Population genetic inference from genomic sequence variation


Genome Res. 2010 20: 291-300 originally published online January 12, 2010
Access the most recent version at doi:10.1101/gr.079509.108

References
This article cites 147 articles, 41 of which can be accessed free at:
http://genome.cshlp.org/content/20/3/291.full.html#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to Genome Research go to:
http://genome.cshlp.org/subscriptions
Population genetic inference from genomic sequence variation

John E. Pool,1,2 Ines Hellmann,3 Jeffrey D. Jensen,1,5 and Rasmus Nielsen1,4,6

1Department of Integrative Biology, University of California, Berkeley, Berkeley, California 94720, USA; 2Center for Population Biology, University of California, Davis, Davis, California 95616, USA; 3Mathematics and Biosciences Group, Max F. Perutz Laboratories, Vienna 1030, Austria; 4Department of Statistics, University of California, Berkeley, Berkeley, California 94720, USA

Population genetics has evolved from a theory-driven field with little empirical data into a data-driven discipline in which genome-scale data sets test the limits of available models and computational analysis methods. In humans and a few model organisms, analyses of whole-genome sequence polymorphism data are currently under way. And in light of the falling costs of next-generation sequencing technologies, such studies will soon become common in many other organisms as well. Here, we assess the challenges to analyzing whole-genome sequence polymorphism data and discuss the potential of these data to yield new insights concerning population history and the genomic prevalence of natural selection.

Population genetics originated in the first half of the 20th century as a field driven by theoretical insights but with very limited empirical data, and for several decades theory remained well ahead of the data available to test its predictions. This situation began to change with the emergence of protein electrophoretic variation (e.g., Harris 1966; Hubby and Lewontin 1966; Lewontin and Hubby 1966; Lewontin 1972). Since the introduction of polymerase chain reaction (PCR) technology, the scale of data has grown exponentially, as restriction fragment length polymorphisms, microsatellites, and small-scale DNA sequencing (e.g., Kreitman 1983) broadened the range of questions open to empirical investigation. With the recent flood of genome-wide single nucleotide polymorphism (SNP) data, and now the advent of fully sequenced population samples of genomes, population genetics has become a fundamentally data-driven discipline.

As the data-generating capacity of population genetics has grown, so has its importance in related disciplines. Population genetics is now at the core of analyses in molecular ecology and conservation biology, where it provides a framework for understanding the distribution of genetic variability among populations and for inferring the demographic histories of natural populations from molecular data. It is also central in studies of molecular evolution, providing a foundation for understanding the contributions of mutation, genetic drift, and natural selection in the evolution of genes and genomes. Finally, with the focus in human genetics on association mapping (Lander and Schork 1994; Risch and Merikangas 1996; Pritchard et al. 2000a), admixture genetic drift, and natural selection in the evolution of genes and genomes. Finally, with the focus in human genetics on association mapping (Lander and Schork 1994; Risch and Merikangas 1996; Pritchard et al. 2000a), admixture mapping (Chakraborty and Weiss 1988; Stephens et al. 1994), relatedness mapping (Cheung and Nelson 1998; Albrechtsen et al. 2009), and related techniques, population genetics has found its way into medical genetics as a core analytical discipline.

Currently, large-scale next-generation sequencing projects are moving forward in a number of organisms including humans, Drosophila, and Arabidopsis. Before the availability of such data, several genome-wide studies have been completed using Sanger sequencing (e.g., Bustamante et al. 2005; Begun et al. 2007) or SNP genotyping (Hinds et al. 2005; The International HapMap Consortium 2005, 2007; Jakobsson et al. 2008; JZ Li et al. 2008). The low-coverage sequencing of six Drosophila simulans genomes by Begun et al. (2007) was an important step forward for population genomics, and yet today one Illumina Genome Analyzer run can produce substantially more data than were present in that study. This expanded data-generating capacity has led to the recent public release of more than 40 Drosophila melanogaster genomes (http://www.dpgp.org), along with the recent published analysis of 40 silkworm genomes (Xia et al. 2009).

