Generation of transcriptomic consensomes

Transcriptomic ‘datasets’ are collections of data from different sources (i.e. different GEO datasets). Experiments, or contrasts in statistical terminology, are pairs of conditions (i.e. control and treatment) within data sets, each of which has multiple observations, that are used to generate nominal p-values and fold changes in expression for each gene (target) represented in the pair. These are all pre-computed and stored in the Signaling Pathways Project database. Experiments are the unit of analysis, of which a single dataset can have one or more. For a consensome (e.g. ERs-Hs-MG-TS) of interest, we retrieve all the ‘experiments’ that mapped to that combination of parameters.

Computation of Target-Specific, Experiment-specific Nominal P-values and Fold Changes

Although RNA-Seq datasets are growing in number, expression arrays remain in use and the vast majority of expression profiling datasets archived in Gene Expression Omnibus are on array platforms. We first therefore set out to develop an algorithm that would establish consensus across array datasets. Although much less than 1% of genes in any particular array experiment are represented by more than 1 probeset, a few genes had 2-5 probesets and a very few had as many as 15 or 20. In such cases, we combined probeset-specific fold changes and probeset-specific p-values to generate gene level fold-changes and p-values. Briefly, we used the fold changes to convert the individual probeset-specific two-tailed nominal p-values into z-scores that capture the direction of the change:

$$Z_p = qnorm\left(\frac{\left|1 + sign\left(log2\left(F_p\right)\right) - P_p\right|}{2}\right)$$

where Z_p is the directional probeset-specific z-score, P_p is the two-tailed probeset-specific p-value, F_p is the probeset-specific fold change, qnorm() is the standard normal inverse CDF, and sign(x) is 1 when x is >= 0 and -1 when x<0. Thus, when F_p is >= 1, this yields Z_p=qnorm(1-P_p/2) (range is [0,∞)), and when F_p <1, this yields the lower tail, Z_p=qnorm(P_p/2) (range is (-∞,0]).

A summary gene-specific p-value was calculated as 2 times the upper tail of the standard normal cumulative distribution function assessed at the absolute value of the average of the probeset-specific Z’s:

$$Z = \frac{\sum_p Z_p}{n}$$

$$P = 2 * \left(1 - pnorm(|Z|)\right)$$

where Z is the average of the probe-specific z-scores, P is the gene specific two-tailed p-value, n is the number of probesets for a gene, and pnorm is the standard normal cumulative distribution function.

The summary gene-specific experiment-specific fold change is calculated by exponentiating the predicted value of log2 fold change from a linear regression of probeset log2 fold changes regressed on probeset z’s, evaluated at the average of the probeset z’s:

$$\widehat{\mathrm{F}} = 2^{(\hat{a}+\hat{b}*Z)}$$

where Fhat is the predicted fold change at Z, ahat is the intercept and bhat is the slope of the linear model of log2(F_p) modeled as a function of Z_p.

Combination of Gene-specific P-values and Fold changes Across Experiments

For each gene, g, in the consensome, we counted the number of experiments, E_g, where the gene has a nominal p-value of 0.05 or less out of Ng experiments where gene-specifc data are not missing. A consensus p-value, P_g, was calculated as the binomial probability of observing E_g or more successes out of N_g trials, when the true probability of success is 0.05. This provides an estimates of the degree to which the fraction of experiments with alterations exceed what might be expected by chance.

The gene-specific consensus fold change, F_g, is geometric mean of the experiment-specific fold changes, expressed as max(F_{ge}, 1/F_{ge}), for the gene of interest. All fold changes are converted to max(F_{ge}, 1/F_{ge}) because some experimental manipulations repress and others enhance and rather than canceling each other out, both should be counted as ‘altered’ so that we can generate a summary measure of magnitude of perturbation.

Supplementary Table 2. Calculation for a hypothetical target. Genes in the consensome analysis are ranked in ascending order by CPValue, with average rank reported in the case of tied CPValues.

Gene	Experiment	Probeset	Fp	Pp	Zp	Fge	Pge	Eg	Ng	CPvalue	GMFC
X	E1	PX1.1	4	0.001	3.29	2.0	0.08	2	4	0.014	2.83
		PX1.2	2	0.05	1.96						
		PX1.3	1	1.0	0						
	E2	PX2	2.5	0.1	1.645	2.5	0.1				
	E3	PX3	0.333	0.02	-2.33	0.333	0.02				
	E4	PX4	0.333	0.02	-2.33	0.333	0.02				