BMC Bioinformatics

Research article

Cross-platform comparison and visualisation of gene expression data using co-inertia analysis

Aedín C Culhane*[1], Guy Perrière[2] and Desmond G Higgins[1]

Address: [1]Department of Biochemistry, Biosciences Institute, University College Cork, Cork, Ireland and [2]Laboratoire de Biométrie et Biologie Évolutive, UMR CNRS n°5558 Université Claude Bernard – Lyon 1, 43, bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Email: Aedín C Culhane* - Aedin.Culhane@ucd.ie; Guy Perrière - Perriere@biomserv.univ-lyon1.fr; Desmond G Higgins - Des.Higgins@ucd.ie

* Corresponding author

Published: 21 November 2003

BMC Bioinformatics 2003, **4**:59

Received: 23 June 2003

Accepted: 21 November 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/59

Abstract

Background: Rapid development of DNA microarray technology has resulted in different laboratories adopting numerous different protocols and technological platforms, which has severely impacted on the comparability of array data. Current cross-platform comparison of microarray gene expression data are usually based on cross-referencing the annotation of each gene transcript represented on the arrays, extracting a list of genes common to all arrays and comparing expression data of this gene subset. Unfortunately, filtering of genes to a subset represented across all arrays often excludes many thousands of genes, because different subsets of genes from the genome are represented on different arrays. We wish to describe the application of a powerful yet simple method for cross-platform comparison of gene expression data. Co-inertia analysis (CIA) is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples. CIA simultaneously finds ordinations (dimension reduction diagrams) from the datasets that are most similar. It does this by finding successive axes from the two datasets with maximum covariance. CIA can be applied to datasets where the number of variables (genes) far exceeds the number of samples (arrays) such is the case with microarray analyses.

Results: We illustrate the power of CIA for cross-platform analysis of gene expression data by using it to identify the main common relationships in expression profiles on a panel of 60 tumour cell lines from the National Cancer Institute (NCI) which have been subjected to microarray studies using both Affymetrix and spotted cDNA array technology. The co-ordinates of the CIA projections of the cell lines from each dataset are graphed in a bi-plot and are connected by a line, the length of which indicates the divergence between the two datasets. Thus, CIA provides graphical representation of consensus and divergence between the gene expression profiles from different microarray platforms. Secondly, the genes that define the main trends in the analysis can be easily identified.

Conclusions: CIA is a robust, efficient approach to coupling of gene expression datasets. CIA provides simple graphical representations of the results making it a particularly attractive method for the identification of relationships between large datasets.

Background

Microarray quantification of global gene expression is becoming a very widely used technique. Microarray technology has developed very rapidly and, as a result, different laboratories have adopted numerous different protocols and technological platforms. This severely impacts on the comparability of microarray results [1]. The value of results from microarray gene expression studies would be much greater if they could be cross-validated and compared with data from similar studies.

Currently, meta-analyses of microarray gene expression data are usually based on cross-referencing the annotation of each probe, that is, each oligonucleotide or cDNA sequence attached to each array, extracting a list of gene probes common to all arrays and comparing the expression data of these. Cross-referencing of expression data is usually achieved using UniGene, where probes are considered matched if the GenBank accession number or IMAGE clone identifier of a probe, map to a common UniGene cluster. Meta-analysis of microarray data obtained using similar commercial platforms, or meta-analysis of small subsets of genes is often very successful [2]. While recent attempts to correlate complete Affymetrix oligonucleotide and spotted cDNA array gene expression datasets have reported some success [3], others have reported remarkably poor correlation [4].

Efforts to standardize and improve array annotation [5] should improve inter-laboratory and inter-technology analysis of gene expression data. Nonetheless, the dependence of meta-analysis of microarray data on annotation is limiting for several reasons. Firstly, the identity of gene transcripts spotted on microarrays may be ambiguous. In this case, cross-referencing genes on arrays based on a gene accession number, clone identifier, or even the sequence of a complete gene, is prone to error. In the case of older microarrays, in particular, only a proportion of clones are fully sequence-verified. Furthermore, probes on different microarray platforms may hybridise to different gene regions with different GC content, which will alter the binding properties. Probes may bind to different splice variants of the gene or to homologous genes. This is particularly true when oligonucleotide and cDNA arrays are compared.

Secondly, many protocols cross-reference genes on arrays to the UniGene database [3,6]. UniGene clusters are generated using automated sequence clustering and contain hundreds of thousands of novel expressed sequence tag (EST) sequences in addition to well-characterized genes. As procedures for automated sequence clustering are still under development, and the data, particular EST data, are continually changing, gene clusters in UniGene are frequently updated, retired or joined. Thus, temporary inaccuracies in UniGene, in addition to any poor quality or inaccurate annotation of genes in several public or private databases, are propagated onto microarray probe annotations. Even though two probe sequences on an array may target the same region of a gene, the annotation of these probes may not concur.

Finally, in the case of many genomes including the human genome, it is not yet technically possible to represent the entire genome together with all possible splice variants on a single microarray chip. Thus, different subsets of genes from the genome are represented on different microarrays. Ideally, given a biological sample that has been subjected to several array analyses, one would like to concatenate and combine results from these in order to get as complete a picture as possible of the gene expression profile of that sample. However, cross-referencing of arrays based on annotation, and filtering expression data to that of genes represented across all arrays, excludes thousands of biologically interesting genes.

In this paper, we wish to describe the application of a powerful yet simple method which allows us to perform cross-platform comparison of gene expression data independent of data annotation. Co-inertia analysis (CIA) [7] is a multivariate method that identifies trends or co-relationships in multiple datasets. CIA is commonly applied to the analysis of relationships between species lists and physico-chemical properties of sites in ecological studies, and has already been applied in bioinformatics to the analysis of amino acid properties [8]. It is used in a similar manner to Canonical Correlation Analysis [9] or Canonical Correspondence Analysis [10]. However, these latter methods have a stringent requirement for more cases than variables and are therefore difficult to apply to microarray datasets. By contrast, CIA can be applied to datasets where the number of variables exceeds the number of observations. This is particularly attractive to the analysis of microarray data, where the number of variables (genes) far exceeds the number of samples (arrays) in most analyses. An important feature of this approach is that it is not limited to the analysis of datasets containing the same number of variables (genes). Thus, CIA does not require annotation or statistically based filtering of data prior to cross-platform analysis.

CIA is accomplished by finding successive orthogonal axes from the two datasets with maximum squared covariance. These axes can be derived by principal components analysis, in which case CIA is closely related to the method of partial least squares (PLS). PLS is, in fact, a particular case of CIA. Although the analyses and diagrams described in this paper could have been produced in a similar manner using PLS, we prefer to derive the axes by correspondence analysis (COA), as this is particularly effective at analysing and visualising relationships in microarray data [11,12]. CIA has further flexibility in that it can be used to analyse multiple sets of qualitative as well as quantitative data [13].

We illustrate the power of CIA for cross-platform analysis of microarray data by using it to identify the main common relationships in expression data on a panel of 60 cell lines from the National Cancer Institute (NCI) which have been subjected to different microarray studies using Affymetrix [14,15] and spotted cDNA array [16] technology.

Mathematical basis of CIA

Ordination is a term used in ecology, where it refers to the representation of objects (sites, stations, variables, etc) as points along one or several axes. These axes are often chosen so as to maximise the variance of the plotted points and so as to be orthogonal to all preceding axes. The axes are usually found as eigenvectors from an eigenvalue decomposition of the original data, after some transformation.

We will briefly describe the underlying mathematical basis of the ordination methods COA and CIA, following the notation of Dolédec and Chessel [7] and of the ADE-4 package [17]. These utilise the model of the duality diagram which is based on the concept of a statistical triplet. A statistical triplet is composed of three matrices ($\mathbf{X}, \mathbf{D_c}, \mathbf{D_r}$), a data matrix $\mathbf{X}$ (having n rows/cases and p columns/variables) with possibly an appropriate transformation, and two diagonal matrices of column and row weights $\mathbf{D_c}$, $\mathbf{D_r}$ which will be defined below. When $n <p$, the principle of the method is the diagonalisation of a $n \times n$ matrix $\mathbf{B}$ defined as:

$$\mathbf{B} = \mathbf{D_c}^{1/2}\mathbf{X}\mathbf{D_r}\mathbf{X}^t\mathbf{D_c}^{1/2} \quad (1)$$

where $\mathbf{X}^t$ is $\mathbf{X}$ transposed and $\mathbf{D_c}^{1/2}$ is $\mathbf{D_c}$ with the square root of each diagonal element along the diagonal. The diagonalisation of $\mathbf{B}$ gives n eigenvalues corresponding to the n principal axes.