The challenges associated with SNP data obtained by genotyping (particularly ascertainment bias) have been discussed extensively elsewhere (e.g., Kuhner et al. 2000; Nielsen 2000, 2004; Marth et al. 2004) and will not be a focus of this review. Instead, we focus on the analysis of next-generation sequencing data, which is likely to be the foundation of many future population genomic studies. Analysis of these data is currently in its infancy. And yet, if the cost of next-generation sequencing continues to decline, genome-wide population genetic data will likely be available not only for humans and the main model organisms, but for most organisms on which active research is being carried out in genetics, ecology, or evolution. Our ability to obtain samples and to propose good biological questions will be the limiting factor—instead of the sequencing costs. In the anticipation of this future, we review some of the fundamental issues relating to the analysis of genome-wide population genetic data.

Next-generation sequencing

Large-scale sequencing (for review, see Shendure and Ji 2008) is now possible using platforms such as Illumina sequencing (Bentley et al. 2008), 454 Life Sciences (Roche) pyrosequencing (Margulies et al. 2005), Applied Biosystems SOLiD sequencing (Fu et al. 2008), and cPAL sequencing (Drmanac et al. 2009). The declining cost of generating such data is transforming the field of population genetics, making large genomic data sets available to most researchers. While the technology has hitherto mostly been used by researchers working on humans and the main model organisms, next-generation sequencing is also emerging as an economical alternative to other methods for generating population genetic data from natural populations of other organisms. Various reduced-representation shotgun sequencing (RRSS) techniques can be used to select a subset of the genome for sequencing (Altschuler et al. 2000; Baird...
et al. 2008). When combined with techniques for labeling reads (e.g., Meyer et al. 2008), so that DNA from many individuals can be analyzed in the same pooled sequencing reaction, RRSS using next-generation sequencing provides an increasingly affordable means for generating population genetic data. Next-generation sequencing is therefore likely to become the standard choice for generating population genetic data in fields such as conservation genetics and molecular ecology, but it will carry new demands for computational infrastructure and statistical and bioinformatics training. While next-generation sequencing may not erase every advantage of genetic model organisms, it can allow for the construction of a genetic map (giving regional estimates of recombination rate) by collecting sequence data from laboratory crosses (Baird et al. 2008) or related wild-caught individuals. This strategy implies an added investment of resources, but knowledge about recombination rates is critical for many population genetic inferences (e.g., Thornton and Andolfatto 2006; Becquet and Przeworski 2007; O’Reilly et al. 2008; Pool and Nielsen 2009). Recombination is not a principal focus of this article, as it has been reviewed elsewhere (Coop and Przeworski 2007).

The special nature of the data produced by next-generation sequencing platforms may entail a new set of challenges for unbiased estimation of population genetic parameters. In contrast to traditional approaches, where a defined fragment is amplified by PCR and then sequenced, sequence reads from next-generation technologies stem from individual DNA molecules and are distributed across the genome in a largely random fashion (although regions with very high or very low GC content may be underrepresented) (Ossowski et al. 2008). Data produced by these technologies are most comparable to single-pass whole-genome shotgun sequences, which suffer from three basic problems: sequence errors, assembly errors, and missing data. The severity of these problems will depend in part on the depth of sequencing, with higher coverage potentially minimizing many errors (e.g., Bentley et al. 2008). But for organisms with large genomes, the trade-off of coverage versus cost and sample size may justify dealing with the statistical complexities of low-coverage data sets (at least until further sequencing improvements and/or cost reductions are achieved). This trade-off may also depend on specific research goals (e.g., the optimal coverage for a study focused on linkage disequilibrium might be higher than for a study based on allele frequencies), but further work is needed to inform this aspect of experimental design.

Sequence errors

Because next-generation sequence reads originate from a single DNA molecule, errors in the sequences can be due to DNA damage, errors introduced during amplification, and sequencing errors. The stage at which errors occur will determine the frequency of that error in the sequenced DNA pool. While it might be assumed that erroneous bases will occur on single reads only, evidence of non-random errors has been reported (Keightley et al. 2009), and so a statistical analysis of error probabilities will be important even for high-coverage data sets. If unaccounted for, errors will inflate nucleotide diversity and skew the allele frequency spectrum (AFS) toward rare alleles, which will mainly be visible as an excess of singletons (e.g., Johnson and Slatkin 2008).

