A distance-based test of association between paired heterogeneous genomic data

Christopher Minas[1,2], Edward Curry[3] and Giovanni Montana[2,*]

[1]Department of Imaging Sciences, Institute of Clinical Sciences, Hammersmith Campus, [2]Statistics Section, Department of Mathematics, South Kensington Campus and [3]Department of Surgery and Cancer, Ovarian Cancer Action Research Centre, Hammersmith Campus, Imperial College London, London W12 0NN, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Due to rapid technological advances, a wide range of different measurements can be obtained from a given biological sample including single nucleotide polymorphisms, copy number variation, gene expression levels, DNA methylation and proteomic profiles. Each of these distinct measurements provides the means to characterize a certain aspect of biological diversity, and a fundamental problem of broad interest concerns the discovery of shared patterns of variation across different data types. Such data types are heterogeneous in the sense that they represent measurements taken at different scales or represented by different data structures.

Results: We propose a distance-based statistical test, the generalized RV (GRV) test, to assess whether there is a common and non-random pattern of variability between paired biological measurements obtained from the same random sample. The measurements enter the test through the use of two distance measures, which can be chosen to capture a particular aspect of the data. An approximate null distribution is proposed to compute *P*-values in closed-form and without the need to perform costly Monte Carlo permutation procedures. Compared with the classical Mantel test for association between distance matrices, the GRV test has been found to be more powerful in a number of simulation settings. We also demonstrate how the GRV test can be used to detect biological pathways in which genetic variability is associated to variation in gene expression levels in an ovarian cancer sample, and present results obtained from two independent cohorts.

Availability: R code to compute the GRV test is freely available from http://www2.imperial.ac.uk/~gmontana

Contact: g.montana@imperial.ac.uk

Supplementary data: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A proliferation in the development and application of high-throughput measurement platforms in biological research has resulted in the increasing availability of multiple levels of molecular data available for the same biological sampling units. For a range of human tumour types, for instance, the Cancer Genome Atlas (TCGA) consortium have made available genome-wide single nucleotide polymorphism (SNP) genotypes, DNA copy number, genome-wide gene (and mRNA) expression, DNA methylation and proteomic profiles, as well as a range of mutation calls arising from deep sequencing. Such comprehensive molecular profiling of biological sampling units provides opportunities for gaining insight into the mechanisms by which different molecular entities interact with one another to influence the overall state of the cell(s) in question. This is possible through comparing the different types of observations on each biological sampling unit, via some means of paired analysis of the individual datasets. A certain degree of complexity is inherent in this type of analysis because the measurements that are being compared are heterogeneous.

One well-studied example of the task of analysing paired heterogeneous biological data is gene expression quantitative trait loci (eQTL) mapping, which seeks to find associations between SNPs and transcript expression levels (Cookson *et al.*, 2009). A particularly interesting application of multivariate eQTL mapping relates to the discovery of association between biomarkers for ovarian cancer, which is treated as a case study in this work.

The application of paired analysis of heterogeneous biological data types is not limited to eQTL mapping, and involves the common problem of discovering shared patterns of variation across different biological measurements. Other areas that have seen considerable interest include the association of genetic copy number variation with gene expression levels (Beroukhim *et al.*, 2007) and identification of genes for which DNA methylation is associated to gene expression (Louhimo and Hautaniemi, 2011). In these and related studies, the problem of comparing heterogeneous data types can be approached using a notion of distance between pairs of biological samples. Distances are scalar-valued measures of dissimilarity between two input observations, with greater values indicating greater dissimilarity between them. For a given application, a suitable semi-metric or metric distance measure can generally be defined depending on the nature of the data and on the specific objectives of the study. Such distances computed between all pair-wise combinations of samples are arranged in square symmetric matrices. These matrices then quantify the variability in the random sample according to that particular distance measure, and can be directly compared using an appropriate statistical test.

Distance-based approaches offer a number of advantages compared with direct comparisons involving the original measurements. Firstly, distances can overcome limitations of

---

*To whom correspondence should be addressed.

traditional multivariate approaches when dealing with high-dimensional random vectors observed on a much smaller number of sampling units (Shannon *et al.*, 2002). Secondly, distance measures can be appropriately defined for data generated under different experimental conditions, and are not confined to data types represented only by vectors. A number of distance measures are readily available for several types of biological data, including DNA sequences (Wu *et al.*, 1997), genetic markers such as SNPs (Goh *et al.*, 2011), gene expression data (Priness *et al.*, 2007), longitudinal gene expression data (Minas *et al.*, 2011) and proteins (Hollich *et al.*, 2005). Moreover, for a specific data type, more than one distance measure can be deployed to capture different aspects of the data. This can ultimately lead to the discovery of different types of associations. It should be noted, however, that in the case of multivariate measurements, interpretation with respect to individual variables is lost. Because we are interested in the combined behaviour of all variables comprising such vectorial measurements, we do not consider this issue further.

Only a handful of statistical procedures are currently available to determine whether there is an association between paired distance matrices, and they all rely on computationally intensive procedures for statistical inference. The most commonly used and well-known procedure is the Mantel test (Mantel, 1967), a generalization of Pearson's correlation. The test has already been found useful in many applications arising in bioinformatics (Beckmann *et al.*, 2005; Shannon *et al.*, 2002; Sun *et al.*, 2011). This testing procedure, however, has some limitations. Firstly, when used to compare paired scalar-valued observations using the Euclidean distance, the test exhibits less power than Pearson's correlation test (Legendre and Fortin, 2010; Peres-Neto and Jackson, 2001), i.e. it is a poor test of the null hypothesis of no correlation between scalar-valued variables. Secondly, computationally intensive Monte Carlo permutations are required to assess significance. This procedure, which assumes exchangeability of the sampling units under the null hypothesis of no association, introduces sampling errors that leads to inaccurate estimation of small *P*-values. This is especially true when too few permutations are used, generally between $O(10^3)$ and $O(10^5)$ (Berry and Mielke, 1983; Knijnenburg *et al.*, 2009; Mielke and Berry, 2007). For example, to obtain a permutation *P*-value within 10^{-5} of a true *P*-value of 0.001, it has been shown that $O(10^7)$ permutations are required, which is computationally infeasible without access to specialized high-performance computational resources. This is the case in all situations where thousands or even millions of tests are required, such as in eQTL mapping studies in which the entire genome is scanned in search of potential associations. In such large-scale applications, it is not uncommon to settle for a much smaller number of Monte Carlo permutations per test, which in turn can lead to inflated family-wise type I error rates (Phipson and Smyth, 2010).

