

Genetics and population analysis

Inferring causal relationships among intermediate phenotypes and biomarkers: a case study of rheumatoid arthritis

Wentian Li^{1,*}, Mingyi Wang², Patricia Irigoyen¹ and Peter K. Gregersen¹

¹The Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, NY, USA and ²College of Computer Science, Zhejiang University, Hangzhou, China

Received on February 8, 2006; revised and accepted on March 13, 2006

Advance Access publication March 21, 2006

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Genetic association analysis is based on statistical correlations which do not assign any cause-to-effect arrows between the two correlated variables. Normally, such assignment of cause and effect label is not necessary in genetic analysis since genes are always the cause and phenotypes are always the effect. However, among intermediate phenotypes and biomarkers, assigning cause and effect becomes meaningful, and causal inference can be useful.

Results: We show that causal inference is possible by an example in a study of rheumatoid arthritis. With the help of genotypic information, the shared epitope, the causal relationship between two biomarkers related to the disease, anti-cyclic citrullinated peptide (anti-CCP) and rheumatoid factor (RF) has been established. We emphasize the fact that third variable must be a genotype to be able to resolve potential ambiguities in causal inference. Two non-trivial conclusions have been reached by the causal inference: (1) anti-CCP is a cause of RF and (2) it is unlikely that a third confounding factor contributes to both anti-CCP and RF.

Contact: wli@nslj-genetics.org

1 INTRODUCTION

In all statistical textbooks, it is clearly stated that statistical correlation should not be equated to causal correlation. If two variables are both caused by the third variable, it does not imply these two correlated variables are in a cause–effect relationship (e.g. wrinkles and cancer risk are both increased with age, but wrinkles do not cause the cancer risk, nor does cancer risk lead to wrinkles). Recent developments in causal inference or causal statistics makes the assignment of cause and effect possible, if the third variable is available and information on conditional correlation can be obtained (Cooper, 1997; Spirtes *et al.*, 2000; Pearl, 2000; Silverstein *et al.*, 2000). Although the current study of causal inference is often within the field of computer science and machine learning, causality has been discussed in biology (Wright, 1921; Niles, 1922; Shipley, 2002), epidemiology (Koopman, 1977; Halloran and Struchiner, 1995; Robin *et al.*, 2000), economics (Granger, 1969, 1980; Hoover, 2001), statistics (Rubin, 1974; Holland, 1986; Cox and Wermuth, 1996), among others, for many years.

Causal inference methods have been applied to microarray analysis and the construction of regulatory pathways (Yoo *et al.*, 2002; Chu *et al.*, 2003; Bay *et al.*, 2004; Xing and Van der Laan, 2005). But it has not yet been applied to genetic analysis. The reason is simple: the causal structure in genetic data is known, i.e. genotype is the cause, phenotype is the effect, and the reverse Lamarckian relationship has been proven to be unlikely. The situation is changed, however, when the two variables under investigation are both intermediate phenotypes, or biomarkers. Either one of the intermediate phenotypes can be produced upstream in a biochemical pathway, whereas another produced downstream. Then the upstream one can be considered as a cause, and the downstream one an effect.

Since the key idea in causal inference is the introduction of the third variable and the subsequent conditional correlation analysis (Dawid, 1979, 1980), a crucial question we ask is what this third variable should be when the two correlated variables are intermediate phenotypes. In this paper, we will show that only when the third variable is a genotype, is it possible that the cause–effect arrow be assigned unambiguously (though still not guaranteed). Interestingly, a similar choice of using a genotype as the third variable to exclude the possibility of reverse causality from the disease status to a risk factor was proposed 20 years ago (Katan, 1986).

Correlation analysis of variables with binary states is based on 2×2 contingency tables. Conditional correlation analysis is based on $2 \times 2 \times 2$ tables, because there will be two 2-by-2 tables, one for each stratified state of the third variable. To fill eight cells with a reasonable number of sample counts, a larger dataset is required. Genetic data, human genetic data in particular, tend to be smaller, owing the difficulties and costs in collecting samples. Fortunately, we have in our possession a large dataset in the study of rheumatoid arthritis (RA) collected under the North American Rheumatoid Arthritis Consortium (NARAC) initiative (Gregersen, 1998).

