Constructing genomic maps of positive selection in humans: Where do we go from here?

Joshua M. Akey

Genome Res. 2009 19: 711-722
Access the most recent version at doi:10.1101/gr.086652.108

Perspective

Constructing genomic maps of positive selection in humans: Where do we go from here?

Joshua M. Akey[1]

Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

Identifying targets of positive selection in humans has, until recently, been frustratingly slow, relying on the analysis of individual candidate genes. Genomics, however, has provided the necessary resources to systematically interrogate the entire genome for signatures of natural selection. To date, 21 genome-wide scans for recent or ongoing positive selection have been performed in humans. A key challenge is to begin synthesizing these newly constructed maps of positive selection into a coherent narrative of human evolutionary history and derive a deeper mechanistic understanding of how natural populations evolve. Here, I chronicle the recent history of the burgeoning field of human population genomics, critically assess genome-wide scans for positive selection in humans, identify important gaps in knowledge, and discuss both short- and long-term strategies for traversing the path from the low-resolution, incomplete, and error-prone maps of selection today to the ultimate goal of a detailed molecular, mechanistic, phenotypic, and population genetics characterization of adaptive alleles.

[Supplemental material is available online at www.genome.org.]

In August 1858, Charles Robert Darwin and Alfred Russel Wallace communicated essays to the Linnean Society of London (Darwin and Wallace 1858) describing their independent discovery of the theory of natural selection and, in doing so, fundamentally altered our understanding of life on Earth. Darwin's meticulously detailed book, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life*, published 1 yr later (Darwin 1859), provided a more comprehensive account of his evidence for natural selection and its role in mediating evolutionary change. The immediate and enduring interest in Darwin's and Wallace's work is not only because of the powerful, unifying, and explanatory theory it provides for biology but also because it is a theory about us as humans and our place in the natural world.

Over the ensuing century and a half, considerable progress has been made in elucidating the molecular and mechanistic details of how natural populations evolve. Integral to this progress was the development and maturation of population genetics, the foundations of which can largely be ascribed to the seminal contributions of Wright, Fisher, and Haldane (Provine 1971), who have been referred to as the "great trinity" of population genetics (Crow 1994). Furthermore, we now recognize the contribution of both chance and natural selection as vehicles of evolutionary change (Kimura 1968; King and Jukes 1969). The very concept of natural selection itself has evolved, with different types of selection being distinguished. Essentially, the various modes of selection follow from whether an allele is advantageous or deleterious and the fitness relationships among genotypes (Box 1). Darwin and Wallace were primarily interested in adaptive evolution, which at the molecular level is governed by selection acting on advantageous alleles. The nomenclature of selection can be daunting for the uninitiated (Boxes 1, 2), and therefore for simplicity I will refer to any form of selection acting on advantageous alleles as positive selection (Nielsen 2005).

[1]**Corresponding author.**
E-mail akeyj@u.washington.edu; fax (206) 685-7301.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.086652.108.

Despite the significant advances made to date, intense research has thus far failed to reveal all of evolution's secrets. In particular, progress in humans has been frustratingly difficult to achieve. However, the tide seems to be turning, and the confluence of dense catalogs of human genetic variation and methodological tools have led to the construction of many genomic maps of positive selection. These maps simultaneously hold great promise and pose important challenges for arriving at a detailed understanding of how positive selection has shaped our genomes and our history.

Here, I will provide an overview of how studies of human evolutionary history have profited from the genomics era and critically assess recent attempts to construct genomic maps of positive selection in humans. My goals are twofold. First, rather than focus on the details of any particular study, which have been extensively reviewed elsewhere (Ronald and Akey 2005; Biswas and Akey 2006; Harris and Meyer 2006; Sabeti et al. 2006; Nielsen et al. 2007), I will discuss several broad themes of recent human evolutionary history emerging from the synthesis of results across studies. Second, I will enumerate both short- and long-term strategies for translating maps of positive selection into a detailed molecular, mechanistic, phenotypic, and population genetics characterization of adaptive alleles.

Making the case for population genomics

Before the genomics era, inferences regarding natural selection were made almost exclusively through candidate gene studies (Sabeti et al. 2006). These single gene analyses have yielded some notable success stories illuminating deep insights into recent human evolutionary history, including compelling evidence for positive selection of *LCT* (Bersaglieri et al. 2004), which allows lactose tolerance to persist throughout adulthood, as well as a number of genes that reduce susceptibility to malaria infection, such as *G6PD* (Tishkoff et al. 2001), *DARC* (Hamblin and Di Rienzo 2000; Hamblin et al. 2002), and *HBB* (Friedman 1978; Currat at el. 2002).

In general, however, candidate gene studies suffer from two key limitations. First, they require an a priori hypothesis about

Box 1. Types of natural selection

A significant amount of insight into the various types of natural selection can be obtained by considering a simple single locus model. Specifically, assume that initially a single allele, A_1, exists, and at some point in time a mutation introduces the allele A_2 into the population. The three possible genotypes are A_1A_1, A_1A_2, and A_2A_2, and the fitness of each genotype is w_{11}, w_{12}, and w_{22}. A straightforward, though naïve, interpretation of genotypic fitness is that it represents the probability that individuals with a particular genotype survive. In reality, fitness has many components, of which survival is just one. In the literature, the absolute fitness values w_{11}, w_{12}, and w_{22} are often converted into relative fitness. For instance, the relative fitness of genotypes A_1A_1, A_1A_2, and A_2A_2 can be denoted as 1, 1 + hs, and 1 + s. Here, the fitness of genotypes A_1A_2 and A_2A_2 are expressed relative to the fitness of genotype A_1A_1 (thus, $1 + hs = w_{12}/w_{11}$ and $1 + s = w_{22}/w_{11}$). The parameters s and h are called the selection coefficient and heterozygote effect, respectively. The various types of selection are described here in terms of relative fitness.

When there are no fitness differences among genotypes ($s = 0$), allele and genotype frequencies are said to evolve neutrally; otherwise natural selection occurs. The specific type of selection depends on whether s is positive or negative and the dominance relationship between alleles as captured by h. For example, directional selection occurs with incomplete dominance ($0 < h < 1$). If $s < 0$, then the newly arisen allele A_2 is deleterious, individuals carrying this allele are less fit, and purifying selection (also known as negative selection) acts to purge A_2 from the population. If $s > 0$, the newly arisen allele A_2 is advantageous, individuals carrying this allele are more fit, and A_2 will ultimately become fixed in the population. Directional selection results in loss of genetic variation and, in general, directional selection of an advantageous allele is most often ascribed to the form of selection Darwin envisioned.

Another type of selection that acts on advantageous alleles is referred to as "overdominant selection," which occurs when the heterogyzote has the highest relative fitness ($s > 0$ and $h > 1$). Overdominant selection (also known as heterozygote advantage) is one specific incarnation of balancing selection, which acts to maintain genetic variation in a population. Note that balancing selection can occur in the absence of overdominance (Gillespie 1991). In the main text, and following Nielsen (2005), I refer to any type of selection acting on an advantageous allele as positive selection. However, directional selection is likely to be the primary form of positive selection identified in genome-wide scans (but see Wang et al. 2006).

The above synopsis is a necessary oversimplification of the full complexities of selection models and the dynamics of fitness influencing alleles. Perhaps the most important point to add is the central role of chance, even in the trajectory of adaptive genetic variation. For example, a classic result of theoretical population genetics is that the fixation probability of a newly arisen advantageous mutation is $\approx 2s$ (Haldane 1927). Thus, the vast majority of advantageous mutations are lost from the population, a phenomenon Gillespie (1998) elegantly referred to as "the quagmire of randomness."

which genes may have been subject to selection, and in order to form this hypothesis, it is necessary to have an understanding about genotype–phenotype relationships. Although this has been feasible for some phenotypes that have a well-defined, almost Mendelian architecture (such as lactose tolerance), the genetic architecture of phenotypic variation remains enigmatic for most traits, constraining our ability to intelligently nominate candidate genes to study. Another avenue for identifying candidate genes has been to focus on loci belonging to classes of genes that appear to be frequent targets of selection, such as those involved in immunity and defense (Bamshad and Wooding 2003). While this approach is intuitively appealing, it potentially leads to a biased set of loci that have been subject to selection. Furthermore, candidate gene studies are an inefficient study design for detecting positive selection in regulatory regions far removed from genic loci.

