and the RCTs they are based on are an accepted dimension of modern medical practice. However, researchers, funding agencies, ethics committees, medical journals, regulators and, above all, the pharmaceutical companies themselves should take a critical look at clinical trial design to maximise relevance and generalisability.

Correspondence to: Professor I Town, University of Canterbury, PB 4800, Christchurch 8001, New Zealand; ian.town@canterbury.ac.nz

Competing interests: None declared.

REFERENCES

1 **Rothwell PM**. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;**365**:82–93.
2 **Charlton BG**, Miles A. The rise and fall of EBM. Q J Med 1998;**91**:371–4.
3 **Global Initiative for Asthma**. Global strategy for asthma management and prevention, NHI Publication Number 02-2659, 2005. www.ginaasthma.org (accessed 15 Jan 2007).
4 **Travers J**, Marsh S, Williams M, et al. External validity of randomised controlled trials in asthma: to whom the results of the trials apply? Thorax 2007;**62**:219–23.
5 **Marsh S**, Aldington S, Williams MU, et al. Physiological associations of computerised tomography lung density: a factor analysis. Int J COPD 2006;**1**:181–7.
6 **Garfield FB**, Garfield JM. Clinical judgement and clinical practice guidelines. Int J Technol Assess Health Care 2000;**16**:1050–60.

Multifactor-dimensionality reduction

....

Looking for a bit of co-action?

John D Blakey

....

MDR, a primary tool for exploratory analyses

Despite the wealth of evidence that many common diseases have a strong heritable component, genetic association studies have provided disappointingly little insight into their pathophysiology. One reason for this is that, buoyed by the success of identifying underlying variants for over 1000 uncommon conditions, researchers have largely continued to ask simple questions of more complex disorders. However, in this issue of Thorax, Park et al[1] (*see page 265*) present an association study that uses one of a number of new statistical methods that have the potential to produce more biologically relevant associations.

A SHORT HISTORY LESSON

For much of the last century, the suggestion of a genetic association study as we now know them might have raised a few eyebrows, regardless of the leap in technology required. "Natura non facit saltum" (nature does not make leaps) was a favourite aphorism of Darwin and greatly influenced mathematical models of adaptation.[2] Accepting such models, which comprise a myriad of elements, each contributing a small overall effect, implies that genetic association studies for complex disorders are fruitless in realistic population sizes. However, since the 1980s experimental evidence has accrued to challenge this notion. In a number of species, individual genetic variants have been shown to account for a large proportion of inherited trait variation,[3][4] although these are markedly in the minority.[5][6]

STUDYING SINGLE POLYMORPHISMS

A plethora of association studies have attempted to locate these few highly influential loci for common, complex traits such as asthma,[7] but there has been little reproducible success. A great deal has been written with a view to improving the design of such studies, including in this journal,[8] but mention of underlying genetic complexity is often made only to support the use of intermediate phenotypes. Under biologically plausible models,[9–13] the chance of a polymorphism in a candidate gene exerting a large effect in isolation remains low, especially in common, heterogeneous conditions. Large populations are therefore required to reliably detect the great majority of contributory effects. By way of example, consider *IL13*, the candidate gene with probably the greatest support[7] for association with atopy. Variation in this gene has recently been studied in relation to IgE levels in more than 3000 adult individuals from the general population. Although a strong association (p = 0.00002) was seen, polymorphisms explained <0.6% of the phenotypic variance.[13] One wonders how many effects of comparable magnitude have not been detected in preliminary studies and so have not gone on to be tested in large populations. Given this low prior probability of finding true associations in most genetic association studies, their "significant" findings are arguably far more likely to be false positives.[14]

STUDYING MULTIPLE LOCI

Several candidate loci considered together are more likely than a single polymorphism to explain enough of the inherited variability of a trait to be reliably detectable. This simple summation of effects will not hold in all situations, given the great complexity of biological processes. However, considering loci concurrently also affords the chance to detect polymorphisms that together overcome the genetic buffering that stabilises phenotypes against the potentially detrimental effects of mutation.[15][16] Loci exhibit epistasis if their collective effect on a trait is greater than that anticipated given their individual influences (which may be negligible in isolation). The number of these epistatic interactions detected in experimental models appears similar to or larger than the number of loci with independent, additive effects,[17] although this can only be a guide to the situation in humans. An apparent paradox therefore exists in that we have a convincing argument to study many loci concurrently, yet most studies have tested association per individual polymorphism or reconstructed haplotype.

LIMITATIONS OF STANDARD APPROACHES

The primary limitation of a multilocus study has previously been technological, but with advances in genotyping technology and cost, and the explosion of in silico resources, statistical hurdles now curtail such endeavours. Simply applying the statistical test that was used for a single pair of alleles to all genotype combinations has the advantage of being easy to perform (if time consuming), but there is no consensus on interpretation of the results of multiple tests that are not fully independent. Standard regression models can accommodate interaction terms, but this leads to an exponential increase in terms to be estimated. There is a resulting commensurate reduction in informative data for each parameter,

therefore errors are large and terms may be erroneously excluded from a model. Both regression modelling and repetitive testing require inferences from the user as to the underlying genetic model.

