Title: Gene clustering coordinates transcriptional output of disparate biological processes in eukaryotes

Authors: Richard I. Joh[1,2*], Michael S. Lawrence[1], Martin J. Aryee[1,3] and Mo Motamedi[1*]

Affiliations:
[1]Center for Cancer Research, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA 02129, USA
[2]Current address: Department of Physics and Massey Cancer Center, Virginia Commonwealth University, VA 23220, USA
[3]Department of Pathology, Massachusetts General Hospital, Charlestown, MA 02129, USA

*Correspondence to: rich.i.joh@gmail.com and mmotamedi@hms.harvard.edu

Abstract

The establishment of distinct transcriptional programs in response to developmental or environmental cues is central to all life. This entails the concordant and discordant transcriptional regulation of several distinct biological processes (BP) often involving thousands of genes. Genomic clustering of genes in a BP is one strategy by which transcriptional coregulation of a BP's genes is achieved. However, whether gene clustering also plays a role in transcriptional coherence of several distinct BPs, often involving thousands of genes, remains unexplored. Here, by analyzing the genomes of eukaryotes ranging from yeast to human, we report the identification of thousands of conserved and species-specific discrete clustered BP pairs, many of which in normal human tissues are transcriptionally correlated. Strikingly, our results reveal that system-level transcriptional coordination is achieved in part by the genic proximity of regulatory nodes of disparate BPs whose coregulation drives the transcriptional coherence of their respective pathways. This, we hypothesize, is one strategy for creating coregulated, tunable modulons in eukaryotes.

Keywords: Genome clustering, Transcriptional coregulation, Transcription factor clustering, Transcriptional Ripple, Modulons, Genome evolution, Eukaryotic genomes, Genomics

Background

To survive stress or deploy developmental programs, cells establish a spectrum of distinct transcriptional states. This involves the rapid, robust and reproducible transcriptional coordination of hundreds or thousands of genes dispersed throughout the genome, governing dozens of disparate biological processes (BPs). To study this type of system-level change in biological outputs, gene regulatory networks (GRNs) were developed to link functionally and temporally the regulatory genes and signaling components of a multitude of BPs [1–6]. These networks are modular and hierarchical [7,8], within which several overrepresented network motifs or sub-circuits [9,10], create positive and/or negative regulatory loops through which biological outputs of up to thousands of genes are coordinated. Redundancies, stratifications and amplifications of motifs within GRNs not only allow the incorporation of signals from a variety of cellular and environmental stimuli [7], but also together produce robust, timely, fine-tuned and stable biological responses. Indeed, the system-level coregulation achieved within GRNs plays a central role in adaptation and development and thus is under strong evolutionary selection [11–13]. Organizationally, GRNs are made up of lower-level sub-circuits such as regulons - a set of genes that are transcriptionally coregulated as a unit - and higher-level sub-circuits, such as modulons, often comprised of several regulons that become transcriptionally linked in response to the same stimuli. Mechanistically, these sub-circuits are interconnected by the combinatorial activities of *trans*-acting factors, such as transcription factors (TF), and chromatin regulatory proteins, targeted to a given set of *cis*-regulatory DNA sequences [12,14]. How the biological outputs of these sub-circuits are organized in time and space is an active area of research.

Gene clustering is one means by which the transcription of large groups of genes can be coordinated. This was first described in prokaryotes where functionally interrelated genes cluster into operons from which poly-cistronic messenger RNAs (mRNAs) are transcribed, ensuring their co-expression [15]. Even though poly- or di-cistronic mRNAs of protein coding genes are rarely found in metazoans except for nematodes [16–18] and flies [19–21], genomic clustering of interrelated genes is a prevalent and conserved feature of eukaryotic genomes in organisms such as yeast [22–24], fly [25–27], worm [28–30], zebrafish [31,32], mouse [33,34] and human [34–41]. In fact, eukaryotic genomes are shaped by multiple domains of transcriptional coregulation [25,27,40], within which a shared chromatin state and *cis*-regulatory elements establish the transcriptional activities of these regions [42–44]. Functionally, gene clustering helps the transcriptional coordination (e.g. housekeeping and highly expressed genes) and temporal regulation (e.g. HOX family of TFs) of interrelated genes [27,35,45–48], and its disruption is linked to several developmental pathologies [49–52]. It also helps coordinate the transcription of unrelated genes at both the micro- (neighboring genes) and macro-scales (A/B compartments and topologically associated domains (TADs)) are transcriptionally correlated [53–59]. However, whether clustering of unrelated genes also has a functional role in coupling the transcription of disparate BPs, spanning many genes dispersed throughout the genome, is not well understood.

Recently, using the fission yeast (*Schizosaccharomyces pombe*), we showed that as cells enter quiescence (G0), the constitutive heterochromatin protein Clr4 - the sole lysine 9 histone H3 (H3K9) methyltransferase in this organism - is deployed to euchromatic parts of the genome to coregulate the expression of hundreds of genes [60]. Unexpectedly, we found that most of these coregulated genes occur in linear gene arrays dispersed throughout the genome, and were together overrepresented in developmental, cell cycle and metabolic BPs, all of which are important for establishing the G0 state [61]. In addition to examples in yeast [62,63], formation of transcriptionally coregulated linear gene arrays are also found in mammals. For example, growth factor-induced exit from quiescence of mouse fibroblast cells also results in the transcriptional activation of immediate early genes (IEG) which spreads in *cis* to neighboring genes via a transcriptional ripple effect [42] creating transcriptionally coordinated linear gene arrays. These and other observations [64,65] suggest that clustering of genes belonging to disparate BPs within a few gene neighborhoods may be a conserved strategy for efficient targeting of the same transcriptional regulatory proteins. This type of genome architecture, we hypothesized, would couple the transcription of disparate BPs, help create modulons and would thus be under strong evolutionary selection.

To test this hypothesis, we devised a statistical framework to quantify genomic clustering between disparate BPs in five of the most commonly used and best-annotated eukaryotic model organisms ranging from yeast to human, spanning over one billion years of evolution [66,67]. Using this platform, we identified thousands of conserved and species-specific clustered BP pairs, many of which are strongly transcriptionally correlated in normal human cells. Interestingly, even in the absence of pathway-wide clustering, BP pairs sharing at least one clustered TF pair display strong transcriptional coherence, suggesting that clustered TFs represents nodes of transcriptional coregulation whose disruption, we find, results in loss of transcriptional coherence. Overall, these data reveal that genic proximity of disparate BPs contributes to their transcriptional coregulation and suggest that such genome organization helps in the rapid and synchronous co-expression of distinct BPs, thus establishing tunable, coregulated modulons in eukaryotes.

Results

Identification of clustered BP pairs in eukaryotic model organisms

To identify significantly clustered BP pairs (Fig. 1A), we developed a statistical framework to quantify clustering using the one-dimensional positioning of protein-coding genes (Fig. 1B) in the fission yeast (*Schizosaccharomyces pombe*), nematode (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*) and human (*Homo sapiens*) genomes. The highly annotated genomes of these organisms endowed our analyses with the power to examine thousands of conserved and species-specific biological pathways across evolution (Fig. 1C). BP terms in the Gene Ontology (GO) consortium [68] were used to define biological pathways in our

analyses. GO provides an up-to-date, pan-organismal classification of BP terms in the aforementioned organisms.

To identify disparate, significantly clustered BP pairs (Fig. 1A-C) (see Methods), we (1) selected BP terms with sufficient number of genes to provide enough statistical power for our analyses in each genome (Table S1), (2) set a threshold distance for gene clustering that captures the majority of the nearby gene-pair transcriptional correlations (Figure S1), and tested its statistical robustness against a range of alternative threshold distances (Figure S2-3), (3) demonstrated the specificity (Figure S4) and performance (Figure S5) of our statistical method using randomly generated gene sets (see Methods), (4) determined background clustering by performing 2,000 random samplings (Figure S6) for each BP-BP analysis in each organism, and (5) estimated *p*-values for each BP-BP analysis in each organism (Table S2). The *p*-values were used to generate quantile-quantile (QQ) plots shown in Fig. 1D-H. The QQ plot for each organism stayed close to the expected diagonal, indicating minimal systematic inflation or deflation. However, the tails of the QQ plots curl upwards in all organisms except the fission yeast indicating that metazoan genomes have thousands of significantly clustered BP pairs. The peculiarity of the fission yeast QQ plot is likely the result of its distinct genome organization relative to metazoans: (A) gene density of protein-coding genes is almost uniform across the fission yeast euchromatic domains, and (B) 50% of all fission yeast protein-coding genes overlap with another gene. Overall, these plots reveal that the statistical framework developed here for capturing BP-BP clustering works effectively in metazoans but may not be ideally suited for the uniformly dense and highly overlapped genomes of yeast species. Next, for each organism, we adjusted the significance of clustering for multiple comparisons using positive false discovery rate [69]. In the end, out of over 2 million BP-BP analyses, 1,855 BP pairs were found to be significantly clustered (*FDR* < 0.05) in the five genomes analyzed (Fig. 1C, Table S2). Fig. 2 shows examples of clustered GO pairs in each organism. Even though some of the clustered BP pairs identified exhibit an overlapping biological function, many BP pairs are functionally disparate (Table S2). Together, these analyses demonstrate that genome clustering of disparate BPs is an organizational feature of eukaryotic genomes.

Several similar BP pairs show conserved clustering in eukaryotic genomes

If gene clustering couples the transcriptional outputs of essential BPs, then it would be under selection and thus maintained in evolution. Accordingly, 17 identical significantly clustered BP pairs were identified in the mouse and human genomes (Table S2). This is largely driven by synteny between these two species which shared a recent common ancestor [70]. Interestingly, we also found that several clustered BP pairs, though not identical, were highly similar to one another.

To determine whether similar BP pairs cluster in these distantly related eukaryotes, we selected BP terms that have a clustered BP partner in at least two organisms. This yielded 393 BP pairs (Table S2). We then used two parallel strategies to answer this question (see Methods). First, using the semantic similarity developed by Lin [71] or Resnik [72], we found that the 393 BP pairs form into 30 highly similar BP groups (Table

S3 and Fig. 3A). The two semantic similarity produced highly overlapping BP groupings (Figure S7A). We then asked in how many different genomes BPs in each group cluster with BPs in other groups. This analysis revealed several highly similar clustered BP pairs in the five eukaryotic genomes (Fig. 3B and Figure S8). In the second parallel strategy, we asked whether a given BP in Table S3 clusters with similar BP terms in the different genomes. For example, if BP1 clusters with BP2 and BP2' in mouse and yeast genomes, respectively, we calculated the BP2-BP2' similarity using semantic similarity by Lin. Our results (Figure S7B) revealed that the frequency with which highly similar (Lin scores >0.9) BP2-BP2' are found is significantly higher than background, supporting the hypothesis that a given BP tends to cluster with highly similar BPs in multiple eukaryotic genomes. Together these data demonstrate that clustering of hundreds of BPs is maintained over long evolutionary timescales, suggesting that this type of genome organization may serve a functional role in linking their transcription. For example, metabolism BPs cluster with stress response BPs in four out of the five genomes, consistent with the well-established link between these two processes in biology [73–75]. Another example is the conservation of clustering between the amino acid metabolism and neural development BPs. Considering the critical role that amino acids and their metabolites play in synaptic transmission [76,77], and learning and memory [78,79], our data suggest that this type of genomic clustering may portend a functional link in their coregulation in organisms ranging from fly to human.

