Allen et al., http://www.jcb.org/cgi/content/full/jcb.200604072/DC1

Supplemental materials and methods

The following is a step-by-step methodology for the microarray statistical ranking analysis. (1) Replace the missing values in each dataset with the median of the present values for each microarray and then rank each microarray. (2) Convert ranks to a value between 0 and 1 and calculate f = (rank − 0.5)/(total number of genes).(3) Use the inverse normal cumulative distribution function and transform into N (0, 1) Z scores for each microarray for each gene for upper and lower. (4) For each microarray, calculate Z lower − Z upper. Call this D. (5) Get the mean of D for all microarrays across each gene. Call this MD. (6) Get standard deviation of D for all microarrays across each gene and divide by square root of number of microarray to get standard deviation of the mean of MD. Call this SMD. (7) Divide MD by SMD to get a final score, which we will call FZ. (8) Use the t CDF with $n - 1$ degrees of freedom for the FZ value and get a two-tailed p-value. (9) Because we are interested in the tail, if $P > 0.5$, then the p-value = 1 − p-value, and if $P \leq 0.5$, then it remains unchanged. At this stage, we have a p-value for each gene for the overall difference between lowers and uppers. (10) Assume the significance cutoff level = 0.001. (11) If the p-value is less than the cutoff level and $FZ > 0$, then class = "lower." If the p-value is less than the cutoff level and $FZ < 0$ then class = "upper." (12) As a result, we have now two gene lists: lower and upper (see Table S1).

Note that having a class equal to "lower" does not imply that the gene is highly expressed for the lowers. It implies that relative to the uppers it had a significantly higher overall expression, which could mean that either the gene is more expressed in the lowers than expected or it is more repressed in the uppers than expected. Nevertheless, in this experiment and similar experiments involving uppers and lowers, if the gene is labeled as lower, it is very likely to be highly expressed.

Also, note that the choice of the level of significance for cutoff purposes is a matter of how many false positives we are willing to accept. For example, let's assume our dataset uses a total of 6,000 genes and that by using a cutoff level of 0.01, we get a total of 200 genes as lowers. Then, of the 6,000 genes, we have an expectation equal to 6,000 × 0.01 = 60 genes that could have been selected just by chance and, thus, we can say that within our 200 selected genes, we expect to have, on average, a total of 60 genes that do not truly belong, leaving us only with 140 genes actually related to the lowers. The difficulty here is that by only using the cutoff described above, we are unable to identify which are the correct 140 genes.