Population genetics models of common diseases
Anna Di Rienzo

The number and frequency of susceptibility alleles for common diseases are important factors to consider in the efficient design of disease association studies. These quantities are the results of the joint effects of mutation, genetic drift and selection. Hence, population genetics models, informed by empirical knowledge about patterns of disease variation, can be used to make predictions about the allelic architecture of common disease susceptibility and to gain an overall understanding about the evolutionary origins of such diseases. Equilibrium models and empirical studies suggest a role for both rare and common variants. In addition, increasing evidence points to changes in selective pressures on susceptibility genes for common diseases; these findings are likely to form the basis for further modeling studies.

Addresses
Department of Human Genetics, University of Chicago, Chicago, Il 60637, USA

Corresponding author: Di Rienzo, Anna (dirienzo@genetics.uchicago.edu)

Introduction
Alleles that contribute to the genetic susceptibility to human diseases are the focus of intense attention in human genetics. They can be studied by a variety of approaches (e.g. linkage analysis and case-control studies) that aim to identify their role in the etiology of clinically defined phenotypes. Disease alleles are also a specific subset of all genetic variation present in human populations. As such, their fate is influenced jointly by the history and strength of the selective pressures acting in human populations as well as by demographic history, including population size changes and geographic structure. Therefore, the properties of disease alleles can also be investigated by population genetics modeling based on evolutionary principles. An important difference between disease mapping and population genetics studies is that the latter focus on the effects of susceptibility alleles on evolutionary fitness rather than focusing on parameters of clinical severity. Fitness is only loosely correlated with clinical severity.

This is because the relevant fitness effects are those that acted in ancestral human populations, whereas clinical phenotypes are assessed in contemporary populations. In addition, clinical severity does not necessarily imply effects on reproduction, viability and survival, especially for diseases with late age of onset.

Disease mapping strategies rely — implicitly or explicitly — on population genetics quantities, such as allelic heterogeneity, allele frequency spectrum and inter-ethnic differentiation; for example, standard association methods are most powerful for detecting common susceptibility variants with limited heterogeneity. Likewise, the evidence for association is most likely to be replicated in additional populations if the susceptibility variants are the same and similarly common across populations. Because these quantities are functions of the underlying population genetics model, modeling studies might help to optimize disease-mapping strategies to the outcomes of specific evolutionary scenarios.

In general, modeling studies aim at analyzing a complex biological process (e.g. the evolution of disease susceptibility) by reducing it to more basic components (e.g. mutation rate or strength of selection etc). The ultimate goal is to understand the properties of the process and to make quantitative predictions about its outcomes. Even though most models are too simple to be realistic, they can still have practical utility. For example, they can generate hypotheses that can be tested by empirical studies of genetic variation or they can point to new approaches to disease mapping. Conversely, as we learn more about genetic susceptibility variants, modeling studies might use the available empirical information to fill in the gaps of knowledge regarding the genetic architecture of common diseases and how it was shaped by natural selection. Hence, population genetics modeling informed by empirical studies of risk variation — and the reverse — will be necessary if a comprehensive understanding of the genetic bases of common diseases is to be obtained.

Here, I review the results of modeling studies performed to date, in light of the available data on susceptibility variants for common diseases. Several hypotheses have also been formulated about how changes in selective pressures acted on susceptibility genes for common diseases. These hypotheses are providing a framework for further empirical and modeling studies.

Mutation–selection balance: models and data
Different types of population genetics models are likely to apply to Mendelian versus common diseases.
Mendelian diseases, which are usually due to highly penetrant and deleterious alleles that segregate in families, probably fit relatively simple equilibrium models of mutation–selection balance in which disease alleles are removed by purifying selection and continually replenished through the mutation process. Mutation–selection balance predicts high allelic heterogeneity and low disease prevalence, which are both common features of Mendelian diseases. Models of positive selection also apply to a minority of Mendelian disease alleles, most notably sickle cell anemia, G6PD deficiency and the thalassemias, which are caused by the pleiotropic effects of disease alleles on resistance to malaria. For other diseases (e.g. cystic fibrosis), a selective advantage has been invoked, but generating definitive evidence has met with considerable difficulties [1–4].

