Gene expression

Exploiting sample variability to enhance multivariate analysis of microarray data

Carla S. Möller-Levet¹,²,* , Catharine M. West² and Crispin J. Miller¹

¹Paterson Institute for Cancer Research, Cancer Research UK and ²Academic Radiation Oncology, The University of Manchester, Christie Hospital, Manchester, M20 4BX, UK

Received on June 12, 2007; revised on July 31, 2007; accepted on August 20, 2007

Advance Access publication September 7, 2007

ABSTRACT

Motivation: Biological and technical variability is intrinsic in any microarray experiment. While most approaches aim to account for this variability, they do not actively exploit it. Here, we consider a novel approach that uses the variability between arrays to provide an extra source of information that can enhance gene expression analyses.

Results: We develop a method that uses sample similarity to incorporate sample variability into the analysis of gene expression profiles. This allows each pairwise correlation calculation to borrow information from all the data in the experiment. Results on synthetic and human cancer microarray datasets show that the inclusion of this information leads to a significant increase in the ability to identify previously characterized relationships and a reduction in false discovery rate, when compared to a standard analysis using Pearson correlation. The information carried by the variability between arrays can be exploited to significantly improve the analysis of gene expression data.

Availability: Matlab script files are available from the author.

Contact: cmoller@picr.man.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

1.1 Background

Microarrays provide a static snapshot of transcript abundance in a particular sample. All probe measurements within the same array share sample-specific biological and technical information. Each array is different, even replicates, because it is impossible to reproduce exactly the biological and technical conditions of the instance at which the snapshot was taken. This produces, to different extents, variability between the arrays.

In microarray experiments the expression profile of a sample is defined by all the probe (or probeset in the case of Affymetrix arrays) values obtained for that sample, while the expression profile of a gene is given by the set of values reported by the probe(s) or probeset(s) targeting that gene across a set of arrays. Data from an entire experiment, consisting of one or more arrays, are represented as a matrix in which the rows correspond to the genes (probes/probesets) and the columns to the samples. Typically, analysis is either gene- or sample-oriented (Armstrong and van de Wiel, 2004). Gene expression profiles are analysed by comparing rows to identify, e.g. similar expression patterns, co-regulated genes and functionally related genes. Samples are analysed by comparing columns in the expression matrix to distinguish between different tumour types, clinical outcomes, cell cycle positions, tumour versus normal samples, etc. Either analysis can be constrained by the other: in two-way analysis methods [e.g. biclustering (Madeira and Oliveira, 2004), also known as simultaneous, subspace and co-clustering], for example, subgroups of genes and samples are identified; such that the genes show highly correlated activities for the subgroup of samples. In this analysis, the data are still represented by two independent dimensions whose elements are generally considered as independent and identically distributed. A particularly common analysis method, popularized by Eisen et al. (1998), uses hierarchical clustering applied independently to the rows and columns to re-order the expression matrix, which is subsequently visualized using a heat map. Although this approach has been used successfully to derive and test diverse biological hypotheses, there is an intrinsic interdependence between gene and sample expression profiles that the method does not take into account.

Principal component analysis (PCA) allows the reproduction of the variance observed in one dimension of the expression matrix using a linear combination of the elements in the second dimension, known as principal components (Alter et al., 2000; Jolliffe, 1986). It is commonly used for dimensionality reduction, because the variance among the variables can be reproduced using a lower number of values per variable. PCA can be very useful in sample-oriented analyses to identify the genes mostly responsible for the principal variance or spatial distribution observed in the samples (Alter et al., 2000). However, given the low number of samples relative to the very large number of genes, it is not suitable for gene-oriented analysis (Yeung and Ruzzo, 2001). Spatial relationships encoding the technical and intrinsic biological variability of the samples are simplified when using fewer dimensions to summarize gene expression variation, loosing important information about sample uniqueness and relative similarity. This is

*To whom correspondence should be addressed.
particularly true if there are more than a few major sources of variation: typical in replicated arrays and clinical samples.