The second key limitation of candidate gene approaches is due to the confounding effects of genetic drift, manifested through population demographic history, and natural selection on extant patterns of genetic variation (Simonsen et al. 1995; Przeworski et al. 2000; Akey et al. 2004; Stajich and Hahn 2005). Thus, interpreting patterns of genetic variation at individual loci is often difficult, making robust inferences of positive selection challenging.

Genomics has offered a new paradigm for detecting signatures of selection, which has been referred to as population genomics. The term population genomics appears to have been introduced into the lexicon of genomics jargon independently by Hedges (2000) and Black IV et al. (2001). In its most general form, population genomics refers to the inference of population genetic and evolutionary parameters from genome-wide data sets (Black IV et al. 2001).

In the context of identifying substrates of positive selection, population genomics offers a potential solution to the two key limitations of candidate gene studies. First, the genome can be surveyed without any a priori assumptions regarding which genes may be under selection, yielding a less biased set of putatively selected loci. Second, population genomics provides a framework for distinguishing, at least in principle, between population demographic history and natural selection. Specifically, the most commonly used population genomics approach involves sampling a large number of loci throughout the genome, calculating

Box 2. Nomenclature of selective sweeps

Effects of positive selection on linked variation

Inferences about natural selection usually rely upon detecting its effects on patterns of linked neutral variation. Genetic hitchhiking refers to the influence that selection of an advantageous allele has on patterns of linked variation (Maynard-Smith and Haigh 1974). When an advantageous allele fixes in a population, it does so on a particular haplotype background. Linked variation is thus swept through the population along with the advantageous mutation, a process referred to as a "selective sweep." Ongoing, or incomplete, sweeps denote any stage prior to the fixation of the advantageous allele. Once it becomes fixed, the sweep is said to be complete.

Hard versus soft sweeps

The classic model of positive selection, implicitly assumed above, is that selection acts upon a newly arisen advantageous mutation. Alternatively, selection could act on preexisting genetic variation that was previously either neutral or deleterious, but has become adaptive due to changes in the environment or genetic background. Recently, selection from standing variation has been referred to as a "soft sweep" (Hermisson and Pennings 2005), to distinguish it from the classic model, or hard sweep. Patterns of genetic variation arising from selection on newly arisen can differ markedly between soft and hard sweeps (Hermisson and Pennings 2005; Przeworski et al. 2005).

a summary statistic that quantifies some aspect of genetic variation, constructing an empirical distribution of this statistic across all loci, and defining putative targets of selection based on "outliers" in the extreme tail of the empirical distribution (Fig. 1). The underlying rationale of this approach is predicated on several implicit assumptions (Fig. 1), the most important of which is that population demographic history is a genome-wide force affecting all loci equally, whereas selection is a locus-specific force acting on a subset of loci (Black IV et al. 2001). If all goes well, selection pulls individual loci into the tails of the empirical distribution, which can be identified as outlier loci. The criteria for defining outliers is often arbitrary, for example, loci falling in 99th percentile of the empirical distribution, although simulations of neutral evolution have also been used to either guide the selection of or evaluate the efficiency of outlier thresholds. Increasingly realistic models of human demographic history, recombination, gene conversion, and mutation rate heterogeneity (Schaffner et al. 2005) will ultimately allow more robust definitions of what constitutes an outlier locus.

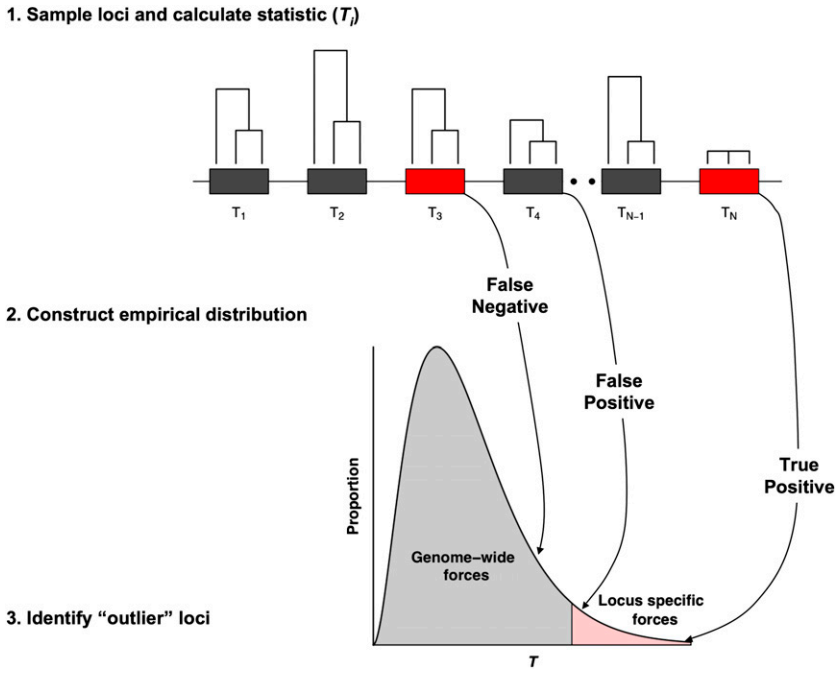

Figure 1. A typical population genomics study design for detecting positive selection. Population genomic studies begin by sampling loci, typically SNPs, throughout the genome. The majority of loci are presumably influenced only by genome-wide forces such as genetic drift (indicated by dark gray boxes). Additional loci, however, may have been subject to locus-specific forces such as selection (indicated by red boxes). Gene genealogies from a sample of three individuals are shown *above* each locus to emphasize that significant variation in genealogies, and thus, patterns of genetic variation are expected throughout the genome. The extent of variation in genealogies depends on many underlying parameters such as population demographic history and local rates of recombination. For each sampled locus, a statistic of interest (denoted here as T_i for the ith locus) is calculated, an empirical distribution is constructed, and outlier loci are identified in the tail of the empirical distribution. Implicit assumptions of a population genomics approach are that loci are independent, drift influences all loci equally, and selection is strong enough to pull individual loci out into the tail of the empirical distribution. It is important to note that simply occurring in the tail of an empirical distribution does not prove that a locus has been influenced by selection; rather, all one can conclude is that the locus simply has patterns of genetic variation that are unusual in some respect relative to the rest of the genome. Indeed, as shown in the empirical distribution, it is inevitable that some selected loci will not appear as outliers (false negatives) and some neutral loci will appear as outliers (false positives). The lighter red and gray shadings of the empirical distribution reflect that each part of the distribution is a mixture of selected and neutral loci.

Enabling resources for human population genomics

The recent advent of population genomics is not due to a great intellectual leap forward. Indeed, the conceptual foundation of the population genomics paradigm was outlined nearly four decades ago (Cavalli-Sforza 1966; Lewontin and Krakauer 1978). Rather, progress in sequencing the human genome (Lander et al. 2001; Venter et al. 2001) and single nucleotide polymorphism (SNP) genotyping technology (and to a lesser extent microsatellite genotyping) allowed dense catalogs of genomic variation to be developed, thus enabling population genomics approaches to become a reality. One of the earliest human population genomics resources was the systematic discovery of over 1.42 million SNPs (Sachidanandam et al. 2001), ~26,500 of which were genotyped by the SNP Consortium in three populations (Thorisson and Stein 2003). This early resource was quickly superseded by the International HapMap Project (International HapMap Consortium 2005) and Perlegen Biosciences data sets (Hinds et al. 2005), which genotyped ~3 million SNPs in 210 unrelated individuals from four populations and 1.6 million SNPs in three populations, respectively. Although the HapMap and Perlegen data sets have served as the primary starting points for much of human population genomics, the continued infusion of additional SNP (Jakobsson et al. 2008; Li et al. 2008), structural variation (Redon et al. 2006; Kidd et al. 2008), and large-scale resequencing data (Livingston et al. 2004) provides a rich repository of raw material to test evolutionary hypotheses. An important characteristic of these data sets that has fueled the rapid growth of human population genomics is their public availability, which has stimulated both methodological development and applications beyond their original purposes in ways that would not have been possible except for open and unfettered access.