When it comes to common diseases with complex patterns of inheritance, however, the formulation of evolutionary models is more complicated. Owing to incomplete penetrance, late age of onset, gene-by-environment interactions and polygenic inheritance, the connection between common disease susceptibility and fitness is not clear. Moreover, although most recessive Mendelian diseases are due to loss-of-function alleles, this expectation might not apply to common diseases, thus adding to the uncertainty regarding the rate at which susceptibility variants are generated by the mutation process. Perhaps more importantly, part of the selective pressures acting on susceptibility alleles for common diseases might have changed dramatically during human history, as a result of major cultural and/or environmental transitions, implying that non-equilibrium models might be needed.

The first models of common disease susceptibility based on population genetics principles were formulated within the mutation–selection balance framework. Pritchard [1] modeled the evolution of a disease susceptibility locus in a randomly mating population of constant size, incorporating the effects of mutation, genetic drift and selection [5]. An important aspect of this analysis is that the total frequency of the susceptibility allele class was treated as an outcome of the evolutionary process and was not fixed to a specified value. Under neutrality and for a broad range of mutation rates, he found that the susceptibility class is unlikely to be polymorphic (i.e. have a frequency between 1% and 99%), but rather it tends to be either rare or near fixation. Accordingly, neutral susceptibility alleles are expected to make only a minor contribution to the genetic variance for the disease. However, if susceptibility alleles are deleterious (and if the mutation rate is relatively high), the probability that the susceptibility class is polymorphic is much higher than under neutrality; hence, deleterious alleles are likely to explain a larger fraction of the genetic variance. These findings suggest that common disease susceptibility is unlikely to be due to selectively neutral variation.

In a related study, Reich and Lander [2] also modeled susceptibility alleles as being (slightly) deleterious and showed that, conditional on a specified total frequency of the susceptibility allele class (20%) at a single locus and for a ‘typical’ disease mutation rate, there is a single predominant allele within the susceptibility class [6]. Importantly, this study focused on the role of population growth on the dynamics of disease alleles. More specifically, the authors showed that, although the number of disease alleles increases very quickly if the frequency of the susceptibility class before the onset of growth is low, the allelic heterogeneity for loci with intermediate frequencies also increases, but much more slowly. For plausible times for the onset of population growth, the increase in allelic heterogeneity for loci with intermediate equilibrium frequency is expected to be negligible. Hence, if indeed common diseases are due to susceptibility loci with intermediate equilibrium frequencies, the allelic heterogeneity is predicted to be relatively low.

Owing to the different parameterization, the conclusions of these two studies cannot be directly compared. Nevertheless, they highlight important issues in modeling the evolution of common disease susceptibility, and they indicate new directions for further investigation. One important issue is the choice of parameter values. For example, an obvious difference between the two studies concerns the assumptions about the rate at which new susceptibility alleles are created by the mutation process. Pritchard [1] considered a broad range of mutation rates (i.e. 2.5 × 10⁻⁶ to 1.25 × 10⁻⁴ per gene per generation), whereas Reich and Lander [2] considered a single and low mutation rate (i.e. 3.2 × 10⁻⁶). The rate at which susceptibility alleles are generated and the rate at which they are repaired or compensated for depend on the functional effects of mutations causing common diseases, which in turn depend on a variety of factors, including the disease pathophysiology and the biological function of the gene. For example, the mutation rate for variants leading to loss of function is likely to be higher than the rate for variants resulting in gain of function. Moreover, partial or complete loss of gene function can have similar phenotypic effects and selection coefficients depending on the redundancy of the gene. More importantly, because the current disease allele frequency and heterogeneity are the result of a long-term history, the phenotypic and fitness effects relevant to population genetics modeling are those existing in ancestral human populations and might not be easily extrapolated from the effects of the same mutations in contemporary populations.

