

Genetics, statistics and human disease: analytical retooling for complexity

Tricia A. Thornton-Wells^{1,2,3}, Jason H. Moore^{3,4} and Jonathan L. Haines^{3,4}

¹Neuroscience Graduate Program, Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN 37240, USA

²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37240, USA

³Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37240, USA

⁴Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37240, USA

Molecular biologists and geneticists alike now acknowledge that most common human diseases with a genetic component are likely to have complex etiologies. Yet despite this belief, many statistical geneticists continue applying, in slightly new and different ways, methodologies that were developed to dissect much simpler etiologies. In this article, we characterize, with examples, the various factors that can complicate genetic analysis and demonstrate their shared features and how they affect genetic analysis. We describe a variety of approaches that are currently available, revealing methodological gaps and suggesting new directions for method development. Finally, we propose a comprehensive two-step approach to analysis that systemically addresses the different genetic factors that are likely to underlie complex diseases.

'If the only tool you have is a hammer, you tend to see every problem as a nail.'

Abraham Maslow, American psychologist, founder of humanistic psychology.

'The difficulty lies, not in the new ideas, but in escaping the old ones.'

John Maynard Keynes, English economist.

Over the past few decades, most of the success in the field of statistical genetics has come from identifying genes with substantial (non-interactive) effects on the disease process. Most statistical tools enabling this success were developed for, and are primarily effective in, the analysis of simple, mendelian diseases such as Huntington disease, cystic fibrosis and early-onset Alzheimer's disease. Molecular biologists and geneticists alike now acknowledge that the most common human diseases with a genetic component are likely to have complex etiologies. However, despite this belief, statistical geneticists continue using primarily traditional methodologies to attack this complex problem. We then decry the failure of our studies to identify and replicate significant genetic effects, all the while failing to see the forest for the trees. Advances in statistical and computational genetic methodology simply have not kept pace with the advance of available sources of data. There have been a few attempts to address complexity directly, including the development of nonparametric

tools, but these have generally limited application. One example is the transmission disequilibrium test that led to the discovery of the insulin receptor gene as a risk factor for diabetes [1].

Going forward, statistical geneticists must not only acknowledge but also directly confront the numerous complicating factors that can be involved in complex genetic diseases, which present significant challenges for traditional statistical methods. Only a fraction of the human genetics literature specifically reports on investigations of such complexity. It is, perhaps, daunting to consider multiple complicating factors, such as allelic heterogeneity, locus heterogeneity, phenocopy, phenotypic variability, trait heterogeneity and gene–gene or gene–environment interactions (for definitions, see [Tables 1–3](#)). However, these must be addressed if we are to have any chance in understanding the genetic legacy of disease left to us by our forebears.

Categorization and analytical approaches

Each of the factors presented in [Tables 1–3](#) complicate statistical analysis in one of two ways – either by creating heterogeneous, or competing, disease models ([Tables 1 and 2](#)), or by creating a multifactorial, interacting disease model ([Table 3](#)). The challenge for modeling the relationship between genetic and environmental risk factors (independent variables) and disease endpoints (dependent variables) is different for these two categories. Of course, what exacerbates the complexity is that none of these competing and interacting models are mutually exclusive. Various combinations of (genetic and/or trait) heterogeneity and interactions might be important in any given disease of interest. Thus, to dissect these factors, we must assemble a toolbox of both tried-and-true and newly constructed genetic analysis methodologies, which together can be used to discover the true underlying etiologies of complex traits.

Many complicating factors can be addressed proactively by a well-considered study design. This is perhaps one of the best investments researchers can make to maximize their ability to discover complex genetic disease models. Because the causally complex relationship between the genotype and phenotype is the object of genetic studies, it is important to collect accurate and abundant phenotypic data. In the absence of phenotypic

Corresponding author: Jonathan L. Haines (jonathan@chgr.mc.vanderbilt.edu). Available online 8 October 2004

Table 1. Heterogeneity-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples^a

	Allelic heterogeneity	Locus heterogeneity	Phenocopy
Definition	When two or more alleles of a single locus are independently associated with the same trait	When two or more DNA variations in distinct genetic loci are independently associated with the same trait	The presence of a disease phenotype that has a non-genetic (random or environmental) basis
Diagram			
Example A	Retinitis pigmentosa (RP, OMIM# 268000) – in the <i>RHO</i> gene (OMIM# 180380), which accounts for 30–40% of autosomal dominant RP, > 100 distinct mutations have been found [51] (http://www.sph.uth.tmc.edu/RetNet)	Retinitis pigmentosa (RP, OMIM# 268000) – genetic variations in at least 15 genes have been associated with RP under an autosomal recessive model; even more have been associated with RP under autosomal dominant and X-linked disease models [51] (http://www.sph.uth.tmc.edu/RetNet)	Parkinson disease (PD, OMIM# 168600) – individuals taking the illicit drug meperidien are sometimes exposed to its by-product MPTP, which causes the destruction of dopaminergic neurons [55,56] and produces the PD phenotype
Example B	Cystic Fibrosis (CF) – > 1000 mutations in the <i>CFTR</i> gene (OMIM# 602421) have been associated with CF [52]	Tuberous sclerosis (TS, OMIM# 191100) – out of families informative for linkage analysis, half have mutations in the <i>TSC1</i> gene (located at 9q34), and the other half have mutations in the <i>TSC2</i> gene (located at 16p13) [52–54]	Epilepsy (OMIM# 600669) – traumatic brain injury can result in posttraumatic epileptic seizures occurring within 24 h or up to several years after the injury [57]

