Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005

Accurate Inference and Estimation in Population Genomics

*Matthew W. Hahn**†*

*Center for Population Biology, University of California, Davis; and †Department of Biology and School of Informatics, Indiana University, Bloomington

Both intra- and interspecific genomic comparisons have revealed local similarities in the level and frequency of mutational variation, as well as in patterns of gene expression. This autocorrelation between measurements leads to violations of assumptions of independence in many statistical methods, resulting in misleading and incorrect inferences. Here I show that autocorrelation can be due to many factors and is present across the genome. Using a one-dimensional spatial stochastic model, I further show how previous results can be employed to correct for autocorrelation along chromosomes in population and comparative genomics research. When multiple hypothesis tests are autocorrelated, I demonstrate that a simple correction can lead to increased power in statistical inference. I present a preliminary analysis of population genomic data from *Drosophila simulans* to show the ubiquity of autocorrelation and applicability of the methods proposed here.

Introduction

One of the major goals of molecular population genetics is to account for the various forces affecting nucleotide variation within and between species. Drift, selection, mutation, and demographic processes can all play important roles in determining the number and frequency of DNA mutations, but determining the relative contribution of each process at any single locus can be challenging (e.g., Tajima 1989; Braverman et al. 1995; Tenaillon et al. 2001; Hahn, Rausher, and Cunningham 2002; Akey et al. 2004; Schmid et al. 2005; Stajich and Hahn 2005). As molecular population genetics has proceeded from the first single-locus study (Kreitman 1983) to two-locus (Hudson, Kreitman, and Aguade 1987), dozen-locus (Begun and Aquadro 1992), and 100-locus (Glinka et al. 2003) studies, researchers have been better able to distinguish the genome-wide effects of drift and demography from the locus- or region-specific effects of natural selection. As sequencing technologies become faster and more affordable, population genetic data sets comprising huge numbers of loci will offer the opportunity to study these evolutionary processes at a genomic scale (population genomics).

Despite this increase in sequence data, the amount of information gained in going from one to even a few loci outweighs that gained in going from hundreds to thousands of loci. This is because the more dense sampling of loci means that many data points are no longer independent: there is an autocorrelation (or "serial correlation") between linked loci in levels of polymorphism, divergence, the allele frequency spectrum, and various summary statistics of variation, even over very large distances (e.g., Tishkoff et al. 1996; Reich et al. 2002; Falush, Stephens, and Pritchard 2003; Gaffney and Keightley 2005; Hinds et al. 2005; Stajich and Hahn 2005). There may be many reasons for this spatial autocorrelation, including shared histories, variation in underlying mutation rates or neutral mutation rates (e.g., in gene-rich or gene-poor regions), linked selection, and demographic events. There is also an autocorrelation in the expression levels of neighboring genes (Hurst, Pal, and Lercher 2004), possibly due to autocorrelated levels of polymorphism (e.g., Kliebenstein et al. 2005) or shared regulatory elements. The sampling of nonindependent loci can mean that patterns present in only a fraction of the genome will be pseudoreplicated and that statistics assuming independent observations will overestimate the true number of independent loci. This leads to underestimated standard errors (SEs) and confidence intervals (CIs) and hence tests that are too liberal (Lehmann 1986). Autocorrelation therefore results in problems with statistical inference and estimation in genomic studies.

Certain data analysis methods—such as sliding window analyses—can also lead to spatial autocorrelation as the data in each window is not independent of its neighbors. This form of autocorrelation is due to statistical nonindependence between measurements, rather than the biological nonindependence induced by shared underlying parameters. Throughout this paper, I refer solely to biological nonindependence, though many of the same problems and solutions may arise when data are autocorrelated due to both biological and statistical causes.

Many situations in population and comparative genomics are affected by an overestimation of the number of independent measurements or simply the assumption that all the data are independent. Here are four scenarios where these effects may occur. First, in tests of the mean between groups, autocorrelated data result in underestimated SEs. Researchers often test for differences in levels of polymorphism between populations (e.g., Andolfatto 2001; Glinka et al. 2003) or in levels of polymorphism or divergence between genes on sex chromosomes and autosomes (e.g., Andolfatto 2001; Betancourt, Presgraves, and Swanson 2002). Such differences provide information about the relative importance of demographic history and natural selection, as well as about genome-wide differences in the efficacy of selection. However, the sampling of autocorrelated loci may lead to misleading results regarding differences between populations or chromosomes. Even tests that are not explicitly spatial—such as comparisons in evolutionary rate among genes in various functional categories—may be affected by autocorrelation because of the frequent occurrence of tandem duplication

Key words: autocorrelation, time series, association study, comparative genomics, natural selection.

