Full text provided by www.sciencedirect.com

Integrating patterns of polymorphism at SNPs and STRs

Bret A. Payseur[1] and Asher D. Cutter[2]

[1]Laboratory of Genetics, University of Wisconsin, Genetics/Biotechnology 2428, 425-G Henry Mall, Madison, WI 53706, USA
[2]Institute of Evolutionary Biology, University of Edinburgh, King's Buildings, W. Mains Rd, Edinburgh EH9 3JT, UK

Single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs) differ in mutation rate and mechanism. As a result of these differences, simultaneous consideration of polymorphism patterns at SNPs and STRs can provide insights that are difficult to obtain from analysis of either marker type in isolation. Here, we use coalescent simulations to model the opposing effects of contrasting mutational dynamics and of shared genealogical history on the correlation between polymorphism at linked SNPs and STRs. Results show that polymorphism patterns are correlated only weakly despite the shared underlying genealogy, underscoring the importance of divergent mutational processes. Examples illustrate how knowledge of these relationships could aid population genetic inference, indicating the need for thorough theoretical studies.

Alternative molecular markers for surveying genetic variation

The ability to survey DNA polymorphism on a genomic scale is accelerating both the pace and scope of genetic research. Based on the question of interest, researchers can choose between classes of molecular markers with different biological characteristics. The two most widely used marker types are single nucleotide polymorphisms (SNPs) (see Glossary) and short tandem repeats (STRs). The choice between SNPs and STRs depends on several factors, including marker density and polymorphism level. Genomes contain many more SNPs than STRs; combined with increased accessibility to SNPs through recent advances in high-throughput genotyping, this observation has catalyzed a shift in marker preference toward SNPs, at least in genetic model organisms. However, STRs typically mutate faster than SNPs [1–4], producing greater levels of variation, and often provide more information per marker. The two marker types also differ in the mutational process: SNPs usually mutate through changes in single base-pair identities, whereas STRs typically mutate by addition or subtraction of repeats. As a result of these differences in mutation rate and mechanism, patterns of variation at SNPs and STRs convey complementary information. This fact, along with the interdigitation of the two marker types throughout genomes, suggests that the joint application of SNPs and STRs can provide biological insight not available from investigation of either marker type in isolation. However, the expected correlation between diversities at linked SNPs and STRs has not been investigated. We provide a preliminary look at this relationship and encourage an interpretative framework that incorporates this information.

Successful applications of SNPs and STRs in combination

Individual STRs usually display more variation than individual SNPs because STRs mutate faster. However, the greater mutability combined with the nature of the mutation process at STRs incurs a cost: alleles are frequently identical by state without being identical by descent. This homoplasy translates into uncertainty about the genealogical processes that generate STR diversity, prompting caution, particularly when comparing

Glossary

Allele excess (*E*): the difference between the observed number of alleles and the number of alleles expected based on the observed heterozygosity [20]. This statistic measures the skew in the frequency spectrum of alleles for STRs.

Effective population size (*N*): the number of breeding individuals in an idealized random-mating population that would have the same population genetic properties as the population under consideration.

Heterozygosity (*H*): the frequency with which two alleles randomly drawn from a population are different. Under the single-step stepwise mutation model at equilibrium [15]: $H = 1 - \frac{1}{\sqrt{8N\mu+1}}$

Homoplasy: a match between alleles that arises from multiple mutations rather than common ancestry.

Infinite sites mutation model: a model of mutation for DNA sequences in which mutations only occur at previously unmutated sites.

Nucleotide diversity (θ_π): the average number of pairwise sequence differences. Under the infinite sites model at equilibrium and with no recombination, $\theta_\pi = 4N\mu$ [49].

Number of segregating sites (*S*): the number of polymorphic sites in a DNA sequence.

Short tandem repeat (STR; *or* microsatellite): a class of molecular marker that exhibits variation in the number of repeats.

Single nucleotide polymorphism (SNP): a class of molecular marker that exhibits variation in the identity of base pairs, typically with only two observed states in a population.

Stepwise mutation model: a model of mutation for STRs in which each new mutation causes an increase or decrease in the number of repeats. Under the original version of this model [15] (used in this article), each new mutation increases or decreases the number of repeats by one, with equal probability.

