 2a and 2b), (2) gene flow for the initial generations after divergence, for 20% of the divergence time (i.e., from 2,000 to 1,600 generations in the past for the recent divergence [model 3b] and from 5,000 to 4,000 generations in the past for the ancient divergence [model 3a]), and (3) gene flow just before the present, for 20% of the divergence time (400 generations in the past to present for the recent divergence [model 4b] and 1,000 in the past to present for the ancient divergence [model 4a]).

We investigated the impact of population growth on estimates of divergence time. Growth was modeled as an exponential process. After population divergence, one of the two populations grew such that population size at $t = 0$ (present) was 10 times the original size (100,000 [models 5a and 5b]) or 100 times the original size (1,000,000 [models 6a and 6b]).

The impacts of a population bottleneck followed by growth on estimates of divergence time were investigated. As with the investigations of population growth, only one of the two populations went through the bottleneck and subsequent growth. The bottleneck occurred immediately after population divergence. A bottleneck consisted of a reduction to 5% of the original population size: postbottleneck population size was 500 (models 7a and 7b). In both cases, subsequent exponential growth led to a population size of 1,000,000 at $t = 0$.

## *Sensitivity Analyses*

We explored the sensitivity of our results for the recent divergence scenario (table 1, model 1b: no growth, bottleneck, or gene flow) to timing of SNP mutation and to mutation model.

All of the above models represented a fairly recent SNP mutation relative to divergence time. We investigated the effect of an older mutation event (divergence time: 2,000 generations in the past; SNP mutation event: 2,700, 3,000, and 8,000 generations in the past) on estimates of divergence time (models 8a, b, and c).

We modeled variation in STR mutation rate and range constraint by randomly choosing mutation rate and range constraint from uniform distributions (mutation rate: 0.0001 to 0.005; range constraint: 10 to 20 allele sizes per locus). Twenty such "realistic loci" were simulated (1,000 observations per locus) to investigate effects on divergence time estimation. Divergence time estimates for the above analyses were based on an average mutation rate of 0.00255 (average of 0.0001 and 0.005). We did not calculate $T_{DL}$ (see *Estimates of Divergence Time*) because initial results revealed that $T_{DL}$ provides more biased estimates of divergence time than do the other statistics.

### Ascertainment

Given the focus on divergence time, we simulated SNPSTR ascertainment by including only simulated data with at least one derived SNP allele in both populations (45 samples each). Simulations for the models described above were repeated to generate 1,000 observations (after ascertainment) per analysis. For model 1a, simulations were repeated to generate observations for 1 (1,000 runs), 2 (2,000 runs), 5 (5,000 runs), 10 (10,000 runs), 20 (20,000 runs), and 100 (100,000 runs) independent loci (where a locus consists of one STR locus or one SNPSTR system). For all other models, simulations were repeated 20,000 times to generate data representing 20 independent loci.

### Estimates of Divergence Time

Three different distance statistics were used to estimate divergence time based on (1) total STR variation and (2) STR variation on the derived SNP allele. We assume that in practice the derived SNP can be identified either through comparison with an outgroup or through examination of the geographic distribution of the two SNP alleles. All samples with the derived SNP share a common ancestor. The STR variation on the derived SNP background is, thus, conditional on the single event of the SNP mutation; such genealogies are known as conditional genealogies (Wuif and Donnelly 1999; Wuif 2000). The STR variation on the derived SNP background represents a genealogically defined subsample of the data, making it appropriate to apply statistics independently to variation on the derived SNP background. Divergence time estimates based on multiple loci were obtained by averaging over loci; the estimate based on 20 loci is the average of 20 single-locus estimates.

$(\delta\mu)^2$ (Goldstein et al. 1995) is a distance statistic appropriate where data fit a stepwise mutation model (e.g., STR allele frequencies):

$$(\delta\mu)^2 = (m_1 - m_2)^2 \tag{2}$$

where $m_1$ and $m_2$ are mean repeat length in populations 1 and 2. $(\delta\mu)^2$ can be used to estimate $\mathrm{T}(\delta\mu)^2$ as follows:

$$T_{(\delta\mu)^2} = \frac{(\delta\mu)^2}{2w} \tag{3}$$

where $w$ is the STR mutation rate. $\mathrm{T}(\delta\mu)^2$ can be calculated (1) for the STR locus (SNP-blind or ignoring the SNP) and (2) based on STR variation on the derived SNP background.

The linearity of $(\delta\mu)^2$ with divergence time is strongly influenced by the presence of range constraints (Feldman et al. 1997; Zhivotovsky, Feldman, and Grishechkin 1997). $D_L$ provides a less biased estimate of divergence time than does $(\delta\mu)^2$ for STR loci with range constraints (Feldman et al. 1997). $D_L$ was estimated as follows:

$$D_L = \log\left\lfloor\left(LM - \sum_{i=1}^{L}(\delta\mu)_i^2\right)\Big/(LM)\right\rfloor. \tag{4}$$

$L$ is the number of loci and $M$ is the maximum possible distance, given by:

$$M = \frac{(R^2-1)}{6} - [4w(N-1)(1-1/R)] \tag{5}$$

where $w$ is the STR mutation rate, $N$ is the population size, and $R$ is the range constraint. $\mathrm{T}_{\mathrm{DL}}$(based on $D_L$) was estimated according to:

$$T_{DL} = \frac{D_L}{-4w(1-\cos(\pi/R))} \tag{6}$$

where $w$ is the STR mutation rate and $R$ is the range constraint. For the simulated population histories (models 1 to 8), we know the value of the range constraint ($R = 20$). As a result, we can calculate $T_{DL}$. For real data, it might be difficult to calculate $T_{DL}$ when $R$ is not known.

Zhivotovsky (2001) developed a distance-based statistic characterized by better performance than $(\delta\mu)^2$ given population growth and/or gene flow after population divergence. $\mathrm{T}_D$ estimates divergence time based on STR variation as follows:

$$T_D = \frac{D_1}{2w} - \frac{V_0}{w} \tag{7}$$

where $w$ is mutation rate, $V_0$ is the STR allelic variance in the ancestral population just before population divergence, and $D_1$ is the average square distance given by:

$$D_1 = \sum_i \sum_j (i-j)^2 x_i y_j \tag{8}$$

where $x_i$ and $y_j$ are the frequencies of repeat lengths $i$ and $j$ in populations $x$ and $y$.

