JPP
Journal of
Pharmacy and Pharmacology

JPP 2010, 62: 161–166
© 2010 The Authors.
Journal compilation © 2010
Royal Pharmaceutical Society
of Great Britain
Received April 20, 2009
Accepted November 3, 2009
DOI 10.1211/jpp/62.02.0002
ISSN 0022-3573

Methodological and statistical issues in pharmacogenomics

Bas J. M. Peters[a], Andrei S. Rodin[b], Anthonius de Boer[a] and Anke-Hilse Maitland-van der Zee[a]

[a]Department of Pharmacoepidemiology & Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, the Netherlands and [b]Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, Texas, USA

Abstract

Pharmacogenomics strives to explain the interindividual variability in response to drugs due to genetic variation. Although technological advances have provided us with relatively easy and cheap methods for genotyping, promises about personalised medicine have not yet met our high expectations. Successful results that have been achieved within the field of pharmacogenomics so far are, to name a few, *HLA-B**5701 screening to avoid hypersensitivity to the antiretroviral abacavir, thiopurine S-methyltransferase (*TPMT*) genotyping to avoid thiopurine toxicity, and *CYP2C9* and *VKORC1* genotyping for better dosing of the anticoagulant warfarin. However, few pharmacogenetic examples have made it into clinical practice in the treatment of complex diseases. Unfortunately, lack of reproducibility of results from observational studies involving many genes and diseases seems to be a common pattern in pharmacogenomic studies.

In this article we address some of the methodological and statistical issues within study design, gene and single nucleotide polymorphism (SNP) selection and data analysis that should be considered in future pharmacogenomic research. First, we discuss some of the issues related to the design of epidemiological studies, specific to pharmacogenomic research. Second, we describe some of the pros and cons of a candidate gene approach (including gene and SNP selection) and a genome-wide scan approach. Finally, conventional as well as several innovative approaches to the analysis of large pharmacogenomic datasets are proposed that deal with the issues of multiple testing and systems biology in different ways.

Keywords bioinformatics; data analysis; methodology; pharmacogenomics; statistics

Introduction

For many decades we have known that patients respond differently to drugs. The contribution of genetic variation to the interindividual response to isoniazid was described by Hughes *et al.* as early as 1954.[1] However, although technological advances have provided us with relatively easy and cheap methods for genotyping, promises about personalised medicine have not yet been met.

A recently published trial investigated the clinical value of screening *HLA-B**5701 for hypersensitivity to the antiretroviral abacavir and showed that genetic screening resulted in a significant reduction in the risk of hypersensitivity to abacavir.[2] However, few pharmacogenetic examples have made it into clinical practice in the treatment of complex diseases, although observational studies have described many pharmacogenetic interactions involving many genes and diseases. For instance, in 1998 Kuivenhoven *et al.* reported a pharmacogenetic interaction between response to pravastatin and the cholesteryl ester transfer protein (*CETP*) TaqIb polymorphism.[3] Homozygous carriers of the B1 allele experienced greatest benefit from pravastatin compared with placebo in terms of progression of coronary atherosclerosis. Ten years and many publications later, this interaction was not replicated in any other study.[4] Unfortunately, the lack of reproducibility seems to be a common pattern in (pharmaco-)genetic studies[5,6] although there have been success stories, such as the aforementioned abacavir–gene interaction. An insightful review article by Evans and Relling provides a thoughtful elaboration on why pharmacogenomics has not reached clinical practice to any significant extent.[7] Briefly, the obstacles include: (i) education of the medical community; (ii) difficulties encountered in conducting definitive clinical

Correspondence:
Anke-Hilse Maitland-van der Zee, Utrecht University,
Faculty of Science, Division of Pharmacoepidemiology and Pharmacotherapy, PO Box 80082,
3508 TB Utrecht,
the Netherlands.
E-mail: a.h.maitland@uu.nl

pharmacogenomics studies (lack of funding, study design issues), and (iii) technical challenges of genetic testing comparable to other molecular diagnostics.

In this article we address some of the methodological and statistical issues within study design, selection of genes and single nucleotide polymorphisms (SNPs) and data analysis that should be considered in future pharmacogenomic research.