Genome-wide scans of selection in humans: Promises and pitfalls

The first genome-wide scans of selection in humans were performed less than a decade ago (Akey et al. 2002; Payseur et al. 2002), albeit with considerably less dense maps of genetic variation compared with what is available today. Since these initial studies, genome-wide scans have been described at a frenetic pace. Specifically, as shown in Table 1, 21 genome-wide scans for recent or ongoing selection (Box 2) have been performed in humans. The majority of these recent analyses have used either the HapMap or Perlegen data, although several studies have been performed in distinct samples (Table 1). Furthermore, these 21 scans for selection

Table 1. Summary of genome-wide scans of positive selection in humans

Study[a]	Data	Statistical method[b]	Sample
Akey et al. (2002)	SNP	Population differentiation	European-American, African-American, Chinese-American
Payseur et al. (2002)	Microsatellite	Site frequency spectrum	European
Kayser et al. (2003)	Microsatellite	Population differentiation	African, European
Storz et al. (2004)	Microsatellite	Population differentiation, site frequency spectrum	African, Asian, European
Shriver et al. (2004)	SNP	Population differentiation	European-American, African-American, East Asian
International HapMap Consortium (2005)	SNP	LD, population differentiation	HapMap
Weir et al. (2005)	SNP	Population differentiation	HapMap, Perlegen
Carlson et al. (2005)	SNP	Site frequency spectrum	Perlegen
Bustamante et al. (2005)	Sequence	Ratio of polymorphism to divergence	European
Mattiangeli et al. (2006)	SNP	Population differentiation	Irish
Wang et al. (2006)	SNP	LD	HapMap, Perlegen
Voight et al. (2006)	SNP	LD	HapMap
Kelley et al. (2006)	SNP	Site frequency spectrum	Perlegen
Tang et al. (2007)	SNP	LD	HapMap, Perlegen
Kimura et al. (2007)	SNP	LD	HapMap
Williamson et al. (2007)	SNP	Site frequency spectrum	Perlegen
Sabeti et al. (2007)	SNP	LD	HapMap
Johansson and Gyllensten (2008)	SNP	Joint analysis of population differentiation and LD	Perlegen
Kimura et al. (2008)	SNP	LD	Melanesian, Polynesian
Oleksyk et al. (2008)	SNP	Population differentiation	European-American, African-American
O'Reilly et al. (2008)	SNP	LD	HapMap, Perlegen

[a]Studies included in the analysis of the integrated map of positive selection are shown in bold. Inclusion criteria were that the study was performed in the HapMap or Perlegen data, lists of all loci deemed as outlier were available as supplemental data, and sufficient information provided information about what genome build was used for the reported map positions.
[b]The general class of statistical test of neutrality is presented; for more details, see Box 3 and original publications.

have used a variety of statistical approaches to detect deviations from neutrality, which will become clear below, is an important factor in interpreting results across studies. A general synopsis of methods used to detect deviations from neutrality is provided in Box 3; detailed reviews can be found elsewhere (Kreitman 2000; Nielsen 2001; Ronald and Akey 2005; Biswas and Akey 2006; Nielsen et al. 2007).

The rapid accumulation of genomic maps of positive selection is an important milestone, increasing the number of loci putatively under selection by several orders of magnitude compared to candidate gene approaches. These maps hold considerable promise in guiding us toward a more detailed understanding of where, and why, positive selection has shaped extant patterns of human genetic variation, but only if they are guiding us toward genuine substrates of selection. It is important to note that the majority of the studies in Table 1 have defined targets of positive selection in the canonical population genomics fashion, namely, the identification of outlier loci. However, being an outlier is not necessarily synonymous with being under selection (Fig. 1).

One way to assess confidence in the results of genome-wide scans is to examine the overlap of outlier loci across studies. To this end, for eight recent studies performed on the HapMap and Perlegen data indicated in Table 1 (which describes inclusion criteria), I obtained genomic positions for all reported autosomal loci under selection, mapped their positions to the same genomic build (UCSC hg18), and merged overlapping loci into a set of nonredundant positions. In addition, I also included results based on a simple genome-wide scan in the HapMap samples using F_{ST}, which is a commonly used test statistic to detect local adaptation (Akey et al. 2002), the type of selection acting upon *LCT* (Bersaglieri et al. 2004; Tishkoff et al. 2007).

The integrated map of positive selection across these nine genome-wide scans is shown in Figure 2. In total, 5110 distinct regions were identified in one or more study. These regions encompass ~409 Mb of sequence (~14% of the genome) and contain 4243 UCSC RefSeq genes (~23% of all genes). Strikingly, only 722 regions (14.1%) were identified in two or more studies, 271 regions (5.3%) were identified in three or more studies, and 129 regions (2.5%) were identified in four or more studies (Fig. 1). Furthermore, the integrated map of positive selection does not include several of the most compelling genes with well-substantiated claims of positive selection, such as *G6PD* and *DARC*. *G6PD* is located on the X chromosome, which generally has not been included in genome-wide analyses. *DARC*, however, is an autosomal gene, but the HapMap and Perlegen samples are not representative of populations in which the signature of selection has been reported by Hamblin and Di Rienzo (2000) and Hamblin et al. (2002). However, even if additional populations were included, it is unclear if *DARC* would emerge as an outlier locus given that its signature of selection is confined to a very small genomic region (~10 kb).

Although the poor concordance among studies is sobering, further inspection suggests some room for optimism. In particular, as noted above, the integrated map of positive selection contains 5110 regions spanning 409 Mb of total sequence. The amount of sequence contained in the 722, 271, and 129 regions identified in

Box 3. Statistical tests of neutrality

Statistical tests of neutrality can broadly be divided into three main classes based on the type of data they use: tests based on within-species polymorphism, tests based on divergence between species, and tests that use both polymorphism within and divergence between species (Biswas and Akey 2006). Below, tests of within-species polymorphism are focused on, as these are most directly relevant to the genome-wide scans of positive selection considered in Table 1.

Site frequency spectrum
This class of tests summarizes the allele frequency distribution of polymorphisms in a region of interest. In general, selective sweeps originating from newly arisen advantageous alleles result in an excess of low frequency alleles (and in the presence of recombination an excess of high frequency derived alleles) relative to neutral expectations and are most powerful in detecting recently completed sweeps (Simonsen et al. 1995). However, they have little power in detecting soft sweeps originating from existing variation (Przeworski et al. 2005). In addition to popular summary statistic methods, such as Tajima's D (Tajima 1989), composite likelihood approaches that make fuller use of the data have been developed (Kim and Stephan 2002; Nielsen et al. 2005; Zhu and Bustamante 2005) and will likely play an increasingly important role in making inferences about selection based on genome-wide patterns of genetic variation.

Linkage disequilibrium
Linkage disequilibrium (LD) refers to the nonrandom association of alleles between two or more loci. An expected signature of an ongoing or incomplete selective sweep is the presence of a high-frequency haplotype with extended LD, because recombination will have little opportunity to occur during the rapid increase in frequency of a haplotype carrying an advantageous allele. Popular LD based tests include rEHH (relative extended haplotype homozygosity) (Sabeti et al. 2002), iHS (integrated haplotype score) (Voight et al. 2006), and LDD (linkage disequilibrium decay test) (Wang et al. 2006). Once a sweep is completed, or nearly complete, LD methods rapidly lose power as little variation is left from which to assess patterns of LD. In addition, a number of test statistics related to rEHH and iHS (Kimura et al. 2007; Sabeti et al. 2007; Tang et al. 2007) have been developed to compare the extent of LD in a particular genomic region between populations, which may be particularly useful in detecting geographically restricted selection (see below) and in some cases retains power to detect population-specific completed sweeps (Kimura et al. 2007).

Population differentiation
Most natural populations exhibit some degree of population structure, potentially allowing geographically restricted selection to occur. In this scenario, an advantageous allele arises only in a subset (or single) subpopulation, or the fitness of an existing allele changes upon being exposed to a new environmental niche. The simplest, and most popular, statistic used to detect local increases in the magnitude of population structure due to geographically restricted selection is F_{ST} (Weir 1996). Several variants of the classic F_{ST} statistic have been developed and applied to genome-wide scans of selection in humans, such as population-specific F_{ST} statistics (Shriver et al 2004; Weir et al. 2005), as well as more computationally sophisticated Bayesian methods of inference (Beaumont and Balding 2004).

two, three, and four or more studies is 245, 148, and 92 Mb, respectively. Thus, ~60%, 36%, and 22% of the total sequence encompassing the integrated map is supported by two, three, and four or more studies, respectively. This paradoxical observation, little overlap in the number of regions but considerably more so of sequence, is due to the marked difference in the average size of regions identified in single versus multiple studies (~80 kb and 300 kb, respectively). I suggest that these results are consistent with at least two mutually compatible explanations. First, loci deemed as an outlier in multiple analyses are more likely to represent the most dramatic selective events, which in general will tend to leave larger footprints in the genome because they are some combination of young, strong, or in regions of low recombination. Second, although multiple studies may demark similar loci, they home in on different positions within the larger genomic interval formed after merging overlapping signatures of selection. In this respect, there seems to be considerable promise in using the chromosomal distribution of various neutrality test statistics to more precisely map targets of positive selection.

