Supplementary methods

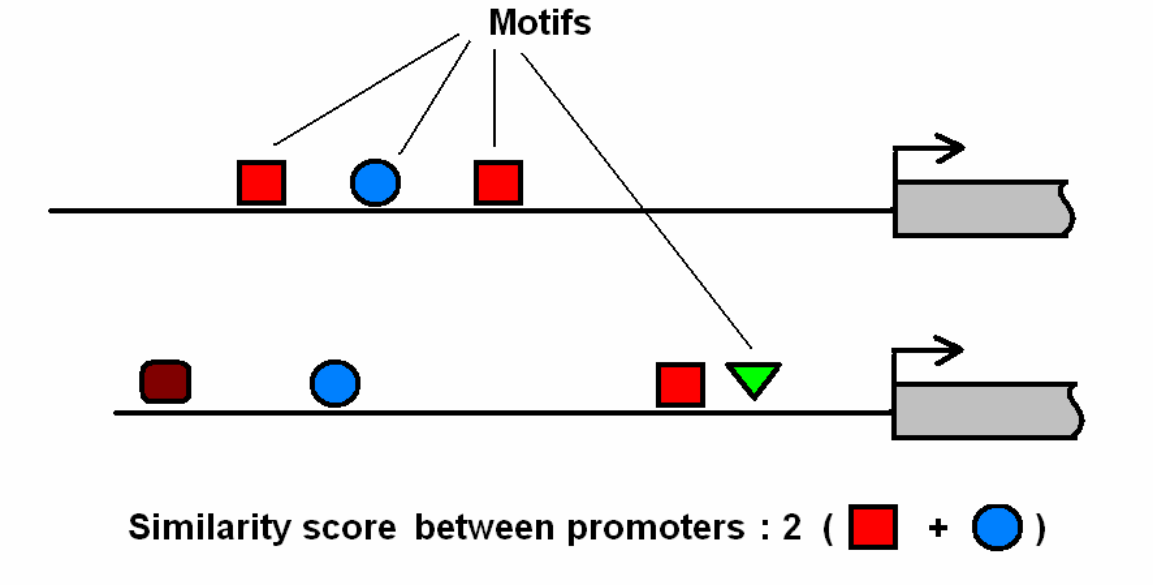
Gene Ontology

The first validation method uses the Gene Ontology (GO), a set of structured, controlled vocabularies to identify functional associations between gene products (Ashburner, et al., 2000). Current GO annotations and external references file were downloaded from the Gene Ontology Annotation resource at EBI (http://www.ebi.ac.uk/GOA/). Each cluster of protein ids was submitted to the High-Throughput GoMiner command-line interface (Zeeberg, et al., 2005). Statistically over-represented GO terms were defined using a Fisher's exact test and corrected for multiple testing by false discovery rate detection (100 permutations). All computation was done on a 400+ core (CPUs) OSCAR compute cluster running Red Hat Enterprise Linux 4.