^aAbbreviation: OMIM, online mendelian inheritance in man (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&tool=toolbar>).

data, there is not even the option of looking for a mapping between genotype and potential clinical subtypes, which could help identify a case of genetic heterogeneity. Established guidelines or protocols concerning data collection should be followed and such data should be made available to others in an accessible format, so as to facilitate future meta-analysis. Information regarding the exposure to potential environmental risk factors should be collected whenever logistically and economically feasible. Even with the best study design with regard to data collection, an ill-advised or incomplete analysis of the data

can still yield disappointing, if not incorrect, results. Thus, we advocate a comprehensive approach to account for both the heterogeneity and the interaction models of disease.

Heterogeneity

For this category of factors, there are multiple independent (predictor) variables or else multiple dependent (outcome) variables that complicate the analysis by creating a heterogeneous model landscape. In the case of allelic or locus heterogeneity or phenocopy, multiple predictor variables (e.g. multiple alleles, multiple loci

Table 2. Heterogeneity-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples^a

	Trait heterogeneity	Phenotypic variability
Definition	When a trait, or disease, has been defined with insufficient specificity such that it is actually two or more distinct underlying traits	Variation in the degree, severity or age of onset of symptoms exhibited by persons who actually have the same trait or disease process
Diagram		
Example A	Autosomal dominant cerebellar ataxia (ADCA, OMIM# 164500) – originally described as a single disease, three different clinical subtypes have been defined based on variable associated symptoms, [58,59] and different genetic loci have been associated with the different subtypes [60]	Holoprosencephaly (HPE, OMIM# 236100) – craniofacial abnormalities associated with HPE can range from a single middle incisor to cyclopia
Example B	Autism (OMIM# 209850) – parents and other relatives of autistic individuals often exhibit one or two, but not all three, of the requisite autistic symptomatology, suggesting autism might be the co-occurrence of three distinct traits; [61] using subset analysis, some success has been achieved identifying genes associated with one of the three symptomatology but not as strongly with the broader autistic phenotype [62,63]	Tuberous sclerosis (TS, OMIM# 191100) – the severity of such TS symptoms as mental retardation, kidney disease and facial angiofibroma differ across affected individuals [64]

^aAbbreviation: OMIM, online mendelian inheritance in man (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&tool=toolbar>).

Table 3. Interaction-related factors complicating analysis of complex genetic disease: definitions, diagrams and examples^a

	Gene-gene interaction	Gene-environment interaction
Definition	When two or more DNA variations interact either directly (DNA-DNA or DNA-mRNA interactions), to change transcription or translation levels, or indirectly by way of their protein products, to alter disease risk separate from their independent effects	When a DNA variation interacts with an environmental factor, such that their combined effect is distinct from their independent effects
Diagram		
Example A	Hirschsprung disease (OMIM# 142623) – variants in the RET (OMIM# 164761) and EDNRB (OMIM# 131244) genes have been shown to interact synergistically, such that they increase disease risk far beyond the combined risk of the independent variants [65]	Bovine spongiform encephalopathy (BSE) – all human cases of BSE, which is commonly known as ‘mad cow disease’ and is transmitted through consumption of contaminated beef, were in individuals who were homozygous for the met129 polymorphism in the PRNP gene (OMIM# 176640). In a cluster of unreported cases of BSE in which individuals had been exposed to contaminated brain electrodes, all but one individual was 129met/met; the remaining person was heterozygous for the polymorphism and had a more protracted course than the others [73,68]
Example B	Creutzfeldt-Jacob Disease (CJD, OMIM# 123400) and Fatal Familial Insomnia (OMIM# 176640.0010) – the Met129Val polymorphism and Asp178Asn mutation in the PRNP gene (OMIM# 176640) interact, such that when the val129 polymorphism is on the same chromosome as the asn178, the phenotype is fatal familial insomnia [66–72]	Unipolar depression (OMIM# 608516) – individuals with one or two copies of the short allele of the serotonin transporter (5-HT T) promoter polymorphism are up to twice as likely to develop depressive symptoms, diagnosable depression and suicidality after stressful life events, than individuals homozygous for the long allele, suggesting a gene-environment interaction [74]

^aAbbreviation: OMIM, online mendelian inheritance in man (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM&tool=toolbar>).

and/or environmental risk factors) are present, some of which might be unmeasured or unobserved and, therefore, unavailable for inclusion in the disease model. In the case of trait heterogeneity or phenotypic variability, multiple outcome variables are present, which cannot or have not been distinguished based on the available phenotypic information.

