Genome-wide association studies and the genetic dissection of complex traits

Paola Sebastiani,[1]* Nadia Timofeev,[1] Daniel A. Dworkis,[2] Thomas T. Perls,[2] and Martin H. Steinberg[2]

The availability of affordable high throughput technology for parallel genotyping has opened the field of genetics to genome-wide association studies (GWAS), and in the last few years hundreds of articles reporting results of GWAS for a variety of heritable traits have been published. What do these results tell us? Although GWAS have discovered a few hundred reproducible associations, this number is underwhelming in relation to the huge amount of data produced, and challenges the conjecture that common variants may be the genetic causes of common diseases. We argue that the massive amount of genetic data that result from these studies remains largely unexplored and unexploited because of the challenge of mining and modeling enormous data sets, the difficulty of using nontraditional computational techniques and the focus of accepted statistical analyses on controlling the false positive rate rather than limiting the false negative rate. In this article, we will review the common approach to analysis of GWAS data and then discuss options to learn more from these data. We will use examples from our ongoing studies of sickle cell anemia and also GWAS in multigenic traits. Am. J. Hematol. 84:504–515, 2009. © 2009 Wiley-Liss, Inc.

Background

Over the past 30 years, about 1,200 disease-causing genes have been identified by studying well characterized phenotypes and by using gene mapping techniques [1,2]. The same approach has not been as successful in identifying the genetic modifiers of common diseases that have a genetic component shown by familial aggregation but do not follow Mendelian laws of inheritance. Examples include many of the common age-related diseases such as hypertension [3], diabetes [4,5], cardiovascular disease [6], and dementia [7], which are presumed to be determined by several genes (epistasis), and their interaction with environmental factors (gene-environment interaction). These common traits are a large public health burden and the discovery of the genetic profiles that can be used for disease risk prediction, prevention or treatment is one of the priorities of modern "personalized" medicine. Genome-wide association studies (GWAS) of common diseases have begun to propel us toward this goal.

Three major factors have made GWAS popular and feasible in a relatively short time and are critically reviewed in [8]. They are the common disease, common variant model (CD-CV) developed in the mid 1990s [9], the catalog of common variants created by the international HapMap project [10], and the rapid development of microtechnology for massive parallel genotyping [11,12]. The CD-CV model hypothesized that the genetic profile of common diseases is determined by genetic variants that are common in the population (frequency > 0.05) and have, individually, a small effect on the disease. This conjecture was based on both theoretical arguments and examples of heterogeneity of disease associated alleles including, for example, *APOE*-ε4 [13]. The CD-CV model made a strong case for the viability of GWAS because if the model was correct, the genetic basis of common diseases could be discovered by searching for common variants with different allele frequencies between cases and controls. To make this approach operational, the genetic community needed access to possibly all common genetic variants [14], and to technology for massive parallel measurements of these variants [15].

The most common genetic variations are single nucleotide polymorphisms (SNPs)—variation of a single base of the genome sequence among individuals—and it was estimated that the human genome has ~10 million SNPs [16]. However, the work of Gabriel in [17] provided the first evidence that, with the exception of hotspots of high recombination, the human genome is characterized by a block structure with sequences of SNPs that are highly correlated with each other in blocks of linkage disequilibrium (LD) (See Table I for technical definitions and Fig. 1). This structure implies that one can reconstruct the majority of variability of these blocks using a small subset of carefully selected tag-SNPs [18]. The International HapMap project was launched in 2003 to characterize common SNPs and to describe the block structure of the human genome that could be used to identify these tag-SNPs [19]. The first comprehensive catalog was published in 2005 and included more than 1 million common SNPs genotyped in 269 nuclear families of the three major ethnic groups [10]. This catalogue was useful to design the SNP microarrays produced by Illumina [12] and the latest SNP microarrays produced by Affymetrix [20].

By 2005, the genetic community was ready to embark on several GWAS. The first published GWAS reported the discovery of a functional SNP in the complement factor H that was associated with age related macular degeneration [21]. The study used a small case control design comparing the allele distribution of ~100,000 SNPs in 96 cases and 50

[1]Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, Massachusetts; [2]Department of Medicine, Boston University School of Medicine, 72 East Concord Street, Boston, Massachusetts

Conflict of interest: Nothing to report.

Contract grant sponsor: NIH/NHLBI; Contract grant numbers: R01-HL87681, R01-HL68970, R01-HL70735.

*Correspondence to: Paola Sebastiani, Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118. E-mail: sebas@bu.edu

Received for publication 5 March 2009; Revised 7 April 2009; Accepted 9 April 2009

Am. J. Hematol. 84:504–515, 2009.

Published online 16 April 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/ajh.21440

TABLE I. A Glossary of Terms Commonly Used in the Article

Term	Description
Autosomal trait	The mutation is on one of the chromosomes 1–22.
Carrier	A subject carrying one copy of the mutation that leads to a recessive disease.
Dominant trait	One copy of the mutated variant is sufficient for the trait to manifest.
False positive rate	The probability of rejecting the null hypothesis of no association when the null hypothesis is true and there is indeed no association.
Genotype	The SNP nucleotides in the chromosome pairs. A biallelic SNP can produce three genotypes, two homozygous genotypes when both chromosomes have the same alleles, and a hetherozygous genotype when the alleles are different. For example, a SNP with alleles A and G can produce the homozygous genotypes AA or GG and the hetherozygous genotype AG.
Identity by state (IBS) and identity by descent (IBD)	SNP alleles are IBS in two subjects when they are the same, and they are IBD when they are inherited from a common ancestor.
Linkage disequilibrium (LD)	Non random association between SNP alleles on the same chromosome.
Minor allele frequency (MAF)	The frequency of the minor allele.
Multiple comparison problem	This refers to controlling the false positive rate when testing many hypotheses simultaneously.
Penetrance	The probability that the genotype manifests in the phenotype.
Population stratification	The situation when allele frequency differences in cases and controls are due to differences in ancestry rather than association between genes and disease.
P-value	In the context of hypothesis testing, probability of observing a result as extreme as that observed in the data given the null hypothesis is true.
r^2	Measure of the allele correlations between two SNPs.
Recessive trait	Two copies of the mutated variant are necessary for the trait to manifest. An X-linked trait is always dominant in males.
SNP	Single nucleotide polymorphisms: variation of a single nucleotide.
Synonymous SNP	SNP that do not change aminoacid.
Type I error	Reject null hypothesis when null hypothesis is true.
X-linked	The mutation is on the X chromosome.

controls. The small sample size was balanced by a very careful ascertainment of cases and controls, gender matching, and stringent quality control rules to reduce the chance for spurious associations. The initial analysis identified one SNP in strong allelic association with the disease and resequencing of the region identified a novel functional variant. This first successful example of a GWAS was soon followed by several other applications to a variety of common diseases including prostate and breast cancers [22–27], Crohn's disease [28–30], coronary artery disease and diabetes [6,31,32], fetal hemoglobin expression [33], and other traits [34]. Table II provides examples of GWAS relating to hematological disease. The catalogue curated by Terri Manolio at the NHGRI provides updated information about results of GWAS [44], and a graphical display of the results mapped on the human genome is available from the HapMap website (http://www.hapmap.org/karyogram/gwas.html). Note that, besides a few exceptions, a caveat of these findings is that they have so far identified chromosomal regions implicated with the traits and more intensive studies will be needed to find the actual genetic variants responsible for the biological process linking genotypes to phenotypes.

The State of the Art

Several reviews describe in great detail the necessary steps to plan, analyze, and report the results of a GWAS [45–47]. These steps are briefly reviewed here.