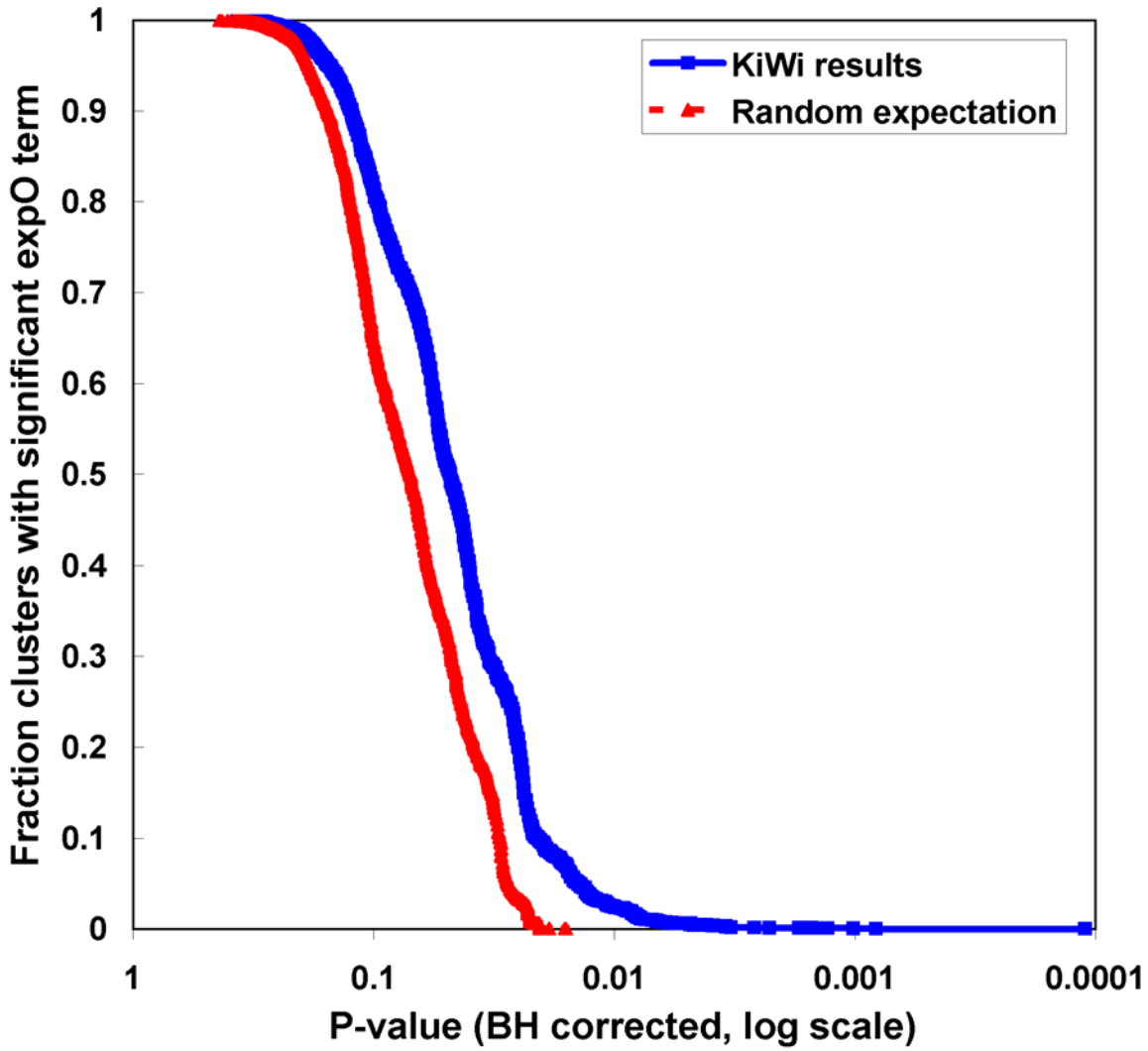
oPossum

The second method uses the oPOSSUM tool to identify statistically over-represented transcription factor binding sites (TFBS) (Ho Sui, et al., 2005). The oPOSSUM API and MySQL database were downloaded and installed locally (http://www.cisreg.ca/cgi-bin/oPOSSUM/opossum). Each cluster of genes was submitted to the software and statistically over-represented TFBSs were defined using the Z-score option.