Another possibility is to represent gene expression profiles as a function of one or more parameters that can be measured for each sample. This is most evident in time-course experiments, in which samples have a specific order and separation over time. In a time-course study expression varies with time, such that genes and samples are not independent. However, commonly used methods do not model order and spacing (intervals) between time points; the set of observations for each gene can therefore be exchanged over time. Pairwise similarity measures, such as Pearson correlation or Euclidean distance are invariant with respect to the order of the observations, i.e. if the temporal order of a pair of series is permuted, their correlation or distance does not change. In (Möller-Levet and Yin, 2005) the co-expression coefficient, a shape-based metric, is defined. In this approach, the modelled gene profiles are time-differentiated to obtain the rate of change over time with the result that the samples are intrinsically weighted relative to the length of the original sampling intervals. Measurements taken at shorter intervals of time have a higher weight than those taken less frequently (the higher the sampling frequency the more information one has to recreate the actual biological process). In this way, gene similarity analysis is enhanced by making use of the information contained in the sampling intervals.

Fundamental to the approach is the idea that timepoint data can be used to provide not only an ordering, but also a notion of relative-distance between samples. This extra information can then be exploited to enhance the analysis. An obvious extension to the method is to substitute time with another continuous variable integral to the experiment. Samples can be ordered by dose level, for example, but in many situations no such variable exists. In its absence, we have found that it is possible to use Multidimensional scaling (MDS), a technique to reduce the dimensionality of a dataset whilst preserving the relative distance between data points, to provide the necessary information.

1.2 Enhanced correlation (EC): a novel method for computing gene expression similarity

The method first uses MDS to recreate sample-similarity relationships on a plane. Here, the distance information in the original space is a function of the correlation coefficient between samples, which is used to quantify their varying degrees of similarity. After reducing the dimensionality of the original samples to two dimensions, samples are placed relative to one another on a plane, with inter-sample distances corresponding to those of the original dataset, such that similar samples are placed close together and dissimilar samples are placed further away (Fig. 1a). A surface (3D model) is then created for each probeset (Fig. 1b and c) by interpolating expression levels over the plane formed by the samples. Next, the surface is smoothed and differentiated to obtain the slopes in both axes of the plane. These slopes define the 3D profile; Figure 2 depicts a flowchart of the procedure. The 3D profiles are then compared using the uncentred Pearson correlation coefficient. Henceforth, we refer to the proposed method as ‘Enhanced correlation’ (EC).

Note that the way MDS is being used here is different from the way it is typically applied to microarray data. Rather than reducing the dimensionality of the expression matrix prior to analysis, here, the original probeset data (and thus the dimensionality of the samples) is preserved. The low-dimensional sample information, reflecting the original sample-similarity relationships, provided by MDS is instead used as
Fig. 2. Generation of the 3D profiles. In the first stage, samples are mapped onto a plane. Next, a grid \([n \times n]\) is obtained by resampling the plane. In the second stage, interpolation of the expression values of a probeset over the resampling grid \((n^2\) sampling points) produces the probeset 3D model. The model is smoothed and the gradient is calculated. The 3D profile is represented by the 2\((n(-1))\) features corresponding to the derivatives with respect to each direction on the projection plane. This is repeated for all probesets and, together, all the 3D profiles form the 3D GEM. Squared boxes represent processes and trapezoid boxes represent data.

an additional source of information with which to augment the analysis.

As with the temporal experiments described above, the samples are intrinsically weighted by their inter-sample distances when the slopes of the 3D models are compared. In this case, distances reflect the similarity of the samples, such that similar samples are close, with the result that the change of absolute distances reflect the similarity of the samples, such that similar samples are intrinsically weighted by their inter-sample distances, which will be naturally placed further apart.