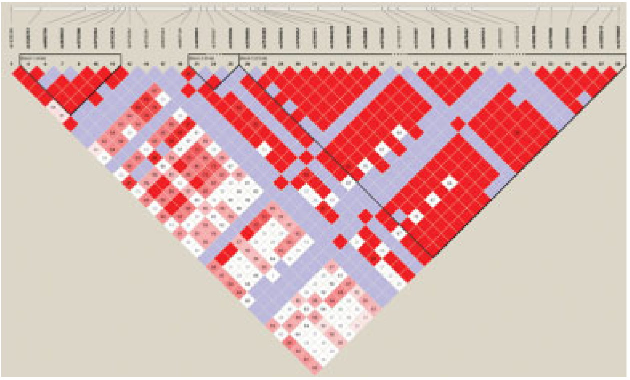

Figure 1. Linkage disequilibrium (LD) map of the gene *HAO2*. The map was generated using the program HaploView and genotype data of 58 SNPs from the 30 trios of the HapMap CEPH. The white bar on top shows the physical position of the SNPs. Each square represents the correlation of two SNPs, and the shades of red indicate the strength of the correlation ranging from no correlation (white) to strong correlation (red). Blue squares represent uncertain situations. The correlation analysis identifies three blocks of LD highlighted by the black outlines. The variability of each of the three blocks can be captured by a small number of tag SNPs: for example the larger block of 23 SNPs on the right can be described by six SNPs.

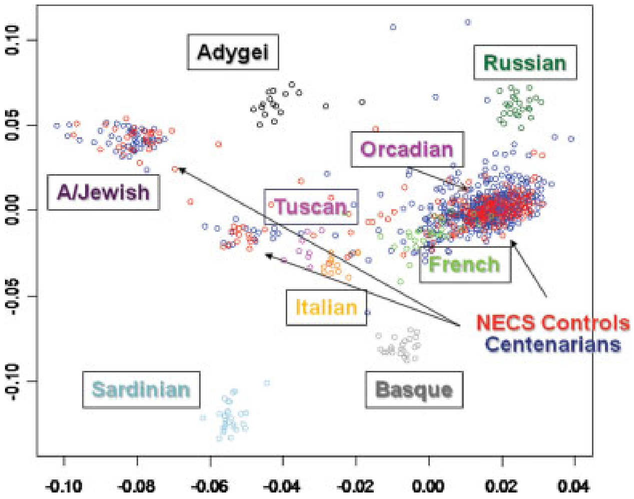

Figure 2. Scatter plot of the first two principal components computed with GWAS data from centenarians and controls of the New England Centenarian Study and the Human Genome Diversity Panel. The principal component analysis summarizes the genome wide genotype data into linear combinations that explain the overall variability of the data and can be ranked by decreasing variability so that the first two principal components explain the largest proportion of the data variability. A plot of the first two principal components can give insight about the level of genetic diversity between cases and controls and uncover possible stratification. In this example, the *x*-axis reports the first principal component and the *y*-axis reports the second principal component. Each point represents a subject and is colored based on either the known geographic ancestry or the trait. Subjects cluster in several groups that represent closed European populations such as Sardinians or Adygeis and three major clusters of North Europeans, South Europeans, and Ashkenazi Jewish that include NECS centenarians and controls. In this example, the overlapping of the three clusters of centenarians and controls suggests no substantial stratification.

Study design

Typically, a GWAS uses a case–control design, in which cases are ascertained based on the trait of interest [48]. The choice of control subjects is often less obvious. A control subject should be disease free, but must also be free of other traits that are not shared by cases. For example, if

TABLE II. GWAS in Hematological Disease

Study	Year	Phenotype	Comment	Ref.
Yang et al.	2007	Multiple including plasma factor levels, hematocrit, viscosity.	Used GWA and linkage studies to analyze multiple-linked phenotypes dealing with hemostasis and hemodynamics.	[35]
Ouwehand	2007	Platelet function.	Mixed candidate gene and GWAS approaches to examine phenotypes in the Wellcome Trust Case Control Consortium.	[36]
Huang et al.	2007, 2008	Cytotoxicity of etoposide and daunorubicin.	Integrated genotype and expression data to study drug toxicity for common treatments of leukemia and lymphoma.	[37,38]
Di Bernardo et al.	2008	Chronic lymphocytic leukemia.	First evidence for the existence of common, low-penetrance susceptibility to a hematological malignancy.	[39]
Sarasquete et al.	2008	Osteonecrosis of the jaw secondary to bisposphonate therapy.	Discovered significant SNPs in *CYP2C8* in one population of patients with multiple myeloma on bispohosphonate.	[40]
Cooper et al.	2008	Warfarin maintenance dose.	Found multiple associations in a single population, validated known genetic association with *VKORC1* and *CYP2C9* in two populations.	[41]
Menzel et al., Uda et al., Sedgewick et al.	2007, 2008, 2008	Fetal hemoglobin levels.	Association of SNP in *BCL11A* with fetal hemoglobin levels in healthy northern Europeans, Sardinians with thalassemia and African Americans with sickle cell anemia.	[33,42,43]

the phenotype of interest is known to manifest within a certain age, it is tempting to choose controls that are much older than the cases to guarantee that they never develop the phenotype. This choice for controls would introduce confounding, and the discovered associations may be related to aging rather than the trait of interest. In general, matching controls on variables that are not of interest in the study, such as exposure to some environmental conditions, is important to avoid introducing spurious associations. However, some caution is necessary to avoid overmatching and losing of generality of the results or missing important associations [49]. Controls selected from families of the cases often offer protection against confounding by matching exposures to risk factors, or genetic background when controls are genetically related to the cases, this being the rationale of family based association studies [50].

Referent cohort subjects used in other studies can also be used as controls. This was the strategy used by the Wellcome Trust Case–Control Consortium that used the same pool of 3,000 controls chosen from the British population to search for genetic modifiers of seven common traits [6]. This strategy is becoming more and more feasible with the increasing availability of GWAS data from dbGaP, the database of genotype–phenotype associations (http://www.ncbi.nlm.nih.gov/gap), and the Illumina control database. However, this approach can introduce confounding due to population stratification. Population stratification has been a well known source of spurious associations [48] and occurs when allele frequency differences in cases and controls are due to differences in ancestry rather than associations between genes and disease. Population stratification is, however, easy to detect in GWAS using multivariate statistical techniques implemented in popular programs such as PLINK [51] and EIGENSOFT [52]. Figure 2 provides an example using data from the New England Centenarian Study [53] and the Human Genome Diversity Panel [54]. The analysis can help to describe the genetic background of cases and controls when reference groups with known ancestry are included in the analysis, and it can identify different levels of genetic diversity between cases and controls that need to be taken into account in subsequent statistical analysis.

When the trait of interest is a quantitative measure, such as blood pressure or fetal hemoglobin concentration, the inclusion criteria for subjects should ensure that sufficient variability of the trait is represented by the sample. Other covariates that could be associated with the trait should be measured so that further analysis can be adjusted and the genetic effect be distinguished from other factors.

Choice of genotyping platform and resource allocation

Despite some different choices of SNPs to be typed—either tag-SNPs chosen to capture substantial variability of the common SNPs in the HapMap, or randomly selected SNPs, or a combination of both strategies—all of the major genotyping platforms that are currently available offer similar coverage of common variants in the HapMap [46,55]. The coverage ranges between 500,000 and 1,000,000 SNPs per sample, and these numbers continue to steadily increase. Initial estimates suggested that, for example, the Illumina humanhap 300 array captured 75% of HapMap common SNPs with an allelic correlation of 80% in subjects of European ancestry, but the coverage was only 28% in subjects of African ancestry. The 80% allelic correlation is a measure of the LD between two SNPs that is based on the correlation coefficient r^2 (see Table I) and measures the correlation of two SNP alleles on the same chromosome [56]. Equivalent arrays produced by Affymetrix had lower coverage of common SNPs for Europeans but higher coverage for Africans [55]. These figures are higher in denser arrays, with a clear gain of coverage in African subjects, so that for example the Illumina 610 provides approximately 60% coverage of the HapMap common SNPs with 80% correlation, meaning that the SNPs in the array correlate with 60% of HapMap common SNPs with $r^2 > 0.8$. Recent studies have suggested that the gain of coverage using denser SNP arrays can almost be recovered by the imputation of untyped SNPs in Europeans and genotyping more samples with less dense arrays may increase the power of a GWAS [57]. Because imputation of untyped genotypes in subjects of African or Asian ancestry is less accurate [58], this strategy may not be useful in studies of non European subjects. These analyses used the HapMap catalogue of common SNPs as a gold standard. Estimates of the coverage of the full set of common and rare variants have been recently reviewed using sequence data and suggest that the initial calculations of the coverage of common platforms were overoptimistic [59]. Larger coverage of the real variants rather than those reported in the HapMap will require additional SNP discovery [60]. It is also important to emphasize that the majority of SNPs in commercially available arrays do not affect protein structure and appear unlikely to affect gene expression so that this design choice limits the discovery power of GWAS to locating chromosomal regions rather than the genetic variants that are responsible for the trait. This initial discovery may be sufficient for prognostic modeling, but understanding the mechanism that leads to the trait of interest will require more