We assumed $V_0 = 0$ for the STR variation on the derived SNP background when calculating $\mathrm{T}_{DdSNP}$. When calculating $\mathrm{T}_{DSTR}$ (based on total STR variation), we assumed $V_0 =$ mutation drift equilibrium allelic variance (given SMM, as suggested in Zhivotovsky [2001]).

We used all three estimators to calculate divergence time for the simplest divergence scenarios (models 1a and 1b) using (1) total STR variation and (2) SNPSTR variation (STR variation on the derived SNP background). The optimal statistics (among the above three statistics) for each marker category (STR and SNPSTR) were identified, and only those statistics were used to estimate divergence time for remaining models.

## Application to Empirical Data

We calculated the divergence time between African and non-African samples for both SNPSTR systems (22SR1 and 5SR1 [Mountain et al. 2002]) using $\mathrm{T}_{DdSNP}$ (formulas 7 and 8, based on the STR variation linked to the derived SNP allele), $\mathrm{T}(\delta\mu)^2_{\mathrm{dSNP}}$ (formulas 2 and 3, based on the STR variation linked to the derived SNP allele), and $\mathrm{T}(\delta\mu)^2_{\mathrm{STR}}$ (formulas 2 and 3, based on the STR). Outgroup comparison including chimpanzees provided information regarding ancestral SNP state for each SNPSTR system (Mountain et al. 2002). Estimates of mutation rate for human STRs range between 0.0006 and 0.003 (Ellegren 2000*b*). We estimated divergence time separately for each system using each of these rates. In addition, we calculated divergence time from an average of estimates from 5SR1 and 22SR1 using an intermediate STR mutation rate of 0.0018.

## Results

### Estimates of Divergence Time

#### *Simple Models*

Table 2 presents estimates of divergence time using $T(\delta\mu)^2_{STR}$, $T_{DSTR}$, $T_{DLSTR}$ (based on STR variation) and $T(\delta\mu)^2_{dSNP}$, $T_{DdSNP}$, $T_{DLdSNP}$ (based on SNPSTR variation, derived SNP only) for data simulated on the basis of a 5,000-generation-old population divergence (model 1a). Comparison of the estimates to the modeled divergence time reveals that $T(\delta\mu)^2$ is the best estimator (among the statistics considered) in the case of total STR variation, whereas $T_D$ is the best estimator in the case of SNPSTR variation. Similar results were observed for model 1b. We assume that optimality under models 1a and 1b carries over to more complex models. To evaluate each type of genetic marker (STR and SNPSTR) under optimal conditions, all comparisons presented henceforth focus on differences between $T(\delta\mu)^2$ (STR) and $T_D$ (SNPSTR-derived SNP). "STR" refers to $T(\delta\mu)^2$ calculated based on STR data alone. "SNPSTR" refers to $T_D$ calculated on the basis of the STR variation on the background of the derived SNP allele for each SNPSTR system.

Table 3 provides a summary of accuracy and precision of STR and SNPSTR estimators for model 1a, given the number of loci. Increasing the number of loci did not significantly change the accuracy but did, as expected, increase the precision in all cases. For example, the 5th to 95th percentile changed from 382 to 16,444 to 2,677 to 8,116 when the number of loci was increased from 1 to 10 for the SNPSTR estimator. Note that increased precision in the context of a strong bias (as for STR's alone) can be highly misleading (5th to 95th percentile, 1 locus: 15 to 18,700 and 5th to 95th percentile, 20 loci: 1,714 to 7,889). Both STR-based and SNPSTR-based estimates of divergence time based on 100 loci were characterized by much lower error (STR: 92 and SNPSTR: 56) than estimates based on fewer loci. This decreased error is primarily the result of increased precision.

The SNPSTR estimator was characterized by lower error (table 4) than the STR estimator for the divergence time estimates based on 20 loci, given simple models of recent and ancient (table 1: models 1a and 1b) population divergence. For both models, the SNPSTR estimator was less biased (fig. 2) (0.3% overestimate for ancient divergence, and 3% overestimate for recent divergence) than the STR estimator (25% underestimate for ancient divergence, and 14% underestimate for the recent divergence). Box plots (fig. 1) illustrate that the SNPSTR estimator also has lower variance for both ancient and recent divergence.

**Table 2**
**Divergence Time Estimates Based on SNPSTR and STRs for Three Statistics**

| Statistic | dSNP | STR |
|---|---|---|
| $T_D$ | 5003 | 2168 |
| $T(\delta\mu)^2$ | 3409 | 3750 |
| $T_{DL}$ | 2155 | 2586 |

NOTE.—For the derived SNP allele (dSNP), divergence time estimate is based on STR variation on the derived SNP background. All estimates are based on 20 loci.

#### *Models with Gene Flow*

Continuous gene flow of one migrant per generation after population divergence (models 2a and 2b) resulted in SNPSTR estimates characterized by lower error (table 4), reflecting higher accuracy and higher precision (fig. 2*a*) than STR estimates. Although both estimators resulted in biased estimates, the bias was greater for the STR estimator (47% and 27% for ancient and recent divergence, respectively) than for the SNPSTR estimator (31% and 16% for ancient and recent divergence, respectively) (fig. 2*a*).

Results for models where gene flow between populations took place during a limited time period immediately after population divergence (table 1: models 3a and 3b) were very similar to those models with continuous gene flow. SNPSTR estimates were characterized by lower error (table 4) resulting from higher accuracy and higher precision than STR estimates. Both SNPSTR and STR estimators underestimated divergence time, although the estimates were slightly less biased than for continuous gene flow. For example, the bias for the SNPSTR estimator was 30% and 15% for ancient and recent divergence, respectively.

Simulations with recent gene flow between populations (table 1: models 4a and 4b) revealed lower error (table 4) and higher precision and accuracy (fig. 2*b*) for the SNPSTR estimator than did the STR estimator (table 4). Although both SNPSTRs and STRs still underestimated divergence time, estimates were less biased (SNPSTR: 0.8% and 6%, respectively, for ancient and recent divergence; STR: 25% and 17%, respectively, for ancient and recent divergence).