Methodological issues

Randomised clinical trials are considered the highest level of evidence and are essential in convincing practising clinicians of the value of genotyping. An example is the forthcoming European Pharmacogenomics Approach to Coumarin Therapy (EU-PACT) trial, which will evaluate the benefit of genotyping to coumarin dosing and the risk of clinical events.[8] However, most commonly used in pharmacogenomics is the case–control design, largely because of its high efficiency: relatively few patients have to be genotyped (researchers can select a certain patient group based on disease status beforehand), the relative ease of patient recruitment, and late-onset diseases can be used as outcome measures without follow-up problems. An issue that may arise in pharmacogenomic studies is the inclusion of subjects. Subjects are asked to give their informed consent to the researcher to collect a sample and analyse the DNA. Privacy is always of high importance and patients' anonymity is therefore guaranteed in publications. However, it is important to realise that as few as 75 independent SNPs could unequivocally lead to an individual person,[9] although in practice we have not experienced this as a major problem in the process of data collection.[10]

In recent years, traditional issues (pros and cons) relating to the different designs of epidemiological studies have been discussed in countless comprehensive reviews.[11,12] For that reason, we focus on study design specific to pharmacogenomics: candidate gene approach (CGA) versus genome-wide scan, and confounding (population structure).

Candidate gene approach and genome-wide association studies

A pharmacogenomic study with a CGA typically involves a couple to tens of SNPs within each candidate gene that is possibly involved in the response to a particular drug. In contrast, a genome-wide association study (GWAS) seeks to identify variants that modify the response to a certain drug throughout the whole genome. CGA and GWAS both have their pros and cons, and differ in several significant ways. To begin with, the number of SNPs for a CGA can range from dozens to thousands, whereas in a GWAS between 100 000 and more than 1 000 000 SNPs are genotyped. Importantly, in contrast to a GWAS that comes with a fixed standard array, custom arrays for CGA studies allow SNPs to be selected by the researcher. A CGA on a genome-wide scale is also possible, although the SNPs in the selected candidate genes will only be those available on the standard arrays.

A GWAS is largely data driven (hypothesis-free) whereas a CGA is hypothesis driven because the selection of genes and SNPs is based on prior (expert) knowledge. The result is that GWAS can detect SNPs in genes that were not considered candidate genes before, or SNPs located outside of genes. It is very unlikely that these SNPs would have been found using a CGA. On the other hand, a CGA may detect associations that would not have been identified in a GWAS because of power issues (discussed below). Lastly, although the costs of a GWAS have plummeted, budget constraints may still only allow a CGA.

For the CGA, the first step is the selection of genes related to the research question. Candidate genes can be genes that have previously been reported to be associated in the research field of interest. In addition, genes involved in the pharmacokinetics (absorption, distribution, metabolism and elimination) and pharmacodynamics (drug targets) of a drug should be considered as candidate genes. Finally, genes related to the underlying disease or intermediate phenotype may be important for the pharmacogenomics of a certain drug. In addition to a straightforward manual literature search, more advanced methods are now available, one example being a method by Hansen *et al.*,[13] who describe a candidate gene-selection method for pharmacogenomic studies that specifically ranks 12 460 genes in the human genome according to the potential relevance to a drug and its indication. Interestingly, it uses gene–drug, gene–gene, and data from drug–drug similarities to construct a network for gene ranking. For several drugs, they were able to identify new candidate genes.[13]

SNPs come in different forms: synonymous, in which the mutation does not change the polypeptide sequence, and non-synonymous, in which the polymorphism results in a different polypeptide sequence. SNPs in non-coding regions may also be important because they can affect processes such as expression and gene splicing. Different types of SNPs should be considered for a CGA: SNPs that were previously associated, SNPs with functional annotation (coding SNPs) and tag SNPs. Coding SNPs can be found in the dbSNP database of NCBI (www.ncbi.nlm.nih.gov/SNP). The main source for many SNPs that are available is the HapMap project (www.hapmap.org) in which 270 individuals with different ancestry have been genotyped for over 3.5 million SNPs. Within the HapMap, so called tag SNPs have been identified in four populations. Tag SNPs are in strong linkage disequilibrium with other SNPs so they can serve as a proxy for the other SNPs, thus tremendously reducing the number of SNPs needed to contain the genetic variance of a gene.[14] Different methodologies for SNP selection are available which not only take advantage of tag SNPs but also give the option of including coding SNPs and defining the size of the flanking region. Two such web-based services based primarily on the international HapMap project are QuickSNP[15] and Tagger.[16] It is important to consider the r^2, which is a measure of the required linkage disequilibrium strength (usually set at 0.8), and the allele frequency of a SNP in the research population, as a low allele frequency may ultimately lead to low power.