Perhaps the most straightforward of the methods for addressing heterogeneity is sample stratification (Table 4). This method subdivides subjects based on any number of genetic, demographic, clinical or environmental factors to create more homogeneous subsets of the data. The premise of this method is that there are two or more underlying disease models, which are conditional on the factor on which the data are being stratified. For example, one genetic model might be associated with disease in the absence of a specific environmental risk factor; however, when that environmental factor is present, a different set of genetic factors are involved. Using different levels of the stratifying factor (e.g. different degrees of environmental exposure), one could perform further analyses, such as logistic regression (discussed in the following section). The main limitation of sample stratification is a reduction in sample size within each stratum and thus a reduction in power.

Some statistical methods that test the hypothesis of locus heterogeneity include the M test [2], the β test [3] and the ADMIXTURE (see Glossary) test (Table 4) [4,5]. Each of these methods is solely applicable to family-based data on which LINKAGE ANALYSIS is performed. The M test uses *a priori* stratification of subjects based on discrete (or discretized) covariates, such as gender, ethnicity or

clinical subtype, and tests for a difference in RECOMBINATION FRACTIONS across the different subsets of families. The β test is a similar but slightly more powerful statistical test than the M test, owing to a difference in their null distributions used to determine statistical significance. The admixture test does not require *a priori* stratification but instead estimates (using maximum likelihood) the degree of admixture present in the sample from two-point or multi-point lod scores between marker and disease loci. It then uses these estimates to evaluate the relative probabilities of linkage with and without heterogeneity. Thus, the M and β tests evaluate a more specific hypothesis, and as a result, have more power than the admixture test. The admixture test also lacks sensitivity and can only account for, not resolve, the underlying heterogeneity.

A more recently developed method to address heterogeneity is the ordered subset analysis (OSA; Table 4) [6,7].

Glossary

Recombination fraction: the probability that a parent will produce a recombinant offspring; the percentage of offspring in a family or dataset who are recombinants; a statistical measure of the distance between two loci.

Admixture: the mixing of two or more subpopulations, having differing characteristics. If the subpopulations have different allele or genotype frequencies and have different disease frequencies it can result in spurious associations.

Lod score: the \log_{10} of the odds in favor of linkage, or the \log_{10} of the ratio of the likelihood of a specific recombination fraction <0.50 to the likelihood of a recombination fraction of 0.50 (no linkage).

Linkage analysis: the analysis of family-based genotypic data to detect linkage of a disease locus with one or more loci within a family

Association analysis: the analysis of population-based (case-control) or family-based genotypic data to detect the association of disease with a specific allele across families (or cases).

Table 4. Summary of analytical approaches to heterogeneity

	Abbreviated description	Example
Sample stratification	Manual sorting by covariate(s)	Stratification by age-of-onset led to the confirmation of APP as a risk factor for early-onset Alzheimer's disease [75]
M test	Manual sorting by covariate(s) + testing difference in recombination fractions across families	Locus heterogeneity across French families with Usher Syndrome was confirmed by the M-test [76]
Beta test	Very similar but slightly more powerful statistical test than the M test	Locus heterogeneity between the F9 and fra(x) loci was confirmed with a near maximal result on the Beta test [3,77]
Admixture test	Estimates population admixture present in the sample and evaluates probability of heterogeneity	Evidence for linkage was significantly increased in the 9q32–34 region of interest for tuberous sclerosis when significant heterogeneity was shown with admixture test and only linked families (with 80% posterior probability) were examined [78]
Ordered subset analysis (OSA)	Ordering of families by covariate(s) and calculation of maximum cumulative lod score	Evidence for linkage was significantly increased in two regions of interest for macular degeneration risk genes (14q13 and 6q14) using intraocular pressure and body mass index as covariates in OSA [79]
Cluster analysis	Clustering of individuals to produce subgroups with high intraclass similarity and low interclass similarity	Clustering by pedigree-specific genomic markers led to identification of linkage region of interest not found after sample stratification by self-reported ethnicity [80]
Latent class analysis and factor analysis	Similar to cluster analysis, except that 'latent' or underlying variables are derived from relationships among known covariates	Factor analysis derived an 'insistence on sameness (IS)' factor from the autism diagnostic interview-revised, which was successfully used as a covariate in OSA to narrow a region of interest for autism susceptibility gene [81]

In OSA, a continuous or ordinal covariate, such as blood-lipid levels or disease age-of-onset, is used to rank order families, and then a cumulative LOD SCORE is iteratively calculated after each family is added (in order) to the sample until the cumulative lod score begins to decrease. Thus, those families included in the linkage analysis all provide support for linkage, and the subset of chosen families is more homogeneous with respect to the covariate and, therefore, hopefully, more genetically homogeneous than the whole dataset.

Other methods aimed at producing more homogeneous subsets of the data include cluster analysis, latent class analysis and factor analysis (Table 4). Unlike the aforementioned statistical tests for heterogeneity that only incorporate linkage analysis, the following methods can also be applied to case-control datasets because they are not tied to any particular statistical analysis of the subsets. There are hundreds of different cluster analysis methods, which operate based on different heuristics and fitness metrics, making them appropriate for particular types of data (continuous versus discrete, low- versus high-dimensional and so on.) They all attempt to produce clusters with either high intraclass similarity or low interclass similarity and have varying degrees of success. Cluster analysis has been widely used for analyzing DNA and protein microarray data [8] and to find more homogeneous subgroups based on genetic background [9].