About 1700 rheumatoid arthritis patients were typed with two biomarkers that can be used for diagnosis of RA: anti-cyclic citrullinated peptide (anti-CCP) antibody (Schellekens *et al.*, 2000) and rheumatoid factor (RF), an antibody that binds other antibodies (Pope and McDuffy, 1979). Anti-CCP level is correlated with the RF level, and both are correlated with RA disease status. For simplicity, both anti-CCP and RF are partitioned into two states: positive or negative.

The third variable available is the genotype at the HLA-DRB1 locus within the major histocompatibility complex (MHC) region on human chromosome 6 (6p21.3) (Beck *et al.*, 1999). This locus

*To whom correspondence should be addressed.

contains the main risk factor known so far for RA (Stastny, 1978; Ollier and Thomson, 1992), and a collection of alleles in this locus that are associated with RA is called ‘shared epitope’ (SE) (Gregersen et al., 1987). Allele-wise, all HLA-DRB1 alleles can be partitioned into SE-positive and SE-negative ones. Genotype-wise, there are three possibilities: SE+/SE+, SE+/SE−, and SE−/SE−. For simplicity, we combine SE+/SE+ and SE+/SE− as one group, resulting in two states (SE+ and SE−) at the genotype level. This SE variable is also highly associated with the RA disease status.

We aim at finding a causal relationship between anti-CCP and RF biomarkers with the help of SE genotype, using the causal inference method proposed in (Cooper, 1997).

2 METHODS AND DATA

Correlation (association) and conditional correlation (association). The terms correlation and association are used interchangeably here. Since the variables we have are discretized into binary states, the correlation in each pair of variables can be analyzed by a 2×2 contingency table $(\{n_{ij}\}, i, j = 1, 2)$. The correlation strength is measured by odds ratio $[OR = n_{11}n_{22}/(n_{12}n_{21})]$. The significance of correlation is measured by the p -value in the Pearson’s χ^2 test.

For the conditional correlation analysis, there are two 2×2 contingency tables stratified by the state of the third variable. The X^2 statistic for each table can be calculated: X_{s1}^2 and X_{s2}^2 . The sum of the two: $X^2 = X_{s1}^2 + X_{s2}^2$ should follow the χ^2 distribution with two degrees of freedom, and the corresponding p -value can be calculated.

Other measures of correlation are also possible, but in general, these are all related. For example, the mutual information, $M = \sum_{ij} p_{ij} \log_2 p_{ij}/(p_i p_j)$, where $n = \sum_{ij} n_{ij}$, $p_{ij} = n_{ij}/n$, $p_i = \sum_j p_{ij}$, $p_j = \sum_i p_{ij}$, is a measure of correlation (Li, 1990). But mutual information multiplied by twice the sample size, $G^2 = 2nM$, is the G^2 -statistic, a likelihood ratio test statistic and G^2 and X^2 statistics are asymptotically equal (Agresti, 2002).

Cooper’s local causality discovery (LCD) rule. (Cooper, 1997) We assume that the three variables x, y, z under investigation do not form a causal relationship loop. Also, we assume that x is not caused by y and/or z , and a fourth hidden variable h is allowed as a potential confounding factor. Suppose x and y are correlated, y and z are correlated, but x and z are uncorrelated conditional on y , then the LCD rule states that only the following causal relationships between x, y, z are possible:

$$x \rightarrow y \rightarrow z \tag{1}$$

$$\begin{array}{c} h \\ \swarrow \searrow \\ x \quad y \rightarrow z \end{array} \tag{2}$$

$$\begin{array}{c} h \\ \swarrow \searrow \\ x \rightarrow y \quad z \end{array} \tag{3}$$

If x is not caused by h , then only Equation (1) is possible.

