# Supplementary Material accompanying:

"Extraction of Transcription Regulatory Signals from
Genome-wide DNA-protein Interaction Data"
by Garten, Kaplan, and Pilpel

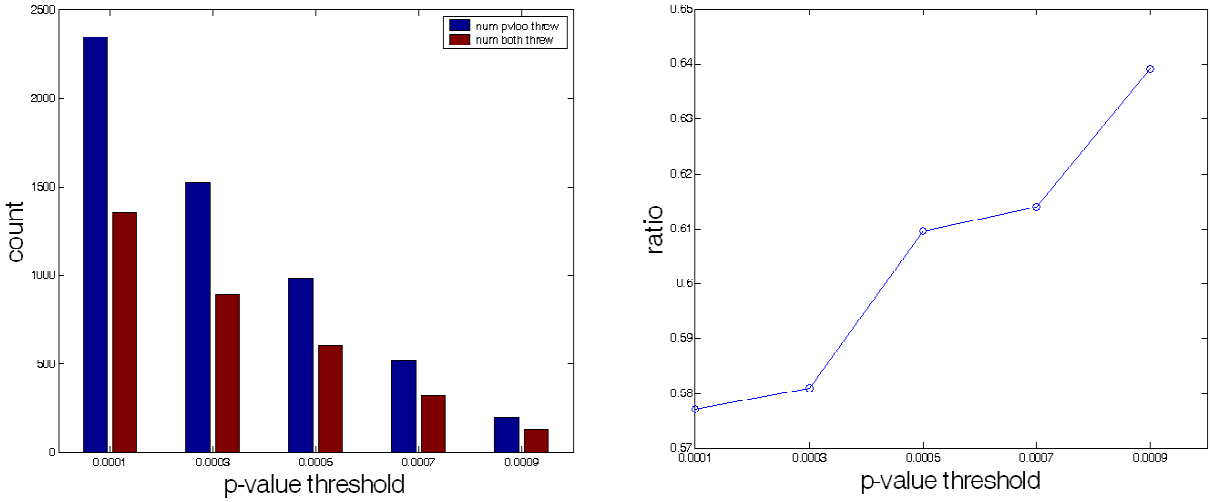
## Supplementary Figures and Legends

Figure S1

We examined the relationship between two conceivable methods of filtering the data:

1. Using a p-value stricter than 0.001 on the binding predictions of the location data
2. Using TF-gene assignments which have the support of at least three filtration methods described in the paper

We show here that our methods save many assignments which have supporting evidence, which would be discarded by using a stricter p-value.
The figure shows the ratio between the number of TF-gene assignments filtered out by both methods and those filtered out of the data by using a stricter p-value alone.




The histogram shows the following, as a function of changing the p-value threshold from 0.001 to the value shown on the x-axis:

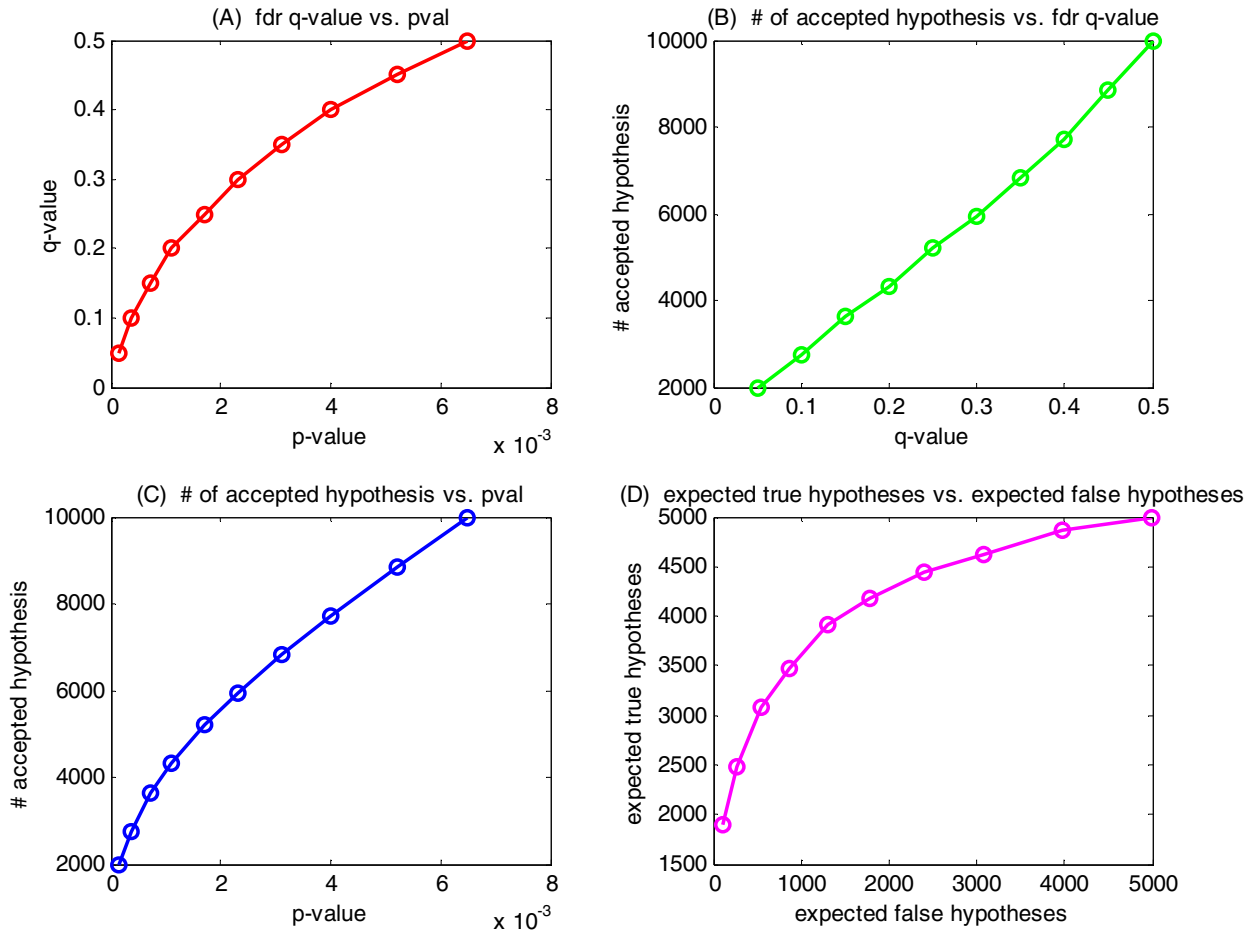
- In red, the number of specific assignments which are filtered out by both methods; using a p-value stricter than 0.001 on the location data predictions, and also by the requirement of support of at least three of the four filtration methods described.
- In blue, the number of assignments lost when filtering only according to p-value.

The plot shows the ratio between these two (red/blue).

It is clear that at all of the p-values thresholds, there is a large number of hypotheses that we rediscover, that filtration using the p-values assigned in the location data would discard. The fraction of such hypotheses grows larger as the p-value selected is more stringent.

Figure S2:

The figure describes an FDR analysis of the original location data.




(A) shows the relationship between the expected false discovery rate (q-value) and the p-value. It can be seen that at a p-value of 0.001, the q-value (FDR) is ~18%.

(B & C) show how the relationship between the number of accepted hypotheses as a function of the q-value and p-value respectively. It can be seen that in order to achieve a false discovery rate of 10%, the number of hypotheses drops significantly to ~3400, significantly lower than the ~4200 hypotheses that were accepted at a p-value threshold of 0.001.

(D) shows the relationship between the number of expected true hypotheses and the number of expected false hypotheses at various p-value thresholds. This graph displays the tradeoff which occurs at the different p-value thresholds. As the p-value threshold is raised (moving from left to right on the x-axis, and from lower to higher values on the y-axis), in the region where the slope of the graph is less than 1, more false hypotheses are added than true hypotheses. On the contrary, when the p-value is strict, in the region where the slope is greater than 1, mostly true hypotheses are gained by relaxing the p-value.

Figure S3:

The figure shows the relationship between the clustering of multiple conditions that correspond to the same TF, compared to that expected from random data.
For each TF, we calculated the number of common genes in the largest cluster, in all pairs of conditions in which the EC score of that TF was significant.
The plot shows the distribution of these pairwise overlaps (blue), compared to that expected by random (red).

