Published in final edited form as:

Clin Genet. 2011 March ; 79(3): 199–206. doi:10.1111/j.1399-0004.2010.01535.x.

Evolutionary evidence of the effect of rare variants on disease etiology

Ivan P. Gorlov^a, Olga Y. Gorlova^b, Marsha L. Frazier^b, Margaret R. Spitz^b, and Christopher I. Amos^b

^aDepartment of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

^bDepartment of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract

The common disease/common variant hypothesis has been popular for describing the genetic architecture of common human diseases for several years. According to the originally stated hypothesis, one or a few common genetic variants with a relatively large effect size control the risk of common diseases. A growing body of evidence, however, suggests that rare single-nucleotide polymorphisms (SNPs), i.e., those with a minor allele frequency of less than 5%, are also an important component of the genetic architecture of common human diseases. In this study, we analyzed the relevance of rare SNPs to the risk of common disease from an evolutionary perspective and found that rare SNPs are more likely than common SNPs to be functional and tend to have a stronger effect size than do common SNPs. This observation, plus the fact that most of the SNPs in the human genome are rare, suggests that rare SNPs are a crucial element of the genetic architecture of common human diseases. We propose that the next generation of genomic studies should focus on analyzing rare SNPs. Further, targeting patients with a family history of the disease, an extreme phenotype, or early disease onset may facilitate the detection of risk-associated rare SNPs.

Keywords

Single Nucleotide Polymorphisms (SNPs); Genome Wide Association Studies (GWAS); Minor Allele Frequency (MAF); negative selection

Introduction

The common disease/common variant (CD/CV) hypothesis has been a popular means of describing the genetic architecture of common human diseases for a number of years (1). According to the original, strict CD/CV hypothesis, genetic control of susceptibility to common human diseases is relatively simple: one or a few common genetic variants have a relatively large effect size that influences the risk of the disease. The CD/CV hypothesis has stimulated and guided the identification of common genetic variants associated with the risk of common human diseases.

Corresponding author: Ivan P. Gorlov, Department of Genitourinary Medical Oncology, Unit 1374, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030-40009, USA. Tel: 713-563-4951; Fax: 713-754-1652; ipgorlov@mdanderson.org.

Conflict of interest statement:

Authors do not report any conflicts of interest

Genome-wide association studies (GWASs) aimed at identifying the associations between common genetic variants [mainly discovered and cataloged in the HapMap project (2)] and the risk of common human diseases were initially based on the postulates of the CD/CV hypothesis. During the last few years, we have witnessed an explosion in the number of GWASs. According to the catalog of published GWAS, available at <http://www.genome.gov/26525384/> (3) (accessed in March 2010), the results of more than 250 GWAS on more than 150 diseases and phenotypes had been published at the time this review was being written.

Although the amount of data generated by GWASs is impressive, our understanding of the genetic control of susceptibility to common human diseases is far from complete. The major reason for this is that the detected effect sizes are too small to explain the interindividual genetic variation in susceptibility [the typical effect size of the risk allele is 1.2 (4)]. In fact, the GWAS-detected genetic variants usually explain only 2–4% of hereditary variation (5, 6), suggesting that a large part of the overall picture of genetic heritability of common human diseases is missing (5, 7, 8). As a result, we and others have hypothesized that rare polymorphisms (9–13) are an important component in genetic susceptibility to common human diseases.

A currently growing body of experimental evidence has indicated that rare single-nucleotide polymorphisms (SNPs)—those with a minor allele frequency (MAF) of less than 5%—might be crucial in the genetic architecture of common human disease (14–16). For example, a recent paper by Nejentsev et al. (16) reported on four rare variants (all with a frequency of < 3%) within the *IFIH1* gene, which affect the risk of childhood-onset type 1 diabetes. All four variants are likely to result in severe functional disruption: one introduces a premature stop signal, two are found in conserved RNA splicing sites, and the fourth alters an evolutionarily conserved site of the gene's encoded protein. It is interesting that the four variants have a larger individual effect on disease risk than do the common variants previously identified.