The idea of weighting expression levels in gene similarity calculations is not new. The microarray data analysis tool ClusterTreeView (http://rana.lbl.gov/manuals/ClusterTreeView .pdf) allows the user to specify weights for samples in the calculation of gene correlations and weights for genes in the calculation of sample correlations. It also provides an automatic weighting based on the inverse of either gene or sample local density. This approach naturally increases the weight of noisy and outlier samples. Weighting based on error and variability of repeated measurements has been reported by Hughes et al. (2000) and Yeung et al. (2003), respectively. The error model developed in Hughes et al. (2000) assigns relatively high error estimates to genes that show greater variation in their repeated expression levels than other genes at similar abundance in their control experiments. Yeung et al. (2003) used variability estimates (standard deviation or coefficient of variation) of repeated measurements to weight similarity measures. In both cases the error or variance is estimated in a gene-based approach using repeated measurements, and sample variability is not considered. Seo et al. (2004) use the complement of detection call \(P\)-value (Liu et al., 2002) to calculate the weight for each expression level in similarity calculations. In this case, repeated measurements are not necessary because the mismatch probes are used instead, however, sample-specific information is also excluded.

The approach described here is appealing because it exploits the fact that all probe measurements within the same array share technical and biological information unique to that sample. It is evaluated using both synthetic and real datasets to compare its performance to a standard analysis using Pearson correlation [henceforth referred to as ‘Pearson correlation’ (PC)]. Two approaches are used for evaluation. The first uses an annotation database to define pairs of probesets expected to hybridize to the same transcript. The second uses prior knowledge of the literature to define genes, and thus probesets, that have been previously reported to be co-expressed. In both cases algorithms are evaluated in terms of their ability to preferentially identify these, known, pairs by assigning them higher scores than corresponding pairs that are not known to be associated. Results show that the gain in information provided by the new approach produces a significant improvement in the analysis of co-expression.

2 MATERIALS AND METHODS

2.1 Datasets

A synthetic dataset and two cancer datasets are used to compare the EC and PC methods. The synthetic dataset simulates a set of replicated samples containing clusters of correlated probesets against a background of random, unrelated, probesets. This dataset provides a controlled environment for exploring the differences between the two methods in terms of co-expression. The generation of the synthetic dataset is described in Section 1 of the Supplementary Material. The real datasets are clinical samples from different cancer tissue types. The first dataset was obtained by microarraying 59 HNSCC obtained prior to any treatment at the time of surgery (Winter et al., 2007). Clustering around the expression of known hypoxia-associated genes yielded a set of 99 up-regulated genes. High expression of the 99 genes was an adverse Exploiting sample variability
formed by 244 probesets targeting 140 genes known to be involved in DNA damage recognition and repair (DDR) (Wood et al., 2005). The second subset (‘hypoxia-probesets’) is formed by 440 probesets targeting 232 genes identified in the literature as hypoxia regulated (Winter et al., 2007).

2.2 Generating the 3D profiles

2.2.1 Mapping the samples onto the plane The microarray samples are mapped onto two dimensions. Mapping onto any number of dimensions, up to the number of probesets minus one could be used, however, the use of two dimensions was found to offer a good compromise between complexity and information gain (further discussion of this can be found in Section 2 of the Supplementary Material). There are several techniques which can be used for data projection (Yin, 2003), the most popular being PCA (Jolliffe, 1986) and MDS (Cox and Cox, 1994).

Consider a set of \( n \) objects, each object represented by a point in a \( g \)-dimensional space. The aim of MDS is to find \( s \) points in a \( d \)-dimensional space (with \( d < g \)), such that the corresponding distances approximate the original ones as well as possible. To obtain the new points, the original data points are moved around in the \( d \)-dimensional space, evaluating how well the dissimilarities between objects are reproduced by the new configuration. There exists a multitude of variants of MDS with slightly different cost functions and optimization algorithms. A popular method is Sammon Mapping (Sammon, 1969), a non-linear method, in which the errors in distance preservation are normalized with the distance in the original space, emphasizing the preservation of small distances.

The starting point of MDS is a matrix consisting of the pairwise dissimilarities of the objects, in this case, the samples. The metric we used to compare the samples is the Pearson correlation coefficient (other metrics could also be considered). Therefore, it is necessary to transform the correlation matrix into a dissimilarity matrix. The straightforward approach is to use one minus the correlation value, however, by using different transformations it is possible to change the relative position of the ‘noisy’ samples with respect to the rest of the samples (examples and further discussion can be found in Section 3 of the Supplementary Material).