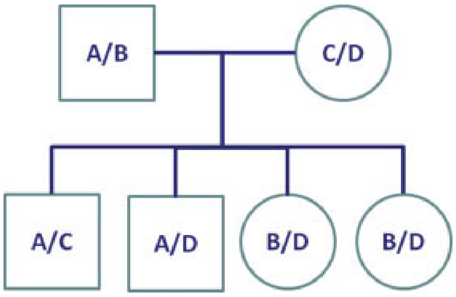

Figure 3. Examples of 0, 1, and 2 alleles that are shared IBD. The pedigree displays two parents who carry the alleles A and B (father) and C and D (mother) of a genetic locus. The four children inherit the allele A from the father and C from the mother (child 1), the allele A from the father and D from the mother (child 2) and the allele B from the father and D from the mother (child 3 and 4). If we compare children 1 and 2, they share the allele A from the father, and hence one allele is shared IBD. If we compare the child 1 and 3 they do not share any allele, and hence zero alleles are shared IBD, while children 3 and 4 share the same allele B from the father and the same allele D from the mother and so two alleles are shared IBD.

TABLE III. Relation Between Probabilities of Genome-Wide Alleles Shared IBD and Relatedness

	IBD probabilities			
Relative pair	0	1	2	P(IBD)
MZ twins	0	0	1	1
Full sibs	0.25	0.5	0.25	0.5
Parent-offspring	0	1	0	0.5
Grandparent-grandchild	0.5	0.5	0	0.25
Half-sibs	0.5	0.5	0	0.25
Avuncular	0.5	0.5	0	0.25
First cousin	0.75	0.25	0	0.125
Unrelated	1	0	0	0

Column 1 describes the type of relation, columns 2–4 report the genome-wide proportion of alleles shared by IBD that can be 0 (column 2), 1 (column 3), and 2 (column 3). The last column indicates the expected probability of alleles shared by IBD for various relations. For example, monozygotic twin (row one) should share the same DNA, and therefore, the probability of any two alleles shared IBD is 1. Unrelated samples with a probability 1 of any two alleles shared IBD can point to sample swaps.

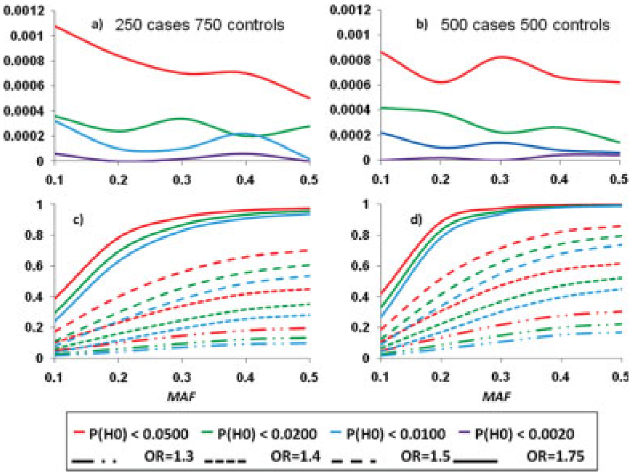

Figure 4. (a) Error rate (*y*-axis) of different Bayes rules for detecting allelic association in an unbalanced design with 250 cases and 750 controls. Each line represents the error rate as a function of the minor allele frequency (MAF) displayed on the x-axis and different colors represent different decision rules (red: posterior probability of the null hypothesis *P*(H0) < 0.05; green: *P*(H0) < 0.02; blue: *P*(H0) < 0.01; and blue: *P*(H0) < 0.002). (b) Same analysis with a balanced design of 500 cases and an equal number of controls. (c) Power (*y*-axis) of the different decision rules for different effects (see legend) and different MAF in cases (*x*-axis) with the unbalanced design. (d) Same analysis as in (c) for the balanced design. All analyses were based on simulating 50,000 data sets for each combination of sample size, threshold on the posterior probability of association, MAF and effect size. The analysis shows that the error rate tends to increase with the rarity of alleles so that the threshold on the posterior probability can be adjusted to the MAF of cases to optimize the power. For example, to test the association of a common variant, a Bayesian rule that rejects the null hypothesis when *P*(H0) < 0.05 has an error rate of approximately 6 × 10^{-4} and a power ranging between 70% in the unbalanced design and 85% in the balanced design. This large power is achieved with half the sample size suggested in other studies that use frequentist methods.

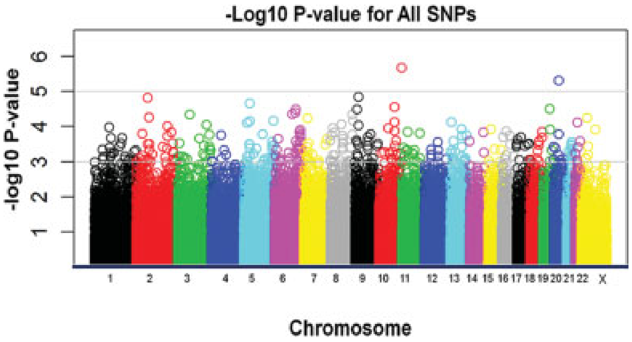

Figure 5. Manhattan plot displaying the −log10(*P*-value) of the association of approximately 270,000 SNPs with response to hydroxyurea in 123 sickle cell anemia patients. The association was tested using linear regression with an additive genetic effect in PLINK. To bound the overall false positive rate to 5%, the Bonferroni correction would require a *P*-value < 10^{-7} and hence −log10(*P*-value) > 7. In this analysis, none of the SNPs reaches genome-wide significance, but there are clearly several regions in the genome that show significant associations. The stringent condition on the significance levels seriously inflates the false negative rate.

intensive studies such as fine mapping or sequencing for the discovery of functional variants [60].

Quality control

Once the DNA samples are hybridized to the arrays and the arrays are read and quantified using some detection method [61], genotypes are called using algorithms that are specific to the different arrays. Some decisions must be made during this phase to trade off sensitivity for specificity and may lead to missing data if the algorithm cannot distinguish among the three genotypes with certainty. Interestingly, it has been noted that excessively stringent conditions on acceptable genotype calls may introduce bias and inflate the false positive rate because the pattern of missing data may be informative rather than random. For example, rare genotypes may be more likely to be missing than common genotypes [46]. On the other hand, excessively relaxed conditions on acceptable genotype calls may introduce erroneous genotypes in the data to be analyzed and cause spurious associations due to technical errors. A good heuristic seems to opt for more relaxed thresholds on genotype calling to maximize the power of the study followed by careful assessment of the quality of the genotype calls of those SNPs identified in the analysis. Inspection of the plots showing the clusters of genotypes can help to detect these technical errors.

Issues with sample quality can be easily identified by the proportion of SNPs that fail to have their genotype called, and samples with a low call rate (usually more than 7% SNP genotypes missing) should be disregarded from the analysis. There are other sources of errors that may impact the analysis and are less evident. Experience shows that inefficient sample tracking—for example not using bar code reading systems to track samples—or sample swaps is unfortunately more frequent than lab technicians would like to admit [62]. There are at least two analyses that can be easily conducted to point to possible errors. The first analysis examines the agreement between gender specification in the phenotypic data and the gender inferred by heterozygosity of the SNPs on chromosome X. The second analysis consists of scoring the genetic similarity between every pair of subjects in the study using estimates of the alleles that are shared identity by descent (IBD). Two alleles are shared IBD when they are identical copies of the same ancestral allele. Mendel's laws of inheritance can be used to estimate the probability that two family members share 0, 1, or 2 alleles IBD. For example, monozygotic twins will always share two alleles IBD with probability 1, whereas siblings will share 0 alleles with probability [1/4], one allele with probability [1/2], and two alleles with probability [1/4] (see examples in Fig. 3). This estimation can be generalized to any relatives [63] and to genome wide alleles that are IBD. The analysis is computationally very intensive and estimates the proportion of alleles that are IBD using frequencies of alleles that are the same, also known as alleles identity by state or IBS [63]. The estimates can be used to quantify the degree of relatedness among subjects in the study as shown in Table III [64], and can identify repeated samples, as well as unknown relations. If the study includes subjects with known familial relations, the analysis can compare expected and observed relations in the sample as a quality control metric. The popular software PLINK for the analysis of GWAS has a module for the estimation of IBS and IBD [51].