In the case of COA, the original $n \times p$ table of genes and arrays is transformed into a table of chi-square values giving the association or correspondence between each gene and each array. Let $\mathbf{M}$ be our matrix containing the raw data, this matrix having n rows and p columns. We can write $\mathbf{M} = [m_{ij}]$ with $1 \leq i \leq n$ and $1 \leq j \leq p$. We denote the row and column sums of $\mathbf{M}$ as $m_{i\bullet}$ and $m_{\bullet j}$ respectively, $m_{\bullet\bullet}$ corresponding to the grand total. The relative contribution or weight of row i to the total variation in the data set is then denoted r_i and is calculated as:

$$r_i = m_{i\bullet}/m_{\bullet\bullet} \quad (2)$$

while the relative contribution of column j is denoted as c_j and is calculated as:

$$c_j = m_{\bullet j}/m_{\bullet\bullet} \quad (3)$$

Similarly, the contribution of each individual element of $\mathbf{M}$ to the total variation in the data set is denoted as f_{ij} and is calculated as:

$$f_{ij} = m_{ij}/m_{\bullet\bullet} \quad (4)$$

The above calculations produce two vectors $\mathbf{R} = [r_i]$ and $\mathbf{C} = [c_j]$ of length n and p respectively, and one matrix $\mathbf{F} = [f_{ij}]$ of dimension $n \times p$. We use these vectors and this matrix to determine the values of x_{ij}, which are calculated as:

$$x_{ij} = \frac{f_{ij}}{r_i c_j - 1} \qquad (5)$$

These values define the matrix $\mathbf{X} = [x_{ij}]$, which along with the diagonal matrices $\mathbf{D_r}$ (an $n \times n$ matrix of zeros with the elements of $\mathbf{R}$ along the diagonal) and $\mathbf{D_c}$ ($p \times p$ matrix with the elements of $\mathbf{C}$ along the diagonal) are used for COA computation as described in equation 1. This analysis results in a series of axes (the eigenvectors of the decomposition) ranked by eigenvalue, on which the arrays can be plotted. COA is of particular interest because one can also add the positions of the variables (the genes) on the plot and examine the relationships between these and the arrays. An array and a gene that have a strong association have a high chi-square value in table $\mathbf{X}$ and will be plotted in a similar direction from the origin of the plot.

With CIA, we have two statistical triplets from two datasets, which we wish to analyse:

$$(\mathbf{X}, \mathbf{D_{cx}}, \mathbf{D_r}) \text{ and } (\mathbf{Y}, \mathbf{D_{cy}}, \mathbf{D_r})$$

These are from two datasets, x and y, which contain the same number of rows (arrays in this case) with the same row weights ($\mathbf{D_r}$), but may have different numbers of columns (two different sets of genes) with different column weights ($\mathbf{D_{cx}}, \mathbf{D_{cy}}$). Tables $\mathbf{X}$ and $\mathbf{Y}$ are the chi-squared tables derived from the two raw datasets as described equation (5). CIA then proceeds by an eigenvalue decomposition of the triplet ($\mathbf{Y}^t\mathbf{D_r}\mathbf{X}, \mathbf{D_{cx}}, \mathbf{D_{cy}}$), using equation (1). The details for deriving the co-inertia axes corresponding to the two datasets and the proof that these are maximally co-variant are given in Dolédec and Chessel [7]. The derivation of these axes is also described by Dray et al., [13,18] and in additional file 6 [see Additional file 6].

This produces two sets of axes, one from each dataset, where the first pair of axes are chosen so as to be maximally co-variant and represent the most important joint trend in the two datasets. The second pair of axes are chosen as to be maximally co-variant but orthogonal to the first pair, and so on for the rest of the axes. We can measure the similarity between the ordinations in two ways. The simplest is to measure the correlation between the data points on any two corresponding axes, one from each ordination. Additionally we measure the overall similarity using a multivariate extension of the Pearson correlation coefficient called the RV-coefficient [19]. The RV-coefficient is calculated as the total co-inertia (sum of eigenvalues of a co-inertia analysis) divided by the square root of the product of the squared total inertias (sum of the eigenvalues) from the individual COAs. It has a range 0 to 1 where a high RV-coefficient indicates a high degree of co-structure.

The main result of the analysis is then a pair of plots, one from each dataset, with the arrays plotted out on the first 2 or 3 axes. These plots should show similar arrangements of the arrays if the datasets have strong joint trends. A simple graphical device is to superimpose the plots for the first two axes of the analysis from the two datasets. If the sample (array) scores are normalised to unit variance along each axis, the standardised scores can be superimposed. Then the location of each data point (each array) can be indicated using an arrow. The tip of the arrow is used to show the location in one plot and the start of the arrow shows the location in the other. If the datasets agree very strongly, the arrows will be short. Equally, a long arrow demonstrates a locally weak relationship between the two sets of variables for that case (array). This is the rationale behind the plots in Figures 1, 2 and 4. In Figure 4, we also plot the locations of the variables (genes) in the two plots.

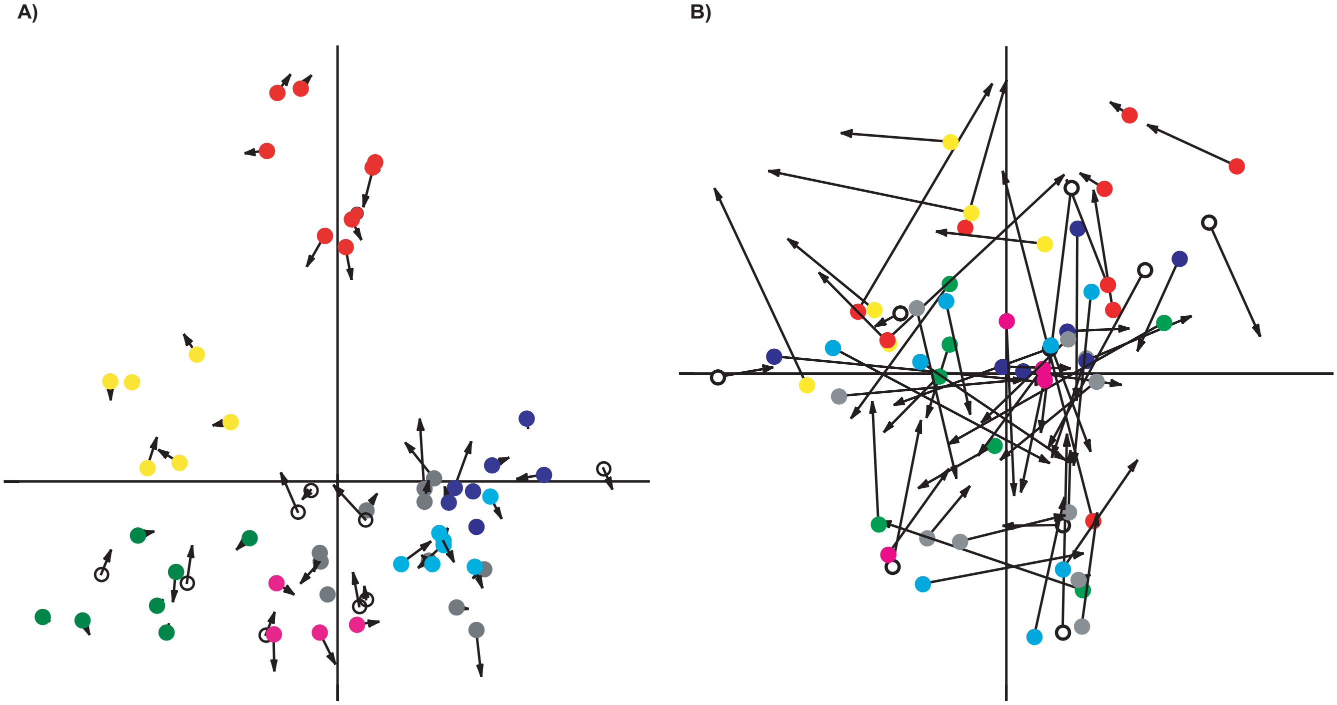

Figure 1
Analysis of very similar and unrelated gene expression datasets using CIA. The first two axes of control CIA studies of very similar (A) and unrelated (B) profiles of Ross spotted cDNA gene expression data of the NCI 60 panel of cell lines are shown. The figure shows results from CIA of A) two random gene subsets of the 1375 gene dataset B) two unrelated datasets composed of 1375 genes, where the 60 cell dataset was duplicated and the arrays in one dataset were randomly permutated. Circles and arrows represent the projected co-ordinates of each dataset, and these are joined by a line, where the length of the line is proportional to the divergence between the datasets. The colours represent the eight NCI60 cell line classes as defined by Blower et al., [21].