Cooper’s LCD rule was obtained by exhaustively listing all possible pairwise causal models (Cooper, 1997), but it can also be understood as follows. Since x and z are uncorrelated conditional on y , there should be no direct causal relationship between x and z , i.e. there is no arrow directly connecting x and z . The remaining relationship among x, y, z (for simplicity, we ignore the part that involves h) has to be chosen among the $x \leftrightarrow y \leftrightarrow z$ configuration. Each \leftrightarrow can be either \leftarrow or \rightarrow , so $x \leftrightarrow y \leftrightarrow z$ covers four possibilities. The relationship $x \leftarrow y \leftarrow z$ and $x \leftarrow y \rightarrow z$ violate our assumption that x is not caused by y or z . The relationship $x \rightarrow y \leftarrow z$ is not consistent with the fact that x and z are conditionally independent. The only possible relationship left is $x \rightarrow y \rightarrow z$.

Table 1. The number of samples in specific anti-cyclic citrullinated peptide (anti-CCP, high/low), rheumatoid factor (RF, high/low) and shared epitope (SE, yes/no) groups

Anti-CCP	RF	SE	No. samples
+	+	+	960
+	+	−	128
+	−	+	84
+	−	−	19
−	+	+	95
−	+	−	74
−	−	+	214
−	−	−	149

This is another form of representation of the $2 \times 2 \times 2$ contingency table.

The LCD rule is extended in Silverstein et al. (2000) by requiring that x and z are correlated unconditional on y (but uncorrelated conditional on y). With this extension there exist pairwise correlation in all three variable pairs, and it is the reason the rule is called the CCC rule allowing for hidden variables. In fact, this extension was recommended in Cooper (1997) by the consideration that the conditional independence between x and z could be falsely claimed owing to a lack of samples. By testing both conditional as well as unconditional correlations between x and z , we have more confidence that an insignificant correlation is not because of a small sample size. Note that the CCC rule in Silverstein et al. (2000) that does not allow for hidden variables is not discussed here, since the assumption for its use is violated in our example.

Data. The information on anti-CCP, RF, and SE status of 1723 Caucasian RA patients are obtained from (Irigoyen et al., 2005). The number of samples in each of the 8 strata are listed in Table 1.

3 RESULTS

The unconditional correlation between the three variables, anti-CCP, RF and SE, are all very strong, with ORs of 22.7 (anti-CCP and RF), 5.1 (anti-CCP and SE) and 2.9 (RF and SE). These ORs are all significant, proven either by the 95% confidence interval (CI) (if both the lower and upper bounds of 95% CI of OR are >1 , or both <1 , the correlation is significant at the 5% level), or by the p -values from the χ^2 test (Table 2).

Since we have a three-way correlation among three variables, and SE as a genotype (variable x) cannot be caused by intermediate phenotypes anti-CCP (y) and RF (z), the LCD rule can be applied. The conditional correlation analysis results are included in Table 2. The conditional correlations between anti-CCP and RF, and that between anti-CCP and SE are still strong and statistically significant (p -values are essentially zero). However, the conditional correlation between RF and SE is absent and not statistically significant: the summed X^2 statistic is 4.3, with p -value of 0.12.

Another way to measure the reduction of correlation from unconditional to conditional situation is to look at the ratio of ORs. When we use the geometric means of the ORs in two strata (OR_1 and OR_2 in Table 2), the ratios of ORs [conditional over unconditional, $(OR_{cond,1} OR_{cond,2})^{1/2}/OR_{uncond}$] are roughly 0.82, 0.83 and 0.43 for anti-CCP–RF, SE–anti-CCP pair and SE–RF pairs. Clearly, SE–RF pair loses their correlation the most by going from unconditional to conditional analysis.

