

Review

Tilting at Quixotic Trait Loci (QTL): An Evolutionary Perspective on Genetic Causation

Kenneth M. Weiss¹

*Department of Anthropology and Integrated Biosciences Genetics Program, Pennsylvania State University,
University Park, Pennsylvania 16802*

ABSTRACT

Recent years have seen great advances in generating and analyzing data to identify the genetic architecture of biological traits. Human disease has understandably received intense research focus, and the genes responsible for most Mendelian diseases have successfully been identified. However, the same advances have shown a consistent if less satisfying pattern, in which complex traits are affected by variation in large numbers of genes, most of which have individually minor or statistically elusive effects, leaving the bulk of genetic etiology unaccounted for. This pattern applies to diverse and unrelated traits, not just disease, in basically all species, and is consistent with evolutionary expectations, raising challenging questions about the best way to approach and understand biological complexity.

THE past 25 years have seen an outpouring of new knowledge in genetics on a scale unprecedented in the history of any science. For important societal reasons the heaviest research investment has been in the genetics of human disease, but there has been comparable progress in understanding normal and abnormal traits in humans and many other species. Numerous approaches, that I generically refer to as “mapping,” have been developed to find statistical association between phenotypes and genotypes. They include searching variation in known candidate genes, genomewide linkage studies in samples of relatives, and genomewide association studies in population samples such as comparing cases and controls (*e.g.*, TERWILLIGER and GORING 2000; MACKAY 2001; RAO and PROVINCE 2001; GEORGES 2007; RAO 2008).

The objective of mapping is reductionistic: to dissect biological traits into enumerable genotypes with estimable effects. Complexity is not a precise concept, but generally means that many genes as well as environmental factors produce a trait, with different combinations of these factors accounting for its variation. Causation is often expressed as probabilistic risk or penetrance, the probability that someone with a given genotype will manifest a particular trait. Whether risk is probabilistic because of the nature of sampling, unmeasured heterogeneity, or because of inherently probabilistic processes is usually not known. Causation takes two faces: to

describe the basis of variation of the trait in populations and to identify the origin of the trait's value in a specific individual. These are philosophically related, but different in practical terms.

Complex phenotypes can usually be viewed in quantitative terms. A trait may be defined quantitatively, like blood pressure, or may be viewed as the qualitative outcome of underlying quantitative risk factors crossing some threshold, as hypertension relative to blood pressure. The quantitative effect may pertain to onset age, severity, or the probability of a stochastic event such as of stroke as a function of blood pressure.

For decades in the history of modern genetics there were few systematic ways to go beyond segregation analysis, a statistical method for testing whether trait variation that clusters in families is consistent with the inherently probabilistic process of Mendelian inheritance. Only in fortuitous exceptions could a specific protein or chromosome anomaly be associated with a disease. Laborious mapping based on recombination among Mendelian traits was possible in experimental plants or animals, but genes and even their number remained largely unidentified until surprisingly recently.

A chromosomal region or gene identified by statistical association can for purposes here be generically referred to as a quantitative trait locus (QTL). The major breakthrough was the advent, just a generation ago, of systematic genomewide mapping techniques that, quite remarkably, could identify QTL without our having to know the biological nature of a trait so long as it could be defined and measured, an advance properly characterized as “a new horizon in human genetics” (BOTSTEIN

¹Address for correspondence: Department of Anthropology, Pennsylvania State University, 409 Carpenter Bldg., University Park, PA 16802.
E-mail: kenweiss@psu.edu

et al. 1980). This is often described as hypothesis-free science, which seems oxymoronic because the scientific method is about testing hypotheses; in fact, mapping does essentially hypothesize genetic causation somewhere in the genome and the objective is to find it. Modern mapping was initially based on RFLP markers, soon supplemented by short tandem repeat (*e.g.*, microsatellite) markers and recently by high-density SNP genotyping.

A long parade of successes quickly followed the availability of genomewide markers. The classic Mendelian pediatric diseases were mapped, led by bellwethers including phenylketonuria (PKU), Duchenne muscular dystrophy, Huntington's disease, retinoblastoma, and cystic fibrosis. The responsible genes were then studied in detail, stuffing *Mendelian Inheritance in Man* beyond printability, forcing it online as OMIM (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim>).

Meanwhile, it has long been recognized that the common chronic diseases that predominate in industrial populations are causally complex. Overall, they do not segregate in families but phenotypes are correlated among relatives, suggesting genetic involvement. Intriguingly, there is usually a subset of families in which cases *do* seem to segregate as if due to a single gene. So, it was natural to ask if such genes could be found by the same mapping methods—and the answer was “Yes.”

The first dramatic successes included the identification of BRCA1, which conferred high risk of breast and ovarian cancer in large multiply affected families (HALL *et al.* 1990; MIKI *et al.* 1994). The subsequent hit parade included genes with major effect on colorectal cancer, Alzheimer's disease, hypercholesterolemia, hemochromatosis, and adult lactase production—just to name a few of the earlier findings. Some pharmacogenetic success has also been achieved (GOLDSTEIN 2008; MALLAL *et al.* 2008).

MUCH ADO ABOUT TOO LITTLE?

These findings affirmed the extension of Mendelian concepts to complex traits, and there has been no looking back. However, poster-child genes do not tell the whole story. Just a few keywords (linkage, mapping, SNP, genomewide association) identified 6866 articles in the PubMed database published in 2007 alone. There has been a comparable burgeoning of online databases; summary, overview, and perspective articles (including this one); new journals; and a sense of urgent competitiveness, with accompanying promotion by journals (with new Errata sections), investigators, companies, and the public media.

A consistent picture has emerged, shown schematically in Figure 1 (ALTMULLER *et al.* 2001; BOWCOCK 2007; KHOURY *et al.* 2007a,b; BODMER and BONILLA 2008; GOLDSTEIN 2008; JANSSENS *et al.* 2008). The poster-child genes explain only a fraction of variation in a trait or a

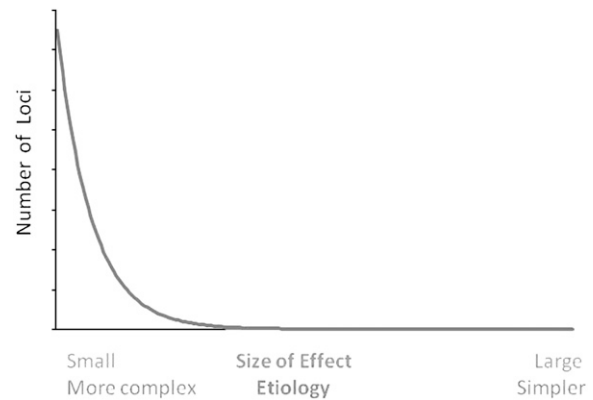


FIGURE 1.—Schematic of the general relationship between number of contributing alleles or loci, their individual effect size, and the consequent degree of complexity of the resulting biological trait. Modified after SING *et al.* (1996).

disease. For almost every tested trait, mapping has identified numerous additional QTL, with lesser or more problematic effects and scattered on many chromosomes (Figure 2).

A tiny sampler from this smorgasbord with selected representative references is given in Table 1. We can add to this feast a chutney of recent reports on various complex diseases (SJOBLOM *et al.* 2006; BENJAMIN *et al.* 2007; WELLCOME TRUST CASE-CONTROL CONSORTIUM 2007; CHEN *et al.* 2008; EMILSSON *et al.* 2008; MANOLIO *et al.* 2008).

One could easily flesh this list out from A to Z, but it is perhaps enough to say that every gene must in some sense be a “disease” gene, because if it has no ill effect when mutated, it will eventually become a pseudogene. More data than you could ever want can easily be found on almost any trait by searching Wikipedia, OMIM, GeneCards (<http://www.genecards.org/>), or the Human Gene Mutation Database (HGMD) (<http://www.hgmd.cf.ac.uk/ac>), among others. Estimates of the number of genes affecting complex traits in populations—or even in simple mouse crosses—range from the tens to hundreds or even thousands (SJOBLOM *et al.* 2006; WANG *et al.* 2006; CHEN *et al.* 2008; REED *et al.* 2008).

This consistently observed pattern applies to traits involving almost every conceivable body part or process, a fact with theoretical, empirical, and epistemological implications. Most information gleaned from mapping consists of numerous small to very small individual effects. The most convincing genes have been found in more than one study, but usually confer relative risks on the order of 1.1–1.3 (Figure 3), which for most complex diseases means small average absolute risk effects (ALTMULLER *et al.* 2001; BODMER and BONILLA 2008; HUNTER *et al.* 2008b). The alleles at these genes usually have both low detectance and low penetrance; that is, the underlying genotype cannot be accurately predicted from the phenotype, and the genotype confers little power to predict the phenotype. Alleles with higher relative risks are more likely to be replicated but are as