E-mail: mwh@indiana.edu.

Mol. Biol. Evol. 23(5):911–918. 2006
doi:10.1093/molbev/msj094
Advance Access publication January 11, 2006

(Friedman and Hughes 2001) and blocks of associated genes (Hurst, Pal, and Lercher 2004) in the genomes of most organisms. Second, evaluating relationships among causal factors may be affected by nonindependent measurements. Regressing measures of variation on factors such as recombination rate (itself an autocorrelated measure; Kong et al. 2002), GC content, or underlying mutation rate may give misleading results as to the causes of patterns of polymorphism and divergence. While ordinary least squares methods are unbiased estimators of the slope of a regression, they underestimate the SE when either one or both of the two variables are themselves autocorrelated (Zeger, Irizarry, and Peng 2004). Third, in cases where one wishes to construct a null model for testing the effects of various evolutionary processes, autocorrelation can lead to biased parameter estimates. For instance, Ometto et al. (2005) generated a null distribution for an out-of-Africa population bottleneck in *Drosophila melanogaster* using 250 noncoding loci on the X chromosome. They identified loci that fell outside of various CIs and were presumably affected by natural selection. However, autocorrelation between sampled loci may cause incorrect identification of the targets of selection because of incorrect parameterization of the null. Finally, any analysis in which a simple distribution of summary statistics is constructed may be misled by autocorrelation between the nonindependent measurements. Looking for outlying loci or for more loci in the tails of the distribution than are expected by chance may result in erroneous inference under the assumption of independence among loci (e.g., Akey et al. 2004; Stajich and Hahn 2005). This occurs because significant results present in a small subset of the data can be pseudoreplicated when autocorrelated observations are made. More significant loci may be found than are expected at random as multiple, autocorrelated loci are counted separately.

Problems with autocorrelated data are an inevitable function of finite genome sizes and will become more common as data sets grow. In this paper, therefore, I present a model that takes into account autocorrelation and allows for more accurate inferences in genomic studies. Using a stochastic model of one-dimensional spatial autocorrelation, I show how previous results can be employed to more accurately estimate the true number of independent observations and, as a consequence, to more accurately estimate the SE in hypothesis tests. I also show that if hypothesis tests on individual loci are not wholly independent, then typical multiplicity correction procedures produce a conservative test and a corresponding loss of power. I present a preliminary analysis of data from light-shotgun genome sequencing of six lines of *Drosophila simulans* (C. Langley and D. Begun, personal communication) in order to show both the ubiquity of autocorrelation in population genomic data sets and the applicability of the methods presented to correct for this autocorrelation.

Materials and Methods

Estimating Autocorrelation

In order to model the autocorrelation present in genomic data, we treat a chromosome as a one-dimensional space with data measured at points along its length. These data points can be equally spaced genes or noncoding loci or in nonoverlapping windows covering the whole chromosome. There are many well-developed methods for the analysis of this type of one-dimensional spatial autocorrelation, generally falling under the statistical field of "time series" (e.g., Shumway 1988; Chatfield 1989; Diggle 1990; Box, Jenkins, and Reinsel 1994; Fuller 1996). However, as I outline below, there is no concept of time or directionality implied in the analyses presented here, and all results are equally valid—as long as stationarity assumptions are met—no matter which end of the chromosome we start from.

For data where successive measurements are correlated (the definition of autocorrelation), a powerful and popular approach is to regress these successive measurements on one another. Such models are called autoregressive and are preferable for most genomic data because they result in correlations that decay gradually rather than steeply: the correlation between measurements i steps apart is ρ^i (where ρ is the autocorrelation coefficient defined in the next paragraph). Other models for autocorrelated data exist (see, e.g., Chatfield 1989) but will not be considered in this first attempt at applying this class of probability models to genomic data. Autoregressive processes are not restricted to regressing neighboring observations—higher order processes that consider regressions at any lag (i.e., number of steps apart) are possible. I largely consider first-order processes here because they are well studied and discuss how to decide whether models of higher order are appropriate below.