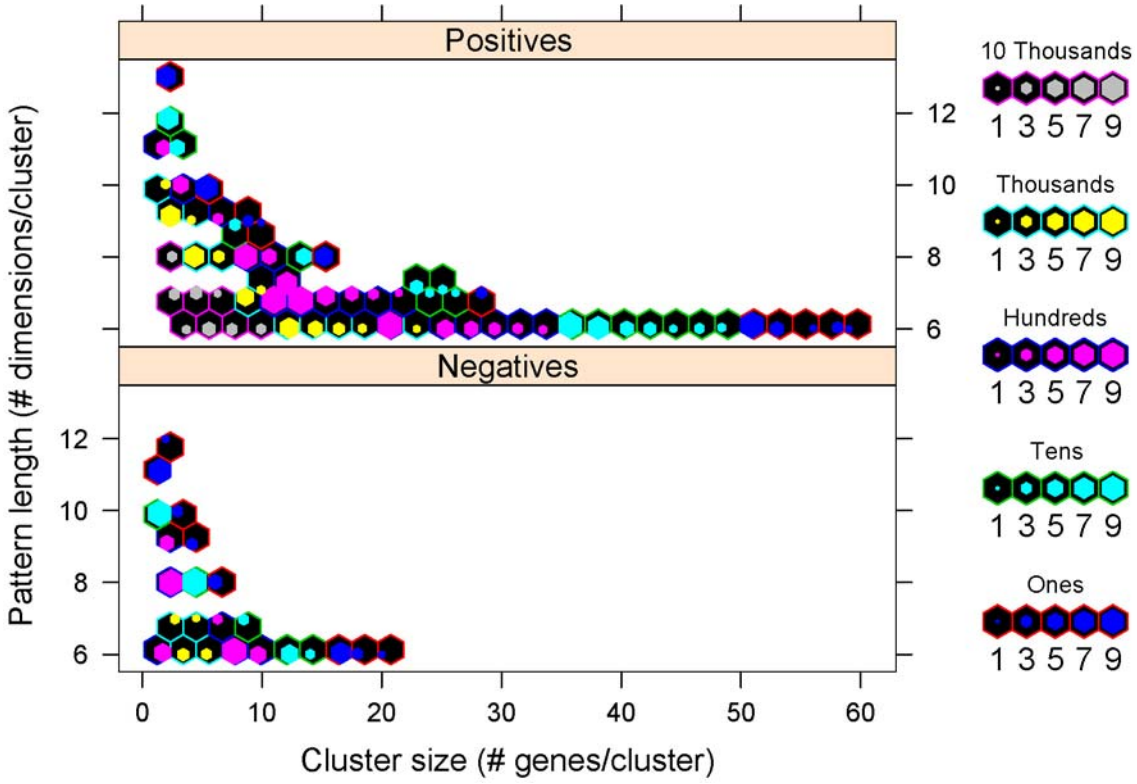
Supplementary figures



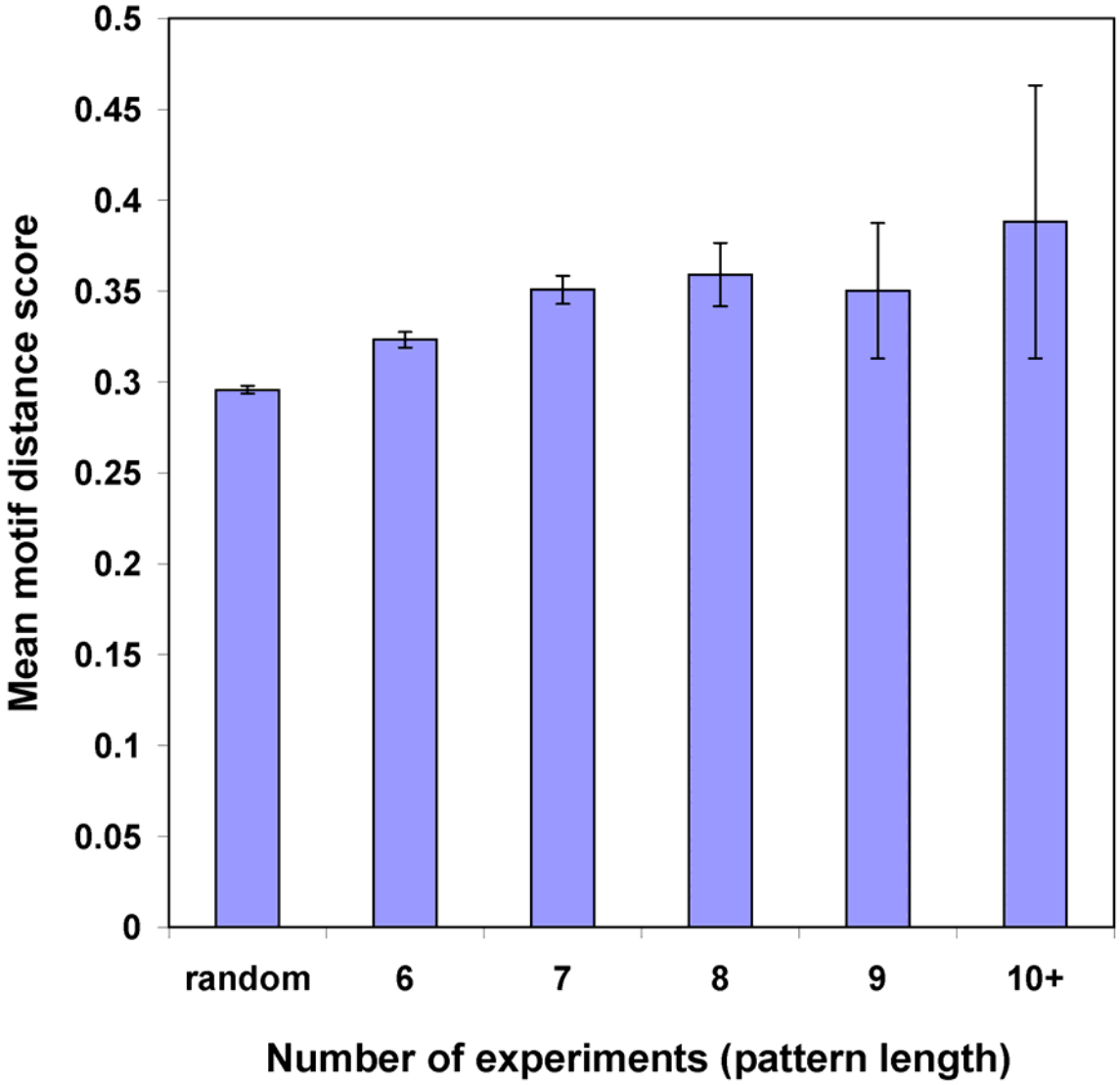
Suppl. Fig. 1. Diagrammatic explanation of "promoter similarity score". For any pair of genes, the upstream region is compared for overlap of annotated TFBS motifs (Sp1, AP-2, etc). These are represented as colored shapes in the diagram above. Each common motif is counted once. The promoters above share two motifs (red square and blue circle). Therefore, the score for this gene pair is 2. Then, the overall promoter similarity score (S) for a cluster of genes is calculated as the sum of the pairwise scores divided by the number of pairs.