In this article we propose a novel test, the generalized RV (GRV) test, to detect association between paired distance matrices. This test is derived as a generalization of the classical multivariate RV test of no correlation between paired random vectors, originally proposed by Escoufier (1973). While inference can be performed via computationally intensive Monte Carlo permutations, we also derive an approximation to the exact null distribution which would be obtained by enumerating all permutations. Using the proposed null distribution, approximate *P*-values can be estimated in closed-form without the need for computationally intensive procedures. The proposed GRV test is introduced in Section 2. In Section 3, we describe power studies showing that the GRV test can offer greater statistical power than that achieved by the Mantel test even when many millions of Monte Carlo permutations are performed. Section 4 presents an application of the GRV test in cancer research, where the interest lies in detecting biological pathways characterized by a non-random association between genetic and gene expression sample variability. Concluding remarks are presented in Section 5.

2 THE GRV TEST

2.1 Problem statement

Assume we have observed paired measurements $\{(x_i, y_i)\}_{i=1}^{N}$ on N independent biological sampling units. Each one of the $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{N}$ measurement sets is represented by a particular data structure, such as continuous or discrete vectors, although these structures do not necessarily need to be vectorial. For instance, the measurements might be structured as graphs or trees (e.g. representing biological networks or ontologies). We also suppose that suitably chosen semi-metric or metric distances d_x and d_y have been identified that capture the dissimilarity between any sample pair $\{(x_i, x_j)\}_{i \neq j}$ and $\{(y_i, y_j)\}_{i \neq j}$, respectively. On evaluating the distances for all possible pairs, we obtain the paired $N \times N$ distance matrices $\mathbf{\Delta}_x = \left\{ d_x(x_i, x_j) \right\}_{i,j=1}^{N}$ and $\mathbf{\Delta}_y = \left\{ d_y(y_i, y_j) \right\}_{i,j=1}^{N}$.

Using the random sample, evidence of non-random association between these paired genomic measurements can be assessed by testing the null hypothesis that $\mathbf{\Delta}_x \neq a\mathbf{\Delta}_y$, for some positive constant a. This constant represents possible scaling differences between the elements of each distance matrix, which may arise because of the chosen distance measures. Under this null hypothesis, the relationship between the elements in $\mathbf{\Delta}_x$ is not maintained by the corresponding elements in $\mathbf{\Delta}_y$. The alternative hypothesis is that the distance matrices are equal up to a constant, in which case every distance $d_x(x_i, x_j)$ is linearly related to the paired distance $d_y(y_i, y_j)$. We are particularly concerned with settings where inferences must be drawn for a large number of such tests simultaneously, and hence the computational cost in obtaining a *P*-value for each test should be kept as low as possible.

2.2 The proposed test statistic

When the observations are vectorial, the mean-centred vectors $\{x_i\}_{i=1}^{N} \in \mathbb{R}^P$ and $\{y_i\}_{i=1}^{N} \in \mathbb{R}^Q$ can be arranged in paired data matrices, X and Y, of dimensions $N \times P$ and $N \times Q$, respectively. In this case, to establish whether the P and Q variables are correlated, the null hypothesis that $\mathbf{\Sigma}_{xy} = \mathbf{0}$ is commonly tested, where $\mathbf{\Sigma}_{xy}$ is the $P \times Q$ covariance matrix containing the true covariances between the P variables observed in $\{x_i\}_{i=1}^{N}$ and Q variables observed in $\{y_i\}_{i=1}^{N}$. The RV statistic has been proposed for this task and arises as a generalization of Pearson's correlation (Escoufier, 1973). As with Pearson's correlation, RV is computed as the ratio of covariance to the square-rooted product of the variances,

$$\mathrm{RV}(X, Y) = \frac{\mathrm{tr}(X^T Y Y^T X)}{\|X^T X\| \|Y^T Y\|}, \qquad (1)$$

where $\mathrm{tr}(\cdot)$ is the trace function and $\|A\| = \sqrt{\mathrm{tr}(A^T A)}$ denotes the Frobenius norm for matrix A. The RV statistic ranges between 0 and 1, with no association when $X^T Y = 0$, and perfect association when there exists a linear mapping that relates every P-dimensional observation in X to every Q-dimensional observation in Y. Thus, larger values provide evidence against the null hypothesis. When $P = Q = 1$, $\mathrm{RV}(X, Y) = \mathrm{cor}^2(X, Y)$, where cor $(\cdot, \cdot)$ denotes Pearson's correlation coefficient.

We first observe that the RV test can be interpreted in terms of the Euclidean distance matrices Δ_x and Δ_y arising by applying the Euclidean distance measure to $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$, respectively. This is because, due to properties of the trace operator, $\mathrm{tr}(X^T Y Y^T X) = \mathrm{tr}(X X^T Y Y^T)$, $\|X^T X\| = \|X X^T\|$ and $\|Y^T Y\| = \|Y Y^T\|$, so that RV statistic (1) can be written in equivalent form as

$$\mathrm{RV}(X, Y) = \frac{\mathrm{tr}(X X^T Y Y^T)}{\|X X^T\| \|Y Y^T\|}. \qquad (2)$$

Note that (1) differs from (2) in that emphasis is placed on the two symmetric $N \times N$ matrices $X X^T$ and $Y Y^T$, instead of the four covariance matrices $X^T Y \in \mathbb{R}^{P \times Q}$, $Y^T X \in \mathbb{R}^{Q \times P}$, $X^T X \in \mathbb{R}^{P \times P}$ and $Y^T Y \in \mathbb{R}^{Q \times Q}$. In settings where $P, Q > N$, which are common when comparing biological measurements, it is computationally convenient to use (2) over (1).