Latent class analysis and factor analysis have a goal similar to cluster analysis but instead of directly clustering or classifying data based on known covariates, such as the scores of different items on a psychological or physical functioning test, these two methods try to derive 'latent' or underlying variables, such as summary scores of various test items, from relationships among the known covariates. These latent variables are then used to classify or

stratify the data. Latent class analysis has been applied to phenotypic data for several diseases, including attention deficit hyperactivity disorder (ADHD) [10], Alzheimer's disease [11], autism [12] and schizophrenia [13].

It should be noted that all of the methods discussed previously, with the exception of the admixture method, depend on covariate data, whether these be known genetic risk factors, demographic data, phenotypic data or endo-phenotypes. Not only must such information be available but also these covariates must actually be relevant to, or be surrogates for, the existing heterogeneity. If the data are incomplete, the performance of many of these methods for dissecting heterogeneity suffers and attempts to correct this problem by imputing data can introduce spurious associations. In the absence of such relevant, complete data, we are left with seemingly few options of how to proceed when we suspect heterogeneity to have a role.

To overcome some of these problems it might be advantageous to adapt the same basic principles of the aforementioned methods to the more complex data. For instance, although clustering methods have been heavily utilized for microarray data, few studies have examined clustering genotypic data from association-based studies to identify multilocus patterns that characterize particular subsets of the data. Some clustering methodologies appropriate for such discrete data include hypergraph clustering [14], bayesian classification [15] and fuzzy k-modes clustering [16].

Interactions

Gene–gene and gene–environment interactions are two complex genetic factors (Table 3) that create a rugged model landscape for statistical analysis. There is clear and convincing evidence that gene–gene interactions, whether synergistic or antagonistic, are not only possible but also

are probably ubiquitous [17,18]. Similarly, gene–environment interactions are likely to be discovered if properly investigated. Thus, it is crucial that complex genetic datasets be properly interrogated for possible underlying interactions.

Analytically it can be difficult to distinguish between heterogeneity and interactions. Many of the methods that address heterogeneity might be equally applicable to uncovering interactions. For instance, the discovery of linkage to a particular locus in only one subset of data produced by sample stratification could be indicative of heterogeneity, or it could be indicative of an interaction between the locus and the covariate used to stratify the data. However, there is also an entirely different set of methods that are particularly well suited to discovering interactions (but not heterogeneity; Table 5).

One traditional approach still widely used today is regression. In particular, logistic regression is used when the outcome variable is discrete, for example, disease status (i.e. you either have the disease or you do not) (Table 5). Logistic regression enables direct modeling of the mathematical relationship of genetic and other risk factors to disease status. However, this ‘workhorse’ suffers from the curse of dimensionality, meaning that as the distribution of data across numerous combinations of factors becomes sparse, the parameter estimates become unreasonably biased, particularly when the ratio of independent variables to sample size exceeds ten to one [19–21]. Thus, when considering a combination of loci, one

or more of which have low minor allele frequencies, the number of individuals with certain multilocus genotype combinations will be so small (or perhaps equal to zero), that one cannot reasonably estimate, or generalize in a population, what is the disease risk for that combination of genotypes. Missing or incomplete data can also create or exacerbate the problem of sparse data. In addition, many standard approaches to implementing logistic regression, such as forward stepwise regression, require significant main effects to be modeled before including interaction effects between factors. This is a major methodological limitation for situations where each locus has relatively small main (non-interactive) effects but more substantial interactive effects because none of those interactive effects would ever be considered.

A more recently developed statistical method for evaluating gene–gene interactions is the S-sum statistic, which is designed to overcome the curse of dimensionality and the multiple-testing problems by reducing any number of independent variable statistics into one sum statistic and then using permutation testing to correct for an experiment-wise Type I error rate, which is the probability of concluding that there is an effect when one does not actually exist [22,23]. ‘Set association’ analysis is the authors’ term for the application of the S statistic to SNP marker data from candidate genes or regions (Table 5). This method selects the ‘best’ set of n number of single nucleotide polymorphisms (SNPs), whose S statistic is statistically significant, leading to the inference that the

