Population genetic inference from genomic sequence variation

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

[1]*Department of Integrative Biology, University of California, Berkeley, Berkeley, California 94720, USA;* [2]*Center for Population Biology, University of California, Davis, Davis, California 95616, USA;* [3]*Mathematics and Biosciences Group, Max F. Perutz Laboratories, Vienna 1030, Austria;* [4]*Department 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 we 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 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 (Altshuler et al. 2000; Baird

[5]Present address: Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA.
[6]Corresponding author.
E-mail rasmus_nielsen@berkeley.edu; fax (510) 643-6264.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.079509.108.

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 nonrandom 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 prefilter. 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.

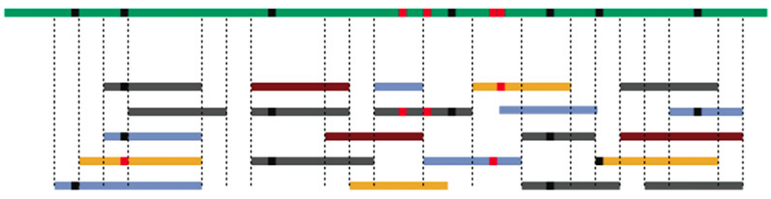

Figure 1. Reads from different individuals are aligned to a reference genome, and SNPs have been called. In this toy example there are seven true segregating sites (black dots) and four false ones (red dots). Also note that segments differ in sample size. In segment 4, for example, five reads from three individuals were sampled; thus, there were at least three chromosomes and at most five sampled in this segment.

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).

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 can be fitted to X-linked diversity data for non-African *D. melanogaster* (e.g., Thornton and Andolfatto 2006), Hutter et al. (2007) found that no simple bottleneck scenario could account for both X-linked and autosomal data, and Pool and Nielsen (2008) then suggested an alternate demographic model that was more compatible with X-linked and autosomal diversity levels. Relatively few genome-wide demographic analyses have incorporated both X-linked and autosomal variation, but in light of the above example, joint consideration of these data sources should produce more accurate inferences of population history.

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 (Huelsenbeck 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

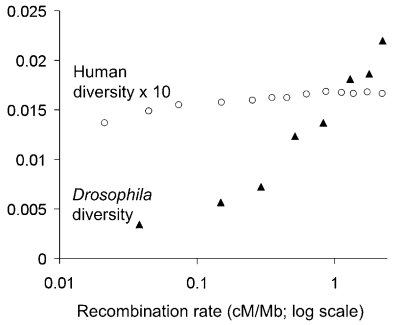

Figure 2. The relationship between recombination rate and nucleotide diversity in data from *D. melanogaster* and humans. The data points shown represent average nucleotide diversity for recombination rate bins. The *D. melanogaster* data (Shapiro et al. 2007) are from 349 loci with ≥50 synonymous sites 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 negative 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.

Box 1. Haplotype sharing within and between populations

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 by 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.

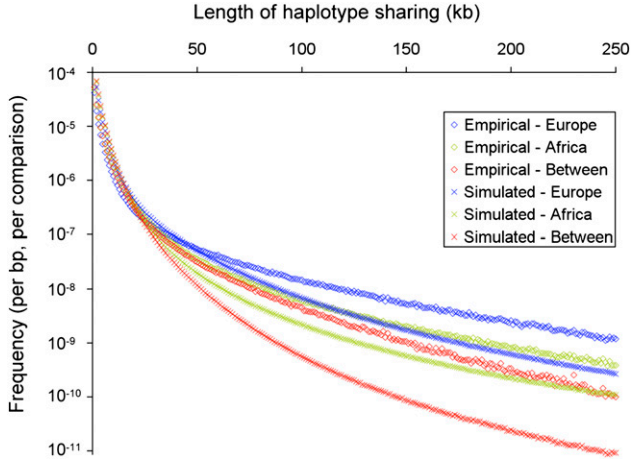

Figure 3. Lengths of haplotypes shared by pairs of alleles within African (green) and European (blue) human populations, and between populations (red), in HapMap SNP data (open boxes) and simulations (×).

larger effective population size of *Drosophila* enabling a more pervasive influence of linked selection, and perhaps also a greater density of functional sites in the more compact *Drosophila* genome. And in general, it has become clear that in *Drosophila*, the assumption of selective neutrality in random portions of the genome is unlikely to hold (for review, see Sella et al. 2009).