Other SNP-specific analyses are often carried out to remove low quality or unreliable data. For example, some groups remove SNPs with a minor allele frequency (MAF) below a fixed threshold or not in Hardy Weinberg Equilibrium before conducting the statistical analysis. Although SNPs with low MAF can inflate the false positive rate (see Discussion below), departure from Hardy Weinberg Equilibrium in cases may actually be consistent with significant associations [46] and these SNPs should be flagged for further inspection but analyzed nevertheless.

Statistical analysis

The core of the analysis is typically a statistical test conducted for each individual SNP in the array with the objective of selecting those SNPs that exhibit a significant association with the trait. Formally, the procedure consists of testing, for each SNP, the null hypothesis that there is no association between the SNP and the phenotype against the alternative hypothesis that there is an association.

Frequentist approaches. The frequentist approach is most commonly used to test these hypotheses and weighs the evidence against the null hypothesis by the *P*-value that is defined as the probability of observing a stronger association than that estimated from the data, when there is indeed no association. For example, if the association is measured by the odds ratio for the disease between exposed (in this case carriers of the risk allele) and unexposed subjects, the *P*-value is the probability of observing an odds ratio more extreme than that estimated from the data, assuming that there is no association. If there is association, the *P*-value should be very small because estimating a large odds ratio would be unlikely to happen. Therefore, a small *P*-value is taken as evidence against the null hypothesis, or evidence of a significant association, and this is the rationale for the decision rule to reject the null hypothesis when the *P*-value is smaller than a fixed threshold. The fixed threshold is called the significance level. This decision rule is not error free and rejecting the null hypotheses when it is true is known as type I error. Type II error is defined as accepting the null hypothesis when the alternative hypothesis is true and it is related to the power of the test [65]. This approach requires a method to estimate the association and a significance level used as the threshold for the *P*-value. We will discuss the former here, and address the latter issue in the section about multiple comparisons.

The methods to estimate the association can be categorized into two groups based on whether the trait is continuous or categorical. A continuous trait following a normal distribution is modeled as a function of the genetic effect using linear regression. General genotype association is tested using an analysis of variance that compares the distribution of the trait in the three genotypes MM, Mm, and mm [65]. A popular alternative is to represent the three genotypes by the variable X taking values $0 = \text{MM}$, $1 = \text{Mm}$, and $2 = \text{mm}$ and use this variable in linear regression. The parameterization is known as the additive genetic model, and the regression coefficient of the variable X represents the average change in the trait for each extra copy of the allele m, with a significant regression coefficient denoting a significant association. The additive model is easy to interpret and therefore the most commonly used. However, note that a lack of association in the additive model does not imply no association between the SNP and the trait. The genetic association can be adjusted for other covariates by adding them in the regression equation, or by modeling the residuals from the regression model that includes only the covariates. Adjustment should be done only for covariates that are significantly associated with the trait to avoid unnecessary loss of power. If the trait does not follow a normal distribution, the *P*-value can be computed using permutation methods, but the computational burden can be very high. Also, permutation methods usually tend to be less powerful [66].

When the trait is categorical and subjects are grouped as cases and controls, there are several association tests that one can use. General genotype association can be tested using the traditional χ^2 test of independence in a 2×3 contingency table. More parsimonious procedures include the Armitage trend test in which genotypes are recoded to model a linear increase in the odds of the disease on the logarithmic scale for each different genotype, or associations of dominant or recessive models in which genotypes are aggregated in two groups [48]. The dominant model for the allele M, tests the associations of genotypes grouped as MM and Mm versus mm, whereas the recessive model for the allele M uses the grouping MM versus Mm and mm. Allelic association can be tested by recoding the data from genotypes into alleles. However, the association is difficult to interpret at the level of individuals. Some care is necessary when testing the associations of rare variants. In this case, the frequency of some genotypes may be too small (<5) for the χ^2 test to be valid and the Fisher exact test should be used. A limitation of these different tests for categorical traits is that they cannot include the effects of covariates. Logistic regression can be used as an alternative by modeling the odds for association in the logarithmic scale using regression. The regression equation can include the genetic effect but also covariates and can be extended to include multiple interacting SNPs as well as gene-environment interactions. Additionally, these common statistical methods rely on large sample approximation, and therefore care is needed when testing the association of rare variants.

Bayesian approaches. Bayesian methods are grounded in a very different conceptual framework and are becoming more popular in genetic epidemiology [67]. The principle of

Bayesian tests of association is to first assume prior probabilities on the two hypotheses of no association and association, then use the data to update the prior probabilities of the two hypotheses into their posterior probabilities. The decision to reject the null hypothesis is then based on appropriate thresholds on the odds of the posterior probabilities or, equivalently, on the posterior probability of the null hypothesis. The choice of the best threshold can be based on trading off sensitivity and specificity of the Bayesian decision rule (see Fig. 4 for an example). The posterior odds are computed by multiplying the prior odds by the Bayes factor that can be calculated in closed form for some of the models described earlier [65]. Nonlinear models usually require sophisticated computational procedures and stochastic computations known as Markov Chain Monte Carlo methods [67]. These methods are very powerful and are commonly used in genetics as they are the engine of very accurate procedures for haplotype reconstruction from unphased data [68] and different imputation methods [69,70]. A Bayesian method has also been proposed to discover the most likely set of functional SNPs in fine mapping or sequence data. This method, known as Bayesian quantitative trait nucleotide (BQTN) analysis, uses Bayesian model selection to test the associations of all possible sets of SNPs with a trait. Recently, it was used to identify four to seven variants that explain the total variability of plasma levels of clotting factor VII (FVII), a risk factor for cardiovascular disease [71].

It is important to emphasize the major theoretical difference between the frequentist and the Bayesian approaches to hypothesis testing. In the frequentist approach, the decision to reject the null hypothesis of no association is typically based on controlling the probability of the type I error and this is done by imposing a threshold on the *P*-value. This procedure does not assess per se whether the null hypothesis is true or false. In the Bayesian approach, the decision to reject the null hypothesis is based directly on the probability that the null hypothesis is false, given the evidence provided by the data. Only Bayesian procedures allow for an explicit assessment of the likelihood of parameters and hypotheses [72].

Family-based studies. The analyses described so far assume that subjects are unrelated. If the study is family based, the analysis has to take into account the correlation between relatives due to their common genetic background. Different approaches have been proposed to account for the family structure. A class of methods known as family based association tests (FBAT) is commonly used in GWAS and reviewed in detail in [50]. The FBAT generalizes the transmission disequilibrium test (TDT) that was introduced in [73] to test for linkage and association in family trios (two parents and an affected child). The intuition of the TDT is to compare the number of alleles that are transmitted to affected offspring from unaffected parents with those expected under the null hypothesis of no association between genotype and phenotype. The method was generalized to accommodate sibships [74], missing genotypes [75], and to include both related and unrelated subjects [76–78]. The FBAT generalizes the TDT by computing the covariance between genotype and phenotype that are centered in such a way to accommodate different sampling designs. The method can be used to study survival traits [79] in addition to qualitative and quantitative phenotypes [80] and general pedigrees. The FBAT is robust to population stratification because of the use of controls within family members who share the same genetic background. Extensions of the TDT that include the founder genotypes to increase the power of the study have been proposed but do not protect against population stratification [81]. The FBAT does not require any parametric assumption about the phenotype, and the null hypothesis of no association can be tested using large sample approximations that appear to work well with at least 10 informative families [50]. Although this model-free feature makes the test robust to model misspecifications, it limits its use to test for associations that can be at most adjusted for the effect of other covariates such as gender or environmental effects, but neither gene–gene interactions nor gene-environment interactions can be tested. Model-based approaches overcome these limitations by directly modeling the effect of the genotype on the phenotype using regression and including extra terms that model explicitly the within family correlation. Examples are the variance component model proposed in [82] and implemented in the package QTDT, its recently proposed extension that uses the family structure to limit the genotyping effort to necessary family members [83], and generalized estimating equations (GEE) that model the variance covariance matrix of the observations by taking into account the family structure [84]. Hybrid approaches that combine family based tests from extended pedigrees with association tests in unrelated individuals can be very powerful and leverage the strengths of both approaches [85]. For example, Uda et al. [42] used a combination of variance components models in extended pedigrees of Sardinians and analysis in unrelated subjects to confirm associations of SNPs in *BCL11A* with fetal hemoglobin expression in β-thalassemia carriers and sickle cell anemia patients. In general, positive associations from both linkage and association studies strengthen the evidence for true positive findings [85].