Clustered BPs are transcriptionally correlated

To ask whether genome clustering predicts the transcriptional coupling of disparate BP pairs in the human genome, we quantified BP-BP transcriptional correlation across 53 different tissues represented in the GTEx dataset [80] (Figure S9). Because clustered gene pairs tend to be transcriptionally correlated (Figure S1), we excluded them to remove this bias from all BP-BP correlation calculations. This permitted us to determine whether a BP-BP correlation spans both pathways or is driven by the clustered genes shared by both pathways. Also, because similar GO terms often impact overlapping functions and are thus transcriptionally coordinated, to focus our analysis on disparate BPs, we separated BP pairs into similar (Lin score >0.1) and disparate (Lin score ≤0.1) pairs. To quantify the BP-BP transcriptional correlation, we calculated Z scores for all genes in each of the 53 different tissues and asked whether clustered BP pairs display higher transcriptional correlation compared to unclustered BP pairs (see Methods). We found that disparate clustered BP pairs (N=278) exhibited significantly higher transcriptional correlation compared to unclustered BP pairs (N=458,999) (Fig. 4A), supporting a model in which clustering of disparate BPs links their transcription. Also, because clustered gene pairs were removed from our calculations, these analyses reveal that the BP-BP correlations are pathway-wide.

BP pairs which share a clustered TF display significant transcriptional correlation

Interestingly, several of the clustered gene pairs in clustered BPs were TFs. Because regulatory nodes of GRNs often consist of *trans*-acting co-expressed genes [81,82], this prompted us to test whether clustered TFs are a strong predictor of the transcriptional coherence of clustered BP pairs. We divided disparate clustered BP pairs (N=278) into

three groups: 1) one or both BPs lacks an annotated TF (N=18); 2) both BPs contain at least one annotated TF, and the TFs are not clustered (N=123); 3) both BPs contain at least one annotated TF, and at least one pair of TFs is clustered (N=142). Our analysis revealed that clustered BPs with clustered TFs display the highest transcriptional correlation of the three groups (Fig. 4B). Fig. 4C shows an example of a clustered BP pair with clustered TFs in human, namely humoral immune response (GO:0006959) and negative regulation of cell adhesion (GO:0007162) (also shown in Fig. 2E). Likewise, similar (Lin score>0.1) clustered BP pairs displayed higher transcriptional correlation relative to unclustered BP pairs (Figure S10A), especially those with a clustered TF pair (Figure S10B). Together, these data support a model in which clustered TFs can be used to couple the transcription of BPs in human.

The strong predictive value of clustered TFs in BP-BP correlation suggested that clustered TFs alone (without pathway-wide BP-BP clustering) could be sufficient to predict the transcriptional correlation of BP pairs. If true, this type of gene placement could couple the transcription of multiple disparate BP pairs. To test for this, we compared the transcriptional correlation of disparate BP pairs that share (N=58,760) or do not share (N=401,118) at least one clustered TF pair. By plotting BP-BP transcriptional correlation versus the strength of clustering, we found that (1) genome clustering positively correlates with BP-BP transcriptional correlation, and that (2) BP-BP pairs with at least one clustered TF have a dramatically higher transcriptional correlation compared to BP pairs without a clustered TF (Fig. 4D). In fact, we also found that the most concordant and discordant BP pairs have the highest TF-TF transcriptional correlation consistent with an activator-activator and activator-repressor TF pairing, respectively (Figure S10D). Likewise, similar BP pairs with at least one clustered TF (N=27,898) have a dramatically higher transcriptional correlation compared to BP pairs without a clustered TF (N=158,016) (Figure S10C). In sum, these analyses demonstrate that beyond pathway-wide BP-BP gene clustering, TF clusters alone are strong predictors of the transcriptional coupling of their associated BPs.

TF clustering is required for transcriptional coregulation of discrete BPs

If clustered TFs couple the transcription of their associated BPs, then their disruption (e.g. by a deletion or translocation) may result in the loss of these transcriptional couplings. To test this prediction, we used the Cancer Cell Line Encyclopedia (CCLE) dataset [83], to identify cell lines in which a deletion or a translocation disrupts a clustered TF pair whose associated BPs are transcriptionally correlated. We then quantified the extent to which such a deletion or translocation impacts the transcriptional correlation of the associated BPs (see Methods). As an example, among the CCLE's lymphocytic cell line collection, SUDHL8 is a cell line that carries a deletion in the IRF8 gene disrupting the IRF8-FOXF1 clustered TF pair. Comparing the SUDHL8 transcriptome versus the other lymphocytic cell lines by using standardized residuals as a measure for transcriptional deviation, we found that even though the overall transcriptome of SUDHL8 is typical for a subset of lymphocytic cell line (Fig. 5A), specifically the transcriptional coherence among the IRF8- and FOXF1-associated BPs is lost (Fig. 5B-C). Similarly, among the transverse colon cancer cell lines, the SNU1033 cell line carries a deletion in the PRDM16 gene, disrupting the PRDM16-TP73 clustered

TF pair. Here also we observed that even though the overall transcriptome of SNU1033 is typical for a transverse colon cancer cell line (Figure S11A), specifically the transcriptional coherence among the PRDM16- andTP73-associated BPs is lost (Figure S11B-C).

Deletion of a TF could impact the expression of many genes in its pathway. To test if genomic clustering of TFs specifically can predict the transcriptional coherence of their associated BPs, we analyzed cell lines carrying translocations between clustered TFs in a similar fashion. Fig. 5D illustrates one example where a translocation in the HCC1419 breast cancer cell line disrupts the clustering of HSF1-FOXH1 TF pair. Consistent with our model, we found that even though the overall transcriptome of this cell line is similar to other breast cancer lines, there is a specific loss of transcriptional coherence among the HSF1/FOXH1-associated BP pairs. Fig. 5E and Table S4 summarize similar findings in 31 other cell lines in which a translocation separating a clustered TF pair co-occurs specifically with a significant loss of transcriptional correlation of their associated BPs. Taken together, these data support a model in which clustered TFs couple the transcriptional output of disparate BPs.

Because genes within TADs are transcriptionally correlated, another predication of our model is that the cooccurrence of the clustered TFs in the same TAD would portend the transcriptional correlations of their associated BPs. To test this prediction (see Methods), we analyzed TAD domains identified from 37 independent samples [53,84] (Figure S12A) and compared the frequency with which a clustered TF is found in the same TAD versus the transcriptional correlation of their associated BPs. As expected, clustered TFs found most frequently in the same TAD displayed the highest TF-TF transcriptional correlations (Figure S12B). Moreover, their associated BP pairs displayed the most highly concordant (activator-activator) and discordant (activator-pressor) transcriptional couplings (Fig. 5F). Together these data further support a model in which transcriptionally coupled clustered TFs link the transcriptional outputs of their associated BPs in the human genome.

Temporal regulation of multiple BPs by clustered TFs

So far, our data suggest that clustered TFs can act as regulatory hubs for transcriptional coordination of discrete BPs. Previous work has shown that transcriptional activation can spread in *cis* to neighboring genes via a transcriptional ripple effect, resulting in the transcriptional coregulation of linear gene arrays [42]. Here we tested whether a ripple mechanism could also activate the transcriptional outputs of TF clusters, providing a possible means for temporal co-regulation of multiple pathways by clustered TFs. To test this, we mined the data sets presented in [85], in which SNAI1, a TF and critical regulator of the epithelial-to-mesenchymal transition (EMT), was overexpressed in the MCF10A human breast cell line. In this study, transcriptional changes caused by SNAI1 overexpression were monitored at regular intervals for up to five days post-overexpression. We selected SNAI1 because it is clustered with two other TFs, CEBPB, and ADNP (Figure S13A), which are transcriptional regulators of other pathways. Because all three TFs are primarily repressors, we asked whether the core SNAI1-,

CEBPB-, and ADNP-regulated genes (see Methods) show any signs of time-dependent repression post SNAI1 overexpression. Consistent with a transcriptional ripple effect initiated at SNAI1, we found that the largest fraction of the early-repressed genes is the core SNAI1-regulated (Figure S13B) genes, whereas CEBPB- and ADNP-regulated genes are enriched in the late-repressed gene set (Figure S13C-E). These data support a model in which a transcriptional ripple, initiated by the activation of a TF, can spread in *cis* to its neighboring TFs coordinating the expression of multiple BPs temporally (Fig. 6).

Discussion

Here, the analysis of the distantly related genome of yeast, fly, worm, mouse and human reveals that clustering of disparate BPs is a conserved feature of eukaryotic genomes, demonstrating a strong evolutionary selection for maintaining this type of genome organization. Clustered BP pairs display strong transcriptional correlation in human cells, especially those which also share a clustered TF. Moreover, TF clustering alone is a strong predictor of BP-BP transcriptional correlation, suggesting that formation and maintenance of TF clusters (which we propose form regulatory nodes) provide an efficient genomic architecture for coupling the transcription of GRN sub-circuits. In support of this, we find that deletion, translocation or overexpression of a TF within a clustered TF pair or group impacts the transcriptional coherence of their associated BPs. Fig. 6 depicts a simple model in which transcriptional coregulation of a clustered TF pair drives the transcriptional coherence of their respective BPs in *trans*. According to our model, eukaryotic genomes contain many regulatory nodes through which the combinatorial activities of *trans*-acting factors, such as TFs, can be efficiently coordinated. This in turn endows organisms with the ability to coregulate several GRN sub-circuits simultaneously or in a temporally ordered manner, which could help the timely establishment of distinct biological outcomes rapidly and robustly (Fig. 6).

BP clustering across evolution

Previously, it was shown that genomic clustering of similarly expressed or functionally interrelated genes provides an efficient strategy for coregulating or temporally ordering the expression of many genes in the same pathway [27,35,45–48,62,65,86]. This helps in the establishment of transient, persistent or temporally ordered adaptive GRN states [87], and is under strong evolutionary selection [11–13]. Accordingly, in support of this, a recent evolutionary survey of 341 fungal species revealed that evolution by vertical and horizontal gene transfer of several metabolic BPs has led to the maintenance of the same gene clusters among these fungal species. Strikingly, this was true even in cases of convergent evolution such that the *de novo* acquirement of some BPs leads to the formation of the same gene clusters [88], demonstrating the importance of clustering of interrelated genes in evolution. Here in this report we show that genomic clustering also extends to disparate BP pairs (Fig. 1-3), many of which in normal human tissue display strong transcriptional correlation (Fig. 4A and Table S2). This correlation is pathway-wide and especially strong among BP pairs which also share a clustered TF pair (Fig. 4B). Moreover, even though individual genes, TFs and lower-level GRN sub-circuits

vary across evolution, our data demonstrate that the maintenance of this type of genome organization for higher level BP-BP coregulation spans hundreds of millions of years of evolutionary time (Fig. 3). In addition to revealing clustering among previously known interdependent BPs (e.g. stress response and metabolism and amino acid metabolism and neural development), some clustered BP pairs illustrate emerging interdependencies. For example, we found that learning and histone methylation and learning and RNA methylation are clustered in worm and fly genomes, respectively (Table S3). Indeed, several recent reports have provided support for the emerging role of these modifications in learning and memory in metazoans [89–92]. Another example is the clustering of sphingolipid metabolism genes with serine/threonine kinase signaling genes in mouse and human genomes. Here also, recent data demonstrate the emerging role of sphingolipid metabolites in intra- and intercellular signaling pathways [93–95], which when combined with their genome clustering with serine/threonine kinase signaling pathways suggest a functional link for their transcriptional coupling. Lastly, we also found that some BP clusters do not have clearly defined links in biology (Table S2). It will be interesting to determine whether these kinds of clustered BP pairs portend previously unappreciated biological interdependencies considering their correlated transcriptional outputs in human cells.

Beyond pathway-wide BP-BP gene clustering, we also found that a strong predictor of transcriptional correlation between disparate BPs is the presence of at least one shared clustered TF pair (Fig. 4D, Figure S10C), suggesting that clustered TFs can be used to couple their transcription. Considering that during the continual shuffling of genes in evolution the probability of forming clustered TFs (a single gene pair) is higher than clustered BP pairs (several gene pairs), clustered TFs may present an efficient evolutionary solution for coregulation of disparate BPs, especially those whose biological outputs are similarly regulated in response to the same developmental or stress conditions.