Therefore, both theoretical and experimental studies have demonstrated that rare SNPs are important to the genetic architecture of common human diseases. It is also well established that mutation-selection balance, genetic drift, and effective population size all play an important role in shaping the frequency and distribution of genetic polymorphisms in the human genome (17–21). In this review, we focus on the analysis of rare SNPs from an evolutionary perspective.

Most SNPs in the human genome are rare SNPs

We define rare SNPs as those with an MAF of less than 5%. Analysis of the SNPs from the dbSNP database <http://www.ncbi.nlm.nih.gov/snp> suggests that most of the validated SNPs in the human genome are rare. Indeed, binning SNPs according to their MAF demonstrates that more than 50% of the SNPs are rare. Fig. 1 shows the distribution of SNPs, by MAF, from the Encyclopedia of DNA Elements (ENCODE) and the International HapMap Database <http://hapmap.ncbi.nlm.nih.gov/>. As the figure makes quite obvious, the proportion of SNPs remains relatively stable from MAF = 0.5 to MAF = 0.1 and increases sharply when we move from MAF = 0.1 to MAF = 0.025. The ENCODE SNPs show a stronger increase than the HapMap SNPs do. The reason for this is likely that ENCODE SNPs were obtained by sequencing, whereas HapMap SNPs were obtained by genotyping previously identified SNPs; thus, the ENCODE data are unlikely to be biased by selection of common variants.

Functional SNPs and purifying selection

We use a broad definition of functionality: a SNP is functional because it changes the expression level of a gene, splicing, protein structure and stability, mRNA processing, among other effects. Direct estimates of the effects of millions of SNPs in the human genome on these functions is not feasible, but indirect estimates of SNP functionality is possible on the basis of quantitation of selection pressure. Because functional SNPs should somehow disturb a normal function, they are expected to be under the pressure of selection; in most cases, negative, and in some rare cases, positive. A stronger effect of a SNP on gene function may be associated with stronger negative selection against it. We therefore hypothesized that disease-associated polymorphisms are mostly “slightly deleterious SNPs” (sdSNPs) because they are deleterious enough to elevate the disease risk but not deleterious enough for negative selection to completely eliminate them from the population (9). A similar idea was put forward earlier by Pritchard (22). In his recent paper Eyre-Walker (21) analyzed the genetic architecture of a complex trait using a model where mutations that affect trait also affect fitness (effects of mutation on trait and fitness are correlated). The author concluded that more deleterious (and therefore rarer mutations) contribute more to the variance of the trait suggesting that effect of rare genetic variants in the control of complex traits may be substantial.

The assumption that the majority of the disease-causing mutations are likely to be slightly deleterious seems to be quite logical for early-onset diseases. Early onset diseases are likely to be under the pressure of negative selection which would keep frequency of the risk variant low. For the genetic variants associated with late-onset diseases (occurring at postreproductive ages) negative selection is unlikely to play a role and those variants might be essentially neutral and therefore more common. We cannot exclude, however, that variants associated with late-onset diseases are also subject to negative selection. Indeed many genes in the human genome are involved in more than one biological function. Mutations in these polyfunctional genes may disturb ontogenesis and if an individual is lucky fortunate to survive during an early period, the mutation may cause a disease at a later stage of ontogenesis. We acknowledge that the above reasoning is largely speculative: we currently do not have enough information yet to prove or reject the hypothesis. One of the few studies that tried to address this issue was conducted by Yue and Moulton (23). The authors demonstrated that genes associated with monogenic human diseases from the human gene mutation database (<http://www.hgmd.cf.ac.uk/ac/index.php>) show a higher level of evolutionary conservation compared to non-disease genes, suggesting stronger functionality and therefore a higher deleterious effect of disease-causing mutations.