It is not surprising that, given the current facility to produce and store large amounts of data, research into data reduction techniques is continuously growing. In this work we use PCA and Sammon Mapping techniques. However, other scaling methods for this application can be the subject of further research.

2.2.2 Creating the surface and calculating the gradient The surfaces (3D models) are generated by interpolating the set of expression values of each probeset over the samples mapped onto the plane. The resolution of the matrix of sampling points used for interpolation should be at least double the minimum interpoint distance. Once the surface is obtained, it is necessary to smooth it to reduce the noise amplification intrinsic to the gradient calculations. Further details and discussion on sampling, interpolation and smoothing can be found in Section 4 of the Supplementary Material. Finally, the gradient of the smoothed surface is calculated; it is given by the partial derivatives with respect to the two axes of the samples’ plane, it can be understood as the derivative in multi-dimensional space. A flowchart of the procedure is shown in Figure 2.

2.3 Comparison of gene expression profiles

The Pearson correlation coefficient, the most common similarity metric used to compare gene expressions, is used throughout this article for similarity calculations of original probesets (PC method). Given that a 3D profile is formed by the partial derivatives of the 3D (smoothed) model (taken with respect to each axis of the projection plane), positive and negative values correspond to increasing and decreasing expression levels, which should be considered as different. Therefore, the uncentred Pearson correlation coefficient (cosine of the angle between the vectors) is used for the 3D profiles (EC method). Spearman correlation of original probesets was also tested as an alternative; the measure is a non-parametric correlation in which the data are converted to ranks before calculating the coefficient. There was an improvement over the Pearson correlation in some of the synthetic datasets, but interestingly, there was a decline in performance for real data (results can be found in Section 5 of the Supplementary Material).

2.4 False discovery rates

To calculate the False discovery rate (FDR) in list enrichment of probesets targeting the same transcript (PTP motifs), the null distribution is generated by bootstrapping \( np \) correlation values from the non-PTP pairs and the real distribution is formed by the total PTP pairs (\( np \) pairs). In the knowledge-based list enrichment, null scores were obtained using random permutations of the expression levels and local FDR (IFDR) is calculated following the algorithm described by Efron et al. (2001).

3 RESULTS

3.1 Multiple targeting for list enrichment comparisons

Affymetrix arrays use a set of short 25-mer oligonucleotide probes to target a transcript, together referred to as a ‘probeset’. Signals from each probe are combined by means of a summarization algorithm [e.g. MAS5 (Affymetrix, 2002), GCRMA (Wu et al., 2003)] to yield a single value for each probeset. In some cases, a probe is capable of forming an exact match with multiple transcripts; in others, a transcript can be targeted by probes from multiple probesets. Together, these lead to a lattice of many-many relationships between probesets and transcripts, and it has been shown that they result in non-independence between measured expression levels (Okoniewski and Miller, 2006). The lattice can be decomposed into two basic elements: probesets targeting the same transcript [Probeset-Transcript-Probeset (PTP motifs)] and transcripts targeted by the same probeset [Transcript-Probeset-Transcript (TPT motifs)].

Here, we exploit these relationships to provide a method for benchmarking different analysis methods. In the absence of other confounding effects, such as alternative splicing, two probesets in a PTP motif would be expected to produce similar expression profiles, since they are measuring the concentration of the same mRNA molecule. Previous work has shown that this is indeed the case; Pearson correlation is significantly greater for probesets within a PTP motif than between randomly selected probeset-pairs (Okoniewski and Miller, 2006).