Thus far, the processing of sequence data and especially the calling of SNPs has been focused on minimizing the false-positive rate, by introducing stringent quality criteria to call SNPs (e.g., Altshuler et al. 2000). Johnson and Slatkin (2006, 2008) noted that stringent SNP-calling criteria will bias diversity estimates by excluding many true SNPs (especially rare alleles) from the data. Therefore, they suggested incorporating quality values directly into the estimation of diversity instead of only using them as a pre-filter. However, this is only possible if the probability of a sequencing error is a known function of the sequence quality value. This relationship has been thoroughly investigated for ABI-Sanger sequencing (Ewing and Green 1998), but is currently much less clear for next-generation sequencing methods. Empirically validated error models for new sequencing platforms that incorporate sequence context and position within reads could improve the correlation between quality scores and error probabilities.

Once error probabilities can be estimated accurately, it is relatively easy to correct for the presence of sequencing errors statistically (e.g., Hellmann et al. 2008; Jiang et al. 2009). Lynch (2008) described a method to estimate error rates and nucleotide diversity in a mixed procedure, where the error rate and nucleotide diversity are first estimated from sites with high coverage using a maximum likelihood approach and then used in a method of moments estimation of nucleotide diversity across the genome. Lynch (2009) then extended this approach to also correct the AFS for missing data and errors, assuming either Hardy-Weinberg equilibrium or a known inbreeding coefficient. While the above methods focus on basic population genetic inferences at the genome-wide level, in the future they might be generalized to more complex demographic models or adapted to search for localized changes in diversity or allele frequencies.

Assembly errors

Next-generation sequence reads are hitherto shorter than in traditional Sanger sequencing (presently up to ~75 bp for Illumina, and up to ~450 bp for 454 Life Sciences sequencing), and this poses serious challenges for assembling reads (e.g., Sundquist et al. 2007; Chaisson and Pevzner 2008; Zerbino and Birney 2008; Bryant et al. 2009), as well as mapping reads to a reference genome (e.g., H Li et al. 2008; R Li et al. 2008; Langmead et al. 2009). These problems can be partially ameliorated via "paired-end" sequencing, which involves short sequence reads on each side of DNA fragments of a particular size class. However, assembly remains challenging in repetitive or highly polymorphic genomic regions, and it is worthwhile to consider the potential biases that imperfect assembly may introduce.

For some mapping algorithms, sequence reads with more than one or two differences from a reference genome will not be placed (e.g., H Li et al. 2008). This makes the mapping of alleles that are different from the reference genome less probable than for a reference-matching allele, causing a bias in allele frequency toward the allele found in the reference sequence. It may additionally reduce the number of SNPs discovered and bias estimates of nucleotide diversity toward smaller values. Moreover, if the reference genome itself is a consensus genome from multiple individuals, this approach will skew the AFS toward high-frequency alleles. The issue of reference sequence bias could be addressed via alignment tools that are more robust to polymorphism, and by incorporating known polymorphisms and their frequencies into the reference sequence. Assembly should ideally take into account the locations of transposable elements in the reference genome (many of which may not exist in other individuals), and allow for indel variation in general.

In the case of ambiguous placements it is common practice to discard those reads. Hence, repetitive and duplicated regions may
have lower coverage. Finally, erroneous alignments of paralogous sequences will inflate nucleotide diversity and could push the AFS toward intermediate frequency alleles. Improved assembly and mapping remain very important and active areas of research, but the most significant improvement to assembly may come from sequencing technology: longer read lengths, and also “paired-end” reads that collect data from each end of fragments of a particular size class. Importantly for population genomic studies, these same advances will increase the haplotype information that can be empirically determined from diploid samples (Bansal et al. 2008; Kidd et al. 2008; Long et al. 2009), along with facilitating the identification of genome rearrangements (Korbel et al. 2007), including copy number variants.

**Missing data**

Another challenge for the analysis of whole-genome sequence polymorphism is missing data. Due to the stochastic placement of sequence reads across the genome, the sampled chromosomes at any particular site may not include all individuals (Figure 1). And unless all samples are sequenced at very high genomic coverage (i.e., >30×) (Bentley et al. 2008), it may not be clear whether both of a diploid individual’s alleles have been sequenced. Sample sizes will therefore vary along the chromosome and will not be known with certainty. This uncertainty increases if the identity of the individual from which a read was sampled is unknown (i.e., for pooled samples) and decreases with coverage per individual. Ignoring missing data will introduce biases in the estimation of population genetic parameters. However, this problem can be circumvented by summing over all possible (unknown) chromosome sample sizes (Hellmann et al. 2008; Lynch 2008; Jiang et al. 2009).