Table 5. Summary of analytical approaches to interactions

	Description	Example
Logistic regression	Mathematical modeling of relationship of discrete genetic and other risk factors to disease status	Three two-locus interactions among three folate-related genes were found by logistic regression to increase risk of neural tube defects [82]
S-sum statistic (set association analysis)	Selects the ‘best’ set of SNPs, whose summary statistic is statistically significant	A significant nine-SNP interaction among 62 candidate genes was found to be associated with restenosis after angioplasty [22,24]
Linear regression	Mathematical modeling of relationship between continuous outcome variable(s) and genetic risk factors to disease status	The nonmodulating hypertension phenotype was found to be associated with an interaction among the angiotensinogen, angiotensin-converting enzyme and aldosterone synthase genes [83]
Multivariate adaptive regression splines (MARS)	Generalization of stepwise linear regression particularly suited for high-dimensional problems with many independent variables	An interaction between two inflammation-related genes – the P-selectin and interleukin-4 genes – was found by MARS to be associated with ischemic stroke [28]
Classification and regression trees (CART)	Iteratively subdivides data to build a hierarchical classification model	Evidence of linkage to cardiovascular disease traits was strengthened in behaviorally-distinct subgroups constructed by CART [84]
Bayesian belief network	Probabilistic reasoning system that builds a hierarchical model of interactions	A network of SNPs and microsatellites in candidate genes was found to predict cervical cancer with reasonable specificity [85]
Combinatorial-partitioning method (CPM)	Utilizes data reduction to investigate gene-gene interactions	CPM found evidence for non-additive effects of the ACE and PAI-1 genes, in addition to additive effects found by linear regression, in the prediction of plasma PAI-1 levels [34]
Restricted-partition method (RPM)	Modification to CPM, which heuristically restricts the search for high-level interactions	This new method was used to analyze ten quantitative measures and ten SNPs in candidate genes related to irinotecan metabolism but did not identify any significant associations [35]
Multifactor dimensionality reduction (MDR)	Utilizes data reduction to investigate gene-gene interactions	A gene-gene interaction between the UCP2 and PPAR-gamma gene was found by MDR to be associated with Type 2 diabetes mellitus [42]
Artificial neural networks	Utilizes pattern recognition to find models for disease risk with multiple gene-gene interactions	Analysis of Type 1 diabetes mellitus data reproduced previously reported results showing highest lod scores for the IDDM1 and IDDM2 loci [45]

entire set of SNPs might be interacting in some way to increase disease risk, or else that they are all contributing independently to disease risk. However, because the summed statistics are all single-marker statistics, set ASSOCIATION ANALYSIS does not look at any specific (non-additive) interactive effects among markers and would be likely to miss nonlinear or antagonistic types of gene–gene interactions. This method has successfully identified a set of seven SNPs, which together were associated with restenosis incidence ($P < 0.0001$) and explained $> 11\%$ of the overall variance [24]. In theory the S statistic can be used with any number of test statistics on discrete or continuous data, but its applications and limitations are still being evaluated [25].

When the outcome variable is continuous, as is the case for a quantitative trait locus (QTL), such as serum prolactin levels, linear regression can be used to model the relationship between risk factors and QTL status (Table 5). However, linear regression faces the same limitations logistic regression does regarding parameter estimation and modeling interactions. Cheverud and Routman [26] developed an alternative parameterization of gene–gene interactions based on its effects on genetic-variance components (additive, dominance and interaction); however, it is limited to evaluating only two loci at a time and all possible genotypes must be present in the sample.

Multivariate adaptive regression splines (MARS) [27,28] is a generalization of stepwise linear regression that is particularly suited for high-dimensional problems in which many independent variables might be modeled. MARS is also similar to classification and regression trees (CART) [28–31], which iteratively subdivide data to build a hierarchical classification model. A bayesian belief network (BBN) [32] is a probabilistic reasoning system that builds a topological (but necessarily hierarchical) model of interactions (joint probabilities) (Table 5). BBN, CART and MARS all suffer from the same problem of sequential conditioning that can plague many other regression-based methods, which makes it difficult to discover interactions (especially higher-order interactions) among predictor variables, depending on the strength of their individual (or lower-order interaction) effects. The binary nature of CART further limits its ability to model any additive interaction. Still, the most troubling limitation that plagues all these methods is their inability to model, much less discover, nonlinear interactions.

Two types of computational methods – data reduction and pattern recognition – that come from the computer science field offer the potential for uncovering such nonlinear interactions, with increased tolerance for missing or incomplete data (Table 5). Nelson *et al.* [33] developed a combinatorial-partitioning method (CPM) that utilizes data reduction to investigate gene–gene interactions. CPM has shown success in building multilocus models with nonlinear interactions to explain and predict variability in plasma triglyceride [33] and plasma plasminogen activator inhibitor 1 levels [34]. Culverhouse *et al.* [35] developed a modification of the CPM method, the restricted-partition method (RPM), which heuristically restricts the exhaustive search used in CPM and thereby reduces its computational load for evaluating interactions.

Multifactor dimensionality reduction (MDR) is one data-reduction method developed specifically for genotypic data that has been successful at finding gene–gene interactions in both simulated data [36–39] and real data [36,40–42]. Artificial neural networks perform pattern recognition and have been applied to genotypic data with varied success [43–47]. However, recent work has improved the reliability of artificial neural networks through their optimization by evolutionary computation (EC) algorithms [48], which use a computational search methodology uniquely suited for rugged model landscapes [49]. One limitation of these computational methods is the potential difficulty of interpreting the biological implications of the resulting predictive models [50].

Concluding remarks: retooling for the future

None of the aforementioned methodologies is superior in all respects for the range of complicating factors that might be present in any given dataset. Given the relative shortcomings of our current analyses in complex diseases, we need to extend greatly the range of available analytical tools. There is a crucial need for extensive reevaluation of existing methodologies for complex diseases, as well as for massive efforts in new method development. It is important that empirical studies be conducted to compare and contrast the relative strengths and weaknesses of methods on specific types of problems. For example, although cluster analysis has shown promise in numerous other scientific and mathematical fields, its use with genetic, particularly discrete genotypic data, has not been adequately explored. Similarly, artificial neural networks modified with evolutionary computation have great potential for discovering nonlinear interactions between genes and environmental factors. However, work is still ongoing to evaluate its limitations with regard to the heritability and effect sizes that can be detected.