A number of investigators have suggested that purifying selection affects SNP frequencies (24-28). One of the first such studies, conducted by Hellmann and colleagues, sequenced more than 5,000 expressed sequence tags from the chimpanzee and compared them with their human counterparts (29). Those investigators estimated that about 40% of sites in protein-coding regions are deleterious and subject to negative selection. When combining SNP data with sequencing data on orthologous genes, Fay et al. (30) estimated that at least 20% of nonsynonymous SNPs with frequencies of 1–10% are sdSNPs and, as a result, are under the pressure of purifying selection. Yampolski et al. (31) estimated the strength of selection against amino acid replacement by combining data on pathogenic missense mutations, non-synonymous SNPs and human-chimpanzee divergence of orthologous proteins. They estimated that about a half of the substitutions reduce fitness by 1%-0.01%. Yue and Moulton (23) developed a method for predicting deleterious SNPs according to the conservation of the amino acid sequence. They demonstrated that approximately one quarter of nonsynonymous SNPs are sdSNPs and subject to purifying selection. Eyre-Walker et al (2006) (32) found that the vast majority of amino acid changing mutations in humans have mild effects of between 1/1000 and 1/10, a conclusion supported by a later study by

Keightley and Eyre-Walker (33). Boyko et al. greatly expanded upon these findings by including results from the study of many more SNPs including 11,404 coding mutations. The authors found that about 28% of mutations were neutral or nearly neutral while 30-42% were moderately deleterious, and only a small fraction were highly deleterious. Synonymous SNPs can also exert an effect on functionality by affecting splicing and mRNA processing (34, 35).

Selection is not the only factor that affects SNP frequency and the probability of the site's being polymorphic (segregating), however; mutability may also affect both. Different sites in the human genome differ by their mutability. One of the highest mutation rates is observed in CpG dinucleotides (36, 37). To estimate the relative effects of mutability and selection on SNP density, we analyzed them simultaneously and found that approximately 87% of the variation in SNP density was due to differences in mutation rate, and the remaining 13% may be explained by variation in selection intensity (38).

Although SNPs located outside coding regions are, overall, less likely to be functional, some of them certainly are. Assessing natural selection in noncoding regions of the human genome is more difficult than it is in coding regions, partly because of the limited availability of sequencing data and validated SNP data from different species (39), but several studies have provided evidence of natural selection's action on noncoding gene-regulatory elements of the human genome (40, 41). A recent study by Torgerson et al. (42) analyzed genetic polymorphism in evolutionarily conserved non-coding sites (CNC) of the human genome and demonstrated that selection on CNCs has played an important role in evolution. Although the overall probability that any single polymorphism in a non-coding region will be functional is much lower compared to polymorphisms in coding regions, the overall contribution of non-coding regions may be substantial because they constitute about 97% of the human genome (43).

Sethupathy et al. (39), using the derived allele frequency distribution test of neutrality, found evidence of positive selection distributed throughout the human genome. A number of other investigators identified loci with evidence of recent positive selection (44-46). It is still uncertain, however, how widespread is positive selection in the human genome (see (47) for review).

Rare SNPs are predicted to more likely be functional than common SNPs

We analyzed the relationship between the MAF and the proportion of nonsynonymous SNPs predicted to be protein damaging (9). We identified such protein-damaging SNPs from among all validated SNPs in the dbSNP database by applying the two most commonly used bioinformatics algorithms, PolyPhen and SIFT (sorting intolerant from tolerant) (48, 49). Fig. 2A illustrates the proportions of protein-damaging SNPs in the different MAF groups according to the PolyPhen algorithm; we found a statistically significant negative association between the MAF and the proportion of SNPs predicted to be functional ($P < 10^{-6}$). We also found that a logarithmic regression explains 79% of the variation and also fitted the data better than did linear regression, which explained 56% of the observed variation. The SIFT algorithm yielded a similar result for the proportion of nsSNPs predicted to be protein damaging (Fig. 2B). As was the case with the PolyPhen-analyzed data, logarithmic regression explained the observed variation and fitted the data better than did linear regression (79% and 54%, respectively).