Under a first-order autoregressive process, the value of the ith observation is given by

$$X_i = \mu + \rho(X_{i-1} - \mu) + \varepsilon_i, \quad |\rho| \leq 1,\ i = 1, 2, \ldots, \quad (1)$$

where parameter μ is the mean of the series of measures, ρ is the autocorrelation at lag 1 (the only correlation considered in models of order 1), and ε is the normally distributed noise term with mean zero and variance σ^2. The parameters of this model can be fit by a number of maximum likelihood estimation procedures (Chatfield 1989). Significance of the autocorrelation parameter (against the null hypothesis that it is equal to zero) can be determined via permutation. A number of approaches exist for choosing the appropriate order of the autoregressive model used, including informal approaches using the partial autocorrelation function and more formal approaches that penalize models with greater numbers of parameters via the Akaike information criterion (Box, Jenkins, and Reinsel 1994).

Assumptions of Autoregressive Processes

Many of the results in time series analysis are based on the assumption that the data are stationary, though methods specific to nonstationary data are available. While the mathematical definition of stationarity is beyond the scope of this paper (see, e.g., Chatfield 1989; Fuller 1996), a few of the requirements of the stationarity assumption are relevant to the analysis of genomic data. First, there is no trend to the data. If there is a trend in the average value or the variance of the observations across the series, transformation

of the data is often recommended in order to make the series stationary (Chatfield 1989; Box, Jenkins, and Reinsel 1994). Second, the correlation between successive measures is constant across the series. This implies that the correlation between X_i and X_{i-1} or X_i and X_{i+1} is the same as it is between any two other neighboring observations. The use of equally spaced observations in a spatial series is largely due to this assumption: if observations are not regular, then the correlations between them can vary. There are a number of ways of dealing with irregularly spaced observations, including sampling of the data to ensure equal spacing and fitting a spline to the data to fill-in missing measurements (Chatfield 1989, p. 199). While sampling will result in a loss of data, fitting a spline assumes an autocorrelated nature to the data and so may not be conservative when testing for the presence of autocorrelation. If both the above assumptions hold, we can see that it does not matter which end of a chromosome we start our series with as there should be no directionality to the data. In the analysis presented below, I do find that the initial data violate assumption 1, largely due to loci at the ends of the chromosome. I removed these data prior to the full analysis to meet both stationarity assumptions.

Correcting for Autocorrelation

For first-order autoregressive processes, there is a simple relationship that can be used to correct for autocorrelation. The number of true independent observations (n^*; also referred to as the "effective sample size") for an autocorrelated series of n measures is given by Dawdy and Matalas (1964):

$$n^* = n[(1 - \rho)/(1 + \rho)] \quad (2)$$

This formula implies that as the autocorrelation goes to zero n^* equals n and that as it goes to unity (complete autocorrelation) there is only one independent observation. Because genomic data are generally positively autocorrelated ($\rho > 0$), such a correction will lower the number of observations in a data set ($n^* < n$) and will increase the SE and CIs.

Drosophila simulans Data

One of the first genome-scale population genetic data sets comes from the shotgun sequencing of six inbred lines of *D. simulans* (Langley et al., personal communication). Each of five lines was sequenced to 1.5× coverage, while the sixth was sequenced to 3× coverage. This means that the whole genome is not covered for any single line and that sequence coverage at any given base has an average of approximately three alleles (Langley et al., personal communication). Additionally, as part of the same project, one inbred strain of *Drosophila yakuba* was sequenced to 10× coverage; using this and the *D. melanogaster* genome sequence (Adams et al. 2000) allow us to estimate divergence on only the lineage leading to *D. simulans*. For the data presented here, only approximately 5 Mb in the middle of the X chromosome was analyzed for both polymorphism and divergence. In order to meet stationarity assumptions, I avoided using the ends of the chromosome as both the mean and variance in polymorphism were lower due to reduced recombination (Langley et al., personal communication).

In the following, I show how to use the framework of spatial autocorrelation to better describe genomic data from *D. simulans* to correct for the autocorrelation inherent in many data sets and to uncover the processes responsible for this autocorrelation.

Results

Data from 10-kb nonoverlapping windows, each required to have at least three alleles across the majority (>50%) of the window, were examined for autocorrelation in π (the average number of nucleotide differences per base; Tajima 1983). Over the length of the 5-Mb sequence, there is an autocorrelation between neighboring windows (fig. 1*a*), with the autocorrelation parameter ρ of a first-order autoregressive process estimated as 0.51 ($P < 1.0 \times 10^{-15}$; all analyses were performed in R [www.r-project.org]). If we examine the decay in autocorrelation measured between windows at lags from 0 and 30 (fig. 2*a*), we see an approximately exponential decline with increasing distance as predicted for first-order autoregressive models. There is an autocorrelation of 1 at lag 0 (as every window is completely autocorrelated with itself), an autocorrelation of 0.51 at lag 1, and decreasing autocorrelations at greater lags. Much of the longer range autocorrelation observed is simply due to the intervening short-range autocorrelations: using a partial autocorrelation analysis (not shown), there is no significant autocorrelation effect at lags greater than 1 after taking into account the autocorrelation at lag 1. While this does not mean that there is no dependence in the data at distances greater than 10 kb (see below), it does support the use of a first-order autoregressive process rather than models of higher order.