While the distribution of Pearson correlation for all probeset-pairs approximately follows a normal distribution centred on zero, the majority of the correlations of PTP probesets are positive, as shown in Figure 3a. Not all PTPs are highly correlated, mainly because of alternative splicing events, artefacts in the sample labelling or hybridization process, or inaccurate annotations. However, given the significant difference in the distributions, this provides a useful quantitative approach for assessing analysis methods: the enrichment of PTP pairs selected at a certain threshold. In all datasets, TPT probesets were filtered out. Figure 3b shows the percentage of PTP pairs found
Fig. 3. Panel (a), probability density distributions of correlations between all probeset-pairs and PTP pairs in the HNSCC dataset. The vertical lines mark the quartiles in each distribution. Panel (b), percentage of PTP pairs found at varying threshold levels of the top \( np \) ranked correlations, where \( np \) is the total number (3323) of PTP pairs in the dataset. The broken lines indicate the percentage of PTP enrichment at 6%, 50% and 100% of the top 3323 correlations. At 6% of the top 3323 correlations (i.e. top 199 correlations), 6% of the PTPs are selected - all 199 correlations correspond to PTP pairs. At 50% of the top 3323 correlations, 28.07% of the PTPs are selected, and at 100% of the top 3323 correlations, 37.88% of the PTPs are selected.

(PTP enrichment) in the top \( np \) ranked correlations, where \( np \) is the total number of PTP pairs in the HNSCC dataset.

3.2 Comparison of mapping techniques

Differences in the results obtained by using PCA and MDS as mapping techniques were explored using the HNSCC dataset. In addition, to examine the relevance of the probesets used to calculate sample similarities, the mappings were obtained using both the DDR-probesets and all-probesets sets. The percentage of PTP enrichment, in all cases, is calculated based on the DDR-probesets. When all-probesets are used for the mapping, the results of EC, using either MDS or PCA, are similar to the results obtained with PC. However, EC produces higher list enrichment when only the DDR-probesets are used for the mapping, and the improvement is greatest for MDS (PC 27.21%, EC PCA 29.25% and EC MDS 34.69% of PTP enrichment). Examination of the eigenvalues reveals that four dimensions are necessary to accommodate more than 50% of the variance of the samples using PCA (four dimensions are necessary to accommodate 57.44%). This could explain the lower improvement of PCA when compared to MDS. When the data have many independent variables, it is not possible to find a good low-dimensional configuration of points using PCA. The non-linearity of MDS appears to cope better with the independence of the variables.

The effect of the detection calls was explored by using a function of the detection call \( P \)-values to provide weights in the calculation of the sample similarities used for the mapping (further details can be found in Section 6 of the Supplementary Material). If information from the detection calls is used for the mapping, EC is best, regardless of which probesets and which techniques are used for the mapping: PC 27.21%, EC PCA (all-probesets) 33.33%, EC PCA (DDR-probesets) 30.61%, EC MDS (all-probesets) 36.05% and EC MDS (DDR-probesets) 36.73% of PTP enrichment. This could indicate that the large number of absent probesets, in the unfiltered all-probesets set, introduce noise into the mapping of the samples, i.e. the signal-to-noise ratio is low.

3.3 Bootstrapping on the number of samples

The influence of the number of samples was explored both on a synthetic dataset and by bootstrapping the HNSCC dataset. We found that in each case, performance (as measured by list enrichment) decreases with fewer samples. The synthetic data simulate a set of replicated samples containing clusters of correlated probesets against a background of random, unrelated, probesets. The percentage of noisy samples is fixed to 20% of the number of samples (Section 1 of the Supplementary Material plots a realization of the synthetic data with five replicates). In the absence of noise, PC and EC produce the same list enrichment and there is a 27% improvement when the number of samples is increased from 5 to 80 (results are shown in Section 7 of the Supplementary Material). When noise is introduced, EC produces up to 40.8% higher list enrichment than PC, and is less affected by the number of samples (Fig. 4). A similar effect can be observed in real data. PTP enrichment for different numbers of samples (the standard error is calculated by bootstrapping) was calculated for the DDR-probesets in the HNSCC dataset. Figure 5 shows that < 20% of the PTP pairs can be identified with five samples. For less than 16 samples EC and PC produce very similar enrichment, from 18 to 39 samples the EC method outperforms PC for a given summarization algorithm, and for a higher number of samples, EC outperforms PC regardless of the summarization algorithm. Using 50 samples, EC produces a 11.27% and 12.45% improvement over PC for GCRMA and MA55, respectively. 2