In addition, the $N \times N$ matrices $X X^T$ and $Y Y^T$ can each be written in terms of the Euclidean distance matrices Δ_x and Δ_y, respectively. This is due to the results that

$$X X^T = -\frac{1}{2} C \Delta_x^2 C \quad \text{and} \quad Y Y^T = -\frac{1}{2} C \Delta_y^2 C, \qquad (3)$$

where $C = (I_N - J_N/N)$ is the symmetric $N \times N$ centring matrix where I_N is the identity matrix of size N and J_N is the square matrix of ones of size N (Borg and Groenen, 2005; Gower, 1966). Also, $-\frac{1}{2} C \Delta_x^2 C = G_x$, where G_x is the square, symmetric and real-valued Gower's centred inner product matrix with (i, j)th element given by

$$g_x(x_i, x_j) = -\frac{1}{2} \left(\frac{1}{N} d_x^2(x_i, x_j) - \frac{1}{N} \sum_{k=1}^N d_x^2(x_i, x_k) \right.$$
$$\left. - \frac{1}{N} \sum_{l=1}^N d_x^2(x_l, x_j) + \frac{1}{N^2} \sum_{l=1}^N \sum_{k=1}^N d_x^2(x_l, x_k) \right),$$

and similarly for G_y, so that $\mathrm{tr}(X X^T Y Y^T) = \mathrm{tr}(G_x G_y)$, $\|X X^T\| = \|G_x\|$ and $\|Y Y^T\| = \|G_y\|$. Thus RV is completely specified by the elements of the Euclidean distance matrices.

Based on these observations, we propose a generalization for the RV statistic that is entirely specified by any pair of distance matrices to test the null hypothesis that $\Delta_x \neq a \Delta_y$. These can be obtained by applying suitable distance measures d_x and d_y to paired observations $\{(x_i, y_i)\}_{i=1}^N$ of any type, i.e. not necessarily the Euclidean distance to vectors. The resulting G_x and G_y matrices can then be used in the GRV test statistic

$$\mathrm{GRV}(G_x, G_y) = \frac{\mathrm{tr}(G_x G_y)}{\|G_x\| \|G_y\|}, \qquad (4)$$

noting the implicit assumption $\|G_x\| \|G_y\| > 0$, which is always satisfied in practice; $\|G_x\| > 0$ because G_x contains real-valued elements that are not all zero, and similarly for G_y.

It is insightful to understand in which ways the proposed GRV statistic is similar to, and differs from, the classical Mantel statistic. The Mantel statistic is computed as the correlation between the upper-triangular distances of each distance matrix. Denote by v_x the column vector containing the $A = N(N-1)/2$ upper-triangular distances $\{d_x(x_i, x_j)\}_{i>j}$, and similarly for v_y. The Mantel statistic is then computed as $r_M(\Delta_x, \Delta_y) = \mathrm{cor}(v_x, v_y)$, i.e.

$$r_M(\Delta_x, \Delta_y) = \frac{1}{A-1} \sum_{i>j} \left(\frac{d_x(x_i, x_j) - \bar{x}}{s_x} \right) \left(\frac{d_y(y_i, y_j) - \bar{y}}{s_y} \right),$$

where $\bar{x} = \sum_{i>j} d_x(x_i, x_j)/A$, $s_x^2 = \sum_{i>j} (d_x(x_i, x_j) - \bar{x})^2/(A-1)$, and similarly for $\bar{y}$ and s_y^2. The statistic therefore standardizes the upper-triangular elements of each distance matrix to have zero mean and unit variance. In this way, the distances between paired distance matrices can be directly compared.

It can be shown that, like Mantel, GRV is also a correlation coefficient. In particular, it can be proved that GRV is equal to the correlation between the vectorized matrices $G_x/\|G_x\|$ and $G_y/\|G_y\|$, i.e. the N^2-dimensional vectors g_x and g_y with elements $\{g_x(x_i, x_j)/\|G_x\|\}_{i,j=1}^N$ and $\{g_y(y_i, y_j)/\|G_y\|\}_{i,j=1}^N$, respectively. That is, $\mathrm{GRV}(G_x, G_y) = \mathrm{cor}(g_x, g_y)$ (see the Supplementary Material). While both Mantel and GRV are correlation coefficients, the difference between them lies in the methods of standardization applied to the distances in each case. In Mantel, the upper-triangular distances are subjected to a classical standardization, where their mean is subtracted and they are divided by their standard deviation. In GRV, however, all distance elements are considered, and they are squared, double-centred and normalized by dividing by their Frobenius norm. This difference leads to greater power of the GRV test to detect association between paired distance matrices than the Mantel test, as demonstrated in Section 3. Further properties of the GRV test statistic are provided in the Supplementary Material. This includes the important property that it overcomes the limitation of Mantel for scalar-valued observations and Euclidean distances. Due to the equality between the RV and GRV statistics when the data are vector-valued and the Euclidean distance measure is applied [from (2) and (3)], if centered data matrices X and Y are orthogonal, then the GRV statistic equals 0. Thus the GRV statistic can be used to test a null hypothesis of no correlation between vector-valued variables, and hence scalar-valued variables, unlike the Mantel statistic.

2.3 Approximate null distribution

Under the null hypothesis, the sampling distribution of GRV is generally unknown. This is because the quantity $T = \mathrm{tr}(G_x G_y)$ in the numerator is completely specified by the elements of Δ_x and Δ_y, which have unknown distributions and are correlated. The null sampling distribution can be generated by using permutations of one of the centered inner product matrices, G_y, say. For each of N_π permutations $\pi \in \Pi$, which are one-to-one mappings of the set $\{1, \ldots, N\}$ to itself, the rows and columns of G_y are simultaneously permuted by π and denoted $G_{y,\pi}$. This generates the set $\{\widehat{\mathrm{GRV}}(G_x, G_{y,\pi})\}_{\pi \in \Pi}$ and the P-value of an observed

GRV statistic, $\widehat{\text{GRV}}(\boldsymbol{G}_x, \boldsymbol{G}_y)$, can be obtained as the proportion of permuted statistic values greater than or equal to $\widehat{\text{GRV}}(\boldsymbol{G}_x, \boldsymbol{G}_y)$. This is a right-tailed test, as larger GRV values provide evidence against the null.