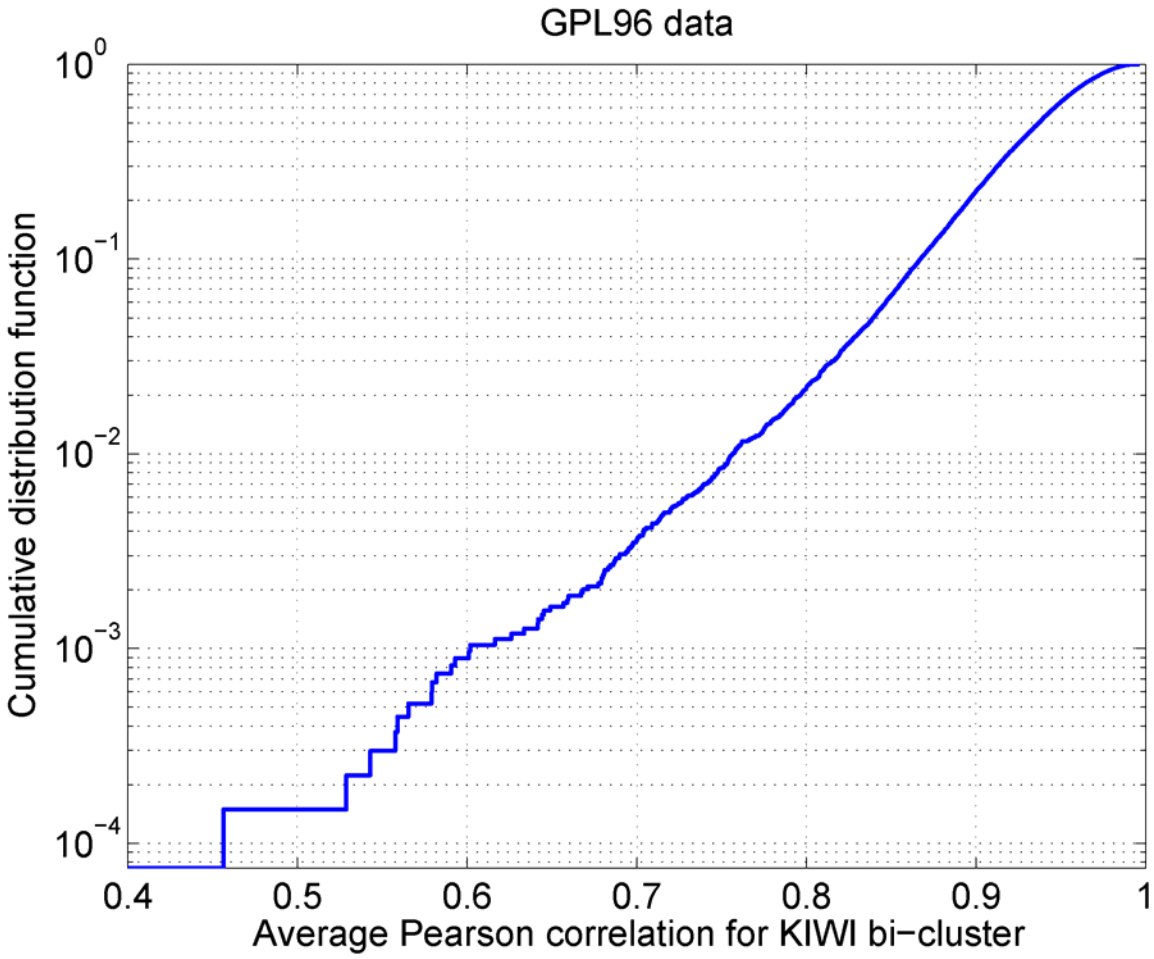
Suppl. Fig. 2. Experimental annotation analysis excluding all annotation terms except tissue source. The fraction of clusters with at least one significantly over-represented tissue source term at each level of significance is shown. Significance was determined by Fisher Exact test. P-values were corrected by the Benjamini and Hochberg method. Kolmogorov-Smirnov test showed a significant difference from random (p=0.005).



Suppl. Fig. 3. KiWi results for Cooper promoter dataset with negative control sequences (Negatives) and real promoter sequences (Positives) clustered separately. The two datasets were submitted to KiWi with identical parameters (k=100000; w=16; min_row=2, min_col=6). The density plot shows that the positive data inherently produces more clusters with longer patterns (more experiments) and greater cluster size (more genes). The density plot was produced using the Bioconductor 'hexbin' library (version 2.3.0).



Suppl. Fig. 4. cisRED analysis comparing different pattern lengths. The mean promoter similarity score for each pattern length is shown. As with cluster size (Fig. 4) the promoter similarity score increases with greater pattern length (number of experiments). Error bars indicate 95% confidence limits.



Suppl. Fig. 5. Distribution of mean Pearson correlations for all KiWi clusters for GPL96 data. This figure shows that the vast majority of subspace clusters (bi-clusters) have very high average pairwise correlations with 90% of clusters having an average r > 0.95.