MULTIFACTOR-DIMENSIONALITY REDUCTION

One method conceived to deal with some of these problems is that used by Park et al[1] in this edition of the journal: Multifactor-dimensionality reduction (MDR). MDR and the related technique of combinatorial partitioning are two of a number of new techniques that seek to handle high-dimensional data to uncover complex relationships, free from assumptions of the genetic model.

The MDR procedure is explained in detail elsewhere,[18] including in the article to which this editorial relates.[1] The approach has three key components: First, multilocus genotype data are collapsed to a single variable with two levels (high and low risk) based on a predetermined threshold for the case-control ratio (eg, >1). Second, the pattern of risk for each possible multilocus genotype is derived from a partition of the data (eg, nine tenths): the "training set". The predictive ability of this pattern is tested in the remaining data. Third, the validity of the best predictor is quantified by the repeated division of the dataset into training and testing sets. The probability of obtaining a classifier of such accuracy and validation consistency can be estimated using a permutation test.

Although no technique is a substitute for adequate subject recruitment, one strength of such approaches is their power to detect epistasis in relatively small populations (hundreds of individuals).[19] However, power appears to be markedly affected in the presence of phenocopy or genetic heterogeneity, and awaits formal assessment for three or more polymorphisms and in other scenarios such as gene–environment interactions or epistasis in the presence of main effects. For these reasons and others (eg, the uncertain relationship between α level and linkage disequilibrium), currently MDR is primarily a tool for exploratory analyses.

Several other, complementary, types of machine learning are emerging and evolving, driven in part by the increase in available microarray data. Consistent results from such analyses of multiple genetic and environmental factors will soon emerge, but each will still require validation with a focused hypothesis in a very large study population. A step toward this has recently been taken with the comparison of meta-analysis results for the β_2 adrenoceptor and asthma with findings in more than 8000 members of the 1958 British birth cohort.[20]

CONCLUSION

Park et al[1] are to be commended on using a new technique to consider the relationship between an intermediate phenotype and both single and multilocus genotypes in a large population. Furthermore, their use of logistic regression to confirm their findings of a synergistic action of polymorphisms in vascular endothelial growth factor 2 and tumour necrosis factor α highlights the use of complementary approaches to seek and quantify association. Clearly no firm inference can be made from a single report, but a clear hypothesis has been generated for a replication study and to guide in vitro studies.

REFERENCES

1 **Park H-W**, Shin E-S, Lee J-E, et al. Multilocus analyses of atopy in Korean children using multifactor-dimensionality reduction. Thorax 2007;**62**:265–9.
2 **Fisher R**. The genetical theory of natural selection: a complete variorum edition. Oxford: Oxford University Press, 2000.
3 **Mackay TF**. The genetic architecture of quantitative traits: lessons from Drosophila. Curr Opin Genet Dev 2004;**14**:253–7.
4 **Shapiro MD**, Marks ME, Peichel CL, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 2004;**428**:717–23.
5 **Orr HA**. The genetic theory of adaptation: a brief history. Nat Rev Genet 2005;**6**:119–27.
6 **Farrall M**. Quantitative genetic variation: a post-modern view. Hum Mol Genet 2004;**13**(Spec No 1):R1–7.
7 **Ober C**, Hoffjan S. Asthma genetics 2006: the long and winding road to gene discovery. Genes Immun 2006;**7**:95–100.
8 **Hall IP**, Blakey JD. Genetic association studies in Thorax. Thorax 2005;**60**:357–9.
9 **Orr HA**. The evolutionary genetics of adaptation: a simulation study. Genet Res 1999;**74**:207–14.
10 **Kimura M**, Ota T. On some principles governing molecular evolution. Proc Natl Acad Sci USA 1974;**71**:2848–52.
11 **Kauffman S**, Levin S. Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol 1987;**128**:11–45.
12 **Gillespie JH**. Natural selection and the molecular clock. Mol Biol Evol 1986;**3**:138–55.
13 **Maier LM**, Howson JM, Walker N, et al. Association of IL13 with total IgE: evidence against an inverse association of atopy and diabetes. J Allergy Clin Immunol 2006;**117**:1306–13.
14 **Sterne JA**, Davey Smith G. Sifting the evidence: what's wrong with significance tests? BMJ 2001;**322**:226–31.
15 **Segre D**, Deluna A, Church GM, et al. Modular epistasis in yeast metabolism. Nat Genet 2005;**37**:77–83.
16 **Stearns SC**. Progress on canalization. Proc Natl Acad Sci USA 2002;**99**:10229–30.
17 **Malmberg RL**, Mauricio R. QTL-based evidence for the role of epistasis in evolution. Genet Res 2005;**86**:89–95.
18 **Moore JH**. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 2004;**4**:795–803.
19 **Ritchie MD**, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 2003;**24**:150–7.
20 **Hall IP**, Blakey JD, Al Balushi KA, et al. Beta2-adrenoceptor polymorphisms and asthma from childhood to middle age in the British 1958 birth cohort: a genetic association study. Lancet 2006;**368**:771–9.