Ultimately, the real power of existing and yet-to-be-developed methods lies in our ability to marry them into a comprehensive approach to genetic analysis, so that their relative strengths and weaknesses can be balanced and few alternative hypotheses are left uninvestigated. We propose routinely taking a two-step approach to analysis because no single method adequately investigates heterogeneity and interaction issues simultaneously. For example, clustering or ordered subset analysis can be used first to uncover genotypic and/or phenotypic heterogeneity and to subdivide the data into more homogeneous groups. Then in a second step, specific tests of interactions, such as the S-sum statistic approach or the multifactor dimensionality-reduction method, could be used to investigate gene–gene or gene–environment interactions within each of the homogenized subgroups. This is still not a perfect approach, but it is an important improvement over the more common alternative of a single-pronged approach to analysis.

Such a combined strategy must be the future of genetic statistical analysis. We must harness our knowledge and experience of existing methods even as we open our minds to newly fashioned techniques and approaches. By thus ‘retooling’ our analyses, we provide the best opportunity for uncovering the genetic basis of common human disease.

Acknowledgements

We thank the reviewers, Marylyn Ritchie and Dan Hahs for their critical reading of this manuscript. This work was supported by NLM training grant T32 MH64913 and by NIH grants HL65234, AI59694, N532830 and A619085.