**Table 3**
**Divergence Time Estimates for Different Numbers of Loci**

| Number of Loci | STR Estimator (Precision) | SNPSTR Estimator (Precision) | Error STR, SNPSTR |
|---|---|---|---|
| 1 | 1,309 (15–18,700) | 3,279 (382–16, 444) | 6639,5764 |
| 2 | 2,350 (160–14,267) | 3,916 (1,046–13,041) | 4621,4170 |
| 5 | 3,308 (846–13,528) | 4,603 (2,047–9,619) | 3417,2517 |
| 10 | 3,487 (1,515–10,023) | 4,715 (2,677–8,116) | 2548,1779 |
| 20 | 3,750 (1,714–7,889) | 5,003 (3,370–7,568) | 1960,1320 |
| 100 | 4,567 (3,420–6,314) | 5,010 (4,236–7,568) | 92,56 |

NOTE.—STR and SNPSTR estimators for 1, 2, 5, 10, 20, and 100 loci for a 5,000-generation-old divergence (model 1a). The 95th and 5th percentiles are indicated in parentheses. Error for STR and SNPSTR estimates are shown.

**Table 4**
**Error of Divergence Estimates for all Models**

| Model Number | Model Description | STR Bias | SNPSTR Bias | STR Precision | SNPSTR Precision | STR Error | SNPSTR Error |
|---|---|---|---|---|---|---|---|
| 1a | Ancient divergence | −0.25 | +0.003 | 1715–7185 | 3370–7568 | 1960 | 1320 |
| 1b | Recent divergence | −0.14 | −0.03 | 817–3424 | 1336–3202 | 857 | 599 |
| 2a | Ancient divergence<br>Gene flow | −0.47 | −0.31 | 1275–5319 | 2154–5216 | 2442 | 1740 |
| 2b | Recent divergence<br>Gene flow | −0.27 | −0.16 | 707–2863 | 1004–2670 | 842 | 584 |
| 3a | Ancient divergence<br>Initial gene flow | −0.36 | −0.30 | 1598–6430 | 2168–5474 | 2093 | 1740 |
| 3b | Recent divergence<br>Initial gene flow | −0.22 | −0.15 | 766–3250 | 1071–2811 | 834 | 579 |
| 4a | Ancient divergence<br>Recent gene flow | +0.25 | +0.008 | 1812–6708 | 3229–7346 | 1934 | 1725 |
| 4b | Recent divergence<br>Recent gene flow | +0.17 | +0.06 | 836–3260 | 1359–3270 | 794 | 623 |
| 5a | Ancient divergence<br>Low growth | +0.44 | −0.067 | 1376–5263 | 3073–6813 | 2361 | 1231 |
| 5b | Recent divergence<br>Low growth | +0.37 | −0.1 | 589–2428 | 1294–3116 | 869 | 582 |
| 6a | Ancient divergence<br>High growth | −0.51 | −0.13 | 1218–4867 | 2963–6702 | 2586 | 1235 |
| 6b | Recent divergence<br>High growth | +0.44 | −0.01 | 535–2178 | 1243–3097 | 942 | 572 |
| 7a | Ancient divergence<br>Bottleneck and growth | +0.25 | −0.06 | 2147–8365 | 2873–6649 | 2822 | 1192 |
| 7b | Recent divergence<br>Bottleneck & growth | +0.75 | −0.05 | 1017–3992 | 1219–2930 | 2068 | 550 |
| 8a | Recent divergence<br>2,700-year-old SNP | −0.16 | +0.13 | 812–3440 | 1100–3900 | 862 | 698 |
| 8b | Recent divergence<br>3,000-year-old SNP | −0.12 | +0.2 | 820–3100 | 1000–4000 | 849 | 770 |
| 8c | Recent divergence<br>8,000-year-old SNP | −0.14 | +0.94 | 817–3430 | 2311–6658 | 854 | 2499 |

NOTE.—Estimates of divergence time (based on 20 loci, $T(\delta\mu)^2{}_{STR}$ for STRs and $T_{DdSNP}$ for SNPSTRs) for all models were compared with true divergence time to estimate error. Higher value of error indicates a more biased and/or less accurate estimate.

### *Models with Growth and Bottlenecks*

Models with population growth (table 1: models 5a and 5b) revealed that the SNPSTR estimator was characterized by lower error (table 4), primarily reflecting lower bias than STR estimates. Simulations demonstrated overestimates of divergence time by the STR estimator (by 44% and 37%, respectively, for ancient and recent population divergence) (fig. 4*a*). The bias was far lower for the SNPSTR estimator (6% underestimate and 3% overestimate, respectively, for ancient and recent divergence). The SNPSTR estimator was characterized by a slightly higher precision than was the STR estimator (fig. 3*a*).

For models with a higher growth rate (table 1: models 6a and 6b), results were similar to those presented above, although the SNPSTR estimator was slightly more biased (13% and 1% underestimate, respectively, for ancient and recent divergence) than for lower growth rates.

Models with a population bottleneck followed by growth (table 1: models 7a and 7b) revealed that the STR estimator was characterized by a higher error (table 4) (almost 4 times as large in the case of recent divergence) than was the SNPSTR estimator, reflecting much higher accuracy and slightly higher precision (fig. 3*b*). STRs overestimated divergence time (25% and 75%, respectively, for ancient and recent divergence) compared with underestimates of divergence time by the SNPSTR estimator (6% and 5%, respectively, for ancient and recent divergence).

## Sensitivity Analyses

### *Timing of Mutation Event Creating SNP*

As the mutation event that created the SNP was pushed back in time relative to divergence, the error characterizing the SNPSTR estimator increased (table 4, error: 698, 770, and 2,499, respectively, for 2,700-generation-old, 3,000-generation-old, and 8,000-generation-old SNP; figure 4, divergence time estimates for 2,400-generation-old, 2,700-gneration-old, and 3,000-generation-old SNP) because of lower accuracy and lower precision. However, even for a 3,000-generation-old SNP, error of the SNPSTR estimator was lower than for the STR estimator (800 versus 900). The bias in estimating divergence time using $T_D$ increased with increasing SNP age (13%, 24%, and 94%, respectively, for mutation timing 2,700, 3,000, and 8,000 generations in the past).

### *Mutational Properties*

Varying mutational properties of the 20 simulated STR loci revealed much higher error for the STR estimator (table 5), reflecting much lower accuracy and higher variance. Using the STR estimator resulted in very significant underestimates of divergence time (77%)

compared with a slight overestimate by the SNPSTR estimator (6%) (table 5). Whereas the 95th percentile to 5th percentile interval was larger for the SNPSTR estimator, this interval for the STR estimator did not even include the true divergence time estimate, so that a comparison of precision is not meaningful in this case.