The comparison of model predictions under constant and growing population size, as discussed by Reich and Lander [2], is an important reminder that susceptibility loci evolve in populations with complex histories and that demographic histories may differ across major ethnic groups. To date, population genetics models of...
disease susceptibility have relied on a small set of highly simplified demographic models. Thus, though much is still to be learnt about plausible models for human history, future modeling should include some of the complexities emerging from recent inferences about human demography. Several lines of evidence point to a history of rapid recent growth from an equilibrium population for sub-Saharan Africans [7–9], whereas genetic variation data from non-African populations are compatible with bottleneck models [8,10] and, possibly, serial founder-events [11,12]. In addition to population expansion, human variation data are also consistent with significant geographic structure on a continental scale as well as on a finer scale [13]. Future modeling studies incorporating some of these demographic complexities might show that the predicted allelic heterogeneity and frequency spectrum vary to an appreciable extent across human populations — even if selective pressures were the same.

Despite the uncertainties regarding plausible parameters values, the results of the above modeling studies have stimulated much empirical and analytical work. Systematic re-examination and meta-analysis of 25 reported disease associations with common variants revealed a significant excess of replications of the original reports, suggesting that some of these associations are real and that common variants with modest effects — odds ratios between 1.07 and 2.28 — do exist [14]. Evidence supporting a role for heterogeneous and rare susceptibility variants is also rapidly growing as strategies for the detection of these alleles are being developed. Testing for heterogeneous and rare susceptibility variants presents considerable analytical challenges, because established disease association methods are tailored to common susceptibility variants and are unlikely to be powerful for the case of multiple rare variants. A strategy specifically developed to meet this challenge relies on performing variation discovery either in affected and unaffected subjects or in the extremes of a distribution (e.g. the top and bottom 5%) for a disease-related phenotype, followed by testing for an excess of a class of variants in one group compared with the other. The class of variants to be tested is defined a priori on the basis of criteria that are likely to identify the subset of all variation with effects on gene function (e.g. nonsynonymous variants). When applied to candidate genes for variation in plasma lipid levels [15**] and risk to colorectal cancer [16**], this approach detected an excess of nonsynonymous variants with frequency lower than 0.5%. Reliable identification of the class of variants with functional effects is a critical aspect of this approach. For this reason, to date it has been applied to coding regions and nonsynonymous variants, for which information on structural and sequence conservation can be used to calculate the probability of an effect on function [17,18]. In principle, this approach could be extended to non-coding sequences, but the prediction of functional effects for variants affecting regulatory sequence elements is far less reliable than for nonsynonymous variants.

Because standard approaches based on typing common variants are thought not to be powerful for the identification of rare risk-alleles, the development of new statistical methods is an area of active investigation. A promising new method exploits the property that rare variants are on average younger than common variants and, hence, are associated with longer homogeneous haplotypes [19]. Because the method considers all haplotypes of all lengths, it uses the information on the long-range haplotype structure typical of rare variants. Computer simulations showed that this method has reasonable power for variants with frequency as low as 1%, as long as the size of the phenotypic effect is at least moderate. Additional work is necessary to investigate the power of this method for less common variants and as a function of the relevant evolutionary parameter values, such as selection coefficient, mutation rate and demographic parameters.

**Advantageous disease variants and changes in selective pressures**

Several additional non-neutral frameworks have been proposed to explain the evolution of common disease susceptibility. Although these were not developed into formal population genetics models of disease susceptibility, they provide a useful set of hypotheses for further investigation. In one scenario, which can be thought of as a case of antagonistic pleiotropy, alleles that increase risk to common late-onset diseases are advantageous in early life [20]. Because the strength of selection decreases with the age of onset of the phenotype, the allele will be favored and selection will maintain the optimum for younger ages. This scenario, originally proposed to explain the evolution of senescence, might also apply to some of the more common late-onset diseases (e.g. type 2 diabetes and heart disease). The finding that mutations affecting the insulin–IGF (insulin-like growth factor) signaling pathway lead to an increased lifespan in worms, flies and mice supports a link between type 2 diabetes and ageing.