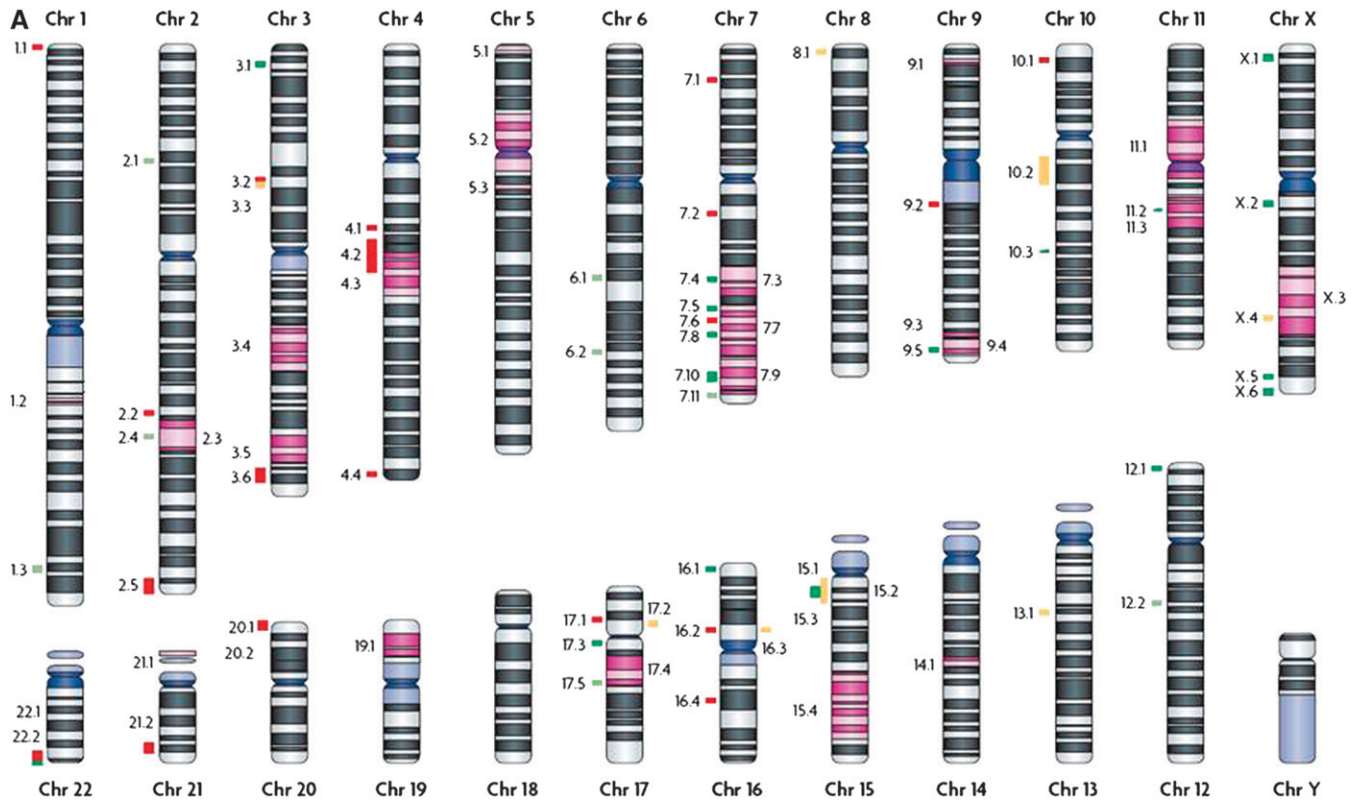


FIGURE 2.—Heuristic examples of genomic mapping hits. The specifics are not important here. (A) A composite of studies of autism (ABRAHAM and GESCHWIND 2008). (B) A few human chromosomes showing obesity-associated hits mapped in mouse or human; the rest of the genome is similarly littered (RANKINEN *et al.* 2006; <http://obesitygene.pbrc.edu>). (The figure is part of Figure 1 of RANKINEN *et al.* 2006 and is adapted by permission from Macmillan Publishers, Ltd: Obesity. Rannekin *et al.* Copyright 2006.)

a rule too rare in the population to have cost effectiveness for variant-targeted therapies.

Statistically, most hits by far have been marginally or only suggestively significant or have not been replicated. Replication is challenging (CHANOCK *et al.* 2007; MCPHERSON *et al.* 2007). This is why Figure 2 is only heuristic: the examples are as of their publication date, subsequent studies always differ, and by no means are all the hits statistically reliable. There is no such thing as “the” true genome map for a complex trait. Even replicated hits are not detected in all studies. In most cases, in both humans and experimental species, the QTL are chromosome regions, often well over 1 Mb long, and may contain tens or hundreds of genes, with no specific gene in the interval statistically implicated (as yet). Replicable QTL usually account for only a fraction of the genetic risk, as estimated by heritability or familial correlation, and a correspondingly smaller fraction of the overall risk that includes environmental effects.

Population isolates such as Finland, Iceland, Sardinia, the Quebecois, or American Hutterites or Amish have been popular sampling frames, on the grounds that due to isolation or founder effect, they will have less variation to sort through. This has helped identify some genes, by reducing background etiological litter, although isolates usually turn out to be less homogeneous

than had been thought—the founding bottlenecks simply were not that severe. Indeed, causation seems comparably complicated even in digenomic crosses between just two strains of inbred mice!

Upon closer inspection it often turns out that different implicated genotypes need not produce exactly the same phenotype. More precise phenotype definition, such as clinical subcategories, can refine mapping and increase the significance or narrow the implicated chromosomal range of some QTL. The price paid is that each subcategory is rarer and the population impact of its QTL less. On the frustrating other hand, sometimes individual traits generate only weak, broad QTL mapping peaks, but multivariate trait analysis of the same data sharpens the peaks. That could indicate that one pleiotropic gene in the chromosomal region affects multiple traits or multiple genes each related to a different trait.

Mapping is essentially made possible by linkage disequilibrium (LD) between marker and chromosomally nearby causal sites, and the fact that markers “tag” causal sites only by indirect, and usually incomplete, statistical association means that once we find the functional site(s) more of the genetic risk may be accounted for (GOLDSTEIN *et al.* 2003). However, the high LD that enables mapping also often makes it impossible to

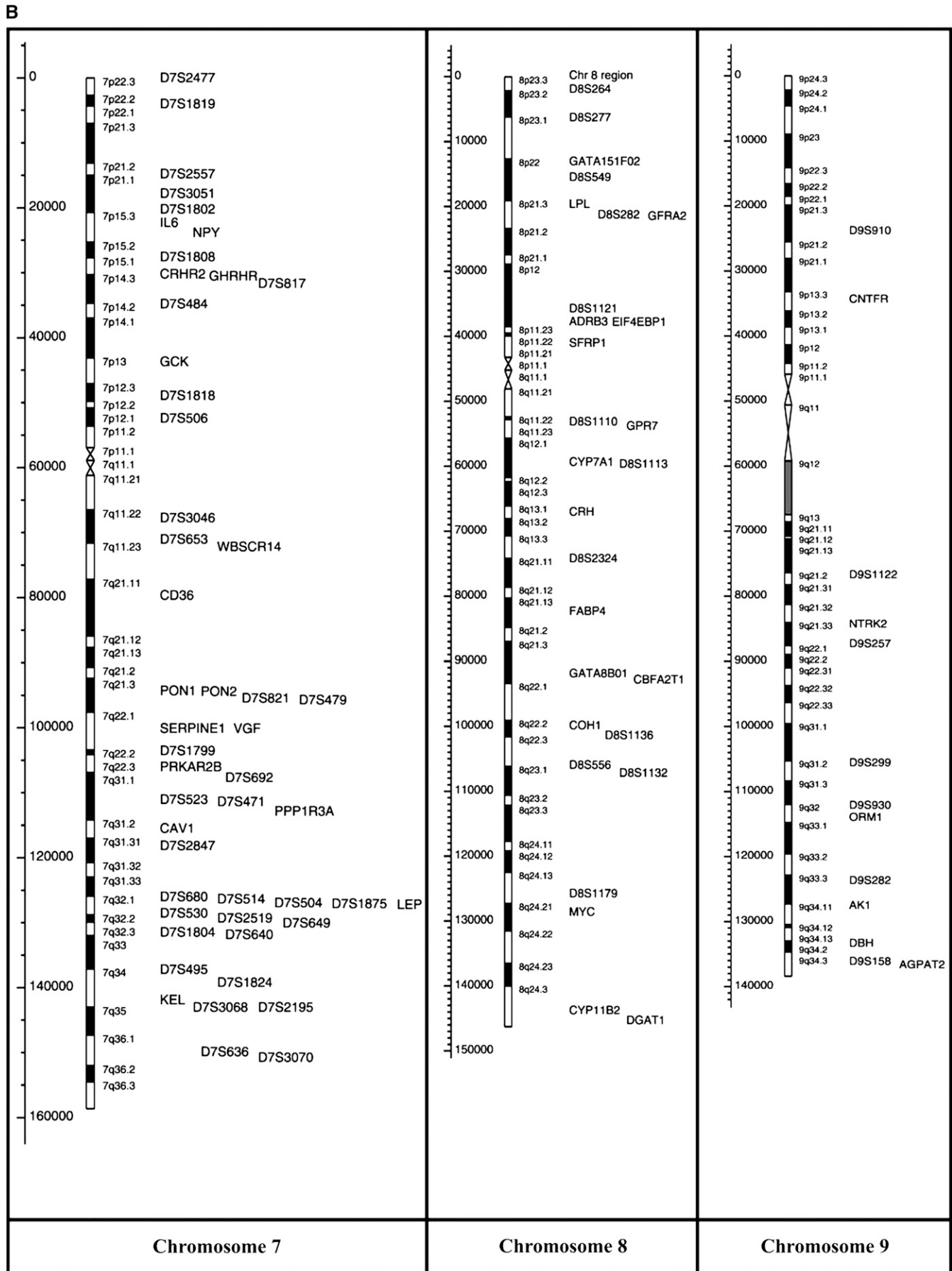


FIGURE 2.—Continued.

TABLE 1

A feast of findings of genes affecting complex traits in one way or another

Asthma	ALTMULLER <i>et al.</i> (2005); OBER and HOFFJAN (2006); DENHAM <i>et al.</i> (2008)
Autism	FOLSTEIN and ROSEN-SHEIDLEY (2001); HAPPE <i>et al.</i> (2006); SZATMARI <i>et al.</i> (2007)
Bone density	HIRSCHHORN and GENNARI (2008)
Coronary heart disease	McPHERSON <i>et al.</i> (2007)
Diabetes	TODD <i>et al.</i> (2007); ZEGGINI <i>et al.</i> (2008)
Epilepsy	MEISLER <i>et al.</i> (2001); CRINO (2007); OTTMAN (2005)
Hypertension	CAULFIELD <i>et al.</i> (2003); KATO <i>et al.</i> (2008)
Infectious diseases	HILL (2006)
Inflammatory bowel disease	XAVIER and PODOLSKY (2007); DUBOIS and VAN HEEL (2008)
Lipoproteins	KATHIRESAN <i>et al.</i> (2008); LUSIS and PAJUKANTA (2008)
Macular degeneration	DAIGER (2005); HUGHES <i>et al.</i> (2006); LI <i>et al.</i> (2006); LOTERY and TRUMP (2007); ALLIKMETS and DEAN (2008)
Obesity	obesitygene.pbrc.edu and RANKINEN <i>et al.</i> (2006); WANG <i>et al.</i> (2006)
Prostate cancer	HAIMAN <i>et al.</i> (2007); ANONYMOUS (2008)
Psoriasis	LIU <i>et al.</i> (2008)
Psychiatric diseases	KELLER and MILLER (2006); ALLEN <i>et al.</i> (2008); BURMEISTER <i>et al.</i> (2008)
Stature	VISSCHER (2008)

dissociate many linked variable sites to identify which are causally relevant, meaning also that the tagged association is closer to capturing the total association at that region. Independent new data, other study designs, and especially the discovery of different alleles at the same locus also associated with the trait strongly reinforce the candidacy of the QTL. When followed up in detail, the gene often turns out to make functional sense relative to the trait. But that the statistically indirect nature of marker-based mapping does not typically account for relatively weak estimated effects or the unmapped fraction of heritability can be seen by follow-up studies of known “causal” SNPs, as shown, for example, by meta-analysis.