### Application to Empirical Data

Estimates of divergence time between African and non-African populations for two SNPSTR systems are given in table 6. For both 22SR1 and 5SR1, the STR estimates (using $T(\delta\mu)^2$) estimates were relatively low: 1,930 ($\mu = 0.003$) and 9,654 ($\mu = 0.0006$) years for 22SR1 and 6,025 ($\mu = 0.003$) and 30,147 ($\mu = 0.0006$) years for 5SR1. The highest estimates of divergence time were based on SNPSTR variation (on the derived SNP allele) and $T_D$: 28,929 ($\mu = 0.003$) and 144,645 ($\mu = 0.0006$) years for 22SR1 and 158,320 ($\mu = 0.003$) and 791,604 ($\mu = 0.0006$) for 5SR1. Estimates based on the 5SR1 system were always higher than for the 22SR1 system.

Estimates of divergence time between African and non-African populations averaged over both SNPSTR systems are given in table 6. The STR estimates (using $T(\delta\mu)^2$) estimates were relatively low (6,632). The highest estimates of divergence time were based on SNPSTR variation (on the derived SNP allele) and $T_D$ (156,041). Estimates based on the 5SR1 system were always higher than estimates based on the 22SR1 system.

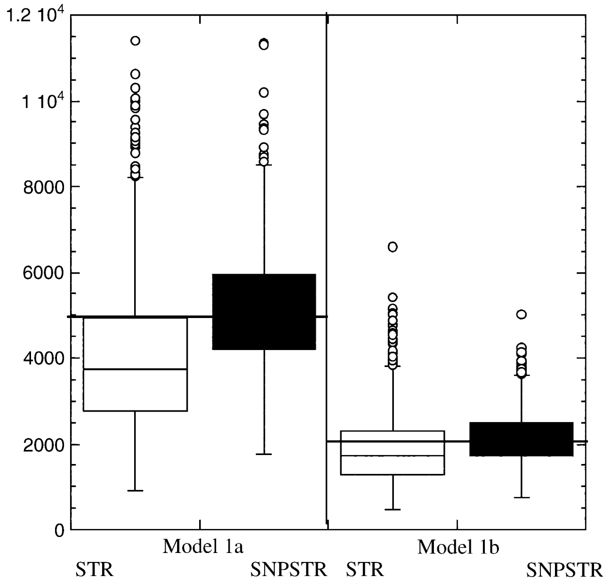

FIG. 1.—Bias and precision for the STR and SNPSTR estimators for two divergence times, 5,000 (model 1a) and 2,000 (model 1b) generations in the past. Estimates are based on 20 loci. The box plots summarize the distribution of divergence time estimates for both statistics. Upper and lower bounds of the box correspond to the 75th and the 25th percentile for the STR and SNPSTR estimators, and the midpoint of the box represents the 50th percentile (bias is the percentage difference between this estimate and the true value, represented by a solid line). Error bars correspond to the 95th and 5th percentile for the STR and SNPSTR estimators (i.e., precision). Circles represent outliers in the STR and SNPSTR estimator distributions.

## Discussion

The coalescent simulations of population divergence presented here demonstrate that estimates of divergence time based on SNPSTR data are characterized by lower error than those based solely on the STR data, irrespective of the demographic history and conditional on the modeled SNP ages. The lower error of SNPSTR estimates is caused by both higher accuracy (lower bias) and higher precision in most cases. We compared the most accurate estimator of divergence time for total STR variation ($T(\delta\mu)^2$) to the most accurate estimator of divergence time for SNPSTR variation ($T_D$) to evaluate each type of marker under optimal conditions, strengthening the above result. The lower error for SNPSTR data is particularly impressive because the number of STR alleles on the derived SNP background is a subset of the total STR variation. (i.e., the sample size on which the SNPSTR estimator are based is much lower than that of the STR estimator). The reduced error reflects the lower levels of STR homoplasy on the derived SNP background.

When STR variation on the derived SNP is used to estimate divergence time, $T_D$ serves as a better estimator than $T(\delta\mu)^2$ or $T_{DL}$. Given a stepwise mutation model for STR evolution, Goldstein et al. (1995) demonstrated that either average square distance (ASD, or $D_1$; formula 8) or $(\delta\mu)^2$ (formula 2) can be used to estimate population divergence time. Estimates based on $(\delta\mu)^2$ have lower variance, making it the preferred statistic. However, estimates of divergence time using $(\delta\mu)^2$ are based on the assumption that the predivergence population is in mutation-drift equilibrium. In the models presented here, because the mutation creating the derived SNP allele is fairly recent compared with population divergence, this assumption is inappropriate for the derived SNP. Hence, $T(\delta\mu)^2$ estimates based on STR variation at the derived SNP allele are biased. Because $T_D$ was not developed based on this assumption, it serves as a more accurate and precise estimator in the context of SNPSTR variation. Additionally, the fact that $T_{DL}$ does not serve as a good estimator suggests that homoplasic variation caused by range constraints is low at the derived SNP allele for this simulated data set.

As expected for both STR and SNPSTR estimators, divergence time estimates for simulations of population divergence scenarios without gene flow, expansion, or bottlenecks (table 1; models 1a and 1b) were less biased than for more complex histories (Zhivotovsky 2001). The STR estimator, however, was consistently more biased than the SNPSTR estimator. Comparing estimates between the two divergence times demonstrated that bias was greater for ancient versus recent divergence (25% versus 14%) for the STR estimator. This result reflects increased homoplasic variation at the STR locus in the case of ancient divergence. On the other hand, for the SNPSTR estimator, the ancient divergence time estimates were less biased than the recent estimates (0.3% versus 3%). The improved estimates for ancient divergence time probably reflect the difference in the proportion of samples with the derived SNP (between model 1a and model 1b). Because the derived SNP allele is younger in the case of the recent