Despite the above considerations, there is no escaping the general conclusion that the overlap among studies is underwhelming. However, this is unsurprising for a number of reasons. For example, a number of neutrality test statistics have been used to scan the genome for signatures of positive selection (Box 3), which are likely recovering selective events from different time periods and for different stages of the selective sweep (Box 2) (see also Biswas and Akey 2006; Sabeti et al. 2006). Furthermore, studies tend to only report the most extreme loci. This conceivably has the effect of reducing overlap among studies if a locus is deemed an outlier in one analysis because it falls in the 99th percentile of the empirical distribution, but is not called an outlier in another study where it falls in the 98th percentile of a different empirical distribution. Finally, and perhaps most importantly, several simulation studies have shown that outlier approaches likely suffer from low power and high false positive rates (Kelley et al. 2006; Teshima et al. 2006). Unfortunately, the actual power and false positive rates depend on a large number of parameters that are difficult to estimate, such as the fraction of the genome under selection; strength of selection; whether adaptive alleles are recessive, dominant, or additive; and whether selection acts on newly arisen versus preexisting variation. Additional theoretical studies of outlier approaches, systematically comparing a broad set of commonly used neutrality test statistics under a range of demographic and selective models, would be invaluable for guiding efficient study designs in genome-wide scans of selection.

In summary, genome-wide analyses of positive selection described to date suggest widespread signatures of positive selection in the human genome. Although these newly constructed maps of selection likely include genuine substrates of positive selection, they also likely possess many false positives and false negatives, and thus considerable caution is needed when interpreting such maps.

What have we learned from recent genome-wide scans of selection?

Although the interpretation of genome-wide scans of selection is hampered by the low-resolution, incomplete, and error-prone maps described above, careful inspection of the results across studies allows several general themes to begin to come into focus. Here, I discuss general insights derived by analyzing the higher confidence selected loci that are supported by two or more studies (Supplemental Table 1).

Familiar friends, new faces

The 722 regions that have been identified in multiple genome-wide scans contain 2465 genes, a number of which have been previously implicated as targets of positive selection. Examples

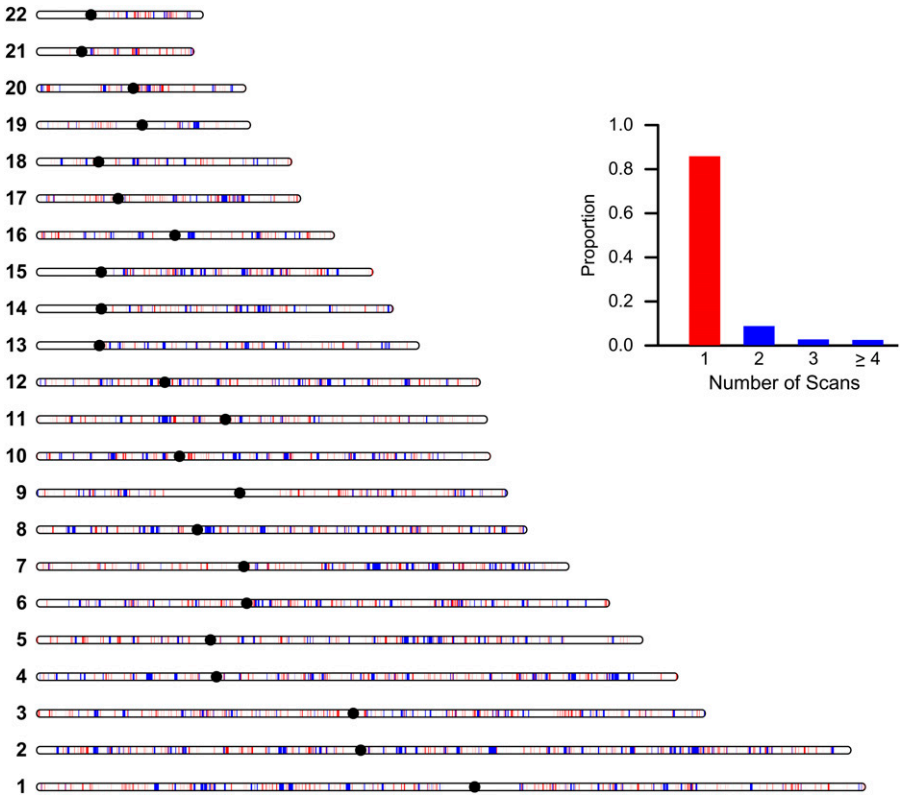

Figure 2. Integrated genomic map of positive selection. Vertical red lines on each autosome indicate loci that were identified in a single genome-wide scan, and blue lines denote regions identified in two or more studies. The histogram shows the proportion of putatively selected loci (*y*-axis) as a function of the number of genome-wide scans in which they were identified (*x*-axis).

include *LCT* (Bersaglieri et al. 2004), *TRPV6* (Akey et al. 2004, 2006), *CYP3A* (Thompson et al. 2004), *CYP1A2* (Wooding et al. 2002), *IL13* (Zhou et al. 2004), and *IL4* (Rockman et al. 2003) among others. Perhaps more interestingly, many new well-supported genes emerge that would not necessarily be strong a priori candidate genes of selection to study. For instance, in a rare example of multiple analyses converging on a single gene, *PCDH15* was identified in six out of the nine genome-wide scans. Mutations in *PCDH15*, which plays a critical role in retinal and cochlear function, can result in Usher syndrome type IF and Autosomal Recessive Deafness 23 (Ahmed et al. 2003). Interestingly, three myosin genes (*MYO1B*, *MYO3A*, and *MYO6*) that are integral in cochlear function (Dumont et al. 2002; Walsh et al. 2002; Sanggaard et al. 2008) are also among the set of loci supported by multiple analyses.

In addition to analyses of individual loci, several interesting observations emerge from examining the general functional classes of these 2465 genes. Table 2 shows PANTHER Biological Process terms (Thomas et al. 2003) that are overrepresented among the set of genes identified in multiple genome-wide analyses. One of the more striking observations in Table 2 is the dominant role that positive selection on metabolic processes seems to have played in recent human evolutionary history. For example, a significant overrepresentation is observed for terms such as protein modification, protein metabolism, carbohydrate metabolism, and phosphate metabolism. Although this observation is in accord with known dramatic shifts in diet during recent human history (Larsen 1995), the pervasive signature of positive selection across so many metabolic processes has not generally been appreciated. The other interesting point gleaned from a cursory examination of Table 2 is that far from acting on a few classes of genes, positive selection appears to have affected a wide variety of biological processes.

Spatially varying selection

A recurring observation of genome-wide studies is that signatures of positive selection are not uniformly distributed across populations, but rather show clear spatial heterogeneity (i.e., local adaptation). In particular, ~80% of the 722 loci observed in multiple scans show evidence of local adaptation. This is consistent with the large number of previous single gene studies describing spatially varying patterns of selection (for review, see Ronald and Akey 2005). The observation of widespread local adaptation is not surprising given the environmental heterogeneity that human populations are confronted with throughout the world.

Although the evidence that local adaptation has played a prominent role in recent human evolutionary history is compelling, some caution is required when comparing signatures of selection across populations. In particular, genome-wide scans are essentially searching for positions in the genome that have a large scaled selection coefficient, $4N_e s$, where N_e is the effective population size and s is the magnitude of selection. As N_e is influenced by population demographic history, the signature of selection is jointly determined by both the strength of selection and demographic history, as well as local rates of recombination and mutation (Kaplan et al. 1989). Thus, population differences in any of these parameters can influence whether a locus appears to be under selection in a single population or multiple populations.

A specific example of the difficulties in interpreting signatures of spatially varying selection is the observation that non-African populations tend to show more evidence for recent positive selection relative to African populations (Akey et al. 2004; Storz et al. 2004; Williamson et al. 2007; but see Voight et al. 2006). While this may be due to increased selection as humans migrated out of Africa and were confronted with new environmental pressures (such as novel climates, diets, and pathogens), differences in demographic history or rates of recombination and mutation between African and non-African populations may obscure the relationship between signatures of selection across populations. Until a wider set of African populations are studied, inferences about the relative frequency of positive selection between African and non-African populations and patterns of shared selective events will remain speculative.