Tajima's D (*D*): the normalized difference between two estimators of $4N\mu$ [21]. This statistic measures skew in the frequency spectrum of polymorphisms for SNPs.

Variance in allele size (*V*): the average squared deviation of STR allele lengths from the population mean. Under the single-step stepwise mutation model at equilibrium, $V = 2N\mu$ [16].

Corresponding author: Payseur, B.A. (payseur@wisc.edu).
Available online 27 June 2006

www.sciencedirect.com 0168-9525/$ - see front matter © 2006 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2006.06.009

Box 1. Covariation between SNP and STR diversity

The expected covariance between the squared difference in allele size (V; assuming the single-step stepwise mutation model) and the number of segregating sites (S; assuming the infinite sites model) for a sample of two chromosomes from an equilibrium population of size N can be calculated using a simple modification of expressions presented by Pritchard and Feldman [41], who considered the covariance between polymorphism levels at linked STRs. This derivation assumes that mutational events at STR and SNP loci occur independently at rates μ_{STR} and μ_{SNP}. The respective times to the most recent common ancestors for each locus are denoted by t_{STR} and t_{SNP}, with particular instances denoted by T_{STR} and T_{SNP}. E denotes expectation, P denotes probability density, Cov denotes covariance and Corr denotes correlation. The covariance between V and S is:

$$\mathrm{Cov}[V,S] = \mathrm{E}[VS] - \mathrm{E}[V]\mathrm{E}[S] \qquad \text{[Eqn I]}$$

Where

$$\mathrm{E}[VS] = \int\limits_{T_{STR}} \int\limits_{T_{SNP}} \mathrm{E}[V|T_{STR}]\cdot \mathrm{E}[S|T_{SNP}]\cdot \mathrm{P}[T_{STR}, T_{SNP}]\partial T_{STR}\partial T_{SNP} \qquad \text{[Eqn II]}$$

which, recalling that $\mathrm{E}[V|T_{\mathrm{STR}}]=2\mu_{\mathrm{STR}}T_{\mathrm{STR}}$ [16] and $\mathrm{E}[S|T_{\mathrm{SNP}}]=4\mu_{\mathrm{SNP}}T_{\mathrm{SNP}}$ [17], can be simplified to:

$$\mathrm{E}[VS] = 8\mu_{\mathrm{STP}}\mu_{\mathrm{SNP}}\mathrm{E}[t_{\mathrm{STR}}t_{\mathrm{SNP}}] \qquad \text{[Eqn III]}$$

Without recombination, the two loci share a genealogy, so that $t=t_{\mathrm{STR}}=t_{\mathrm{SNP}}$. t follows an exponential distribution, with $\mathrm{E}[t]=N$ and $\mathrm{E}[t^2]=2N^2$. Therefore, we can simplify the expression for the covariance to:

$$\mathrm{Cov}[V,S] = 8\mu_{STR}\mu_{SNP}\mathrm{E}[t^2] - (2\mu_{STR}\mathrm{E}[t])(4\mu_{SNP}\mathrm{E}[t]) = 8\mu_{STR}\mu_{SNP}N^2 \qquad \text{[Eqn IV]}$$

This covariance formula can be used in conjunction with expressions for the variances of V and S [41,17] to generate the expected correlation in diversities at the SNP and STR loci. After some simplification, this yields:

$$\mathrm{Corr}[V,S] = 2N\sqrt{\frac{2\mu_{STR}\mu_{SNP}}{(1+10N\mu_{STR})(1+4N\mu_{SNP})}} \qquad \text{[Eqn V]}$$

Results from simulations with samples of two chromosomes indicate that this theory performs well. The formula can also be generalized to account for recombination by allowing genealogies at the two loci to be incompletely correlated (i.e. $t_{\mathrm{STR}} \neq t_{\mathrm{SNP}}$) [50].

populations that diverged in the distant past [5]. On a population-genetic timescale, SNPs generally suffer no such problem, but might not offer enough polymorphism to unravel recently acting evolutionary mechanisms, particularly in species with small effective population sizes (N). Several investigators have demonstrated that information from the two marker types can be combined to ‘anchor’ the large rate of polymorphism at STRs with low-homoplasy SNPs. For example, methods that jointly consider polymorphism from one STR completely linked to one SNP return more accurate and precise estimates of population divergence time than approaches based solely on STR variation [6]. Joint consideration of SNP and STR polymorphism from the human Y chromosome provides useful information for reconstructing phylogeographic history (e.g. Ref. [7]). In addition, a likelihood model aimed at distinguishing between gene flow and population divergence as causes of observed levels of differentiation was modified to analyze haplotypes composed of SNPs and STRs [8,9]. Hey *et al.* [9] applied this method to two species of cichlid fishes from Lake Malawi and found clear evidence for ongoing gene flow between species.

