

## S7 | Box 2. Large-scale comparison of microarray data.

Our analysis is a relatively simple attempt to collate and compare expression data from multiple sources. Other examples of large-scale comparisons of microarray data can be found in references 1 and 2. We discuss below some of the strengths and limitations of the method we used for our large-scale comparison.

### Strengths

1. It enables the comparison of more samples than would be feasible in a single experimental study.
2. It helps define both common and specific expression changes.
3. It reveals the extent to which different studies of the same system agree.
4. We can be more confident of expression changes observed in many different studies of the same process.
5. Genes that cluster based on a shared expression pattern over many hundreds of samples are likely to be co-regulated.

In our analysis, we compare data from diverse studies, from those that follow expression changes in a single cell type over time to those that compare multiple samples of the same tumour type. This variance between datasets is not as problematic as first appears. Regardless of how the expression ratios are generated, cluster analysis will group together genes that share the same expression pattern. For each dataset, genes that are expressed more strongly in pathogen-exposed cells compared with control cells have positive expression ratios and vice versa. The subsequent visualization of the gene expression patterns (as a red/green display) allows the clusters to be interpreted. For example, it is clear that the common host response cluster (Fig. 1 and **supplementary information S6**) contains the majority of genes commonly upregulated in cells in response to pathogens.

### Limitations

Biases can be introduced into the clustering of expression data from a number of different sources:

#### 1. Unequal numbers of samples.

By default, the clustering algorithm applies equal weight to each sample and so will be biased towards groups of samples that are over-represented in the dataset. For example, our analysis is biased towards macrophages and away from neurons.

#### 2. Unequal numbers of genes.

A number of datasets do not contain data for certain genes, either because the data were not made available, removed by filtering or not represented on the array. Samples containing a lot of missing data have less influence on the structure of the resulting clusters. The advent of MIAME (Minimum information about a microarray experiment) guidelines and the use of greater capacity arrays will help minimize this issue.

#### 3. Variations in the magnitude of expression changes.

Due to differences in array technology and data processing, identical changes in transcript abundance may be reported differently between studies. The magnitude of any expression change is also affected by the experimental design, for example the multiplicity of infection used. This has two effects; firstly it biases the clustering towards samples containing genes with large expression ratios and, secondly, it means that differences in the magnitude of expression changes between studies may not be significant. For example, we cannot say that ISGs are induced less strongly by HCV in liver in vivo than they are by influenza in DCs in vitro. That said however, differences in gene expression can still be identified if other patterns are shared. For example, we can see that PBMCs induce an additional set of genes compared to macrophages (blue box, Figure 1) because other clusters (red box, Figure 1) show equal upregulation in these two cell types. A requirement for investigators to analyse a number of standardised RNA samples in parallel with their experimental samples would help normalise inter-study differences in expression measurements.

#### References:

1. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA* **101**, 9309–9314 (2004).
2. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nature Genet.* **36**, 1090–1098 (2004).