Regulatory versus protein adaptive evolution

An ongoing debate, now over three decades old and still going strong (King and Wilson 1975; Hoekstra and Coyne 2007; Wray

Table 2. Enriched PANTHER biological process terms

Biological process	*P*-value[a]
Protein modification	9.12×10^{-15}
Signal transduction	1.93×10^{-11}
Protein phosphorylation	4.83×10^{-11}
Protein metabolism and modification	7.19×10^{-11}
Developmental processes	2.66×10^{-8}
Olfaction	4.46×10^{-6}
Chemosensory perception	1.53×10^{-6}
Cell adhesion–mediated signaling	3.3×10^{-5}
Nucleoside, nucleotide, and nucleic acid metabolism	1.42×10^{-4}
Cell cycle	1.95×10^{-4}
Cell adhesion	2.61×10^{-4}
Mesoderm development	4.00×10^{-4}
Other metabolism	6.81×10^{-4}
Cell communication	1.06×10^{-3}
Intracellular protein traffic	1.11×10^{-3}
Other intracellular signaling cascade	1.64×10^{-3}
Cation transport	2.17×10^{-3}
Proteolysis	2.41×10^{-3}
Ion transport	2.61×10^{-3}
Neuronal activities	3.72×10^{-3}
Synaptic transmission	6.35×10^{-3}
Transport	6.44×10^{-3}
Carbohydrate metabolism	8.34×10^{-3}
Cell cycle control	9.29×10^{-3}
Other carbohydrate metabolism	9.79×10^{-3}
Protein acetylation	1.21×10^{-2}
Cell surface receptor mediated signal transduction	1.73×10^{-2}
Cell proliferation and differentiation	2.81×10^{-2}
Phosphate metabolism	3.12×10^{-2}

[a]*P*-values were adjusted for multiple testing by Bonferroni corrections.

2007), is the relative contribution of changes in gene regulation versus protein structure as mechanisms of evolutionary change. The poor resolution of current genomic maps of selection precludes definitive inferences about the proportion of recent positive selection in humans due to regulatory versus coding evolution. However, we can gain some insight by considering the genic content of the 722 regions supported by multiple studies. In total, 101 of the 722 regions (~14%) contain no UCSC RefSeq protein coding genes and are therefore attractive candidates for harboring adaptive regulatory variation. Obviously, selected regions that overlap protein coding genes may also be the result of regulatory changes, so 14% provides a rough lower bound on the number of well-supported regions in the integrated map of positive selection that are driven by regulatory evolution.

An interesting example of a nonprotein-containing region is a 150-kb interval on chromosome 20 downstream from *BMP2*, which is a member of the transforming growth factor beta superfamily involved in bone and cartilage formation (Wang et al. 1990). In mice, an enhancer in the 3′ region of *BMP2* has been identified that influences expression in osteoblast progenitor cells (Chandler et al. 2007), and in humans, SNPs in the 3′ region are associated with otosclerosis (Schrauwen et al. 2008), a progressive disease of the temporal bone that can lead to hearing loss. Additional bioinformatics analyses of the *BMP2* downstream region, and the remaining 100 nongenic regions, may provide valuable insights into distinguishing characteristics of these loci, guide experimental studies, and generate testable evolutionary hypotheses.

Genetic draft and "off-target" effects

Genomic maps of selection suggest widespread genetic hitchhiking (Box 2) throughout the genome. Although the veracity of this statement is subject to the limitations described above, it is fair to say that the number of strong selective events thought to exist in the human genome today is considerably more than that imagined less than a decade ago. Again, restricting our attention to the 722 loci identified in two or more genome-wide scans, ~245 Mb (~8%) of the genome has been influenced by positive selection, and an even larger fraction may have been subject to more modest selective pressure.

If a substantial fraction of the genome has indeed been influenced by positive selection, this would have important theoretical and practical implications. For example, it may be necessary to consider models, such as genetic draft (Gillespie 2000), where the stochastic population genetic dynamics of neutral variation is governed more by the indirect effects of selection on adaptive alleles than by genetic drift (the primary force governing levels of variation within and between populations in neutral and nearly neutral models) (Kimura 1983; Ohta and Gillespie 1996). In this vein, it will also be important to incorporate background selection (the effect of purifying selection on linked patterns of variation) (Charlesworth et al. 1993; Reed et al. 2005) into drift and draft models to better understand the causes and consequences of genomic patterns of human genetic variation.

From a more practical perspective, it will be of considerable interest to determine how often genetic hitchhiking has influenced the population genetic characteristics of neutral or nearly neutral alleles that contribute to phenotypic variation. For instance, ~10% of the significant results from recent genome-wide association studies (GWAS) summarized in the NHGRI GWAS catalog (Hindorff et al. 2009) are located in the higher confidence selected regions. As this is not significantly more than expected by chance (*P*-value >0.05), it seems unlikely that all significant GWAS results have been direct targets of selection, implying that the frequency and distribution of at least some alleles contributing to human phenotypic variation and disease susceptibility may have been indirectly influenced by positive selection.

Outliers are simple, their evolutionary history may not be

The typical outcome of a genome-wide scan for positive selection is a list of loci that have been categorized as either "an outlier" or "not an outlier." This simple binary classification belies the complex evolutionary history that outlier loci may have experienced. Previous candidate gene analyses provide a glimpse into the types of complexity that may be found when the signature of selection at individual outlier loci is studied in more detail. Examples include convergent evolution (Lamason et al. 2005; Tishkoff et al. 2007; Enattah et al. 2008), selection on standing variation (Hamblin et al. 2002; Enattah et al. 2008; Magalon et al. 2008), evidence for both directional and balancing selection acting on the same locus (Tishkoff et al. 2001; Verrelli et al. 2002), and epistatic selection (Williams et al. 2005).

Hints of complexity are already apparent in the results from genome-wide scans. For instance, ~8% of the regions supported by multiple studies show evidence of selection in both European and African samples, suggesting these loci may have experienced independent selective pressures. Furthermore, we have recently performed a detailed population genetics analysis of *ALMS1*, which occurs in a region identified in seven out of the nine genome-wide scans, and found compelling evidence for geographically restricted selection, selection from standing variation, and three ancient and divergent haplogroup lineages (Scheinfeldt et al. 2009). Thus, when subjected to further scrutiny, outlier loci

will not be uniformly explained by classic models of a newly arisen advantageous mutation sweeping to fixation (Maynard-Smith and Haigh 1974). Rather, they will likely be the result of a range of evolutionary models, a deeper understanding of which will provide basic insights into the mechanistic details of how natural populations adapt.

Back to the future: A return to candidate gene studies?

Earlier, I espoused the virtues of population genomics over candidate gene approaches. However, after the dust settles from genome-wide scans of selection, we are left with many regions that possess unusual patterns of variation consistent with the hypothesis of selection and perhaps a rough estimate of when and where selection occurred. Genome-wide scans are a powerful beginning but are clearly not the end toward developing a detailed, precise, and mechanistic understanding of human evolutionary history. In other words, genome-wide scans are a hatchet, whereas what we need now is a scalpel. In-depth follow-up studies of individual outlier loci can be one such scalpel, more precisely defining important population genetic parameters such as the timing and magnitude of selection, the geographic distribution of selected variation, the interaction of population demographic history, recombination, and selection in shaping patterns of variation, and the functional form of selection acting on individual outlier loci.

However, follow-up studies of outlier loci offer several new methodological challenges. For instance, it is necessary to carefully consider how hypothesis testing is performed, as the study of outlier loci introduces an ascertainment bias that needs to be properly taken into account (Kreitman and Di Rienzo 2004; Thornton and Jensen 2007). As a concrete example, consider a hypothetical locus that is an outlier in the empirical distribution of F_{ST} derived from the HapMap data. In a follow-up resequencing study in the same samples, additional neutrality test statistics, such as Tajima's D, are calculated. Because F_{ST} and Tajima's D at individual loci are not independent, it would be misleading to evaluate the statistical significance of the latter without taking into account the initial ascertainment on strong population structure. Approaches have already been developed to address issues of ascertainment bias encountered in follow-up studies of selected loci (Thornton and Jensen 2007), and additional work in this area would provide further insights into how best to design, analyze, and interpret follow-up studies of outlier loci.

In short, although delving into the minutiae of individual outlier loci is perhaps less glamorous than the initial genome-wide analysis, it is a necessary step toward developing a coherent principled narrative of recent human evolutionary history. The analysis of individual outlier loci differs, however, in important conceptual and methodological ways from earlier candidate gene approaches and therefore should not be viewed as simply a return to candidate gene studies.