Meta-analysis that jointly analyzes multiple or pooled studies often achieves sample sizes adequate to support the candidacy of replicated SNPs and/or to see how geographically widespread similar associations are, although the relative risks typically converge toward a small overall effect, and many or even most candidates fail to survive the test (*e.g.*, McPHERSON *et al.* 2007;

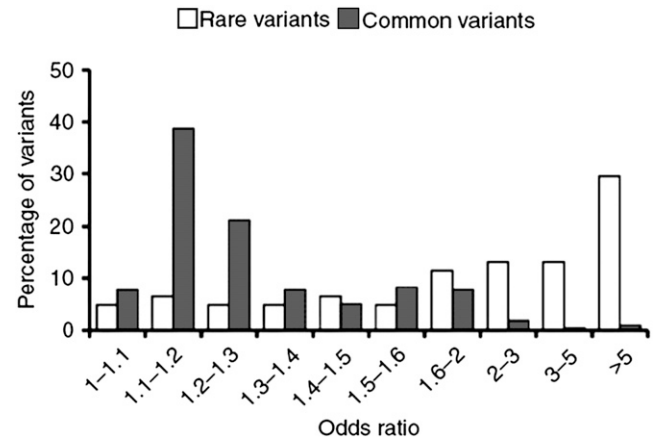


FIGURE 3.—Typical odds ratios for rare (0.1–3% minor allele frequency, MAF) and common (>5% MAF) variants in genomewide association studies (from BODMER and BONILLA 2008) (The figure is Figure 2 from BODMER and BONILLA 2008 and is reprinted by permission from Macmillan Publishers, Ltd: Nature Genetics. Bodmer and Bonilla. Copyright 2008.)

ALLEN *et al.* 2008). Meta-analysis presents a number of analytic challenges (*e.g.*, IOANNIDIS *et al.* 2004; IOANNIDIS 2007; KAVVOURA and IOANNIDIS 2008), not the least of which is upward biases in risk estimates, especially the “winner’s curse” of first reports (GÖRING *et al.* 2001; BEGG 2002; ZOLLNER and PRITCHARD 2007), which often are based on studies intentionally biased to optimize detection (TERWILLIGER and WEISS 2003). There is another subtle bias, in that meta-analysis is a candidate-gene design, testing the effects of a known allele rather than searching for unknown effects, but it usually includes the often-biased first report, but not mapping studies that did not find a “hit” in the candidate’s chromosomal region (this would admittedly be hard to do for various reasons of comparability between mapping scores and candidate gene tests, made worse perhaps by reluctance of journals to publish negative results). Some meta-analyses find statistical evidence of risk heterogeneity among studies, a warning that QTL that are *not* consistently replicated may not all be false positives. By the same token, there must also be many false negatives.

Even after a gene has been identified there is still more mapping to be done. And it turns out that there is no truly free lunch, as “simple” traits are not so simple after all (SCRIVER and WATERS 1999). Many different alleles are found in the normal population (see dbSNP at <http://www.ncbi.nlm.nih.gov/projects/SNP/> or HapMap at <http://www.hapmap.org/>), and tens to hundreds of alleles are found among patients: >560 for PAH, 1400 for CFTR, and 1300 for Dystrophin (HGMD), usually served one to a haplotype. A few miner’s canaries that enabled the gene to be mapped are of high penetrance and relatively common (among patients, though usually rare in the general population), but subsequent resequencing of the gene in

patients reveals a long tail of increasingly rare alleles, most of which have been observed only once. Even the relatively common alleles are often restricted to a single geographic region, and the allelic spectrum may have no overlap across continental regions. Therefore, unless the mutation clearly knocks out function in the gene, for example by causing a frameshift, singleton or near-singleton alleles can legitimately be considered disease related only on the *assumption* that the gene is responsible for the disease in the person in which it is found.

This variation shows that Mendelian notions such as recessiveness have clung on beyond their sell-by date, because many or even most cases of some classical “recessive” diseases are actually heterozygotes at the sequence level, which close inspection shows have quantitative rather than dichotomous genotype–phenotype associations. Indeed, Mendel himself probably could not have succeeded had he had to work by mapping in samples of wild peas rather than carefully choosing traits segregating dichotomously in inbred lines that he could study experimentally. However, we seem to hunger to make things categorical, and these findings have led to new clinical entities like mild PKU or nonphenylalanine hyperphenylalaninemia, to supplement classical PKU (<http://www.PAHdb.mcgill.ca>).

Some traits appear to be a mix of genetic complexity and simplicity: many different genes have been implicated, but most families seem to be segregating a highly penetrant allele at only one of them. Examples include nonsyndromic deafness (125 genes; <http://webh01.ua.ac.be/hhh>) and retinitis pigmentosa (>100 genes; HARTONG *et al.* 2006) or epilepsies (MEISLER *et al.* 2001; CRINO 2007). This is *multiple unilocus* etiology, quite different from classical polygenic traits in which variants at many different genes are thought to contribute to the trait in each affected individual. Yet these genes together account for only a fraction of cases, and as with all other genes many different mutations are found among cases, with a spectrum of frequency and phenotypic effect within and among families.

Along with this cornucopia of data, the menu of the Mapper’s Café has also greatly expanded: it is becoming clear that our search for candidates must go beyond the few percent of the genome that comprises the exons of our paltry 20,000 genes. Genes often have multiple context-specific splicing variants (HILLER and PLATZER 2008), their expression controlled by multiple alternative or sometimes countervailing regulatory elements (DAVULURI *et al.* 2008), phenotypes affected by multiple *cis*-alleles (perhaps on the same haplotype) that may have compensating phenotypic effects relative to each other (KONDRASHOV *et al.* 2002; KWIATKOWSKI 2005; HUGHES *et al.* 2006; LI *et al.* 2006), and *trans*-haplotypic effects as well as other kinds of epistasis that are important (*e.g.*, MOORE and WILLIAMS 2005; TSAI *et al.* 2007). For the bulk of common complex traits such as

chronic diseases that allow successful embryogenesis and develop gradually or strike decades later in life, gene regulation rather than coding differences may be the more important source of phenogenetic variation (MANOLIO *et al.* 2008), yet regulatory regions are currently largely unknown in number and location and can even be *trans* to the affected gene (*e.g.*, CHEN *et al.* 2008).

It also now appears that a large fraction of genomes are transcribed into noncoding RNA, whose conserved pattern, along with subtle genome structural or copy number variation (CNV) (projects.tcag.ca/variations and eichlerlab.gs.washington.edu/database.html), suggests phenotypic relevance (BIRNEY *et al.* 2007; PHEASANT and MATTICK 2007; STRANGER *et al.* 2007; AMARAL *et al.* 2008; HURLES *et al.* 2008; WEISS *et al.* 2008). Best understood at present are miRNA “genes” that regulate expression via effects on chromosomal packaging or mRNA translation via RNA interference, which can have disease consequences (VAN ROOIJ *et al.* 2008). Epigenetic chromosome modification such as by sequence-specific nucleotide methylation or histone acetylation that affects gene expression, along with CNV, can be polymorphic even between MZ twins (WONG *et al.* 2005; BRENA *et al.* 2006; PETRONIS 2006; BRUDER *et al.* 2008). Parent-specific or monoallelic expression turns out to be widespread in autosomal as well as X-linked genes and which allele is activated is at least partly stochastic (KRUEGER and MORISON 2008). This choice is made cell by cell early in embryogenesis and is mitotically remembered in cell lineages thereafter, creating somatic mosaics within and thus phenotypic variation among nominally identical heterozygotes at the locus. Even the humble mitochondria have surprising proliferating effects that can be somatic as well as inherited (WALLACE 2008). Somatic changes, including mutations, contribute phenogenetic variation that is not transmitted across generations and hence is cryptic to mapping strategies (WEISS 2005).

The ability to test a given cell type for expression of a high fraction of genes in the genome has led to a new type of mapping search, for expression QTL (eQTL). The idea is to find sequence variation whose effect is to alter the timing or expression of other genes (MORLEY *et al.* 2004; CHEUNG *et al.* 2005; STRANGER *et al.* 2007). Cluster analysis identifies sets or networks of genes whose correlated expression is altered by variation in a mapped region elsewhere in the genome (WANG *et al.* 2006; CHEN *et al.* 2008) and animal studies can be tested for consistency with human disease etiology (EMILSSON *et al.* 2008). Similarly, correlated expression-level changes involving large numbers of genes characterize cell-specific expression profiles, or before-and-after effects of experimental treatment, or, in the case of diseases like cancer, effects related to prognosis (NEVINS and POTTI 2007). Because most genes are pleiotropic, expressed in and affecting multiple traits (BUCHANAN *et al.* 2009), this approach, while highly promising,

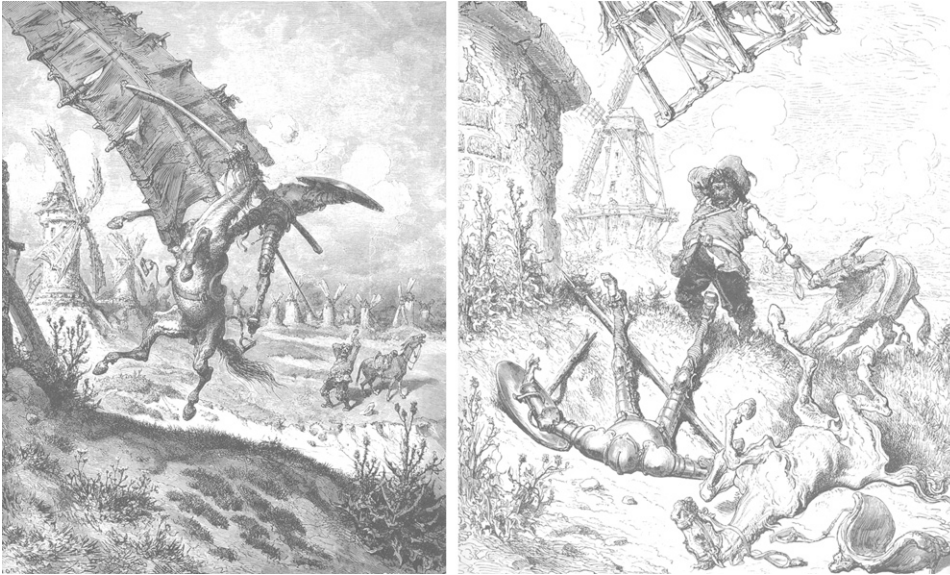


FIGURE 4.—Don Quixote, undaunted in assaulting elusive evil. Drawings are by Gustav Doré (1869), inspired by *Don Quixote* (Miguel Cervantes, 1605).

requires access to the right types of cells at the right time. This is not a trivial issue especially when the trait is complex or developmental: What cells do you look at if you want to understand craniofacial development, diabetes, or schizophrenia, and how do you get your hands on the cells?