Table 2. Testing of pairwise correlation unconditional or conditional on the third variable (anti-CCP and RF, with SE as the third variable; anti-CCP and SE, with RF as the third variable; RF and SE with anti-CCP as the third variable)

	Unconditional			Conditional			
	X^2	OR 95% CI	p -value	$X_{s1}^2 + X_{s2}^2 = X^2$	OR ₁ 95% CI	OR ₂ 95% CI	p -value
Anti-CCP and RF	661.7	22.7 (17.3, 29.8)	0	520.2 + 103.8 = 624.0	25.7 (18.5, 35.8)	13.6 (7.8, 23.7)	0
Anti-CCP and SE	190.7	5.1 (4.0, 6.5)	0	111.2 + 17.8 = 129.0	5.8 (4.1, 8.3)	3.1 (1.8, 5.3)	0
RF and SE	80.5	2.9 (2.3, 3.7)	10^{-19}	3.9 + 0.4 = 4.3	1.7 (1, 2.9)	0.9 (0.6, 1.6)	0.12

The correlation/association is tested by the X^2 test-statistic, using the χ^2 distribution with 1 (unconditional) or 2 (conditional) degrees of freedom. Also listed are the ORs and its 95% confidence intervals.

By Cooper's LCD rule, the three causal relationships as represented by Equations (1)–(3) are possible, all concluding that anti-CCP is a 'cause' of RF. If the hidden variable is an environmental factor or another intermediate phenotype, the causal relationships represented by Equations (2) and (3) are impossible. Then we have another conclusion that SE is a cause of anti-CCP.

To apply Cooper's LCD rule (extended version) requires all three variables to have pairwise unconditional correlations. If out of the three variables, two are genotypes and one is a phenotype, the correlation between the two genotypes are usually missing, unless the two genes are located nearby on the chromosome and are in linkage disequilibrium. One may re-acquire a correlation between the two genotypes conditional on the phenotype. This is the so-called CCU rule discussed in Silverstein *et al.* (2000). However, no new causal relationship will be learned as we had known that genotypes are causes and phenotypes are effects.

It is clear why the third variable has to be a genotype instead of another intermediate phenotype when inferring the causal relationship between two intermediate phenotypes. If the third variable is a phenotype, even when a pair of variable (x and z) loses correlation conditional on the the third variable (y), we can only infer $x \leftrightarrow y \leftrightarrow z$, but we will be unable to exclude causal models within the equivalent class to reach a unique model. Only when we require the x variable to have no cause among y and z , can the causal direction between y and z be unambiguously determined.

As all samples in our dataset are RA patients, it is a case-only design. The conclusion reached concerning the causal relationship between anti-CCP and RF may not apply to a control-only dataset. We note that if the association between SE and anti-CCP, SE and RF, SE and RF is absent in the control group, their presence in the case group is an indication of 'interaction' among them (Clayton and McKeigue, 2001).

4 DISCUSSION AND CONCLUSIONS

Our study is different from some recent applications of Bayesian network to the genotype–phenotype mapping (Rodin and Boerwinkle, 2005; Schadt *et al.*, 2005). Instead of examining many genes, their expressions and the connectivity of the gene network, we focus locally on only three variables. For gene network studies with expression levels of hundreds of genes, it is not guaranteed that any three genes are all pairwise correlated so that the extended version of LCD rule can be applied. Here, we choose the two intermediate phenotypes and one genotype that are most strongly correlated with the disease RA, and the application of

LCD rule is guaranteed. Also, obtaining values for three genotypes–phenotypes for thousands of samples is much easier than getting the values for hundreds of variables with the same sample size. Yet another potential problem with analyzing large number of genes is that because of biological feedback loops, causal loops might be present in the gene network, thus violating the assumption required for our causal inference.

Even though the specific gene used in this paper, HLA-DRB1 locus, is the major genetic contributor to the RA, it is not certain if the true causal gene is included in our dataset. From what we know about the genetics of RA, there are a few recently discovered genes whose polymorphisms are significantly associated with the RA, such as PTPN22, CTLA4, PADI4, in at least certain populations (Plenge *et al.*, 2005). These genes could be used as the third variable besides anti-CCP and RF to investigate the causal relationship. However, preliminary analysis on the data in (Lee *et al.*, 2005) showed that PTPN22-RF correlation is much weaker than that of SE-RF ($X^2 = 2.8665$, p -value = 0.09).