Combining diversities at SNPs and STRs can also yield insight about natural selection. For example, Tishkoff *et al.* [10] used STR variation nested within SNP alleles to characterize the strength of selection on and age of a human *G6PD* variant that confers resistance to malaria. The utility of linked pairs of SNP and STR loci has led some groups to develop procedures to rapidly locate and genotype them (called ‘SNPSTRs’) [11].

Expected correlations between SNP and STR diversities

The examples in the previous section show that jointly considering diversities at linked SNPs and STRs can improve inference in population genetics. The success of these approaches and the accumulation of SNP and STR polymorphism data on a genomic scale raise the question of how polymorphism at linked SNPs and STRs should covary in general. As we will demonstrate, a theoretical context for such patterns would be useful for several problems in population genetics.

The expected correlation between the number of segregating sites (S; at SNPs mutating according to an infinite sites mutation model) and the squared difference in allele size (V; at STRs mutating according to a stepwise mutation model) can be calculated for a sample of size of two chromosomes from an equilibrium population (Box 1). These results suggest two biological contributors to the relationship between S and V. First, the correlation grows with the effective population size (N). Second, the correlation increases with the mutation rates (μ) at the two locus types. Therefore, holding other parameters constant and given complete linkage, species with greater variation (which is products of N and μ) should exhibit stronger correlations between these particular summaries of polymorphism.

Obtaining analytic expressions for samples of more than two chromosomes or for other measures of polymorphism is challenging, primarily because of STR homoplasy. Further results are possible by assuming that each STR mutation produces a unique allele [12], but the stepwise mutation process seems more appropriate for STRs [13]. Therefore, to examine briefly the relationships between variation at linked SNPs and STRs under more general conditions, we conducted computer simulations. We modified the output of a coalescent simulation program [14] to generate STR variation with a single-step stepwise mutation model and SNP variation with an infinite sites mutation model applied to the same genealogy. We confirmed that results of the simulations matched theoretical expectations for means and variances at both locus types [15–17]. Correlations between some commonly used polymorphism statistics, each based on 10 000 simulated samples of 50 chromosomes, are

Table 1. Correlations between polymorphism summary statistics for completely linked SNPs and STRs[a]

	H				*V*				*E*			
θ_{STR}	1	10	100	1000	1	10	100	1000	1	10	100	1000
S	0.128[b]	0.108	0.098	0.099	0.163	0.158	0.184	0.186	−0.042	0.056	0.029	−0.016[d]
θ_π	0.197	0.158	0.141	0.135	0.230	0.219	0.261	0.267	−0.115	0.020[c]	0.001[d]	−0.028
D	0.173	0.147	0.112	0.111	0.168	0.175	0.188	0.201	−0.141	−0.036	−0.025[c]	−0.025[c]

[a] Abbreviations: H = STR heterozygosity, V = STR variance in allele size, E = STR allele excess, S = SNP number of segregating sites, θ_π = SNP pairwise nucleotide diversity, D = SNP Tajima's D.

[b] Values are Pearson product–moment correlations. All correlations have $P<0.006$ except

[c] $0.01 \leq P \leq 0.05$ and

[d] $P>0.05$.

presented in Table 1. Because of space limitations, we present results for one scaled mutation rate for SNPs ($\theta_{SNP}=1$) and four scaled mutation rates for STRs ($\theta_{STR}=$ 1, 10, 100 and 1000).

Although results (not shown) for simulated samples of size two verify the accuracy of predictions from the theory in Box 1, the simulations also suggest that these predictions do not necessarily generalize to larger samples or to other summary statistics of polymorphism. For example, correlations between STR heterozygosity (H) and measures of SNP variation decrease slightly as the scaled STR mutation rate increases (Table 1).