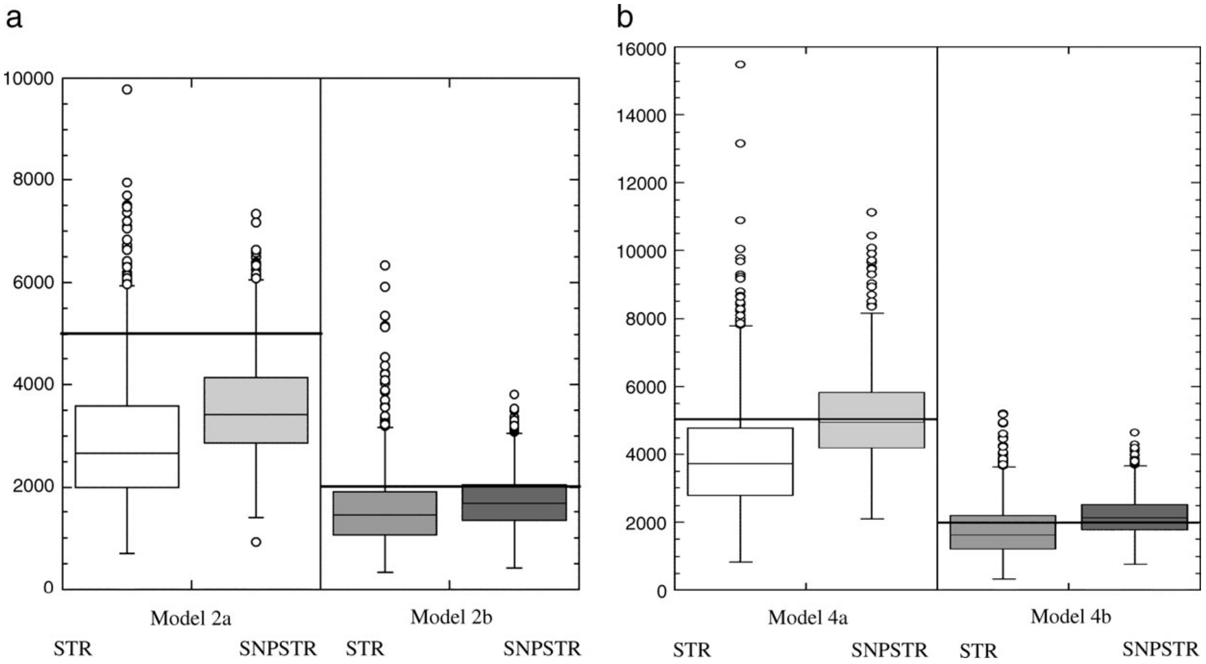

FIG. 2.—Bias and precision for the STR and SNPSTR estimators four models of population history. Estimates are based on 20 loci. The box plots summarize the distribution of divergence time estimates for both statistics. Upper and lower bounds of the box correspond to the 75th and the 25th percentile for the STR and SNPSTR estimators, and the midpoint of the box represents the 50th percentile (bias is the percentage difference between this estimate and the true value, represented by a solid line). Error bars correspond to the 95th and 5th percentile for the STR and SNPSTR estimators (i.e., precision). Circles represent outliers in the STR and SNPSTR estimator distributions. (*a*) Two divergence times, 5,000 (model 2a) and 2,000 (model 2b) generations in the past, followed by continuous gene flow (1 migrant per generation); (*b*) two divergence times, 5,000 (model 4a) and 2,000 (model 4b) generations in the past, followed by recent gene flow (1 migrant per generation).

divergence (2,400 generations versus 6,000 generations), a smaller proportion of the samples have the derived allele, leading to greater sampling error. Because improved estimates of divergence time reflect a combination of decreased homoplasy and increased sample size (individuals with STR variation on the derived SNP background), the contributions of these factors cannot be estimated individually.

Models with gene flow between populations (table 1: models 2a–4b) resulted in very biased estimates, reflecting the high level of gene flow modeled here. Estimates will be less biased for lower levels of gene flow, especially in cases where gene flow occurred soon after the divergence event. The small change in bias between continuous gene flow and ancient gene flow may be the result of the smaller number of mutations generated at the tips of a coalescent genealogy. Hence, recent gene flow does not have as significant an impact on STR variation as does ancient gene flow, at least for the level of gene flow modeled in our simulations. Higher levels of recent gene flow might significantly impact bias and variance of divergence time estimates.

Demographic scenarios incorporating growth or a bottleneck followed by growth (table 1: models 5a–7b) always resulted in biased estimates of divergence time for the STR, as expected (Zhivotovsky 2001). Such bias is less evident when STR variation on the derived SNP background is used to estimate divergence time: the bias is always lower than 13% when using the SNPSTR estimator. On the other hand, the STR estimator resulted in bias between 25% and 75%.

Sensitivity analyses indicated that the earlier the mutation event that generated the SNP, the greater the bias in SNPSTR estimates of divergence time. However, only when the mutation generating the SNP occurred 50% earlier than the divergence time is the bias comparable to STR estimates. We also investigated models for much older SNPs (between 3,000 and 8,000 generations in the past). We found that for very old SNPs, $T_D$ estimates are characterized by a bias of +95% and that bias increases linearly with SNP age. The increased bias is consistent with our understanding that for $T_D$, the older the SNP mutation, the less valid is the assumption that $V_0 = 0$ (Zhivotovsky 2001). Additionally, as the SNP gets older, the value of $T(\delta\mu)^2$ based on the derived SNP allele asymptotically approaches the STR estimate of divergence time.

Our results demonstrate that the relevance of a given SNPSTR system depends on the timing of the population events in question. The same dependency arises in the case of the Y chromosome. Each unique event polymorphism (UEP) on the Y chromosome, having arisen in a particular geographic location at a particular time, has the potential to reveal information regarding a subset of population events. The UEP defines a "haplogroup" and associated STR variation provides insight into population history (Stumpf and Goldstein 2001). Although the Y chromosome provides the advantage of multiple STRs linked to any given UEP, SNPSTR systems provide a different advantage: multiple, independently evolving systems relevant in the context of any given population event. Using a set of SNPSTR systems with a range of SNP ages (including