To estimate the *P*-value without permutations, we adopt a moment matching approach where the exact null distribution that would be obtained if all *N*! permutations were used is approximated by a continuous distribution. In particular, we approximate the null distribution by the Pearson type III distribution, which has been shown to capture skewed characteristics of null sampling distributions (Josse *et al.*, 2008; Mielke and Berry, 2007). To use this distribution, we require the first few moments of the exact permutation distribution of *T*; the mean is $\mu = \frac{1}{N!}\sum_{\pi \in \Pi} \hat{T}_\pi$, the variance is $\sigma^2 = \frac{1}{N!}\sum_{\pi \in \Pi} \hat{T}_\pi^2 - \mu^2$ and the skewness is

$$\gamma = \frac{\frac{1}{N!}\sum_{\pi \in \Pi} \hat{T}_\pi^3 - 3\mu\sigma^2 - \mu^3}{\sigma^3},$$

where $\hat{T}_\pi = \text{tr}(\boldsymbol{G}_x \boldsymbol{G}_{y,\pi})$ and Π contains all *N*! permutations. No permutations are needed to compute $\{\mu, \sigma, \gamma\}$, as closed-form expressions of these quantities are retrievable via the analytical results of Kazi-Aoual *et al.* (1995). These results require that $\boldsymbol{G}_x$ and $\boldsymbol{G}_y$ are square, symmetric and centered; properties satisfied by definition.

On obtaining closed-form expressions for the mean, variance and skewness of the exact permutation distribution of *T*, we standardize *T* by subtracting μ and dividing by σ. The resulting $T_s = (T - \mu)/\sigma$ is then assumed to have the Pearson type III distribution under the null with probability density function (PDF) denoted $f_{T_s}(t; \gamma)$; see Mielke and Berry (2007) for the full definition of the distribution. We denote the CDF of T_s by $\mathcal{F}_{T_s}(t; \gamma)$. The approximate null distribution of GRV $(\boldsymbol{G}_x, \boldsymbol{G}_y) = T/\|\boldsymbol{G}_x\|\|\boldsymbol{G}_y\|$ can then be derived by a simple transformation of the distribution of T_s. The CDF of GRV, $\mathcal{F}_{\text{GRV}}(x; \gamma)$, is given by

$$\mathcal{F}_{\text{GRV}}(x; \gamma) = \mathcal{F}_{T_s}\left(\frac{x\|\boldsymbol{G}_x\|\|\boldsymbol{G}_y\| - \mu}{\sigma}; \gamma\right),$$

and this is a valid CDF because $\mathcal{F}_{T_s}(; \gamma)$ is a valid CDF. The *P*-value of an observed GRV statistic $\widehat{\text{GRV}}(\boldsymbol{G}_x, \boldsymbol{G}_y)$ is then estimated by $1 - \mathcal{F}_{\text{GRV}}(\widehat{\text{GRV}}(\boldsymbol{G}_x, \boldsymbol{G}_y); \gamma)$. The PDF $f_{\text{GRV}}(x, \gamma)$ is found by differentiation as

$$f_{\text{GRV}}(x, \gamma) = \left(\frac{\|\boldsymbol{G}_x\|\|\boldsymbol{G}_y\|}{\sigma}\right)f_{T_s}\left(\frac{x\|\boldsymbol{G}_x\|\|\boldsymbol{G}_y\| - \mu}{\sigma}; \gamma\right),$$

because $\|\boldsymbol{G}_x\|\|\boldsymbol{G}_y\| > 0$ and $\sigma > 0$.

3 POWER STUDIES

In this section we assess the power of the proposed GRV test and compare it with the classical Mantel test. A Monte Carlo procedure is used to estimate the power of the GRV test, and to estimate the power of the Mantel test we use the recently proposed algorithm of Gandy and Rubin-Delanchy (2013). This algorithm estimates the power of a Monte Carlo testing procedure and states a confidence interval around this estimate, boasting a guaranteed coverage probability. It also states the average

number of Monte Carlo permutations required to achieve the given power estimate. It requires specification of the maximum length of the resulting confidence interval and the desired coverage probability, and then runs as many permutations as required to yield a power estimate with a confidence interval of length no greater than specified. Using this approach, we are able to demonstrate that, at least for the experimental set-ups considered, many millions of permutations are needed for the Mantel test to achieve power of the same precision as the GRV test, and that the GRV test achieves greater power.

We use simulated data that mimics the experimental data of a typical eQTL application where transcriptional measurements are paired with SNPs. We first generate an $N \times P$ matrix $\boldsymbol{X}$ containing *N* simulated SNPs (i.e. minor allele counts), denoted $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iP})^T$ for $i = 1, \ldots, N$, with varying minor allele frequencies at each of the *P* SNPs. The minor allele frequency of SNP *p*, m_p, is first simulated from a uniform distribution $U(0.1, 0.5)$, and the allele count is then simulated from a multinomial distribution with probabilities $(1 - m_p)^2$, $2m_p(1 - m_p)$ and m_p^2 of observing 0, 1 and 2 minor alleles, respectively. The paired data matrix $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)^T$ is then generated as follows. The $N \times 1$ vector $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{I}_P = (z_1, \ldots, z_N)^T$ containing the row sums of $\boldsymbol{X}$, i.e. the minor allele counts across the *P* SNPs, is computed, and $\boldsymbol{y}_i = z_i\boldsymbol{I}_Q + \boldsymbol{e}_i$ for $i = 1, \ldots, N$, where $\boldsymbol{e}_i \sim N_Q(\mu, \Sigma)$, with $\mu = (\mu_1, \ldots, \mu_Q)^T$ and $\mu_q \sim U(0, 1)$ for $q = 1, \ldots, Q$, and Σ a random $Q \times Q$ Wishart matrix. We then consider the widely used identity-by-state (IBS) distance for the simulated genetic markers in $\boldsymbol{X}$, which measures the degree to which alleles are shared across the SNPs [see, for instance, Goh *et al.* (2011)], in addition to the Sokal and Sneath (SS), and Rogers and Tanimoto I (RTI) distances (Selinski and Ickstadt, 2005). For the observations in $\boldsymbol{Y}$ we use the Euclidean, Manhattan and Mahalanobis distances. See the Supplementary Material for details on these distances.

For $P = 2$, $Q = 10$ and each of $N = \{30, 50, 70, 90, 100\}$, the power of the GRV test is computed using 50 Monte Carlo runs, and each time generating 50 paired datasets as above. For each genetic and gene expression distance, the GRV test is applied and the proportion of *P*-values less than or equal to the significance level of 0.1% is recorded. The mean power across all 50 Monte Carlo runs for each distance combination and value of *N* is reported in Table 1, in addition to the standard deviation of these estimates. As expected, the power increases with *N*.