The sobering fact about these many genomic functions is that they each add to, and never subtract from, the sequence elements that may affect disease. How much each of these new functional genomic features actually contributes to phenotypic variation is anybody's guess at present. But together they comprise a large DNA target for mutation and variation. The point is not that noncoding variation is unmappable, but rather that to find functional candidates, we need to search whole megabase-long QTL regions rather than just the protein-coding genes they contain. At ~ 1000 SNPs/Mb of candidate region (even just in a pairwise comparison), this raises many problems for statistical and causal inference.

The patterns I have described tell an empirical tale without any over-arching theoretical framework, so one is free to believe that to resolve the incompleteness all we need are larger-scale, longer-term studies. This argument rests on a subtle *assumption* that the signal-to-noise (S/N) ratio will improve with sample size, number of markers or complete genome sequence, phenotype details, and number of environmental variables. Yet it is not obvious that S/N will behave as expected. Will genetic or other sources of heterogeneity rise as fast as sample size? Scaling up mapping studies will not detect alleles too rare to generate significance in the sample, and common or large-effect alleles can already be replicated and, once identified, their effects estimated directly from small samples. So the huge longitudinal biobanks being launched can be expected mostly to refine estimates of modestly common alleles

with modest effects and discover a scattering of minor effects (Figure 3).

The demand for increased sequence and sample size may itself be evidence that we are approaching diminishing returns. Scaled-up studies move us ever more toward tilting at *quixotic* trait loci, chasing the effects most difficult to replicate, hardest to discriminate between true and false positives, or from which to make accurate risk estimates. Like the man of La Mancha (Figure 4), we have perhaps misperceived the sails of QTL windmills: they are not standing giants waiting to be lanced, but elusively whirling, sometimes ephemeral targets. Despite the unquestioned exceptions, quixotic trait loci are the rule, and we have known to expect them for nearly a century.

Indeed, it is more than a little remarkable that the same phenogenetic pattern pertains to traits that are genomically, functionally, histologically, and adaptively unrelated, in plants, animals, and microbes. Surely there must be meaning there! It is a central lesson, for which an evolutionary perspective provides a kind of theoretical support that has otherwise been missing.

EVOLUTIONARY UNDERPINNINGS

The broad outlines of what we see today were predicted by early geneticists with little direct understanding of the nature of genes, based almost exclusively on the phenotypic similarity among related individuals and Mendelian principles backed by basic biology and experimental breeding and population genetics (WRIGHT 1931, 1978; WADDINGTON 1957; PROVINE 1971, 1986). A benchmark article was R. A. Fisher's 1918 turgid demonstration that the combined effects of many "Mendelian" (discretely segregating) loci could account for both quantitative inheritance (of continuously varying traits) and the observed correlation

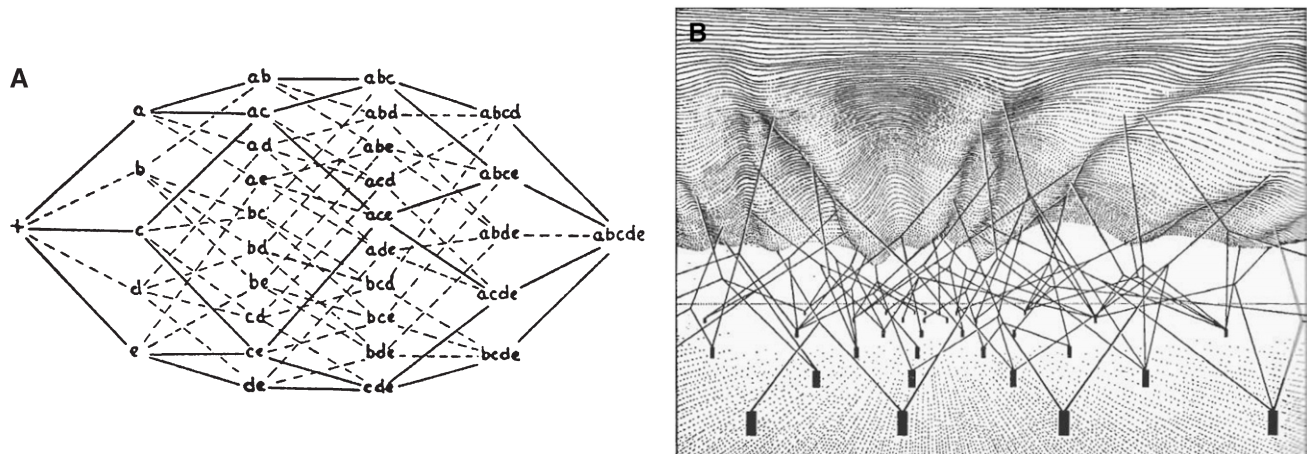


FIGURE 5.—Complex trait architecture, then and now: early conceptual diagrams of interactive complexity of genes. (A) Wright's schematic indicator of combinations of genotypes that can affect phenotypes affected by just five pairs of alleles with potential differences among the interactions depending on which alleles are present (WRIGHT 1931). (B) Waddington's metaphor of multiple genes tugging on various parts of a developmental landscape (WADDINGTON 1957). (C) Elements of a genetic network of ~200 diabetes-related genes correlating gene expression and human variation; red and green indicate, respectively, genes with positively or negatively correlated expression (courtesy of Joanne Curran, Southwest Foundation for Biomedical Research, unpublished research). (D) Hypothetical diabetes-related metabolic syndrome network suggested by mapping and experimental data in a cross between B6 and C3H laboratory mice, in which changes in one mapped location (left) affect a whole network of genes (center), ultimately modifying the final disease-related phenotype (right) (CHEN *et al.* 2008). (D is Figure 4C from CHEN *et al.* 2008 and is reprinted by permission from Macmillan Publishers, Ltd: Nature. Chen *et al.* Copyright 2008.)

among relatives (FISHER 1918). A central fact about this model is *genotypic equivalence*, that different genotypes—different combinations of alleles—can confer effectively the same phenotype. Each gene is a “will-o-the-wisp” (WRIGHT 1934), but together they constitute the classical *polygenes* that contribute effects to, rather than cause, traits; their individual effects are individually small—in the theoretical limit, infinitely many identical loci each contribute infinitesimal effects. This is why complex traits can aggregate but not segregate in families—and can at the same time be “genetic” and yet not Mendelian.

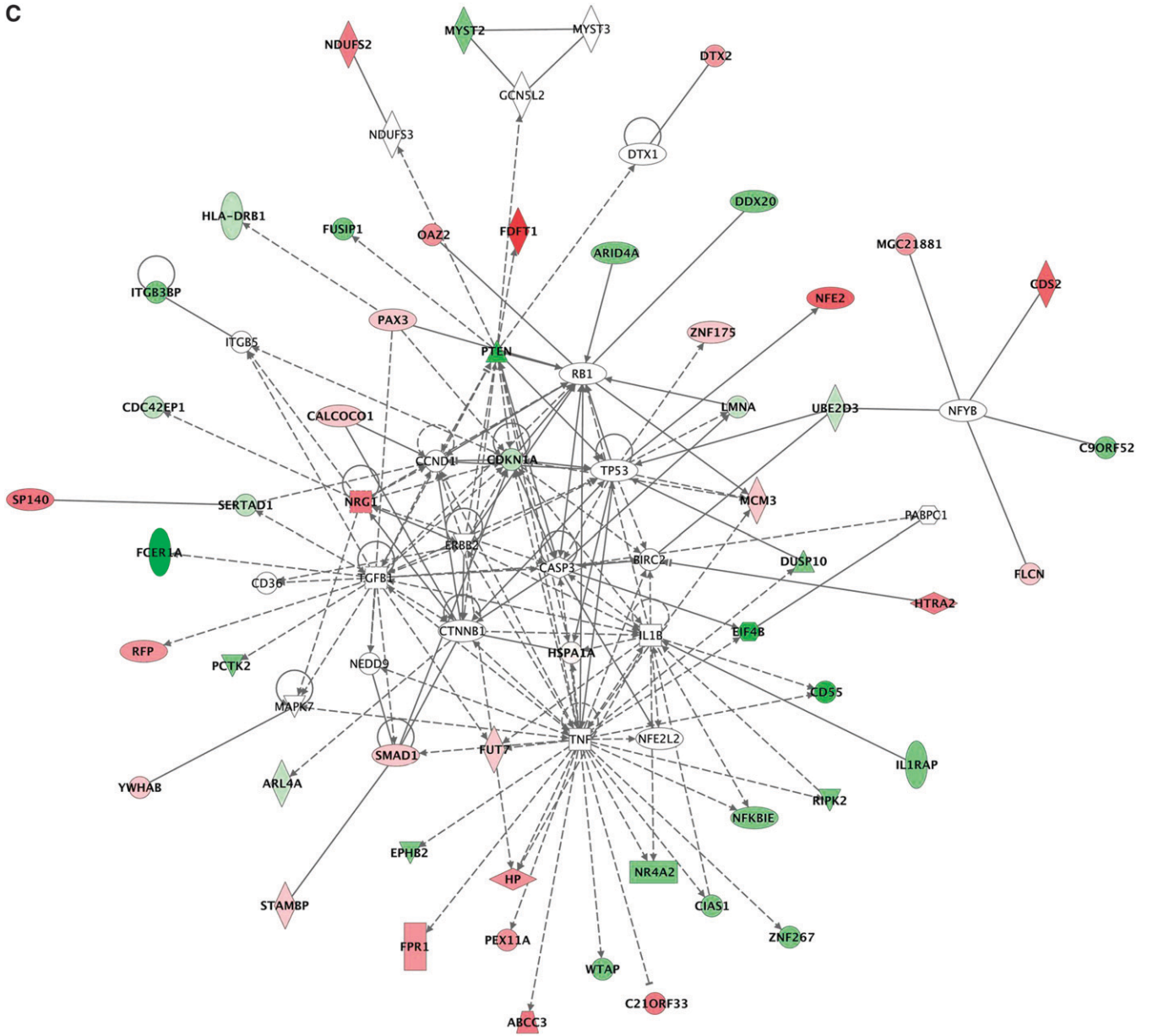
The recent response to the quixotic trait locus landscape is the catch phrase “systems biology.” But long before their molecular nature was known, leading biologists said in strikingly modern ways that complex traits are the result of networks of multiple contributing, interacting genes (MORGAN 1917; WADDINGTON 1957; WRIGHT 1931, 1978) (Figure 5). Volume 1 of Wright's great retrospective is especially instructive of the early recognition of the fundamental facts (WRIGHT 1968). What we have been discovering in disease genetics was also predictable before the outpouring of modern data (WEISS 1993; WEISS and TERWILLIGER 2000), which are now providing at least preliminary documentation (Figure 5, C and D).