The first pharmacogenomic GWASs are starting to emerge.[17,18] When designing a GWAS, one should be aware of the computational burden, as up to more than 1 million variables are available in the epidemiological dataset. Furthermore, the recommendations made by the Wellcome Trust Case Control Consortium (WTCCC) should be considered.[19] First, the WTCCC stresses the importance of careful quality control, as small systematic differences can easily produce effects that may obscure true associations. Second, the potential for hidden population structure is a phenomenon that should be recognised (discussed below). Third, even with many cases and controls (2000 and 3000, respectively), the study power is limited to the detection of common variants with large effects only. Therefore, meta-analysis of existing GWAS is encouraged if possible. Furthermore, the WTCCC underlines the importance of replication studies to confirm true associations, and functional studies to gain more insight into, and mechanistic understanding of, the underlying biological molecular mechanisms.[19] Because tag SNPs that show an association are likely to be in linkage disequilibrium with the causal variants, re-sequencing of this region is of major importance for the identification of causal variants.

Population structure

Dealing with confounding in epidemiology is a huge challenge and has been the subject of many discussions.[20] Besides the conventional confounding in epidemiology, pharmacogenomic research is faced with other potential confounding such as 'hidden population structure' (or population stratification). This phenomenon is present when genetically incompletely mixed distinct subpopulations exist within the research population.[21] Associations may then reflect confounding due to the different prevalence of a variant allele and prevalence or magnitude of the outcome of interest. Moreover, the exposure to a drug can be unevenly distributed among genetically different subgroups. Therefore, the gene–drug interactions may be biased by the population structure.[21]

Minimising irrelevant allelic differences in groups can be achieved by sampling cases and controls from the same population and/or by matching cases and controls on the basis of genetic background using surrogate markers such as geographic proximity, physical characteristics and self-reported ethnicity.[22]

Most widely used methods to detect and adjust for this problem are genomic control, structured association methods and the EIGENSTRAT method.[23] The genomic control method, developed by Devlin and Roeder, corrects for variance inflation caused by population structure, using SNPs that are unrelated to the outcome (case or control).[24] The variance inflation factor, denoted by λ, is based on the assumption that λ is the same across the genome for all null SNPs, and can be calculated by dividing the median of the Armitage test statistic for the 'null' SNPs by 0.456 (the median of a chi squared with one degree of freedom [χ^2 distribution, df = 1]). λ is expected to be larger than 1 but in the absence of population stratification may also be smaller than 1 (if this is the case, it is suggested to be set to 1[25]). Subsequently, the Armitage test statistic for the candidate SNPs are divided by λ. This method has also been extended for continuous outcome measures.[26]

In addition, Pritchard and colleagues developed a two-phase structured association method that can test for association in the presence of population structure.[27] The first phase uses the 'null' SNPs to identify the presence of population structure – assuming any of the associations to be the result of population structure – to subsequently assign the individuals to putative subpopulations. In the second phase, associations are tested conditionally on the subpopulation allocation.[27] Of note, the result of this computationally demanding method is highly sensitive to the assumed and unknown number of subpopulations.

Finally, a popular tool to detect and correct for population structure is the EIGENSTRAT method,[28] based on principal component analysis. First, principal component analysis is applied to genotype data ('null' SNPs) to infer continuous axes of genetic variation. Second, using the residuals of linear regression, the observed genotypes and phenotypes are continuously adjusted by the amounts attributable to ancestry along each axis. Finally, use of the ancestry-adjusted genotypes and phenotypes used to calculate association statistics takes into account the population structure.

Data analysis

The relationship between variation in DNA sequence and clinical endpoints is likely to involve gene–gene (epistatic) interactions. The term epistasis is not unequivocal,[29,30] as it is used in different contexts. Generally, epistasis can be defined as either biological or statistical. Biologically, epistasis is the physical interactions among proteins or other biomolecules that affect the phenotype. Statistically, epistasis is generally defined in terms of deviation from a model of additive effects.[30] Gene–gene interactions may actually be a plausible explanation for non-replication of positive associations, since these interactions may vary between populations.

Traditional statistics is not well suited to deal with gene–gene and gene–environment interactions on a large scale. In pharmacogenomics we are faced with an additional challenge – the primary goal of our analyses is not the genetic association with the phenotype, but rather the effect of genetics on the association between a certain drug and the phenotype.