We then estimate the power of the Mantel test for each of the above settings using the algorithm of Gandy and Rubin-Delanchy (2013). For each setting, we bound the confidence interval length by twice the standard deviation of the corresponding estimate for the GRV test. That is, we consider one standard deviation on either side of the power estimate as an empirical indication of the precision achieved by the GRV test. We monitor the number of Monte Carlo permutations required to obtain power estimates with such confidence intervals with a coverage probability of 95%. These results are also given in Table 1, where the power is stated alongside the confidence interval, and the number of Monte Carlo permutations required is stated on a separate line below the confidence interval.

In all settings, the GRV test achieves greater power than the Mantel test, even when using large numbers of Monte Carlo permutations. Note that while the Mantel power estimates

Table 1. Power of the GRV and Mantel tests at a significance level of 0.1% for $N = \{30, 50, 70, 90, 100\}$ and all combinations of the genetic (gen) IBS, SS and RTI distances and gene expression (gex) Euclidean (Euc), Manhattan (Man) and Mahalanobis (Mah) distances

Gen dist	Gex dist	Test	N									
			30		50		70		90		100	
IBS	Euc	GRV	0.126	(0.042)	0.351	(0.064)	0.566	(0.074)	0.712	(0.049)	0.780	(0.051)
		Mantel	0.081	[0.042,0.126] 3.081	0.116	[0.056,0.183] 1.575	0.225	[0.145,0.291] 9.921	0.275	[0.228,0.325] 13.51	0.322	[0.272,0.374] 19.61
	Man	GRV	0.139	(0.049)	0.375	(0.068)	0.590	(0.066)	0.730	(0.059)	0.772	(0.065)
		Mantel	0.069	[0.023,0.120] 1.859	0.151	[0.091,0.226] 2.134	0.240	[0.173,0.306] 6.290	0.302	[0.238,0.355] 13.80	0.378	[0.314,0.444] 8.794
	Mah	GRV	0.394	(0.061)	0.846	(0.047)	0.952	(0.030)	0.978	(0.020)	0.990	(0.014)
		Mantel	0.100	[0.049,0.169] 2.818	0.311	[0.262,0.355] 28.53	0.542	[0.513,0.572] 135.9	0.715	[0.694,0.735] 424.2	0.776	[0.761,0.789] 666.8
SS	Euc	GRV	0.038	(0.029)	0.157	(0.042)	0.299	(0.067)	0.470	(0.064)	0.556	(0.063)
		Mantel	0.035	[0.006,0.064] 0.971	0.111	[0.069,0.154] 5.318	0.180	[0.112,0.245] 3.032	0.268	[0.202,0.329] 7.445	0.319	[0.254,0.381] 10.69
	Man	GRV	0.045	(0.028)	0.160	(0.046)	0.308	(0.068)	0.503	(0.073)	0.556	(0.068)
		Mantel	0.047	[0.018,0.073] 2.879	0.121	[0.080,0.171] 10.27	0.195	[0.133,0.269] 4.371	0.310	[0.242,0.387] 7.325	0.363	[0.290,0.424] 9.232
	Mah	GRV	0.092	(0.041)	0.481	(0.074)	0.784	(0.059)	0.907	(0.037)	0.927	(0.034)
		Mantel	0.072	[0.030,0.112] 5.567	0.266	[0.194,0.340] 8.635	0.510	[0.448,0.564] 21.92	0.652	[0.613,0.687] 74.12	0.723	[0.689,0.756] 40.01
RTI	Euc	GRV	0.044	(0.027)	0.159	(0.047)	0.340	(0.062)	0.551	(0.067)	0.654	(0.063)
		Mantel	0.032	[0.010,0.065] 1.327	0.106	[0.061,0.154] 5.141	0.205	[0.142,0.265] 8.112	0.302	[0.233,0.366] 1.421	0.339	[0.272,0.397] 16.36
	Man	GRV	0.055	(0.035)	0.172	(0.052)	0.363	(0.067)	0.540	(0.076)	0.639	(0.072)
		Mantel	0.051	[0.019,0.089] 1.373	0.124	[0.067,0.171] 4.512	0.216	[0.154,0.288] 4.371	0.360	[0.282,0.432] 9.289	0.401	[0.321,0.465] 7.233
	Mah	GRV	0.173	(0.055)	0.675	(0.068)	0.925	(0.038)	0.987	(0.015)	0.993	(0.010)
		Mantel	0.062	[0.011,0.119] 0.8066	0.241	[0.180,0.314] 11.89	0.515	[0.478,0.554] 151.0	0.688	[0.672,0.702] 654.4	0.765	[0.755,0.776] 1068

Note: For the GRV test, the standard deviation of the power estimate is given in brackets. For the Mantel test, the confidence interval obtained with 95% coverage probability is given in square brackets. The average number of Monte Carlo permutations required to achieve the power estimates for Mantel are stated below the confidence interval in millions.

improve with N, as expected, the number of Monte Carlo permutations varies between $O(10^5)$ and $O(10^9)$ non-linearly with N. This is an artefact of the algorithm; the expected number of permutations depends on the length of the confidence interval sought and the region of the confidence interval in $[0,1]$ [see Gandy and Rubin-Delanchy (2013) for more details]. An example of how the standardization used by GRV yields greater power than that used by Mantel is presented in the Supplementary Material, in addition to simulation results demonstrating that the GRV test attains the nominal significance level of a given test.

4 APPLICATION TO OVARIAN CANCER

Around 225 000 women were estimated to have developed ovarian cancer in 2008, of which over 80% are epithelial ovarian cancer (EOC). In the UK, EOC has particularly poor prognosis, with only 34% of patients surviving 5 years after diagnosis.

In many cancers, the ability to measure some biomarker that can predict the likely response to a chemotherapeutic agent could prevent patients unnecessarily suffering side effects from ineffective treatments, and may suggest the best course of action where multiple treatments are available. Unfortunately, while high-throughput gene expression profiling has been successful in identifying prognostic 'signatures' for many cancers (van't Veer and Bernards, 2008), solid tumour gene expression profiles have practical drawbacks preventing their use as prognostic or predictive biomarkers in the clinic, especially in terms of obtaining mRNA from the tumour and the transitive nature of gene expression (Sawyers, 2008). However, circulating tumour DNA can be isolated from serum samples in ovarian cancer patients (Hickey, 1999), which reflect an attractive source for genetic biomarkers both because of the relative ease of accessibility of the patient material and the long-term stability of the genetic sequence. Additionally, treating biological pathways as multidimensional biomarkers may provide substantial improvements in terms of robustness over single agents (Dai *et al.*, 2011).