The random data was formed as follows: for each pair of conditions, the percent of expected common genes was obtained by randomly choosing two sets of genes (the sizes of the major cluster of each of the two conditions), out of the genes assigned to the TF, and calculating the percent of their overlap.

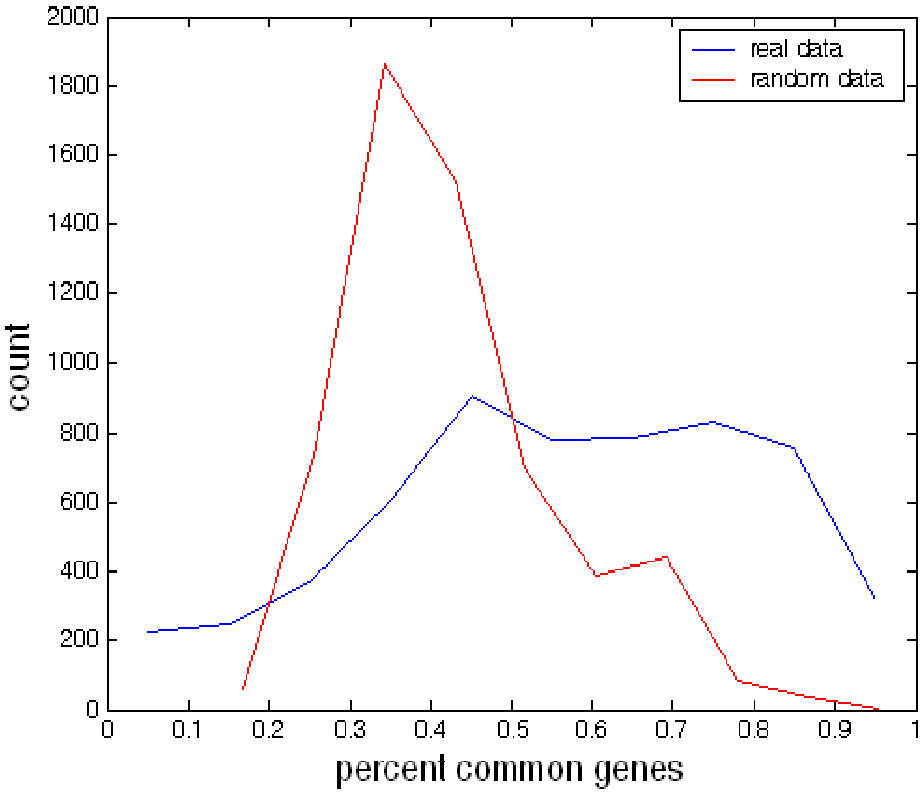


Figure S4:
The fraction (A) and absolute numbers (B) of genes discarded by the various filtration methods, per TF.

S4A:
The image shows in each of the 4 filtration methods, and in their union and intersection, per TF, the fraction of genes discarded by the filtration, relative to the total number of genes assigned to the TF in the location data.