In addition to protein-coding genes, clustering and coregulation of *trans*-acting noncoding factors, such as micro RNAs (miRNAs) also may play important roles in coregulation of multiple BPs, consistent with the presence of their target sequences in multiple transcripts [96]. These, together with TFs, provide additional layers of combinatorial regulation through which system-level changes to transcriptomes can be achieved. For example, TF-miRNA and miRNA-miRNA clustering also may serve as efficient strategies to regulate expression of a wide range of target genes. With better classification and identification of ncRNAs and their targets, the application of our clustering framework may expose novel BP interconnections and other networks motifs involving various coding and noncoding elements.

Clustered TFs as nodes for modulon regulation

Our data reveal that transcriptional coregulation of neighboring genes, such as clustered TFs, is a key driver of BP-BP coherence (Fig. 5 and Figure S9-10). This is conserved in organisms ranging from bacteria to human and plays a critical role in coregulation and thus establishment of GRN sub-circuits [9–11]. In fact, in human,

genes that display similar expression changes across different tissues occur in clusters, suggesting that tissue-level GRN regulation occurs at gene-cluster or chromosomal-domain scale [97]. There are several molecular mechanisms that help coregulate gene neighbors in *cis* such as Transcription interference [98,99], divergent transcription [100], and regulation of local chromatin states [101,102]. Another is the so-called transcriptional ripple effect observed in yeast [62,63] and mammalian cells [42]. In mammals, transcriptional activation of immediate early genes (IEGs) causes time-lagged ripples of gene expression and active chromatin marks (histone H3 and H4 acetylation) that spread in *cis* from IEGs to neighboring genes. Similarly, transcriptional silencing demarcated by H3K9 and/or H3K27 methylation can spread in *cis* into neighboring regions [103,104] to repress the transcriptional output of a group of linked genes. In this paper, we showed that SNAI1 induction causes time-lagged coregulation of CEBPB- and ADNP-associated genes (Fig. 5E-F). Interestingly, both CEBPB and ADNP play important roles in EMT [105,106], suggesting that neighboring TFs like SNAI1, CEBPB, and ADNP may represent a regulatory node for coordinating the expression of multiple eukaryotic regulons, facilitating EMT during development [107]. Based on these observations, we hypothesize that eukaryotic genomes contain multiple regulatory nodes through whose coregulation or temporally ordered transcriptional activation, the transcriptional output of multiple distinct BPs can be coupled. We hypothesize that this, in turn, helps create modulons, orchestrating the coordinate establishment of GRNs in eukaryotes (Fig. 6). Indeed, such a model is consistent with the recent observations that combinations of TFs [108–110] and enhancers [111] define the various cell states in mammals. Because in the human genome, gene clusters showing the highest transcriptional coupling across different tissues occur in regions of high gene density [97], we hypothesize that TF clusters in high gene density regions present promising targets for probing novel biological interrelationships between BPs.

Conclusions

Overall, based on these analyses we hypothesize that as eukaryotic genomes expanded in size and complexity during evolution, modulons were maintained and created by the formation of *trans*-acting regulatory nodes (such as TF clusters) through which the biological outputs of disparate BPs were coupled. This, we propose, is a conserved organizing principle of eukaryotic genomes. Additionally, the transcriptionally linked disparate BP pairs identified in this study not only support the well-known interdependencies among some disparate BPs, but also suggest the existence of new BP interconnections which future studies may uncover their molecular links. Finally, because the functional, hierarchical and temporal organization of these sub-circuits underlie the complex developmental and adaptive programs deployed in eukaryotes, in future studies it will be informative to ask whether and how the loss or inappropriate gain of these coregulations impacts disease states such as cancer.

Methods

Gene lists and gene ontology (GO) terms

The gene ontology (GO) annotation files for *S. pombe* and other organisms (*D. melanogaster*, *C. elegans*, *M. musculus*, and *H. sapiens*) were obtained from PomBase [112] and Gene Ontology Consortium [68,113], respectively. To focus our analyses on biological pathways, we chose Biological Process (BP) GO terms only. We also limited our analyses to BP GO terms with gene numbers suitable for generating sufficient statistical power in our analyses. In each organism, these gene number cutoffs were set after accounting for the total number of genes with at least one associated BP GO term and the number of chromosomes. For *S. pombe*, BP GO terms which contain10-250 genes, for *C. elegans* and *D. melanogaster* BPs with 20-250 genes, and for *M. musculus* and *H. sapiens* BPs with 50-250 genes were considered for analysis (for total BPs in each organism, see Fig. 1C). The list of genes and their coordinates are from PomBase [112] for *S. pombe* and Ensemble Genomes [114] for the other model organisms. We also removed largely redundant BP terms as defined by those containing 75% or more identical genes. In such instances, the BP term with more genes was retained and the other was eliminated from analyses. This reduced the number of BP-BP analyses between largely redundant BP pairs. The number of genes and BP terms analyzed in this work can be found in Fig. 1C. The list of all BP terms in each organism can be found at https://figshare.com/projects/Transcriptional_coherence/72644.

Identification of clustered BP pairs

Setting the threshold distance for clustering and testing its statistical robustness

Because in most eukaryotes *cis*-regulatory elements and local gene-gene interactions driving transcriptional correlation of gene pairs tend to occur within five times (5X) the average intergenic distance [53,115], we assessed whether this distance (1MB in human) also captures the majority of proximity-driven gene pair transcriptional correlations found in the human genome. We used the human GTEx datasets [80] and quantified transcriptional correlation among all protein-coding gene pairs across different tissues (see below for details on how transcriptional correlations were calculated). Consistent with previous work [25,27], our data revealed that (1) transcriptional correlation is a function of the gene pair's intergenic distance, and (2) 5X mean intergenic distance (1Mb) is a good threshold distance capturing the majority of *cis* transcriptional correlations (Figure S1).

To confirm the statistical robustness of this threshold distance, we performed our clustering (see below) analyses and *p*-value quantifications for a range of different threshold distances (1X, 2X, 7X, 10X, 15X and 20X) and compared these *p*-values versus those generated using the 5X intergenic distance. This comparison was done for all organisms for all BP pairs. Figure S2 shows our analyses for the human genome as an example. In sum, these analyses revealed that significantly clustered BP pairs are robust at several different threshold distances in all model genomes analyzed in this report.

Also, we compared the overlap of highly clustered BP pairs ($p<0.001$) among the different threshold values in each organism (Figure S3). Based on these data, we set

5X mean intergenic distance as the threshold distance for gene-pair clustering for the genomes used in this study.

Determining the number of clustered gene pairs

To quantify clustering between two BP terms (e.g., BP1 and BP2), we determined the number of instances genes from BP1 lie within the threshold distance of genes from BP2. Gene coordinates were used to count instances of gene pair clustering in BP1 and BP2. To eliminate clustering overestimation, if identical genes were found in the two BPs, they were removed from consideration in our analyses.

Specificity and performance of the statistical method used to quantify genome clustering

To test the specificity of our statistical framework, we generated several random gene sets by selecting genes from the pool of BP-associated genes used in this study in each organism. Similar to the gene number limits used to select BPs in each organism (see above), the number of randomly selected genes were 10-250 in yeast, 20-250 in worm and fly, and 50-250 in mouse and human. Next, all pairwise clustering analyses were performed and QQ plots were generated as depicted in Figure S4.

To monitor the performance of our statistical method, we selected 100 artificially generated gene sets and analyzed how the *p*-value for clustering changes as more clustered gene pairs are added incrementally to each analysis. As expected, the artificial addition of clustered gene pairs decreases the *p*-value for clustering (Figure S5). We also found that the magnitude of decrease in *p*-value caused by the incremental addition of clustered gene pairs to a given BP-BP analysis varies depending on the distribution of each gene set pair.

Determining background clustering using Poisson distribution

Background clustering was calculated for each BP-BP analysis in each organism. In total, over 2 million clustering analyses were performed in this study. To illustrate how background clustering between two BP terms was calculated, we use the following example: BP1 has 20 and BP2 has 30 genes. To determine background clustering between BP1 and BP2, first the 20 genes in BP1 were fixed in the genome and 30 genes were randomly selected 1,000 times from among the pool of genes with at least one ascribed BP designation. The numbers of clustered gene pairs generated from these 1,000 random samplings were then fit into a Poisson distribution. Next, the process was repeated by fixing the 30 genes in BP2, and 20 genes were picked randomly 1,000 times from among the pool of genes with at least one ascribed BP designation. The numbers of clustered gene pairs generated from these 1,000 random samplings were also fit into a Poisson distribution. In total, over 4 billion random samplings were performed to generate the Poisson graphs used in our study.

To calculate the significance of clustering for each pairwise analysis, we compared the observed number of clustered gene pairs versus the Poisson distribution generated by the background clustering as described above. Figure S6 shows simulated and fitted Poisson distributions for four representative sets of BP-BP analyses in the human

genome. Note that the Poisson distribution can overestimate the *p*-value when the number of clustered gene pairs is large. In these instances, the fitted distribution is wider than the actual distribution (Figure S6D), thus this method is a stringent way to calculate *p*-values. This generates two *p*-values for each clustering analysis of which we used the smaller of the two. These values were used to generate the quantile-quantile (QQ) plots depicted in Fig. 1D-H.

Adjusting *p*-values and selecting highly clustered BP pairs

Positive false discovery rate (pFDR) [69] was used to deal with the multiple comparison problem using a cutoff of 0.05. Also, we noted that some BP terms tended to cluster with many BP terms. To reduce this bias in our analyses, we kept only the top five most significantly clustered BP pairs in these instances. The list of significantly clustered BP pairs in all organisms can be found in Table S2. Significantly clustered BP pairs that had a Lin similarity score (see below for details) of 0.1 or less were considered disparate.

Conservation of clustering

To ask whether similar BP pairs cluster in the distantly related eukaryotes analyzed in this study, we selected BP terms which have clustered counterparts in at least two model organisms. The cutoff of two also permits the identification of mammalian-specific (mouse and human) or metazoan-specific (fly and worm) clustered BP pairs. This generated a list of 396 BP terms. Of the 396 BP terms 3 were absent in the R package and were not considered further. Using semantic similarity measurement scheme developed by Lin [71] and the R package GOSemSim [116], we found that the 393 BP terms form 30 highly similar BP groups (Table S3) (Fig. 3A and Figure S7A). Highly similar BP groupings were defined by performing hierarchical clustering analysis on the Lin's similarity scores (ranging from 0-1.0) with the Euclidean pairwise distance (Fig. 3A). The robustness of these BP groupings were tested against the Resnik semantic similarity method [72] revealing highly similar results (Figure S7A). We then asked in how many different genomes BPs in each group cluster with BPs in the other groups (Fig. 3B and Figure S8).

We also used a parallel strategy in which we asked whether a given BP in Table S3 clusters with similar BP terms in the different genomes. For example, if BP1 clusters with BP2 and BP2' in mouse and yeast respectively, we used Lin to calculate the BP2-BP2' similarity score. The distribution of these values and the frequency with which highly similar (Lin scores >0.9) BP2-BP2' are found are depicted in Figure S7B.