The most general feature of Table 1 is that indices of diversity at linked SNPs and STRs are correlated weakly, an observation that underscores the importance of differences in mutational processes at SNPs and STRs. For example, adding mutations at equivalent rates ($\theta_{SNP}=\theta_{STR}=1$) along the same genealogy produces a correlation of only 0.197 between nucleotide diversity at a SNP locus (θ_π) and H at a STR. This correlation is statistically significant, but illustrates that variation at one locus does not strongly predict variation at the other locus, despite their identical genealogical history. For comparison, the analogous correlation between θ_π values at two completely linked SNP loci is twice as large (0.393; estimated from simulations given $\theta_{SNP1}=\theta_{SNP2}=1$) whereas the correlation between H values at two completely linked STRs is 0.132 (estimated from simulations given $\theta_{STR1}=\theta_{STR2}=1$). In another example, allele excess is only weakly correlated with an analogous measure for SNPs (Tajima's D), and statistical support for these associations fades as the difference in mutation rates grows.

Several factors probably contribute to the weakness of these correlations. First, STR summary statistics have large sampling variances [18], sometimes poorly reflecting underlying genealogies. The increased uncertainty surrounding STR polymorphism estimates relative to SNP estimates comes from the extra dimension of STR variation (repeat length) and the contribution of multiple sites to SNP polymorphism measures. Second, linked SNPs and STRs evolve independently because of contrasting mutational dynamics, despite sharing the same historical pattern of inheritance. This line of reasoning suggests that alternative STR mutational models might produce different results. As a preliminary investigation of this idea, we conducted simulations with the parameters in Table 1, but using a two-phase mutation model [19]. In these simulations, a fraction (f) of mutations resulted in single-step changes, whereas the remaining mutations ($1-f$) changed allele size by two or more repeats, with repeat size drawn from a geometric distribution with probability of success 0.5. Using values of f ranging from 0.7 to 0.95, observed correlations between STR and SNP diversity are generally stronger than those seen under the strict stepwise mutation model. These results might reflect a reduction in STR homoplasy caused by a greater fraction of unique alleles (relative to the strict stepwise mutation model), illustrating the importance of assumptions about the STR mutational process in interpreting patterns of covariation with SNP loci and indicating the need for investigation of additional mutational models.

Mining joint patterns of SNP and STR diversity

We now describe specific examples to illustrate how understanding relationships between polymorphism at SNPs and STRs can be helpful for several problems in population genetics.

Nonequilibrium processes

Nonequilibrium processes alter the genealogy at a locus, leading to changes in patterns of polymorphism. For example, population bottlenecks and directional selection both reduce diversity, yet different measures of diversity are affected in disparate ways, allowing inference about these events by comparing polymorphism indices [20–23]. Because mutations at linked SNPs and STRs fall along the same genealogy, relationships between patterns of variation at these markers should also contain information about bottlenecks and selective events. To test this idea, we simulated the recovery of V and θ_π ($\theta_{STR}=10$, $\theta_{SNP}=1$) following a ten-times reduction in population size that lasted 4N generations (Figure 1). Both V and θ_π recovered from the bottleneck in similar ways in terms of average diversity (the greater STR mutation rate generates a faster rate of recovery, but mutation–drift equilibrium is also further away; Figure 1a,b). However, the correlation between V and θ_π changed with the passage of time, decreasing soon after the end of the bottleneck and then rising slowly to its equilibrium level (Figure 1c). Similar results were obtained using H instead of V (not shown). These patterns suggest that joint functions of diversity statistics at linked SNPs and STRs might provide new information for characterizing population size changes. Combining information from linked SNPs and STRs could also help detect other departures from equilibrium, including population structure and selection on linked loci.