Multiple comparisons in regression models

As the number of SNPs increases, data analysis becomes a statistical challenge because of the multiple testing (comparisons) problem. Generally, the *P*-value threshold that is considered significant in biomedical research is set at 0.05. This *P* value is not appropriate when testing many variables, as the frequency of type I errors will increase. Testing 20 random variables will give a 64% chance of finding one significantly associated SNP at random (P ($\geq$ 1 significant result) = 1 − P (no significant results) = 1 − $(1 - 0.05)^{20}$ = approx 0.64). There are different ways of dealing with this issue. The Bonferroni correction can be applied by setting the significance cut-off at the *P* value for one test (i.e. 0.05) divided by the number of tests.[31] When testing 100 SNPs, the null hypothesis will then only be rejected when the *P* value is below 0.05/100 = 0.0005. In (pharmaco)genomics, the Bonferroni correction can be considered too stringent, as it may wipe out many small effects that one may actually expect (increased rate of type II errors). One of the reasons Bonferroni correction is too conservative is that many SNPs are not independent.

Another way of dealing with multiple testing is the increasingly popular false discovery rate (FDR) approach.[32] The FDR estimates the expected proportion of false positives among the tests declared significant, expressed as a *q* value. In the case where the FDR gives a *q* value of 0.2 for 50 significantly associated SNPs, the proportion of false positives would be 20% (10 SNPs). This approach is very different from using a threshold *P* value. Only a few of the 50 SNPs that were associated using the FDR would have been associated when applying a Bonferroni correction. There is no threshold *q* value that is considered standard. Depending on the study (number of patients, number of SNPs, biological plausibility), different *q* values might be chosen.[33]

Performing numerous tests that are necessary to analyse the large number of SNPs, epistasis, the exposure to a drug, and drug–/gene–environment interactions may have a detrimental effect on the ability to detect small effects because of the avalanche of multiple testing issues. It may therefore be necessary to deviate from conventional methods to other methods. Data analysis within the Bayesian framework, for example, may be of great value as there is no penalty for multiple analyses of the data. After all, the prior probability of an association should not be affected by the tests that the investigator chooses to carry out.[23]

An alternative multi-stage analysis strategy proposed for pharmacogenomic data exploration is shown in Figure 1. The first step entails variable (SNP) selection and ranking, where the number of potentially predictive SNPs is significantly reduced. Second, a set of SNPs with high predictive potential is refined, and a descriptive/predictive model is fitted, in order to reverse-engineer the biological relationships underlying the system in question. Finally, traditional statistical methods are used to calculate odds ratios or relative risks for the specific associations between SNPs, phenotypes, exposures and other epidemiological factors.

For **step 1**, conventional univariate methods (such as logistic regression or simple contingency tables) can be used to rank SNPs and to reduce them to a smaller subset (several hundred SNPs) based on the association strength. The choice of the *P*-value cut-off point is somewhat arbitrary (e.g. 'top 100'), but should ideally be a function of the data itself (i.e. how many SNPs actually carry the signal, as opposed to noise) and thus should probably depend on the number of SNPs that were tested. This univariate approach (known as 'filtering' in computer science vernacular[34]) does not take into account the interactions that may play a role in predicting the outcome. An alternative variable ranking/selection method is a random forests (RF) classifier.[35] RF is capable of accounting for some epistatic interactions, because it aggregates many (thousands, usually) single classification and regression trees (CARTs[36]). A single decision tree is generated by recursively partitioning the data set into subsets. In the whole data set the best possible predictor of the case status is selected to split the root node into two 'child' nodes (e.g. smoking, non-smoking). In the next steps, recursively, these child nodes are split again, using the best remaining predictors. This process continues until either all cases and controls are separated, or the terminal nodes are too small to split. To build an RF, two randomisation mechanisms are added. First is bootstrapping, where a number of randomised samples are generated from the original dataset by using resampling with replacement.[37] The second randomisation mechanism is the selection of a random (and small) subset of predictors (SNPs) to build each single tree. Once a 'forest' (consisting of thousands of randomised single decision trees) is built to classify a new observation, each tree in the forest classifies it separately; the class that gets the most votes predicts the class of the new observation. RF is capable of accounting for variable interactions because many possible variable combinations are encountered repeatedly within the forest. Another aspect of RF that makes it particularly attractive for large-scale studies is that it is more computationally efficient than comparable classifiers when the number of variables is high. Numerous RF implementations are available.[38–40]

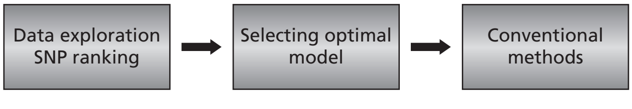

Figure 1 A strategy for analysing large (pharmaco)genomic datasets.