Human transcriptome analysis

To quantify the transcriptional correlation between two BPs, we used the datasets available at the GTEx portal (v7 dataset) which contain transcriptomes of 53 different human tissues [80] from healthy individuals. The GTEx data were converted to mean $\log_2$ Transcript Per Million (TPM) values for each gene in every tissue (Figure S9A). Then the mean log2(TPM+1) expression for each gene was calculated across all 53 tissues. For each gene, these values were then normalized by subtracting the mean expression value in each tissue (Figure S9B). The normalized TPM values were used to calculate Z-scores (deviation from the mean log TPM) for each gene across tissues

(shown in Figure S9C). Using these data, we calculated transcriptional correlation between two BP terms as the correlation between mean tissue expression in BP1 and BP2 by summing all the Z-scores in each BP (53 mean tissue-specific Z-scores for each BP term as shown in Figure S9D).

Calculating transcriptional correlations

For transcriptional correlation calculations, we excluded genes that are present both in BP1 and BP2 from our analysis. Also, because clustered gene pairs are transcriptionally correlated (Figure S1), to remove this bias from our correlation analyses, we excluded the clustered gene pairs from all BP-BP correlation calculations. This permitted us to determine if the observed BP-BP correlations are driven by gene clusters or the entire pathway. Also, because similar GO terms often impact overlapping functions and are thus transcriptionally coordinated, to focus our analysis on disparate BPs, we separated BP pairs into two groups defined based on their semantic similarity score: (1) similar (Lin score >0.1) and (2) disparate BP pairs (Lin score <=0.1) [71]. List of all BP pairs, the presence of shared clustered TFs, semantic similarity and transcriptional correlation scores can be found here (https://figshare.com/projects/Transcriptional_coherence/72644). Fig. 4 and Figure S10 illustrate the impact of clustered TF pairs on the transcriptional correlation of disparate (Lin score <=0.1) and similar (Lin score >0.1) pairs, respectively. 'circlize' package from the Comprehensive R Archive Network was used to draw the graph in Fig. 4C [117]

Human transcription factors

The list of human TFs were obtained from UniProt [118].

Cancer cell line analysis

Cancer Cell Line Encyclopedia (CCLE) [83] provide both gene expression and copy number variation data. For our analyses, we used 806 cell lines for which both transcriptome and copy number data were available. To estimate if a cancer cell line is an outlier in terms of the transcriptional coherence among cancer cell lines of the same type, we used the standardized residual of linear regression from a BP pair (BP1/BP2) within a specific tumor type. We calculated the mean expression of two BP terms for every cell line (excluding overlapping genes), and then used it to estimate two standardized residuals based on BP1 and BP2 as independent variables. The root mean squared values of two standardized residuals for TF-associated BP pairs (e.g. TF1 is associated with BP1, and TF2 is associated with BP2) and background (10,000 random BP pairs) are shown in Fig. 5C-D and Figure S11C. This analysis was performed for cell lines that carry a deletion which eliminates one of the two clustered TFs or a translocation which physically separates a clustered TF pair.

Topologically associated domains (TADs)

To ask if the genes in a clustered TF pair occur in the same TAD, we used the 3D Genome Browser data and associated studies [53,84]. These datasets (http://promoter.bx.psu.edu/hi-c/publications.html) predict TADs by applying the Dixon et al. [53] pipeline to analyze their and other published datasets [55,119]. For each clustered TF pair, we calculated the number of times the clustered TFs were found in

the same TAD in the 37 samples analyzed in this study. For Fig. 5F and Figure S12, all the values sorted by x-values and y-values represent 50-points moving average.

SNAI1 overexpression data

To study the effect of SNAI1 overexpression on neighboring TFs, we mined the datasets presented in the Javaid et al., 2013 paper, in which SNAI1, a TF and a critical regulator of epithelial to mesenchymal transition (EMT), was overexpressed in the MCF10A breast cancer cell line. In this study, transcriptional changes caused by SNAI1 overexpression were monitored at regular intervals for up to five days post-overexpression. We used the early/late/transiently up/down-regulated gene sets reported to create the early- and late-repressed gene sets used in this paper. The early-repressed gene set was based on the combined transcriptomes of samples collected 3, 6 and 12hr after SNAI1 overexpression; late-repressed gene set was based on the combined transcriptomes of samples collected 72 and 120hrs after SNAI1 overexpression.

To identify core genes that are regulated by SNAI1 and TFs which lie in its vicinity, namely ADNP and CEBPB (Figure S13A), we calculated the number of times a gene is associated with a SNAI1-, CEBPB -, or ADNP-associated BP term. We reasoned that the higher the number of times that a gene belongs to a BP regulated by the TF, the more likely that this gene is one of the core genes regulated by the TF. For Figure S13B and S13C, we used genes which belong to seven SNAI1-associated, 10 CEBPB-associated and 8 ADNP-associated BP terms each containing roughly 75 genes (Table S5).

Availability of data and materials

Data for all pairwise analyses and MATLAB codes for this project can be found at https://figshare.com/projects/Transcriptional_coherence/72644.

Acknowledgement

We thank Meeta Mistry for her help with the semantic similarity analysis. We are grateful to Russell Jenkins for helpful suggestion regarding sphingolipid metabolism and kinase signaling. We also thank the members of the Motamedi Lab and Nick Dyson for critical reading of this manuscript.

Author contribution

MM conceived the concepts and ideas. RJ developed the statistical framework in consultation with ML and MA. RJ performed all data analyses. ML and MA vetted the analyses. RJ and MM wrote the manuscript which was edited by ML and MA. All authors have read and approved the final manuscript.

Funding

This work was supported by NIH (R01GM125782), American Cancer Society Research Scholar Award and Howard M. Goodman Fellowship to MM.

Supplementary information

Additional file 1: Figures S1-13.
Additional file 2: Table S1: List of all BP terms
Additional file 3: Table S2: List of clustered BP pairs, including the number of observed gene pairs and background, p and pFDR values
Additional file 4: Table S3: Groups of BP terms based on hierarchical clustering
Additional file 5: Table S4: List of translocations between clustered TFs in CCLE collection, which cooccur with the loss of transcriptional correlation of the associated BPs.
Additional file 6: Table S5: List of SNAI1/CEBPB/ADNP-associated genes

Reference

1. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, et al. A genomic regulatory network for development. Science. 2002. p. 1669–78.
2. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science. 2002;298:799–804.
3. Stathopoulos A, Levine M. Genomic regulatory networks and animal development. Dev Cell. 2005;9:449–62.
4. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22:281–5.
5. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. J Comput Biol; 2000;7:601–20.
6. Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature. Nature Publishing Group; 2000;408:307–10.
7. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nat Rev Genet. Nature Publishing Group; 2009;10:141–8.
8. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics. Narnia; 2001;17:126–36.
9. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science. 2002;298:824–7.
10. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet. 2002;31:64–8.
11. Thompson D, Regev A, Roy S. Comparative analysis of gene regulatory networks: From network reconstruction to evolution. Annu Rev Cell Dev Biol. Annual Reviews; 2015;31:399–428.
12. Halfon MS. Perspectives on gene regulatory network evolution. Trends Genet. Elsevier Ltd; 2017. p. 436–47.
13. Peter IS, Davidson EH. Evolution of gene regulatory networks controlling body plan development. Cell. NIH Public Access; 2011;144:970–85.
14. Maeso I, Irimia M, Tena JJ, Casares F, Gómez-Skarmeta JL. Deep conservation of cis-regulatory elements in metazoans. Philos Trans R Soc B Biol Sci. Royal Society of London; 2013;368.
15. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961;3:318–56.
16. Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell. 1993;73:521–32.
17. Brogna S, Ashburner M. The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. EMBO J. European Molecular Biology Organization; 1997;16:2023–31.
18. Blumenthal T, Davis P, Garrido-Lecca A. Operon and non-operon gene clusters in the C. elegans genome. WormBook. Various; 2015;1–20.
19. Crosby MA, Sian Gramates L, dos Santos G, Matthews BB, St. Pierre SE, Zhou P, et al. Gene model annotations for Drosophila melanogaster: The rule-benders. G3 Genes, Genomes, Genet. Genetics Society of America; 2015;5:1737–49.
20. Michalak K, Orr WC, Radyuk SN. Drosophila peroxiredoxin 5 is the second gene in

a dicistronic operon. Biochem Biophys Res Commun. NIH Public Access; 2008;368:273–8.
21. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, et al. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. Genome Biol. BioMed Central; 2002;3:RESEARCH0083.
22. Cohen BA, Pilpel Y, Mitra RD, Church GM. Discrimination between paralogs using microarray analysis: application to the Yap1p and Yap2p transcriptional networks. Yamamoto KR, editor. Mol Biol Cell. 2002;13:1608–14.
23. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature. Nature Publishing Group; 2003;424:194–7.
24. Poyatos JF, Hurst LD. The determinants of gene order conservation in yeasts. Genome Biol. BioMed Central; 2007;8:R233.
25. Spellman PT, Rubin GM. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol. 2002;1:5.
26. Mezey JG, Nuzhdin S V, Ye F, Jones CD. Coordinated evolution of co-expressed gene clusters in the Drosophila transcriptome. BMC Evol Biol. BioMed Central; 2008;8:2.
27. Weber CC, Hurst LD. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. Genome Biol. BioMed Central; 2011;12:R23.
28. Blumenthal T, Gleason KS. Caenorhabditis elegans operons: form and function. Nat Rev Genet. Nature Publishing Group; 2003;4:110–8.
29. Roy PJ, Stuart JM, Lund J, Kim SK. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature. Nature Publishing Group; 2002;418:975–9.
30. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature. Nature Publishing Group; 2003;421:231–7.
31. Tsai HK, Huang PY, Kao CY, Wang D. Co-expression of neighboring genes in the zebrafish (Danio rerio) genome. Int J Mol Sci. Molecular Diversity Preservation International; 2009;10:3658–70.
32. Ng YK, Wu W, Zhang L. Positive correlation between gene coexpression and positional clustering in the zebrafish genome. BMC Genomics. BioMed Central; 2009;10:42.
33. Li Q, Lee BT, Zhang L. Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics. BioMed Central; 2005;6:7.
34. Singer GAC, Lloyd AT, Huminiecki LB, Wolfe KH. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol. 2005;22:767–75.
35. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, et al. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science. American Association for the Advancement of Science; 2001;291:1289–92.
36. Lee JM, Sonnhammer ELL. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. Cold Spring Harbor Laboratory Press; 2003;13:875–82.
37. Fukuoka Y, Inaoka H, Kohane IS. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics. BioMed Central; 2004;5:4.