With larger genome-wide data sets, it will become increasingly possible to move beyond qualitative conclusions about selection in the genome and obtain quantitative estimates of parameters such as the rate of selective sweeps and the strength of selection. Several recent polymorphism-based inference methods of this type have been developed and applied to data from *Drosophila* (Li and Stephan 2006; Andolfatto 2007; Macpherson et al. 2007; Jensen et al. 2008). These estimators differ statistically (likelihood vs. Bayesian), by the type of data analyzed (polymorphism and/or divergence) and in general framework, with some depending on the genomic variance created on differing spatial scales between models (e.g., Macpherson et al. 2007), and others taking a McDonald and Kreitman (1991)–based approach (e.g., Andolfatto 2007). Perhaps because of differences in methodology and the spatial scale of analysis, published estimates using these different approaches have been far from consistent. In particular, the mean estimates of average genomic selection coefficients for beneficial mutations in *Drosophila* range from very weak (s = 0.00001) to strong selection (s = 0.01). Whole-genome sequence polymorphism data will be instrumental in differentiating between these scenarios, since weak sweeps should leave narrow footprints (e.g., a high variance in diversity on a fine chromosomal scale) that may only be detectable from the densest data.

While the distribution of selection coefficients for adaptive mutations remains unclear (aside from a few microbial experimental evolution studies) (for review, see Eyre-Walker and Keightley 2007), the distribution of fitness effects for deleterious mutations can be inferred based on comparisons of allele frequencies at synonymous and nonsynonymous sites. Keightley and Eyre-Walker (2007) and Boyko et al. (2008) found that roughly half of human nonsynonymous mutations were neutral or weakly deleterious, while in *Drosophila* the vast majority were more strongly deleterious ($N_e s > 10$) (Keightley and Eyre-Walker 2007). This difference may again reflect the larger N_e and increased efficiency of selection in *Drosophila*. Population sizes may also vary

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 (UO1HL084706) to R.N.

References

Adams A, Hudson RR. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168:** 1699–1712.

Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol* **33:** 266–274.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19:** 1655–1664.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437:** 1149–1152.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* **17:** 1755–1762.

Andolfatto P, Przeworski M. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158:** 657–665.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3:** e3376. doi: 10.1371/journal.pone.0003376.

Bansal V, Halpern AL, Axelrod N, Bafna V. 2008. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res* **18:** 1336–1346.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. *Genetics* **162:** 2025–2035.

Becquet C, Przeworski M. 2007. A new approach to estimate parameters of speciation models with application to apes. *Genome Res* **17:** 1505–1519.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356:** 519–520.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh YP, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol* **5:** e310. doi: 10.1371/journal.pbio.0050310.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Bhangale TR, Rieder MJ, Nickerson DA. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* **40:** 841–843.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4:** e1000083. doi: 10.1371/journal.pgen.1000083.

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140:** 783–796.

Bryant D, Wong W, Mockler T. 2009. QSRA—a quality-value guided de novo short read assembler. *BMC Bioinformatics* **10:** 69. doi: 10.1186/1471-2105-10-69.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, et al. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* **3:** e163. doi: 10.1371/journal.pgen.0030163.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* **15:** 1553–1565.