The missing (phenotype) link

A significant impediment to understanding and interpreting signatures of positive selection, and ultimately identifying adaptive alleles, is that we are often ignorant about the phenotype the selected locus influences. This outcome is a direct consequence of the "bottom-up" strategy used in genome-wide scans for selection, where inferences are made directly from patterns of genetic variation (Fig. 3). However, the direct substrate of selection is phenotypic variation that influences fitness. Thus, the relationship between positive selection and patterns of genetic variation depends on the underlying genetic architecture of phenotypes, which is increasingly being cast in a systems biology framework (Benfey and Mitchell-Olds 2008; Ellegren and Sheldon 2008). A better understanding of how genetic variation influences variation in molecular networks, which interact with each other and the environment to shape patterns of phenotypic variation, would significantly accelerate the interpretation of signatures of positive selection (Fig. 3; Ellegren and Sheldon 2008).

An excellent recent example of how phenotypic context can facilitate a deeper understanding of selection is *SLC24A5* (Lamason et al. 2005). In a mutant screen, the investigators showed the zebrafish homolog of *SLC24A5* affects pigmentation. Next, they examined patterns of genetic variation for this gene in the HapMap data and found a dramatic signature of positive selection (Lamason et al. 2005). Guided by the zebrafish data and evidence for positive selection in humans, they went on to demonstrate that a nonsynonymous SNP in *SLC24A5* influenced differences in pigmentation levels between individuals of West

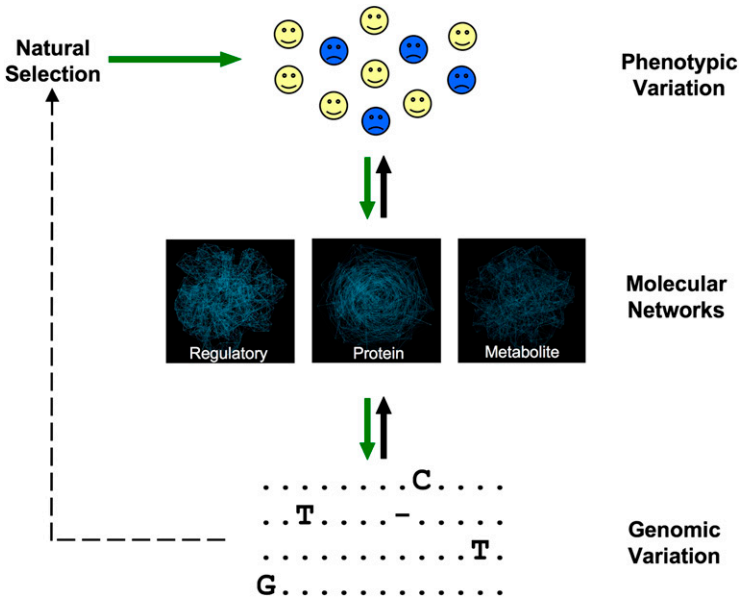

Figure 3. Bottom-up population genomics. Genome-wide scans of positive selection are agnostic to phenotypic data and make inferences of selection directly from patterns of genetic variation (dashed black arrow). However, selection acts directly on phenotypic variation and only indirectly on DNA sequence variation (dark green arrows). Solid black arrows show that the path from genetic to phenotypic variation runs through dynamic molecular networks (such as regulatory, protein, and metabolite). Scale-free molecular networks were simulated with the R package igraph and visualized in CytoScape (Cline et al. 2007).

African and European ancestry. It is worth noting that *SLC24A5* is among the most well-supported loci in the integrated map of positive selection described above (Fig. 2), having been identified in six out of the nine genome-wide scans. Without prior knowledge that *SLC24A5* was a pigmentation gene, there would have been little impetus for performing the subsequent association analyses demonstrating its phenotypic effect in humans, leaving this locus as an anonymous outlier without all of the deep biological and evolutionary insights that now exist. Although the details differ, a common theme underlying almost every other well-understood instance of positive selection in humans (i.e., *LCT*, *G6PD*, *HBB*, *DARC*, etc.) is that something was known about the phenotype these genes influenced.

Aside from facilitating interpretations of positively selected loci, a deeper understanding of phenotypic variation in a systems biology framework will also expand the scope of evolutionary inferences that are possible. Specifically, genome-wide scans for recent positive selection only have reasonable power to detect fairly strong selective effects ($4N_e s \sim 400$) (Kelley et al. 2006; Teshima et al. 2006). Thus, the current catalog of positively selected loci identified through genome-wide scans represents only the tip of the selective iceberg. At this time, there seems little reason for optimism in detecting weak positive selection acting at a single locus. However, it may be possible to increase the power to detect more subtle selection by combining information across loci, allowing inferences about the influence of adaptive evolution on specific pathways or modules (Hancock et al. 2008).

Is Darwinian evolution enough?

The basic tenets that contemporary evolutionary studies, including genome-wide scans of selection, operate under are encapsulated in what is referred to as the modern synthesis. Excellent and detailed reviews on the origins and scope of the modern synthesis, can be found elsewhere (Provine 1971), but can succinctly be described as the coalescence of Darwinian selection, theoretical population genetics, and Mendelian principles into a unified account of how populations and species evolve. As new insights into basic biological mechanisms and data accumulated over the decades, particularly in the genomics era, a slow but steady call has been made to extend the modern synthesis (e.g., Kutschera and Niklas 2004; Pigliucci 2007, and references therein). General areas that are fueling the call for an expanded evolutionary synthesis include the evolution of evolvability, epigenetic inheritance, phenotypic plasticity, and the origins of complexity (for review, see Pigliucci 2007). Although these ideas, and more specifically their interpretation, are controversial (Pennisi 2008) and ultimately may not necessitate a major extension to the modern synthesis, they should not be categorically dismissed. Rather, they should be subjected to increased scrutiny, investigating on a case-by-case basis their contribution to evolutionary processes and relationships with currently held evolutionary paradigms.

To date, genome-wide scans have provided little insight into issues that potentially extend beyond the framework of the modern synthesis. Whether this is because they are not well suited for addressing such issues or because they have not been used to ask relevant questions remains to be determined. For instance, the recent observation that SNPs disrupting CpG sites can have unexpectedly large influences on the methylation status of a region (Kerkel et al. 2008) provides a link between heritable genetic variation and epigenetic variation. Perhaps some of these methylation altering SNPs are adaptive and thus could potentially be identified if explicitly looked for in genome-wide scans for positive selection. Although this is admittedly a nebulous example, it highlights the point that detailed analyses of outlier loci, guided by specific hypotheses, may be a powerful avenue for elucidating fundamental mechanisms of evolutionary change.

Future directions

Genomics has unquestionably and profoundly changed the field of human evolutionary genetics. Genome-wide scans have provided coarse maps of positive selection in humans, maps that may ultimately yield a deeper understanding into mechanisms of adaptive change. However, a key now is to begin traversing the path from the low-resolution, incomplete, and error-prone maps of selection today to the ultimate goal of a detailed molecular, mechanistic, phenotypic, and population genetics characterization of adaptive alleles. How do we get from here to there?

Looking ahead, we can expect a continuing deluge of data, perhaps the most exciting of which are whole-genome sequences from thousands of geographically diverse individuals (Wise 2008). Although whole-genome sequences from thousands of individuals will approach the limits of complete genetic information, how much closer will it get us to our ultimate goals? Many of the already performed genome-wide analyses (Table 1) will and should be repeated on these data sets, because of their greater information content and lack of ascertainment bias, which has hampered evolutionary analyses of SNP data (Akey et al. 2003; Clark et al. 2005). These analyses will allow in-depth population genetic studies of already and soon to be identified outlier loci and a better understanding of how putatively selected variation is apportioned within and among populations.

Although important, these lines of investigation, in and of themselves, will not suffice for at least two reasons. First, there is considerable need for further theoretical and methodological research. Specific areas of fruitful inquiry include a more comprehensive statistical characterization of neutrality tests to a wider range of selective and demographic models, developing more realistic models of how selection operates in natural populations and methods to detect it (Orr and Betancourt 2001; Wakeley 2004; Hermisson and Pennings 2005; Przeworski et al. 2005), and models and methods of analysis for understanding adaptive evolution in the context of systems biology.