MAF is not the only predictor of SNP functionality, though. For example, synonymous SNPs are less likely than nsSNPs to be functional (38, 50, 51). We thus separately compared the proportions of SNPs predicted to be functional for nsSNPs that produce radical missense mutations and for those that produce conservative missense mutations (Fig. 3). A radical

missense mutation replaces the wild-type amino acid with an amino acid that is chemically different, whereas a conservative mutation replaces the wild-type amino acid with a chemically similar one. Therefore, the overall proportion of functional substitutions is expected to be greater among radical missense mutations than among conservative ones. The overall probability that a SNP is functional is almost two times greater for nsSNPs producing radical missense mutations than for nsSNPs producing conservative missense mutations. We also found that the logarithmic regression curves of the proportion of functional SNPs on MAF were very similar for these two types of SNPs, suggesting that the same factors influence MAF–functionality relationships for SNPs having different prior probabilities to be functional.

Rare SNPs tend to have a larger effect size than common SNPs

It is reasonable to assume that nsSNPs differ by their effect on protein structure or function. Amino acid substitutions in functional sites of a protein are expected to have a stronger effect on protein function than SNPs located in other parts of the protein have. The effect of the amino acid substitution is also expected to be stronger when it is located close to the active site, it affects phosphorylation, or it is located in a region important for protein folding or transport to a specific cellular location.

Rare SNPs may have a stronger disturbing effect on protein function than common SNPs and therefore might be under stronger negative selection. We evaluated this hypothesis by correlating the change in accessible surface area—*dprop*—and MAF. Change in accessible surface resulting from substitution is one of the most important predictors of SNP functionality in PolyPhen (49). The effect of *dprop* is strongest when the proportion the accessible area in wild-type protein is low (< 0.02) because in this case, the accessible surface is a limiting factor (49). Fig. 4 shows the relationship between *dprop* for proteins with accessible area in wild-type < 0.02 and MAF. We noted a statistically significant negative correlation between MAF and *dprop* (Spearman's correlation coefficient, 0.85; $N = 17$; $P = 0.00002$). Therefore, there is a negative association between the MAF and the degree of damage to the protein structure: rarer SNPs more strongly impair protein structure than do the common ones, and one can expect that they will be also have a larger effect size on disease risk. In fact, the negative association between effect size and allele frequency is also predicted by modeling the evolution of disease susceptibility (52).

We reviewed the recent catalog of published GWASs (3) (accessed in January 2010) to see if there is evidence for a negative association between MAF and effect size. If the same SNP had been reported as being associated with the same disease in several studies, we used the average odds ratio (OR) and MAF weighted by sample size. The results of our analysis (Fig. 5) suggest a negative correlation between MAF and OR (Pearson correlation coefficient, -0.29 ; $N = 339$, $P < 10^{-6}$). The correlation remained statistically significant after we limited the analysis to SNPs with a MAF > 0.1 ($r = -0.21$; $N = 299$; $P < 10^{-4}$), suggesting that it is unlikely that the correlation is driven only by rare SNPs.

To account for a possible effect of the sample size, we estimated partial correlation coefficients between the OR and MAF after controlling for the sample size and found -0.29 ; $N = 339$; $P < 10^{-6}$. Our use of a multiple linear regression model with the OR as the response and the MAF and total sample size as predictors also identified a significant negative association between MAF and OR, with a coefficient beta of -0.28 and an associated P value of $< 10^{-6}$. Therefore, our analysis suggests that rare SNPs tend to have a larger effect size than do common variants. However because the correction we have made for sample size is based on an assumption of a linear effect we cannot exclude that the observed correlation (at least partially) results from the limited power to detect rarer variants. Of note, the regression curve in Figure 4 looks similar to the one that describes the

relationship between MAF and OR in Figure 5. We believe that this similarity is not accidental, that SNPs that have a strong effect on the protein structure also more strongly (to a greater extent) disrupt protein function, which in turn tends to be associated with a higher OR. Our results are consistent with those published earlier using smaller sample sizes (53). It is likely that the observation that (i) rare SNPs are more likely to be functional and that (ii) rarer SNPs tend to have larger effect sizes may be different manifestations of the same association between allelic frequency and deleterious effect.

