Human Molecular Genetics, 2011, Vol. 20, Review Issue 2 **R182–R188**
doi:10.1093/hmg/ddr378
Advance Access published on August 25, 2011

Genetic risk prediction in complex disease

Luke Jostins and Jeffrey C. Barrett*

Statistical and Computational Genetics, Wellcome Trust Sanger Institute, Cambs CB10 1HH, UK

Received July 21, 2011; Revised and Accepted August 22, 2011

Attempting to classify patients into high or low risk for disease onset or outcomes is one of the cornerstones of epidemiology. For some (but by no means all) diseases, clinically usable risk prediction can be performed using classical risk factors such as body mass index, lipid levels, smoking status, family history and, under certain circumstances, genetics (e.g. *BRCA1/2* in breast cancer). The advent of genome-wide association studies (GWAS) has led to the discovery of common risk loci for the majority of common diseases. These discoveries raise the possibility of using these variants for risk prediction in a clinical setting. We discuss the different ways in which the predictive accuracy of these loci can be measured, and survey the predictive accuracy of GWAS variants for 18 common diseases. We show that predictive accuracy from genetic models varies greatly across diseases, but that the range is similar to that of non-genetic risk-prediction models. We discuss what factors drive differences in predictive accuracy, and how much value these predictions add over classical predictive tests. We also review the uses and pitfalls of idealized models of risk prediction. Finally, we look forward towards possible future clinical implementation of genetic risk prediction, and discuss realistic expectations for future utility.

EPIDEMIOLOGY AND RISK PREDICTION

Attempting to predict the onset and progression of disease is one of the cornerstones of epidemiology. Accurate risk prediction can enable targeted preventative treatments, such as fitness regimens for patients at risk of cardiovascular disease, or increased mammogram frequency for patients with high breast cancer risk. Traditional epidemiological risk prediction incorporates a small number of environmental and clinical factors known to be associated with disease, such as body mass index and lipid levels for type 2 diabetes (1) or the various Framingham risk scores for predicting cardiovascular outcomes (2). Some of these predictions (e.g. type 2 diabetes) are accurate enough to be clinically useful, but for many diseases (e.g. Crohn's) the prediction is barely better than chance. Evaluating the accuracy of environmental prediction is further complicated by recall-bias and the potential for reverse causality when data are retrospectively collected.

The importance of genetic factors in risk prediction has long been appreciated, and is exemplified in a simple form by the value of family history in predicting many complex diseases. Nevertheless, only a few specific molecular genetic variables have played an important role in historical risk predictions [e.g. BRCA1 and BRCA2 in familial breast and ovarian cancer (3)]. The success of recent genome-wide association studies (GWAS), however, has rapidly changed the outlook for genetic risk prediction. These studies have unlocked thousands of clearly validated genetic associations to complex diseases, but their generally weak effects have left their predictive value and clinical utility subject to hot debate.

The principal outputs of the GWAS revolution have been the new insights into the biological mechanisms of disease (4–6), but it is also possible to use the fruits of GWAS extend genetic prediction from single large factors to aggregations of individually weak effects. In order to explore post-GWAS risk prediction, we first discuss the relative merits of different statistical summaries of prediction. We next consider the state of prediction from current GWAS knowledge and consider possible insights from idealized models of prediction—as well as their potential pitfalls. Finally, we look forward towards possible future clinical implementation of genetic risk prediction.

QUANTIFYING PREDICTIVE ACCURACY

The widespread interest in risk prediction has led to the development of an entire ecosystem of classification metrics. The most appropriate statistics to use depend on the

*To whom correspondence should be addressed at. Tel: +44 1223834244; Fax: +44 1223494919; Email: barrett@sanger.ac.uk

circumstances and the question being asked. We discuss below the approaches to quantifying prediction that are most relevant to the subsequent discussion in this paper. More detailed discussions of many of these metrics, how they are related and how they can be classified are available elsewhere (7,8).

Predictive tests can produce either a binary classification of each individual as high or low risk, or a quantitative risk score that represents the degree of risk for each individual. Optimally, such a risk score is equal to the posterior probability of developing the disease (e.g. from logistic regression), although some widely used scores (such as risk allele counting) do not meet this criterion. Scores can be transformed into binary outcomes by defining high risk to be individuals with a score greater than a threshold T, and all others as low risk. Analysis of risk for quantitative traits is even more straightforward, as such transformations are unnecessary, and much of the discussion below is equally applicable to such non-categorical scenarios.