38. Sémon M, Lobry JR, Duret L. No evidence for tissue-specific adaptation of synonymous codon usage in humans. Mol Biol Evol. Narnia; 2006;23:523–9.
39. Makino T, McLysaght A. Interacting gene clusters and the evolution of the vertebrate immune system. Mol Biol Evol. Narnia; 2008;25:1855–62.
40. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. Academic Press; 2008;91:243–8.
41. Al-Shahrour F, Minguez P, Marqués-Bonet T, Gazave E, Navarro A, Dopazo J. Selection upon genome architecture: conservation of functional neighborhoods with changing genes. Eisen JA, editor. PLoS Comput Biol. Public Library of Science; 2010;6:e1000953.
42. Ebisuya M, Yamamoto T, Nakajima M, Nishida E. Ripples from neighbouring transcription. Nat Cell Biol. Nature Publishing Group; 2008;10:1106–13.
43. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. American Society of Plant Biologists; 2009;150:535–46.
44. Feuerborn A, Cook PR. Why the activity of a gene depends on its neighbors. Trends Genet. Elsevier; 2015;31:483–90.
45. Lercher MJ, Urrutia AO, Hurst LD. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet. Nature Publishing Group; 2002;31:180–3.
46. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, et al. Zebrafish hox clusters and vertebrate genome evolution. Science. 1998;282:1711–4.
47. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. Science. American Association for the Advancement of Science; 2006;313:1918–22.
48. Versteeg R, van Schaik BDC, van Batenburg MF, Roos M, Monajemi R, Caron H, et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. Cold Spring Harbor Laboratory Press; 2003;13:1998–2004.
49. Boulding H, Webber C. Large-scale objective association of mouse phenotypes with human symptoms through structural variation identified in patients with developmental disorders. Hum Mutat. John Wiley & Sons, Ltd; 2012;33:874–83.
50. Doelken SC, Kohler S, Mungall CJ, Gkoutos G V., Ruef BJ, Smith C, et al. Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. Dis Model Mech. 2013;6:358–72.
51. Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S, et al. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature. NIH Public Access; 2012;485:363–7.
52. Andrews T, Honti F, Pfundt R, de Leeuw N, Hehir-Kwa J, Vulto-van Silfhout A, et al. The clustering of functionally related genes contributes to CNV-mediated disease. Genome Res. Cold Spring Harbor Laboratory Press; 2015;25:802–13.
53. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. Nature Publishing Group; 2012;485:376–80.
54. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-

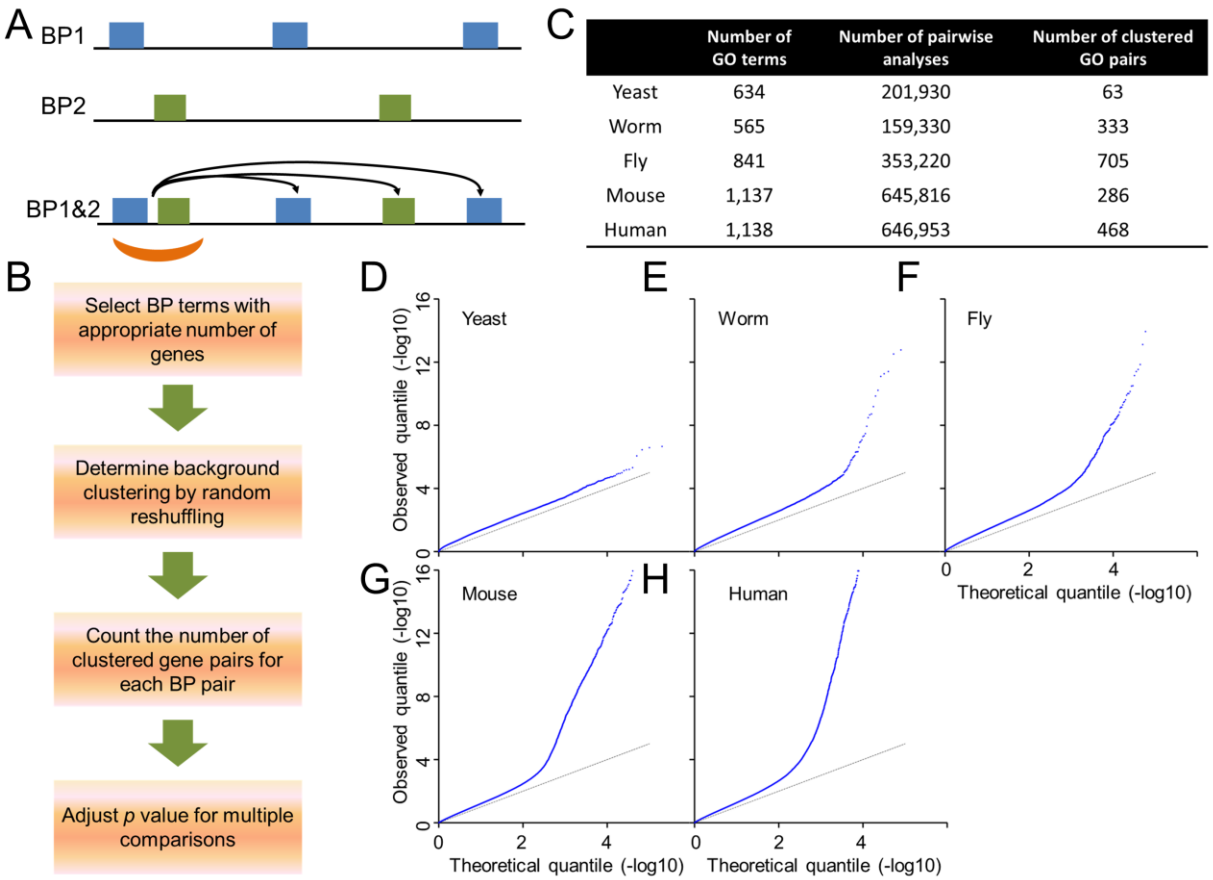
enhancer interactions. Cell. Cell Press; 2015;161:1012–25.
55. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. American Association for the Advancement of Science; 2009;326:289–93.
56. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. Nature Publishing Group; 2015;47:598–606.
57. Le Dily F, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, et al. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. Cold Spring Harbor Laboratory Press; 2014;28:2151–62.
58. Irimia M, Maeso I, Roy SW, Fraser HB. Ancient cis-regulatory constraints and the evolution of genome architecture. Trends Genet. Trends Genet; 2013. p. 521–8.
59. Irimia M, Tena JJ, Alexis MS, Fernandez-Miñan A, Maeso I, Bogdanović O, et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. Cold Spring Harbor Laboratory Press; 2012;22:2356–67.
60. Joh RI, Khanduja JS, Calvo IA, Mistry M, Palmieri CMCM, Savol AJAJ, et al. Survival in quiescence requires the euchromatic deployment of Clr4/SUV39H by Argonaute-associated small RNAs. Mol Cell. 2016;64:1088–101.
61. Valcourt JR, Lemons JMS, Haley EM, Kojima M, Demuren OO, Coller HA. Staying alive. http://dx.doi.org/104161/cc19879. Taylor & Francis; 2012;
62. Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet. 2000;26:183–6.
63. Janicki SM, Tsukamoto T, Salghetti SE, Tansey WP, Sachidanandam R, Prasanth K V, et al. From silencing to gene expression: real-time analysis in single cells. Cell. 2004;116:683–98.
64. Purmann A, Toedling J, Schueler M, Carninci P, Lehrach H, Hayashizaki Y, et al. Genomic organization of transcriptomes in mammals: Coregulation and cofunctionality. Genomics. Academic Press; 2007;89:580–7.
65. Eldabagh RS, Mejia NG, Barrett RL, Monzo CR, So MK, Foley JJ, et al. Systematic identification, characterization, and conservation of adjacent-gene coregulation in the budding yeast Saccharomyces cerevisiae. mSphere. American Society for Microbiology Journals; 2018;3.
66. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. Nature Publishing Group; 2016;1:16048.
67. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006;311:1283–7.
68. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
69. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q -value. Ann Stat. Institute of Mathematical Statistics; 2003;31:2013–35.
70. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520–62.

71. Lin D. An information-theoretic definition of similarity. Proc Fifteenth Int Conf Mach Learn. Morgan Kaufmann Publishers; 1998;296–304.
72. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proc 14TH Int Jt Conf Artif Intell. 1995;1:448--453.
73. Hotamisligil GS, Davis RJ. Cell signaling and stress responses. Cold Spring Harb Perspect Biol. 2016;8:a006072.
74. Wellen KE, Thompson CB. Cellular metabolic stress: Considering how cells respond to nutrient excess. Mol Cell. 2010;40:323–32.
75. Green DR, Galluzzi L, Kroemer G. Metabolic control of cell death. Science. 2014;345:1250256–1250256.
76. Fernstrom JD, Fernstrom MH. Tyrosine, phenylalanine, and catecholamine synthesis and function in the brain. J Nutr. 2007;137:1539S-1547S.
77. Fernstrom JD. Aromatic amino acids and monoamine synthesis in the central nervous system: influence of the diet. J Nutr Biochem. Elsevier; 1990;1:508–17.
78. Jakeman PM. Amino acid metabolism, branched-chain amino acid feeding and brain monoamine function. Proc Nutr Soc. 1998;57:35–41.
79. Usuda K, Kawase T, Shigeno Y, Fukuzawa S, Fujii K, Zhang H, et al. Hippocampal metabolism of amino acids by L-amino acid oxidase is involved in fear learning and memory. Sci Rep. Nature Publishing Group; 2018;8:11073.
80. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. NIH Public Access; 2013;45:580–5.
81. Mao L, Van Hemert JL, Dash S, Dickerson JA. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. BioMed Central; 2009;10:346.
82. Tornow S, Mewes HW. Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res. Oxford University Press; 2003;31:6283–9.
83. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–7.
84. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D genome browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. BioMed Central Ltd.; 2018;19:151.
85. Javaid S, Zhang J, Anderssen E, Black JC, Wittner BS, Tajima K, et al. Dynamic chromatin modification sustains epithelial-mesenchymal transition following inducible expression of Snail-1. Cell Rep. Howard Hughes Medical Institute; 2013;5:1679–89.
86. Wang G-Z, Chen W-H, Lercher MJ. Coexpression of linked gene pairs persists long after their separation. Genome Biol Evol. 2011;3:565–70.
87. Peter IS, Davidson EH. Implications of developmental gene regulatory networks inside and outside developmental biology. Curr Top Dev Biol. Academic Press Inc.; 2016;117:237–51.
88. Marcet-Houben M, Gabaldón T. Evolutionary and functional patterns of shared gene neighbourhood in fungi. Nat Microbiol. Nature Publishing Group; 2019;4:2383–92.
89. Li J, Yang X, Qi Z, Sang Y, Liu Y, Xu B, et al. The role of mRNA m6A methylation in the nervous system. Cell Biosci. BioMed Central; 2019;9:66.
90. Widagdo IS, Pratt NL, Roughead EE. The association between frailty and medicines use over time: an analysis using the Australian Longitudinal Study on Ageing population. J Pharm Pract Res. John Wiley & Sons, Ltd; 2018;48:405–15.

91. Jarome TJ, Lubin FD. Histone lysine methylation: critical regulator of memory and behavior. Rev Neurosci. 2013;24:375–87.
92. Collins BE, Greer CB, Coleman BC, Sweatt JD. Histone H3 lysine K4 methylation and its role in learning and memory. Epigenetics Chromatin. BioMed Central; 2019;12:7.
93. Cartier A, Hla T. Sphingosine 1-phosphate: Lipid signaling in pathology and therapy. Science. American Association for the Advancement of Science; 2019;366:eaar5551.
94. Montefusco DJ, Matmati N, Hannun YA. The yeast sphingolipid signaling landscape. Chem Phys Lipids. 2014;177:26–40.
95. Hannun YA, Obeid LM. Sphingolipids and their metabolism in physiology and disease. Nat Rev Mol Cell Biol. 2018;19:175–91.
96. Hausser J, Zavolan M. Identification and consequences of miRNA–target interactions — beyond repression of gene expression. Nat Rev Genet. Nature Publishing Group; 2014;15:599–612.
97. Ghanbarian AT, Hurst LD. Neighboring genes show correlated evolution in gene expression. Mol Biol Evol. 2015;32:1748–66.
98. Rosa S, Duncan S, Dean C. Mutually exclusive sense–antisense transcription at FLC facilitates environmentally induced gene repression. Nat Commun. 2016;7:13031.
99. Gullerova M, Proudfoot NJ. Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. Nat Struct Mol Biol. Europe PMC Funders; 2012;19:1193–201.
100. Seila AC, Core LJ, Lis JT, Sharp PA. Divergent transcription: A new feature of active promoters. Cell Cycle. 2009;8:2557–64.
101. de Laat W, Grosveld F. Spatial organization of gene expression: the active chromatin hub. Chromosom Res. Kluwer Academic Publishers; 2003;11:447–59.
102. Talbert PB, Henikoff S. Spreading of silent chromatin: inaction at a distance. Nat Rev Genet. Nature Publishing Group; 2006;7:793–803.
103. Bintu L, Yong J, Antebi YE, McCue K, Kazuki Y, Uno N, et al. Dynamics of epigenetic regulation at the single-cell level. Science. 2016;351:720–4.
104. Erdel F, Greene EC. Generalized nucleation and looping model for epigenetic memory of histone modifications. Proc Natl Acad Sci. 2016;113:E4180–9.
105. Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer. BioMed Central; 2007;7:55.
106. Li J, Shan F, Xiong G, Chen X, Guan X, Wang J-M, et al. EGF-induced C/EBP participates in EMT by decreasing the expression of miR-203 in esophageal squamous cell carcinoma cells. J Cell Sci. 2014;127:3735–44.
107. Wang Y, Shi J, Chai K, Ying X, Zhou BP. The role of Snail in EMT and tumorigenesis. Curr Cancer Drug Targets. NIH Public Access; 2013;13:963–72.
108. Braun T, Gautel M. Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. Nat Rev Mol Cell Biol. 2011;12:349–61.
109. Martello G, Smith A. The nature of embryonic stem cells. Annu Rev Cell Dev Biol. Annual Reviews; 2014;30:647–75.
110. Waardenberg AJ, Ramialison M, Bouveret R, Harvey RP. Genetic networks governing heart development. Cold Spring Harb Perspect Med. 2014;4:a013839.
111. Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, et al. Interactome

maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell. 2013;155:1507–20.
112. Wood V, Harris MA, McDowall MD, Rutherford K, Vaughan BW, Staines DM, et al. PomBase: a comprehensive online resource for fission yeast. Nucl Acids Res. 2012;40:D695-9.
113. The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. Oxford University Press; 2017;45:D331–8.
114. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. Oxford University Press; 2018;46:D754–61.
115. Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD, et al. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. Nature. 2014;516:432–5.
116. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. Oxford University Press; 2010;26:976–8.
117. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30:2811–2.
118. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. Oxford University Press; 2017;45:D158–69.
119. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. Nature Publishing Group; 2015;518:350–4.

Main Figures



	Number of GO terms	Number of pairwise analyses	Number of clustered GO pairs
Yeast	634	201,930	63
Worm	565	159,330	333
Fly	841	353,220	705
Mouse	1,137	645,816	286
Human	1,138	646,953	468

Fig. 1.
Hundreds of disparate Gene Ontology (GO) biological process (BP) pairs are clustered in the genomes of organisms ranging from yeast to human. (A) Scheme depicting how clustered BP pairs were identified. Blue and green boxes depict genes in BP1 and BP2, respectively. (B) Flow chart of all analytical steps in this study. (C) Table summarizing the numbers of BP terms, pairwise analyses, and significantly clustered BP pairs identified in each organism. (D-H) Quantile-quantile (Q-Q) plots of observed (Y-axis) versus theoretical (X-axis) of *p*-value distributions for BP-BP clustering in (D) yeast (*S. pombe*), (E) worm (*C. elegans*), (F) fly (*D. melanogaster*), (G) mouse (*M. musculus*) and (H) human (*H. sapiens*).

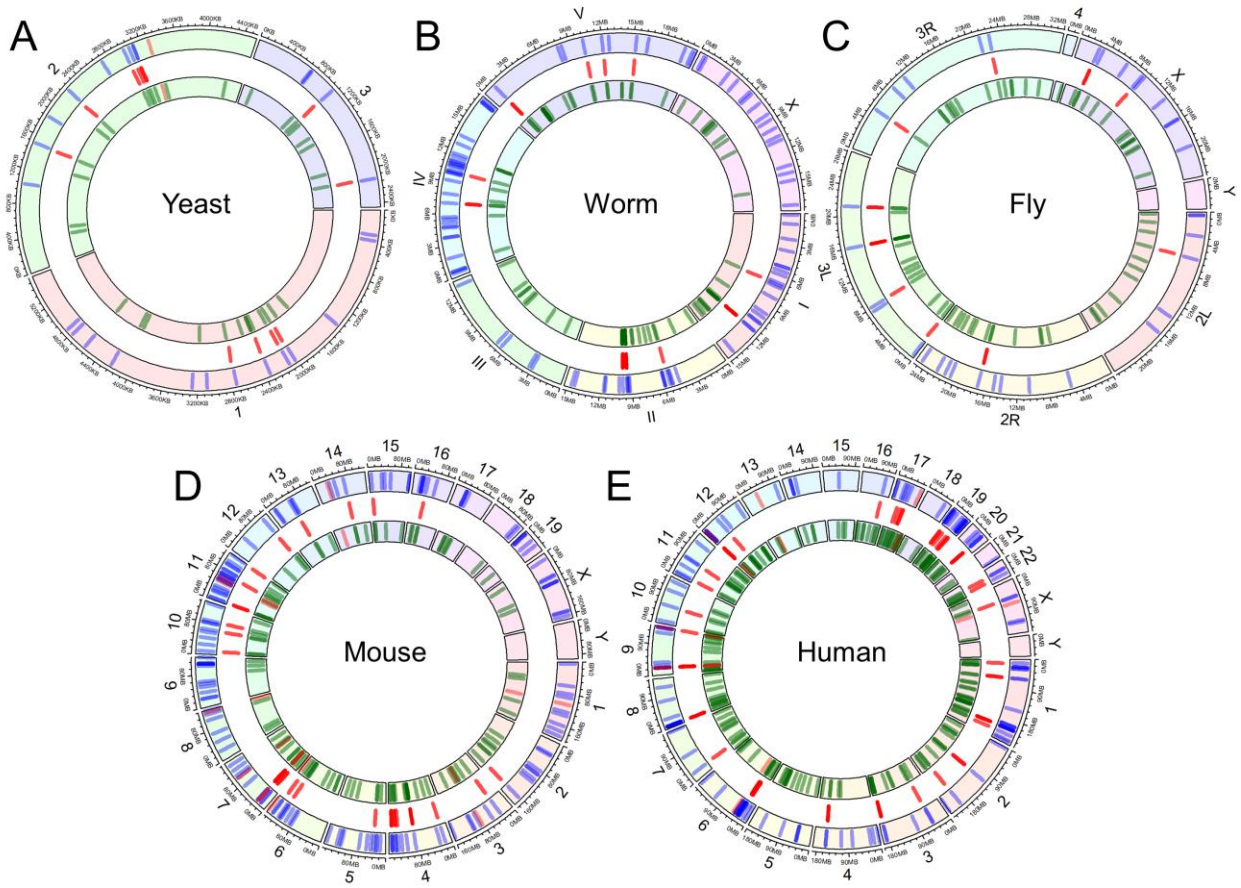


Fig. 2.

Clustered BPs are found in organisms ranging from yeast to human. Ring plots depicting representative clustered BP pairs in (A) yeast (*S. pombe* GO:0032956 (regulation of actin cytoskeleton organization) and GO:2001251 (negative regulation of chromosome organization), (B) worm (*C. elegans* GO:0007169 (transmembrane receptor protein tyrosine kinase signaling pathway) and GO:1901136 (carbohydrate derivative catabolic process)), (C) fly (*D. melanogaster* GO:0045747 (positive regulation of Notch signaling pathway)and GO:0046486 (glycerolipid metabolic process)), (D) mouse (*M. musculus* GO:0030308 (negative regulation of cell growth) and GO:0033135 (regulation of peptidyl-serine phosphorylation)) and (E) human (*H. sapiens* GO:0006959 (humoral immune response) and GO:0007162 (negative regulation of cell adhesion)). Ring segments with a number on the outside denote a different chromosome. Blue and green lines in outer and inner rings indicate the genomic location of each gene in their respective pathways. Red lines in central gap indicate locations of clustered gene pairs.

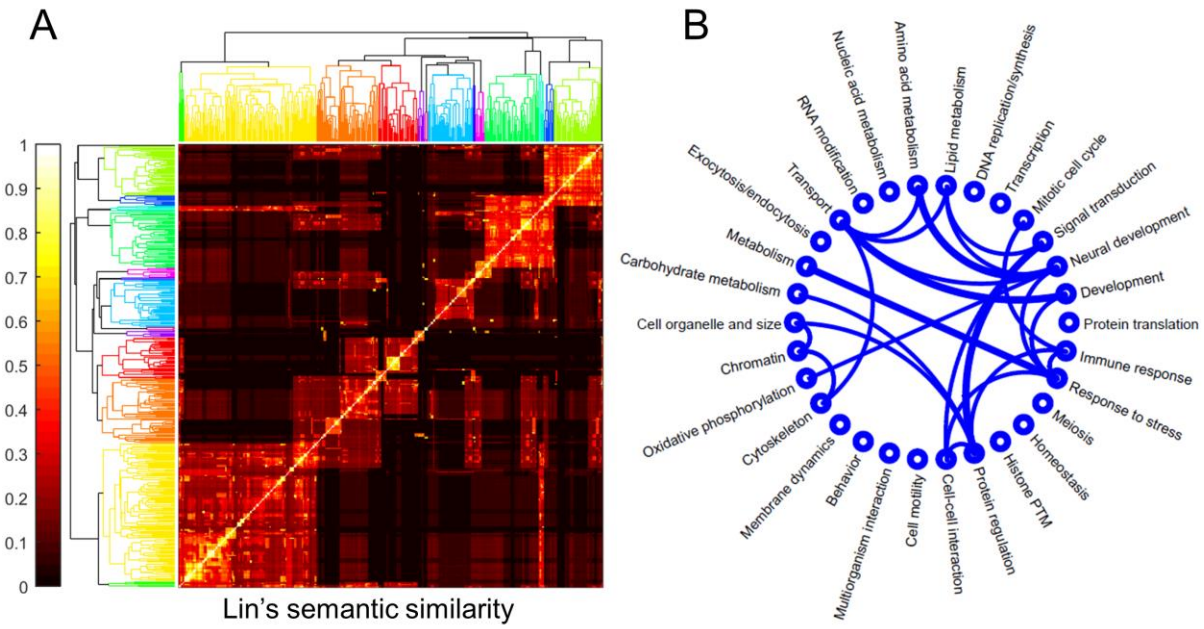


Fig. 3.

Several similar BP pairs show conserved clustering in eukaryotes. (A) Heatmap depicting hierarchical clustering of BP terms with clustered BP partners in at least two model organisms. Color is proportional to the magnitude of Lin's semantic similarity score (0-1.0). (B) Visual representation of clustering among different BP groups in yeast, worm, fly, mouse, and human. Lines connecting two BP groups indicate clustering in at least three genomes. Thickness of lines is proportional to the number of organisms in which pairwise clustering occurs.