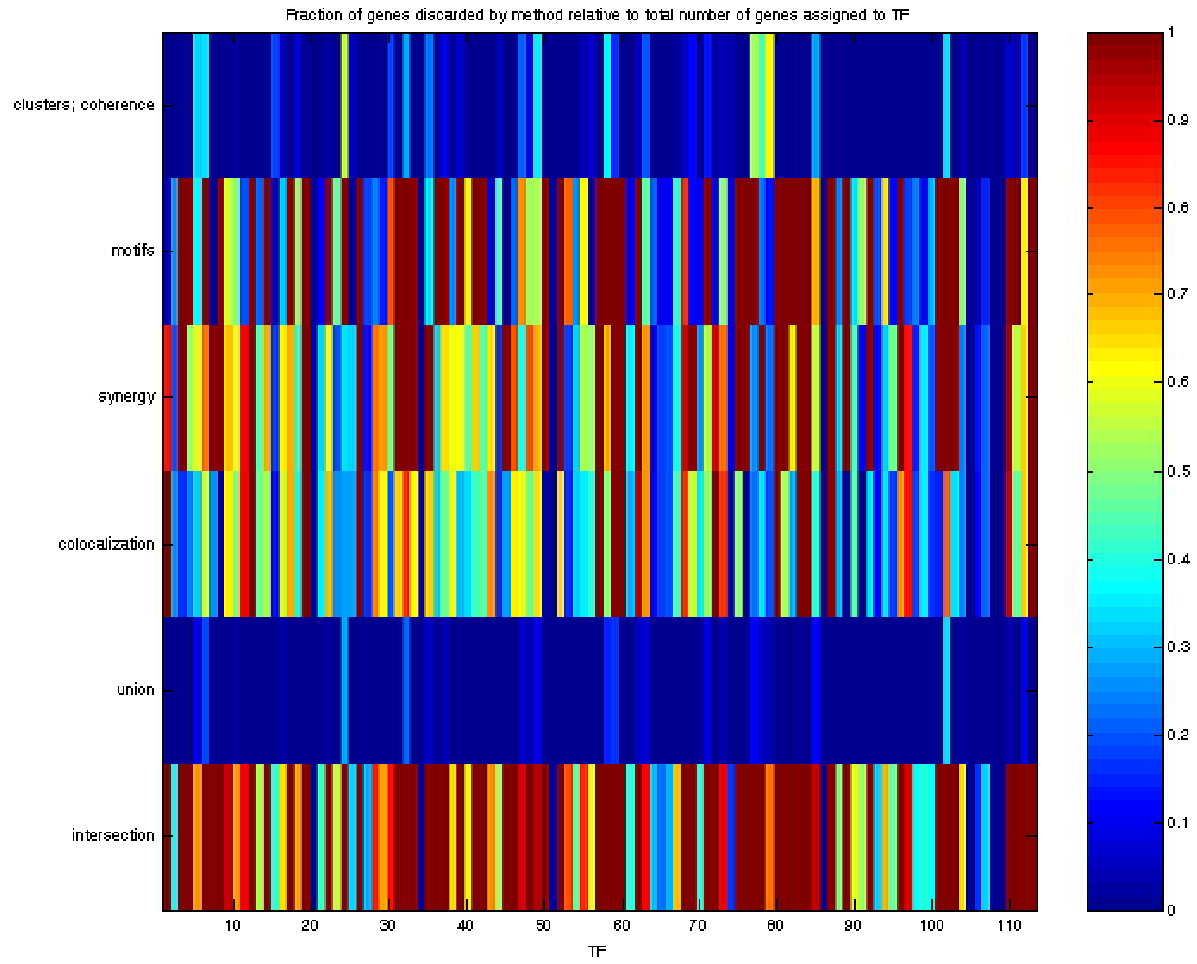


Figure S4B

The bar diagrams below show in each of the 4 filtration methods, and in their union and intersection, per TF, the number of genes discarded by the filtration, and number of genes saved; for each TF the sum of these two numbers is the total number of genes assigned to the TF in the location data.