In **step 2**, a set of SNPs with high predictive potential will be selected, and the relationships between these SNPs and other factors will be ascertained. Methods like Bayesian or belief network (BN) modelling, multifactor-dimensionality reduction (MDR), boosted classifiers and also RF are suitable for this step.

A BN provides a systems biology analytic approach for identifying interactions between genetic, physiological and environmental variables, including the outcome of interest.[41,42] A biological network modelling genotype-to-phenotype relationship is represented visually as a graph consisting of nodes (indicating discrete and continuous variables, such as SNPs, environmental factors, metabolite concentration, phenotypes, etc.) and directed edges (or arrows) that link mutually dependent nodes. Absence of an edge between two nodes indicates their conditional independence. The edge directionality is somewhat arbitrary and is not intended to imply causation; rather, it is used for mathematical convenience to distinguish between the 'parent' and the 'child' nodes. The edge strength indicates the relative magnitude of the dependency between the two variables, given the other interrelationship, and is measured as the marginal likelihood ratio test of the BN with the edge versus the otherwise identical BN without the edge. An edge between two SNPs is indicative of strong linkage disequilibrium; therefore, BN can simultaneously take into account linkage disequilibrium while doing genotype–phenotype association analyses. Since we are interested in predicting the efficacy of a drug on the outcome of interest, the BN can be reduced to a sub-network of the outcome of interest and a limited number of immediately predictive variables. A formal conceptualisation of such a sub-network is the Markov 'blanket'.[43] By definition, the Markov blanket of node A consists of the parents of A, the offspring of A and the nodes that share an offspring with A (Figure 2). Given its Markov blanket, the outcome variable is independent of all the other variables in the network. Dependencies within the Markov blanket may be checked for statistical robustness using bootstrapping or subsequent standard statistical tests. Because the BN data analysis can be carried out

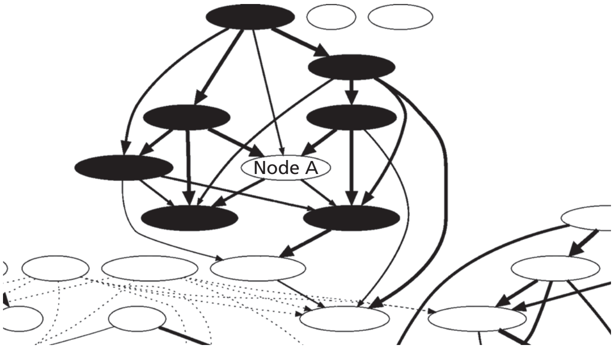

Figure 2 Bayesian network with Markov Blanket. In a Bayesian network, the Markov blanket of node A includes the black nodes (parents, children and the other parents of all of its children).

simultaneously with other analysis methods, the level of overfitting (sensitivity/specificity balance) can be adjusted so that the number of predictive variables (SNPs) generated is roughly the same across the whole palette of analysis methods, and the predictive variable rankings generated by the different analysis methods can be compared directly. Alternatively, simulation studies can be performed to ascertain the optimal balance explicitly.

Recently, BNs were used to study the pharmacogenomics of short-acting bronchodilator medication.[44] Himes *et al.* reported 15 of 426 SNPs in 15 of 254 genes to be predictive of bronchodilator response using a BN model with fair accuracy. They compared the BN (multivariate) model with a single-gene approach, and demonstrated that the BN model was much better at predicting bronchodilator response, suggesting that some of the relationships among SNPs and bronchodilator response were potentially true biological relationships. Interestingly, they found two relationships between two SNPs and bronchodilator response, a relationship that would not have been captured using traditional statistics.[44]