The simplest measures of classification accuracy are the sensitivity and specificity of the test, respectively defined as the proportion of individuals who develop the disease who were classified as high risk, and the proportion of healthy individuals classified as low risk. These values vary with the choice of T, which represents the unavoidable trade-off between sensitivity and specificity: predicting everyone will become ill will guarantee complete sensitivity, but without any specificity. A plot of the sensitivity against 1-specificity for all possible choices of T is known as a receiver-operating characteristic (ROC) curve (9). The **a**rea **u**nder the ROC **c**urve (the AUC, sometimes called the C statistic) has the pleasing property of being equal to the probability that a randomly selected individual with the disease has a higher score than a randomly selected healthy individual.

ROC statistics are not without their disadvantages, however. They are not dependent on the prevalence of the disease, with the result that even a high AUC predictor of a very rare disease is often of little practical use. For instance, consider a predictor of a disease with a prevalence of 1%: even with specificity and sensitivity of 0.93 (typical of a test with an AUC of around 0.98) only 12% of the individuals who test positive will go on to develop the disease. Alternate statistics such as the positive and negative predictive values account for prevalence. The positive predictive value (PPV) is the proportion of people who test positive for the disease who go on to develop it, and the negative predictive value (NPV) is the proportion of people who test negative who remain healthy. Note that, like sensitivity and specificity, the positive and negative predictive values are dependent on the risk score threshold T. These statistics can be used to tune the parameter T: for instance, while in the example above a test of a 1% disease with an AUC of 0.98 had a PPV of 12% and a sensitivity of 93%, by raising the threshold we could also produce a test with a sensitivity of only 40%, but a PPV of 75%. This test would miss a larger proportion of people with the disease, but would have much higher confidence that those it caught will go on to show symptoms.

PREDICTION IN THE POST-GWAS ERA

The most basic means of incorporating genetic information in risk prediction is via family history, which has predictive accuracy proportional to both the heritability and prevalence of disease (the AUC of a single sibling family history is $1/2+K(\lambda_S-1)/2(1-K)$, where K is the prevalence and λ_S the sibling relative risk, see Supplementary Methods for derivation). The information from a single affected sibling, for instance, predicts Crohn's disease with AUC of 0.56 (10). Prediction based on molecular measurements of genotype began with common loci of unusually large effect, such as the *HLA* effect in autoimmune diseases like type 1 diabetes [which alone gives an AUC of 0.85 (11)], rheumatoid arthritis and lupus or the effect of *APOE* in Alzheimer's. Many of these loci were identified in the pre-GWAS era by linkage studies, which further mark them as exceptions in complex disease genetics. Rare high penetrance mutations (such as *BRCA1* and *BRCA2* for breast cancer), while obviously very important to the families who carry them, are of surprisingly little value in population-level prediction (12,13). *BRCA1/2* population screening has an AUC of only 0.52, assuming a mutation frequency of 1% (14), a penetrance of 50% (15) and a breast cancer lifetime risk of 12% (16) (see Supplementary Methods for derivation).

Hundreds of GWAS and ever-larger meta-analyses have discovered a lengthening list of variants associated with complex disease. Figure 1 shows the AUC of predictors