Once the modular nature of the genome and its evolution became known, we could understand where polygenes and networks—the plethora of contributing functional elements described above—come from. Biological traits are built up over eons by episodic mutation and duplication events. Their genetic architecture is not

as internally homogenized as textbook polygenic models suggest. Basic functions such as core molecules in signaling or metabolic networks, rate-limiting genes in physiologic systems, or core protein domains are phylogenetically deep and widespread. Subsequent components arise that modify but must be compatible with these earlier ones. Interactions build up in this way so that, viewed retrospectively, we name the result a “network,” often after the “hub” genes (*e.g.*, “Hedgehog signaling”), with more numerous less centrally connected downstream “spoke” genes. The pervasiveness of such systems shows that, while the predominant image of life is the Darwinian one of winner-take-all competition, the predominant nature of life itself is of cooperative interactions among multiple components: signals and receptors, proteins and each other or with DNA, and so on (WEISS and BUCHANAN 2008a,b). So it is no surprise that multiple genes inherently affect interestingly complex traits.

There are always exceptions, because life is a contingent, highly stochastic evolutionary phenomenon, but generally there is reason to think that the coding regions of early developmental or hub genes evolve relatively slowly (KIM *et al.* 2007; H. A. LAWSON, unpublished results). Partly this is because such genes are typically pleiotropic, rather than being evolved for some specific function (BUCHANAN *et al.* 2009). Mutations with large effect (almost always negative) may be quickly removed by the cold hand of selection, having little chance to become geographically widespread. However, most functional genetic elements do tolerate variation

C



D Complex of closely linked QTL

Gene network

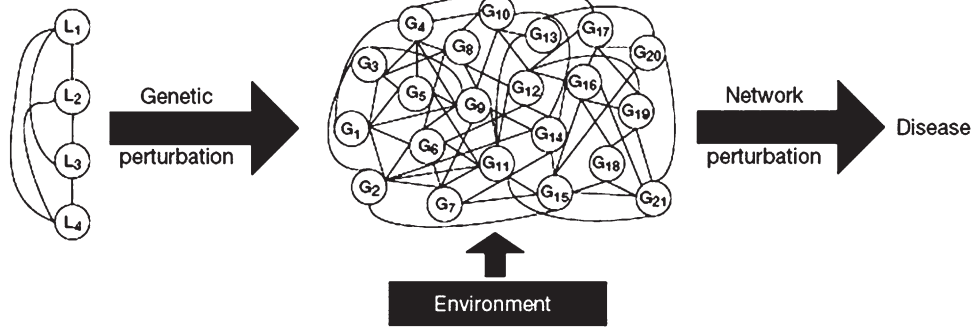


FIGURE 5.—Continued.

and most new mutations have little effect either on phenotype or on Darwinian fitness (OHTA 2002; EYRE-WALKER and KEIGHTLEY 2007; KEIGHTLEY and EYRE-WALKER 2007; LYNCH 2007b; BODMER and BONILLA 2008; BOYKO *et al.* 2008). These alleles are evolutionarily neutral or nearly neutral, and their frequencies change predominantly by genetic drift. Drift generates distributions of allele frequencies that are skewed toward rare, local alleles with only a few common, widely distributed ones at a given locus.

Surprisingly, natural selection, even balancing selection, leaves qualitatively similar results. The numerous genetic elements that affect complex traits present a large DNA target for mutation, and many alleles arise that have a comparable adaptive effect, and their frequencies evolve by drift relative to each other (HARTL and CAMPBELL 1982). For every common allele maintained even by heterosis, such as sickle cell hemoglobin in malarial environments, there are many rare, geographically local alleles (at the same or other genes) also maintained by the same selective pressure, and even the major alleles are often found only within a geographic region (KWIATKOWSKI 2005).

Evolution generally molds complex traits to have modal distributions, in which most individuals are within the trait's historically accepted fitness range. This generates a subtle confounding of frequency and effect size that is reflected implicitly in Figure 1 (SING *et al.* 1996). Like fitness, allelic effects are essentially measured relative to a population mean, so that a "large" effect must almost by definition be far from the mean and hence rare: were it too common, it would *be* the mean, with zero effect.

These points together comprise one of the underappreciated implications of *gradualism*, a cornerstone of evolutionary theory. Evolution, working through phenotypes, but only indirectly on underlying genotypes (WEISS and BUCHANAN 2003, 2004), has led to the complexity that buffers organisms against devastating mutation. Interestingly, while major mutations and hub genes with easily studied effects understandably receive the preponderance of experimental investigation, their importance and constrained evolution may mean that adaptive evolution usually occurs through small-effect mutation in the *less* vital, but more numerous and more nearly neutral downstream or peripheral genes with causation less driven by strong selective adaptation and more internally heterogeneous (*e.g.*, LYNCH 2007a,b; WEISS and BUCHANAN 2008a,b).

Why does this inform us what to expect in genetic association studies? Geographically dispersed or common risk alleles are older and more likely to be repeatedly detected (CHAKRAVARTI 1999). But, their widespread dispersion indicates that those alleles are benign (at least in regard to fitness history), so if they are associated with disease the causal finger actually points to recent *environmental* change rather than primarily to

genetic etiology. Rapid environmental change and secular changes in incidence of complex traits are characteristic of our age, and most common chronic disorders in the developed world are possible because of reduced risks of infectious or early onset disease, plus widespread exposure to sedentary lifestyles and old age (NEEL 1962; TROWELL and BURKITT 1981; POLLARD 2008). Yet ironically, these largely environmentally induced diseases have become the most intensely studied by geneticists.

A few years ago the idea was promulgated that common variants were commonly going to be found to make major contribution to our common diseases and hence to public health: the common variants/common diseases ("CV/CD") notion (*e.g.*, REICH and LANDER 2001; GOLDSTEIN *et al.* 2003). If alleles with large effect are rare, yet a disease is common, then it might seem that the few contributing small-effect alleles must have high frequency (CHAKRAVARTI 1999)—otherwise there would be too few risk genotypes for the disease to be common. Common variants are economically attractive pharmaceutical targets because large numbers of people would be affected by them. Of course, "common" is a moveable target, and CV/CD had an element of hopeful thinking from the beginning (WEISS and CLARK 2002). It has not generally been borne out by experience (Figure 3, BODMER and BONILLA 2008), even though the exceptions properly receive special attention (*e.g.*, GOLDSTEIN 2008; MALLAL *et al.* 2008).

Nonetheless, association mapping can help identify unsuspected pathways even when segregating variants *per se* do not have major public health impact. CV/CD probably has the most promise in specific molecular-recognition interactions, such as autoimmune or infectious disease, chemical exposures, or pharmaceutical agents (GOLDSTEIN *et al.* 2003), but again these are environmental interactions. Indeed, a potential future surprise could be that more chronic diseases have an immune or infectious component than has been suspected.

We still have to ask how a trait with substantial heritability that is produced by alleles at several genes could be common if those alleles are rare. One answer may lie in the size of the *aggregate* of alleles that affect complex trait values, mostly rare as the empirical data and theory (*e.g.*, PRITCHARD *et al.* 2000) suggest. The biomedical ascertainment system of registries and specialty clinics collects cases from populations numbering hundreds of millions, thus ascertaining very rare alleles. They are geographically local and difficult to replicate, but they may provide a sufficient pool of risk polygenotypes to make the trait common. These are the quixotic trait loci that populate current mapping data.

EPISTEMOLOGICAL AND BIOETHICAL ISSUES

Genetic association studies reconstruct the history of today's phenotypes. This must perforce be done retro-

spectively, in terms of the sampled persons' exposures to their inherited genotypes at conception and to subsequent environmental factors. Yet risk estimation is a *prospective* enterprise, to predict future genotype-specific phenotypes, and here the devil is in the nongenetic component. The lesson should be chilling. Among the most undisputed disease risk alleles are those at BRCA1 and BRCA2 associated with breast and ovarian cancer. Yet, their estimated risk varies by roughly twofold depending on many factors including birth cohort (FODOR *et al.* 1998; KING *et al.* 2003; CHEN *et al.* 2006) and lifestyle differences.

For alleles as dangerous as those in BRCA1/2, causation seems to be real and screening can lead to prevention without worrying about the fine points of the risk estimates. But risks associated with lesser genotypes will more likely be enhanced, reduced, or even disappear under future environmental conditions, and new ones will appear. Yet we have no way to know how those exposures will change in the future, especially if we must ascertain them in exceedingly subtle ways from conception onward (DOBLHAMMER and VAUPEL 2001; GLUCKMAN *et al.* 2008), although we can be confident *that* they will change. Things always look simpler after the space of unseen possibilities is narrowed by the quixotic turns of history down to one—what actually happened. This is a major conceptual weakness of prospective biobank studies because, once done, their estimated risks will again be backward looking.

A profound epistemological challenge is that environmental risks are, if anything, even more problematic to identify and estimate, much less to predict, to the point that environmental epidemiologists have been rushing to the genetics bandwagon expecting to be bailed out by causal factors that are more tangible, not realizing that we face very similar problems (BUCHANAN *et al.* 2006).

The instability or unpredictability of genotype-specific risks raises obvious ethical issues. Reviews of association studies reflect understandable enthusiasm; caveats are usually offered, but often seem unconvincing or stated largely in passing (*e.g.*, BLANGERO 2004; DAIGER 2005; CARDON 2006; EVANS and CARDON 2006; JAQUISH 2007; KHOURY *et al.* 2007a,b; CHEN *et al.* 2008; EMILSSON *et al.* 2008; HUNTER *et al.* 2008b; JANSSENS *et al.* 2008; MANOLIO *et al.* 2008; PEARSON and MANOLIO 2008; YESUPRIYA *et al.* 2008). I would not be the first to note that the literature often reflects at least potential corporate, professional, or institutional conflicts of interest. There are also bioethical implications of the reluctance of leading journals, despite the known issues reviewed here, to publish negative studies: they are just not exciting enough. Yet from a biological point of view, good negative results may be some of the *best*, most positively instructive evidence about genetic architecture, and tentative positive studies, though headline grabbing, could be the most misleading.