Examining the autocorrelation in divergence across the same 10-kb windows, we see a similar pattern (fig. 1*c*): a relatively high value of ρ estimated across the sequence (0.38; $P < 1.0 \times 10^{-15}$). The similarity in patterns of autocorrelation is not surprising as many of the same forces responsible for autocorrelation in polymorphism also affect divergence. Indeed, there is a correlation between levels of polymorphism and divergence for this region (Pearson's $r = 0.44$; $P < 2.2 \times 10^{-16}$), as expected under the neutral theory of molecular evolution (Kimura 1968). Autocorrelation appears to be consistently higher for polymorphism than for divergence, most likely because demographic processes and linked selection affect polymorphism but not divergence (Birky and Walsh 1988). Nonetheless, it can be seen from the polymorphism and divergence data that there is a long-range dependence in both measures, with measured correlation present over distances upward of 300 kb for π and almost 100 kb for divergence (fig. 2*a* and *b*).

One question raised by the above analyses is how to determine the most appropriate window size to use. Choosing a window size can be arbitrary (even a "day" or a "month" is arbitrary in standard time series analysis), and depending on the scale of the expected effect different window sizes may be appropriate. To examine the consistency

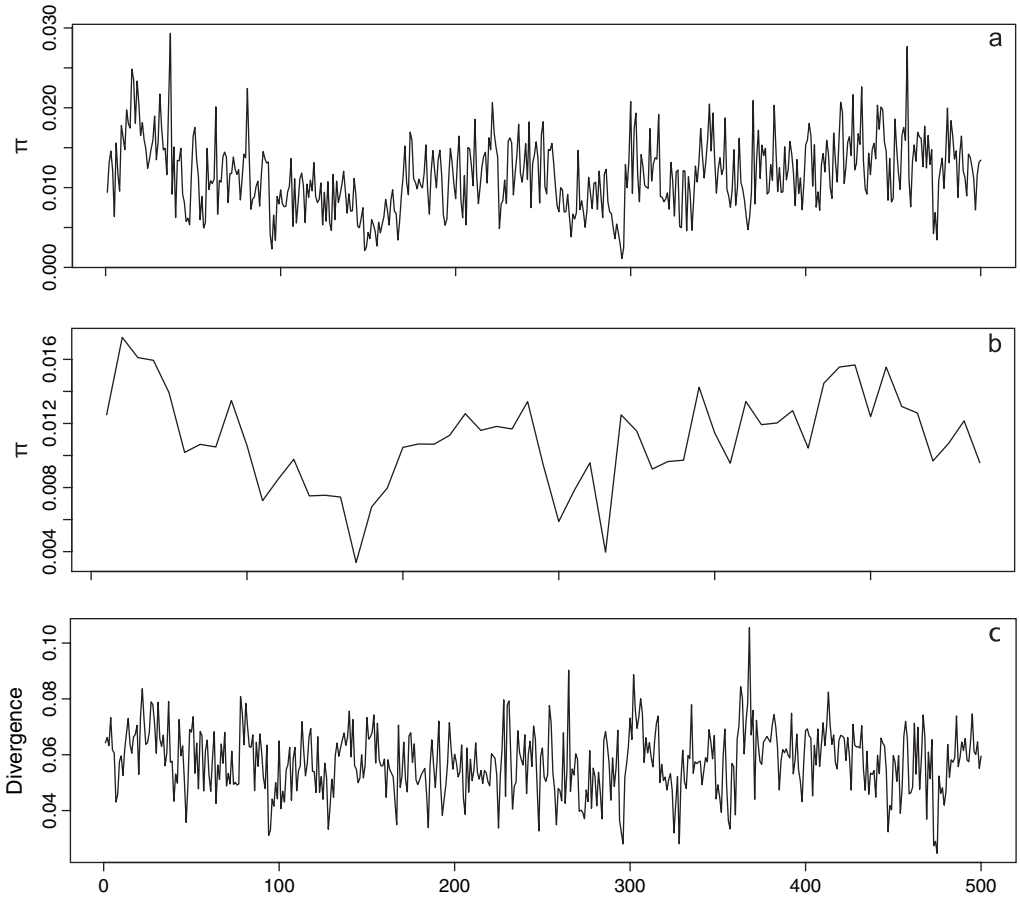

FIG. 1.—Levels of polymorphism and divergence along 5 Mb of the X chromosome. (*a*) Polymorphism measured in 500 nonoverlapping 10-kb windows. (*b*) Polymorphism measured in 50 nonoverlapping 100-kb windows. (*c*) Divergence measured in 500 nonoverlapping 10-kb windows. These windows correspond to the same windows in (*a*).