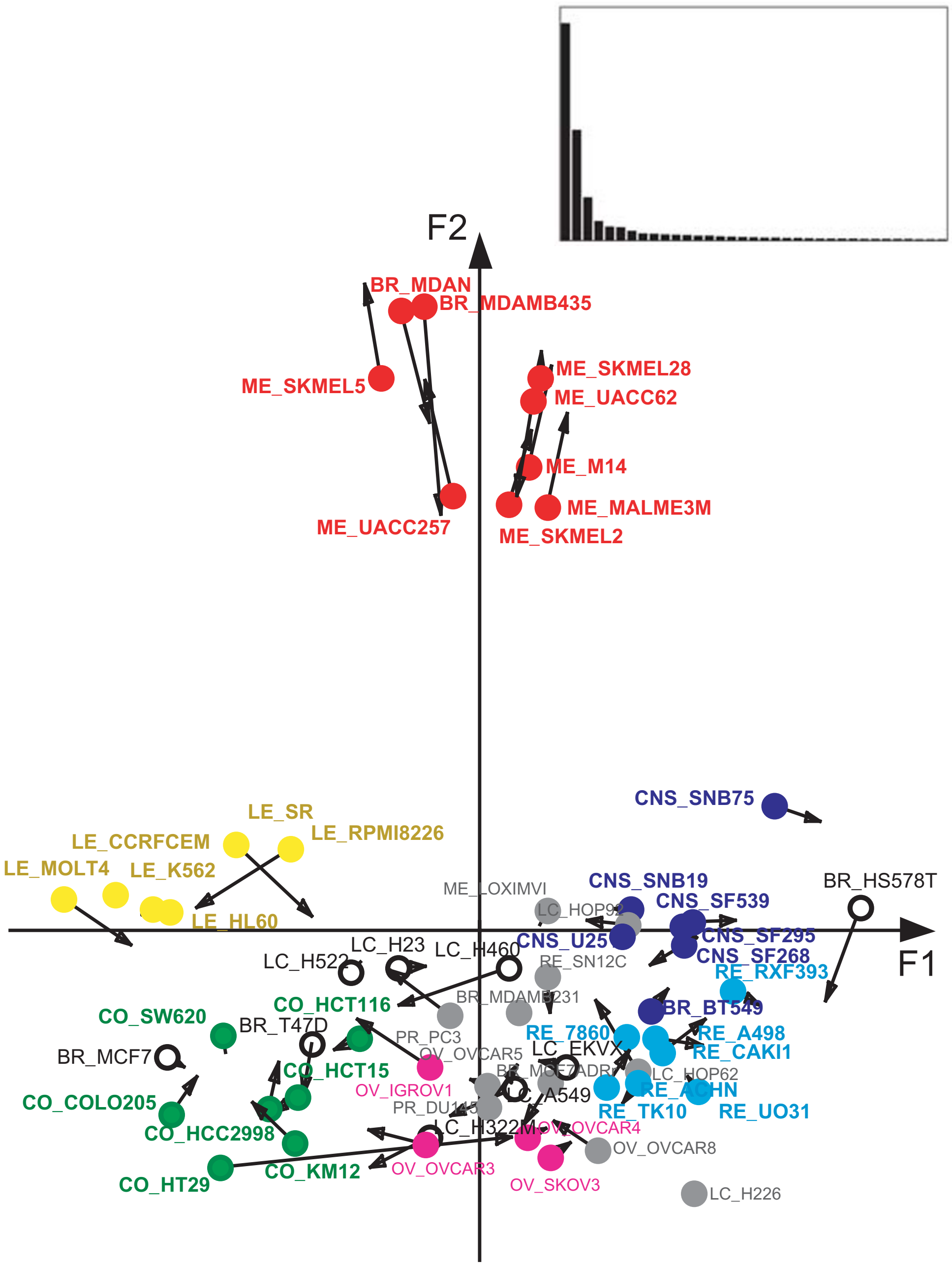

Figure 2

Cross-platform comparison of Affymetrix and spotted cDNA expression profiles using CIA. The first two axes of a CIA of gene expression profiles of the complete gene set from the Ross spotted cDNA array dataset (closed circles) and 1517 genes from the Staunton Affymetrix dataset (arrows) are shown. Circles and arrow represent the projected co-ordinates of each dataset, and these are joined by a line, where the length of the line is proportional to the divergence between the different gene expression profiles. The cell lines are coloured as in Figure 1. The cell lines are derived from breast (BR), melanoma (ME), colon (CO), ovarian (OV), renal (RE), lung (LC), central nervous system (CNS, glioblastoma), prostate (PR) cancers and leukaemia (LE). Colon and leukaemia cells were separated from those with mesenchymal or stromal features (glioblastoma and renal tumour cell lines) on the first axis (F1, horizontal), and melanoma cell lines were distinguished from the other cell lines on the second axis (F2, vertical). A histogram of the main factors which explain the total variability of this CIA is superimposed on the top right corner. The first three axes represented 42%, 21% and 8% of the inertia.