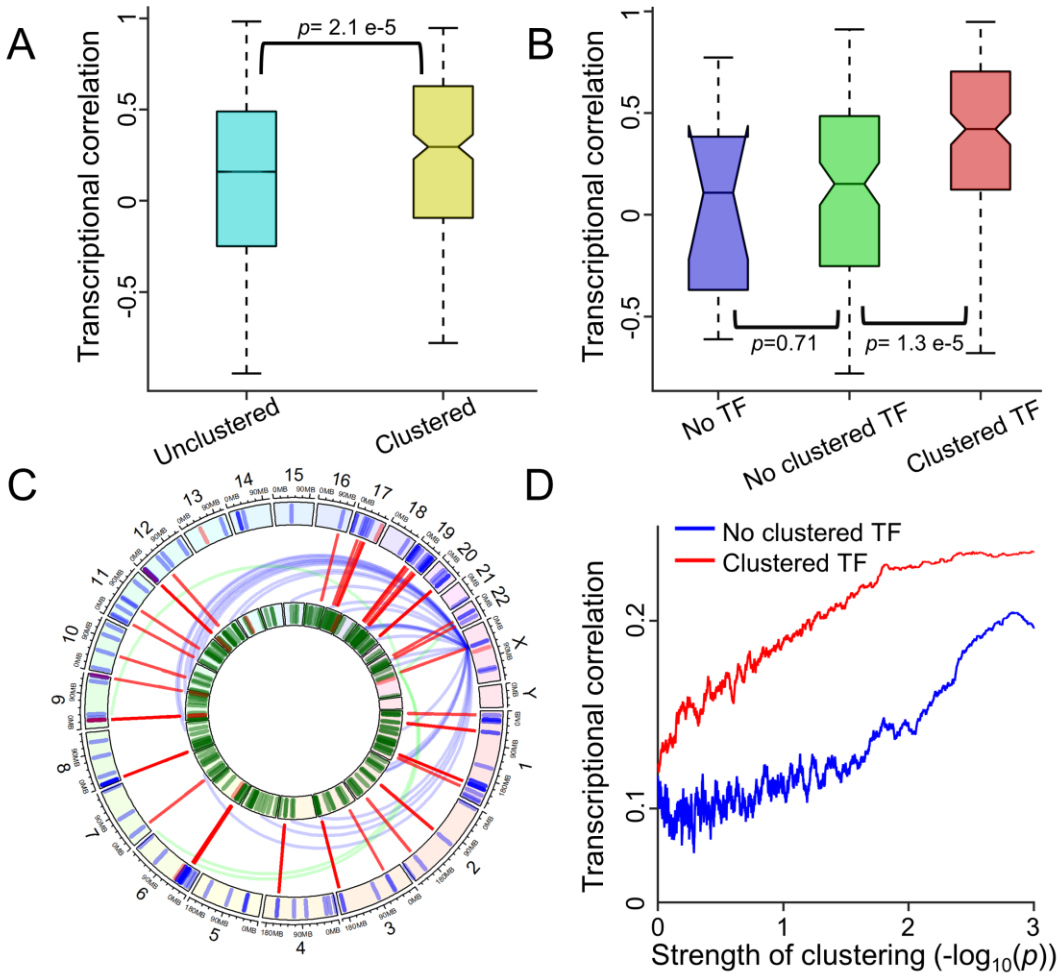
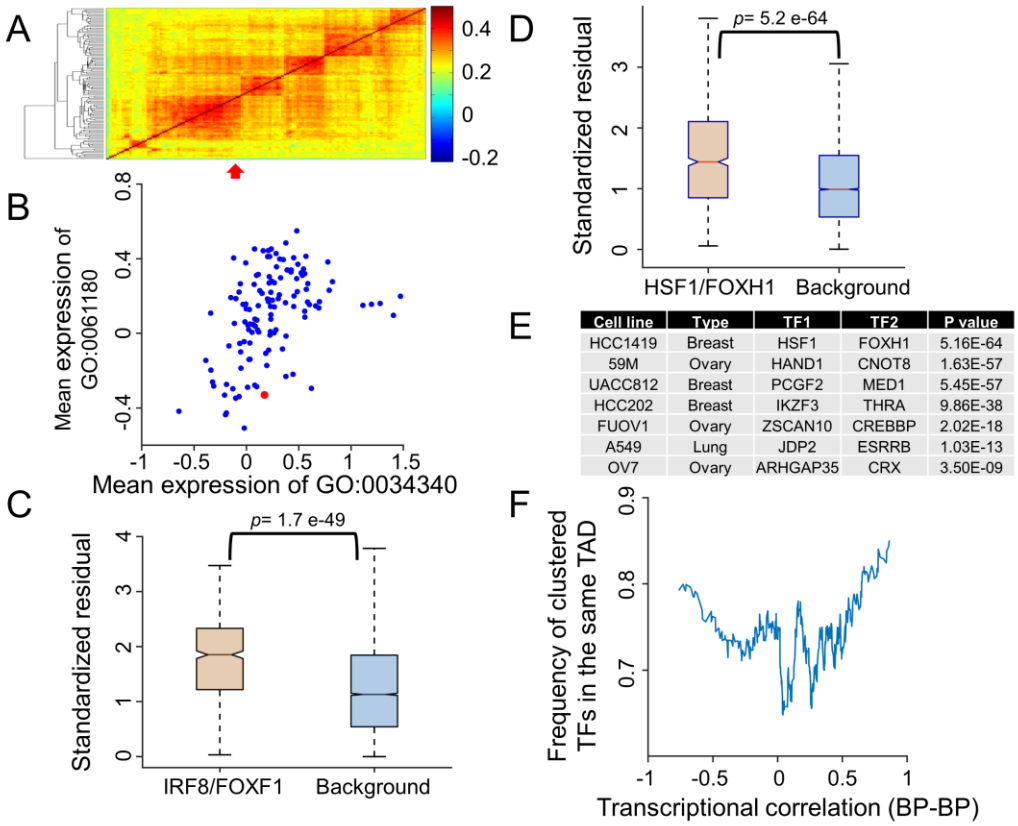


Fig. 4.

Disparate (Lin Score ≤0.1) clustered BP pairs, especially those with clustered TFs, are transcriptionally correlated in human. (A) Box plots depicting transcriptional correlation of clustered and unclustered BP pairs. (B) Box plots depicting transcriptional correlation of clustered BP pairs assigned to three groups based on the presence and clustering of TFs. 'No TF' refers to clustered BP pairs in which one or both of the two BP terms is missing a TF; 'No clustered TF' refers to clustered BP pairs in which both BP terms have a TF, but their TFs are not clustered; and 'Clustered TF' refers to clustered BP pairs in which both BP pairs have a TF and share at least one clustered TF pair. *p*-values were calculated by the two-sample t-test. (C) Ring plot of human genome, drawn to the same specification as Fig. 2E, depicting the transcriptional correlation of a representative clustered BP pair with clustered TFs. Green and blue curves depict high transcriptional correlation (>0.5) between a clustered TF and genes in the other BP. The two BP terms are GO:0006959 humoral immune response (blue) and GO:0007162 negative regulation of cell adhesion (green) in human. (D) Graph depicting the transcriptional correlation of all disparate human BP pairs as a function of the strength of their genomic clustering. Red and blue lines represent transcriptional correlation of BP pairs with or without clustered TFs, respectively. BP pairs were sorted by the strength of clustering. The correlation plotted represents the moving average of 5,000 points.



Cell line	Type	TF1	TF2	P value
HCC1419	Breast	HSF1	FOXH1	5.16E-64
59M	Ovary	HAND1	CNOT8	1.63E-57
UACC812	Breast	PCGF2	MED1	5.45E-57
HCC202	Breast	IKZF3	THRA	9.86E-38
FUOV1	Ovary	ZSCAN10	CREBBP	2.02E-18
A549	Lung	JDP2	ESRRB	1.03E-13
OV7	Ovary	ARHGAP35	CRX	3.50E-09

Fig. 5.

Genomic clustering of TFs can couple the transcription of associated BPs. (A) Hierarchical clustering of transcriptomes of EBV-transformed lymphocytic cancer cell lines (N=126). Red arrow indicates SUDHL8, the cell line with IRF8 deletion. The color indicates the overall pairwise correlation between two cancer cell lines. (B) Graph depicting the mean expression of two IRF8- and FOXF1-associated BPs (GO:0034340 response to type I interferon and GO:0061180 mammary gland epithelium development, respectively) among lymphocytic cell lines. Red dot represents SUDHL8. (C) Boxplot depicting the distribution of standardized residuals of BP pairs in SUDHL8 versus other lymphocytic cancer cell lines. The left and right bars depict this distribution for the IRF8/FOXF1-associated BP pairs and random BP pairs, respectively. (D) Boxplot depicting the distribution of the standardized residuals of BP pairs in HCC1419 cell line, carrying a translocation disrupting HSF1-FOXH1 genomic clustering, versus other breast cancer cell lines. The left and right bars depict this distribution for the HSF1/FOXH1-associated BP pairs and 10,000 random BP pairs, respectively. (E) Examples of cancer cell lines where the occurrence of a translocation, physically separating a clustered TF pair, co-occurs with greater standard residuals of the TF-associated BPs versus random BP pairs. Similar to C and D, standard residual of translocation-bearing cell lines was compared against other cancer cell lines from the same tissue. *p*-values are from two-sample *t*-tests. (F) Plot depicting the frequency that clustered TFs occur in the same TAD versus the transcriptional correlation of the associated BPs. For each clustered TF pair, the mean correlation of associated BP pairs was plotted. The frequency of TF occurring in the same TAD is the moving average of 50 points.

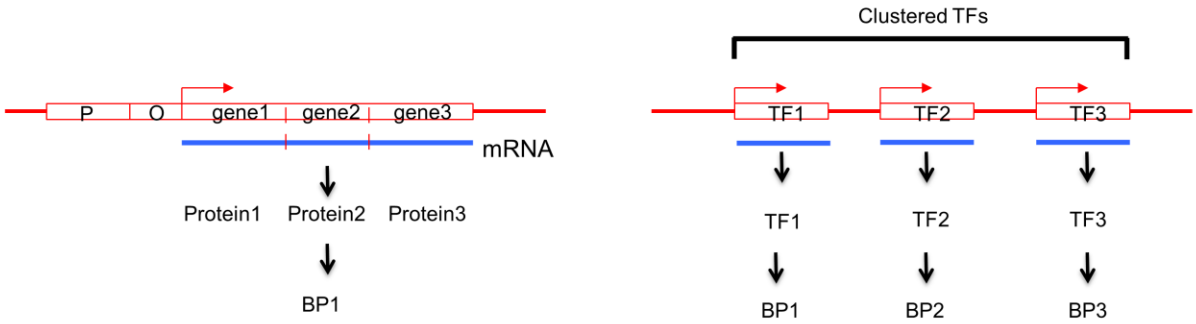


Fig. 6.
Model for how clustered TFs drive transcriptional coherence of disparate BPs in eukaryotes. Schematic model of prokaryotic operons (left) and eukaryotic modulons regulated by clustered TFs (right). According to our model, temporally ordered activation of clustered TFs can coordinate the expression of hundreds of genes belonging to disparate BPs dispersed throughout the genome.

Supplementary Figures

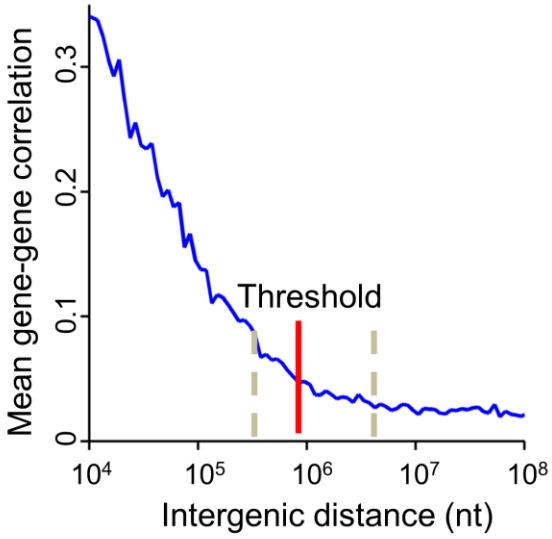


Fig. S1.
Transcriptional correlation of human gene pairs is a function of their intergenic distance. The graph depicts the transcriptional correlation of all possible human gene pairs as a function of the pairwise intergenic distance using the human GTEx data. 200 intergenic distance bins were created ranging from 20-230MB and the mean transcriptional correlation of all gene pairs within each bin was calculated and plotted (blue line). The red line represents the threshold distance (five times (5X) mean intergenic distance) chosen to define clustered gene pairs in the human genome in this paper (see Materials and Methods). Left and right dashed gray lines represent 2X and 20X mean intergenic distances, respectively, shown for comparison.