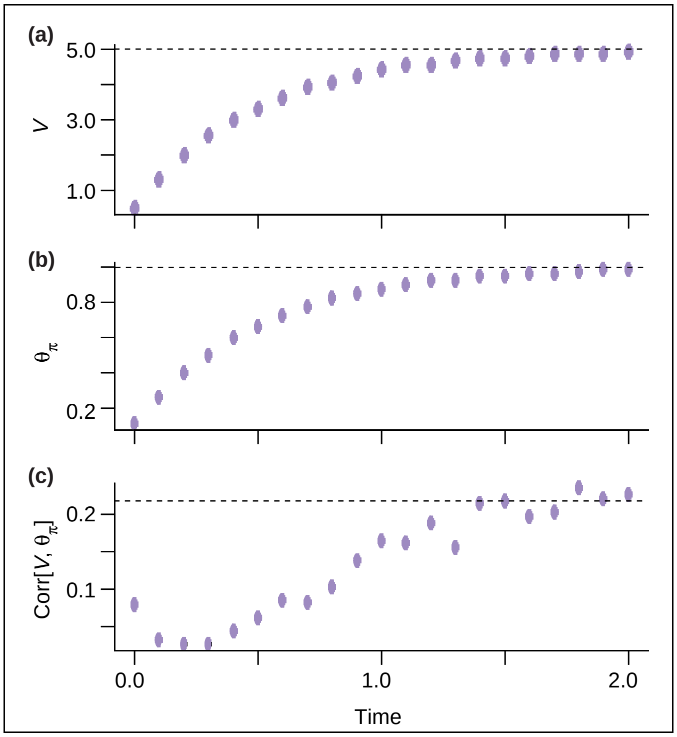

Figure 1. Joint recovery of SNP and STR diversity levels from a population bottleneck. The population size is decreased ten times for 4*N* generations and then allowed to recover to the initial size. The STR mutation rate ($\theta_{\mathrm{STR}}=10$) is an order of magnitude greater than that for SNPs ($\theta_{\mathrm{SNP}}=1$). Time (on the *x*-axis) is measured in units of 4*N* generations following the bottleneck. **(a)** Variance in allele size, *V*, at an STR. **(b)** Average number of pairwise differences, θ_π, at a SNP locus. **(c)** Correlation between *V* and θ_π at linked STR-SNP pairs. Each point in (a) and (b) is an average calculated from 10 000 simulated samples of 50 chromosomes. Each point in (c) is the estimated correlation across the 10 000 simulated samples. The dashed lines indicate points of mutation–drift equilibrium.

Identifying genomic regions that show extreme patterns

Recent advances in high-throughput genotyping technology have enabled whole-genome scans for associations between genotypic and phenotypic variation in groups of unrelated individuals [24,25]; several marker attributes affect the power and false-positive rate of these tests [26]. Although considerable attention has focused on the expected replicability of associations discovered using different sets of SNPs, little is known about expected results for contrasting classes of markers [27,28]. Similar issues arise in the context of genomic scans for natural selection, when the identification of unusual (potentially selected) genomic regions using one marker type is often followed by more detailed surveys of these regions using the other marker type [29].

Does the identification of outlier loci in one genomic scan require adjustment of the interpretation of subsequent surveys of polymorphism at linked markers? As an example, we simulated 1000 linked pairs of STR and SNP loci under neutrality ($\theta_{\mathrm{STR}}=10$, $\theta_{\mathrm{SNP}}=1$), identified the STR at the 1% quantile in the distribution of *V*, and then asked where θ_π at the linked SNP locus fell within the distribution of θ_π values. We repeated this experiment 1000 times and recorded the distribution of θ_π values and quantiles from across genomic scans. In this example, the average θ_π was 0.80, a value significantly less than the expected value of 1.0 for randomly chosen SNP loci ($t_{999}=12.65$; $P<10^{-15}$). The average quantile was 33% (compared with the expected 50%). This example illustrates two factors that affect replicability of genomic scans. First, successive scans with markers placed in the same genomic regions are not statistically independent. Shared genealogical history causes loci linked to those identified as unusual in an initial scan to behave as outliers in a second effort, even under the null model of no locus-specific forces. Second, as in previous examples, some of this nonindependence is ameliorated by the different mutational processes at STRs and SNPs. Similar calculations can be used to adjust significance thresholds for formal tests of genotype–phenotype association and tests of neutrality.