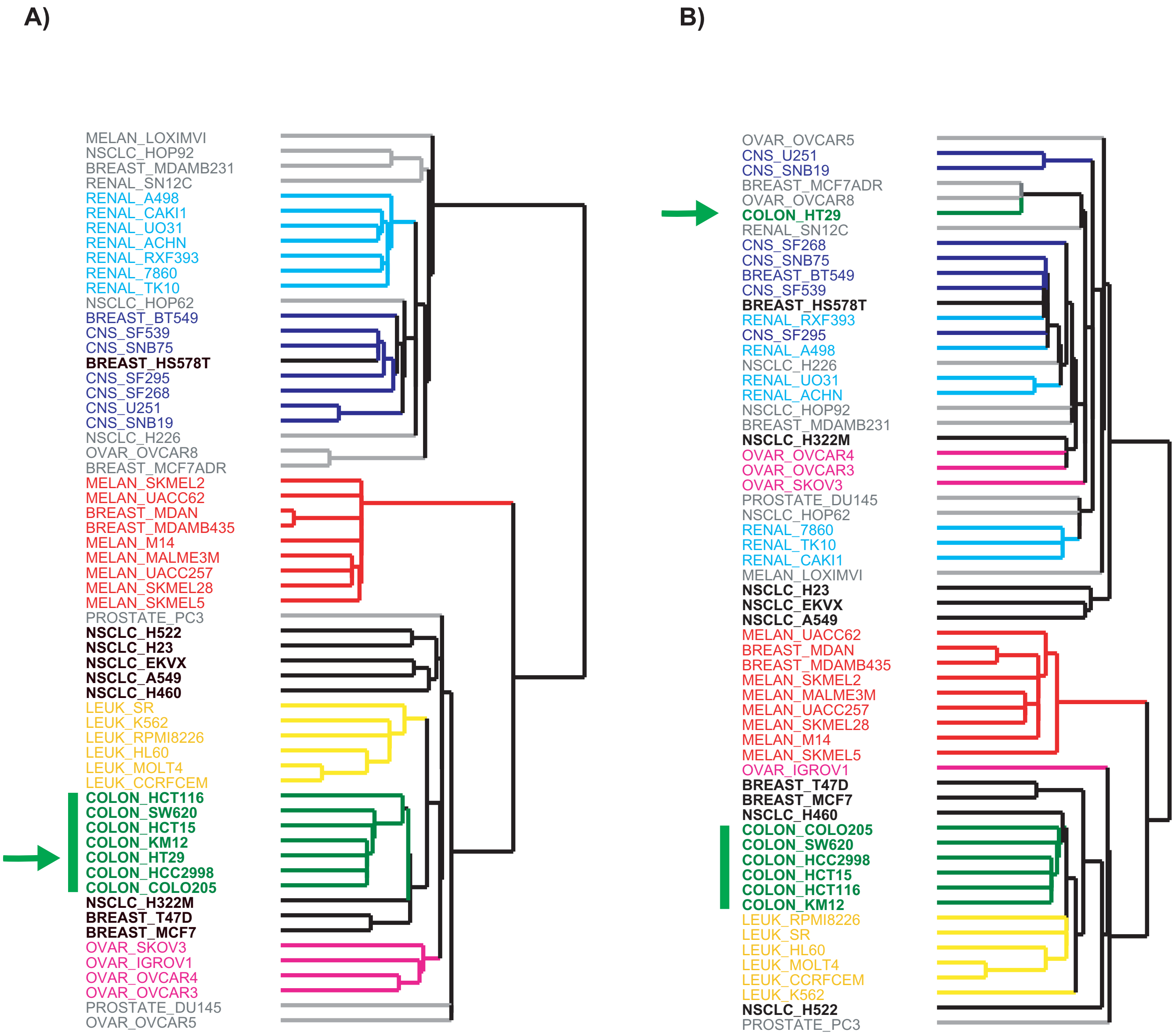

Figure 3

Hierarchical clustering of Affymetrix and spotted cDNA expression profiles of 60 cell lines. Dendrograms showings average linkage hierarchical clustering of NCI60 human cancer cell lines using Spearman Rank correlations. Cluster analyses of the 60 cell lines based on A) gene expression profiles of 1415 genes from the Ross spotted cDNA array dataset and B) 1517 genes from the Staunton Affymetrix dataset are shown. The cell lines are coloured as in Figure 1. The colon tumour cell line HT29 and cluster of colon tumour cell lines are highlighted by a green arrow and bar respectively.

Results

CIA of randomised datasets

A number of control studies of CIA of gene expression data were performed. We wished to establish what happens when datasets that are artificially similar or artificially distinct are compared. Firstly, we took the 1375 gene subset of the Ross dataset (described in the Methods section) and split it in two (by randomly assigning genes to one split or the other). This provided two datasets which have different collections of genes but which are expected to show similar patterns and trends. A graphical representation of results from this CIA of these datasets is shown in Figure 1a. Each sample (array of a cell line) is defined by an arrow where the head of the arrow marks

Table 1: Results of CIA of different subsets of gene expression datasets

Number of genes in each dataset		Matchminer results	Results of Coinertia Analysis on two datasets				
Ross-cDNA*	Staunton-Affymetrix**	Number of "matched" genes ±	RV coefficient	% Inertia		Correlation of ordinations	
				F1	F2	F1	F2
5643	3144	1416	0.85	40	61	0.96	0.97
	2455	1169	0.86	40	61	0.96	0.97
	1517	776	0.88	42	63	0.96	0.98
3748	3144	786	0.86	30	49	0.96	0.97
	2455	625	0.87	31	50	0.97	0.97
	1517	388	0.86	32	51	0.97	0.97
1415	3144	-	0.83	38	62	0.95	0.96
	2455	-	0.85	38	62	0.95	0.97
	1517	-	0.86	40	64	0.95	0.97
1375	3144	433	0.83	38	62	0.95	0.96
	2455	370	0.84	37	62	0.95	0.97
	1517	269	0.86	40	64	0.95	0.97

Gene expression data subsets from *spotted cDNA [16] and **Affymetrix [15] were subjected to CIA, where COA was performed on the Affymetrix dataset, and row weighted COA on spotted cDNA array dataset. Results of the co-inertia analysis show the RV co-efficient, accumulated inertia (% of total sum of eigenvalues of co-inertia analysis), and correlation between the coordinates on first pair (F1) and second pair (F2) of axes. ± Probes (sequence spots on each array) were matched using MatchMiner [22]. The 1415 cDNA subset contained the 1375 cDNA geneset and 40 extra genes for which no image identifier was given, thus matchminer counts for these 40 extra genes could not be determined, but the number of probes matched should be similar to the 1375 cDNA gene set results.

the position of the sample according to one ordination, and the end of the arrow indicates the sample position in the second ordination. The arrows are short and randomly oriented. The two pairs of projection coordinates are highly correlated (R = 0.99 between the two sets of coordinates on the first axes F1). The overall similarity in the structure of the datasets was very high resulting in a RV coefficient of 0.97. Clearly, CIA is able to detect and highlight the similarity between these subsets, despite the fact that they have practically no variables in common.

Secondly, the effect of comparing two unrelated datasets using CIA was assessed. The same Ross dataset of 60 arrays and 1375 genes was duplicated and the arrays (cell lines) of one of these datasets were randomly permuted. Thus more or less all of the rows in these two datasets should be unrelated. The results of CIA analysis of these datasets are shown in Figure 1b. Long randomly orientated arrows connected samples and the RV coefficient was only 0.30 reflecting the lack of joint structure in these datasets.

Cross-platform comparison of gene expression data using CIA

Matching genes common across arrays using annotation

Currently, meta-analyses of microarray gene expression data are usually based on cross-referencing each spot represented on the arrays, extracting genes common to all arrays and examining the correlation between the expression profiles of only these genes. Several subsets of the Ross spotted cDNA expression dataset have been selected in different studies [16,20,21]. The number of genes common across these and the subsets of the Staunton Affymetrix datasets (described in more detail in the Methods section), were compared using MatchMiner [22]. MatchMiner matched the IMAGE clone identifiers of genes represented on the cDNA arrays with GenBank accession numbers of oligonucleotide sequences attached to the Affymetrix array. The number of "matched" or common genes across each of the data subsets is given in Table 1. Only 1416 genes were matched between the largest Ross (5643 genes) and Staunton (3144 genes) datasets.

Identifying the most covariant gene expression data subsets using CIA

The disadvantage of only examining genes present across all arrays is that data from biologically significant genes may be lost if a gene is not represented on all DNA microarray platforms examined. CIA does not require pre-filtering of genes to those present in all datasets. We applied CIA to compare gene expression profiles from the Ross and Staunton datasets. Each of the Ross datasets; the complete dataset of 5643 genes, along with the Blower [21] subset of 3748 genes, and the two Scherf [20] subsets of 1375 and 1415 genes, were compared to different subselections of genes from the Staunton dataset using CIA. These preprocessed data which were used to perform these analyses are available [see Additional file 1,2,3,4, 5].

The relationships between these datasets as described by the RV co-efficient after CIA is shown in Table 1. The correlations between the pairs of ordinations along the first (F1, horizontal axis) and second pair of axes (F2, vertical axis) are also shown. The results in Table 1 show that between 49% and 64% of the total variance (sum of the eigenvalues) are represented by the F1 and F2 in each analysis, and there is a high correlation between pairs of ordinations on each axis. CIA of the complete Ross dataset and the smaller Staunton subset of 1517 genes resulted in the highest RV co-efficient (0.88) among these data subsets examined. CIA results from this analysis are examined in detail below.

Visualising cross-platform consistencies and divergences using CIA

In Figure 2, the results of CIA co-structure analysis between the gene expression profiles of the two datasets are shown. According to the eigenvalue histogram, the first three axes accounted for 42%, 21% and 8% of the explained variance respectively. Thus 63% of the variance of the co-inertia analysis was accounted for by the first and second co-inertia axes and thus presented a good initial summary of the co-structure between the two datasets. The correlation (R value) between the first axes (F1) of the two ordinations was 0.96, and it was 0.98 between second axes (F2) of the two ordinations. These high values partly result from the maximisation of the covariance, ie the product of the correlation and the squared variances projected onto the co-inertia axes. Thus a Monte Carlo permutation test, where the rows of one matrix are randomly permutated followed by a re-computation of the total inertia [23] was used to check the significance of co-structure of this CIA. A total of 1000 co-inertia analyses using random matching of the two tables were processed. Permutation analysis of these 1000 datasets showed that the observed inertia was much greater than that of the simulated datasets. The probability of obtaining a total inertia equal to that observed, using the hypothesis of independence between the gene expression datasets, was less than 0.001. This underlines that the two tables are significantly related and a co-structure exists.