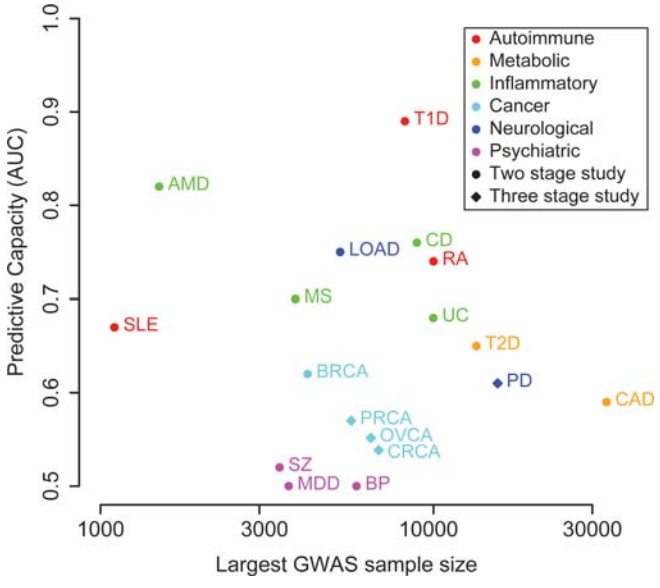

Figure 1. The predictive accuracy of variants discovered by GWAS, as a function of the effective sample size $[=2/(1/N_{\text{case}}+1/N_{\text{control}})]$, adjusted for the number of stages in the study (three stage studies have a smaller fraction of samples with GWAS data, and thus have lower power). Risk prediction is performed using logistic regression evaluated on data sets simulated from allele frequencies and odds ratios taken from replication data. PD: Parkinson's disease (45,46), AMD, age-related macular degeneration (47); T1D, type 1 diabetes (11); T2D, type 2 diabetes (48); UC, ulcerative colitis (49); CD, Crohn's disease (18,50); RA, rheumatoid arthritis (51); CAD, coronary artery disease (52); BRCA, breast cancer (53); LOAD, late-onset Alzheimer's disease (54,55); MS, multiple sclerosis (56); MDD, major depressive disorder (57); BP, bipolar disorder (58); SLE, systemic lupus erythematosus (59); SZ, schizophrenia (29); CRCA, colorectal cancer (60); PRCA, prostate cancer (61); OVCA, ovarian cancer (62,63).

based on the current genetic knowledge of 18 diseases. Several of these have rapidly improved the prospects of genetic prediction via GWAS. For example, good genetic prediction of age-related macular degeneration was quickly enabled by multiple large-effect variants identified by relatively small GWAS. Another notable GWAS success story (17,18), Crohn's disease, can be reasonably well predicted by a large number of weak effects. Note that the range of AUCs for these diseases is very similar to the range found in classical prediction (1,19–22).

The wide spectrum of AUC values shown in Figure 1 is attributable to a number of factors. Sample size and study design (such as the 'three stage' designs used in several cancer studies) play important roles in the process of variant discovery which feeds into risk prediction. Clinical heterogeneity and the complexity of affected tissues likely contribute to the recalcitrance of psychiatric illness to genetic prediction (and GWAS more generally). Highly heritable diseases are unsurprisingly usually easier to predict [a review of heritabilities is in Wray *et al*. (23)]. These principles are exemplified in Crohn's disease, which has been subject to large GWAS, is more heritable than the etiologically similar (but harder to predict) ulcerative colitis and has a definitive clinical diagnosis. Other differences are harder to explain using any of these arguments, such as the significantly greater predictive accuracy of type 2 diabetes compared with coronary artery disease, which is more heritable and has been subjected to larger GWAS meta-analyses.

GWAS-based predictions can be further improved by returning to first principles and incorporating family history conditional on genotype at known loci (10,24). For instance, AUC from a [now out-dated (18)] list of 30 Crohn's loci is 0.71 (10), much higher than the 0.56 mentioned above for family history alone, but less than the AUC of 0.74 for family history and GWAS combined. Since collecting family history is an important part of standard medical assessment, and can contribute independent genetic information beyond GWAS variants, it seems sensible to incorporate it into genetic risk prediction.

For diseases where non-genetic prediction is already well established, it is important to evaluate the information added by genetic loci. Clearly, if classical prediction is strong and genetic prediction is weak, little additional value is added. Furthermore, GWAS risk factors are not necessarily independent of the classical predictors. For instance, if a risk variant increases the risk of developing a disease through increasing the level of a blood biomarker, and that blood biomarker is part of the classical test, then the genetic factor will substantially increase the predictive accuracy. Even this example is more complex than it may appear, as genetic variants that influence lipid levels do grant some increase in prediction even when lipid levels are measured (25), likely due to the fact that they can predict lipid production over longer time periods than a blood lipid measurement at a single time point can. Prospective studies are required to disentangle these issues, and recent examples run the gamut from success stories, such as using common variants to increase the AUC of risk prediction from 0.76 to 0.83 in age-related macular degeneration (21), to negligible improvements for prediction of metabolic diseases (1,26).

MODELLING CONCERNS

The predictive accuracies mentioned in the previous section are all based on replicated risk variants discovered by GWAS to date. However, established GWAS loci typically explain only a small fraction of the heritability of complex diseases [an observation known as 'missing heritability' (27)]. Regardless of the explanation for this phenomenon, it raises questions about broader methods of risk prediction using the entire genome. A number of approaches have been developed to address this issue, but the answers they provide depend on the assumptions inherent in different models of as-yet undiscovered genetic risk.