Another efficient descriptive/predictive modelling method is boosted classifiers.[45] While RF is a robust and scalable classifier, the complexity of the generated model (thousands of single decision trees) means that it is hardly interpretable by a human expert. On the other hand, single decision trees such as CART are not particularly robust. An attractive compromise is the boosted classifier in which there is more than one tree but their construction is adaptive rather than random (each new tree is aimed primarily at the observations misclassified by the preceding tree), and the number of trees is low (no more than a dozen, typically). The model then can be expressed as a set of 'if/then' rules that also cluster the sample into groups of similar individuals, for example, 'out of 200 individuals, 24 have SNP1 = AA, SNP2 = AG and SNP3 = AC; out of these 24 individuals, 22 are cases, and 2 are controls'. Therefore, by representing the resulting decision tree models as sets of rules ('rule sets'), we perform both classification and sample clustering, thus accounting (to some extent) for genetic heterogeneity within the sample. Various implementations of boosted classifiers are available.[40,46,47]

MDR is also a computer-science based method, developed by Ritchie *et al.*[48] for the explicit identification and characterisation of high-order gene–gene and gene–environment interactions in relatively small-scale studies. MDR is capable of doing so by reducing genotype predictors from multiple dimensions to one by pooling multilocus genotypes into high- and low-risk groups. In other words, a one-dimensional multilocus-genotype variable is computed for each model (combination of predicting variables).

A good example of how MDR has been applied in pharmacogenomic research is that of Motsinger *et al.,* who investigated the effect of variants in a set of selected genes on the pharmacokinetics and treatment response to efavirenz. They showed that combinations of variants in *CYP2B6* and *ABCB1* were the most predictive for the 24-h area under the plasma–concentration time curve for efavirenz and for virologic failure and toxicity failure.[49] Unfortunately, the number of variables that can be included in the model is limited; this is the price one pays for addressing the non-additive interaction issue explicitly. MDR software is publicly available at www.epistasis.org/software.html.

It should be mentioned that many other novel (mostly computer science-derived) methods can be used for variable selection and descriptive and predictive modelling. A useful internet resource and a convenient starting point for further exploration of data mining software can be found at www.kdnuggets.com/software/index.html.

Step 3 completes the analyses by using conventional statistical methods to calculate odds ratios or hazard ratios, allowing a direct epidemiological interpretation. This is beyond the scope of this article.

Future perspectives

Although there are several well-established examples,[2,50–52] pharmacogenomics is still relatively uncommon in clinical practice. Concomitant to the issues we have discussed in this paper, future research should benefit from the technical advantages that modern technology has to offer. Pharmacogenomics is a staggeringly complex research field that requires a multi-disciplinary approach. Therefore, genome-wide methods at the level of expression, genotype scans and proteomics should be combined with what is already known about a drug. In addition, bioinformatics and ontology-based approaches should play important roles in sorting through the large amounts of data currently available.

Declarations

Conflict of interest

The Author(s) declare(s) that they have no conflicts of interest to disclose.

Funding

Anke-Hilse Maitland-van der Zee is funded by a Veni grant from the Netherlands Organization for Scientific Research (NWO). Andrei S. Rodin is funded by grants from the NIH, Gilson-Longenbaugh Foundation and UTHSCH Office of Biotechnology Seed Grant Program.

