**Microarray data processing.** The steps required for processing data from each paper into a format allowing comparison between them. The datasets were acquired from various websites (addresses for which can be found in supplementary information S2 (table). Each dataset was processed separately and then combined at the end. The datasets can be divided into 9 groups according to the format of the available data and the processing involved. Data from groups 1 and 2 required little data processing, other than converting to log base 2 for group 2. Data from group 3 were taken from the Stanford microarray database (SMD). These datasets were filtered in SMD to remove probes flagged as missing by the array scanner with subsequent processing steps carried out locally in Excel. Median centering the arrays sets the median log2 expression ratio for each array to 0. This is performed so that the data from each array is scaled equivalently. Median centering genes does the same thing for genes and is performed for studies utilizing a common reference so that expression changes become ratios of the median value for that gene across all samples. Data in groups 4-6 were generated with Affymetrix arrays and were provided as absolute signal intensities. The data processing for these datasets reflects the need to create log2 expression ratios. The data for many of these studies were also filtered to remove genes that were often or always called absent by the Affymetrix software. In all cases, data were processed in a manner that best matches the data processing pipeline originally used by the authors. From those studies where the information was not included, HUGO gene IDs were retrieved using the ID provided, for example GenBank accession number, LocusLink or Affymetrix IDs. Data from multiple probes representing a single gene (single HUGO ID) were averaged to provide one expression value per HUGO ID per experiment. For each dataset, those genes that are also present on the Affymetrix Hu6800 array were isolated. This array was used to provide the final gene set because it has been used extensively by our lab and others to produce expression data – the group 7 studies). When a gene present on the Affymetrix Hu6800 array was not found in another dataset a gap appears in the data (grey in Figs. 1 and supplementary information S6). Data in each dataset were then aligned so that each row represents a single HUGO ID. The genes were then ordered with a 1-dimensional self-organising map and grouped using average-linkage hierarchical clustering.