Recent studies have estimated that a large proportion of heritability can be explained by common variation, based on identity-by-descent sharing of distantly related individuals (28). It has been shown that ∼3% of variance (corresponding to an AUC of 0.65) (23) in schizophrenia risk can be explained by a polygenic model, including a large number of loci that did not achieve genome-wide significance (29). Various attempts have been made to use highly polygenic risk scores based on these non-significant loci for prediction in different diseases, with varying degrees of success (30–32).

A natural extension of considering genome-wide risk prediction is the theoretical accuracy one might achieve if the genetic architecture of a disease were completely described. Some diseases might be difficult to predict due to poor current understanding of the underlying genetics, whereas others might never be tractable to genetic prediction. Epidemiological estimates of heritability (23) can be used to create such theoretically complete risk models. Three models have been proposed (33), each of which corresponds to a different assumption about the distribution of disease probability in the population. One of these models (the log model) is an analytically tractable but relatively unrealistic, assuming that probabilities are log-normally distributed, which can create disease probabilities greater than 1. The other two models are more complex but also more realistic, with each making different but apparently equally valid assumptions. The logit (or logistic) model assumes that odds ratios are log-normally distributed, and as a result has similar properties to, and is easy to integrate with, the logistic regression techniques used in GWAS. The probit (or liability threshold) model is a generalization of the variance component models used in quantitative trait modelling, and assumes a continuous distribution of a disease phenotype (called the liability) in the population, with heritable and non-heritable components. An individual who develops the disease is assumed to have a liability value above a certain threshold. Clayton (11) derived expressions for the AUC and the ROC curves given the log model, which showed surprisingly small AUCs for most scenarios. However, Wray and Goddard (33)'s calculations based on the probit model showed very high AUCs for a similar range of heritabilities. Similar expressions cannot be easily derived for the logistic model, although they can be calculated numerically.

Figure 2 shows the predicted ROC curves for diseases with a prevalence of $K = 1/200$ and $K = 1/20$, and a sibling relative risk of $\lambda_S = 9$ and $\lambda_S = 3$ for the three models. These values represent conservative parameter estimates for uncommon

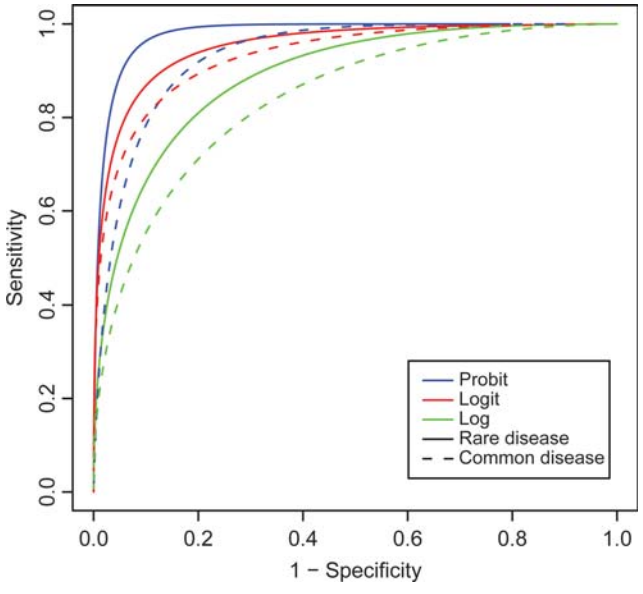

Figure 2. The ROC curves for the log, logit and probit models of disease risk for a rare disease with a prevalence $K = 1/200$ and sibling relative risk of $\lambda_S = 9$, and a common disease with $K = 1/20$ and $\lambda s = 3$, given that has been explained. The corresponding AUCs are 0.89, 0.96 and 0.98, respectively, for the rare disease, and 0.84, 0.93 and 0.93 for the common disease.