In the CIA plot of Figure 2, the co-ordinates of the 60 cell lines from both the Ross (circles) and Staunton (arrows) datasets are connected by a line, the length of which indicates divergence between the two datasets. The first axes (the horizontal F1 axes from the two data sets) separated leukaemia cells and colon cells with epithelial characteristics, from cells with mesenchymal or stromal features such as the glioblastoma and renal tumour cell lines. We inferred that the second axis (F2, vertical) is the melanoma axis, separating the melanoma cell lines from the other cell lines.

Cell lines from non-small cell lung carcinomas and breast cancers were distributed in multiple clusters indicating that their gene expression patterns were more heterogeneous. For example, we observed that the breast cancer cell line Hs578T clustered (was geometrically close to) with the stromal/mesenchymal cluster of glioblastoma and renal tumour cell lines at the positive end of the F1 axes. By contrast, the breast cancer cells MCF-7 and T47D were projected at the opposite end of the F1 axes, closer to the colon cancer cells which have an epithelial phenotype. These observations agree with previous findings [16].

For most cell lines, the divergence between the Ross and Staunton gene expression profiles was little above background noise. However the colon tumour cell line HT29 was represented by a long arrow, indicating that there were significant cross-platform differences between the expression profiles of this cell line. In the Ross ordination, the cell line HT29 clustered with the other colon tumour cell lines, but in the Staunton ordination it shifted significantly. Hence, we performed an independent evaluation using hierarchical cluster analysis (Figure 3). This analysis verified that the HT29 cell line clustered within the colon cell lines cluster when the Ross data but not the Staunton data were analysed. No single gene was responsible for the shift between ordinations of HT29.

Each projection of cell lines was defined by the expression of specific genes. A summary of a number of genes that were identified using CIA on each of the axes is given in Table 2 and plots showing the coordinates of genes that defined the first two axes of the CIA are shown in Figure 4. The genes most responsible for defining the axes are located at the ends of the axes. Genes and cell lines which project in the same direction from the origin have a strong association and represent genes whose expression is increased or upregulated in these cell lines. Equally genes projected in the opposite direction from the origin to cell lines are frequently genes that are lost or down regulated in those cell lines.

In Figure 4 the most extreme genes from the ends of each axis are labelled. Genes labelled in red are those that were present in the top 30 genes at the ends of F1 and F2 and were "matched" across platforms, that is where an IMAGE clone identifier of a spotted cDNA clone and a GenBank accession number of an Affymetrix oligonucleotide probe set mapped to the same UniGene cluster. Of the 1416 genes "matched" between these two datasets (Table 1),

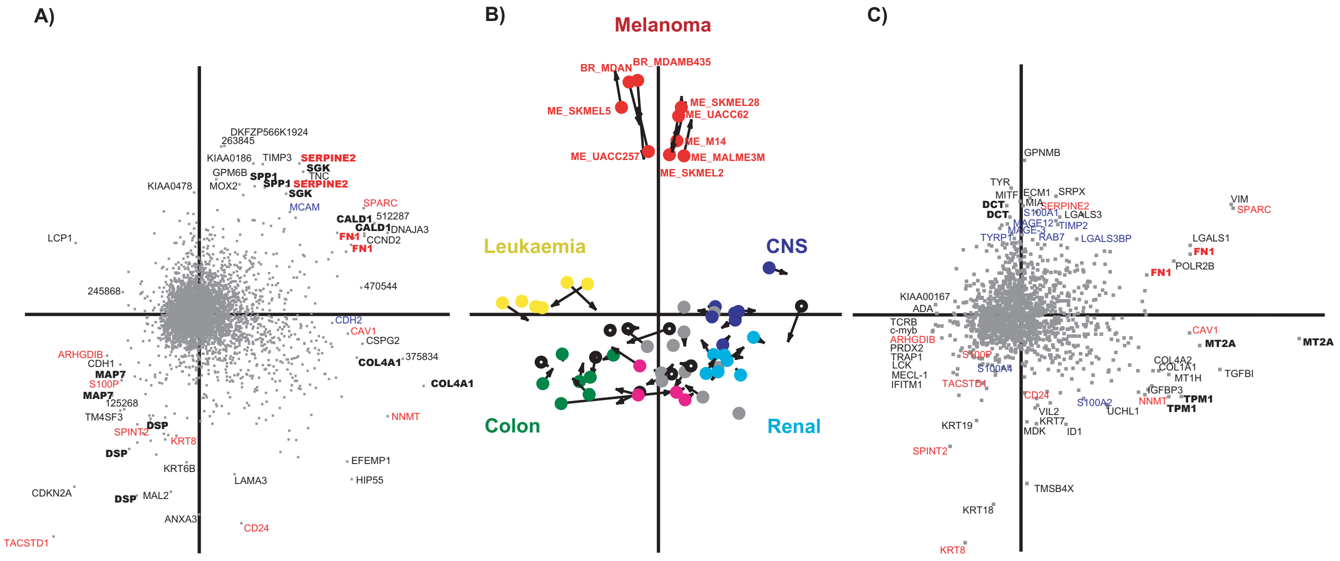

Figure 4

Detecting genes defining major trends identified using CIA. The central panel (B) is the CIA from Figure 2. The co-ordinates of the genes in each ordination are shown in the side panels A) Ross cDNA and C) Staunton Affymetrix. The top ten genes at the end of axes F1 and F2 are labelled, where red gene labels indicate genes that were present in both datasets. Genes labelled in bold describe genes that were replicated on the microarray. Genes labelled in blue represent genes that were not contained in the top ten genes, but were in the top thirty genes at the end of each axes and are of biological interest.

only 11 "matched" genes were projected within the top 30 genes at the ends of the F1 and F2 axes in both ordinations. Although only 11 of 120 genes at the ends of the F1 and F2 axes were matched, many top genes of one ordination were present in the second dataset, but were not projected at the ends of these axes. Among the top 120 genes in the Staunton Affymetrix ordination, 53 were present in the Ross spotted cDNA dataset. Equally 40 of the top 120 genes detected in the Ross ordination were present in the Affymetrix dataset. This observation that several genes present on both arrays were only associated with trends in one ordination, could highlight annotation problems, differences in binding properties between the oligonucleotide and cDNA probes representing these genes or measurement error in one or more datasets.

The observation that the majority of genes associated with trends were represented on only one array type is significant, as these would have been excluded from analysis if standard "annotation based" methods were used. Thus gene expression data from each platform are co-visualised using CIA. We examined the genes defining each axis in the Ross or Staunton ordinations in more detail.

Epithelial versus mesenchymal clusters of cell lines on the first axis

The first axis clearly distinguished cells with epithelial versus mesenchymal characteristics. The epithelial to mesenchymal transition (EMT) is an ancient pathway integral to normal embryonic development and is implicated in the progression of malignancy of epithelial cancers such as breast and colon carcinomas [24]. During EMT, cells acquire a morphology that is appropriate for migration and thus understanding the processes that trigger EMT may help in refining our knowledge of the biological basis of tumour progression to metastasis.

Epithelial genes were projected in the same direction as the less invasive carcinoma cell lines. The breast carcinoma cell lines MCF-7 and T47D, which have a pure luminal phenotype, were projected onto the epithelial side of the F1 axis, whereas the more invasive breast cancer MDA MB231 was projected onto the mesenchymal end of the F1 axis. This ordination agrees with recent immunohistochemical studies on these tumour cell lines [25].