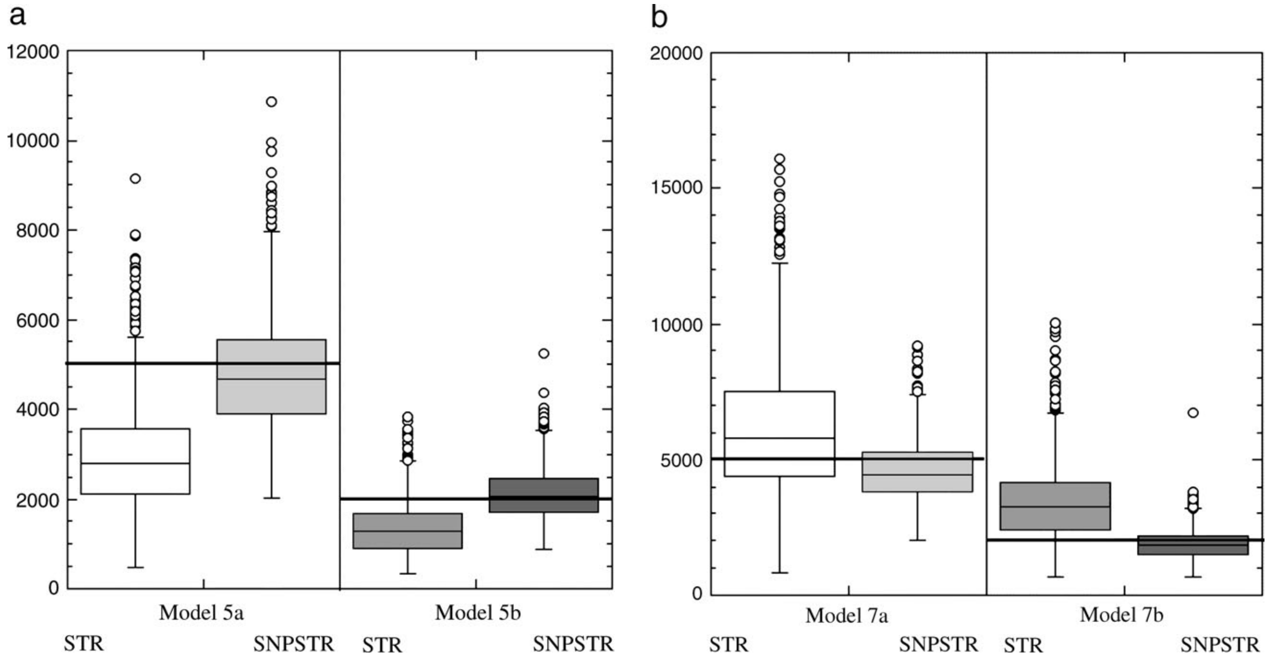

FIG. 3.—Bias and precision for the STR and SNPSTR estimators for two divergence times, 5,000 (model 5a) and 2,000 (model 5b) generations. Estimates are based on 20 loci. The box plots summarize the distribution of divergence time estimates for both statistics. Upper and lower bounds of the box correspond to the 75th and the 25th percentile for the STR and SNPSTR estimators and the midpoint of the box represents the 50th percentile (bias is the percentage difference between this estimate and the true value, represented by a solid line). Error bars correspond to the 95th and 5th percentile for the STR and SNPSTR estimators (i.e., precision). Circles represent outliers in the STR and SNPSTR estimator distributions. (*a*) Two divergence times, 5,000 (model 5a) and 2,000 (model 5b) generations in the past, followed by exponential population growth in one of the two diverging population, and (*b*) two divergence times, 5,000 (model 7a) and 2,000 (model 7b) generations in the past, followed by a bottleneck and exponential population growth in one of the two diverging population.

some recent) greatly increases the benefits of using SNPSTR (derived SNP) estimates of divergence time.

In most data analysis scenarios, estimates of range constraint and mutation rate require knowledge or assumptions about population history, making them difficult to apply. Extensive molecular analysis has indicated high levels of variation in mutational properties among STR loci (Huang et al. 2002). Given the high levels of homoplasy expected under such conditions, we can expect estimates based on SNPSTR variation to be closer to the true divergence time than those based on total STR variation: there is less homoplasy in the SNSPTR case because the STR variation on the derived SNP allele background is more recent.

Simulations described here reveal more accurate and precise estimation from SNPSTR variation than from STR variation for a given number of independent loci. Given that in some cases SNPSTR systems may be more costly to genotype than STRs, we compared estimates of divergence time based on five SNPSTR systems with those based on 10 STRs. We found that estimates based on five SNPSTR systems are still characterized by lower error (because of lower bias and lower variance) than those estimates based on 10 STR systems. Divergence time estimates based on 20 SNPSTR systems remain less biased than estimates from 100 STR loci. The variance of estimates based on 20 SNPSTR systems, however, is higher than that of estimates based on 100 STRs. Novel methods for more cost-effective genotyping of SNPSTR systems are currently under development (A. Knight, personal communication) and will greatly facilitate the use of SNPSTRs in the study of humans and of other species.

One additional important statistical measure is that of power. Because the levels of bias are so high, estimates of power for both the SNPSTR and the STR estimators do not illustrate relevant statistical properties of these estimators correctly. We note, however, that with as few as 20 loci, the power to distinguish between null models involving no population divergence and alternative models involving both recent and ancient population divergence was close to its maximum value of 1 for both the SNPSTR and the STR estimators (models 1a and 1b, details not shown). Shorter divergence times would require a greater number of loci to attain power close to 1.

The initial divergence between African and non-African populations is estimated to have taken place on the order of 95,000 to 170,000 years ago (for example, Zhivotovsky et al. [2003], based on 377 STR loci). We interpret divergence time estimates based on STR and SNPSTR variation (table 6) in the light of those dates and our simulation results. As expected, $T(\delta\mu)^2$ based on the STR variation alone greatly underestimated divergence time. Variation at the two SNPSTR systems revealed more evidence for homoplasy at 22SR1 than did 5SR1 (Mountain et al. 2002), consistent with lower $T(\delta\mu)^2$ and $T_D$ estimates from 22SR1 variation than from 5SR1 variation. Simulations where both mutational properties of the STR (mutation rate and range constraint) and timing of SNP varied across loci demonstrated that the lowest estimates of divergence time (based on STR variation on

