Confirmation of links within data sets

Results: Many genes are represented more than once per data set (though often by different sequences), giving such genes 'multiple chances' to generate links per data set. For example, there were 755 genes that were represented twice or more on average per data set and that were represented in at least 3 data sets. Because of this potential source of bias, we used a statistical multiple test correction to penalize links that were tested multiple times in a given data set, resulting in the rejection of many links during our initial stage of analysis (see Methods). Despite this correction, we observed a low but significant correlation between the number of probe representations per data set and the number of confirmed links for each gene (Spearman's rho = 0.37 for 3+ confirmed links; $p<0.001$; $R^2 = 0.03$; Supplementary Figure A).

We examined the issue of *intra*-dataset reproducibility of links in more detail. Of our original pool of 11 million links, there were 6 million that had been tested more than once in at least one data set. Of these, on average 7.4% per data set showed perfect agreement among tests, meaning that all of 2 or more tests of the same link in the same data set met our link selection criteria (about 373,000 links in total; 561,000 were reproduced in >0.5 of the tests; Supplementary Figure B). The extreme cases are illustrated by a small number of links that were perfectly reproducible (11,466 links) or highly reproducible (at least 75% of tests; 41,146 links) in 4 or more tests in the same data set, and a larger number of links (505,417) that are not reproduced in 10 or more tests in the same data set (observed just once in a data set).

Interestingly, intra-dataset reproducibility of links was correlated with inter-dataset confirmation (Supplementary Figure C). The links that had incidences of high non-reproducibility within a single data set as defined above were 3+ confirmed at a much lower rate than the average (0.6% compared to 2.2%, $p < 10^{-15}$, Wilcoxon signed-rank test), while links that were highly reproducible within data sets had a very high 3+ confirmation rate (9.6%; $p< 10^{-15}$). Exclusion of

occurrences of poor intra-dataset confirmation as defined above results in the loss of 20,096 3+ links, but does not reduce the correlation of link count to probe representation. This suggests that the increased number of links detected for over-represented genes is not simply due the inclusion of unreliable links.

We finally note that the “intra-dataset irreproducible” links we identified in the previous paragraph do not have worse semantic similarity measures than links that had the highest intra-dataset reproducibility (Supplementary Figure D).

Discussion: We found that genes that have multiple representations within microarray designs have a tendency to have more confirmed links. This effect, though statistically significant, accounts for a small fraction of the variance in link confirmation level. This effect is not accounted for by statistical multiple testing (which we corrected for), nor could it be corrected for by increasing the stringency of acceptance for links that are tested multiple times on a microarray. Intra-data set reproducibility was also not a good predictor of semantic similarity of annotations for pairs of genes, in contrast to our findings for inter-data set confirmation. We hypothesize that the overrepresentation effect is in part due to varying specificities for different probes for the same gene on the same array (e.g., targeting different splice variants). Thus a single positive finding, if confirmed across microarray studies, may be “real” despite many apparent failures to confirm the finding in the same data set. An additional possible explanation is that genes that are highly represented on microarrays tend to be “important” in the view of the microarray designer – a type of selection bias. Such genes might be expected to have more coexpression “interactions” with other genes. We leave further examination of this issue as a topic for future work.