The evolutionary scenarios discussed so far envision selection regimes in which selective pressures do not change substantially over time; hence, they are essentially equilibrium models. However, decades of anthropological research have established that the environment in which we now live is radically different from the one that ancestral human populations adapted to. A number of transitions have resulted in shifts in the selective pressures acting on the biological processes that, in contemporary populations, underlie major classes of common diseases. Crucial junctures include the dispersal of modern humans out of Africa, and the shift away from a hunting and gathering life-style to a subsistence based on agriculture and animal husbandry, in addition to the more recent transitions associated with industrialization.
and globalization. These transitions were associated with major changes in environmental factors — including diet, pathogen exposure, life style and social organization etc — that are likely to have shaped the evolution of susceptibility genes for common diseases.

The impact of these selective shifts on susceptibility genes for many common diseases has been formalized in a series of hypotheses, which are based on epidemiological trends and the disease pathophysiology. Perhaps the most popular is the ‘thrifty genotype’ hypothesis first proposed in 1962 to explain the high prevalence of type 2 diabetes and obesity [21]. This hypothesis posits that since ancestral hunter-gatherer populations underwent seasonal cycles of feast and famine, they would have benefited from being ‘thrifty’ (i.e. having extremely efficient fat and carbohydrate storage). With changes in subsistence strategies, this ‘thriftiness’ became detrimental as food production, processing and storage resulted in more reliable availability across seasons and changes in dietary patterns. More recently, it was proposed that inter-ethnic differences in the prevalence of type 2 diabetes and obesity, in particular the low prevalence in Europeans, might be due to adaptations to cold climates, whereby adaptive variation in a subset of susceptibility genes results in increased thermogenesis and lower risk to type 2 diabetes and obesity [22]. Consistent with the idea that cold stress was an important selective force during human dispersal, there is substantial evidence that body mass index (BMI) and basal metabolic rate (BMR) vary clinically around the globe [23,24]; high BMI and low BMR are correlated with increased risk of type 2 diabetes and obesity. A related scenario, referred to as the ‘sodium retention’ hypothesis, was proposed to explain the inter-ethnic differences in the prevalence of hypertension, particularly the salt-sensitive type [25]. Under this scenario, ancestral populations living in equatorial Africa adapted to hot and humid climates by increasing the rate of sodium and water retention; as populations moved out of Africa into cooler, drier environments, this high level of sodium retention probably became deleterious. Energy metabolism and sodium homeostasis are not the only biological processes associated with currently common diseases that have undergone major selective shifts. Parasite loads and pathogen exposure are likely to have increased dramatically with the switch to the sedentary life-style, higher population densities and greater proximity with domesticated animals resulting from the spread of agriculture and animal husbandry [26]. The reduced exposure to childhood infections, and the overall higher sanitary conditions typical of industrial societies are thought to be responsible for the rapid increase in inflammatory and allergic diseases, such as asthma and inflammatory bowel disease [27].

What is common to the above scenarios is the notion that the observed increase in disease prevalence associated with the adoption of a Western life-style and diet is too rapid to be due to a change in the gene pool. Rather, it is likely that the environmental change simply uncovers the latent genetic susceptibility of human populations to diseases that today account for a disproportionate fraction of the public health burden. Consistent with the idea that common disease susceptibility might be the result of ancient human adaptations to a long-term steady environment, a number of alleles that increase risk to common diseases are ancestral (i.e. they are identical to the allele found at the orthologous position in chimpanzee [28,29]). The new (derived) allele at these variants confers protection against the disease. A possible population genetics model for this class of variants is that the ancestral alleles were maintained by purifying selection in the ancestral environment, whereas the derived alleles were slightly deleterious (see Figure 1). Upon the change in selective pressures, the ancestral alleles increase risk to disease whereas the derived alleles become either neutral or advantageous.