Power and multiple comparisons

Several articles report the power analysis of GWAS for given sample sizes and effect sizes. For example, Wang and colleagues showed that a sample size of 1,000 cases and 1,000 controls allows for estimation of an allelic odds ratio of 1.5 with 80% power when the disease allele is common, with a frequency between 0.4 and 0.5 [86]. The sample size necessary to detect the same effect when the disease allele is less common, for example, a frequency of about 10%, is 2,000 cases and 2,000 controls and increases almost exponentially with smaller disease allele frequencies [8]. These calculations make two assumptions: (1) they assume stringent conditions on the *P*-value to control the overall probability of type I error, and (2) they assume that the analysis is conducted using standard logistic regression. The two assumptions are not independent, because both the *P*-value and power are relative to the statistical method used for the analysis and not only to the sample size. Imposing stringent conditions on the significance of individual tests has become the popular approach to control the global significance of multiple tests. This number is related to the so called "family-wise error rate" that represents the probability of one or more type I errors in testing multiple hypotheses. The rationale is that, if the standard threshold α is adopted to accept a significant association when testing one single hypothesis, then the probability of one or more type I errors in testing N hypotheses is given by the formula

$$\text{probability (number Type I error} > 0) = 1 - (1 - \alpha)^N$$

This probability is essentially 1 when more than 100 tests are conducted with an individual significance level of 0.05.

An equivalent way of assessing the magnitude of the problem is that the number of false positive associations that are expected by chance when testing 500,000 null hypotheses, assuming all of the null hypotheses are true, is $\alpha \times 500{,}000$. For example, this number is 25,000 when $\alpha = 0.05$. The Bonferroni correction attempts to limit this number by reducing the individual significance of each test so that the overall number of expected false positive associations is 5% of the number of tests that are conducted [48]. In practice, the Bonferroni correction consists of dividing the usual significance level by the number of tests that are conducted and requires dramatic P-values $< 10^{-6}$ or smaller to meet genome-wide significance [47]. Figure 5 shows an example taken from a study of the response of fetal hemoglobin to treatment with hydroxyurea in patients with in sickle cell anemia [87].

It is well known that this correction is too conservative and reduces the power dramatically and unnecessarily [88]. Controlling the false discovery rate rather than the overall false positive rate has been proposed as a less conservative method. The false discovery rate is the proportion of false positive associations among the detected significant associations and can be controlled for using a simple algorithm [89]. Work conducted in the past few years to reduce the number of falsely significant associations in microarray data analysis also provides a variety of solutions even in small samples with correlated data [88,90–92]. Neither procedure changes the rank of the *P*-values, but simply provides additional guidance as to which associations are most significant across the entire study. Some statistical methods are more powerful than others and often a substantial increase in power can be accomplished by adopting more sophisticated statistical analyses without needing to increase sample size.

When using a frequentist approach to control the type I error or the family wise error rates in multiple testing, one can increase the power only by increasing the sample size. This is not the case with Bayesian procedures. Bayesian hypothesis testing does not base the decision to reject the null hypothesis on the significance level and every hypothesis is tested independently from the others. However, the threshold used on the posterior probability of the hypotheses implies that every time a null hypothesis is rejected because the posterior probability of the null hypothesis P(H0) is smaller than a chosen threshold, there is a probability P(H0) of error. One can use this number to compute the probability of one or more errors, and assess the global error rates through simulations. Figure 4 shows examples of different Bayesian decision rules to genome-wide hypothesis testing and their sensitivity and specificity in a small sample study. The prior probability of the hypotheses can also incorporate information about the number of hypotheses that are expected to be true [93,94].

Besides philosophical differences and some gain in power of Bayesian procedures, testing many hypotheses will inevitably increase the probability of errors, both type I and II, and relying solely on statistical methods is not sufficient. It is becoming clear that by controlling the probability of the type I error we may be ignoring the majority of the biologically important findings [95] and we need methods to be able to look at associations with less stringent *P*-values or posterior probabilities. To this end, we introduced a Bayesian procedure for analysis of GWAS that uses a hierarchical set of filters to reduce the false positive rate without imposing unnecessary stringent thresholds [96]. This procedure leverages the patterns of linkage disequilibrium in the human genome to accept as significant only those associations that are supported by associations of SNPs in the same LD blocks, and our experimental evaluation suggests that these filters help reducing the false positive rate by 50% [96]. Several authors have proposed strategies that leverage properties of the human genome or knowledge of disease mechanisms and pathways more likely to be involved in the disease to prioritized genes [97].

Validation and replication

The inflated false positive rate due to multiple testing, the issue of population stratification that can confound associations and technical errors that can be committed during the collection of DNA samples, and the analysis require the replication of the results from GWAS in at least one independent study to guarantee their validity [98]. Replication should not be confused with technical validation of the genotype data that requires genotyping a small set of SNPs with a different technology. This strategy is highly recommended in [98] but not very often adopted or reported. It has been suggested that a convincing replication should use a larger sample from an independent study population, with the same genetic background of the primary study population, the same definition of the phenotype, and should report the association of the same SNPs with the same genetic model and show the same genetic effect. Inclusion of proxy SNPs that are in high LD is still an open issue: some authors recommend this practice as further evidence of a real association [98] while others recommend against it because it is unnecessary [46]. Replication of findings in a population with different genetic backgrounds can strengthen the evidence of true associations and identify variants that are robust to different genetic background and environmental exposures [43]. However, failure to reproduce an association in a genetically different population should not be taken as evidence of a false positive. Also the requirement of a larger size sample for the replication study when compared with the primary study is arguable. The large sample size of the primary population is necessary to achieve sufficient power with genome-wide significance levels. However, replication usually is limited to a small selection of SNPs in which case there is no multiple testing problem and traditional significance levels should be acceptable.

Another emerging approach to replication of GWAS is the use of meta-analysis. By combining the results of different studies, statistical meta-analysis can also provide additional power for the discovery of new associations. For example, meta-analysis has recently produced the discovery of additional loci associated with BMI [99], lipid traits [100,101], and was crucial for the discovery of robust associations with diabetes [102]. Recent papers that assemble the results from different GWAS using meta-analysis often rely on imputation-based analyses because the original studies used different genotyping platforms with different SNPs [99–101]. Imputation of untyped SNPs is a convenient and often accurate procedure to synchronize genotype data of different arrays [58], and the accuracy can be very high in studies of populations that are well represented in the HapMap project. However, because imputation of data in different cohorts is based on the same reference haplotypes from the HapMap projects, there could be an intrinsic bias toward positive replications and initial results based on imputed data should be followed by actual genotyping of the missing data to confirm real effects. The efficacy of imputation of data from populations that are not well represented in the HapMap project, such as African Americans, is still an open question [58], and similar caution should be used when imputed data of these populations are used for meta-analysis.

Replication of findings in different studies can strengthen the evidence for a real association. However, this practice appears to be overrated. The majority of SNPs in the commercially available arrays are chosen to conveniently tag regions of the human genome with the consequence that very few SNP determine changes of aminoacids, and very few are located in known regulatory regions of the genome. Therefore, associations discovered through GWAS identify chromosomal regions that need finer mapping or sequencing studies to find the functional variants responsible for the disease. This intrinsic limitation of GWAS should imply that the identification of the same chromosomal region in independent studies is sufficient to move from statistical association of convenient SNPs to the discovery of the true variants or other regulatory elements that can lead to novel biological insights.

Looking Ahead

In February 2009, the catalogue of genome-wide association studies at the NHGRI listed more than 250 publications with results of GWAS and more than 1,000 SNPs that were discovered as associated with a variety of traits and disease [44]. This large number proves that the GWAS approach works and can indeed discover common variants related to common diseases, but it is much smaller than expected. Several issues have become apparent [8,95] and will have a substantial impact on the best way to follow up these initial GWAS.