Irrespective of the methods, it is important to notice the difference in list enrichment relative to the number of samples used to calculate the correlation. In particular, with 10 samples or less, even the best approach (EC, GCRMA) finds less than one-third of the possible PTP pairs in the data, and the FDR is correspondingly high (data not shown). Thus, it is important to assess the statistical significance of correlation, and to consider the number of samples required to identify co-expression and co-regulation with high confidence.

3.4 List enrichment in diverse cancer tissue types

PTP enrichment of the HNSCC dataset1 was calculated based on two subsets of known, related, probesets, as well as on the set of all probesets in the array. Figure 6 shows that in all three sets, EC outperforms PC with the same, or lower FDR. The tests were extended to different cancer tissue types, using the expO cancer dataset. 1 For these tests, seven datasets were generated as follows: 50 breast samples, 50 colon samples, 50 kidney samples, 50 endometrium samples, 42 different tissue samples (called mixed) and two 50 samples datasets (called combined 1 and combined 2) formed by 10 breast, 10 colon, 10 kidney, 10 endometrium and 10 from the mixed dataset. Table 1 compares the PTP enrichment obtained with EC and PC for the different datasets and different probeset subsets. The results indicate that

1Data filtered for absent calls. MDS is used as the mapping technique.
2A comparison of sample variability between MAS5 and GCRMA data is shown in Section 8 of the Supplementary Material.
3PTP enrichment plots, based on all probesets in the array, for all seven datasets, are shown in Section 9 of the Supplementary Material.
the improvement of EC over PC is statistically significant ($\chi^2$ test $P$-values: DDR-probesets 0.006, hypoxia-probesets 0.005 and all-probesets 0.028). Similar to the HNSCC dataset, the FDR is lower for EC (data not shown) - even for the datasets in which EC and PC produce similar list enrichment. This lowering of the FDR occurs because there is a non-monotonic relationship between the distributions of the correlations produced by PC and EC; the dissimilarity between the distributions of the correlations of all probeset-pairs and PTP pairs is different in EC and PC; the dissimilarity between the distributions of the correlations produced by PC and EC is lower for EC (data not shown) - even for the datasets in which EC and PC produce similar list enrichment. This lowering of the FDR occurs because there is a non-monotonic relationship between the distributions of the correlations produced by PC and EC; the dissimilarity between the distributions of the correlations of all probeset-pairs and PTP pairs is different in each case.

3.5 Knowledge-based list enrichment

A knowledge-based list enrichment of genes validated in the literature as regulated by hypoxia was carried out by clustering around CA9 gene expression (Winter et al. 2007). Carbonic anhydrase IX (CA9) is a hypoxia-inducible gene (up-regulated by hypoxia) that has emerged as an important surrogate marker for hypoxia in solid tumours and has been associated with poor outcome in many epithelial cancers (Harris, 2002; Lal et al., 2001; Winter et al., 2007). The correlation of CA9 gene expression (205199_at) with all the probesets in the array was calculated using PC and EC. The relative enrichment of hypoxia-related probesets selected at varying levels of FDR (see Materials and Methods section) is shown in Figure 7. The use of EC increases the proportion of literature-based hypoxia-related probesets correlating with CA9 gene expression.

Lists generated by selecting probesets with IFDR (see Materials and Methods section) < 10% were processed through FatiGO (Al-Shahrour et al., 2004), from the Babelomics suite (Al-Shahrour et al., 2005), to find significant differences in the distribution of GO terms between the two groups of genes, taking into account multiple testing. At an adjusted FDR $P$-value of <0.01, EC had a significantly higher percentage of defence response (GO:0006952) and immune response (GO:0006955) associated genes than PC (Defence response, EC 21.79% and PC 2.58%; Immune response, EC 19.23% and PC 2.58%). The genes associated to these processes in each list and Gene Ontology

---

**Fig. 4.** Percentage of same-cluster pairs (vertical axis) found in the top ranked correlations (horizontal axis), relative to the total number of same-cluster pairs. The results are based on the synthetic dataset and $n_s$ is the number of samples. EC produces up to 40.8% higher list enrichment than PC.

