Figure 1 (a) Hierarchical clustering of the patterns of variation in the expression of 6849 cDNA clones in 27 cell lines. The data are shown in a table format, in which rows represent individual genes and columns represent individual cell lines. The color in each cell reflects the expression level of the corresponding gene in the corresponding cell line, relative to its mean expression level across the entire set of cell lines. The scale (lower right corner) extends from fluorescence ratios of 0.25 to 4 relative to the mean level for all samples. Grey indicates missing or excluded data. (b) to (f) Features of the variation in the gene expression patterns that can be related to specific physiological or histological features of the cell lines. (b) epithelial cell cluster; (c) B lymphocytes cluster; (d) T lymphocytes cluster; (e) endothelial cell cluster; (f) fibroblast cell cluster. Due to limited space, only a few selected gene names are shown. See Supplementary Information for full data.

Materials and Methods: A total of 27 cell lines were used in this study. These included: twelve gastric cancer cell lines (AGS, KATO-3, SNU1, SNU5, SNU16, RF1, RF48, NCI-N87, NUGC3, MKN45, BGC823, PCAM82), two T cell lines (Jurkat, Molt4), two B cell lines (LAM, HFI1), 1 APL-like cell line (NB4+RA), one monocyte-like cell line (U937+PMA), one pancreas cancer cell line (Bxpc3), four colon cancer cell lines (colo205, HCT116, SW620, HCT15), one breast cancer cell line (MCF7), one primary fibroblast and two primary endothelial cell lines. The detailed information for these cell lines is available through the web supplement. All cell lines are cultured to 80% confluence, harvested and frozen in −80°C until ready to be isolated. mRNA was extracted directly from the frozen cell pellet using FastTrack (Invitrogen) mRNA isolation kit. For the microarray production, 43,000 cDNA clones, representing about 36,000 unique genes, were mechanically printed onto treated glass microscope slides, as previously described (http://cmgm.stanford.edu/pbrown/array.html) (Perou et al., 2000). For RNA labeling, a common reference, which consisted of mixture of eleven cell lines
was used (Perou et al., 2000). The hybridization procedures were performed as previously described (Alizadeh et al., 2000). A detailed protocol is available at: [http://cmgm.stanford.edu/pbrown/protocols/5_hyb_human.html](http://cmgm.stanford.edu/pbrown/protocols/5_hyb_human.html). Primary data collection and analysis were carried out using GenePix Pro 3.0 (Axon Instruments). Areas of the array with obvious blemishes were manually flagged and excluded from subsequent analysis. The raw data were deposited into Stanford Microarray Database (Sherlock et al., 2001) at: [http://genome-www4.stanford.edu/MicroArray/SMD/index.html](http://genome-www4.stanford.edu/MicroArray/SMD/index.html). For the generation of the cluster, all non-flagged array elements for which the fluorescent intensity in either channel was greater than 2.5 times the local background were considered well measured. Genes for which fewer than 60% of measurements across all the samples in this study met this standard were excluded from further analysis. We selected for further analysis genes whose expression level differed by at least four fold, in at least one sample, from their mean expression level across all samples. We applied a hierarchical clustering algorithm both to the genes and arrays using the Pearson correlation coefficient as the measure of similarity, and average linkage clustering, as described (Eisen et al., 1998). The results were visualized and analyzed with TreeView (M. Eisen; [http://rana.lbl.gov](http://rana.lbl.gov)).