Are common variants sufficient to discover the genetic bases of many complex traits?

Most common variants that have been found associated with disease through GWAS typically have very small effects on the variability of the trait and explain a rather small portion of the heritability [8]. These initial findings suggested that many GWAS may have not been sufficiently powered to discover associations with such small effects and therefore stimulated the creation of consortia to merge results from several GWAS [99–101,103,104] in order to reach sufficient statistical power to identify smaller and smaller genetic effects. Increasing sample size indeed provides the required power, but the clinical significance of these findings remains an open question.

It is expected that functional variants discovered through resequencing of regions implicated by tag SNPs will uncover larger genetic effects [95]. Another conjecture that is finding increasing support is that rare rather than common variants, or a combination of both, might account for the unexplained variability of complex traits [105]. This conjecture could reinvigorate interest in family based studies, which are more powerful to detect rare variants with high penetrance [50]. Because most of the rare variants are probably unknown and the SNP arrays that are available are mainly designed to capture common variants catalogued through the HapMap project, only further SNP discovery through fine mapping and deep sequencing will unveil the truth [106].

Recently, several GWAS have begun to consider genomic copy number variations (CNVs) in addition to SNPs as possible targets for association with a phenotype [107,108]. CNVs are defined as inherited duplications and deletions of kilo- to mega-base lengths of DNA and they have been shown to be present in various numbers in all individuals. CNVs have been detected in locations covering $\sim$12% of the genome [109,110]. On a nucleotide by nucleotide basis, CNVs have therefore a higher polymorphic yield than the set of SNPs. Recent technological advances in both the hardware and software required to detect and analyze CNVs have begun to make the consideration of CNVs by a GWAS significantly more mainstream [107,111,112]. First round studies have shown CNVs to be associated with various phenotypes and disease states, including glomerulonephritis [108,113], sub-arachnoid hemorrhage [108], BMI [99], and even cultural dietary preferences [114]. These associations, if causative, may be due to the effects of CNVs on gene dosing, or due to their possible disruption or alteration of transcription factor binding sites, micro-RNAs, or local chromatin architecture. Additionally, several authors have commented on the importance of identifying areas of CNV for proper genotyping of a SNP in the context of a GWAS [115]. As the technology and theory surrounding CNVs continue to improve, and as higher-density, more reliable maps of the frequencies of CNVs in various populations become available, CNVs may take an increasingly prominent role alongside SNPs as targets of a GWAS.

Can the results of GWAS be translated into personalized medicine?

Many complex traits should be predictable once the genes that modulate their course are known and one of the promises of GWAS is to provide the decoder of these complex genetic diseases that, when coupled with information about environmental exposure, can be used to compute individual risk for a disease and to suggest appropriate prophylactic treatments or lifestyle changes. We share with multiple other investigators the belief that data from GWAS are largely unexploited [46,95], and may contain the information to decipher the genetics of complex diseases. However, building models that predict the outcome of individual patients based on their genetic profile challenges investigators with a plethora of computational issues.

The majority of papers reporting findings from GWAS list SNPs that are significantly associated with a trait, when analyzed one at a time, and do not attempt to integrate them into a risk prediction model [5,6,26,27,116]. Exceptions are the few efforts to develop risk scores that are based on simple linear functions of the genetic profiles [100]. Although several investigators see this initial selection as the first step to prognostic modeling [102], the reductionist approach has two limitations: it may identify too many associations because of dependencies between genetic variants that are the results of evolution [117] and it is unable to discover associations that are due to interdependent multiple genotypes [118]. For example, Hoh and Ott [118] describe a situation in which the simultaneous presence of three genotypes at different loci leads to a disease. The three genotypes have the same marginal penetrance and would not be found associated with the disease in a one-at-a-time search but only when examined simultaneously. Multivariate statistical models, such as linear or logistic regression, can circumvent these limitations by examining the overall dependency structure between genotypes, phenotype, environmental, and clinical variables. However, traditional regression models require large sample sizes and/or experimental and control samples that are sufficiently different in terms of the phenotype of interest to confer significant power [46]. The amount of data produced by the new genotyping technology requires novel techniques that go beyond "traditional statistical thinking" in order to accommodate the potential complexity of genetic models.

Several machine learning methods used in data mining may be more appropriate to discover and describe the genetic base of complex traits [119]. Classification and regression trees (CART) [120], random forests [121], and Bayesian networks [122] have been proposed for modeling complex gene-environment interactions when the phenotype is a well defined variable [123,124]. CART is a multivariate statistical technique that creates a set of if-then rules linking combinations of genotypes and environmental

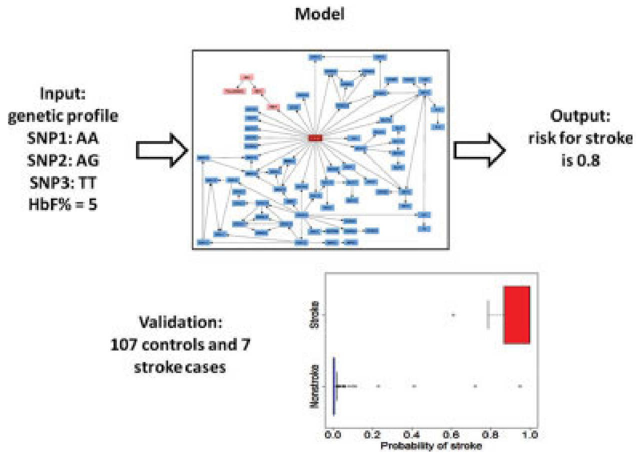

Figure 6. Illustration of the use of a Bayesian network model for risk prediction. The model in the middle is the Bayesian network that describes the interrelationships between genetic variants, clinical variables and stroke in sickle cell anemia. Given the genetic profile of a patient with sickle cell anemia, the network can be used to compute the risk for stroke. We applied this model to compute the risk for stroke in 114 subjects and the boxplots show in blue the predicted risk of the 107 disease free subjects, and in red the predicted risk of the seven stroke cases. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

exposures to the phenotypes. The if-then rules are created with a recursive procedure that groups data into sets to maximize the overall information. Random forests are an expansion of CART that have shown particular promise for the analysis of genotype–phenotype correlations, taking into account gene–gene interactions. The intuition is to use permutation methods and bootstrapping techniques to create thousands of CART models. From the analysis of these models, one can produce an importance measure for each SNP that takes into account interactions with other SNPs that affect the phenotype [125]. It has been shown that when unknown interactions among SNPs and genetic heterogeneity exist, random forest analysis can be substantially more powerful than standard univariate screening methods [126].

Bayesian networks represent the association between many variables using conditional probability distributions and rely on Bayes' theorem to show how changes in one or more variables in the networks affect other variables [127]. In this way, they can be used prognostically to compute the probability of the outcome (say a specific fetal hemoglobin range) of an individual given his genetic profile. They can also be used diagnostically, to discover the genetic profiles that maximize the probability of a particular outcome, and therefore be used to study how patterns of behavior can interact with the genetic profiles to shape the phenotype [128]. In this way, they appear to be able to simultaneously cast light on novel biological findings and be a prognostic tool for personalized medicine [46]. The Bayesian network that we developed for the genetic dissection of stroke in sickle cell anemia (see Fig. 6) offers an example of the power of these models. The network captures the interaction between 31 SNPs in 12 genes that, together with fetal hemoglobin, modulate the risk for stroke. We showed the prognostic accuracy of this model by predicting with 98.2% accuracy the occurrence of stroke in 114 subjects not included in the primary analysis and showed that this approach outperformed statistical models based on logistic regression [122]. We used the same approach to develop a model of severity of sickle cell disease defined by survival. Phenotypic heterogeneity is a well known characteristic of this disease. Patients have different rates of complications, such as pulmonary hypertension, painful episodes, acute chest syndrome, and osteonecrosis, as well as variations in levels of laboratory variables. To integrate individual disease variables into a global measure of severity, we developed a network that describes the complex associations of 25 clinical and laboratory variables, deriving a score that we used to define disease severity (0, least severe to 1, most severe) as the risk of death within 5 years [129]. This initial network was validated in 140 patients whose disease severity was assessed by expert clinicians and 210 adults where severity was also assessed by the echocardiographic diagnosis of pulmonary hypertension and death. We implemented a version of this calculator of disease severity on the internet using Java (http://www.bu.edu/sicklecell/downloads/Projects/). Although Bayesian networks are more difficult to generate and more challenging to communicate than traditional regression models, they are slowly becoming more accepted in the genetic community [130,131].