**Fig. 5.** PTP enrichment produced with different number of samples. The standard error reported is calculated by bootstrapping. Results based on the DDR-probesets in the HNSCC dataset.

**Fig. 6.** Left panels show percentage of PTP pairs found in the top $np$ ranked correlations, where $np$ is the total number of PTP pairs in the set. Right panels show FDRs for different values of correlation. Results based on the HNSCC data.
The enrichment of the list produced by EC can be found in Section 10 of the Supplementary Material. The EC approach can reveal new potential associations at the transcriptional level that are not identified with a standard analysis of expression levels. These associations can improve the derivation of biologically relevant gene signatures based on unsupervised analyses.

4 DISCUSSION AND CONCLUSIONS

A major issue with microarray experiments is the need to account for the biological and technical variability within the data. This is generally done by performing the experiment multiple times to generate replicate arrays, which are then used to characterize the variability in the dataset. Typically, this information is used to determine the minimum size an effect (e.g. a correlation or a fold-change) must be before it is unlikely to have occurred simply by chance. The approach described here comes from a slightly different perspective. It was developed in recognition that since much of the variability between samples is due to real biological differences (Bakay et al., 2002; Klebanov and Yakovlev, 2007) it therefore has the potential to represent useful information. Rather than simply account for the variability within replicates, the method aims to actively exploit it. This is done by calculating the inter-sample similarities within an experimental dataset, and then integrating this additional information into the original gene expression data in order to produce an enhanced measure of correlation between genes. The modelling of gene expression levels as a function of sample similarity yields a new, 3D, representation that maximizes the use of sample information.

Alternative methods (functions and/or parameters) were evaluated at each step of the approach and it was shown that the proposed, Enhanced correlation, generates better results than a standard analysis using Pearson correlation, regardless of the choice of specific methods. However, the degree of improvement is relative to the selection of these. Recommendations of particular functions and/or parameters are made based on the results of synthetic and real datasets.

The Enhanced correlation was compared to a standard analysis from two different perspectives: theoretical- and knowledge-based list enrichment. In the theoretical-based evaluation, enrichment of probeset-pairs targeting the same gene selected at varying threshold levels of correlation was calculated. In the knowledge-based evaluation, the proportion of known biologically related genes identified by correlation at varying threshold levels of FDR was calculated. CA9 is a hypoxia-inducible gene used as a surrogate marker for hypoxia. Probesets correlating with CA9 with an FDR lower than certain threshold formed the list, and the enrichment was given by the proportion of probesets in the set that targeted genes that have been previously reported in the literature as being hypoxia related. The use of the Enhanced correlation yielded significantly higher list enrichment, with lower FDR, in both theoretical- and knowledge-based evaluations.

The Enhanced correlation could be extended to the analysis of differential gene expression. In differential expression analysis, the method would not classify the samples into replicates and conditions in a discrete manner, instead the samples would be allowed to move continuously in both directions (replicatewise and conditionwise), avoiding a discrete categorization of the replicates. For this case, PCA offers a suitable mapping because PCA axes are not arbitrary (as is the case with MDS) and only a few major sources of variation would be expected.
The work presented in this article is based on Affymetrix data, allowing us to use PTP motifs as part of the evaluation (a second method for evaluation was also used, based on list enrichment of literature-based relationships). However, biological variability is always present regardless of the array platform; the Enhanced correlation could be applied to other expression technologies such as 2-colour arrays.

ACKNOWLEDGEMENTS
This work was founded by Cancer Research, UK. Sara Bhana provided the list of DDR-related genes. We thank S. Bhana, F. Buffa, M. Okoniewski, M. Rattray and A. Sims for useful discussions.

Conflict of interest: none declared.

REFERENCES