diseases, such as Crohn's disease or type 1 diabetes, and more common diseases such as cardiovascular disease. For the rarer disease, all the models give divergent answers, with the probit model giving an AUC of 0.98, a logit model an AUC of 0.96 and the log model an AUC of 0.89. For the common disease, the logit and probit models agree on an AUC of 0.93, although with a different sensitivity–specificity trade-off, and the log model gives a much lower AUC of 0.84. Part of this discrepancy is explained by an assumption in the log model that $K\lambda_s << 1$, to avoid troublesome risk probabilities >1 (33). In practice, we have found that the log model is inaccurate if $K\lambda_s > 0.01$ (Supplementary Material, Fig. S1) and therefore agree with Wray and Goddard (33) that this model should not be used for common diseases. A plausible maximum AUC for rare diseases therefore likely lies between 0.96 and 0.98, and common diseases around 0.93, as predicted by the logit and probit models. If future genetic studies are able to account for a significant proportion of the heritability of common diseases, then genetic prediction has the potential to become much more powerful.

THE FUTURE OF GENETIC RISK PREDICTION

GWAS results available today allow prediction for a large number of common diseases, with accuracies ranging from slight to moderately high, similar to the range of predictive accuracies found in classical prediction (although the specific diseases that can be predicted well differ by method). Indeed, genetic prediction has already been incorporated into clinical practice in situations where relatively rare, but powerful predictors have been discovered, such as HLA-B*701 mediated hypersensitivity reaction to the antiretroviral abacavir (34). In addition, larger meta-analyses and future sequencing studies will identify further risk variants, possibly including lower frequency variants of large effect size. Such studies could bring risk prediction ever closer to the high-accuracy theoretical predictions described above.

Irrespective of predictive power, there are a number of benefits of such genetic prediction over classical alternatives. For instance, unlike classical risk prediction, genetic risk prediction is highly stable over time, as a person's genetic sequence is essentially constant throughout their life. Some currently used clinical biomarkers, in contrast, are powerful predictors of disease risk in the near term but less valuable in assessing lifetime risk. A study in type 2 diabetes (35) showed that the AUC for clinical predictors declines from 0.76 to 0.64 as mean follow-up time increases from 16 to 28 years, but genetic prediction improves from 0.57 to 0.62 over the same timescale (see also the lipid level example above). This allows risk prediction to be performed on a much longer time scale than is currently plausible. Such stable risk stratification could be especially important when the proposed interventions are more effective if started at an early age, or continued over a long time period.

The inherent value of any disease predictor, however, is a function not just of predictive power, but also the cost and invasiveness of the prediction procedure, and the cost and effectiveness of the interventions available. This balance of practicality and predictive power is central to the incorporation of any predictor to routine medical practice, including genetics. Genetic risk prediction is currently not straightforward, as it requires obtaining a blood or saliva sample and ordering a bespoke genotyping assay for a locus of interest. However, the marginal cost of prediction could be very low in the future if full genome sequences are available and infrastructure is developed to interpret it. The continuing plunge in the cost of sequencing individual genomes (36) is making this scenario increasingly likely. Cheap and readily available genome sequencing is already being used in clinical genetics practice to diagnose genetic disease (37,38), to guide cancer treatment (39) and as a cost-effective form of carrier testing (40). Once a patient's genome is on file, risk prediction can essentially be performed for free and can then be used to inform diagnostics, screening and preventative measures in an automated way.

Attaching a patient's genome to an electronic medical record will enable a variety of prediction scenarios dependent on disease aetiology, prevalence and prevention and treatment options. For some diseases, such as age-related macular degeneration, the high accuracy of genetic prediction could be applied to entire populations so that regular ophthalmological examinations for at-risk individuals could allow early detection and treatment of this degenerative disease. For rarer diseases, population screening is less useful due to low positive predictive values, but genetic prediction could be applied when patients present with early symptoms. For instance, while the rarity of Crohn's disease results in a low positive predictive value in the population, genetic data could aid in the diagnosis of a patient who presents with early symptoms such as abdominal pain, diarrhoea and weight loss. Complex risk prediction also interacts with clinical genetics because some diseases, such as diabetes,

have similar presentation of both complex and monogenic forms (41). An accurate prediction of either must take into account the possibility of different underlying genetic models: conditional on disease symptoms, the probability of having a monogenic mutation increases as complex disease risk decreases.