In association studies it is common practice to impute missing data from the surrounding haplotype patterns (Marchini et al. 2007; Servin and Stephens 2007). This technique could be useful if the goal is to identify putative disease causing SNPs. However, imputation is likely to introduce bias in most population genetic analyses. For example, since singleton polymorphisms cannot be imputed, the use of imputation would lead to downwardly biased nucleotide diversity estimates and a bias against singletons in the AFS. Additional bias may result if the sampled alleles represent only a subset of the population’s haplotype diversity (as found for human “tag-SNPs” by Bhangale et al. [2008]).

Next-generation sequencing technologies are evolving with great speed, but the development of appropriate analysis tools is lagging behind. It takes time to characterize the occurrences of sequencing errors and biases with respect to nucleotide content (for example) and then develop appropriate estimators that take such problems into account. Because population genetic inferences are particularly susceptible to sequencing errors and missing data, researchers who use next-generation sequencing data for inferences about demography and selection should always keep these problems in mind. Fortunately, most of the bias introduced by sequencing errors and missing data can be mitigated using appropriate statistical corrections.

**Prospects for demographic inference from whole-genome sequence polymorphism**

Inference of population history is a central aim of population genetic studies, whether this knowledge is sought for its own sake or to strengthen the conclusions of genome-wide scans for positive selection or genotype–phenotype associations. Currently, demographic analysis of genome-wide SNP data sets often focuses on clustering methods that assign individuals’ genomes to one or more populations, or methods that analyze genetic distances between individuals and/or populations (e.g., Jakobsson et al. 2008; JZ Li et al. 2008; Novembre et al. 2008). In some sense, such studies are less ambitious than some traditional methods based on a single or a few loci (e.g., Kuhner et al. 1998; Nielsen and Wakeley 2001; Beaumont et al. 2002) in that rather than estimating demographic parameters directly, they merely aim to quantify the relationship between individuals without a population genetic model or an explicit demographic context. Methods that do infer population parameters from large data sets often focus on the AFS or the genomic means of summary statistics and their variances across the genome. However, many uniquely informative aspects of genome-wide data—such as long-range haplotype patterns—have not been fully utilized. Analysis of whole-genome sequence polymorphism is clearly no less computationally intensive, but compared to SNP data, its advantages for demographic inference include better haplotype information, inclusion of rare population- and region-specific variants, and an unbiased AFS.

**Historical inference from allele frequencies and summary statistics**

One of the simplest ways to summarize population genetic data is via the AFS. Examples of the use of SNP allele frequency data for demographic inference are provided by Nielsen (2000), Wooding and Rogers (2002), Polanski and Kimmel (2003), Marth et al. (2004), and Williamson et al. (2005), who all modeled the expected AFS under different models of changing population sizes. These methods can also be applied to more than one population and more complex demographic models using the so-called multidimensional frequency spectrum (e.g., Caicedo et al. 2007; Gutenkunst et al. 2009; Nielsen et al. 2009). Although some of the early analyses were limited to a relatively small data set, inference based on the AFS is also computationally tractable for larger analyses. For example, Williamson et al. (2005) used a genome-wide data set of directly sequenced human protein-coding regions. However, while the AFS does contain significant information about past changes in population size, it fails to capture much of the relevant information from population genetic data (such as haplotype structure and variance across the genome), and it may not contain sufficient information for historical inference in more complicated models (Adams and Hudson 2004; Myers et al. 2008).

![Figure 1](image-url). Reads from different individuals are aligned to a reference genome, and SNPs have been called. In this toy example there are seven true segregating sites (black dots) and four false ones (red dots). Also note that segments differ in sample size. In segment 4, for example, five reads from three individuals were sampled; thus, there were at least three chromosomes and at most five sampled in this segment.
Several studies have used multiple statistics to compare empirical data against simulations with varying demographic histories. For instance, Schaffner et al. (2005) used several summary statistics (based on allele frequencies, linkage disequilibrium, and population differentiation) to jointly infer historical and recombination models for human populations. Voight et al. (2005) and Thornton and Andolfatto (2006) each used three different statistics to fit population bottleneck models for non-African populations of humans and *D. melanogaster*, respectively. Examining a different type of model—that of a population split with subsequent migration—Becquet and Przeworski (2007) used numbers of shared variants between populations, private alleles, and fixed differences to estimate demographic parameters for ape populations.