Second, and more fundamentally, the statistical analysis of DNA sequence variation, or any single approach, cannot provide a complete description of human evolutionary history. Indeed, as evolution is an inherently stochastic process and the result of a series of historical contingencies (Lewontin 1966; Jacob 1977), there are likely questions that may never be satisfactorily answered. However, an account of what is possible to know will be incomplete until the functional consequences of genetic variation can be determined in a high-throughput and comprehensive manner; until a deeper appreciation of how genetic variation perturbs regulatory and protein networks is attained; until the genetic and environmental architecture of phenotypic variation is elucidated; and until cultural and ecological aspects of human populations past and present are better delimited. In other words, it will require the continued efforts from all branches of science in increasingly synergistic and interdisciplinary ways.

Acknowledgments

I thank members of the Akey laboratory, particularly Laura Scheinfeldt and Shameek Biswas. I thank Matt Maurano for

performing the analysis of the GWAS data. This work was supported by a research grant (1R01GM076036-01A1) from the NIH and a Sloan Fellowship in Computational Biology to J.M.A.

References

Ahmed, Z.M., Riazuddin, S., Ahmad, J., Bernstein, S.L., Guo, Y., Sabar, M.F., Sieving, P., Griffith, A.J., Friedman, T.B., Belyantseva, I.A., et al. 2003. *PCDH15* is expressed in the neurosensory epithelium of the eye and ear and mutant alleles are responsible for both USH1F and DFNB23. *Hum. Mol. Genet.* **12:** 3215–3223.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12:** 1805–1814.

Akey, J.M., Zhang, K., Xiong, M., and Jin, L. 2003. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20:** 232–242.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., and Kruglyak, L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* **2:** e286. doi: 10.1371/journal.pbio.0020286.

Akey, J.M., Swanson, W.J., Madeoy, J., Eberle, M., and Shriver, M.D. 2006. *TRPV6* exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Hum. Mol. Genet.* **15:** 2106–2113.

Bamshad, M. and Wooding, S.P. 2003. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4:** 99–111.

Beaumont, M.A. and Balding, D.J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* **13:** 969–980.

Benfey, P.N. and Mitchell-Olds, T. 2008. From genotype to phenotype: Systems biology meets natural variation. *Science* **320:** 495–497.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74:** 1111–1120.

Biswas, S. and Akey, J.M. 2006. Genomic insights into positive selection. *Trends Genet.* **22:** 437–446.

Black IV, W.C., Baer, C.F., Antolin, M.F., and DuTeau, N.M. 2001. Population genomics: Genome-wide sampling of insect populations. *Annu. Rev. Entomol.* **46:** 441–469.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J., and Nickerson, D.A. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15:** 1553–1565.

Cavalli-Sforza, L. 1966. Population structure and human evolution. *Proc. R. Soc. Lond. Ser. B.* **164:** 362–379.

Chandler, R.L., Chandler, K.J., McFarland, K.A., and Mortlock, D.P. 2007. *Bmp2* transcription in osteoblast progenitors is regulated by a distant 3′ enhancer located 156.3 kilobases from the promoter. *Mol. Cell. Biol.* **27:** 2934–2951.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15:** 1496–1502.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2:** 2366–2382.

Crow, J.F. 1994. Foreword. In *Population genetics, molecular evolution and the neutral theory: Selected papers* (ed. M. Kimura), pp. xiii–xv. University of Chicago Press, Chicago, IL.

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R.M., Clegg, J.B., Langaney, A., and Excoffier, L. 2002. Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta^S Senegal mutation. *Am. J. Hum. Genet.* **70:** 207–223.

Darwin, C. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life,* 1st ed. John Murray, London, UK.

Darwin, C.R. and Wallace, A.R. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Proc. Linnean Soc. Lond. Zool.* **3:** 46–50.

Dumont, R.A., Zhao, Y.D., Holt, J.R., Bahler, M., and Gillespie, P.G. 2002. Myosin-I isozymes in neonatal rodent auditory and vestibular epithelia. *J. Assoc. Res. Otolaryngol.* **3:** 375–389.

Ellegren, H. and Sheldon, B.C. 2008. Genetic basis of fitness differences in natural populations. *Nature* **452:** 169–175.

Enattah, N.S., Jensen, T.G., Nielsen, M., Lewinski, R., Kuokkanen, M., Rasinpera, H., El-Shanti, H., Seo, J.K., Alifrangis, M., Khalil, I.F., et al. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am. J. Hum. Genet.* **82:** 57–72.

Friedman, M.J. 1978. Erythrocytic mechanism of sickle cell resistance to malaria. *Proc. Natl. Acad. Sci.* **75:** 1994–1997.

Gillespie, J.H. 1991. *The causes of molecular evolution.* Oxford University Press, New York.

Gillespie, J.H. 1998. *Population genetics: A concise guide.* The Johns Hopkins University Press, Baltimore, MD.

Gillespie, J.H. 2000. Genetic drift in an infinite population: The pseudohitchhiking model. *Genetics* **155:** 909–919.

Haldane, J.B.S. 1927. The mathematical theory of natural and artificial selection. *Proc. Camb. Philol. Soc.* **23:** 838–844.

Hamblin, M.T. and Di Rienzo, A. 2000. Detection of the signature of natural selection in humans: Evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66:** 1669–1679.

Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70:** 369–383.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G., and Di Rienzo, A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* **4:** e32. doi: 10.1371/journal.pgen.0040032.

Harris, E.E. and Meyer, D. 2006. The molecular signature of selection underlying human adaptations. *Am. J. Phys. Anthropol.* **131** (Suppl. 43): 89–130.

Hedges, S.B. 2000. Human evolution. A start for population genomics. *Nature* **408:** 652–653.

Hermisson, J. and Pennings, P.S. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169:** 2335–2352.

Hindorff, L.A., Junkins, H.A., Mehta, J.P., and Manolio, T.A. 2009. A catalog of published genome-wide association studies. www.genome.gov/26525384.

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

Hoekstra, H.E. and Coyne, J.A. 2007. The locus of evolution: Evo devo and the genetics of adaptation. *Evolution Int. J. Org. Evolution* **61:** 995–1016.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

Jacob, F. 1977. Evolution as tinkering. *Science* **196:** 1161–1166.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451:** 998–1003.

Johansson, A. and Gyllensten, U. 2008. Identification of local selective sweeps in human populations since the exodus from Africa. *Hereditas* **145:** 126–137.

Kaplan, N.L., Hudson, R.R., and Langley, C.H. 1989. The "hitchhiking effect" revisited. *Genetics* **123:** 887–899.

Kayser, M., Brauer, S., and Stoneking, M. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol. Biol. Evol.* **20:** 893–900.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., and Akey, J.M. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* **16:** 980–989.

Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E., et al. 2008. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.* **40:** 904–908.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453:** 56–64.

Kim, Y. and Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160:** 765–777.

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

Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, New York.

Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. 2007. A practical genome scan for population-specific strong selective sweeps that have