Widespread clinical incorporation of genetic risk prediction will also require development of sophisticated infrastructure to perform prediction and deliver interpretable results to patients and health care professionals. Such systems will need to be designed both to provide evidence at the bedside to doctors, and to enable communication of these results to patients in a maximally beneficial way. Furthermore, complete genome sequences will inevitably lead to potentially worrisome incidental findings, such as *APOE* homozygosity for a high Alzheimer's risk allele. Procedures will need to be put in place to handle such discoveries in a consistent manner which avoids unintended psychological harm. While early reports have not shown drastic behavioural consequences of genetic testing (42–44), a more detailed understanding of the psychological response to genetic risk prediction, and how best to communicate such predictions to patients, is required. In the longer term, clinical trials will be required to learn how effective these applications are at improving outcomes, as well as how much of a cost burden is associated with them.

The promise of risk prediction has (sometimes ominously) hovered over the study of disease genetics since the initial sequencing of the human genome. Genetic risk prediction has never become as powerful as some early hype suggested it would be, but neither is it as useless as some detractors claim. Genetic risk prediction can already improve upon classical prediction, in some cases substantially so. However, as is true for classical predictors, the utility of genetic risk prediction is dependent not just on predictive accuracy, but also on cost and the ability of clinicians and patients to effectively use this information. The falling cost of whole-genome sequencing will drive the marginal cost of prediction lower and lower, but further progress in gene-mapping research, infrastructure and medical practice will be needed to take full advantage of genetic risk prediction.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors would like to thank Katherine Morley and Yang Luo for comments on the manuscript.

Conflict of Interest statement. None declared.

FUNDING

We also thank the Wellcome Trust for funding this work (WT089120/Z/09/Z). Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

REFERENCES

1. Buijsse, B., Simmons, R.K., Griffin, S.J. and Schulze, M.B. (2011) Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. *Epidemiol. Rev.*, **33**, 46–62.
2. Pencina, M.J., D'Agostino, R.B., Larson, M.G., Massaro, J.M. and Vasan, R.S. (2009) Predicting the 30-year risk of cardiovascular disease: the Framingham heart study. *Circulation*, **119**, 3078–3084.
3. Lu, K., Kauff, N., Powell, C.B., Chen, L.M., Cass, I., Lancaster, J., Karlan, B., Berchuck, A. and Mutch, D. (2009) ACOG Practice Bulletin No. 103: hereditary breast and ovarian cancer syndrome. *Obstet. Gynecol.*, **113**, 957–966.
4. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
5. Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.*, **10**, 43–55.
6. Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C. and Daly, M.J. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.*, **7**, e1001273.
7. Seliya, N., Khoshgoftaar, T.M. and Van Hulse, J. (2009) A study on the relationships of classifier performance metrics In *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*, pp. 59–66. IEEE Computer Society, Washington, DC, USA. doi:10.1002/gepi.20600.
8. Ferri, C., Hernández-Orallo, J. and Modroiu, R. (2009) An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.*, **30**, 27–38.
9. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.
10. Ruderfer, D.M., Korn, J. and Purcell, S.M. (2010) Family-based genetic risk prediction of multifactorial disease. *Genome Med.*, **2**, 2.
11. Clayton, D.G. (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.*, **5**, e1000540.
12. Newman, B., Mu, H., Butler, L.M., Millikan, R.C., Moorman, P.G. and King, M.C. (1998) Frequency of breast cancer attributable to BRCA1 in a population-based series of American women. *JAMA*, **279**, 915–921.
13. Anglian Breast Cancer Study Group. (2000) Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br. J. Cancer*, 83, 1301–1308.
14. Risch, H.A., McLaughlin, J.R., Cole, D.E., Rosen, B., Bradley, L., Fan, I., Tang, J., Li, S., Zhang, S., Shaw, P.A. and Narod, S.A. (2006) Population BRCA1 and BRCA2 mutation frequencies and cancer penetrances: a kin-cohort study in Ontario, Canada. *J. Natl Cancer Inst.*, **98**, 1694–1706.
15. Chen, S. and Parmigiani, G. (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.*, **25**, 1329–1333.
16. Howlader, N, Noone, A.M, Krapcho, M, Neyman, N, Aminou, R, Waldron, W, Altekruse, SF, Kosary, CL, Ruhl, J and Tatalovich, Z (2011) SEER cancer statistics review, 1975–2008. http://seer.cancer.gov/csr/1975_2008/.
17. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat. Genet.* **40**, 955–962.
18. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.*, **42**, 1118–1125.
19. Lloyd-Jones, D.M., Liu, K., Tian, L. and Greenland, P. (2006) Narrative review: assessment of C-reactive protein in risk prediction for cardiovascular disease. *Ann. Intern. Med.*, **145**, 35–42.
20. Cassidy, A., Myles, J.P., van Tongeren, M., Page, R.D., Liloglou, T., Duffy, S.W. and Field, J.K. (2008) The LLP risk model: an individual risk prediction model for lung cancer. *Br. J. Cancer*, **98**, 270–276.
21. Seddon, J.M., Reynolds, R., Maller, J., Fagerness, J.A., Daly, M.J. and Rosner, B. (2009) Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest. Ophthalmol. Vis. Sci.*, **50**, 2044–2053.