The genes at the mesenchymal end of the first pair of CIA axes included *TGFβ*, *N-cadherin*, along with several mus-

Table 2: Selection of genes identified using CIA

Axis	Cell lines	Genes*	Description	Spotted cDNA	Affymetrix
F1 (mesenchymal)	All CNS, Renal cells and the breast cancer cell line BR-Hs578T	*COL1A1*	Collagen marker	-	+
		COL4A1	Collagen marker	+	-
		COL4A2	Collagen marker	-	+
		TPM1	Muscle marker	-	+
		VIM	Vimentin	-	+
		FN1	Fibronectin 1	+	+
		TGFβ	Inducer of EMT	-	+
		CDH2	N-cadherin	+	-
		MT2A	Metallothionein A2-associated with invasive breast cancer	-	+
F1 (epithelial)	All colon cells and the breast cancer cells MCF-7 and TR7D	*CDH1*	E-cadherin, primary epithelial marker	+	-
		SPINT2	Serine protease inhibitor, Kunitz type, 2 an inhibitor of hepatocyte growth factor	+	+
		KRT8	Keratin 8, epithelial marker	+	+
		KRT18	Keratin 18, epithelial marker	-	+
		KRT19	Keratin 19, epithelial marker	-	+
		DSP	Desmoplakin I, epithelial marker	+	-
		S100A2	Loss of S100A2 early event in melanoma development	-	+
F1 (colon cell markers)		*TACSTD1*	Ep-Cam. Target antigen in colorectal carcinoma	+	+
		CDKN2A	Target antigen in colorectal carcinoma	+	-
F1 (Leukaemia)	All leukaemia cell lines	*ARHGDIB*	A lymphoid-specific guanosine diphosphate dissociation inhibitor	+	+
		LCP1	Lymphocyte cytosolic protein 1, L-plastin	+	-
		IFITM1	An interferon induced transmembrane protein	-	+
F2 (Melanoma)	All melanoma cells and the breast cancer cells BR_MDA and BR_MDAMB435	*MITF*	Microphthalamia-associated transcription factor	-	+
		TYR	Tyrosinase	-	+
		DCT	Dopachrome tautomerase	-	+
		TYRP1	Tyrosinase-related protein 1	-	+
		RAB7	Ras-associated protein 7	-	+
		MIA	Melanoma inhibitory activity	-	+
		MCAM	MUC18, melanoma cell adhesion molecule MCAM	+	-
		MAGE 3	Melanoma-associated antigen 3	-	+
		MAGE 12	Melanoma-associated antigen 12	-	+
		GPNMB	Glycomembrane protein nmb	-	+
		TIMP2	Tissue inhibitor of metalloproteinase 2	-	+
		TIMP3	Tissue inhibitor of metalloproteinase 3	+	-

Genes identified on the first (F1) and second (F2) axes, where + or - indicated whether a gene was detected or not detected within the top 30 genes at the ends of each of these axes in CIA of Affymetrix and spotted cDNA array gene expression profiles of the NCI60 cell lines. These genes are graphically presented in Figure 4 and further details on these genes are available in the Results section. *Official gene symbol names are used for each gene.

cle, collagen and mesenchymal markers, such as *vimentin* and *fibronectin* (Table 2). At the opposite end of this axis, several markers of epithelially-derived genes, including *E-cadherin*, the *cytokeratins* 8, 18 and 19, as well as *desmoplakin I* were observed.

Although a number of these genes were present in both the Staunton and Ross ordinations, the majority were in one of the two datasets only (Table 2). In the Ross ordination, *E-cadherin* and *N-cadherin* were projected at opposite ends of the F1 axis. *E-cadherin* maintains the integrity of epithelial tissue and is considered the primary "caretaker" gene of the epithelial phenotype. Loss of *E-cadherin* is heavily implicated in EMT. Loss of *E-cadherin* is accompanied by loss of epithelial keratins and gain of mesenchymal *vimentin* and *fibronection*, as well as progression of malignant carcinoma [24]. *N-cadherin* is gained in some carcinomas that have lost *E-cadherin* and this has been associated with reduced five year survival in patients with non-small cell lung cancer [26]. We also observed that *metallothionein A2* was strongly associated with the mesenchymal side of the F1 axis in the Affymetrix dataset ordination and this has shown to be implicated with invasive ductal breast carcinoma [27]. Both hepatocyte growth factor (*HGF*), and *TGF*β have been shown to induce EMT, and colon cancers that lack receptors to *TGF*β have a better prognosis [28]. *TGF*β and *vimentin* were identified in the Staunton Affymetrix data. *SPRINT2*, an inhibitor of an inhibitor of *HGF*, was detected at the epithelial end of the F1 axis in both ordinations. These genes are integral to EMT and thus the merging of such information from both of these datasets using CIA is noteworthy.

Genes associated with the colon cell and leukaemia cell line clusters

The first axis distinguished CNS/renal tumour tissue derived cell lines from those having their origin in either leukaemia or colon cancer. Although the leukaemia and colon tumour cell lines appear close together on the first axis, these were separated to either end of the third axis, thus, genes defining each of these cell types could be identified.

Two genes, tumor-associated calcium signal transducer 1 (*TACSTD1*) and cyclin-dependent kinase inhibitor 2A (*CDKN2A, p16*), a tumour suppressor gene, were strongly associated with the colon tumour cell lines in the Ross spotted cDNA array data ordination. *TACSTD1* also featured on the Staunton Affymetrix ordination. *TACSTD1* is a cell adhesion molecule expressed on the majority of tumour cells in most patients with colorectal carcinoma and, interestingly, was the target of one of the first mouse monoclonal antibodies produced for therapeutic use. Several clinical trials are ongoing using *TACSTD1/CO17-1A/EpCam* as a target antigen in colorectal carcinoma [29]. We observed that increased gene expression of *CDKN2A* was associated with the colon tumour cell lines, although hypermethylation of *CDKN2A* has been correlated with poor prognosis of patients with colorectal cancer [30].

Genes that are expressed preferentially in haematopoietic tissues defined the leukaemia cluster. *ARHGDIB*, a lymphoid-specific guanosine diphosphate dissociation inhibitor, was strongly associated with the leukaemia cell line cluster and was present on both microarray platforms. In addition, a number of genes that distinguished the leukaemia cluster were only present on one of the two DNA microarray platforms. Lymphocyte cytosolic protein 1 (*L-plastin, LCP1*) was represented in the spotted cDNA array dataset, but not in the Affymetrix array subset. *LCP1* encodes an actin-binding protein and is situated at 3q27, a locus associated with a translocation event t(3;13)(q27;q14) found in various types of non-Hodgkin's lymphoma [31]. In the Affymetrix ordination, T-cell receptor *TRCB*, and an interferon induced transmembrane protein (*IFITM1*) which has been implicated in the control of cell growth and deregulation, were among the genes associated with the leukaemia cluster.

Melanoma cell lines clustered with two metastases BR_MDAN and BR_MDAMB435

We observed an interesting trend within the melanoma cell line cluster, which contained seven melanoma cell lines, as well as BR_MDAN and BR_MDAMB435, two melanoma metastases which were derived from a patient diagnosed with breast cancer. In the ordination of the Ross dataset, these two "breast cancer" cell lines were furthest along the second axis. However, the melanoma cell lines were projected further along this axis in the ordination of the Staunton gene expression data. This indicated that the Affymetrix gene expression profiles contained more information on the melanoma cell lines compared to the two metastases which were not as discriminated on the axis. Thus, we examined the melanoma-specific genes represented in each dataset.

Diagnosis of melanoma is normally associated with a neoplasm that is *keratin* negative, and is positive for *vimentin, S100* and *HMB-45*, though *MITF* and *Melan-A* were reported recently to be superior markers to *S-100* and *HMB-45* [32].

These melanoma-specific genes were very well represented on the Staunton Affymetrix ordination. We observed expression of *vimentin* and *MITF*, as well as other genes associated with pigmentation/differentiation (*TYR, DCT, TYRP1, MITF, RAB7*), several serum markers of melanoma progression (*MIA, MAGE 3* and *MAGE 12*) and glycomembrane protein nmb (*GPNMB*) in this ordination. Expression of *GPNMB* has been shown to be inversely correlated with the metastatic potential of melanoma cell lines [33]. In addition, on the negative end of this axis, *keratins* 8, 18 and 19, along with *S100A2* were observed. Absence of these keratins is used in clinical diagnosis of melanoma and loss of *S100A2* gene expression has been implicated as an early event in melanoma development [34]. Thus, the melanoma phenotype was well represented on the Affymetrix ordination.

By contrast, there were considerably less melanoma-specific genes in the Ross dataset. Expression of melanoma cell adhesion molecule *MCAM* (also called *MUC18*), which reportedly correlates directly with the metastatic potential of human melanoma cells, was detected in the Ross cDNA ordination. In addition, *keratin* 8 was projected onto the negative end of the F1 axis in the Ross ordination. Although Ross et al. [16] identified *TYR, S100*β and *DCT* as melanoma associated genes, these were subsequently excluded in the revised release of their dataset (see Methods section) and were thus not identified in this analysis.