Estimating rates of mutation

Mutation rates at STRs vary widely across species, genomes and repeat types [30,31], with good estimates often coming from studies of pedigrees or mutation-accumulation experiments [2,3]. Estimating STR mutation rates from species divergence in allele size is generally unreliable because of homoplasy; mutation rates can instead be inferred from diversities at STR loci by assuming mutation–drift equilibrium and a particular value of *N* (because diversity is a function of *N* and the neutral mutation rate). However, *N* also is difficult to estimate, and can effectively vary from one locus to another as a result of selection at linked sites [32]. We consider an alternative approach that exploits a SNP locus completely linked to a STR. At the SNP locus, $\theta_\pi=4N\mu_{\mathrm{SNP}}$, and at the STR, $V=2N\mu_{\mathrm{STR}}$, suggesting that

$$\mu_{STR}=\frac{2\mu_{SNP}V}{\theta_\pi} \qquad \text{[Eqn 1]}$$

Thus, estimation of the SNP mutation rate (μ_{SNP}) by comparison with a sequence from a closely related species, which usually yields accurate results, can be used to obtain an absolute estimate of the STR mutation rate (μ_{STR}). Although the performance of this *ad hoc* estimator needs to be tested, such an approach does not depend on *N*, suggesting that it might be robust to departures from mutation–drift equilibrium.

Understanding the effects of selection on a genomic scale

Theory shows that selection on beneficial or deleterious mutations can erode linked, neutral variation [33,34]. Empirical evidence for such effects can be seen in the correlation between nucleotide polymorphism and recombination rate, a pattern now known to characterize several species [35,36]. Because the removal of deleterious mutations and linked neutral variants (‘background selection’) is an equilibrium process, the expected reduction in variation is independent of the neutral mutation rate, suggesting that the pattern should be visible at both SNPs and STRs [37–39]. By contrast, levels of diversity at markers with high mutation rates are less affected by positive selection on linked alleles, which could lead to different patterns for SNPs and STRs [37,40]. Hence, comparisons of variation at linked SNPs and STRs can help gauge the relative importance of these two selective processes.

Measuring linkage disequilibrium

Several linkage disequilibrium measures are available for markers with two segregating alleles (such as SNPs). Comparable measures for loci harboring more variants (such as STRs) are more difficult to construct and interpret. Pritchard and Feldman [41] suggested that correlations in V could be used to estimate linkage disequilibrium between STR loci. Linkage disequilibria and correlations between polymorphism levels both measure the covariance in coalescence times at different markers [41–43]. Therefore, correlations between polymorphism statistics at SNPs and STRs might also provide a useful metric of association. The statistical significance of these correlations could be assessed using resampling methods [41]. Importantly, such measures would not require inference of haplotypes, a challenging exercise that contributes its own statistical uncertainty (especially for markers harboring many alleles).

Future research

Building a general theoretical framework for covariation between polymorphism at SNPs and STRs will require detailed consideration of several variables. Although genomic sequences have facilitated the rapid identification of neighboring SNPs and STRs, the effects of recombination will presumably weaken all correlations and need to be evaluated. In addition, sources of mutational covariation between SNPs and STRs (such as local base composition) should be investigated. Perhaps the most significant challenge to constructing the desired framework is accurate modeling of the STR mutational process. Here we have primarily focused on a mutational model that provides a reasonable fit to many aspects of STR variation [13,44,45], but contributions of additional processes, including non-stepwise mutations, allele-size-dependent mutation rates, and range constraints, could also be important [46,47]. In terms of characterizing genomic patterns, it will also be useful to ascertain the influence of heterogeneity in mutation rates. For example, if variable STR mutation rates are permitted, then expected correlations between measures of SNP and STR diversity could be smaller than shown in our simulations. Inference might benefit from comparisons among populations, in which the effects of interlocus variation in mutation rates can be muted [48]. Finally, extension of likelihood and Bayesian procedures that use polymorphism data more efficiently than summary statistics to incorporate information from SNPs and STRs simultaneously [9] is an exciting avenue for future work.

Acknowledgements

We are grateful to Scott Williamson for suggesting the simulation scheme. We thank Aida M. Andrés, Deborah Charlesworth and three anonymous reviewers for comments on the manuscript. We thank Chip Aquadro, Jeff Jensen, Rasmus Nielsen, Razeesh Sainudiin, Kevin Thornton, John Wakeley and Travis Wheeler for helpful discussions. B.A.P. is supported by start-up funds from the University of Wisconsin Medical School. A.D.C. is supported by NSF International Research Fellowship Program grant #0401897.