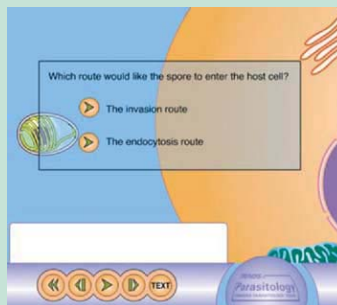
References

- 1 Spielman, R.S. *et al.* (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506–516
- 2 Morton, N.E. (1955) Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7, 277–318
- 3 Risch, N. (1988) A new statistical test for linkage heterogeneity. *Am. J. Hum. Genet.* 42, 353–364
- 4 Smith, C.A.B. (1963) Testing for heterogeneity of recombination fraction values in human genetics. *Ann. Hum. Genet.* 27, 175–182
- 5 Ott, J. (1992) Strategies for characterizing highly polymorphic markers in human gene mapping. *Am. J. Hum. Genet.* 41, 283–290
- 6 Hauser, E.R. *et al.* (1998) Stratified linkage analysis of complex genetic traits using related covariates. *Am. J. Hum. Genet.* 63, A45
- 7 Hauser, E.R. *et al.* (2004) Ordered subset analysis in genetic linkage mapping of complex traits. *Genet. Epidemiol.* 27, 53–63
- 8 Slonim, D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet.* 32, 502–508
- 9 Mountain, J.L. and Cavalli-Sforza, L.L. (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* 61, 705–718
- 10 Neuman, R.J. *et al.* (1999) Evaluation of ADHD typology in three contrasting samples: a latent class approach. *J. Am. Acad. Child Adolesc. Psychiatry* 38, 25–33
- 11 Neuman, R.J. *et al.* (2000) Clustering methods applied to allele sharing data. *Genet. Epidemiol.* 19, S57–S63
- 12 Pickles, A. *et al.* (1995) Latent class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Am. J. Hum. Genet.* 57, 717–726
- 13 Sham, P.C. *et al.* (1996) Further exploration of a latent class typology of schizophrenia. *Schizophr. Res.* 20, 105–115
- 14 Han, E.H. *et al.* (1997) Clustering based on association rule hypergraphs. In SIGMOD'97 Workshop on Research Issues on Data Mining and Knowledge Discovery (J. Peckham ed.), ACM Press, Tuscon Arizona
- 15 Hanson, R. *et al.* (1991) Bayesian classification with correlation and inheritance. In Proceedings of the 12th international Joint Conference on Artificial Intelligence, (Vol. 2) pp. 629–698 Morgan Kaufmann Publishers, San Francisco
- 16 Huang, Z. and Ng, M.K. (1999) A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans. Fuzzy Syst.* 7, 446–452
- 17 Moore, J.H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* 56, 73–82
- 18 Tong, A.H. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science* 303, 808–813
- 19 Concato, J. *et al.* (1993) The risk of determining risk with multivariable models. *Ann. Intern. Med.* 118, 201–210
- 20 Moore, J.H. and Williams, S.M. (2002) New strategies for identifying gene-gene interactions in hypertension. *Ann. Med.* 34, 88–95
- 21 Peduzzi, P. *et al.* (1996) A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49, 1373–1379
- 22 Hoh, J. *et al.* (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.* 11, 2115–2119
- 23 Ott, J. and Hoh, J. (2003) Set association analysis of SNP case-control and microarray data. *J. Comput. Biol.* 10, 569–574
- 24 Zee, R.Y. *et al.* (2002) A prospective evaluation of the angiotensin-converting enzyme D/I polymorphism and left ventricular remodeling in the 'Healing and Early Afterload Reducing Therapy' study. *Clin. Genet.* 61, 21–25
- 25 Wille, A. *et al.* (2003) Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. *Genet. Epidemiol.* 25, 350–359
- 26 Cheverud, J.M. and Routman, E.J. (1995) Epistasis and its contribution to genetic variance components. *Genetics* 139, 1455–1461
- 27 Friedman, J. (1991) Multivariate adaptive regression splines. *Ann. Stat.* 19, 1–141
- 28 Cook, N.R. *et al.* (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Stat. Med.* 23, 1439–1453
- 29 Morgan, J.N. and Sonquist, J.A. (1963) Problems in the analysis of survey data and a proposal. *J. Am. Stat. Assoc.* 58, 415–434
- 30 Province, M.A. *et al.* (2001) Classification methods for confronting heterogeneity. *Adv. Genet.* 42, 273–286
- 31 Shannon, W.D. *et al.* (2001) Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. *Genet. Epidemiol.* 20, 293–306
- 32 Good, I.J. (1961) A causal calculus. *Br. J. Philos. Sci.* 11, 305–318
- 33 Nelson, M.R. *et al.* (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11, 458–470
- 34 Moore, J.H. *et al.* (2002) A comparison of combinatorial partitioning and linear regression for the detection of epistatic effects of the ACE I/D and PAI-1 4G/5G polymorphisms on plasma PAI-1 levels. *Clin. Genet.* 62, 74–79
- 35 Culverhouse, R. *et al.* (2004) Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.* 27, 141–152
- 36 Ritchie, M.D. *et al.* (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147
- 37 Hahn, L.W. *et al.* (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19, 376–382
- 38 Ritchie, M.D. *et al.* (2003) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, phenocopy and genetic heterogeneity. *Genet. Epidemiol.* 24, 150–157
- 39 Hahn, L.W. and Moore, J.H. Ideal discrimination of discrete clinical endpoints using multilocus genotypes. *In Silico Biol.* (in press)
- 40 Williams, S.M. *et al.* (2004) The use of animal models in the study of complex disease: all else is never equal, or why do so many human studies fail to replicate animal findings? *BioEssays* 26, 170–179
- 41 Tsai, C.T. *et al.* (2004) Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation* 109, 1640–1646
- 42 Cho, Y.M. *et al.* (2004) Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47, 549–554
- 43 McCulloch, W. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133
- 44 Lucek, P.R. and Ott, J. (1997) Neural network analysis of complex traits. *Genet. Epidemiol.* 14, 1101–1106
- 45 Lucek, P. *et al.* (1998) Multi-locus nonparametric linkage analysis of complex trait loci with neural networks. *Hum. Hered.* 48, 275–284
- 46 Marinov, M. and Weeks, D. (2001) The complexity of linkage analysis with neural networks. *Hum. Hered.* 51, 169–176
- 47 Sherriff, A. and Ott, J. (2001) Applications of neural networks for gene finding. *Adv. Genet.* 42, 287–298
- 48 Fogel, G.B. and Corne, D.W. (2002) *Evolutionary Computation in Bioinformatics*, Elsevier Science, San Francisco
- 49 Ritchie, M.D. *et al.* (2003) Optimization of neural network architecture improves the power to identify gene-gene interaction in common diseases. *BMC Bioinformatics* 4, 28
- 50 Moore, J.H. and Ritchie, M.D. (2004) The challenges of whole-genome approaches to common diseases. *JAMA.* 291, 1642–1643
- 51 Rivolta, C. *et al.* (2002) Retinitis pigmentosa and allied diseases: numerous diseases, genes, and inheritance patterns. *Hum. Mol. Genet.* 11, 1219–1227
- 52 Kulczycki, L.L. *et al.* (2003) A clinical perspective of cystic fibrosis and new genetic findings: relationship of CFTR mutations to genotype-phenotype manifestations. *Am. J. Med. Genet.* 116A, 262–267
- 53 Povey, S. *et al.* (1994) Two loci for tuberous sclerosis: one on 9q34 and one on 16p13. *Ann. Hum. Genet.* 58, 107–127
- 54 Young, J. and Povey, S. (1998) The genetic basis of tuberous sclerosis. *Mol. Med. Today* 4, 313–319
- 55 Langston, J.W. *et al.* (1984) MPTP-induced parkinsonism in human and non-human primates – clinical and experimental aspects. *Acta Neurol. Scand.* 100 (Suppl.), 49–54
- 56 Langston, J.W. and Ballard, P. (1984) Parkinsonism induced by