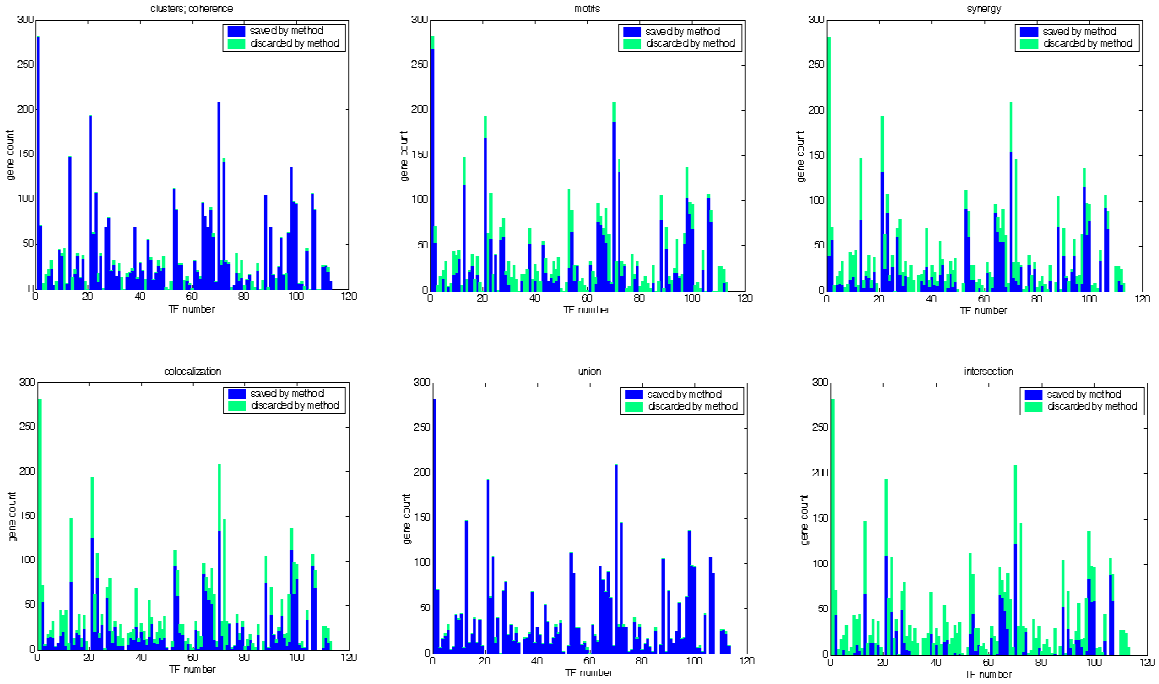


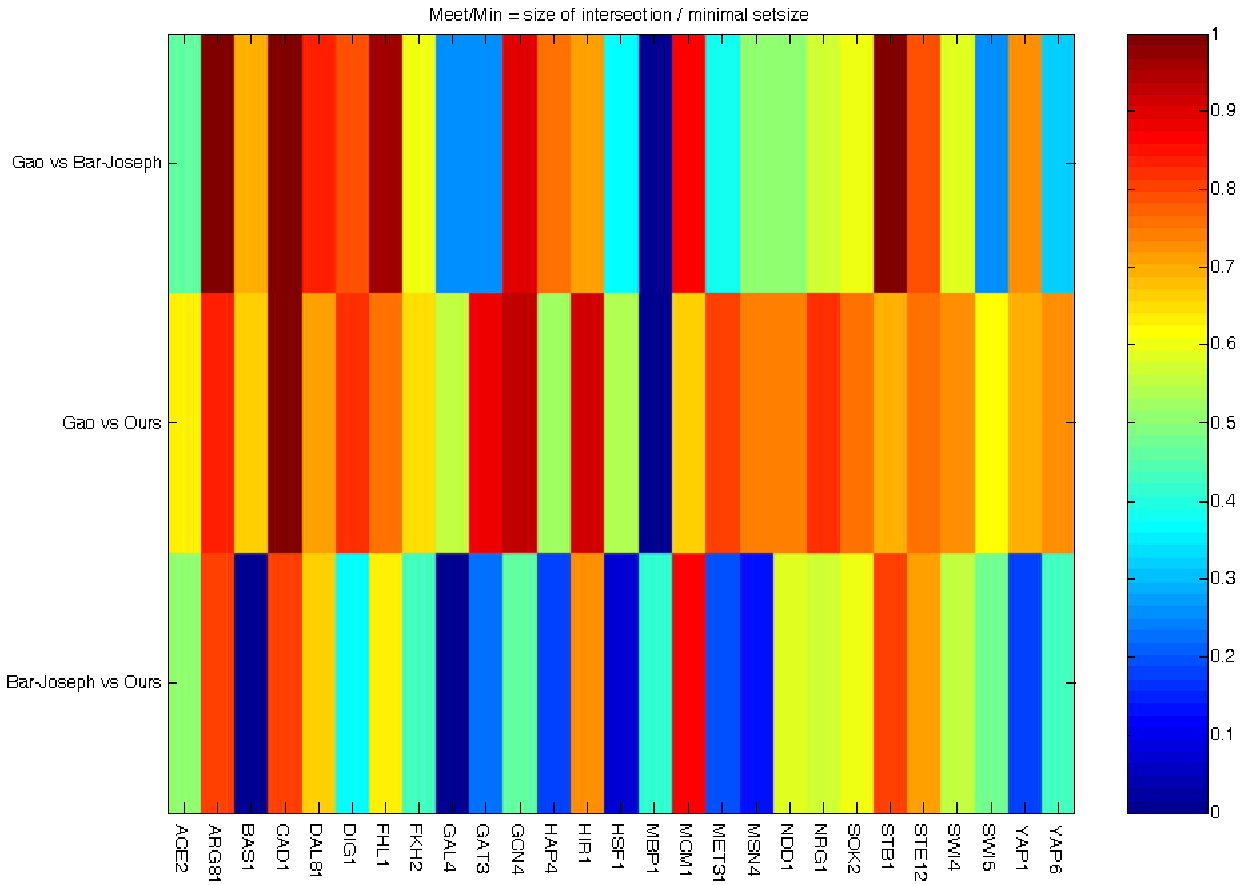
Figure S5

Comparison of our work with that of Gao et al. and Bar-Joseph et al.:
We calculated the Meet/Min and Jaccard coefficients (1), two measures of overlap between sets, between the lists of genes assigned to each TF by Gao et al. (2), Bar-Joseph et al. (3), and by the intersection of our four methods (intersection matrix). These coefficients are respectively defined as the size of the intersection of two sets divided by the size of the smaller of the two sets, and the size of the intersection of the two sets divided by the size of their union.

We have only analyzed TFs for which there exists at least one gene assignment by all three works (Gao, Bar-Joseph, and our own). The figures color-code the Meet/Min and Jaccard coefficients between each pair of studies.

| | Average Meet/Min | Average Jaccard |
|---|---|---|
| Gao vs Bar-Joseph | 0.6181 | 0. 2795 |
| Gao vs Ours | 0.7082 | 0. 3848 |
| Bar-Joseph vs Ours | 0.4575 | 0. 2106 |

Meet/Min matrix:
It is clear from the figure that the work of Gao and ourselves are highly congruent and the average Meet/Min coefficient across 28 TFs is 0.7082. On the other hand, the Bar-Joseph assignments show somewhat lower congruence with Gao's study and an even lower similarity with the present work.
Note that the Meet/Min coefficient minimizes the differences between the sets of genes assigned to a TF by each pair of studies, which stem directly from the addition by Bar-Joseph et al. of gene targets assigned p-values greater than 0.001 in the location data.


MeetMin = size of intersection / minimal setsize
Gao vs Bar-Joseph
Gao vs Ours
Bar-Joseph vs Ours
ACE2
ARG81
BAS1
CAD1
DAL81
DIG1
FHL1
FKH2
GAL4
GAT3
GCN4
HAP4
HIR1
HSF1
MBP1
MCM1
MET31
MSN4
NDD1
NRG1
SOK2
STB1
STE12
SWI4
SWI5
YAP1
YAP6
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