What risk estimates (if any) should be given to customers of DNA testing companies or posted on public websites (*e.g.*, dbGAP at ncbi.nlm.nih.gov, or HuGe at <http://www.cdc.gov/genomics/hugenet/>, and see YU *et al.* 2008)? Should censoring or self-censoring be imposed? These are active, but difficult questions, especially in an impatient, market-driven age (MCGUIRE *et al.* 2007). One need think only of the cost to health care systems of using problematic associations as the basis of clinical or lifestyle intervention or of the potential social consequences of intervening in regard to genes widely treated as if they were “for” various kinds of behavior but that in most if not all cases involve complex interactions and only vaguely specifiable environmental conditions. As of this writing, services that use genotypes to give individualized risk are under legal scrutiny concerning what constitutes medical practice *vs.* “informational service” (WADMAN 2008).

MISSION ACCOMPLISHED? MOORE'S LAW AND MURPHY'S? WHAT DO WE WANT TO KNOW?

We have been served a feast of proverbial low-hanging phenogenetic fruit (BLANGERO 2004), genes with clear effects on normal or pathological traits in humans or countless other organisms. Findings to date have not had a major impact on public health; that may be sobering (think of sickle cell, known for >60 years), but should not be discouraging, since there is no reason to expect genetic engineering to be easy or quick just because a gene is known. The facts in hand also support our understanding of the evolutionary origin of genetic diversity. The knowledge we have gained constitutes substantial, positive, and reassuring scientific success. But in the face of quixotic trait loci, it is too early to declare “mission accomplished.”

The pace of biotechnological growth may have exceeded even Moore's law that computing power doubles every 2 years. It may soon reach its human threshold of 6+ billion nucleotides affordably identified in each diploid individual. With larger samples with cheaper and better DNA sequencing, many statistical power issues will fade as limiting factors. But another law seems to apply: Murphy's law, that whatever can go wrong will go wrong. Complex biological traits have many redundant, interlaced, stochastic, interacting, variable, emergent properties. Consistent with this, disease-related genes show every conceivable type of mutation. To the extent that each instance of a phenotype is etiologically unique, it can be resistant to science that depends on replication. Yet the strategies currently proposed are for even more technologically intense enumerative reductionism.

By contrast, quantitative genetics has been the basis of agricultural breeding formally or empirically, for thousands of years. Artificial selection basically works by aggregate empiricism, without needing to identify



FIGURE 6.—Mississippi River drainage to New Orleans. (A) The major contributing rivers. (B) The entire drainage system (courtesy of R. K. Weiss, ESRI Geographic Information Systems).

specific genes (FALCONER and MACKAY 1996; LYNCH and WALSH 1998; GRIFFITHS *et al.* 2004). The object is to change genetic architecture across generations, a speedup of the natural evolutionary process that produces organisms. But the need in biomedical genetics is the more specific challenge to address *individual* risk, within the individual's lifetime.

Proponents of systems biology suggest that “targeting of whole networks” (CHEN *et al.* 2008; HUNTER *et al.* 2008a) will be the answer. In principle this is analogous

to quantitative genetics applied to individuals within their lifetimes rather than to populations across generations. Computational and experimental approaches in simple model systems have nibbled at causal networks (*e.g.*, MOORE and WILLIAMS 2005; MOORE *et al.* 2005; KELLER *et al.* 2008; ZHU *et al.* 2008) or related individuals' relative position in genotype space to complex phenotypes (NIEVERGELT *et al.* 2008). But whether such approaches can tractably yield substantial, stable individual risk estimates or account for the bulk of

genetic risk, or will again run up against a few modestly predictive network genotypes, probably mostly rare, and a long tail of ephemeral ones, remains to be seen.

Unfortunately, I think the latter is the predictable outcome, even if the number of risk loci is small (PHAROAH *et al.* 2008). An important network should be identifiable without needing huge biobank studies. But its individual multilocus risk genotypes will almost automatically be rare to exceedingly rare, even if the number of loci is on the order of 10, much less hundreds, and the component alleles are individually common. So while mapping naturally occurring variation may be able to identify pathways that can then be followed up in *other* ways for biological understanding, the persistent hope for major individual risk prediction from that approach remains problematic at best.

Note that this discussion is about inferring biological mechanism and causation from naturally occurring variation. The story *may* be quite different in experimental biology where these things can be studied in replicable model systems at the cell and developmental level. Model systems provide mechanistic stereotypes, which may be quite useful in, for example, designing therapeutic agents to target stable biological pathways. But the degree to which that really reduces complexity or addresses the problem of connecting variation to outcome—individual genotypes to individual phenotypes—such as in the medical setting, is doubtful.

Given the present situation, and that the genetic architecture of no biological trait is as yet fully known, I think it is important if not urgent to try to get a much better understanding of the phenogenetic lay of the land that we are trying to infer. Computing power now makes possible very flexible, forward evolutionary simulation approaches that may be a key tool in such investigation. Useful simulation can be based on natural evolution-by-phenotype approaches (LAMBERT *et al.* 2008) or ones more centered on genomic questions (*e.g.*, HOGGART *et al.* 2007; PENG *et al.* 2007; EDWARDS *et al.* 2008).

But perhaps we also need a clearer goal. What would it mean to say that we understand the genetic basis of a biological trait? Must we know all the genes that contribute? All their variation? In all populations—or all species—and their frequencies? Specifically for each instance? These are literally hopeless ideals, since new variation is always arising by mutation and recombination and being lost to selection and drift, and environments are mostly unmeasured but surely always changing.

Let us look at this with an analogy, as in Figure 6. Suppose we are in New Orleans and wish to predict each instance of a flood of the Mississippi River (not counting hurricanes coming from the other direction). Figure 6A shows the major rivers that contribute to the Mississippi flow past New Orleans. Is this enough to monitor, or do we need to enumerate and measure all the streamlets in the entire drainage system, shown on the right, to make

our prediction? Their courses continually vary and are subject to the vagaries of local weather and human activity, and their contributions differ greatly from flood to flood. They are the fluvial equivalent of quixotic trait loci. They are real, but they are elusive and ephemeral. How far upriver do we have to go before we know enough?

Don Quixote is a satire whose hero is usually treated as the object of ridicule, but in fact he was a sympathetic hero who knew he was illusional, but remained undaunted in his quest against evil. In our quest to understand genetic architecture, it is possible to imagine that its complex appearance is an illusion and that the low-hanging fruit really does tell the biological tale. But if the exceptions really are exceptions, we risk being lured to struggle vainly for the rest of the bunch, which may remain out of our grasp until we have a more biologically grounded approach than to enumerate the quixotic fraction of nature. Meanwhile we should at least do our best to discriminate carefully before we tilt at windmills, thinking that they are giants.

I appreciate the editors giving me the freedom to express this perspective on a complex but important problem. The references cited are my quixotic attempt to exemplify issues fairly, by recent examples and overviews where back references can be found, but with no attempt to be comprehensive. I apologize to the countless authors whose equally worthy work I do not know or could not explicitly acknowledge. I thank Anne Buchanan, Allan Spradling, Barak Cohen, and an additional reviewer for helpful criticism of the manuscript. My work in this area is supported by grants from the National Institutes of Health (MH063749) and the National Science Foundation (BCS 0343442 and BCS 0725227) and by my Penn State Evan Pugh Professor's research fund.

LITERATURE CITED

- ABRAHAMS, B. S., and D. H. GESCHWIND, 2008 Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.* **9**: 341–355.
- ALLEN, N. C., S. BAGADE, M. B. MCQUEEN, J. P. IOANNIDIS, F. K. KAVVOURA *et al.*, 2008 Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.* **40**: 827–834.
- ALLIKMETS, R., and M. DEAN, 2008 Bringing age-related macular degeneration into focus. *Nat. Genet.* **40**: 820–821.
- ALTMULLER, J., L. J. PALMER, G. FISCHER, H. SCHERB and M. WJST, 2001 Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**: 936–950.
- ALTMULLER, J., C. SEIDEL, Y. A. LEE, S. LOESGEN, D. BULLE *et al.*, 2005 Phenotypic and genetic heterogeneity in a genome-wide linkage study of asthma families. *BMC Pulm. Med.* **5**: 1.
- AMARAL, P. P., M. E. DINGER, T. R. MERCER and J. S. MATTICK, 2008 The eukaryotic genome as an RNA machine. *Science* **319**: 1787–1789.
- ANONYMOUS, 2008 Risk loci, biological candidates and biomarkers. *Nat. Genet.* **40**: 257.
- BEGG, C. B., 2002 On the use of familial aggregation in population-based case probands for calculating penetrance. *J. Natl. Cancer Inst.* **94**: 1221–1226.
- BENJAMIN, E. J., J. DUPUIS, M. G. LARSON, K. L. LUNETTA, S. L. BOOTH *et al.*, 2007 Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med. Genet.* **8**(Suppl. 1): S11.
- BIRNEY, E., J. A. STAMATOYANNOPOULOS, A. DUTTA, R. GUIGO, T. R. GINGERAS *et al.*, 2007 Identification and analysis of functional

- elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.
- BLANGERO, J., 2004 Localization and identification of human quantitative trait loci: king harvest has surely come. *Curr. Opin. Genet. Dev.* **14**: 233–240.
- BODMER, W., and C. BONILLA, 2008 Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**: 695–701.
- BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**: 314–331.
- BOWCOCK, A. M., 2007 Genomics: guilt by association. *Nature* **447**: 645–646.
- BOYKO, A., S. WILLIAMSON, A. INDAP, J. DEGENHARDT, R. HERNANDEZ *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* **4**: e1000083.
- BRENA, R. M., T. H. HUANG and C. PLASS, 2006 Toward a human epigenome. *Nat. Genet.* **38**: 1359–1360.
- BRUDER, C. E., A. PIOTROWSKI, A. A. GJSBERS, R. ANDERSSON, S. ERICKSON *et al.*, 2008 Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**: 763–771.
- BUCHANAN, A. V., K. WEISS and S. M. FULLERTON, 2006 Dissecting complex disease: The quest for the philosopher's stone? *Int. J. Epidemiol.* **35**: 562–571.
- BUCHANAN, A. V., S. SHOLTIS, J. RICHTSMEIER and K. WEISS, 2009 What are genes for? Knowns we thought we knew. *BioEssays* (in press).
- BURMEISTER, M., M. G. MCINNIS and S. ZOLLNER, 2008 Psychiatric genetics: progress amid controversy. *Nat. Rev. Genet.* **9**: 527–540.
- CARDON, L. R., 2006 Genetics. Delivering new disease genes. *Science* **314**: 1403–1405.
- CAULFIELD, M., P. MUNROE, J. PEMBROKE, N. SAMANI, A. DOMINICZAK *et al.*, 2003 Genome-wide mapping of human loci for essential hypertension. *Lancet* **361**: 2118–2123.
- CHAKRAVARTI, A., 1999 Population genetics—making sense out of sequence. *Nat. Genet.* **21**: 56–60.
- CHANOCK, S. J., T. MANOLIO, M. BOEHNKE, E. BOERWINKLE, D. J. HUNTER *et al.*, 2007 Replicating genotype-phenotype associations. *Nature* **447**: 655–660.
- CHEN, S., E. S. IVERSEN, T. FRIEBEL, D. FINKELSTEIN, B. L. WEBER *et al.*, 2006 Characterization of BRCA1 and BRCA2 mutations in a large United States sample. *J. Clin. Oncol.* **24**: 863–871.
- CHEN, Y., J. ZHU, P. Y. LUM, X. YANG, S. PINTO *et al.*, 2008 Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**: 429–435.
- CHEUNG, V. G., R. S. SPIELMAN, K. G. EWENS, T. M. WEBER, M. MORLEY *et al.*, 2005 Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**: 1365–1369.
- CRINO, P. B., 2007 Gene expression, genetics, and genomics in epilepsy: some answers, more questions. *Epilepsia* **48**(Suppl. 2): 42–50.
- DAIGER, S. P., 2005 Genetics. Was the Human Genome Project worth the effort? *Science* **308**: 362–364.
- DAVULURI, R. V., Y. SUZUKI, S. SUGANO, C. PLASS and T. H. HUANG, 2008 The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.* **24**: 167–177.
- DENHAM, S., G. H. KOPPELMAN, J. BLAKEY, M. WJST, M. A. FERREIRA *et al.*, 2008 Meta-analysis of genome-wide linkage studies of asthma and related traits. *Respir. Res.* **9**: 38.
- DOBLHAMMER, G., and J. W. VAUPEL, 2001 Lifespan depends on month of birth. *Proc. Natl. Acad. Sci. USA* **98**: 2934–2939.
- DUBOIS, P. C., and D. A. VAN HEEL, 2008 New susceptibility genes for ulcerative colitis. *Nat. Genet.* **40**: 686–688.
- EDWARDS, T., W. BUSH, S. TURNER, S. DUDEK, E. TORSTENSON *et al.*, 2008 Generating linkage disequilibrium patterns in data simulations using genome SIMLA. Lecture notes in computer science: evolutionary computation. *Mach. Learn. Data Min. Bioinform.* **4973**: 24–35.
- EMILSSON, V., G. THORLEIFSSON, B. ZHANG, A. S. LEONARDSON, F. ZINK *et al.*, 2008 Genetics of gene expression and its effect on disease. *Nature* **452**: 423–428.
- EVANS, D. M., and L. R. CARDON, 2006 Genome-wide association: a promising start to a long race. *Trends Genet.* **22**: 350–354.
- EYRE-WALKER, A., and P. D. KEIGHTLEY, 2007 The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* **8**: 610–618.
- FALCONER, D., and T. MACKAY, 1996 *Introduction to Quantitative Genetics*. Longman, Essex, UK.
- FISHER, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**: 399–433.
- FODOR, F. H., A. WESTON, I. J. BLEIWEISS, L. D. MCCURDY, M. M. WALSH *et al.*, 1998 Frequency and carrier risk associated with common BRCA1 and BRCA2 mutations in Ashkenazi Jewish breast cancer patients. *Am. J. Hum. Genet.* **63**: 45–51.
- FOLSTEIN, S. E., and B. ROSEN-SHEIDLEY, 2001 Genetics of autism: complex aetiology for a heterogeneous disorder. *Nat. Rev. Genet.* **2**: 943–955.
- GEORGES, M., 2007 Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annu. Rev. Genomics Hum. Genet.* **8**: 131–162.
- GLUCKSMAN, P. D., M. A. HANSON, C. COOPER and K. L. THORNBURG, 2008 Effect of in utero and early-life conditions on adult health and disease. *N. Engl. J. Med.* **359**: 61–73.
- GOLDSTEIN, D. B., 2008 Genomics and biology come together to fight HIV. *PLoS Biol.* **6**: e76.
- GOLDSTEIN, D. B., K. R. AHMADI, M. E. WEALE and N. W. WOOD, 2003 Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**: 615–622.
- GÖRING, H. H., J. BLANGERO and J. D. TERWILLIGER, 2001 Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69**: 1357–1369.
- GRIFFITHS, A. J. F., D. T. SUZUKI, R. C. LEWONTIN, J. MILLER and W. GELBART, 2004 *Introduction to Genetic Analysis*. W. H. Freeman, San Francisco.
- HAIMAN, C. A., N. PATTERSON, M. L. FREEDMAN, S. R. MYERS, M. C. PIKE *et al.*, 2007 Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**: 638–644.
- HALL, J. M., M. K. LEE, B. NEWMAN, J. E. MORROW, L. A. ANDERSON *et al.*, 1990 Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**: 1684–1689.
- HAPPE, F., A. RONALD and R. PLOMIN, 2006 Time to give up on a single explanation for autism. *Nat. Neurosci.* **9**: 1218–1220.
- HARTL, D. L., and R. B. CAMPBELL, 1982 Allele multiplicity in simple Mendelian disorders. *Am. J. Hum. Genet.* **34**: 866–873.
- HARTONG, D. T., E. L. BERSON and T. P. DRYJA, 2006 Retinitis pigmentosa. *Lancet* **368**: 1795–1809.
- HILL, A. V., 2006 Aspects of genetic susceptibility to human infectious diseases. *Annu. Rev. Genet.* **40**: 469–486.
- HILLER, M., and M. PLATZER, 2008 Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Genet.* **24**: 246–255.
- HIRSCHHORN, J. N., and L. GENNARI, 2008 Bona fide genetic associations with bone mineral density. *N. Engl. J. Med.* **358**: 2403–2405.
- HOGGART, C. J., M. CHADEAU-HYAM, T. G. CLARK, R. LAMPARIELLO, J. C. WHITTAKER *et al.*, 2007 Sequence-level population simulations over large genomic regions. *Genetics* **177**: 1725–1731.
- HUGHES, A. E., N. ORR, H. ESFANDIARY, M. DIAZ-TORRES, T. GOODSHIP *et al.*, 2006 A common CFH haplotype, with deletion of CFHR1 and CFHR3, is associated with lower risk of age-related macular degeneration. *Nat. Genet.* **38**: 1173–1177.
- HUNTER, D. J., D. ALTSHULER and D. J. RADER, 2008a From Darwin's finches to canaries in the coal mine—mining the genome for new biology. *N. Engl. J. Med.* **358**: 2760–2763.
- HUNTER, D. J., M. J. KHOURY and J. M. DRAZEN, 2008b Letting the genome out of the bottle—Will we get our wish? *N. Engl. J. Med.* **358**: 105–107.
- HURLES, M. E., E. T. DERMITZAKIS and C. TYLER-SMITH, 2008 The functional impact of structural variation in humans. *Trends Genet.* **24**: 238–245.
- IOANNIDIS, J. P., 2007 Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64**: 203–213.
- IOANNIDIS, J. P., E. E. NTZANI and T. A. TRIKALINOS, 2004 'Racial' differences in genetic effects for complex diseases. *Nat. Genet.* **36**: 1312–1318.
- JANSSENS, A. C., M. GWINN, L. A. BRADLEY, B. A. OOSTRA, C. M. VAN DUIN *et al.*, 2008 A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am. J. Hum. Genet.* **82**: 593–599.