We sought to discover if it might be possible to use genetic biomarkers as surrogates for the transcriptional activation of pathways. Such pathway-based genetic biomarkers may suggest

starting points for the development of clinical diagnostic tests to predict the outcome of a therapeutic exploiting a related pathway.

To identify pathways for which the profile of SNP genotypes across genes in the pathway closely reflects the state of expression of the genes in the pathway, we applied the GRV test to evaluate the significance of the degree of similarity in distance matrices derived from SNP genotypes and gene expression levels from a large set of selected pathways, for two independent cohorts.

4.1 Datasets

For this study, we use two independent cohorts of ovarian cancer patients: one collected by the TCGA consortium, and an independent cohort for validation. For the TCGA cohort, raw data CEL files from Affymetrix SNP 6.0 and from Affymetrix HT-HGU133A microarray profiling 494 primary high grade serous ovarian tumours were obtained from the TCGA web repository. For the Affymetrix SNP 6.0 arrays, normalization and genotype calling were performed using the CRLMM method (Carvalho et al., 2007). Current annotations were obtained from Affymetrix's NetAffx Analysis Center and SNPs were annotated with their associated Gene Symbols. Once SNP probes were mapped to Gene Symbols, lists of SNPs corresponding to each ConsensusPathDB (CPDB) (Kamburov et al., 2011) pathway were obtained by selecting all SNP probes mapping to Gene Symbols that were included in the pathway's list of hgnc_symbol_ids as downloaded from CPDB. For the Affymetrix HT-HGU133A arrays, data normalization was performed using the RMA method (Irizarry et al., 2003). Probe-set mapping to Gene Symbols was performed using the hthgu133a.db package from Bioconductor. Once probe-sets were mapped to Gene Symbols, lists of probe-sets corresponding to each CPDB pathway were obtained by selecting all probe-sets mapping to Gene Symbols that were included in the pathway's list of hgnc_symbol_ids as downloaded from CPDB. The tables of genotype calls and normalized gene expression values, mapped to pathways that were used in this analysis, are provided in Supplementary Tables S1 and S2, respectively.

As an independent validation dataset, raw data CEL files from both Affymetrix SNP 6.0 and Affymetrix HGU133plus2 microarrays profiling 60 primary high-grade serous ovarian tumours were obtained from the Hammersmith Biobank Resource Centre. The datasets were provided courtesy of the Genome Institute of Singapore. Raw data for this independent tumour cohort was processed identically to those obtained from TCGA, although mapping of gene expression probe-sets to Gene Symbols was performed using the hgu133plus2.db package from Bioconductor. The tables of genotype calls and normalized gene expression values, mapped to pathways that were used in this analysis, are provided in Supplementary Tables S3 and S4, respectively.

We wanted to reduce the possibility that any identified associations simply reflected genome-wide similarity structures between the patients (e.g. due to molecular subtypes of disease) or involved so few measurements that they were not indicative of the entire pathway. From the full list of 4119 CPDB pathways we filtered out all those with less than 5 or more than 200 probe sets in either gene expression or SNP genotype array datasets.

A total of 479 pathways remained after this filtering, and these pathways are listed in Supplementary Table S5.

4.2 Experimental results

Because different distance measures highlight different relationships among genetic markers and features of gene expression data (Priness et al., 2007; Song et al., 2012), we sought to perform a meta-analysis to combine the results of similarity testing across a range of distance measures for both the genotypes and gene expression values. The rationale being that if a given pathway shows similar distance matrices at the genotype and gene expression level for a number of different distance measures, the observed similarity is unlikely to be due to a quirk of one particular distance measure. For each of the 479 pathways, sample-wise distance matrices were calculated using eight different distance measures (Euclidean, Mahalanobis, Manhattan, Maximum, Bray-Curtis, Pearson's correlation, Spearman's correlation and Cosine angle distances) for the gene expression data and five distance measures (IBS, SS, RTI, Simple Matching and Hamman I) for the SNP genotypes; see the Supplementary Material for details on these distances. The association between the paired distance matrices, one derived from the genotypes of SNPs mapping to genes in the pathway and the other derived from the expression levels of genes in the pathway, was evaluated using the GRV test. This resulted in 40 individual P-values for each pathway, one for each combination of distance measures used. An illustration of how the null distribution compares with the permutation distribution when applied to this cancer data is presented in the Supplementary Material. Meta-analysis was performed to yield a combined significance estimate for each pathway as follows. On denoting the P-values of the ith pathway by $\{p_{ij}\}_{j=1}^{40}$, the Fisher summary statistic $-2\sum_{j=1}^{40}\log(p_{ij})$ was calculated for each pathway. The corresponding P-value was estimated using 10 000 permutations of the individual P-values across all pathways. All P-values were subjected to the Benjamini–Hochberg false discovery rate adjustment and used to rank pathways by significance of association. The meta-analysis was repeated for the validation cohort. Ranks for the TCGA and validation cohorts are given in Supplementary Tables S6 and S7, respectively.

To obtain a perspective of the robustness/repeatability of the associations detected between genotype and gene expression at the pathway level across the two independent cohorts, we compared the degree of agreement between the ranked lists of pathways arising from the meta-analysis using the normalized Canberra distance (Jurman et al., 2008). The general principle is to quantify the distance between the rankings of the top k terms of two ranked lists, with each term's influence on the overall distance weighted by its position in its respective list. By evaluating the normalized Canberra distance for the top-ranked k pathways, the relationship between overlap and ranking can be explored: smaller values indicate a lesser discrepancy in rankings and therefore more similar rankings. Figure 1 shows the normalized Canberra distance plotted against k. It can be seen that the overlap of the meta-analysis results from the two cohorts is greatest for small k, implying that the highest-ranking pathways in each list are more similar than the lower-ranking pathways. Furthermore, for each k we used 5000 permutations of
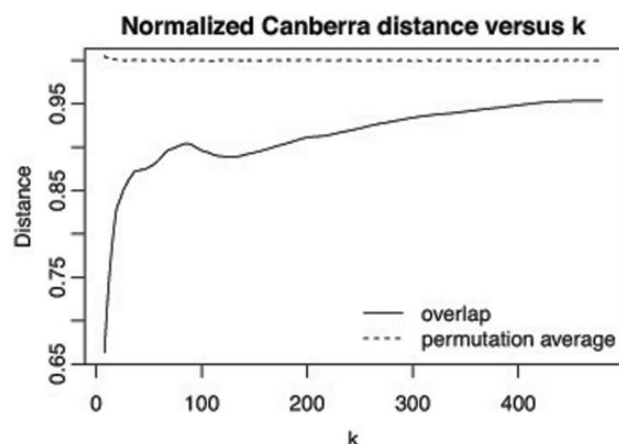
Normalized Canberra distance versus k