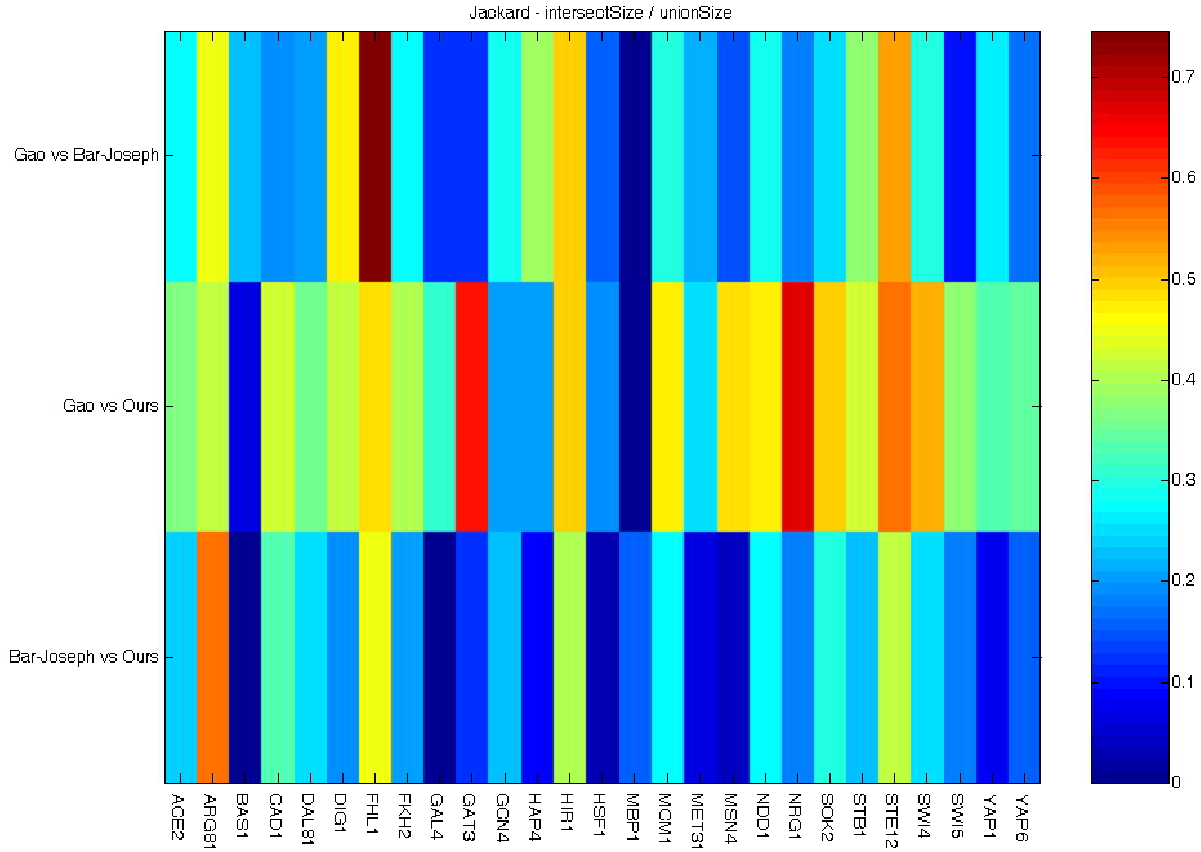

Jackard - intersectSize / unionSize
Gao vs Bar-Joseph
Gao vs Ours
Bar-Joseph vs Ours
ACE2
ARG81
BAS1
CAD1
DAL81
DIG1
FHL1
FKH2
GAL4
GAT3
GCN4
HAP4
HIR1
HSF1
MBP1
MCM1
MET31
MSN4
NDD1
NRG1
SOK2
STB1
STE12
SWI4
SWI5
YAP1
YAP6
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure S6

The figure shows the most populated cluster obtained by clustering the genes assigned to the TF Bas1, when using the QT_clust algorithm and the adaptive clustering algorithm by De Smet et al. The size of the most populated cluster obtained with the adaptive algorithm is constant across the entire range [0-1] of the significance parameter S. In contrast using QT_clust with a range of input diameter values yields a wider dynamic range of cluster sizes, corresponding to significantly coherent gene sets, as seen in the figure.

The figure shows the largest cluster obtained with various values of input parameters as input to both algorithms. The first three plots, A-C, are the result of the QT_clust algorithm, with diameters calculated directly from the expression data (obtained with 5th, 20th and 50th distance percentiles for A-C respectively, see Material and Methods). The fourth plot, D, shows results obtained with the adaptive algorithm that were obtained with any of 20 significance values that evenly span the range [0-1]. We report here that regardless of the significance level the adaptive algorithm generates a largest cluster of constant size of 16 genes. On the other hand, QT_clust obtains this size (with same set of genes), yet in addition it also obtains other sizes that correspond to alternative thresholds. Thus, use of the QT_clust algorithm allows additional results to be obtained which cannot be obtained by the adaptive algorithm, namely a larger major cluster of coherent genes.

Qualitatively similar results were obtained with other transcription factors (not shown).