of the observed autocorrelation in polymorphism, I repeated the same analysis across the 5 Mb of X chromosome with nonoverlapping windows of 100 kb (fig. 1*b*). Again there was a significant autocorrelation in levels of polymorphism, with $\rho = 0.59$ ($P < 1.0 \times 10^{-4}$), which implies that measurement of autocorrelation is not very sensitive to window size. However, this result still does not indicate the appropriate window size to use: choosing large window sizes may overwhelm small-scale effects, while choosing small window sizes may obscure larger effects by introducing noise.

As discussed above, autocorrelation in both polymorphism and divergence will cause us to overestimate the true number of biologically independent samples taken. Using equation (2), we can correct for this overestimation, taking into account autocorrelation to estimate the true number of independent measurements. For example, the standard error of the mean is given by $\sigma/\sqrt{n}$, where σ is the standard deviation and n is the number of measurements. Taking a random stretch of 20 consecutive 10-kb windows of π from the *D. simulans* data used here, the mean is 0.016. Correcting for autocorrelation using $\rho = 0.51$ gives $n^* = 6.5$. Plugging in this adjusted number of measurements results in a SE of 0.0019 rather than 0.0011 without correction, an increase of almost 100% in the error around the estimate of the mean. The simple calculation presented in equation (2) leads to an adjustment in the sample size and, consequently, more accurate hypothesis testing and inference. There are many instances in genomic studies where this correction will help to better quantify the amount of evidence in favor of competing hypotheses.

Discussion

One of the most important assumptions of standard statistical analyses is that multiple observations are independent of one another. Confidence in estimates of the mean and other parameters is often expressed as a function of the number of independent observations obtained, and hypothesis tests are often based on the error around these estimates (Lehmann 1986). When data are autocorrelated, however, this independence assumption is violated—there may be fewer truly independent observations than is believed. Overcounting of observations will then lead to an inflated confidence in parameter estimates, evident as underestimation of the SE about the mean or of CIs. The result is overly permissive hypothesis testing and false rejection of the null (Lehmann 1986).

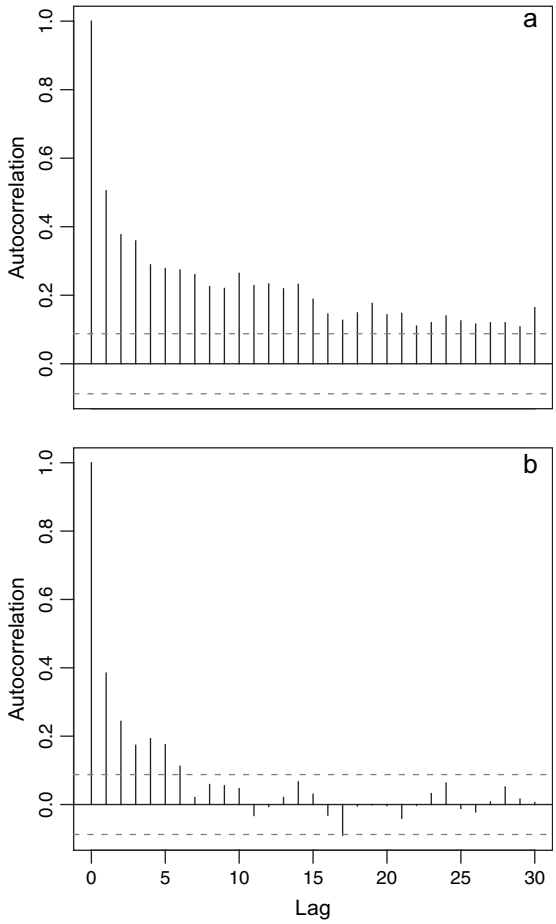

FIG. 2.—Autocorrelation in polymorphism and divergence. (*a*) Autocorrelation in polymorphism, measured in 10-kb windows. The autocorrelation at lag 0 is by definition equal to 1. Dashed lines represent the 95% CIs around an autocorrelation coefficient equal to zero. (*b*) Autocorrelation in divergence, measured in 10-kb windows.