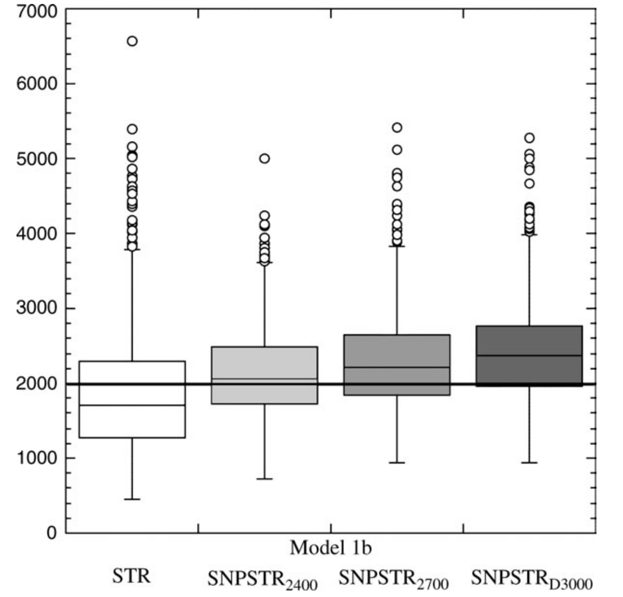

FIG. 4.—Bias and precision of the SNPSTR estimator when the SNP is 2,400-generations-old, 2,700-generations-old, and 3,000-generations-old. Bias and precision for the STR estimator is the same in all three cases. Estimates are based on 20 loci. The box plots summarize the distribution of divergence time estimates for both statistics. Upper and lower bounds of the box correspond to the 75th and the 25th percentile for the STR and SNPSTR estimators and the midpoint of the box represents the 50th percentile (bias is the percentage difference between this estimate and the true value, represented by a solid line). Error bars correspond to the 95th and 5th percentile for the STR and SNPSTR estimators (i.e., precision). Circles represent outliers in the STR and SNPSTR estimator distributions.

the derived SNP allele and $T_D$) were the most accurate for a set of loci; older SNPs led to consistently higher, positively biased estimates of divergence time. Divergence time estimates from 22SR1 variation are, therefore, expected to better reflect true divergence time than are estimates from 5SR1 variation. Additionally, the average divergence time estimate based on STR variation on the derived SNP background and $T_D$ (156,041) is within the range estimated by 377 STRs. In this sense, the two SNPSTR systems provide a better estimate than do the corresponding STR systems.

We have demonstrated that SNPSTR systems provide more accurate and precise estimates of population divergence time than do STR polymorphisms, particularly when the mutation event generating the SNP took place shortly before population divergence. In practice, it is possible to investigate whether a particular SNPSTR system is informative for estimating divergence times for the sampled populations. In fact, the ratio of $T_D$ (in the context of SNPSTR systems) to $T(\delta\mu)^2$ (based on the total STR variation) can be used as a qualitative measure. The value of $T_D$ increases rapidly with SNP age. On the other hand, the value of $T(\delta\mu)^2$ does not change with SNP age. We found that the ratio $T_D/T(\delta\mu)^2$ increased with the age of the SNP. When the SNP is pushed back from 2,400 generations to 3,000 generations before present (given population divergence 2,000 generations ago), this ratio goes from 1.1 to 1.4, and is even higher for a very old SNP (1.9 for 6,000-generation-old SNP). This ratio provides a qualitative way to assess the age of the SNP in the SNPSTR system. For the SNPSTR data presented on human populations, $T_D/T(\delta\mu)^2$ for 5SR1 was twice as high as the value for 22SR1, suggesting that the 22SR1 SNP mutation is younger. Alternatively, analytical methods (e.g., Slatkin and Rannala 2000) and coalescent-based Bayesian methods such as BATWING (Wilson et al. 2003) can be used to estimate the age of the SNP.

**Table 5**
**Bias, Precision, and Error of Estimates of Divergence Time for 20 Realistic Loci (Model 1b)**

| | Bias | Precision | Error |
|---|---|---|---|
| SNPSTR | +0.06 | 1,289–3,118 | 577 |
| STR | −0.77 | 245–864 | 1,500 |

NOTE.—Bias (represented as difference between the median estimated divergence time and the true divergence time), precision (the 5th and 95th percentile of divergence time estimates) and error (square root of mean square difference between estimated and true divergence time) of the STR and SNPSTR estimators of divergence time for 20 realistic loci (mutation rate: 0.0001–0.005; rate constraint: 10–20).

We expect that SNPSTR variation will provide more accurate and precise estimates not only of divergence time but also of other population genetic parameters such as rates of gene flow and population growth. Recent work by Hey et al. (2004) suggests that SNPSTR systems are more informative regarding gene flow than are STRs alone. Further work is necessary to evaluate the accuracy and precision of such estimates.

In models presented here, we assume complete linkage between the SNP and the STR constituting the SNPSTR system. We expect this assumption to be valid in the case of the human genome when the distance between the SNP and the STR is less than 500 bp (as is the case for previously described SNPSTR systems [Mountain et al. 2002]). In simplified terms, the human genome can be characterized as tightly linked sequence blocks separated by recombination hotspots. Characterization of linkage disequilibrium in the human genome demonstrates that blocks of DNA of length 5,000 bp are essentially completely linked (Reich et al. 2001). The probability of recombination between the SNP and the STR is very low if the SNPSTR system falls within a tightly linked block. Given the distribution of recombination hotspots in the human genome, the chance that a SNPSTR will overlap with a recombination hotspot is low (~3% [Mountain et al. 2002]). Models incorporating recombination will be more important when considering SNPSTR systems of other species or for human SNPSTR systems where the SNP and STR are further apart. We expect that as the distance between the SNP and the STR increases, the rate of recombination will increase, and SNPSTR variation will provide less additional information beyond the STR variation. If the SNP and the STR locus are completely unlinked, we expect estimates of divergence time based on the derived SNP background to be similar to estimates based on the STR variation alone. The overall probability that recombination between the SNP and STR has introduced haplotypes distinct from the parent haplotypes depends not only on the physical distance separating the

**Table 6**
**Estimates of Divergence Time Between African and non-African Human Populations from STR Allele Frequencies or SNPSTR Haplotype Frequencies**

| | $T(\delta\mu)^2_{STR}$ | | $T(\delta\mu)^2_{dSNP}$ | | $T_{DdSNP}$ | |
|---|---|---|---|---|---|---|
| SNPSTR | m = 0.0006 | m = 0.003 | m = 0.0006 | m = 0.003 | m = 0.0006 | m = 0.003 |
| 22SR1 | 9,654 | 1,930 | 31,428 | 6,285 | 144,645 | 28,929 |
| 5SR1 | 30,147 | 6,029 | 72,2015 | 14,403 | 791,604 | 158,320 |
| Average | 6,632 | | 17,270 | | 156,041 | |

NOTE.—$T(\delta\mu)^2$ $_{STR}$: from total STR variation. $T(\delta\mu)^2$ $_{dSNP}$: from STR variation on the derived SNP background. $T_{DdSNP}$: from STR variation on the derived SNP background. Estimates represent years before present for a high ($\mu = 0.003$) and low ($\mu = 0.0006$) STR mutation rate. Average estimates represent years before present for average STR mutation rates ($\mu = 0.0018$) averaged over loci. Generation time = 25 years. Source: Mountain et al. (2002).