A signature of recent positive selection has been reported for several derived alleles that have protective effects against heart disease and diabetes [30,31,32*,33]; this observation is somewhat surprising because, given the late age of onset of these diseases, protective alleles are expected to have little or no fitness consequences. One possible explanation is that alleles that protect against the primary disease phenotypes also protect against phenotypes with stronger fitness effects, such as those that affect reproductive function more directly. For example, the ancestral Thr235 allele in the Angiotensinogen (AGT) gene was reported to increase risk to both hypertension [34,35] and pre-eclampsia [36]; accordingly, the haplotype carrying the derived (protective) allele exhibits a haplotype structure consistent with the action of recent positive selection [31]. Moreover, as predicted by the sodium retention hypothesis, the worldwide distribution of allele frequency at the AGT Thr235Met variant, as well as other variants influencing risk to hypertension, is consistent with the action of spatially varying selection: the protective alleles decrease gradually in frequency with increasing distance from the equator [32*,37*]. A selective advantage for protective alleles through direct effects on reproduction might also be at work in type 2 diabetes. Over the past 25 years, a decline in diabetes prevalence has been reported in the Pacific island of Nauru, which experienced a dramatic epidemic starting in the 1950s. The lower number of live births observed in Nauruan women with diabetes or impaired glucose tolerance suggests a link between diabetes risk and reduced fertility that is consistent with the observed rapid decline in diabetes prevalence [38].

The above frameworks open the appealing prospect that approaches based on detecting the signature of selection might be added to the arsenal of disease mapping.
strategies. Because population genetics studies and disease association studies are based on independent theoretical frameworks, evidence converging on an overlapping set of candidate susceptibility variants will increase the confidence that the true causative variation has been identified. Consistent with this proposal, a scan for selection signals in 132 genes involved in inflammation, blood clotting, and blood pressure regulation found signatures of selection in several genes previously implicated in common disease risk [39]. A practical advantage of population genetics studies over disease mapping studies is that they can be performed on random samples of unrelated individuals and do not require the costly and time-consuming collection of extensive phenotypic data. For example, to the extent that susceptibility to common diseases might be related to phenotypes for which selection varies according to measurable aspects of geography (e.g. phenotypes related to sodium homeostasis or energy metabolism), mapping susceptibility alleles could be conducted by genotyping individuals from geographically diverse populations and testing for correlations between population allele frequencies and environmental variables. Along the same lines, a recent genome-wide scan for signals of recent positive selection developed a set of SNPs that tag the haplotypes carrying the strongest signatures of selection; it was proposed that this set of tag SNPs be included in whole-genome scans for complex traits [40]. There is, however, uncertainty regarding the assessment of statistical significance for signals of selection [41].

Conclusions
Although the development of formal population genetics models for common diseases is still in its infancy, it is already clear that no single model can explain the evolution of susceptibility alleles even for the same disease or the same susceptibility gene. For example, analyses aimed at detecting rare susceptibility variants for a given phenotype showed that there is strong evidence for such variants in some candidate genes, but not in others [15**,16**]. Moreover, both common and rare variants have been shown to contribute to the susceptibility to the same common disease. This might even apply to variants within the same gene; for example, the \textit{ABCA1} (\textit{ATP-binding cassette A1}) gene, which codes for a cholesterol efflux pump, was shown to harbor multiple rare variants as well as a common variant (Ile883Met) that influence plasma HDL cholesterol levels [15**]. Given the complex history of selective pressures acting on humans, and the composite pathways underlying the pathophysiology of common diseases, it is not surprising that a diverse spectrum of plausible evolutionary models is emerging. Further modeling studies rooted in population genetics theory and data are necessary to achieve a more complete understanding of the evolutionary causes of common diseases. Moreover, by broadening the range
of plausible models, they might lead to mapping approaches that are powerful for the identification of the full gamut of susceptibility variants.

Acknowledgements
I would like to thank M Przeworski, J Cohen, N Freimer, H Hobbs, D Nicolae and J Pritchard for comments on the manuscript. I was supported in part by National Institutes of Health grant DK56670.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

* of special interest
** of outstanding interest