Chaisson MJ, Pevzner PA. 2008. Short read fragment assembly of bacterial genomes. *Genome Res* **18:** 324–330.
Chakraborty R, Weiss KM. 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci* **85:** 9119–9123.
Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.
Cheung VG, Nelson SF. 1998. Genomic mismatch scanning identifies human genomic DNA shared identical by descent. *Genomics* **47:** 1–6.
Coop G, Przeworski M. 2007. An evolutionary view of human recombination. *Nat Rev Genet* **8:** 23–34.
Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9:** 539. doi: 10.1186/1471-2105-9-539.
Davison D, Pritchard J, Coop G. 2009. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor Popul Biol* **75:** 331–345.
Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. 2009. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327:** 78–81.
Ewing B, Green P. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res* **8:** 186–194.
Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* **8:** 610–618.
Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164:** 1567–1587.
Fay JC, Wu C-I. 1999. A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol Biol Evol* **16:** 1003–1005.
Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405–1413.
Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915–925.
Fu Y, Peckham HE, McLaughlin SF, Rhodes MD, Malek JA, McKernan KJ, Blanchard AP. 2008. SOLID sequencing and Z-Base encoding. In *The Biology of Genomes Meeting, Cold Spring Harbor Laboratory*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Gillespie JH. 1973. Natural selection with varying selection coefficients—a haploid model. *Genet Res* **21:** 115–120.
Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. 2009. Whole population, genome-wide mapping of hidden relatedness. *Genome Res* **19:** 318–326.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5:** e1000695. doi: 10.1371/journal.pgen.1000695.
Hahn MW. 2006. Accurate inference and estimation in population genomics. *Mol Biol Evol* **23:** 911–918.
Harris H. 1966. Enzyme polymorphisms in man. *Proc R Soc Lond B Biol Sci* **164:** 298–310.
Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a copying model. *PLoS Genet* **4:** e1000078. doi: 10.1371/journal.pgen.1000078.
Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72:** 1527–1535.
Hellmann I, Mang Y, Gu Z, Li P, de la Vega FM, Clark AG, Nielsen R. 2008. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18:** 1020–1029.
Hermisson J, Pennings PS. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169:** 2335–2352.
Hey J, Harris E. 1999. Population bottlenecks and patterns of human polymorphism. *Mol Biol Evol* **16:** 1423–1426.
Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.
Hoggart C, Shriver M, Kittles R, Clayton D, McKeigue P. 2004. Design and analysis of admixture mapping studies. *Am J Hum Genet* **74:** 965–978.
Hubby JL, Lewontin RC. 1966. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* **54:** 577–594.
Hudson RR, Kaplan NL. 1988. The coalescent process in models with selection and recombination. *Genetics* **120:** 831–840.
Huelsenbeck JP, Andolfatto P. 2007. Inference of population structure under a Dirichlet process model. *Genetics* **175:** 1787–1802.
Huerta-Sanchez E, Durrett R, Bustamante CD. 2008. Population genetics of polymorphism and divergence under fluctuating selection. *Genetics* **178:** 325–337.
Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. *Genetics* **177:** 469–480.
Innan H, Kim Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Natl Acad Sci* **101:** 10667–10672.
Innan H, Stephan W. 2003. Distinguishing the hitchhiking and background selection models. *Genetics* **165:** 2307–2312.
The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.
The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.
Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H-C, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451:** 998–1003.
Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. 2007. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* **176:** 2371–2379.
Jensen JD, Thornton KR, Andolfatto P. 2008. An approximate Bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genet* **4:** e1000198. doi: 10.1371/journal.pgen.1000198.
Jiang R, Tavare S, Marjoram P. 2009. Population genetic inference from resequencing data. *Genetics* **181:** 187–197.
Johnson PLF, Slatkin M. 2006. Inference of population genetic parameters in metagenomics: A clean look at messy data. *Genome Res* **16:** 1320–1327.
Johnson PLF, Slatkin M. 2008. Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* **25:** 199–206.
Kaplan NL, Hudson RR, Langley CH. 1989. The 'hitchhiking effect' revisited. *Genetics* **123:** 887–899.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177:** 2251–2261.
Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* **19:** 1195–1201.
Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39:** 1251–1255.
Kelley JL, Swanson WJ. 2008. Positive selection in the human genome: From genome scans to biological significance. *Annu Rev Genomics Hum Genet* **9:** 143–160.
Kidd JM, Cheng Z, Graves T, Fulton B, Wilson RK, Eichler EE. 2008. Haplotype sorting using human fosmid clone end-sequence pairs. *Genome Res* **18:** 2016–2023.
Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167:** 1513–1524.
Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241–247.
Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.
Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304:** 412–417.
Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149:** 429–434.
Kuhner MK, Beerli P, Yamato J, Felsenstein J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156:** 439–447.
Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Science* **265:** 2037–2048.
Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.
Lewontin RC. 1972. The apportionment of human diversity. In *Evolutionary biology* (ed. TH Dobzhansky et al.), pp. 381–398. Kluwer Academic Publishers, New York.
Lewontin RC, Hubby JL. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* **54:** 595–609.
Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2:** e166. doi: 10.1371/journal.pgen.0020166.
Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165:** 2213–2233.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.
Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: Short Oligonucleotide Alignment Program. *Bioinformatics* **24:** 713–714.
Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451:** 994–997.
Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182:** 217–231.
Long Q, MacArthur D, Ning Z, Tyler-Smith C. 2009. HI: Haplotype Improver using paired-end short reads. *Bioinformatics* **25:** 2436–2437.
Lynch M. 2008. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25:** 2409–2419.
Lynch M. 2009. Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182:** 295–301.
Macpherson JM, Sella G, Davis JC, Petrov DA. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* **177:** 2083–2099.
Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39:** 906–913.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genomes sequencing in open microfabricated high density picoliter reactors. *Nature* **437:** 376–380.
Marth GT, Czabarka E, Murvai J, Sherry ST. 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166:** 351–372.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.
McVean G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics* **175:** 1395–1406.
Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* **201:** 786–792.
Meyer M, Stenzel U, Hofreiter M. 2008. Parallel tagged sequencing on the 454 platform. *Nat Protoc* **3:** 267–278.
Mustonen V, Lässig M. 2007. Adaptations to fluctuating selection in *Drosophila*. *Proc Natl Acad Sci* **104:** 2277–2282.
Myers S, Fefferman C, Patterson N. 2008. Can one learn history from the allelic spectrum? *Theor Popul Biol* **73:** 342–348.
Nielsen R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154:** 931–942.
Nielsen R. 2004. Population genetic analysis of ascertained SNP data. *Hum Genomics* **3:** 218–224.
Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics* **158:** 885–896.
Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* **8:** 857–868.
Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19:** 838–849.
Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* **40:** 646–649.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelsen MR, et al. 2008. Genes mirror geography within Europe. *Nature* **456:** 98–101.
Ohta T. 1972. Population size and rate of evolution. *J Mol Evol* **1:** 305–314.
O'Reilly PF, Birney E, Balding DJ. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res* **18:** 1304–1313.
Orr HA, Betancourt AJ. 2001. Haldane's sieve and adaptation from standing genetic variation. *Genetics* **157:** 875–884.
Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18:** 2024–2033.
Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, et al. 2004. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74:** 979–1000.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2:** e190. doi: 10.1371/journal.pgen.0020190.
Pennings PS, Hermisson J. 2006. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol Biol Evol* **23:** 1076–1084.
Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet* **2:** e105. doi: 10.1371/journal.pgen.0020105.
Polanski A, Kimmel M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165:** 427–436.
Pool JE, Nielsen R. 2007. Population size changes reshape genomic patterns of diversity. *Evolution* **61:** 3001–3006.
Pool JE, Nielsen R. 2008. The impact of founder events on chromosomal variability in multiply mating species. *Mol Biol Evol* **25:** 1728–1736.
Pool JE, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181:** 711–719.
Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5:** e1000519. doi: 10.1371/journal.pgen.1000519.
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000a. Association mapping in structured populations. *Am J Hum Genet* **67:** 170–181.
Pritchard JK, Stephens M, Donnelly P. 2000b. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.
Przeworski M. 2002. The signature of positive selection at randomly chosen loci. *Genetics* **160:** 1179–1189.
Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* **59:** 2312–2323.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.
Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.
Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.
Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449:** 913–918.
Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15:** 1576–1583.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet* **5:** e1000495. doi: 10.1371/journal.pgen.1000495.
Servin B, Stephens M. 2007. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet* **3:** e114. doi: 10.1371/journal.pgen.0030114.
Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA, Fang S, Wang H-Y, Hudson RR, Nielsen R, et al. 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci* **104:** 2271–2276.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135–1145.
Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* **23:** 23–35.
Stephan W, Wiehe THE, Lenz MW. 1992. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theor Popul Biol* **47:** 237–254.
Stephan W, Song YS, Langley CH. 2006. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172:** 2647–2663.
Stephens JC, Briscoe D, O'Brien SJ. 1994. Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. *Am J Hum Genet* **55:** 809–824.
Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S. 2007. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS One* **2:** e484. doi: 10.1371/journal.pone.0000484.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Takahata N, Ishii K, Matsuda H. 1975. Effect of temporal fluctuation of selection coefficient on gene frequency in a population. *Proc Natl Acad Sci* **72:** 4541–4545.
Thornton K, Andolfatto P. 2006. Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172:** 1607–1619.
Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci* **102:** 18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72. doi: 10.1371/journal.pbio.0040072.
Wall JD, Andolfatto P, Przeworski M. 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162:** 203–216.
Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci* **102:** 7882–7887.
Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3:** e90. doi: 10.1371/journal.pgen.0030090.
Wooding SA, Rogers A. 2002. The matrix coalescent and application to human single-nucleotide polymorphisms. *Genetics* **161:** 1641–1650.
Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. 2005. The effects of artificial selection on the maize genome. *Science* **308:** 1310–1314.
Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z, Dai F, Li Y, Cheng D, Li R, et al. 2009. Complete resequencing of 40 genomes reveals domestication events and genes in Silkworm (*Bombyx*). *Science* **326:** 433–436.
Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18:** 821–829.