Supplemental Preprocessed Data to Culhane et al., (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis

Brief Summary of additional data files 1,2, 4 and 5

Additional data file 1"Ross_5643.zip" and Additional data file 2 "Staunton_7129.zip" are provided in Microsoft Excel format, and are compressed using Winzip.

Additional data file 4 "Ross_5643_KNN.txt" and Additional data file 5 "Staunton_1517_CS.txt" are tab delimited text files. Ross_5643_KNN.txt is worksheet 4 "5643_sorted_arrays" from Ross_5643.zip (see below for description). Staunton_1517_CS.txt is worksheet 3 "Centred data 1517 sorted label" from Staunton_7129.zip (see below for description). Probe ID/clone ID are in the first column and array names in the first row.

Additional Data File 1 Ross_5643.zip:

Ross_5643.zip in an excel file that contains 4 worksheet and a readme.

Worksheet Label	Description
Array_labels	Labels of the cell lines (note these were edited so they would be consistent with those from Staunton et al.,)
Clone ID	Image clones ID's of the spots on the cDNA array
Clone_ID_5643_Gene_Info	Further information about the spots (data retrieved using SOURCE, at Stanford, Unigene August 2003)
5643_sorted_arrays	Gene Expression data of 5643 Spots, where rows >15% missing data are removed, and remaining missing values are imputed. Arrays are sorted so they are in same order as that used from Staunton et al.

Preprocessing of the Ross dataset

The data was provided as log ratio (from http://discover.nci.nih.gov/datasets/Nature2000.jsp). We retrieved the dataset (updated 12/19/01)This downloaded data set contained expression data on 9,703 spots, but there were many missing values.

Data (rows of gene expression values) were not removed if >15% of gene values were missing across the 60 cell lines. This reduced the dataset to 5643 spots.

The remaining missing values were imputed using KNN impute, a K-nearest neighbour method, (using parameters 16 neighbours, Euclidean distance metric).

The columns (arrays) names were edited so the cell line names on this agreed with Staunton et al., The columns were sorted so that arrays were in the same order in both datasets.

Subsequent processing of data for our own analysis

A positive scalar (10) was added to data to make positive, and the data was transposed so that it could be for processed using co-inertia analysis in ADE4.

Additional Data File 2 Staunton_7129.zip:

Staunton_7129.zip is an excel file that contains 7 worksheet, where the first sheet is a readme and the remaining sheets contain:

Worksheet Label	Description
Array_labels (sorting)	Cell line (arrays or column) names from both the Ross et al., and Staunton dataset were compared. The were edited and sorted for conformity
Staunton 7129	The raw average difference values as downloaded annotated with gene information downloaded from SOURCE (UniGene August 2003).
Centred data 1517 sorted label	Data from worksheet "Staunton 7129", columns (cell line order) are sorted so they are consistent with Ross et al., Genes (rows) are filtered to the subset where the max-min average difference value across all 60 cell lines >500. The data are logged (base 2) and median centred
Centred data 2455 sorted label	Data from worksheet "Staunton 7129",, columns (cell line order) are sorted so they are consistent with Ross et al., Genes (rows) are filtered to the subset where the max-min average difference value across all 60 cell lines >200. The data are logged (base 2) and median centred
Centred data 3144 sorted label	Data from worksheet "Staunton 7129", columns (cell line order) are sorted so they are consistent with Ross et al., Genes (rows) are filtered to the subset where the max-min average difference value across all 60 cell lines >100. The data are logged (base 2) and median centred

Preprocessing of the Staunton dataset

The data was provided as average difference (pm-mm) values (from http://www.genome.wi.mit.edu/mpr/NCI60). This downloaded data set contained expression data on 7129 spots. Gene expression values were thesholded at 100, rows which were invariant (max-min value=0) across the 60 cell lines were removed. This reduced the dataset to 4615 values.

As this still included substantial numbers of relatively invariant genes, filters of max-min > 100, 200 and 500 were made. This produced gene sets of 3144, 2455 and 1517 probe sets.

The columns (arrays) names were edited so the cell line names on this agreed with Staunton et al., The columns were sorted so that arrays were in the same order in both datasets.

Subsequent processing of data for our own analysis

The data was median centred, a positive scalar (10) was added to data to make positive, and the data was transposed so that it could be for processed using co-inertia analysis in ADE4.