Fig. 1. Plot of the normalized Canberra distance against k for the top k elements in two lists of pathways. Each pathway list is ranked according to meta-analysis estimate of significance of association between genotype and gene expression over 40 combinations of distance measures, one ranked list per cohort. The dotted line represents the average normalized Canberra distance between the TCGA data-derived pathway list and each of 5000 random permutations of the ranks in the second list

the ranks in the second list to obtain a P-value estimate for the distance observed between the two rankings. These P-values across the range of k were less than 0.005 after false discovery rate adjustment, reflecting the fact that the overlap between the two ranked lists of pathways was significantly greater than that expected by chance. These results suggest that the association between distance matrices based on the levels of expression of genes within certain pathways and distance matrices based on SNP genotypes associated with genes in the corresponding pathways is robust and reliably evaluated by the GRV test. The meta-analysis of the GRV test result across multiple distance measures demonstrates that these associations are not a result of peculiarities of particular distance measures, and certain pathways are shown to exhibit a strong association in both cohorts.

A particularly striking result of this study is the CPDB pathway '5-aminoimidazole ribonucleotide biosynthesis', which is ranked #1 in the meta-analysis of GRV test results from the TCGA dataset and ranked #2 in the meta-analysis of GRV test results from the validation dataset. The 5-aminoimidazole ribonucleotide is a precursor to de novo purine biosynthesis, which is essential for rapidy proliferating tissue (Firestine and Davisson, 1993). Pharmacological inhibition of de novo purine biosynthesis has proved to be a successful strategy in the clinic for treating cancers and inflammatory disorders (Christopherson *et al.*, 2002). It has been shown through a clinical trial that genetic polymorphisms could be biomarkers to predict response to methotrexate therapy, which inhibits de novo purine biosynthesis, in patients with rheumatoid arthritis (Dervieux *et al.*, 2004). The 'methotrexate pathway' itself showed significant genotype-expression association in the TCGA cohort (ranked #17), but not in the validation cohort (ranked #358). The '5-aminoimidazole ribonucleotide biosynthesis' pathway may therefore suggest additional or alternative SNPs to use as genetic biomarkers for predicting outcome of inhibiting de novo purine biosynthesis.

Another result of note is that the CPDB pathway 'Platinum Pathway, Pharmacokinetics/Pharmacodynamics' yields highly significantly similar tumour-wise distance matrices across all 40 applications of the GRV test on this paired ovarian cancer data, for both cohorts investigated (ranked #20 from the TCGA dataset and #20 from the validation dataset). This is particularly interesting because platinum-based chemotherapy is the standard treatment for ovarian cancer, but unlike for many other solid cancers, survival rates in ovarian cancer have not improved noticeably in the 30 years since this treatment was introduced (Vaughan *et al.*, 2011). The principal reason for this is the high frequency of platinum-resistant relapse, which may occur through multiple mechanisms (Agarwal and Kaye, 2003; Stronach *et al.*, 2011). Therefore, being able to predict the response to the platinum-based DNA damaging agents could be of great clinical value in the treatment of ovarian cancer.

Additionally, all genotype-derived and expression-derived distance matrices for the Base Excision Repair pathway are highly significantly similar in the TCGA cohort (#25). This is of interest as the Base Excision Repair pathway is involved in repairing platinum-induced DNA damage. In fact, effective inhibition of this pathway (via PARP) in patients with BRCA1/2 mutations is associated with sensitivity to platinum chemotherapy (Fong *et al.*, 2010).

5 CONCLUSION

In this work, we have presented the GRV test as a novel procedure to detect association between paired distance matrices, which is applicable in any setting where suitable distance measures can be defined. Similarly to the widely used Mantel test, the GRV test can be directly applied whether the distance measures are metric or semi-metric, but overcomes the limitation of the Mantel test that, where paired data is vector-valued, a hypothesis of no correlation between the vectors of interest can be tested. As a further advantage of the GRV test over existing distance-based tests, an approximate null distribution is proposed such that inference can be performed without expensive Monte Carlo permutation procedures. Through extensive simulations we have demonstrated that the GRV test achieves greater power than the popular Mantel test. This greater power is exhibited even when many millions of permutations are used for the Mantel test, thus demonstrating its suitability as a tool for use in multiple-testing settings were many tests need to be simultaneously performed.

The GRV test was applied to paired ovarian cancer datasets in which genome-wide profiles of SNP genotypes and gene expression levels were available for each patient across two independent cohorts ($N = 494$ and $N = 60$). The paired datasets were split into subsets in which the features all mapped to the same pathway (using mappings from the Consensus Pathway DB), so that each GRV test result would indicate the statistical significance of the similarity between the (tumour-wise) distance matrix derived from the genotype data and the distance matrix derived from the gene expression data. For each cohort a meta-analysis was performed across the GRV test results from 40 different combinations of distance measures, and the pathways were ranked according to their significance across all distance measures. Permutation-based estimates of the significance of the overlap

between the ranked lists of pathways from the two independent cohorts revealed a highly significant agreement in the two rankings, indicating a degree of robustness in the results of our analysis.

This application of the GRV test has demonstrated that the transcriptional activity of a number of pathways seem to be reliably predicted by sequencing a limited set of genetic markers that could be detected in circulating tumour DNA obtained from patients' serum. This relationship was particularly pronounced for the pathway '5-aminoimidazole ribonucleotide biosynthesis', for which the genotypes and the gene expression patterns were strikingly associated in two independent cohorts of patients. The association between genotypes and pathway-level gene expression profile is of particular interest in the case of the platinum response pathway, which is clearly clinically important in EOC given that the only standard treatment involves platinum-based chemotherapy. Following further investigation, these observations could potentially lead to development of non-invasive biomarkers that could predict patient response to front-line chemotherapy, and serve to stratify patients for selection into clinical trials involving alternative treatments.