Discussion

CIA is a particularly attractive method for visually relating multiple microarray gene expression datasets. CIA is a data coupling approach that identifies trends or patterns in tables of data that contain the same samples. In this paper CIA is applied to the cross-platform analysis of relationships in gene expression profiles of 60 cell lines, rather than to the analysis of specific genes. This is an attractive feature of CIA. Since CIA maps two gene expression datasets at the data, not the annotation level, it is not limited by the immaturity of gene annotation. Secondly as CIA can accept data where the number of variables exceeds the number of individuals, filtering of data to those genes represented on all arrays is not required, and thus more genes are available for analysis. An earlier report which attempted to correlate these datasets reported disappointingly poor correlations between gene datasets [4]. Kuo and colleagues [4] used the BLAST algorithm to sequence match genes represented on both array platforms. Of the 9,703 cDNA probes on the spotted cDNA array, in question, and 7,245 probes sets of the Hu6800 Affymetrix arrays, 2,895 spots/probe sets were found to be sequence-matched. However analysis of this filtered set of data showed poor cross-platform concordance.

In our analyses, the divergence between the Ross and Staunton gene expression profiles of most cell lines was little above background noise, however, we detected a large variation between the expression patterns of the colon tumour cell line HT29. The melanoma cell lines were more defined in the Affymetrix ordination than in the ordination from the Ross dataset. This may be due to the increased numbers of melanoma associated genes in this dataset. Thus, CIA can be used to highlight lack or presence of co-structure between datasets. Moreover, CIA can assist in the selection of the strongest features from each datasets for subsequent analysis.

Several clinically significant genes were detected in the CIA of the Ross and Staunton data. The first axes were associated with the characteristics of epithelial and mesenchymal phenotypes. Mesenchymal cells possess migratory and invasive properties typical of malignant metastasising cancer, and thus the transition between epithelial and mesenchymal phenotypes is a key field in cancer biology [24]. Carcinoma cell lines with more invasive phenotypes were associated with the mesenchymal end of the axis. We were easily able to identify several of the most important genes associated with both the epithelial (*keratins 8, 18 and 19, E-cadherin, SPINT2*) and mesenchymal (*TGF*β, *vimentin and fibronectin*) cell types. Although a number of defining genes were present on both arrays (*keratin 8, fibronectin*), the majority of genes were present only on one array (Table 2). Thus, given a strong association, CIA provides an opportunity to assimilate data from different gene expression sources. Equally, on the second axes of the ordinations, which defined the melanoma phenotype, and the third axes, which distinguished the leukaemia cells, nearly all of the genetic markers detected were only present in one rather than both datasets, and thus these would have been lost if we had filtered our data to those genes present across all arrays.

CIA is very flexible and extensible [13]. It is suitable for analysis of quantitative, qualitative or even fuzzy variables. It allows coupling of two tables which can be subjected to various transformations and/or centering (COA, PCA etc) with the only constraint being that the samples (arrays) are weighted in the same way for the two analyses.

Conclusion

We believe CIA is a very useful method for cross-platform comparison of gene expression profiles where the same tissue or cell lines have been arrayed multiple times. Consensus and divergence between gene expression profiles from different DNA microarray platforms are graphically visualised. Importantly, this method is not dependent on probe or sequence annotation, and thus it can extract important genes even when there are not present across all datasets.

Methods

Datasets

The NCI 60 series consists of a panel of 60 human tumour cell lines derived from patients with leukaemia, melanoma, along with, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers. This panel has been subjected to three different DNA microarray studies using Affymetrix [14,15] and spotted cDNA array [16] technology. We compared two of these studies, one cDNA spotted [16] and one Affymetrix [15] study and refer to them as the Ross and Staunton datasets respectively. These pre-processed data are available in additional data files [see Additional file 1,2,3,4].

The Ross Dataset

The Ross dataset contained gene expression profiles of each cell lines in the NCI-60 panel, which were determined using spotted cDNA arrays containing 9,703 human cDNAs. The data were downloaded from **The NCI Genomics and Bioinformatics Group Datasets resource** http://discover.nci.nih.gov/datasetsNature2000.jsp. The updated version of this dataset (updated 12/19/01) was retrieved. Data were provided as log ratio values. In this study, rows (genes) with greater than 15% of values missing were deemed unreliable and were removed from analysis, reducing the dataset to 5643 spot values per cell line. Remaining missing values were imputed using a K nearest neighbour method, with 16 neighbours and a Euclidean distance metric [35]. This set of 5643 genes, along with subsets of 1375, 1415 [20] or 3748 genes [21] that were used in previous reports, were used.

The Staunton Dataset

These data were derived using high density Hu6800 Affymetrix microarrays containing 7129 probe sets. The dataset was downloaded from the **Whitehead Institute Cancer Genomics supplemental data to the paper from Staunton et al.**, http://www-genome.wi.mit.edu/mpr/NCI60/, where the data were provided as average difference (perfect match-mismatch) values. As described by Staunton et al., [15], an expression value of 100 units was assigned to all average difference values less than 100. Genes whose expression was invariant across all 60 cell lines were not considered, reducing the dataset to 4515 probe sets. Gene subsets where the minimum change in gene expression across all 60 cell lines was greater than 100, 200 and 500 average difference units were selected resulting in subsets of 3144, 2455, and 1517 probe sets. Data were logged (base 2) and median centred.

Computation of CIA

Computation of CIA was performed using the ADE-4 package [17], a general-purpose package for multivariate statistical analysis, which has been widely used in the analysis of environmental and ecological data. It runs under MacOS 7 or Windows operating systems and can be downloaded from **The ADE-4 homepage** http://pbil.univ-lyon1.fr/ADE-4/. In addition, ADE-4 is available as routines written in the R statistical computing language. These can be downloaded from **The R homepage** http://cran.r-project.org/src/contrib/PACKAGES.html#ade4 or **The ADE-4 for R homepage** http://pbil.univ-lyon1.fr/ADE-4. R scripts to run CIA are available on request.

The ADE-4 modules required to perform CIA are ADEtrans, FilesUtil (using the Transpose option), PCA (Correlation Matrix PCA, Covariance matrix PCA options), COA (Correspondence Analysis, Row weighted COA), CoInertia (Match two statistical triplets, coinertia test, coinertia analysis). ADE-4 can be run interactively or in batch mode. Graphical displays were obtained using the ADE-4 modules Scatters and Scatterclass.

Cross-platform comparison of two microarray datasets using CIA

The labelling of the NCI-60 cell lines varied between the Ross and Staunton studies. The cell line labels were verified, matched and sorted so that the order of the arrays was the same in each analysis. Within the ADE4 implementation of CIA, it assumes that the row weights of both datasets are the same, thus for analysis of microarray data, the data was transposed. All data points in each dataset were made positive by the addition of a constant, as done by Fellenberg et al., [11] and Culhane et al., [12].

CIA was used to determine the main relationships between the gene expression profiles from the same 60 cell lines, but which were derived using two different microarray technologies. Each of the four subsets of the spotted Ross data and the three subsets of the Staunton data were subjected to analysis. COA was performed on each Ross dataset, and row weighted COA was performed on the gene expression data from the Staunton data, where row weights from the Ross analysis were used. The covariance of the rows (arrays) of the two chi-squared tables were then analysed using CIA.

Cross-platform comparison of two microarray datasets using annotation methods

A list of gene transcripts represented on both array platforms was determined by using BLAST [36] to compare sequences represented on each array. In addition IMAGE clone identifiers of spotted cDNA elements and GenBank accession numbers of genes detected by Affymetrix oligonucleotide probe sets were "annotation matched" via UniGene ID using MatchMiner [22]. SOURCE [37] was used to retrieve and update gene annotation.

Hierarchical clustering

Before applying clustering, rows and columns (genes and cell lines) of datasets were median centred and normalised to unity. We used average linkage cluster analysis to cluster cell lines and genes using the Spearman Rank correlation measure of similarity. Analyses were accomplished using the Cluster and Treeview programs [38].

Authors Contributions

AC conceived this study and carried out the analysis as a postdoctoral researcher in the group of DH. DH supervised the study and provided input both in the design of the study and drafting of the final manuscript. GP provided input regarding the interpretation of the methodology and results. All authors read and approved the manuscript.