It is clear that many aspects of genomic data show autocorrelation (e.g., Tishkoff et al. 1996; Reich et al. 2002; Gaffney and Keightley 2005; Hinds et al. 2005; Stajich and Hahn 2005). In order to take full advantage of the information contained within thousands of measurements, we therefore need to control for the autocorrelated nature of these observations. How can we both correct for autocorrelation and estimate the actual number of independent observations? There are a number of ways of correcting for the bias introduced by spatial autocorrelation, as addressed previously in fields like ecology, hydrology, and climatology (e.g., Legendre 1993). Here I proposed using a method that explicitly estimates the true number of independent observations (eq. 2). While this approach does not obviate all of the problems introduced by autocorrelated data (such as higher order dependencies), it does help to correct for problems in a common class of analyses. As an alternative to the approach advocated here, data can be selected or sampled to maximize the additional information they provide. Data spaced farther apart or completely beyond the range of autocorrelation will contain more information per observation than densely spaced data. In this way the problems of autocorrelation may be avoided, though at the cost of much information. The results and models used in this paper are clearly just a first attempt at resolving a very complex problem.

Adjusting for Multiple Hypothesis Tests

While dependence among measurements can result in overly liberal hypothesis tests and spurious significance, it may also have a conservative effect on hypothesis testing. This effect occurs because of the manner in which one corrects for multiple hypothesis tests: when multiple tests are carried out, significance levels must be adjusted to account for the increased probability of type I errors (false positives). Generally, Bonferroni or other corrections are made such that α, the family-wise error rate, is adjusted downward for the number of tests: $\alpha^* = \alpha/n$ (Sokal and Rohlf 1995). The adjusted value, α^*, is then used as the level below which P values must fall for tests to be considered significant (Sokal and Rohlf 1995). However, if tests are autocorrelated, then correcting for an overinflated number of apparently independent tests will make this P value cutoff too low, resulting in conservative hypothesis testing (e.g., McIntyre et al. 2000; Nyholt 2004; Stajich and Hahn 2005).

Problems with overcorrecting for autocorrelated hypothesis tests can occur in many situations in population genomics, only two of which I outline here. First, autocorrelation in test statistics such as Tajima's D (Tajima 1989) may result in conservative hypothesis tests. Because of selective sweeps, various demographic scenarios, or simply shared histories between neighboring loci, the mutation frequency spectrum (and consequently Tajima's D) can be similar over large regions of the genome. If one is calculating Tajima's D for many loci in order to find unusually evolving genes or regions, autocorrelation can result in testing the same or similar data many times, thereby reducing power. Second, association studies aimed at mapping loci responsible for quantitative variation (Lander and Schork 1994; Risch and Merikangas 1996) may be affected by autocorrelation. In association studies, multiple markers may be in linkage disequilibrium with one another, resulting in autocorrelated tests. Indeed, in their seminal paper on interval mapping of quantitative trait loci, Lander and Botstein (1989) used a diffusion process to model and correct for the autocorrelation between significant, closely spaced markers. When large, genome-wide association studies are conducted using hundreds or thousands of markers, the problem of multiple testing is exacerbated because of the large number of tests performed (Hirschhorn and Daly 2005). Even in smaller scale association studies, however, such as those that aim to find the functional variant responsible for variation at a single locus, multiple markers may be tested that are autocorrelated with one another (e.g., Genissel et al. 2004). This will be a larger problem in humans because of the large distances (and therefore increased number of alleles) that must be examined to find functional regulatory variants (Rockman and Wray 2002).

One way to adjust for the actual number of independent tests (i.e., measures of a test statistic) performed would be to use the correction provided by first-order autoregressive

processes outlined above (eq. 2). Such a correction would account for the repeated nature of autocorrelated tests and ensures that family-wise error rates are not overly conservative. It would do so by adjusting Bonferroni calculations so that $\alpha^* = \alpha/n^*$. (This calculation would have to be made chromosome by chromosome in order to meet the assumptions of one-dimensional models.) Unless repeated measures are negatively autocorrelated (an unlikely scenario in genomic data), Bonferroni corrections calculated in this way result in a higher value for α^* and less stringent P value cutoffs. For instance, if 100 tests are conducted with an autocorrelation of $\rho = 0.5$ between tests, then the adjusted P value for a nominal 0.05% false positive rate is 0.0015 (=0.05/33), according to the method proposed here, rather than 0.0005 when considering the tests to be independent. This is a threefold increase in the P value that must be reached for tests to be considered significant (see Cheverud 2001 and Nyholt 2004 for alternative approaches to such a correction).