References

1 Henderson, S.T. and Petes, T.D. (1992) Instability of simple sequence DNA in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 12, 2749–2757

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

3 Schug, M.D. *et al*. (1998) The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol. Biol. Evol.* 15, 1751–1760

4 Dallas, J.F. (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mamm. Genome* 3, 452–456

5 Estoup, A. *et al*. (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* 11, 1591–1604

6 Ramakrishnan, U. and Mountain, J.L. (2004) Precision and accuracy of divergence time estimates from STR and SNPSTR variation. *Mol. Biol. Evol.* 21, 1960–1971

7 Zegura, S.L. *et al*. (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol. Biol. Evol.* 21, 164–175

8 Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158, 885–896

9 Hey, J. *et al*. (2004) Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Mol. Ecol.* 13, 909–919

10 Tishkoff, S.A. *et al*. (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293, 455–462

11 Mountain, J.L. *et al*. (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res.* 12, 1766–1772

12 Kimura, M. and Crow, J.F. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49, 725–738

13 Valdes, A.M. *et al*. (1993) Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133, 737–749

14 Hudson, R.R. (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18, 337–338

15 Ohta, T. and Kimura, M. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22, 201–204

16 Moran, P.A. (1975) Wandering distributions and the electrophoretic profile. *Theor. Popul. Biol.* 8, 318–330

17 Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276

18 Zhivotovsky, L.A. and Feldman, M.W. (1995) Microsatellite variability and genetic distances. *Proc. Natl. Acad. Sci. U. S. A.* 92, 11549–11552

19 Di Rienzo, A. *et al*. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. U. S. A.* 91, 3166–3170

20 Kimura, M. and Ohta, T. (1975) Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc. Natl. Acad. Sci. U. S. A.* 72, 2761–2764

21 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595

22 Fu, Y.X. and Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* 133, 693–709

23 Kimmel, M. *et al*. (1998) Signatures of population expansion in microsatellite repeat data. *Genetics* 148, 1921–1930

24 Hinds, D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079

25 Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108

26 Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* 5, 89–100

27 Evans, D.M. and Cardon, L.R. (2004) Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am. J. Hum. Genet.* 75, 687–692

28 Schaid, D.J. *et al*. (2004) Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci. *Am. J. Hum. Genet.* 75, 948–965

29 Harr, B. *et al*. (2002) Hitchhiking mapping: A population-based fine-mapping strategy for adaptive mutations in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12949–12954

30 Chakraborty, R. *et al*. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1041–1046

31 Neff, B.D. and Gross, M.R. (2001) Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. *Evolution Int. J. Org. Evolution* 55, 1717–1733

32 Hill, W.G. and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res.* 8, 269–294

33 Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35

34 Charlesworth, B. *et al*. (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303

35 Begun, D.J. and Aquadro, C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356, 519–520

36 Cutter, A.D. and Payseur, B.A. (2003) Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.* 20, 665–673

37 Slatkin, M. (1995) Hitchhiking and associative overdominance at a microsatellite locus. *Mol. Biol. Evol.* 12, 473–480

38 Schug, M.D. *et al*. (1998) Mutation and evolution of microsatellites in *Drosophila melanogaster*. *Genetica* 102-103, 359–367

39 Payseur, B.A. and Nachman, M.W. (2000) Microsatellite variation and recombination rate in the human genome. *Genetics* 156, 1285–1298

40 Wiehe, T. (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theor. Popul. Biol.* 53, 272–283

41 Pritchard, J.K. and Feldman, M.W. (1996) Statistics for microsatellite variation based on coalescence. *Theor. Popul. Biol.* 50, 325–344

42 McVean, G.A. (2002) A genealogical interpretation of linkage disequilibrium. *Genetics* 162, 987–991

43 Nordborg, M. and Tavare, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18, 83–90

44 Shriver, M.D. *et al*. (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* 134, 983–993

45 Banchs, I. *et al*. (1994) New alleles at microsatellite loci in CEPH families mainly arise from somatic mutations in the lymphoblastoid cell lines. *Hum. Mutat.* 3, 365–372

46 Ellegren, H. (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* 24, 400–402

47 Sainudiin, R. *et al*. (2004) Microsatellite mutation models: insights from a comparison of humans and chimpanzees. *Genetics* 168, 383–395

48 Schlötterer, C. (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160, 753–763

49 Nei, M. and Li, W.H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273

50 Hudson, R.R. (1990) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (Futuyma, D.J. and Antonovics, J., eds), pp. 1–44, Oxford University Press