Designing future GWASs: GWASs targeting rare SNPs

It is interesting to draw an emerging picture of the architecture of common human diseases. Experimental evidence supports the idea that common human diseases are mostly polygenic (54-56). The complexity of the genetic architecture of common human diseases may partly reflect poor clinical definitions of diseases so that the diseases are in fact a mix of several distinct diseases with different underlying molecular mechanisms (57). The complexity of common human diseases can also result from complexity at the molecular level. Cancer development, for example, is a multistep process which often is viewed as an evolutionary process (58, 59). A number of basic cellular processes, including cell-cycle control, apoptosis, angiogenesis, and cell-microenvironment interactions have been shown to be hallmarks of carcinogenesis (60). Each process is controlled by multiple genes, so it is not surprising that the number of genes shown to be associated with carcinogenesis is more than 4,000 (61). Of course, this number will be much smaller when we consider a specific cancer, but it will still be high (62-64). Different genes may carry differing spectra of allelic frequencies with some genes harboring mostly common and other genes harboring mostly rare variants. A single locus also can be heterogeneous, harboring both common and rare variants (e.g. (65, 66). The relative contribution of rare and common variants is likely to be different for different diseases, e.g. there is convincing evidence that common polygenic variation plays a crucial role in the risk of schizophrenia and bipolar disorder (67). One may conclude, as recent studies suggest, that the originally postulated dichotomy between pure Mendelian diseases and diseases controlled by common genetic variants is no longer valid: rather there is a whole spectrum of disease-associated variants between these extremes (11).

The multigenic nature of common human diseases suggests the existence of multiple disease-associated polymorphisms. It is difficult to estimate how many genetic polymorphisms underlie disease susceptibility. Pawitan and colleagues recently (68) investigated how many SNPs are required to explain a common disease with 40% heritability. They considered models with different underlying genetic architecture of disease susceptibility and found that the minimal required number of susceptibility alleles varied from 80, for the model in which causal variants had frequencies between 0.7% and 5% and ORs between 1.6 and 4.0, to more than 3,000, for the model in which causal variants had frequencies between 1% and 10% and ORs between 1.05 and 1.15. A recent study by Dickson et al. (2010) (69) suggests that at least some GWAS-detected common variants can be in fact synthetic associations of common haplotypes surrounding rare causal alleles.

Concluding remarks

We hypothesize that rare polymorphisms play an important role in the underlying genetic architecture of many common diseases. Because multiple genes may underlie disease susceptibility, one can expect that a considerable fraction of affected individuals will have multiple disease-associated rare genetic polymorphisms. We found that rare alleles are more likely to be functional and confer a larger effect size than common ones are. These observations may provide guidance for designing association studies to detect rare disease-associated SNPs. In particular, selection of cases from families with a strong history of disease may be beneficial for selecting rare alleles and improve our power to detect them

(70). On the other hand, severity of the disease may be an indication of the presence of an allele with a strong disrupting effect on the function of the disease-associated protein that is a characteristic of rare SNPs. Therefore, selection of patients with severe prefer the term extreme phenotypes, such as early onset of disease, rapid disease progression, or poor survival, may also indicate the presence of rare causal polymorphisms with strong deleterious effects on the function of the causal gene. Future association studies to detect rare disease-associated SNPs should probably focus on cases with a strong family history and on more-severe phenotypes.