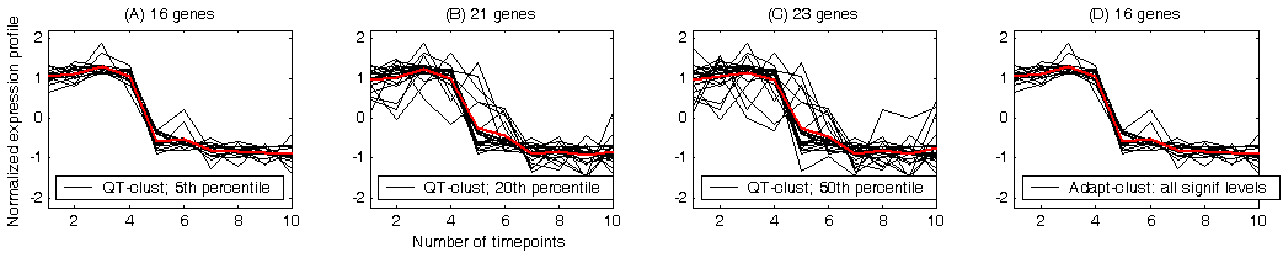
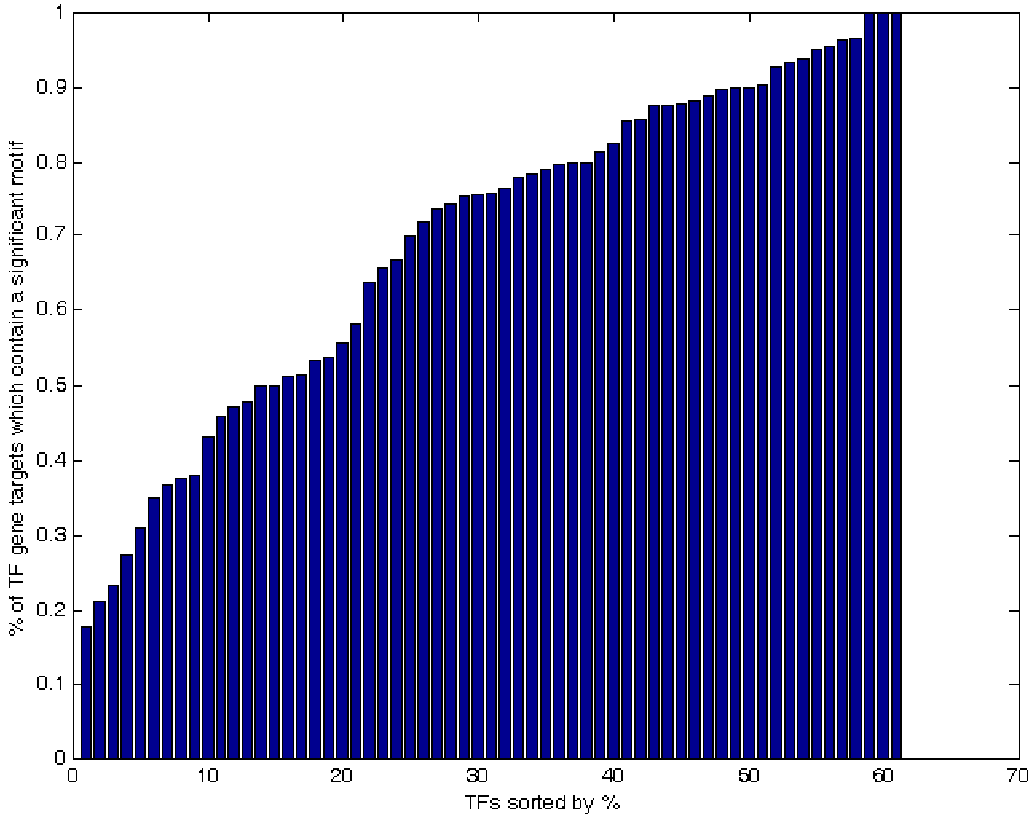


Figure S7

For the 61 TFs for which significant motifs were found, the figure shows the percent of genes assigned to a TF which contain at least one significant motif in their promoter.




References:

1. Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci U S A, 100, 4372-4376.
2. Gao, F., Foat, B.C. and Bussemaker, H.J. (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics, 5, 31.
3. Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. et al. (2003) Computational discovery of gene modules and regulatory networks. Nat Biotechnol, 21, 1337-1342.