In addition to using different summary statistics, the studies cited above illustrate different methods for comparing summary statistics from empirical and simulated data, including a root mean squared error approach (Schaffner et al. 2005), combining summary statistic P-values (Voight et al. 2005), an approximate Bayesian rejection sampling approach (Thornton and Andolfatto 2006), and an approximate Bayesian Markov chain Monte Carlo likelihood approach (Becquet and Przeworski 2007). None of these methods were applied to genome-scale polymorphism data, and certainly one key to their potential scalability will be computational efficiency. Another issue is the transition from short, independent loci to full genomic coverage. At the simplest, this could be achieved by slicing chromosomes into mostly independent windows of some arbitrary length; but preferably, analyses should account for the nonindependent nature of sequence variation by statistically correcting for the effect of autocorrelation on *P*-values (Hahn 2006) and confidence intervals (e.g., Keinan et al. 2007).

When genome-scale polymorphism data are available, historical inference can be improved by accounting for both autosomal and X-linked patterns of diversity. The X chromosome will typically have a different effective population size than the autosomes, and will thus operate on a different population genetic time scale. Because the X chromosome will therefore be affected differently by events such as population size changes (Fay and Wu 1999; Hey and Harris 1999; Wall et al. 2002; Pool and Nielsen 2007), it represents a complementary source of information for demographic inference. For example, although a bottleneck model that of a population split with subsequent migration—Becquet and Przeworski (2007) used numbers of shared variants between populations, private alleles, and fixed differences to estimate demographic parameters for ape populations.

Population structure and historical inference from haplotypes

One goal of population genetic analysis is to identify the genetic structure that exists within a set of genotyped individuals, which may give insight into population relationships and help to minimize false-positive results in association mapping studies. Principle components analysis (PCA) was introduced to population genetics more than 30 yr ago (Menozzi et al. 1978) but experienced renewed interest following its implementation by Patterson et al. (2006) in a form allowing statistical validation of inferred structure. The computational tractability of PCA makes it applicable to large data sets, as demonstrated by Novembre et al. (2008), who found that principle components inferred from genome-wide SNP data essentially reconstructed the geographic map of Europe. However, interpretation of principle components in terms of population history is far from clear (Novembre and Stephens 2008). PCA is therefore typically a first analysis aimed at defining the genetic relationships among groups.

Population structure can also be analyzed using clustering methods such as STRUCTURE (Pritchard et al. 2000b; Falush et al. 2003). STRUCTURE is relatively computationally intensive, and care must be taken to verify that results have converged, but it has been applied to fairly large data sets. Faster-converging MCMC methods for analyzing genetic structure are now available (Huelskenz and Andolfatto 2007; Corander et al. 2008; Alexander et al. 2009). Jakobsson et al. (2008) applied STRUCTURE to more than 500,000 SNPs in worldwide human populations. Supporting the demographic utility of linkage information, this study found that haplotypes were far more likely than individual SNPs to be geographically region-specific, and STRUCTURE analysis of haplotypes enabled detection of additional genetic structure within Africa.

The linkage model of STRUCTURE (Falush et al. 2003) uses “admixture linkage disequilibrium” to estimate ancestry along chromosomes, and a recent method (Price et al. 2009) accounts for local linkage disequilibrium as well. This type of information opens up new possibilities for demographic inference, as demonstrated by methods that infer both ancestry along chromosomes and parameters relevant to recent admixture history (e.g., Hoggart et al. 2004; Patterson et al. 2004), and by a method that uses the lengths of migrant DNA tracts to test for a recent change in migration rate (Pool and Nielsen 2009). By extension, methods that infer genomic tracts of relatedness between individuals (e.g., Purcell et al. 2007; Albrechtsen et al. 2009; Gusev et al. 2009) may also provide relevant information for inferring recent demographic events.

Hellenthal et al. (2008) also used linkage patterns to infer population relationships, implementing an approach based on the copying model of Li and Stephens (2003) to estimate the ancestry sources of human populations. Rather than directly modeling the ancestry process that gives rise to haplotypes along recombining chromosomes, the copying model (also referred to as the “product of approximate conditionals” or the PAC likelihood model) builds samples sequentially by copying segments of existing chromosomes. A second demographic application of the PAC model is provided by Davison et al. (2009), who used it to estimate parameters of a population split model. Because it does not deal with the complexity of ancestral recombination graphs, the copying model is computationally much faster than coalescent-based approaches with recombination. However, the need to correct parameter estimates obtained by this approach (Davison et al. 2009) emphasizes that the PAC model is an approximation that may have significant differences from the true evolutionary process.