22. Wacholder, S., Hartge, P., Prentice, R., Garcia-Closas, M., Feigelson, H.S., Diver, W.R., Thun, M.J., Cox, D.G., Hankinson, S.E., Kraft, P. *et al.* (2010) Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.*, **362**, 986–993.
23. Wray, N.R., Yang, J., Goddard, M.E. and Visscher, P.M. (2010) The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.*, **6**, e1000864.
24. So, H.C., Kwan, J.S., Cherny, S.S. and Sham, P.C. (2011) Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.*, **88**, 548–565.
25. Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L. *et al.* (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.*, **358**, 1240–1249.
26. Companioni, O., Rodríguez Esparragón, F., Fernández-Aceituno, A.M. and Rodríguez Pérez, J.C. (2011) Genetic variants, cardiovascular risk and genome-wide association studies. *Rev. Esp. Cardiol.*, **64**, 509–514.
27. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
28. Lee, S.H., Wray, N.R., Goddard, M.E. and Visscher, P.M. (2011) Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, **88**, 294–305.
29. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F. and Sklar, P. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
30. Kang, J., Kugathasan, S., Georges, M., Zhao, H. and Cho, J.H. (2011) Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum. Mol. Genet.*, **20**, 2435–2442.
31. Wei, Z., Wang, K., Qu, H.Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J.T., Chiavacci, R. *et al.* (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.*, **5**, e1000678.
32. Machiela, M.J., Chen, C.Y., Chen, C., Chanock, S.J., Hunter, D.J. and Kraft, P. (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epidemiol.*, **35**, 506–514.
33. Wray, N.R. and Goddard, M.E. (2010) Multi-locus models of genetic risk of disease. *Genome Med.*, **2**, 10.
34. Lalonde, R.G., Thomas, R., Rachlis, A., Gill, M.J., Roger, M., Angel, J.B., Smith, G., Higgins, N. and Trottier, B. (2010) Successful implementation of a national HLA-B*5701 genetic testing service in Canada. *Tissue Antigens*, **75**, 12–18.
35. Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., Berglund, G., Altshuler, D., Nilsson, P. and Groop, L. (2008) Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.*, **359**, 2220–2232.
36. Wetterstrand, K.A. (2011) DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. http://www.genome.gov/sequencingcosts/.
37. Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D., Bonacci, B.B., Decker, B., Serpe, J.M., Dasu, T., Tschannen, M.R., Veith, R.L. *et al.* (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.*, **13**, 255–262.
38. Rios, J., Stein, E., Shendure, J., Hobbs, H.H. and Cohen, J.C. (2010) Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum. Mol. Genet.*, **19**, 4313–4318.
39. Link, D.C., Schuettpelz, L.G., Shen, D., Wang, J., Walter, M.J., Kulkarni, S., Payton, J.E., Ivanovich, J., Goodfellow, P.J., Le Beau, M. *et al.* (2011) Identification of a novel tp53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *JAMA*, **305**, 1568–1576.
40. Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., Mudge, J., Langley, R.J., Zhang, L., Lee, C.C., Schilkey, F.D. *et al.* (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.*, **3**, 65ra4.
41. Molven, A. and Njølstad, P.R. (2011) Role of molecular genetics in transforming diagnosis of diabetes mellitus. *Expert Rev. Mol. Diagn.*, **11**, 313–320.
42. Green, R.C., Roberts, J.S., Cupples, L.A., Relkin, N.R., Whitehouse, P.J., Brown, T., Eckert, S.L., Butson, M., Sadovnick, A.D., Quaid, K.A. *et al.* (2009) Disclosure of APOE genotype for risk of Alzheimer's disease. *N. Engl. J. Med.*, **361**, 245–254.
43. Marteau, T.M., French, D.P., Griffin, S.J., Prevost, A.T., Sutton, S., Watkinson, C., Attwood, S. and ollands, G.J. (2010) Effects of communicating DNA-based disease risk estimates on risk-reducing behaviours. *Cochrane Database Syst. Rev.*, **10**, CD007275.
44. Bloss, C.S., Schork, N.J. and Topol, E.J. (2011) Effect of direct-to-consumer genomewide profiling to assess disease risk. *N. Engl. J. Med.*, **364**, 524–534.