Acknowledgments

This work was supported by the David H. Koch Center for Applied Research of Genitourinary Cancers; the NIH Prostate SPORE grant, 1 P50 CA140388-01 the NIH Cancer Center Support Grant, 5 P30 CA16672; NIH grant R03 CA133885, NIH grants R01CA127219 R01CA070759, R01CA133996-01 and P01CA34936;.

References

1. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 2001; 17:502–510. [PubMed: 11525833]
2. Consortium TIH. The International HapMap Project. *Nature.* 2003; 426:789–796. [PubMed: 14685227]
3. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. *BMC Med Genet.* 2009; 10:6. [PubMed: 19161620]
4. Khoury MJ, Bertram L, Boffetta P, et al. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *Am J Epidemiol.* 2009; 170:269–279. [PubMed: 19498075]
5. Bogardus C. Missing heritability and GWAS utility. *Obesity (Silver Spring).* 2009; 17:209–210. [PubMed: 19169219]
6. Lango H, Palmer CN, Morris AD, et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes.* 2008; 57:3129–3135. [PubMed: 18591388]
7. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics.* 2009; 182:845–850. [PubMed: 19416939]
8. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461:747–753. [PubMed: 19812666]
9. Gorlov IP, Gorlova OY, Sunyaev SR, et al. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet.* 2008; 82:100–112. [PubMed: 18179889]
10. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008; 40:695–701. [PubMed: 18509313]
11. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* 2008; 322:881–888. [PubMed: 18988837]
12. Need AC, Goldstein DB. Whole genome association studies in complex diseases: where do we stand? *Dialogues Clin Neurosci.* 2010; 12:37–46. [PubMed: 20373665]
13. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 2010; 11:415–425. [PubMed: 20479773]
14. Jakkula E, Leppä V, Sulonen AM, et al. Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. *Am J Hum Genet.* 86285–291
15. Ma D, Salyakina D, Jaworski JM, et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet.* 2009; 73:263–273. [PubMed: 19456320]
16. Nejentsev S, Walker N, Riches D, et al. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science.* 2009; 324:387–389. [PubMed: 19264985]
17. Davidson S, Starkey A, MacKenzie A. Evidence of uneven selective pressure on different subsets of the conserved human genome; implications for the significance of intronic and intergenic DNA. *BMC Genomics.* 2009; 10:614. [PubMed: 20015390]