The Davison et al. (2009) study also illustrates that linkage patterns carry historic information beyond recent migration events. A second example is provided by Lohmueller et al. (2009), who used the joint distribution of haplotype number and major haplotype frequency in empirical and simulated data to estimate population size changes from human SNP data. In addition, Plagnol and Wall (2006) used linked clusters of mutations to detect signals of archaic structure in human populations. Thus, while
long-range haplotype patterns carry unique information about the history of recent migration, short-range haplotype patterns can be strong signals of more ancient gene flow and other demographic events. In light of these studies, the haplotype information provided by next-generation sequencing data will offer a significant advantage over SNP data for detecting historical population events and fine population structure. The ability to detect rare population- or region-specific polymorphisms (which will often be missed in SNP studies) may also improve such inferences.

A final illustration of the potential demographic informativeness of haplotype patterns is shown in Box 1. The particular enrichment of long haplotypes shared between European and African humans could reflect a relatively high rate of recent migration between continents. However, we point out that haplotype patterns, like all population genetic summaries, are potentially influenced by other evolutionary processes such as selection and recombination. Some progress has been made in jointly analyzing natural selection and population history (e.g., Williamson et al. 2005; Wright et al. 2005; Li and Stephan 2006), but the development of realistic evolutionary models for population genomic analysis remains largely an unsolved problem.

Identifying locus-specific and genome-wide effects of selection

One of the most exciting prospects of whole-genome polymorphism data is the increased power to characterize not only the recent adaptive history of natural populations, but also the genomic prevalence of positive and negative natural selection. Negative selection reduces variation in the genome by eliminating some mutations, holding others to low frequency, and also causing the loss of variants linked to deleterious alleles (background selection) (Charlesworth et al. 1993). Positive selection leads to local reductions in genetic diversity via the “genetic hitchhiking” effect of Smith and Haigh (1974). As a favorable mutation increases in frequency in a population, linked neutral variants will either become fixed along with it or be lost from the population. The size of the region of the genome affected by such a “selective sweep” is determined mainly by the strength of selection and the rate of recombination (Smith and Haigh 1974; Hudson and Kaplan 1988; Stephan et al. 1992).

A large literature has arisen characterizing the expected polymorphism patterns resulting from selective sweeps—ranging from a deficit of variation and an excess of rare alleles around the selected site (Hudson and Kaplan 1988; Tajima 1989; Braverman et al. 1995; Fu 1997), to an excess of high-frequency derived alleles in flanking regions (Fay and Wu 2000), to effects on linkage disequilibrium (e.g., Przeworski 2002; Kim and Nielsen 2004; McVean 2007). These signals have been incorporated into methods that scan population genomic data for loci affected by recent selective sweeps. For example, several studies (e.g., Carlson et al. 2005; Williamson et al. 2007; Nielsen et al. 2009) have used the distribution of human SNP frequencies along chromosomes to scan for completed sweeps. Whole-genome sequence polymorphism data should include many rare SNPs absent from previous data sets, and may thus increase the power of these methods to detect selection.

The improved haplotype information of next-generation sequencing data will also augment efforts to detect selection. Selective sweeps produce a distinct spatial pattern of linkage disequilibrium (Stephan et al. 2006) that may represent a unique signal of hitchhiking as opposed to stochastic patterns from population bottlenecks (for example, see Jensen et al. 2007). Linkage patterns can also provide a clear signal of partial selective sweeps, based on the imbalance of haplotype homozygosity between a favored allele class and other variants in the same sample (Sabeti et al. 2002; Voight et al. 2006). By comparing haplotype homozygosity between samples, this approach can also identify population-specific selective sweeps (Sabeti et al. 2007).

The addition of interspecies divergence data to polymorphism within species can allow detection of recurrent selective fixations. For example, comparison of polymorphism and divergence at synonymous versus nonsynonymous sites (McDonald and Kreitman 1991) has been used to identify coding sequences subject to recurrent positive selection (e.g., Bustamante et al. 2005) and to establish the importance of regulatory sequences in adaptive evolution (e.g., Andolfatto 2005). The future availability of genome-wide polymorphism data from multiple closely related species will expand the range of possible analyses and improve our basic understanding of molecular evolution.