45. International Parkinson's Disease Genomics Consortium and Wellcome Trust Case Control Consortium 2. (2011) A two-stage meta-analysis identifies several new Loci for Parkinson's disease. *PLoS Genet.*, **7**, e1002142.
46. Nalls, M.A., Plagnol, V., Hernandez, D.G., Sharma, M., Sheerin, U.M., Saad, M., Simon-Sanchez, J., Schulte, C., Lesage, S., Sveinbjornsdottir, S. *et al.* (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*, **377**, 641–649.
47. Chen, W., Stambolian, D., Edwards, A.O., Branham, K.E., Othman, M., Jakobsdottir, J., Tosakulwong, N., Pericak-Vance, M.A., Campochiaro, P.A., Klein, M.L. *et al.* (2010) Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl Acad. Sci. USA*, **107**, 7401–7406.
48. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G. *et al.* (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.*, **42**, 579–589.
49. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A. *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.*, **43**, 246–252.
50. Yazdanyar, S., Weischer, M. and Nordestgaard, B.G. (2009) Genotyping for NOD2 genetic variants and Crohn disease: a metaanalysis. *Clin. Chem.*, **55**, 1950–1957.
51. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A., Zhernakova, A., Hinks, A. *et al.* (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.*, **42**, 508–514.
52. Schunkert, H., Konig, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F., Barbalic, M., Gieger, C. *et al.* (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.*, **43**, 333–338.
53. Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S. *et al.* (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.*, **42**, 504–517.
54. Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A. *et al.* (2009) Genome-wide association study identifies variants at CLU and PICALM associated with alzheimer's disease. *Nat. Genet.*, **41**, 1088–1093.
55. Corneveaux, J.J., Myers, A.J., Allen, A.N., Pruzin, J.J., Ramirez, M., Engel, A., Nalls, M.A., Chen, K., Lee, W., Chewning, K. *et al.* (2010) Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. *Hum. Mol. Genet.*, **19**, 3295–3301.
56. De Jager, P.L., Jia, X., Wang, J., de Bakker, P.I., Ottoboni, L., Aggarwal, N.T., Piccio, L., Raychaudhuri, S., Tran, D., Aubin, C. *et al.* (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.*, **41**, 776–782.
57. Shyn, S.I., Shi, J., Kraft, J.B., Potash, J.B., Knowles, J.A., Weissman, M.M., Garriock, H.A., Yokoyama, J.S., McGrath, P.J., Peters, E.J. *et al.* (2009) Novel loci for major depression identified by genome-wide association study of Sequenced Treatment Alternatives to Relieve Depression and meta-analysis of three studies. *Mol. Psychiatr.*, **16**, 202–215.
58. Scott, L.J., Muglia, P., Kong, X.Q., Guan, W., Flickinger, M., Upmanyu, R., Tozzi, F., Li, J.Z., Burmeister, M., Absher, D. *et al.* (2009)

Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl Acad. Sci. USA*, **106**, 7501–7506.
59. Harley, J.B., Alarcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K. *et al.* (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.*, **40**, 204–210.
60. Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S. *et al.* (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.*, **40**, 1426–1435.
61. Eeles, R.A., Kote-Jarai, Z., Al Olama, A.A., Giles, G.G., Guy, M., Severi, G., Muir, K., Hopper, J.L., Henderson, B.E., Haiman, C.A. *et al.* (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
62. Goode, E.L., Chenevix-Trench, G., Song, H., Ramus, S.J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R.A., Larson, M.C., Kjaer, S.K. *et al.* (2010) A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat. Genet.*, **42**, 874–879.
63. Song, H., Ramus, S.J., Tyrer, J., Bolton, K.L., Gentry-Maharaj, A., Wozniak, E., Anton-Culver, H., Chang-Claude, J., Cramer, D.W., DiCioccio, R. *et al.* (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat. Genet.*, **41**, 996–1000.