Further Uses for Spatial Stochastic Processes

The use of stochastic models for spatial series suggests a number of additional approaches to the analysis of population genomic data. Recent interest in finding targets of adaptive evolution has focused on carrying out so-called genomic scans of selection on large data sets (reviewed in Storz 2005). The general approach taken is to either use a sliding window analysis along each chromosome or to simply look for outlying loci within the whole genome. While both of these methods may suffer from problems due to autocorrelation among loci, they are also statistically unsatisfying because of the many heuristics that must be employed (for instance, in the size of the window and the size of the step taken when sliding the window). We recently proposed (Turner, Hahn, and Nuzhdin 2005) using population genetic hidden Markov models (PopGenHMMs; cf. Felsenstein and Churchill 1996; Siepel and Haussler 2004) to find regions of interest regardless of predetermined window size. Similar approaches have been used to define regions of high linkage disequilibrium in humans (Daly et al. 2001; Falush, Stephens, and Pritchard 2003) and to find regulatory sequences evolving under constraint in yeast (Chin, Chaung, and Li 2005). PopGenHMMs can be used to determine windows that are significantly different from the surrounding sequence by any number of summary statistics or per-nucleotide measurements and may suggest a natural scale for the size of windows used in other analyses (A. Kern and M. Hahn, personal communication).

Even when regions with extreme patterns of polymorphism or divergence are identified, however, we lack a framework that provides a genome-wide expectation of the size and frequency of such regions under the hypothesis of neutral evolution. Spatial stochastic processes may provide the framework needed by explicitly modeling the rise and fall of nucleotide variability along a chromosome. Results on the "exceedance" of a series can be used to give a distribution of the expected number and size of regions exceeding (or falling below) particular values of a statistic or measure (e.g., Leadbetter 1995). This is analogous to a series going above or below some threshold level in polymorphism or divergence in its random walk along a chromosome. Such distributions could be used as a null against which the observed number of such regions—for example, stretches with $\pi = 0$—can be tested. Exceedance analysis and population genetic HMMs are just two ways in which spatial stochastic models may be extended to deal with population genomic data.

Conclusions

Autocorrelation between neighboring loci can lead to violations of the assumption of independent sampling and, as a consequence, to invalid or misleading inferences in genomic studies. While the autocorrelation due to shared histories on nonrecombining chromosomes such as the mitochondrion has long been recognized, I have shown here that dependencies between loci can be due to many biological factors and can be present across the genome. Treating a chromosome as a one-dimensional space allows for the use of well-known results in stochastic processes, as well as extensions of these previous results to problems unique to genomics. Though I have stressed the importance of correcting for autocorrelation in population-based studies, the ideas presented here are just as valid for interspecific comparisons or single-genome studies. As data sets grow and autocorrelation is recognized in many spheres, the analysis of genomic data will come hard upon the unforeseen—and possibly unimagined—problem of too many loci.

Acknowledgments

The author would like to thank R. Shumway for clarifying many points in time series analysis, L. Moyle for helpful suggestions on the substance and presentation of the research, and D. Begun, J. Gillespie, A. Kern, C. Langley, J. Mezey, M. Rockman, K. Simonsen, M. Turelli, and W. Wilson for comments, suggestions, and encouragement. Two anonymous reviewers helped to greatly improve the manuscript. I am also indebted to M. Wayne, L. Katz, and the Society for Molecular Biology and Evolution for providing this forum and to C. Langley, D. Begun, K. Stevens, A. Kern, and the Washington University Genome Sequencing Center for making the *D. simulans* data available and usable.

Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt et al. (192 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol. **2**:e286.

Andolfatto, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18**:279–290.

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

Betancourt, A. J., D. C. Presgraves, and W. J. Swanson. 2002. A test for faster X evolution in Drosophila. Mol. Biol. Evol. **19**:1816–1819.

Birky, C. W. Jr., and J. B. Walsh. 1988. Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA **85**:6414–6418.

Box, G. E. P., G. M. Jenkins, and G. Reinsel. 1994. Time series analysis: forecasting & control. Prentice Hall, Englewood Cliffs, N.J.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140**:783–796.

Chatfield, C. 1989. The analysis of time series. Chapman & Hall, London.

Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. Heredity **87**:52–58.

Chin, C. S., J. H. Chuang, and H. Li. 2005. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res. **15**:205–213.