In the last section, we had concluded that Equation (1), i.e. $SE \rightarrow anti-CCP \rightarrow RF$ was the only possible causal relationship among the three variables, if the potential confounding variable h is either an environmental or a phenotypic factor. When h is another gene that is in linkage disequilibrium (LD) with the HLA-DRB1 gene (and genotype SE), both Equations (2) and (3) are possible. These models have very simple explanation: Equation (2) represents the situation when another gene in the HLA region is responsible for the anti-CCP phenotype and that gene is in LD with HLA-DRB1 gene. Equation (3) represents the situation when both that gene and HLA-DRB1 gene are joint causes of anti-CCP. To really narrow down the disease-causing mutation from a region known to be linked to the disease is a notoriously difficult task and may require extra information and novel study design (Jawaheer *et al.*, 2002)

One causal model between SE, anti-CCP and SE that was thought to be biologically possible is the following:

$$\begin{array}{c} \text{environment}(h) \\ \swarrow \searrow \\ SE(x) \rightarrow \text{anti-CCP}(y) \quad RF(z), \end{array} \quad (4)$$

where some environmental factor contributes to both anti-CCP and RF, whereas there is no causal link between anti-CCP and RF. This model is the line 50 in Table 4 of Cooper (1997). Interestingly, this model is rejected because it would imply a correlation between SE and RF conditional on the anti-CCP variable, whereas such correlation is absent in our data. In fact, our data reject any causal models

that contain a confounding hidden variable to be the cause of both anti-CCP and RF. For a hidden variable to be a cause of only anti-CCP (or only RF), on the other hand, is possible.

Equation (1) indicates a causal relationship among the variables under investigation, and it does not prevent other variables that are not included in our study as intermediates along the arrows. For example, the following causal relationship is consistent with Equation (1):

$$\text{SE}(x) \rightarrow \begin{array}{c} \text{gene}(h') \\ \downarrow \\ \dots \end{array} \text{anti-CCP}(y) \rightarrow \begin{array}{c} \text{env}(h'') \\ \downarrow \\ \dots \end{array} \text{RF}(z), \quad (5)$$

where the dots represent other unspecified intermediate phenotypes or biomarkers. Both anti-CCP and RF can have genetic or environmental contributions [as indicated by h' and h'' in Equation (5)] but not in a confounding fashion [as indicated by h in Equation (4)].

Since cause always precede effects in time, our conclusion of anti-CCP \rightarrow RF predicts that patients with RA should first develop high anti-CCP level (anti-CCP positive) before developing high RF level (RF positive). Both anti-CCP and RF biomarkers are of diagnostic and prognostic value (Zendman *et al.*, 2004; Rantapää-Dahlqvist, 2005) because a high proportion (roughly 50%) of the RA patients test positive for either one of the two biomarkers before the onset of the disease (Aho *et al.*, 1991; Schellekens *et al.*, 2000), and they tend to belong to a more severe form of the disease.

In a recent survey, 22, 32 and 39 RA patients (out of total 79 patients, so 27.8, 40.5 and 49.4%) became RF-positive, anti-CCP positive and both RF and anti-CCP positive before the onset of RA (Nielen *et al.*, 2004). The median time for a positive RF or anti-CCP test result with respect to the onset of RA is -2 and -4.8 years. This at least provides some evidence that averaging over a group of RA patients, anti-CCP becomes positive before RF becomes positive. Further evidence is needed to test whether for individual RA patients the anti-CCP is usually positive prior to the onset of positive RF.