reached fixation. *PLoS One* **2:** e286. doi: 10.1371/journal.pone.0000286.
Kimura, R., Ohashi, J., Matsumura, Y., Nakazawa, M., Inaoka, T., Ohtsuka, R., Osawa, M., and Tokunaga, K. 2008. Gene flow and natural selection in oceanic human populations inferred from genome-wide SNP typing. *Mol. Biol. Evol.* **25:** 1750–1761.
King, J.L. and Jukes, T.H. 1969. Non-Darwinian evolution. *Science* **164:** 788–798.
King, M.C. and Wilson, A.C. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116.
Kreitman, M. 2000. Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1:** 539–559.
Kreitman, M. and Di Rienzo, A. 2004. Balancing claims for balancing selection. *Trends Genet.* **20:** 300–304.
Kutschera, U. and Niklas, K.J. 2004. The modern theory of biological evolution: An expanded synthesis. *Naturwissenschaften* **91:** 255–276.
Lamason, R.L., Mohideen, M.A., Mest, J.R., Wong, A.C., Norton, H.L., Aros, M.C., Jurynec, M.J., Mao, X., Humphreville, V.R., Humbert, J.E., et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310:** 1782–1786.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Larsen, C.S. 1995. Biological changes in human populations with agriculture. *Annu. Rev. Anthropol.* **24:** 185–213.
Lewontin, R.C. 1966. Is nature probable or capricious? *Bioscience* **16:** 25–27.
Lewontin, R.C. and Krakauer, J. 1978. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74:** 175–195.
Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.
Livingston, R.J., von Niederhausern, A., Jegga, A.G., Crawford, D.C., Carlson, C.S., Rieder, M.J., Gowrisankar, S., Aronow, B.J., Weiss, R.B., and Nickerson, D.A. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14:** 1821–1831.
Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L., and Heyer, E. 2008. Population genetic diversity of the *NAT2* gene supports a role of acetylation in human adaptation to farming in central Asia. *Eur. J. Hum. Genet.* **16:** 243–251.
Mattiangeli, V., Ryan, A.W., McManus, R., and Bradley, D.G. 2006. A genome-wide approach to identify genetic loci with a signature of natural selection in the Irish population. *Genome Biol.* **7:** R74. doi: 10.1186/gb-2006-7-8-r74.
Maynard-Smith, J.M. and Haigh, J. 1974. The hitchhiking effect of a favourable gene. *Genet. Res.* **23:** 23–35.
Nielsen, R. 2001. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86:** 641–647.
Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39:** 197–218.
Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., and Bustamante, C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* **15:** 1566–1575.
Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. 2007. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8:** 857–868.
Ohta, T. and Gillespie, J.H. 1996. Development of neutral and nearly neutral theories. *Theor. Popul. Biol.* **49:** 128–142.
Oleksyk, T.K., Zhao, K., De La Vega, F.M., Gilbert, D.A., O'Brien, S.J., and Smith, M.W. 2008. Identifying selected regions from heterozygosity and divergence using a light-coverage genomic data set from two human populations. *PLoS One* **3:** e1712. doi: 10.1371/journal.pone.0001712.
O'Reilly, P.F., Birney, E., and Balding, D.J. 2008. Confounding between recombination and selection, and the Ped/Pop method for detecting selection. *Genome Res.* **18:** 1304–1313.
Orr, H.A. and Betancourt, A.J. 2001. Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157:** 875–884.
Payseur, B.A., Cutter, A.D., and Nachman, M.W. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19:** 1143–1153.
Pennisi, E. 2008. Evolution. Modernizing the modern synthesis. *Science* **321:** 196–197.
Pigliucci, M. 2007. Do we need an extended evolutionary synthesis? *Evolution Int. J. Org. Evolution* **61:** 2743–2749.
Provine, W.B. 1971. *The origins of theoretical population genetics*. University of Chicago Press, Chicago, IL.
Przeworski, M., Hudson, R.R., and Di Rienzo, A. 2000. Adjusting the focus on human variation. *Trends Genet.* **16:** 296–302.
Przeworski, M., Coop, G., and Wall, J.D. 2005. The signature of positive selection on standing genetic variation. *Evolution Int. J. Org. Evolution* **59:** 2312–2323.
Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.
Reed, F.A., Akey, J.M., and Aquadro, C.F. 2005. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. *Genome Res.* **15:** 1211–1221.
Rockman, M.V., Hahn, M.W., Soranzo, N., Goldstein, D.B., and Wray, G.A. 2003. Positive selection on a human-specific transcription factor binding site regulating *IL4* expression. *Curr. Biol.* **13:** 2118–2123.
Ronald, J. and Akey, J.M. 2005. Genome-wide scans for loci under selection in humans. *Hum. Genomics* **2:** 113–125.
Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419:** 832–837.
Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. 2006. Positive natural selection in the human lineage. *Science* **312:** 1614–1620.
Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449:** 913–918.
Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Sanggaard, K.M., Kjaer, K.W., Eiberg, H., Nürnberg, G., Nürnberg, P., Hoffman, K., Jensen, H., Sørum, C., Rendtorff, N.D., and Tranebjaerg, L. 2008. A novel nonsense mutation in *MYO6* is associated with progressive nonsyndromic hearing loss in a Danish *DFNA22* family. *Am. J. Med. Genet. A* **146A:** 1017–1025.
Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15:** 1576–1583.
Scheinfeldt, L.B., Biswas, S., Madeoy, J., Connelly, C.F., Schadt, E.E., and Akey, J.M. 2009. Population genomics analysis of *ALMS1* in humans reveals a surprisingly complex evolutionary history. *Mol. Bio. Evol.* (in press). doi: 10.1093/molbev/msp045.
Schrauwen, I., Thys, M., Vanderstraeten, K., Fransen, E., Dieltjens, N., Huyghe, J.R., Ealy, M., Claustres, M., Cremers, C.R., Dhooge, I., et al. 2008. Association of bone morphogenetic proteins with otosclerosis. *J. Bone Miner. Res.* **23:** 507–516.
Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., and Jones, K.W. 2004. The genomic distribution of population substructure in four populations using 8525 autosomal SNPs. *Hum. Genomics* **1:** 274–286.
Simonsen, K.L., Churchill, G.A., and Aquadro, C.F. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141:** 413–429.
Stajich, J.E. and Hahn, M.W. 2005. Disentangling the effects of demography and selection in human history. *Mol. Biol. Evol.* **22:** 63–73.
Storz, J.F., Payseur, B.A., and Nachman, M.W. 2004. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol. Biol. Evol.* **21:** 1800–1811.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.
Tang, K., Thornton, K.R., and Stoneking, M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5:** e171. doi: 10.1371/journal.pbio.0050171.
Teshima, K.M., Coop, G., and Przeworski, M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* **16:** 702–712.
Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13:** 2129–2141.
Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. 2004. *CYP3A* variation and the evolution of salt-sensitivity variants. *Am. J. Hum. Genet.* **75:** 1059–1069.
Thorisson, G.A. and Stein, L.D. 2003. The SNP Consortium website: Past, present and future. *Nucleic Acids Res.* **31:** 124–127.
Thornton, K.R. and Jensen, J.D. 2007. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics* **175:** 737–750.
Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. 2001. Haplotype diversity and linkage disequilibrium at human *G6PD*: Recent origin of alleles that confer malarial resistance. *Science* **293:** 455–462.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39:** 31–40.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Verrelli, B.C., McDonald, J.H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A.N., Loiselet, J., and Tishkoff, S.A. 2002. Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. *Am. J. Hum. Genet.* **71:** 1112–1128.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* **4:** e72. doi: 10.1371/journal.pbio.0040072.

Wakeley, J. 2004. Recent trends in population genetics: More data! More math! Simple models? *J. Hered.* **95:** 397–405.

Walsh, T., Walsh, V., Vreugde, S., Hertzano, R., Shahin, H., Haika, S., Lee, M.K., Kanaan, M., King, M.C., and Avraham, K.B. 2002. From flies' eyes to our ears: Mutations in a human class III myosin cause progressive nonsyndromic hearing loss DFNB30. *Proc. Natl. Acad. Sci.* **99:** 7518–7523.

Wang, E.A., Rosen, V., D'Alessandro, J.S., Bauduy, M., Cordes, P., Harada, T., Israel, D.I., Hewick, R.M., Kerns, K.M., LaPan, P., et al. 1990. Recombinant human bone morphogenetic protein induces bone formation. *Proc. Natl. Acad. Sci.* **87:** 2220–2224.

Wang, E.T., Kodama, G., Baldi, P., and Moyzis, R.K. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc. Natl. Acad. Sci.* **103:** 135–140.

Weir, B.S. 1996. *Genetic data analysis II*. Sinauer Associates, Sunderland, MA.

Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., and Hill, W.G. 2005. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* **15:** 1468–1476.

Williams, T.N., Mwangi, T.W., Wambua, S., Peto, T.E., Weatherall, D.J., Gupta, S., Recker, M., Penman, B.S., Uyoga, S., Macharia, A., et al. 2005. Negative epistasis between the malaria-protective effects of alpha$^+$-thalassemia and the sickle cell trait. *Nat. Genet.* **37:** 1253–1257.

Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., and Nielsen, R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3:** e90. doi: 10.1371/journal.pgen.0030090.

Wise, J. 2008. Consortium hopes to sequence genome of 1000 volunteers. *BMJ* **336:** 237. doi: 10.1136/bmj.39472.676481.DB.

Wooding, S.P., Watkins, W.S., Bamshad, M.J., Dunn, D.M., Weiss, R.B., and Jorde, L.B. 2002. DNA sequence variation in a 3.7-kb noncoding sequence 5′ of the *CYP1A2* gene: Implications for human population history and natural selection. *Am. J. Hum. Genet.* **71:** 528–542.

Wray, G.A. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* **8:** 206–216.

Zhou, G., Zhai, Y., Dong, X., Zhang, X., He, F., Zhou, K., Zhu, Y., Wei, H., Yao, Z., Zhong, S., et al. 2004. Haplotype structure and evidence for positive selection at the human IL13 locus. *Mol. Biol. Evol.* **21:** 29–35.

Zhu, L. and Bustamante, C.D. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170:** 1411–1421.