Daly, M. J., J. D. Rioux, S. E. Schaffner, T. J. Hudson, and E. S. Lander. 2001. High-resolution haplotype structure in the human genome. Nat. Genet. **29**:229–232.

Dawdy, D. R., and N. C. Matalas. 1964. Statistical and probability analysis of hydrologic data, part III: analysis of variance, covariance and time series. Pp. 68–90 *in* V. T. Chow, ed. Handbook of applied hydrology, a compendium of water-resources technology. McGraw-Hill Book Company, New York.

Diggle, P. 1990. Time series: a biostatistical introduction. Clarendon Press, Oxford.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics **164**:1567–1587.

Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

Friedman, R., and A. L. Hughes. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. **11**: 373–381.

Fuller, W. A. 1996. Introduction to statistical time series. Wiley-Interscience, New York.

Gaffney, D. J., and P. D. Keightley. 2005. The scale of mutational variation in the murid genome. Genome Res. **15**: 1086–1094.

Genissel, A., T. Pastinen, A. Dowell, T. F. C. Mackay, and A. D. Long. 2004. No evidence for an association between common nonsynonymous polymorphisms in Delta and bristle number variation in natural and laboratory populations of *Drosophila melanogaster*. Genetics **166**:291–306.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165**:1269–1278.

Hahn, M. W., M. D. Rausher, and C. W. Cunningham. 2002. Distinguishing between selection and population expansion in an experimental lineage of bacteriophage T7. Genetics **161**: 11–20.

Hinds, D. A., L. L. Stuve, G. B. Nilsen, E. Halperin, E. Eskin, D. G. Ballinger, K. A. Frazer, and D. R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. Science **307**:1072–1079.

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

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Hurst, L. D., C. Pal, and M. J. Lercher. 2004. The evolutionary dynamics of gene order. Nat. Rev. Genet. **5**:299–310.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624–626.

Kliebenstein, D. J., M. A. L. West, H. van Leeuwen, K. Kim, R. W. Doerge, R. W. Michelmore, and D. A. St.Clair. 2005. Genomic survey of gene expression diversity in *Arabidopsis thaliana*. Genetics 10.1534.

Kong, A., D. F. Gudbjartsson, J. Sainz et al. (13 co-authors). 2002. A high-resolution recombination map of the human genome. Nat. Genet. **31**:241–247.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304**:412–417.

Lander, E. S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**:185–199.

Lander, E. S., and N. J. Schork. 1994. Genetic dissection of complex traits. Science **265**:2037–2048.

Leadbetter, M. R. 1995. On high level exceedance modeling and tail inference. J. Stat. Plan. Inference **45**:247–260.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? Ecology **74**:1659–1673.

Lehmann, E. L. 1986. Testing statistical hypotheses. Springer-Verlag, New York.

McIntyre, L. M., E. R. Martin, K. L. Simonsen, and N. L. Kaplan. 2000. Circumventing multiple testing: a multilocus Monte Carlo approach to testing for association. Genet. Epidemiol. **19**:18–29.

Nyholt, D. R. 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am. J. Hum. Genet. **74**:765–769.

Ometto, L., S. Glinka, D. De Lorenzo, and W. Stephan. 2005. Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. Mol. Biol. Evol. **22**:2119–2130.

Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. Nat. Genet. **32**:135–142.

Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science **273**:1516–1517.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for *cis*-regulatory evolution in humans. Mol. Biol. Evol. **19**:1991–2004.

Schmid, K. J., S. Ramos-Onsins, H. Ringys-Beckstein, B. Weisshaar, and T. Mitchell-Olds. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics **169**:1601–1615.

Shumway, R. H. 1988. Applied statistical time series analysis. Prentice Hall, Englewood Cliffs, N.J.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**:468–488.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. W.H. Freeman, New York.

Stajich, J. E., and M. W. Hahn. 2005. Disentangling the effects of demography and selection in human history. Mol. Biol. Evol. **22**:63–73.

Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol. Ecol. **14**:671–688.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

———. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–595.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp *mays* L.). Proc. Natl. Acad. Sci. USA **98**:9161–9166.

Tishkoff, S. A., E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonne-Tamir, A. S. Santachiara-Benerecetti, P. Moral, and M. Krings. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271**:1380–1387.

Turner, T. L., M. W. Hahn, and S. V. Nuzhdin. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. **3**:e285.

Zeger, S., R. Irizarry, and R. Peng. 2004. On time series analysis of public health and biomedical data. Johns Hopkins Department of Biostatistics Working Paper Series No. 1054.

Marta Wayne, Associate Editor

Accepted December 30, 2005