As individual distance measures can yield different results in the analysis of associations across gene expression patterns, unless a single distance measure is selected *a priori* as the most appropriate for each dataset, combining results of multiple tests of association using different distance measures may provide more reliable results than a single test. A key advantage of the GRV test in our application to multivariate eQTL mapping of pathways in ovarian cancer is that it enables a meta-analysis of the associations between genetic distance and gene expression distance for different combinations of distance measures. Having a test for these associations that does not require expensive permutations enables a fast estimation of the robustness of the genotype-expression pathway level associations against quirks of particular definitions of distance. Given the robust associations discovered in this analysis, we suggest that it may be possible to predict the mode of activation of a pathway (in terms of which members are transcriptionally activated) in cancer patients, based on the genotype of the tumours' DNA across a limited panel of SNPs.

ACKNOWLEDGEMENTS

REFERENCES

Agarwal,R. and Kaye,S.B. (2003) Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nat. Rev. Cancer*, **3**, 502–516.

Beckmann,L. *et al.* (2005) Haplotype sharing analysis using mantel statistics. *Hum. Heredity*, **59**, 67–78.

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *PNAS*, **104**, 20007–20012.

Berry,K.J. and Mielke,P.W. (1983) Moment approximations as an alternative to the F test in analysis of variance. *Br. J. Math. Stat. Psychol.*, **36**, 202–206.

Borg,I. and Groenen,P.J.F. (2005) *Modern multidimensional scaling: theory and applications*. 2nd edn. Springer Science+Business Media, Inc., New York, USA.

Carvalho,B. *et al.* (2007) Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics*, **8**, 485–499.

Christopherson,R.I. *et al.* (2002) Inhibitors of de novo nucleotide biosynthesis as drugs. *Acc. Chem. Res.*, **35**, 961–971.

Cookson,W. *et al.* (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.

Dai,W. *et al.* (2011) Systematic CpG islands methylation profiling of genes in the Wnt pathway in epithelial ovarian cancer identifies biomarkers of progression-free survival. *Clin. Cancer Res.*, **17**, 4052.

Dervieux,T. *et al.* (2004) Polyglutamation of methotrexate with common polymorphisms in reduced folate carrier, aminoimidazole carboxamide ribonucleotide transformylase, and thymidylate synthase are associated with methotrexate effects in rheumatoid arthritis. *Arthritis Rheum.*, **50**, 2766–2774.

Escoufier,Y. (1973) Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

Firestine,S.M. and Davisson,V.J. (1993) A tight binding inhibitor of 5-Aminoimidazole Ribonucleotide Carboxylase. *J. Med. Chem.*, **36**, 3484–3486.

Fong,P.C. *et al.* (2010) Poly (ADP)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. *J. Clin. Oncol.*, **28**, 2512–2519.

Gandy,A. and Rubin-Delanchy,P. (2013) An algorithm to compute the power of monte carlo tests with guaranteed precision. *Ann.Appl. Stat.*, **41**, 125–142.

Goh,L. *et al.* (2011) Assessing matched normal and tumor pairs in next-generation sequencing studies. *PLoS One*, **6**, e17810.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Hickey,K.P. *et al.* (1999) Molecular detection of tumour DNA in serum and peritoneal fluid from ovarian cancer patients. *Br. J. Cancer*, **80**, 1803–1808.

Hollich,V. *et al.* (2005) Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Mol. Biol. Evol.*, **22**, 2257–2264.

Irizarry,R. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Josse,J. *et al.* (2008) Testing the significance of the RV coefficient. *Comput. Stat. Data Anal.*, **53**, 82–91.

Jurman,G. *et al.* (2008) Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, **24**, 258–264.

Kamburov,A. *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.

Kazi-Aoual,F. *et al.* (1995) Refined approximations to permutation tests for multivariate inference. *Comput. Stat. Data Anal.*, **20**, 643–656.

Knijnenburg,T.A. *et al.* (2009) Fewer permutations, more accurate P-values. *Bioinformatics*, **25**, i161–i168.

Legendre,P. and Fortin,M.J. (2010) Comparison of the mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour.*, **10**, 831–844.

Louhimo,R. and Hautaniemi,S. (2011) CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, **27**, 887–888.

Mantel,N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res.*, **27**, 209–220.

Mielke,P.W. and Berry,K.J. (2007) *Permutation methods: a distance function approach*. Springer Science+Business Media, LLC, New York, USA.

Minas,C. *et al.* (2011) Distance-based differential analysis of gene curves. *Bioinformatics*, **27**, 3135–3141.

Peres-Neto,P.R. and Jackson,D.A. (2001) How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178.

Phipson,B. and Smyth,G.K. (2010) Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, **9**, 39.

Priness,I. *et al.* (2007) Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, **8**, 111.

Sawyers,C.L. (2008) The cancer biomarker problem. *Nature*, **452**, 548–552.

Selinski,S. and Ickstadt,K. (2005) Similarity Measures for Clustering SNP Data. *Technical Report / Universitt Dortmund, SFB 475 Komplexittsreduktion in Multivariaten Datenstrukturen*.

Shannon,W.D. *et al.* (2002) Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genet. Epidemiol.*, **23**, 87–96.

Song,L. *et al.* (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, **13**, 328.

Stronach,E.A. *et al.* (2011) HDAC4-regulated STAT1 activation mediates platinum resistance in ovarian cancer. *Cancer Res.*, **71**, 4412.

Sun,Y.V. *et al.* (2011) Identification of genes associated with complex traits by testing the genetic dissimilarity between individuals. In: *BMC Proceedings.* Vol. 5, p. S120, Licensee BioMed Central Ltd.

van't Veer,L.J. and Bernards,R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**, 564–570.

Vaughan,S. *et al.* (2011) Rethinking ovarian cancer: recommendations for improving outcomes. *Nat. Rev. Cancer*, **11**, 719–725.

Wu,T.J. *et al.* (1997) A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. *Biometrics*, **53**, 1431–1439.