Characterizing genomic parameters of adaptation

While many studies have identified specific loci with evidence for positive selection (reviewed extensively elsewhere; e.g., Nielsen et al. 2007; Kelley and Swanson 2008), it is increasingly possible to analyze genome-wide signals of hitchhiking. One example is the correlation between recombination and diversity, which was originally observed in D. melanogaster (Begun and Aquadro 1992), and suggested the influence of linked selection. While such a correlation could result from “background selection” against linked deleterious variation (Charlesworth et al. 1993), subsequent analyses have favored the genetic hitchhiking model as a primary explanation (Andolfatto and Przeworski 2001; Innan and Stephan 2003). Hellmann et al. (2008) recently verified that this correlation exists for human data (beyond the effect of mutation rate differences), and likewise found that a hitchhiking model fit the data best. However, the human and Drosophila correlations are of strikingly different magnitudes (Fig. 2), which may reflect the

**Figure 2.** The relationship between recombination rate and nucleotide diversity in data from D. melanogaster and humans. The data points shown represent average nucleotide diversity for recombination rate bins. The D. melanogaster data (Shapiro et al. 2007) are from 349 loci with \( \approx 50 \) synonymous sites sequenced in up to 15 African lines (as analyzed by Sella et al. 2009). The human data are from the whole-genome shotgun sequence data analyzed by Hellmann et al. (2008), analyzed in 100-kb windows and processed as described in that study. Human nucleotide diversity is corrected for differences in interspecific divergence (to account for differences in mutation rate) (Hellmann et al. 2003). *Drosophila* diversity is not corrected for divergence—the correlation between recombination and divergence was weakly positive for this data set (Sella et al. 2009). For both data sets, recombination rate bins were defined manually based on data availability and log-scale bin width.
An underutilized evolutionary signal in whole-genome diversity data is the frequency of long shared haplotypes. Localized excess of long shared haplotypes has been used to identify targets of positive selection (e.g., Sabeti et al. 2002), but the genome-wide abundance of long identical tracts shared across population boundaries may also shed light on recent rates of gene flow. This is similar to the logic underlying methods that infer admixture parameters (e.g., Falush et al. 2003) or changes in migration rate (Pool and Nielsen 2009) based on the sizes of introgressed chromosomal segments, except that here no inference of population ancestry along chromosomes is required.

To examine this pattern in the human genome, we compared the long shared haplotypes found within and between the African and European HapMap populations (SNP data with known phasing from HapMap release 23) against that predicted from simulations from the demographic and recombination model estimated by Schaffner et al. (2005), using the coalescent simulation program COSI. Data were simulated for 10,000 regions of length 1 Mb. Ascertainment correction was made using the two-dimensional frequency spectrum (for both populations together), by retaining variable sites from the simulated data with a probability equal to the ratio of the per-base-pair frequencies of the 2D frequency class between the HapMap and unfiltered simulated data. Because the Schaffner et al. (2005) model uses regional recombination rates drawn from the genetic map of Kong et al. (2002), only HapMap data within the bounds of this map were included, and centromeric regions were excluded. Singletons (polymorphisms observed in only one allele) were excluded from both data sets. Finally, to conservatively eliminate regions of low SNP coverage from the HapMap data, gaps of >10 kb between non-singleton SNPs were excluded from the analyzed regions.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

An underutilized evolutionary signal in whole-genome diversity data is the frequency of long shared haplotypes. Localized excess of long shared haplotypes has been used to identify targets of positive selection (e.g., Sabeti et al. 2002), but the genome-wide abundance of long identical tracts shared across population boundaries may also shed light on recent rates of gene flow. This is similar to the logic underlying methods that infer admixture parameters (e.g., Falush et al. 2003) or changes in migration rate (Pool and Nielsen 2009) based on the sizes of introgressed chromosomal segments, except that here no inference of population ancestry along chromosomes is required.