SNP and STR but also on the age of the SNP; if recombination is relatively rare, we expect the probability that such distinct haplotypes have been generated to be lower for younger than for older SNPs.

Here we have focused on simple two-population models. Inference of more complex population histories will benefit from evaluation of many SNPSTR systems with a range of SNP ages. Fortunately, each genome harbors thousands of potential SNPSTR systems: for humans, SNPs have been identified within 500 bp upstream or downstream of STRs in about 25% of the cases examined (Mountain et al. 2002). We can, therefore, expect these systems to contribute substantially to our understanding of evolutionary history.

## Acknowledgments

We thank Marcela Miyazawa for assistance with calculating summary statistics. We thank Melissa Brown, David Degusta, Marc Feldman, Matt Jobin, Alec Knight, Karen Martien, Charles Roseman, Peter Underhill, Lev Zhivotovsky. We also thank David Posada and one anonymous reviewer for comments on earlier versions of this manuscript. This research was supported by NIH grant GM 28428.

## Literature Cited

Arbogast, B. S., S. W. Edwards, J. Wakeley, P. Beerli, and J. B. Slowinsky. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu. Rev. Ecol. Syst. **33**:707–740.

de Kniff, P. 2002. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am. J. Genet. **67**:1055–1061.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, and M. Slatkin. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91**:3166–3170.

Ellegren, H. 2000*a*. Heterogeneous mutation processes in human microsatellite DNA sequences. Nature Genet. **24**:400–402.

———. 2000*b*. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. **16**:551–558.

Estoup, A., P. Jarne, and J. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. **11**:159–1604.

Excoffier, L., J. Novembre, and S. Schneider. 2000. SIMCOAL: a general coalescent program for simulation of molecular data in interconnected populations with arbitrary demography. J. Hered. **91**:506–509.

Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics **145**:207–216.

Goldstein, D. B., A. R. Linares, M. W. Feldman, and L. L. Cavalli-Sforza. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Nat. Acad. Sci. USA **92**:6723–6727.

Grant, T., and A. G. Kluge. 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics **19**:379–418.

Grimaldi, M. C., and B. Crouau-Roy. 1997. Microsatellite allelic homoplasy due to variable flanking sequences. J. Mol. Evol. **44**:336–340.

Hey, J., Y-J. Won, A. Sivasundar, R. Nielsen, and J. A. Markert. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. Mol. Ecol. **13**:909–919.

Huang, Q. Y., F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker, and H. W. Deng. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. **70**:625–634.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7**:1–44.

Hurles, M. E., R. Vietia, E. Arroyo, M. Armenteros, J. Bertraboetit, A. Perez-Lezaun, E. Bosch, M. Shluurnukova, A. Cambon-Thomsen, and K. McElreavey. 1999. Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. Am. J. Hum. Genet. **65**:1437–1488.

Jin, L., M. L. Baskett, L. L. Cavalli-Sforza, L. A. Zhivotovsky, M. W. Feldman, and N. A. Rosenberg. 2000. Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. Ann. Hum. Genet. **64**:117–134.

Kayser, M., S. Brauer, G. Weiss, W. Schiefenhovel, P. Underhill, P. Shen, P. Oefner, M. Tommaseo-Ponzetta, and M. Stoneking. 2003. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am. J. Hum. Genet. **72**:281–302.

Kimura M., and T. Ohata. 1978. Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Nat. Acad. Sci. USA **75**:2868–2872.

Knight, A., P. A. Underhill, H. M. Mortensen, L. A. Zhivotovsky, M. Ruhlen, and J. L. Mountain. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr. Biol. **13**:464–473.

Makova, K. D., J. C. Patton, E. Y. Krysanov, R. K. Chesser, and R. J. Baker. 1998. Microsatellite markers in wood mouse and striped field mouse (genus *Apodemus*). Mol. Ecol. **7**:247–249.

Mountain, J. L., A. Knight, M. Jobin, C. Gignoux, A. Miller, A. A. Lin, and P. A. Underhill. 2002. SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. Genome Res. **12**:1766–1772.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature **411**:199–204.

Slatkin, M., and B. Rannala. 2000. Estimating allele age. Annu. Rev. Genomics Hum. Genet. **1**:225–249.

Stumpf, M. P. H., and D. B. Goldstein. 2001. Genealogical and evolutionary inferences with the human Y chromosome. Science **291**:1738–1741.

Tautz, D., and C. Schlotterer. 1994. Simple sequences. Curr. Opin. Genet. Dev. **4**:832–837.

Tishkoff S. A., A. J. Pakstis, M. Stoneking, J. R. Kidd, G. Destro-Bisol, A. Sanjantila, R. Lu, A. Deinard, G. Sirugo, T. Jenkins, K. K. Kidd, and A. G. Clark. 2000. Short tandem-repeat polymorphism/Alu haplotype variation at the plat locus: implications for modern human origins. Am. J. Hum. Genet. **67**:901–925.

Tishkoff, S. A., G. Ruano, J. R. Kidd, and K. K. Kidd. 1996. Distribution and frequency of a polymorphic Alu insertion at the PLAT locus in humans. Hum. Gen. **97**(6):759–764.

Van Oppen, J. H., C. Rico, G. F. Turner, and G. M. Hewitt. 2000. Extensive homoplasy, nonstepwise mutations and ancestral polymorphism at a complex microsatellite locus in the lake Malawi Cichlids. Mol. Biol. Evol. **17**:489–498.

Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

Wilson, I. J., M. E. Weale, and D. J. Balding. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. J. R. Stat. Soc. Ser. A **166**:155–188.

Wuif, C. 2000. On the genealogy of a sample of rare neutral alleles. Theor. Pop. Biol. **58**:61–75.

Wuif, C., and P. Donnelly. 1999. Conditional genealogies and the age of a neutral mutant. Theor. Pop. Biol. **56**:183–201.

Zhivotovsky, L. A. 2001. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol. Biol. Evol. **18**:700–709.

Zhivotovsky, L. A., M. W. Feldman, and S. A. Grishechkin. 1997. Biased mutations and microsatellite variation. Mol. Biol. Evol. **14**:926–933.

Zhivotovsky, L. A., N. A. Rosenberg, and M. W. Feldman. 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am. J. Hum. Genet. **72**:1171–1186.

Scott Edwards, Associate Editor

Accepted July 5, 2004