Another causality triangle appeared in the literature is the 'Mendelian randomization' (MR) discussed in epidemiology. The purpose of MR is to clarify the causal nature of an association between an environmentally influenceable intermediate phenotype and the disease, with a help of a genotypic polymorphism (Davey-Smith and Ebrahim, 2003). There are many differences between MR and the LCD application discussed here. One superficial difference is that MR concerns the gene, intermediate phenotype, disease (G, IP, D) triangle, whereas our application concerns the (G, IP, IP) triangle; this also leads a contrast of the case-control design versus the case-only design. The MR aims at verifying consistency between the data and a particular causal model by, for example, checking the propagation of the risks (Keavney, 2004). This consistency check assumes the pathway structure from the gene to the intermediate phenotype, and finally to the disease is a simple one. The participation of other risk factors makes the consistency check invalid. The local causality discovery rule, on the other hand, does not make assumption on variables or risk factors that are not included in the analysis, and it does not require a quantitative consistency of the risks. A more fundamental difference between MR and LCD is that MR is not a true causal inference: the goal of MR is limited to verifying consistency with a particular

causal model, and it may not be able to distinguish different causal models (Thomas and Conti, 2004).

In conclusion, with two intermediate phenotypes (biomarkers) and one genotype all associated with a disease, it is possible to determine the causal relationship between the two intermediate phenotypes. We have successfully applied a local causality discovery rule (Cooper, 1997; Silverstein *et al.*, 2000) to the three-variable set of two biomarkers for RA, anti-CCP antibody and RF, and one genotype known to be associated with the RA, the HLA-DRB1 allele. Non-trivial conclusions have been inferred from the data that anti-CCP is upstream in a biochemical pathway with respect to RF, and it is unlikely that a confounding factor causes both anti-CCP and RF. Based on this success, we recommend a routine use of causal inference in genetic analysis when stratified data are available, though the analysis needs to be handled with care if the sample size is only moderate, or if biological feedback loops are present.

ACKNOWLEDGEMENTS

The authors are grateful for support from the Eileen Ludwig Greenland Center for Rheumatoid Arthritis. The authors thank Franak Batliwalla, Annette Lee and Hye-Soon Lee for helpful discussions. Support for this work was provided by NIH grants RO1-AR44422 and NO1-AR-2-2263.

Conflict of Interest: none declared.

REFERENCES

- Agresti, A. (2002) *Categorical Data Analysis*, 2nd edn. Wiley & Sons, Hoboken, New Jersey.
- Aho, K. *et al.* (1991) Rheumatoid factors antedating clinical rheumatoid arthritis. *J. Rheumatol.*, **18**, 1282–1284.
- Bay, S.D. *et al.* (2004) Temporal aggregation bias and inference of causal regulatory networks. *J. Comput. Biol.*, **11**, 971–985.
- Beck, S. *et al.* (The MHC Sequencing Consortium) (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature*, **401**, 921–923.
- Chu, T. *et al.* (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147–1152.
- Clayton, D. and McKeigue, P.M. (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, **358**, 1356–1360.
- Cooper, G. (1997) A simple constraint-based algorithm for efficiently mining observational databases. *Data Min. Knowl. Disc.*, **1**, 203–224.
- Cox, D.R. and Wermuth, N. (1996) *Multivariate Dependencies*. Chapman & Hall, London.
- Davey-Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of diseases? *Int. J. Epidemiol.*, **32**, 1–22.
- Dawid, A.P. (1979) Conditional independence in statistical theory. *J. R. Statist. Soc. Ser. B*, **41**, 1–31.
- Dawid, A.P. (1980) Conditional independence in statistical operation. *Ann. Stat.*, **8**, 598–617.
- Granger, C.W.J. (1969) Investigating causal relations by econometric methods and cross-spectral methods. *Econometrica*, **37**, 424–438.
- Granger, C.W.J. (1980) Testing for causality. A personal viewpoint. *J. Econ. Dyn. Control*, **2**, 329–352.
- Gregersen, P.K. (1998) The North American Rheumatoid Arthritis Consortium—bringing genetic analysis to bear on disease susceptibility, severity, and outcome. *Arthritis Care Res.*, **11**, 1–2.
- Gregersen, P.K. *et al.* (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.*, **30**, 1205–1213.
- Halloran, M.E. and Struchiner, C.J. (1995) Causal inference in infectious diseases. *Epidemiology*, **6**, 142–151.