References

1. Hughes HB *et al*. Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. *Am Rev Tuberc* 1954; 70: 266–273.
2. Mallal S *et al*. HLA-B*5701 screening for hypersensitivity to abacavir. *N Engl J Med* 2008; 358: 568–579.
3. Kuivenhoven JA *et al*. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. The Regression Growth Evaluation Statin Study Group. *N Engl J Med* 1998; 338: 86–93.
4. Boekholdt SM *et al*. Cholesteryl ester transfer protein TaqIB variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13,677 subjects. *Circulation* 2005; 111: 278–287.
5. Hirschhorn JN *et al*. A comprehensive review of genetic association studies. *Genet Med* 2002; 4: 45–61.
6. Peters BJM *et al*. Pharmacogenetics of cardiovascular drug therapy. *Clin Cases Miner Bone Metab* 2009; 6: 55–65.
7. Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. *Nature* 2004; 429: 464–468.
8. van Schie RMF *et al*. Genotype-guided dosing of coumarin derivatives. The EUropean Pharmacogenetics of AntiCoagulant Therapy (EU-PACT) Trial Design. *Pharmacogenomics* 2009; 10: 1687–1695.
9. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science* 2004; 305(5681): 183.
10. van Wieren-de Wijer DB *et al*. Recruitment of participants through community pharmacies for a pharmacogenetic study of antihypertensive drug treatment. *Pharm World Sci* 2009; 31: 158–164.
11. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; 2: 91–99.
12. Dempfle A *et al*. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. *Eur J Hum Genet* 2008; 16: 1164–1172.
13. Hansen NT *et al*. Generating genome-scale candidate gene lists for pharmacogenomics. *Clin Pharmacol Ther* 2009; 86: 183–189.
14. Frazer KA *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; 449: 851–861.
15. Grover D *et al*. QuickSNP: an automated web server for selection of tagSNPs. *Nucleic Acids Res* 2007; 35(Web Server issue): W115–W120.
16. de Bakker PI *et al*. Efficiency and power in genetic association studies. *Nat Genet* 2005; 37: 1217–1223.
17. Link E *et al*. SLCO1B1 variants and statin-induced myopathy – a genomewide study. *N Engl J Med* 2008; 359: 789–799.
18. Cooper GM *et al*. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* 2008; 112: 1022–1027.
19. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; 447: 661–678.
20. Rothman KJ, Greenland S. *Modern Epidemiology*, 2nd edn. Greenland: Lippincott-Raven, 1998.
21. Umbach DM. Invited commentary: on studying the joint effects of candidate genes and exposures. *Am J Epidemiol* 2000; 152: 701–703.
22. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; 361: 598–604.
23. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006; 7: 781–791.
24. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; 55: 997–1004.
25. Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; 66: 1933–1944.
26. Bacanu SA *et al*. Association studies for quantitative traits in structured populations. *Genet Epidemiol* 2002; 22: 78–93.
27. Pritchard JK *et al*. Association mapping in structured populations. *Am J Hum Genet* 2000; 67: 170–181.
28. Price AL *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904–909.
29. Phillips PC. Epistasis – the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 2008; 9: 855–867.
30. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 2005; 27: 637–646.
31. Shaffer JP. Multiple hypothesis testing. *Ann Rev Psychol* 1995; 46: 561–584.
32. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; 100: 9440–9445.
33. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statistical Soc, Series B (Methodological)* 1995; 57: 289–300.
34. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Machine Learning Res* 2003; 3: 1157–1182.
35. Breiman L. Random forests. *Machine Learning* 2001; 45: 5–32.
36. Breiman L *et al. Classification and regression trees*. Belmont, CA, USA: Wadsworth International, 1984.
37. Efron B. Bootstrap methods: another look at the jackknife. *Ann Statistics* 1979; 7: 1–26.
38. Breiman L, Cutler A. Random forests. www.stat.berkeley.edu/users/breiman/RandomForests/ [accessed 11 November 2008].
39. Breiman L, Cutler A. Salford Systems Software: Random Forests. www.salfordsystems.com/randomforests.php [accessed 11 November 2008].
40. Witten IH, Frank E. *Data Mining: Practical machine learning tools and techniques,* 2nd edn. San Francisco, CA, USA: Morgan Kaufmann, 2005.
41. Pearl J. *Probabilistic reasoning in intelligent systems.* San Francisco, CA, USA: Morgan Kaufman, 1988.
42. 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.
43. Judea P. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA, USA: Morgan Kaufmann, 1988.
44. Himes BE *et al*. Predicting response to short-acting bronchodilator medication using Bayesian networks. *Pharmacogenomics* 2009; 10: 1393–1412.
45. Freund Y, Schapire RE. A short introduction to boosting. *J Japan Soc Artif Intel* 1999; 14: 771–780.
46. Cohen W. The SLIPPER Rule Learning System. www.cs.cmu.edu/~wcohen/slipper/ [accessed 11 November 2008].
47. The R Project for Statistical Computing. www.r-project.org [accessed 11 November, 2008].
48. Ritchie MD *et al*. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; 69: 138–147.
49. Motsinger AA *et al*. Multilocus genetic interactions and response to efavirenz-containing regimens: an adult AIDS clinical trials group study. *Pharmacogenet Genomics* 2006; 16: 837–845.
50. Lennard L *et al*. Pharmacogenetics of acute azathioprine toxicity: relationship to thiopurine methyltransferase genetic polymorphism. *Clin Pharmacol Ther* 1989; 46: 149–154.
51. Muss HB *et al*. c-erbB-2 expression and response to adjuvant therapy in women with node-positive early breast cancer. *N Engl J Med* 1994; 330: 1260–1266.
52. Weinshilboum R. Inheritance and drug response. *N Engl J Med* 2003; 348: 529–537.