18. Hawks J, Wang ET, Cochran GM, et al. Recent acceleration of human adaptive evolution. *Proc Natl Acad Sci U S A*. 2007; 104:20753–20758. [PubMed: 18087044]
19. Hughes AL, Packer B, Welch R, et al. High level of functional polymorphism indicates a unique role of natural selection at human immune system loci. *Immunogenetics*. 2005; 57:821–827. [PubMed: 16261383]
20. Liu J, Zhang Y, Lei X, et al. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. *Genome Biol*. 2008; 9:R69. [PubMed: 18397526]
21. Eyre-Walker A. Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc Natl Acad Sci U S A*. 2010; 107(Suppl 1):1752–1756. [PubMed: 20133822]
22. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*. 2002; 11:2417–2423. [PubMed: 12351577]
23. Yue P, Moulton J. Identification and analysis of deleterious human SNPs. *J Mol Biol*. 2006; 356:1263–1274. [PubMed: 16412461]
24. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 1999; 22:231–238. [PubMed: 10391209]
25. Arbiza L, Duchi S, Montaner D, et al. Selective pressures at a codon-level predict deleterious mutations in human disease genes. *J Mol Biol*. 2006; 358:1390–1404. [PubMed: 16584746]
26. Asthana S, Noble WS, Kryukov G, et al. Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci U S A*. 2007; 104:12410–12415. [PubMed: 17640883]
27. Gorlov IP, Kimmel M, Amos CI. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum Mol Genet*. 2006
28. Parnley JL, Chamary JV, Hurst LD. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol*. 2006; 23:301–309. [PubMed: 16221894]
29. Hellmann I, Zollner S, Enard W, et al. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res*. 2003; 13:831–837. [PubMed: 12727903]
30. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. *Genetics*. 2001; 158:1227–1234. [PubMed: 11454770]
31. Yampolsky LY, Kondrashov FA, Kondrashov AS. Distribution of the strength of selection against amino acid replacements in human proteins. *Hum Mol Genet*. 2005; 14:3191–3201. [PubMed: 16174645]
32. Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*. 2006; 173:891–900. [PubMed: 16547091]
33. Keightley PD, Eyre-Walker A. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*. 2007; 177:2251–2261. [PubMed: 18073430]
34. Macaya D, Katsanis SH, Hefferon TW, et al. A synonymous mutation in TCOF1 causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. *Am J Med Genet A*. 2009; 149A: 1624–1627. [PubMed: 19572402]
35. Vidal C, Cachia A, Xuereb-Anastasi A. Effects of a synonymous variant in exon 9 of the CD44 gene on pre-mRNA splicing in a family with osteoporosis. *Bone*. 2009; 45:736–742. [PubMed: 19580891]
36. Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat*. 2003; 21:12–27. [PubMed: 12497628]
37. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics*. 2000; 156:297–304. [PubMed: 10978293]
38. Gorlov IP, Gorlova OY, Amos CI. Relative effects of mutability and selection on single nucleotide polymorphisms in transcribed regions of the human genome. *BMC Genomics*. 2008; 9:292. [PubMed: 18559102]
39. Sethupathy P, Giang H, Plotkin JB, et al. Genome-wide analysis of natural selection on human cis-elements. *PLoS One*. 2008; 3:e3137. [PubMed: 18781197]

40. Hahn MW. Detecting natural selection on cis-regulatory DNA. *Genetica*. 2007; 129:7–18. [PubMed: 16955334]
41. Bush EC, Lahn BT. Selective constraint on noncoding regions of hominid genomes. *PLoS Comput Biol*. 2005; 1:e73. [PubMed: 16362073]
42. Torgerson DG, Boyko AR, Hernandez RD, et al. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet*. 2009; 5:e1000592. [PubMed: 19662163]
43. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002; 420:520–562. [PubMed: 12466850]
44. Cheng F, Chen W, Richards E, et al. SNP@Evolution: a hierarchical database of positive selection on the human genome. *BMC Evol Biol*. 2009; 9:221. [PubMed: 19732458]
45. Kelley JL, Swanson WJ. Positive selection in the human genome: from genome scans to biological significance. *Annu Rev Genomics Hum Genet*. 2008; 9:143–160. [PubMed: 18505377]
46. Pickrell JK, Coop G, Novembre J, et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*. 2009; 19:826–837. [PubMed: 19307593]
47. Nielsen R, Hellmann I, Hubisz M, et al. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 2007; 8:857–868. [PubMed: 17943193]
48. Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform*. 2005; 6:44–56. [PubMed: 15826356]
49. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 2002; 30:3894–3900. [PubMed: 12202775]
50. Carlini DB, Genut JE. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. *J Mol Evol*. 2006; 62:89–98. [PubMed: 16320116]
51. Yue P, Melamud E, Moul J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*. 2006; 7:166. [PubMed: 16551372]
52. Peng B, Amos CI, Kimmel M. Forward-time simulations of human populations with complex diseases. *PLoS Genet*. 2007; 3:e47. [PubMed: 17381243]
53. Iles MM. What Can Genome-Wide Association Studies Tell Us about the Genetics of Common Disease? *PLoS Genet*. 2008; 4:e33. [PubMed: 18454206]
54. Freimer NB, Sabatti C. Human genetics: variants in common diseases. *Nature*. 2007; 445:828–830. [PubMed: 17293879]
55. Mukherjee O, Saleem Q, Purushottam M, et al. Common psychiatric diseases and human genetic variation. *Community Genet*. 2002; 5:171–177. [PubMed: 14960887]
56. Rich SS, Concannon P. Challenges and strategies for investigating the genetic complexity of common human diseases. *Diabetes*. 2002; 51(Suppl 3):S288–294. [PubMed: 12475765]
57. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. *Nat Rev Cancer*. 2007; 7:778–790. [PubMed: 17882278]
58. Hendriksen PJ, Dits NF, Kokame K, et al. Evolution of the androgen receptor pathway during progression of prostate cancer. *Cancer Res*. 2006; 66:5012–5020. [PubMed: 16707422]
59. Klein CA. Gene expression signatures, cancer cell evolution and metastatic progression. *Cell Cycle*. 2004; 3:29–31. [PubMed: 14657660]
60. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000; 100:57–70. [PubMed: 10647931]
61. Kumar GR, Subazini TK, Subha K, et al. CanGeneBase (CGB) - a database on cancer related genes. *Bioinformation*. 2009; 3:422–424. [PubMed: 19759863]
62. Essack M, Radovanovic A, Schaefer U, et al. DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer*. 2009; 9:219. [PubMed: 19580656]
63. Levine AE, Steffen DL. OrCGDB: a database of genes involved in oral cancer. *Nucleic Acids Res*. 2001; 29:300–302. [PubMed: 11125119]
64. Su WH, Chao CC, Yeh SH, et al. OncoDB.HCC: an integrated oncogenomic database of hepatocellular carcinoma revealed aberrant cancer target genes and loci. *Nucleic Acids Res*. 2007; 35:D727–731. [PubMed: 17098932]

65. Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet.* 2008; 40:189–197. [PubMed: 18193044]
66. Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet.* 2008; 40:161–169. [PubMed: 18193043]
67. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009; 460:748–752. [PubMed: 19571811]
68. Pawitan Y, Seng KC, Magnusson PKE. How Many Genetic Variants Remain to Be Discovered? *PLoS ONE.* 2009; 4:e7969. [PubMed: 19956539]
69. Dickson SP, Wang K, Krantz I, et al. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8:e1000294. [PubMed: 20126254]
70. Peng B, Li B, Han Y, et al. Power analysis for case-control association studies of samples with known family histories. *Hum Genet.* 2010

Fig. 1. Distribution of single-nucleotide polymorphisms (SNPs) from the Encyclopedia of DNA Elements (ENCODE) (orange bars) and of all SNPs reported in the International HapMap Database (blue bars) by minor allele frequency (MAF).

Fig. 2.

Proportion of nonsynonymous single-nucleotide polymorphisms (nsSNPs) predicted to be protein damaging, by minor allele frequency (MAF). Each point represents the proportion of functional nsSNPs in a given MAF category. Error bars indicate standard error. (A) The proportion predicted by using the PolyPhen method. The black line is the unadjusted logarithmic regression curve, and the orange line is the curve adjusted for PolyPhen's sensitivity and specificity. (B) The proportion predicted by using the sorting intolerant from tolerant (SIFT) method.

Fig. 3.

The proportions of functional single-nucleotide polymorphisms (SNPs) among radical (blue line) and conservative (green line) amino acid substitutions are shown. Predictive curves (gray) and equations are shown separately for radical and conservative substitutions. Error bars indicate standard error.

Fig. 4. The relationship between the change in accessible surface propensity (*dprop*) and minor allele frequency (MAF). Each dot represents a SNP.

Fig. 5.

The association between odds ratio (OR) and minor allele frequency (MAF). The orange line represents the regression curve. The variation in MAF explains up to 26% of the variation in the OR. On the figure, the maximum OR is limited to 4: several single-nucleotide polymorphisms with ORs higher than 4 are not shown, although they were included in the computation of correlation coefficients.