Is pleiotropy the explanation?

One of the interesting findings from the first series of results of GWAS is that several genes, and often the same SNPs, are associated with multiple traits. Some make immediate, intuitive sense, such as the associations of the SNP rs1051730 in *CHRNA3* with both lung cancer [132] and nicotine dependence [133], while others are less obvious. For example rs10484554 in *HLA-C* was found associated with AIDS nonprogression [134] and susceptibility to early onset psoriasis [135]. The SNP rs2476601 in *PTPN22* was found associated with Crohn's disease [30], rheumatoid arthritis [136], and type 1 diabetes [4,32]. LDL cholesterol, triglycerides, and Alzheimer's disease are another set of traits that were found associated with rs4420638 in *APOE* in independent studies [101,137,138]. These associations suggest a pleiotropic effect of the genes involved that may affect many different traits. An alternative explanation is that the different phenotypes associated with the same gene may be the endpoints of disease progression sharing a common mechanism that is regulated by the gene. For example, we observed that several well known aging genes, including *Klotho*, were associated with vasoocclusive complications of sickle cell anemia [139,140]. Based on this observation, we searched for common genes that simultaneously affect the overall severity of sickle cell anemia and aging in GWAS of sickle cell anemia phenotypes and exceptional longevity. Preliminary findings show that both traits are associated with the same several SNPs and these association would not be detected by studying the two traits alone [141]. It may be valuable to study different phenotypes not in a vacuum but transversely to increase the chance of identifying biologically important genes [142]. Tools developed to describe phenome-genome networks may be valuable to link phenotypes through common physiological mechanisms [143,144].

Conclusions

Four years after the publication of the first article reporting a positive finding from a GWAS, we have learned that GWAS can be effective to discover novel genetic modifiers of common diseases. However, much work remains to be done to fully extract information from the massive amount of data produced by these studies. Integration of genetic with other gene product data, follow up of preliminary results through informative experiments, and deep modeling of data can help translate GWAS and other genomic data into better understanding of the mechanism leading to disease and tools for disease prevention.

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265:2037–2048.
2. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003;33Suppl:228–237.
3. Wang Y, O'Connell JR, McArdle PF, et al. From the cover: Whole-genome association study identifies *STK39* as a hypertension susceptibility gene. Proc Natl Acad Sci USA 2009;106:226–231.
4. Hakonarson H, Grant SF, Bradfield JP, et al. A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. Nature 2007;448:591–594.
5. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881–885.
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678.
7. Waring SC, Rosenberg RN. Genome-wide association studies in Alzheimer disease. Arch Neurol 2008;65:329–334.
8. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881–888.
9. Lander ES. The new genomics: Global views of biology. Science 1996;274: 536–539.
10. International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299–1320.
11. Kennedy GC, Matsuzaki H, Dong S, et al. Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233–1237.
12. Gunderson KL, Steemers FJ, Lee G, et al. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005;37:549–554.
13. Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. Genome Biol 2001;2: COMMENT2007.
14. Collins FS, Guyer MS, Chakravarti A. Variations on a theme: Cataloging human DNA sequence variation. Science 1997;278:1580–1581.
15. Lander ES. Array of hope. Nat Genet 1999;21:3–4.
16. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
17. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225–2229.
18. Sebastiani P, Lazarus R, Weiss ST, et al. Minimal haplotype tagging. Proc Natl Acad Sci 2003;100:9900–9905.
19. The International Hapmap Consortium. The international HapMap project. Nature 2003;426:798–796.
20. Nishida N, Koike A, Tajima A, et al. Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. BMC Genomics 2008;9:431.
21. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–389.
22. Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. Nat Genet 2007;39:977–983.
23. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39:865–869.
24. Gudmundsson J, Sulem P, Manolescu A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007;39:631–637.
25. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39:870–874.
26. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39:645–649.
27. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–1093.
28. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007;39:596–604.
29. Parkes M, Barrett JC, Prescott NJ, et al. Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 2007;39:830–832.
30. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008; 40:955–962.
31. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443–453.
32. Todd JA, Walker NM, Cooper JD, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007;39:857–864.
33. Menzel S, Garner C, Gut I, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nat Genet 2007;39:1197–1199.
34. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Med Genet 2009;10:6.
35. Yang Q, Kathiresan S, Lin J-P, et al. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. BMC Med Genet 2007;8:S12.
36. Ouwehand WH. Platelet genomics and the risk of atherothrombosis. J Thromb Haemost 2007;5:188–195.
37. Huang RS, Duan S, Bleibel WK, et al. A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity. Proc Natl Acad Sci USA 2007;104:9758–9763.
38. Huang RS, Duan S, Kistner EO, et al. Genetic variants contributing to daunorubicin-induced cytotoxicity. Cancer Res 2008;68:3161–3168.
39. Di Bernardo MC, Crowther-Swanepoel D, Broderick P, et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat Genet 2008;40:1204–1210.
40. Sarasquete ME, Garcia-Sanz R, Marin L, et al. Bisphosphonate-related osteonecrosis of the jaw is associated with polymorphisms of the cytochrome P450 *CYP2C8* in multiple myeloma: A genome-wide single nucleotide polymorphism analysis. Blood 2008;112:2709–2712.
41. Cooper GM, Johnson JA, Langaee TY, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 2008;112:1022–1027.
42. Uda M, Galanello R, Sanna S, et al. Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. Proc Natl Acad Sci USA 2008;105:1620–1625.
43. Sedgewick AE, Timofeev N, Sebastiani P, et al. *BCL11A* is a major HbF quantitative trait locus in three different populations with beta-hemoglobinopathies. Blood Cells Mol Dis 2008;41:255–258.
44. Hindorff LA, Junkins HA, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at www.genome.gov/2652384. Accessed 1 29 2009.
45. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. Nat Protoc 2007;2:2492–2501.
46. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–369.
47. Pearson TA, Manolio TA. How to interpret a genome-wide association study. JAMA 2008;299:1335–1344.
48. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001;2:91–99.
49. Sorensen HT, Gillman MW. Matching in case-control studies. BMJ 1995;310: 329–330.
50. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7:385–394.
51. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.
52. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–909.
53. Perls T, Kohler IV, Andersen S, et al. Survival of parents and siblings of supercentenarians. J Gerontol A Biol Sci Med Sci 2007;62:1028–1034.
54. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100–1104.
55. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet 2006;38:659–662.
56. Ardlie K, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 2002;3:299–309.
57. Anderson CA, Pettersson FH, Barrett JC, et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet 2008;83:112–119.
58. Zhao Z, Timofeev N, Hartley SW, et al. Imputation of missing genotypes: An empirical evaluation of IMPUTE. BMC Genet 2008;9:85.
59. Bhangale TR, Rieder MJ, Nickerson DA. Estimating coverage and power for genetic association studies using near-complete variation data. Nat Genet 2008;40:841–843.
60. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10:241–251.
61. Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. Nat Rev Genet 2006;7:632–644.
62. Bonin A, Bellemain E, Bronken Eidesen P, et al. How to track and assess genotyping errors in population genetics studies. Mol Ecol 2004;13:3261–3273.
63. Milligan BG. Maximum-likelihood estimation of relatedness. Genetics 2003; 163:1153–1167.
64. Sun L, Wilder K, McPeek MS. Enhanced pedigree error detection. Hum Hered 2002;54:99–110.
65. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–791.
66. Kvam PH, Vidakovic B. Nonparametric Statistics with Applications to Science and Enjgineering. New York: Wiley; 2007.
67. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet 2004;5:251–261.
68. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 2003;73: 1162–1169.
69. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–989.

70. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39: 906–913.
71. Soria JM, Almasy L, Souto JC, et al. The *F7* gene and clotting factor VII levels: Dissection of a human quantitative trait locus. Hum Biol 2005;77:561–575.
72. Bernardo JM, Smith AFM. Bayesian Theory. New York, NY: Wiley; 1994.
73. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516.
74. Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. Am J Hum Genet 1998;62: 450–458.
75. Sebastiani P, Abad MM, Alpargu G, Ramoni MF. Robust transmission/disequilibrium test for incomplete family genotypes. Genetics 2004;168:2329–2337.
76. Epstein MP, Veal CD, Trembath RC, et al. Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 2005;76:592–608.
77. Allen-Brady K, Wong J, Camp NJ. PedGenie: An analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 2006;7:209.
78. Curtin K, Wong J, Allen-Brady K, Camp NJ. PedGenie: Meta genetic association testing in mixed family and case-control designs. BMC Bioinformatics 2007;8:448.
79. Lange C, Blacker D, Laird NM. Family-based association tests for survival and times-to-onset analysis. Stat Med 2004;23:179–189.
80. Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000;50:211–223.
81. Havill LM, Dyer TD, Richardson DK, et al. The quantitative trait linkage disequilibrium test: A more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification. BMC Genet 2005;6Suppl 1:S91.
82. Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000;66:279–292.
83. Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. Am J Hum Genet 2007;81:913–926.
84. Hancock DB, Martin ER, Li YJ, Scott WK. Methods for interaction analyses using family-based case-control data: Conditional logistic regression versus generalized estimating equations. Genet Epidemiol 2007;31:883–893.
85. Cupples LA. Family study designs in the age of genome-wide association studies: Experience from the Framingham Heart Study. Curr Opin Lipidol 2008;19:144–150.
86. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: Theoretical and practical concerns. Nat Rev Genet 2005;6:109–118.
87. Timofeev N, Sebastiani P, Hartley SW, et al. Fetal hemoglobin in sickle cell anemia: A genome-wide association study of the response to hydroxyurea. Blood (ASH Annual Meeting Abstracts) 2008;112:2471.
88. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98: 5116–5121.
89. Benjamini Y, Drai D, Elmer G, et al. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125:279–284.
90. Pawitan Y, Calza S, Ploner A. Estimation of false discovery proportion under general dependence. Bioinformatics 2006;22:3025–3031.
91. Yang H, Churchill G. Estimating p-values in small microarray experiments. Bioinformatics 2007;23:38–43.
92. Scheid S, Spang R. A stochastic downhill search algorithm for estimating the local false discovery rate. IEEE/ACM Trans Comput Biol Bioinform 2004; 1:98–108.
93. Gopalan R, Berry DA. Bayesian multiple comparisons using Dirichlet process priors. J Am Stat Assoc 1998;93:1130–1139.
94. Scott JG, Berger JO. An exploration of aspects of Bayesian multiple testing. J Stat Plann Infer 2005;136:2144–2162.
95. Donnelly P. Progress and challenges in genome-wide association studies in humans. Nature 2008;456:728–731.
96. Sebastiani P, Zhao Z, Abad-Grau MM, et al. A hierarchical and modular approach to the discovery of robust associations in genome-wide association studies from pooled DNA samples. BMC Genet 2008;9:6.
97. Hutz JE, Kraja AT, McLeod HL, Province MA. CANDID: A flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol 2008;32:779–790.
98. Chanock SJ, Manolio T, Boehnke M, et al. Replicating genotype-phenotype associations. Nature 2007;447:655–660.
99. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 2009;41:25–34.
100. Aulchenko YS, Ripatti S, Lindqvist I, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 2009;41:47–55.
101. Kathiresan S, Willer CJ, Peloso GM, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 2009;41:56–65.
102. McCarthy MI, Hirschhorn JN. Genome-wide association studies: Potential next steps on a genetic journey. Hum Mol Genet 2008;17:R156–R165.
103. Kathiresan S, Voight BF, Purcell S, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 2009;41:334–341.
104. Thorleifsson G, Walters GB, Gudbjartsson DF, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 2009;41:18–24.
105. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008;40:695–701.
106. O'Donnell CJ, Elosua R. [Cardiovascular risk factors. Insights from Framingham Heart Study]. Rev Esp Cardiol 2008;61:299–310.
107. McCarroll SA. Extending genome-wide association studies to copy-number variation. HumMolGenet 2008;17:R135–R142.
108. Bae JS, Cheong HS, Kim J-O, et al. Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. Biochem Biophys Res Commun 2008;373:593–596.
109. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444–454.
110. Perry GH, Ben-Dor A, Tsalenko A, et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 2008;82:685–695.
111. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008; 40:1166–1174.
112. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253–1260.
113. Aitman TJ, Dong R, Vyse TJ, et al. Copy number polymorphism in *FCGR3* predisposes to glomerulonephritis in rats and humans. Nature 2006;439: 851–855.
114. Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007;39:1256–1260.
115. Ionita-Laza I, Rogers AJ, Lange C, et al. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 2009;93:22–26.
116. Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 2007;8:657–662.
117. Chakravarti A. Population genetics–making sense out of sequence. Nat Genet 1999;21:56–60.
118. Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 2003;4:701–709.
119. Hand DJ, Mannila H, Smyth P. Principles of Data Mining. Cambridge, MA: MIT Press; 2001.
120. Quinlan JR. Improved use of continuous attributes in C4.5. J Art Intell Res 1996;4:77–90.
121. Breiman L. Random forests. Mach Learn 2001;45:5–32.
122. Sebastiani P, Ramoni MF, Nolan V, et al. Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nat Genet 2005;37:435–440.
123. Cook NR, Zee RY, Ridker PM. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 2004;23:1439–1453.
124. Nolan V, Wilcox M, Sebastiani P, et al. Gene-Gene interactions and the pathophysiology of sickle cell disease: Modeling the effects of SNPs on sickle cell-associated vasoocclusive events using classification and regression trees and stochastic gradient boosting. Blood 2005;Supplement: ASH 2005:3183.
125. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: Exploiting interactions using random forests. BMC Genet 2004;5:32.
126. Bureau A, Dupuis J, Falls K, et al. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 2005;28:171–182.
127. Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ. Probabilistic Networks and Expert Systems. New York: Springer Verlag; 1999.
128. Perls TT, Sebastiani P. Genetics of exceptional longevity. In: Guarente LP, Partridge L, Wallace DC, editors. Molecular Biology of Aging. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 2008.
129. Sebastiani P, Nolan VG, Baldwin CT, et al. A network model to predict the risk of death in sickle cell disease. Blood 2007;110:2727–2735.
130. Rodin AS, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma *APOE* levels). Bioinformatics 2005;21:3273–3278.
131. Ramoni RB, Himes BE, Sale MM, et al. Predictive genomics of cardioembolic stroke. Stroke 2009;40(3 Suppl):S67–S70.
132. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452:633–637.
133. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452: 638–642.
134. Limou S, Le Clerc S, Coulonges C, et al. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by *HLA* genes (ANRS Genomewide Association Study 02). J Infect Dis 2009;199:419–426.
135. Liu Y, Helms C, Liao W, et al. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet 2008;4: e1000041.

136. Plenge RM, Seielstad M, Padyukov L, et al. *TRAF1-C5* as a risk locus for rheumatoid arthritis–a genomewide study. N Engl J Med 2007;357:1199–1209.
137. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316: 1331–1336.
138. Webster JA, Myers AJ, Pearson JV, et al. *SORL1* as an Alzheimer's disease predisposition gene?. Neurodegener Dis 2008;5:60–64.
139. Baldwin C, Nolan VG, Wyszynski DF, et al. Association of *KLOTHO*, bone morphogenic protein 6, and annexin A2 polymorphisms with sickle cell osteonecrosis. Blood 2005;106:372–375.
140. Steinberg MH, Adewoye AH. Modifier genes and sickle cell anemia. Curr Opin Hematol 2006;13:131–136.
141. Sebastiani P, Timofeev N, Hartley SW, et al. Genome-wide association studies suggest shared polymorphisms are associated with severity of sickle cell anemia and exceptional longevity. Blood (ASH Annual Meeting Abstracts) 2008;112:1446.
142. Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. Mol Syst Biol 2007;3:124.
143. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. Nat Biotechnol 2006;24:55–62.
144. Goh KI, Cusick ME, Valle D, et al. The human disease network. Proc Natl Acad Sci USA 2007;104:8685–8690.