Additional material

Additional File 1

Ross_5643.zip *is a Microsoft Excel file that is compressed using winzip 8.0. It contains the pre-processed Ross (spotted microarray) data subsets described in this manuscript. The excel file contains 5 worksheets; the first is a readme which gives further details of the data. In addition details of the data contained in this file are given in additional file 2 'Readme.txt'.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S1.zip]

Additional File 2

Staunton_7129.zip *is a Microsoft Excel file that is compressed using winzip 8.0. It contains the pre-processed Staunton (Affymetrix) data subsets described in this manuscript. The excel file contains 7 worksheets; the first is a readme which gives further details of the data. In addition details of the data contained in this file are given in additional file 2 'Readme.txt'.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S2.zip]

Additional File 3

Readme.pdf *is a pdf (abode) file, which describes the pre-processed data, contained in additional files 1,2,4.5*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S3.pdf]

Additional File 4

Ross_5643_KNN.txt *is a tab delimited plain text file. Ross_5643_KNN.txt is worksheet 4 from Ross_5643.zip. This 5643 gene subset of the Ross data is described in detail in the manuscript. The IMAGE clone identifiers are in the first column, and sample (array) names in the first row.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S4.txt]

Additional File 5

Staunton_1517_CS.txt *is a tab delimited plain text file. Staunton_1517_CS.txt is worksheet 3 from Staunton_7129.zip. This 1517 gene subset of the Staunton data is described in detail in the manuscript. The Affymetrix probe identifiers are in the first column, and sample (array) names in the first row.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S5.txt]

Additional File 6

Further details on the mathematical model of CIA
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-4-59-S6.pdf]

Acknowledgements

We wish to thank the three anonymous referees for their very helpful comments. The authors would also like to thank Dr. William Gallagher for useful discussions. We are grateful to the Health Research Board, Ireland for funding.

References

1. Holloway AJ, van Laar RK, Tothill RW, Bowtell DD: **Options available--from start to finish--for obtaining data from DNA microarrays II.** *Nat Genet* 2002, **32 Suppl:**481-489.
2. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat Genet* 2003, **33:**49-54.
3. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Res* 2002, **62:**4427-4433.
4. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18:**405-412.
5. Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, Cavalieri D, Gaasterland T, Hingamp P, Holstege F, Ringwald M, Spellman P, Stoeckert C. J., Jr., Stewart JE, Taylor R, Brazma A, Quackenbush J: **Standards for microarray data.** *Science* 2002, **298:**539.
6. Kulkarni AV, Williams NS, Lian Y, Wren JD, Mittelman D, Pertsemlidis A, Garner HR: **ARROGANT: an application to manipulate large gene collections.** *Bioinformatics* 2002, **18:**1410-1417.
7. Dolédec S, Chessel D: **Co-inertia analysis: an alternative method for studying species-environment relationships.** *Freshwater Biology* 1994, **31:**277-294.
8. Thioulouse J, Lobry JR: **Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package.** *Comput Appl Biosci* 1995, **11:**321-329.
9. Gittins R: **Canonical analysis, a review with applications in ecology. Vol.12 of Biomathematics.** *Berlin, Springer- Verlag*; 1985.
10. Ter Braak CJF: **Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis.** *Ecology* 1986, **69:**1167-1179.
11. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data.** *Proc Natl Acad Sci U S A* 2001, **98:**10781-10786.
12. Culhane AC, Perriere G, Considine EC, Cotter TG, Higgins DG: **Between-group analysis of microarray data.** *Bioinformatics* 2002, **18:**1600-1608.
13. Dray S, Chessel D, Thioulouse J: **Co-inertia analysis and the linking of ecological tables.** *Ecology* 2003, **84:**3078-3089.
14. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.** *Proc Natl Acad Sci U S A* 2000, **97:**12182-12186.
15. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci U S A* 2001, **98:**10787-10792.
16. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
17. Thioulouse J, Chessel D, Dolédec S, Olivier JM: **ADE-4: a multivariate analysis and graphical display software.** *Statistics and Computing* 1997, **7:**75-83.
18. Dray S, Chessel D, Thioulouse J: **Procrustean co-inertia analysis for the linking of ecological tables.** *Ecoscience* 2003, **10:**110-119.
19. Robert P, Escoufier Y: **A unifying tool for linear multivariate statistical methods: the RV-coefficient.** *Appl. Statist.* 1976, **25:**.
20. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN: **A gene expression database for the molecular pharmacology of cancer.** *Nat Genet* 2000, **24:**236-244.
21. Blower PE, Yang C, Fligner MA, Verducci JS, Yu L, Richman S, Weinstein JN: **Pharmacogenomic analysis: correlating molecular

substructure classes with microarray gene expression data. *Pharmacogenomics J* 2002, **2:**259-271.

22. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein JN: **MatchMiner: a tool for batch navigation among gene and gene product identifiers.** *Genome Biol* 2003, **4:**R27.
23. Thioulouse J, Cadet P, Albrecht A: **The use of permutation tests in co-inertia analysis : application to the study of nematode-soil relationships.** *Biometric Bulletin* 1996, **13:**10.
24. Thiery JP: **Epithelial-mesenchymal transitions in tumour progression.** *Nat Rev Cancer* 2002, **2:**442-454.
25. Fuchs IB, Lichtenegger W, Buehler H, Henrich W, Stein H, Kleine-Tebbe A, Schaller G: **The prognostic significance of epithelial-mesenchymal transition in breast cancer.** *Anticancer Res* 2002, **22:**3415-3419.
26. Nakashima T, Huang C, Liu D, Kameyama K, Masuya D, Kobayashi S, Kinoshita M, Yokomise H: **Neural-cadherin expression associated with angiogenesis in non-small-cell lung cancer patients.** *Br J Cancer* 2003, **88:**1727-1733.
27. Jin R, Chow VT, Tan PH, Dheen ST, Duan W, Bay BH: **Metallothionein 2A expression is associated with cell proliferation in breast cancer.** *Carcinogenesis* 2002, **23:**81-86.
28. Watanabe T, Wu TT, Catalano PJ, Ueki T, Satriano R, Haller DG, Benson A. B., 3rd, Hamilton SR: **Molecular predictors of survival after adjuvant chemotherapy for colon cancer.** *N Engl J Med* 2001, **344:**1196-1206.
29. Frodin JE, Fagerberg J, Hjelm Skog AL, Liljefors M, Ragnhammar P, Mellstedt H: **MAb17-1A and cytokines for the treatment of patients with colorectal carcinoma.** *Hybrid Hybridomics* 2002, **21:**99-101.
30. Maeda K, Kawakami K, Ishida Y, Ishiguro K, Omura K, Watanabe G: **Hypermethylation of the CDKN2A gene in colorectal cancer is associated with shorter survival.** *Oncol Rep* 2003, **10:**935-938.
31. Galiegue-Zouitina S, Quief S, Hildebrand MP, Denis C, Detourmignies L, Lai JL, Kerckaert JP: **Nonrandom fusion of L-plastin(LCP1) and LAZ3(BCL6) genes by t(3;13)(q27;q14) chromosome translocation in two cases of B-cell non-Hodgkin lymphoma.** *Genes Chromosomes Cancer* 1999, **26:**97-105.
32. Sheffield MV, Yee H, Dorvault CC, Weilbaecher KN, Eltoum IA, Siegal GP, Fisher DE, Chhieng DC: **Comparison of five antibodies as markers in the diagnosis of melanoma in cytologic preparations.** *Am J Clin Pathol* 2002, **118:**930-936.
33. Degen WG, Weterman MA, van Groningen JJ, Cornelissen IM, Lemmers JP, Agterbos MA, Geurts van Kessel A, Swart GW, Bloemers HP: **Expression of nma, a novel gene, inversely correlates with the metastatic potential of human melanoma cell lines and xenografts.** *Int J Cancer* 1996, **65:**460-465.
34. Maelandsmo GM, Florenes VA, Mellingsaeter T, Hovig E, Kerbel RS, Fodstad O: **Differential expression patterns of S100A2, S100A4 and S100A6 during progression of human malignant melanoma.** *Int J Cancer* 1997, **74:**464-469.
35. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
36. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
37. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
38. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.