- 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP): implications for treatment and the pathogenesis of Parkinson's disease. *Can. J. Neurol. Sci.* 11, 160–165
- 57 Frey, L.C. (2003) Epidemiology of posttraumatic epilepsy: a critical review. *Epilepsia* 44, 11–17
- 58 Harding, A.E. (1993) Clinical features and classification of inherited ataxia. *Adv. Neurol.* 61, 1–14
- 59 Rosenberg, R.N. (1995) Autosomal dominant cerebellar phenotypes: the genotype has settled the issue. *Neurology* 45, 1–5
- 60 Devos, D. *et al.* (2001) Clinical features and genetic analysis of a new form of spinocerebellar ataxia. *Neurology* 56, 234–238
- 61 Tager-Flusberg, H. and Joseph, R.M. (2003) Identifying neurocognitive phenotypes in autism. *Philos. Trans. R. Soc. B. Biol. Sci.* 358, 303–314
- 62 Bradford, Y. *et al.* (2001) Incorporating language phenotypes strengthens evidence of linkage to autism. *Am. J. Med. Genet.* 105, 539–547
- 63 Shao, Y. *et al.* (2002) Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *Am. J. Hum. Genet.* 70, 1058–1061
- 64 Lendvay, T.S. and Marshall, F.F. (2003) The tuberous sclerosis complex and its highly variable manifestations. *J. Urol.* 169, 1635–1642
- 65 Carrasquillo, M.M. *et al.* (2002) Genome-wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat. Genet.* 32, 237–244
- 66 Doh-ura, K. *et al.* (1989) Pro-to-leu change at position 102 of prion protein is the most common but not the sole mutation related to Gerstmann-Straussler syndrome. *Biochem. Biophys. Res. Commun.* 163, 974–979
- 67 Owen, F. *et al.* (1990) A codon 129 polymorphism in the PRIP gene. *Nucleic Acids Res.* 18, 3103
- 68 Collinge, J. *et al.* (1991) Genetic predisposition to iatrogenic Creutzfeldt–Jakob disease. *Lancet* 337, 1441–1442
- 69 Palmer, M.S. *et al.* (1991) Homozygous prion protein genotype predisposes to sporadic Creutzfeldt–Jakob disease. *Nature* 352, 340–342
- 70 De Silva, R. *et al.* (1994) Neuropathological phenotype and 'prion protein' genotype correlation in sporadic Creutzfeldt–Jakob disease. *Neurosci. Lett.* 179, 50–52
- 71 Doh-ura, K. *et al.* (1991) CJD discrepancy. *Nature* 353, 801–802
- 72 Goldfarb, L.G. *et al.* (1992) Creutzfeldt–Jakob disease cosegregates with the codon 178Asn PRNP mutation in families of European origin. *Ann. Neurol.* 31, 274–281
- 73 Aguzzi, A. and Weissmann, C. (1996) A suspicious signature. *Nature* 383, 666–667
- 74 Caspi, A. *et al.* (2003) Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science* 301, 386–389
- 75 Goate, A. *et al.* (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 349, 704–706
- 76 Larget-Piet, D. *et al.* (1994) Genetic heterogeneity of Usher syndrome type 1 in French families. *Genomics* 21, 138–143
- 77 Brown, W.T. *et al.* (1987) Further evidence for genetic heterogeneity in the fragile X syndrome. *Hum. Genet.* 75, 311–321
- 78 Haines, J.L. *et al.* (1991) Localization of one gene for tuberous sclerosis within 9q32–9q34, and further evidence for heterogeneity. *Am. J. Hum. Genet.* 49, 764–772
- 79 Schmidt, S. *et al.* (2004) Ordered subset linkage analysis supports a susceptibility locus for age-related macular degeneration on chromosome 16p12. *BMC Genet.* 5, 18
- 80 Grigull, J. *et al.* (2001) Clustering of pedigrees using marker allele frequencies: impact on linkage analysis. *Genet. Epidemiol.* 21(Suppl.), S61–S66
- 81 Shao, Y. *et al.* (2003) Fine mapping of autistic disorder to chromosome 15q11–q13 by use of phenotypic subtypes. *Am. J. Hum. Genet.* 72, 539–548
- 82 Relton, C.L. *et al.* (2004) Gene–gene interaction in folate-related genes and risk of neural tube defects in a UK population. *J. Med. Genet.* 41, 256–260
- 83 Kosachunhanun, N. *et al.* (2003) Genetic determinants of nonmodulating hypertension. *Hypertension* 42, 901–908
- 84 Costello, T.J. *et al.* (2003) Use of tree-based models to identify subgroups and increase power to detect linkage to cardiovascular disease traits. *BMC Genet.* 4, S66
- 85 Horng, J.T. *et al.* (2004) Identifying the combination of genetic factors that determine susceptibility to cervical cancer. *IEEE Trans. Inf. Technol. Biomed.* 8, 59–66

Getting animated with parasites!

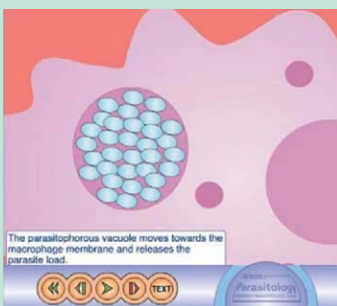
Interested in the molecular cell biology of host–parasite interactions? Then take a look at the online animations produced by *Trends in Parasitology*, one of our companion *TRENDS* journals. The pictures below are snapshots from two of our collection of animations revealing the latest advances in understanding parasite life cycles. Check them out today!



Microsporidia: how can they invade other cells?

By C. Franzen [(2004) *Trends Parasitol.* 20, 10.1016/j.pt.2004.04.009]

<http://archive.bmn.com/supp/part/franzen.html>



Interaction of *Leishmania* with the host macrophage

By E. Handman and D.V.R. Bullen [(2002) *Trends Parasitol.* 18, 332–334]

<http://archive.bmn.com/supp/part/swf012.html>