- JAKUSH, C. E., 2007 The Framingham Heart Study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.* **8**: 63.
- KATHIRESAN, S., O. MELANDER, C. GUIDUCCI, A. SURTI, N. P. BURTT *et al.*, 2008 Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40**: 189–197.
- KATO, N., T. MIYATA, Y. TABARA, T. KATSUYA, K. YANAI *et al.*, 2008 High-density association study and nomination of susceptibility genes for hypertension in the Japanese National Project. *Hum. Mol. Genet.* **17**: 617–627.
- KAVVOURA, F. K., and J. P. IOANNIDIS, 2008 Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum. Genet.* **123**: 1–14.
- KEIGHTLEY, P. D., and A. EYRE-WALKER, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.
- KELLER, M. C., and G. MILLER, 2006 Resolving the paradox of common, harmful, heritable mental disorders: Which evolutionary genetic models work best? *Behav. Brain Sci.* **29**: 385–404, 405–452.
- KELLER, M. P., Y. CHOI, P. WANG, D. B. DAVIS, M. E. RABAGLIA *et al.*, 2008 A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* **18**: 706–716.
- KHOURY, M. J., M. GWINN and M. S. BOWEN, 2007a Genomics and public health research. *J. Am. Med. Assoc.* **297**: 2347–2348.
- KHOURY, M. J., J. LITTLE, J. HIGGINS, J. P. IOANNIDIS and M. GWINN, 2007b Reporting of systematic reviews: the challenge of genetic association studies. *PLoS Med.* **4**: e211.
- KIM, P. M., J. O. KORBEL and M. B. GERSTEIN, 2007 Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc. Natl. Acad. Sci. USA* **104**: 20274–20279.
- KING, M. C., J. H. MARKS and J. B. MANDELL, 2003 Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**: 643–646.
- KONDRASHOV, A. S., S. SUNYAEV and F. A. KONDRASHOV, 2002 Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* **99**: 14878–14883.
- KRUEGER, C., and I. M. MORISON, 2008 Random monoallelic expression: making a choice. *Trends Genet.* **24**: 257–259.
- KWIATKOWSKI, D. P., 2005 How malaria has affected the human genome and what human genetics can teach us about malaria. *Am. J. Hum. Genet.* **77**: 171–192.
- LAMBERT, B. L., J. D. TERWILLIGER and K. M. WEISS, 2008 ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. *Bioinformatics* (in press).
- LI, M., P. ATMACA-SONMEZ, M. OTHMAN, K. E. BRANHAM, R. KHANNA *et al.*, 2006 CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nat. Genet.* **38**: 1049–1054.
- LIU, Y., C. HELMS, W. LIAO, L. C. ZABA, S. DUAN *et al.*, 2008 A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet.* **4**: e1000041.
- LOTERY, A., and D. TRUMP, 2007 Progress in defining the molecular biology of age related macular degeneration. *Hum. Genet.* **122**: 219–236.
- LUSIS, A. J., and P. PAJUKANTA, 2008 A treasure trove for lipoprotein biology. *Nat. Genet.* **40**: 129–130.
- LYNCH, M., 2007a The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. USA* **104**(Suppl. 1): 8597–8604.
- LYNCH, M., 2007b *The Origin of Genome Architecture*. Sinauer Associates, Sunderland, MA.
- LYNCH, M., and B. WALSH, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
- MACKAY, T. F., 2001 The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**: 303–339.
- MALLAL, S., E. PHILLIPS, G. CAROSI, J. M. MOLINA, C. WORKMAN *et al.*, 2008 HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358**: 568–579.
- MANOLIO, T., L. D. BROOKS and F. S. COLLINS, 2008 A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**: 1590–1605.
- MCGUIRE, A. L., M. K. CHO, S. E. MCGUIRE and T. CAULFIELD, 2007 Medicine. The future of personal genomics. *Science* **317**: 1687.
- MCPHERSON, R., A. PERTSEMLIDIS, N. KAVASLAR, A. STEWART, R. ROBERTS *et al.*, 2007 A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**: 1488–1491.
- MEISLER, M. H., J. KEARNEY, R. OTTMAN and A. ESCAYG, 2001 Identification of epilepsy genes in human and mouse. *Annu. Rev. Genet.* **35**: 567–588.
- MIKI, Y., J. SWENSEN, D. SHATTUCK-EIDENS, P. A. FUTREAL, K. HARSHMAN *et al.*, 1994 A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**: 66–71.
- MOORE, J. H., and S. M. WILLIAMS, 2005 Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* **27**: 637–646.
- MOORE, J. H., E. M. BOCZKO and M. L. SUMMAR, 2005 Connecting the dots between genes, biochemistry, and disease susceptibility: systems biology modeling in human genetics. *Mol. Genet. Metab.* **84**: 104–111.
- MORGAN, T. H., 1917 The theory of the gene. *Am. Nat.* **51**: 513–544.
- MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.
- NEEL, J. V., 1962 Diabetes mellitus: a “thrifty genotype” rendered detrimental by “progress.” *Am. J. Hum. Genet.* **14**: 353–362.
- NEVINS, J. R., and A. POTTI, 2007 Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.* **8**: 601–609.
- NEVERGELT, C. M., O. LIGIGER and N. J. SCHORK, 2008 Generalized analysis of molecular variance. *PLoS Genet.* **3**: e51.
- OBER, C., and S. HOFFJAN, 2006 Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun.* **7**: 95–100.
- OHTA, T., 2002 Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. USA* **99**: 16134–16137.
- OTTMAN, R., 2005 Analysis of genetically complex epilepsies. *Epilepsia* **46**: 7–14.
- PEARSON, T. A., and T. A. MANOLIO, 2008 How to interpret a genome-wide association study. *J. Am. Med. Assoc.* **299**: 1335–1344.
- PENG, B., C. I. AMOS and M. KIMMEL, 2007 Forward-time simulations of human populations with complex diseases. *PLoS Genet.* **3**: e47.
- PETRONIS, A., 2006 Epigenetics and twins: three variations on the theme. *Trends Genet.* **22**: 347–350.
- PHAROAH, P. D., A. C. ANTONIOU, D. F. EASTON and B. A. PONDER, 2008 Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.* **358**: 2796–2803.
- PHEASANT, M., and J. S. MATTICK, 2007 Raising the estimate of functional human sequences. *Genome Res.* **17**: 1245–1253.
- POLLARD, T., 2008 *Western Diseases: An Evolutionary Perspective*. Cambridge University Press, Cambridge, UK.
- PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.
- PROVINE, W., 1971 *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago.
- PROVINE, W. B., 1986 *Sewall Wright and Evolutionary Biology*. University of Chicago Press, Chicago.
- RANKINEN, T., A. ZUBERI, Y. C. CHAGNON, S. J. WEISNAGEL, G. ARGYROPOULOS *et al.*, 2006 The human obesity gene map: the 2005 update. *Obesity* **14**: 529–644.
- RAO, D. C., 2008 An overview of the genetic dissection of complex traits. *Adv. Genet.* **60**: 3–34.
- RAO, D. C., and M. A. PROVINCE (Editors), 2001 *Genetic Dissection of Complex Traits*. Academic Press, San Diego.
- REED, D. R., M. P. LAWLER and M. G. TORDOFF, 2008 Reduced body weight is a common effect of gene knockout in mice. *BMC Genet.* **9**: 4.
- REICH, D. E., and E. S. LANDER, 2001 On the allelic spectrum of human disease. *Trends Genet.* **17**: 502–510.
- SCRIVER, C. R., and P. J. WATERS, 1999 Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**: 267–272.
- SING, C. F., M. B. HAVILAND and S. L. REILLY, 1996 Genetic architecture of common multifactorial diseases. *CIBA Found. Symp.* **197**: 211–232.
- SJOBLOM, T., S. JONES, L. D. WOOD, D. W. PARSONS, J. LIN *et al.*, 2006 The consensus coding sequences of human breast and colorectal cancers. *Science* **314**: 268–274.

- STRANGER, B. E., M. S. FORREST, M. DUNNING, C. E. INGLE, C. BEAZLEY *et al.*, 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848–853.
- SZATMARI, P., A. D. PATERSON, L. ZWAIGENBAUM, W. ROBERTS, J. BRIAN *et al.*, 2007 Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.* **39**: 319–328.
- TERWILLIGER, J. D., and H. H. GORING, 2000 Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum. Biol.* **72**: 63–132.
- TERWILLIGER, J. D., and K. M. WEISS, 2003 Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann. Med.* **35**: 532–544.
- TODD, J. A., N. M. WALKER, J. D. COOPER, D. J. SMYTH, K. DOWNES *et al.*, 2007 Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**: 857–864.
- TROWELL, H., and D. BURKITT, 1981 *Western Diseases: Their Emergence and Prevention*. Harvard University Press, Cambridge, MA.
- TSAL, C. T., J. J. HWANG, M. D. RITCHIE, J. H. MOORE, F. T. CHIANG *et al.*, 2007 Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis* **195**: 172–180.
- VAN ROOIJ, E., N. LIU and E. N. OLSON, 2008 MicroRNAs flex their muscles. *Trends Genet.* **24**: 159–166.
- VISSCHER, P. M., 2008 Sizing up human height variation. *Nat. Genet.* **40**: 489–490.
- WADDINGTON, C. H., 1957 *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. George Allen & Unwin, London.
- WADMAN, M., 2008 Gene-testing firms face legal battle. *Nature* **453**: 1148–1149.
- WALLACE, D. C., 2008 Mitochondria as chi. *Genetics* **179**: 727–735.
- WANG, S., N. YEHYA, E. E. SCHADT, H. WANG, T. A. DRAKE *et al.*, 2006 Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2**: e15.
- WEISS, K. M., 1993 *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge University Press, Cambridge, UK.
- WEISS, K. M., 2005 Cryptic causation of human disease: reading between the (germ) lines. *Trends Genet.* **21**: 82–88.
- WEISS, K. M., and A. V. BUCHANAN, 2003 Evolution by phenotype: a biomedical perspective. *Perspect. Biol. Med.* **46**: 159–182.
- WEISS, K. M., and A. V. BUCHANAN, 2004 *Genetics and the Logic of Evolution*. Wiley-Liss, New York.
- WEISS, K. M., and A. V. BUCHANAN, 2008a The cooperative genome: organisms as social contracts. *Int. J. Dev. Biol.* (in press).
- WEISS, K. M., and A. V. BUCHANAN, 2008b *The Mermaid's Tale: Four Billion Years of Cooperation in the Making of Living Things*. Harvard University Press, Cambridge, MA (in press).
- WEISS, K. M., and A. G. CLARK, 2002 Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**: 19–24.
- WEISS, K. M., and J. D. TERWILLIGER, 2000 How many diseases does it take to map a gene with SNPs? *Nat. Genet.* **26**: 151–157.
- WEISS, L. A., Y. SHEN, J. M. KORN, D. E. ARKING, D. T. MILLER *et al.*, 2008 Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**: 667–675.
- WELLCOME TRUST CASE-CONTROL CONSORTIUM, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678.
- WONG, A. H., I. I. GOTTESMAN and A. PETRONIS, 2005 Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Hum. Mol. Genet.* **14**(Spec. No. 1): R11–18.
- WRIGHT, S., 1931 Evolution in Mendelian populations. *Genetics* **16**: 97–159.
- WRIGHT, S., 1934 Genetics of abnormal growth in the guinea pig. *Cold Spring Harbor Symp. Quant. Biol.* **2**: 137–147.
- WRIGHT, S., 1968 *Genetic and Biometric Foundations* (Evolution and the Genetics of Populations, Vol. 1). University of Chicago Press, Chicago.
- WRIGHT, S., 1978 *Variability Within and Among Natural Populations* (Evolution and the Genetics of Populations. Vol. 4). University of Chicago Press, Chicago.
- XAVIER, R. J., and D. K. PODOLSKY, 2007 Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**: 427–434.
- YESUPRIYA, A., E. EVANGELOU, F. K. KAVVOURA, N. A. PATSOPOULOS, M. CLYNE *et al.*, 2008 Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. *BMC Med. Res. Methodol.* **8**: 31.
- YU, W., M. GWINN, M. CLYNE, A. YESUPRIYA and M. J. KHOURY, 2008 A navigator for human genome epidemiology. *Nat. Genet.* **40**: 124–125.
- ZEGGINI, E., L. J. SCOTT, R. SAXENA, B. F. VOIGHT, J. L. MARCHINI *et al.*, 2008 Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**: 638–645.
- ZHU, J., B. ZHANG, B. DREES, R. B. BREM, L. KRUGLYAK *et al.*, 2008 Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**: 854–861.
- ZOLLNER, S., and J. K. PRITCHARD, 2007 Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80**: 605–615.

Communicating editor: A. SPRADLING