- Holland, P.W. (1986) Statistics and causal inference. *J. Am. Stat. Assoc.*, **81**, 945–970.
- Hoover, K.D. (2001) *Causality in Macroeconomics*. Cambridge University Press.
- Irigoyen, P. et al. (2005) Regulation of anti-cyclic citrullinated peptide antibodies in rheumatoid arthritis. *Arthritis Rheum.*, **52**, 3813–3818.
- Jawaheer, D. et al. (2002) Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am. J. Hum. Genet.*, **71**, 585–594.
- Katan, M.B. (1986) Apolipoprotein E isoforms, serum cholesterol and cancer. *Lancet*, **1**, 507–508.
- Keavney, B. (2004) Commentary: Katan's remarkable foresight: genes and causality 18 years on. *Int. J. Epidemiol.*, **33**, 11–14.
- Koopman, J.S. (1977) Causal models and sources of interaction. *Am. J. Epidemiol.*, **106**, 439–443.
- Lee, A.T. et al. (2005) The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status. *Genes Immun.*, **6**, 129–133.
- Li, W. (1990) Mutual information functions versus correlation functions. *J. Stat. Phys.*, **60**, 823–837.
- Nielen, M.M. et al. (2004) Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. *Arthritis Rheum.*, **50**, 380–386.
- Niles, H.E. (1922) Correlation, causation, and Wright's theory of 'path coefficients'. *Genetics*, **7**, 258–273.
- Ollier, W. and Thomson, W. (1992) Population genetics of rheumatoid arthritis. *Rheum. Dis. Clin. North Am.*, **18**, 741–759.
- Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK.
- Plenge, R.M. et al. (2005) Replication of putative candidate-gene associations with rheumatoid arthritis in >4000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am. J. Hum. Genet.*, **77**, 1044–1060.
- Pope, R.M. and McDuffy, S.J. (1979) IgG rheumatoid factor. Relationship to seropositive rheumatoid arthritis and absence in seronegative disorders. *Arthritis Rheum.*, **22**, 988–998.
- Rantapää-Dahlqvist, S. (2005) Diagnostic and prognostic significance of autoantibodies in early rheumatoid arthritis. *Scand. J. Rheumatol.*, **34**, 83–96.
- Robins, J.M. et al. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 561–570.
- Rodin, A.S. and Boerwinkle, E. (2005) Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). *Bioinformatics*, **21**, 3273–3278.
- Rubin, D.B. (1974) Estimating causal effect of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
- Schadt, E.E. et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.
- Schellekens, G.A. et al. (2000) The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. *Arthritis Rheum.*, **43**, 155–163.
- Shipley, B. (2002) *Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, Cambridge, UK.
- Silverstein, C. et al. (2000) Scalable techniques for mining causal structures. *Data Min. Knowl. Disc.*, **4**, 163–192.
- Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*. Cambridge University Press, Cambridge, UK.
- Stastny, P. (1978) Association of the B-cell alloantigen DRw4 with rheumatoid arthritis. *New Engl. J. Med.*, **298**, 869–871.
- Thomas, D.C. and Conti, D.V. (2004) Commentary: the concept of 'Mendelian randomization'. *Int. J. Epidemiol.*, **33**, 21–25.
- Wright, S. (1921) Correlation and causation. *J. Agric. Res.*, **20**, 557–585.
- Xing, B. and Van der Laan, M.J. (2005) A causal inference approach for constructing transcriptional regulatory networks. *Bioinformatics*, **21**, 4007–4013.
- Yoo, C. et al. (2002) Discovery of causal relationships in a gene regulation pathway from a mixture of experimental and observational DNA microarray data. *Pac. Symp. Biocomput.*, **16**, 498–509.
- Zendman, A.J. et al. (2004) Autoantibodies to citrullinated (poly)peptides: a key diagnostic and prognostic marker for rheumatoid arthritis. *Autoimmunity*, **37**, 295–299.