To examine this pattern in the human genome, we compared the long shared haplotypes found within and between the African and European HapMap populations (SNP data with known phasing from HapMap release 23) against that predicted from simulations from the demographic and recombination model estimated by Schaffner et al. (2005), using the coalescent simulation program COSI. Data were simulated for 10,000 regions of length 1 Mb. Ascertainment correction was made using the two-dimensional frequency spectrum (for both populations together), by retaining variable sites from the simulated data with a probability equal to the ratio of the per-base-pair frequencies of the 2D frequency class between the HapMap and unfiltered simulated data. Because the Schaffner et al. (2005) model uses regional recombination rates drawn from the genetic map of Kong et al. (2002), only HapMap data within the bounds of this map were included, and centromeric regions were excluded. Singletons (polymorphisms observed in only one allele) were excluded from both data sets. Finally, to conservatively eliminate regions of low SNP coverage from the HapMap data, gaps of >10 kb between non-singleton SNPs were excluded from the analyzed regions.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.

Results of this comparison (see Fig. 3) show at least two notable patterns. First, long shared haplotypes within populations are more abundant in the HapMap data than predicted by the Schaffner et al. (2005) model. For example, tracts within a 10-kb range centered on 200 kb are 3.7 times more abundant within the African HapMap data, and 3.9 times more abundant in Europe. Evolutionary processes that could account for this difference include (1) recombination rates more heterogeneous than modeled, (2) additional recent bottlenecks in both populations, and (3) selective sweeps. Second, it is apparent that long haplotypes shared between populations are more greatly enriched (by a factor of 12.6 in the same window) than within-population tracts. This pattern could be a signal of recently elevated migration between continents, but further analysis is needed to evaluate this and other hypotheses.
within species, as suggested by Lohmueller et al. (2008) to explain the higher proportion of deleterious variants inferred for European Americans relative to African-Americans (as expected if Europeans have had historically smaller population sizes). Because the study of deleterious variation often focuses on rare alleles, generation of whole-genome sequence polymorphism data from large population samples will be instrumental in refining our understanding of selective constraint in the genome and the genetic load of natural populations.

**The need for improved models of selection**

Understanding the relative roles of natural selection and neutral forces in shaping genetic diversity is a central but unresolved issue in population genetics. However, our ability to accurately model the joint effects of demography and both positive selection and negative selection in a recombining genome is largely restricted to simulations. Most models of positive selection make strong simplifying assumptions, such as constant selection pressure over time, and/or all selection acting on new variants. There has been some progress in developing alternative models of selection, such as the case of a sweep from standing variation (Orr and Betancourt 2001; Innan and Kim 2004; Hermisson and Pennings 2005; Przeworski et al. 2005; Pennings and Hermisson 2006). Other potential departures from the basic recurrent hitchhiking model (Kaplan et al. 1989; Stephan et al. 1992) include variation in selection coefficients through space and/or time (e.g., Ohta 1972; Gillespie 1973; Takahata et al. 1975; Mustonen and Lässig 2007; Huerta-Sanchez et al. 2008). Even with growing population genomic data sets, testing among alternative models of selection in the presence of nonequilibrium demography will present a formidable challenge. Instead of generating very complex parametric models, it may be useful to concentrate on specific aspects of the data that can help distinguish between models, such as the effect of recombination rate on summary statistics, comparisons of markers with different modes of inheritance, and the distribution of shared haplotype lengths. While population genetic theory once far exceeded the data available to test it, today it is the models and methods that must catch up with the data.

**Conclusions**

Genome-wide data are becoming readily available in a number of organisms. It is clear that population genetics is increasingly moving toward genome-wide analyses, especially in organisms such as humans and *Drosophila*. But even ecological and evolutionary studies of natural populations may increasingly turn to genome-wide sequencing based on RRSS to cheaply and effectively generate large data sets. Analyses of genome-wide data will allow us to use new tools for understanding the ecology and evolution of natural populations. For example, we may use shared haplotypes to make inferences about very recent migration between populations. The study of genome-wide patterns of variability may also greatly improve our understanding of molecular evolution and the relative contributions of mutation, recombination, genetic drift, and natural selection. However, it will be important in such studies to take the special nature of the data into account: a high sequencing error rate, possible assembly errors, and missing data. While several of these problems can be addressed by using very high coverage, this is usually not cost-effective. Instead, we must increasingly rely on a statistical analysis of the data that takes all of these challenges into account.

**Acknowledgments**

This research was supported by a National Institutes of Health (NIH) Kirschstein-NRSA postdoctoral fellowship (F32 HG004182) to J.E.P., a Human Frontier Science Program postdoctoral fellowship (LT00794/2006-L) to I.H., and a NIH research grant (U01HL084706) to R.N.

**References**