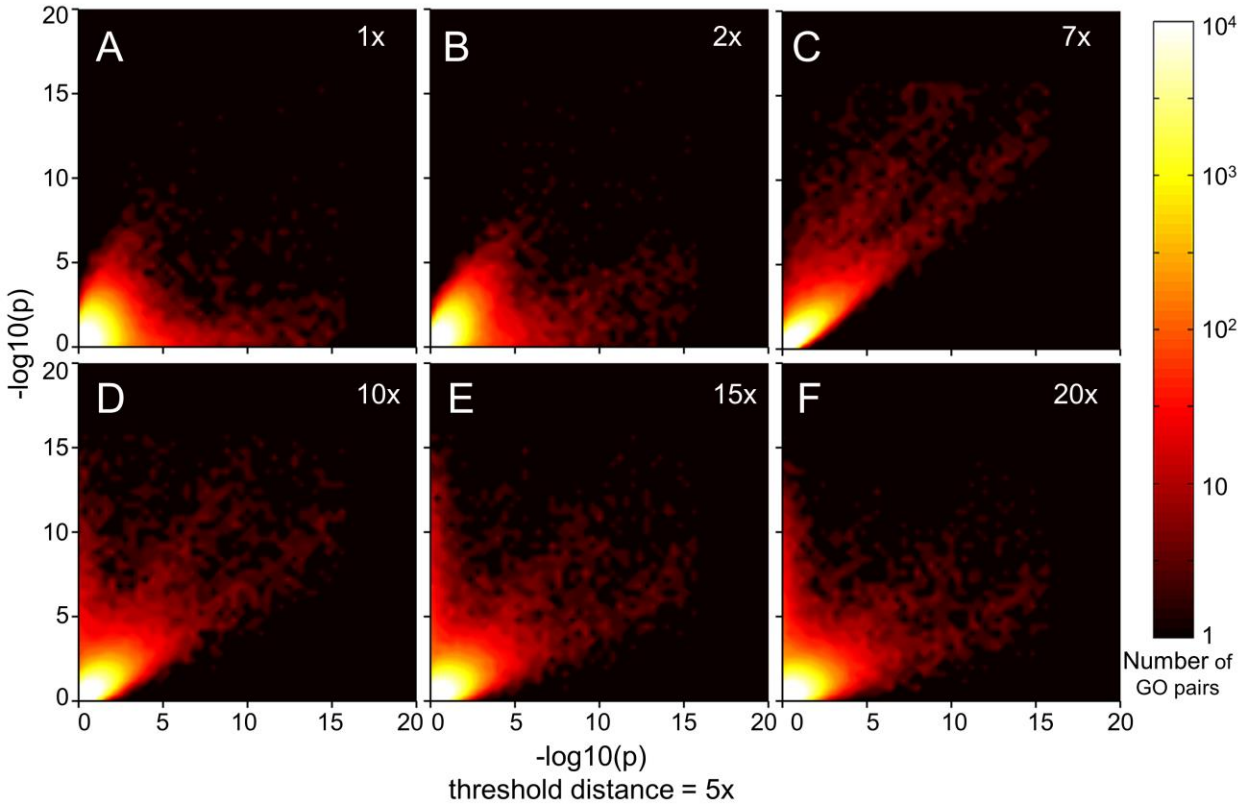


Fig. S2.

5X mean intergenic distance is robust for identifying clustered BP pairs in the human genome. Heatmaps depicting distribution of *p*-values (log10(p)) for identifying significantly clustered BP pairs calculated for five times (5X) mean intergenic distance versus the *p*-values calculated for the indicated threshold distances (1X, 2X, 7X, 10X, 15X, 20X mean intergenic distance). Color is proportional to the number of BP pairs in each bin. The data for other model organisms are available at https://figshare.com/projects/Transcriptional_coherence/72644.

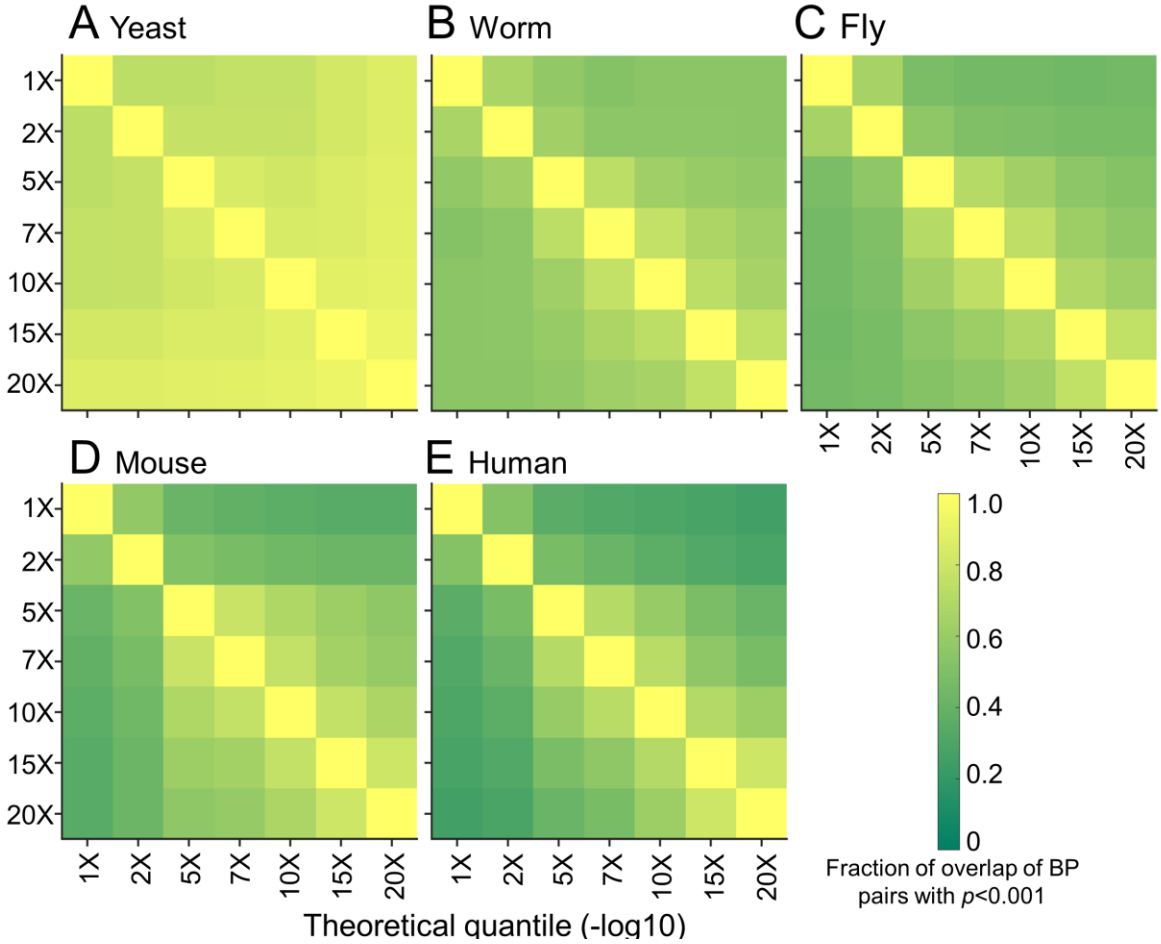


Fig. S3.
5X threshold distance captures a cross section of significantly clustered GO pairs in all organisms. Heat map depicting the fraction of overlap of significantly clustered BP pairs with *p* values less than 0.001 for the indicated threshold distances (1X, 2X, 5X, 7X, 10X, 15X, 20X mean intergenic distance) in each organism. Color is proportional to the fraction of overlap of significantly clustered BP pairs at each threshold distance.

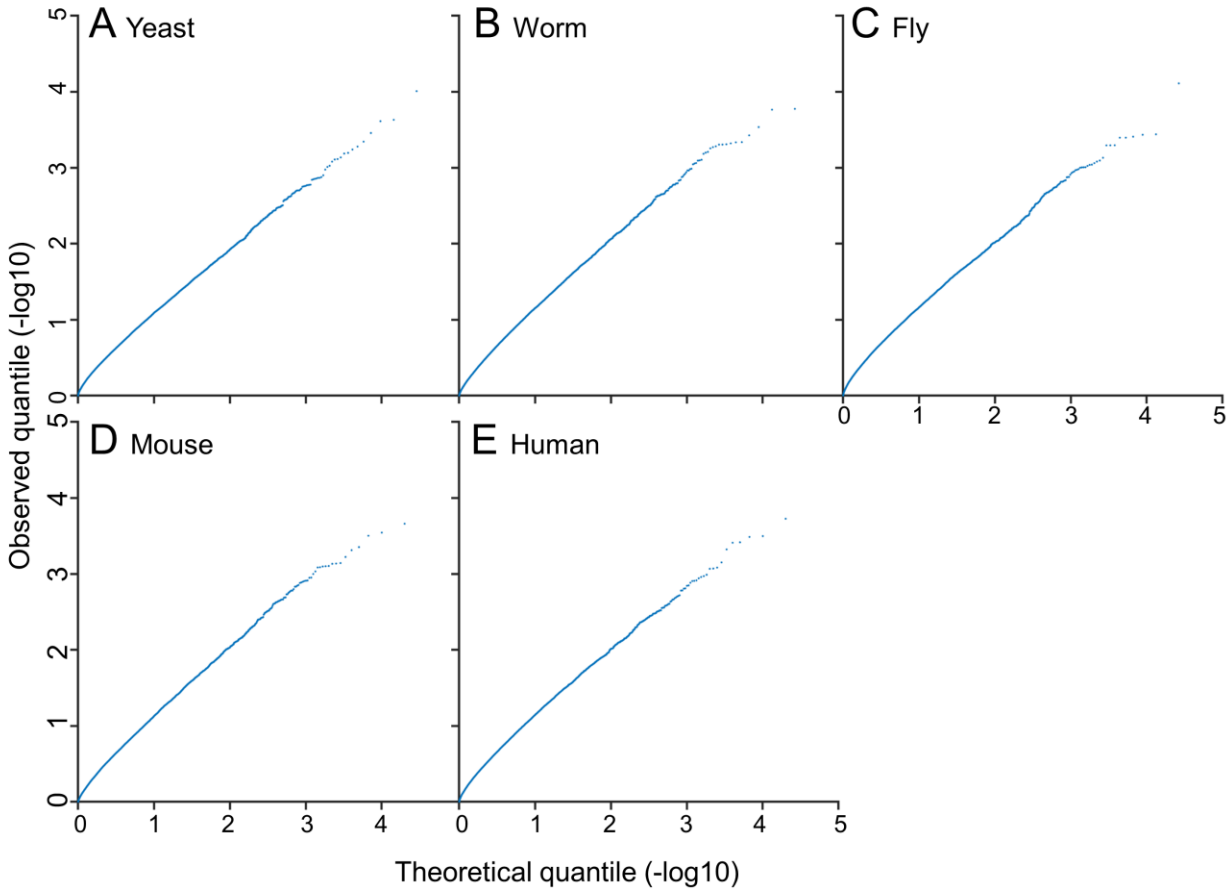


Fig. S4.
Randomly selected gene sets do not display significant genome clustering. Quantile-quantile (QQ) plots of observed (Y-axis) versus theoretical (X-axis) *p*-value distributions for genome clustering between gene sets generated by randomly selecting genes from the pool of BP-associated genes in each organism (see Materials and Methods). The QQ plots in (A) yeast (*S. pombe*), (B) worm (*C. elegans*), (C) fly (*D. melanogaster*), (D) mouse (*M. musculus*) and (E) human (*H. sapiens*) are depicted. In contrast to naturally occurring BPs (Fig. 1D to H), these QQ plots do not curl up, thus fail to exhibit significant clustering. These analyses reveal that our statistical method specifically identifies significantly clustered BP pairs in each genome.

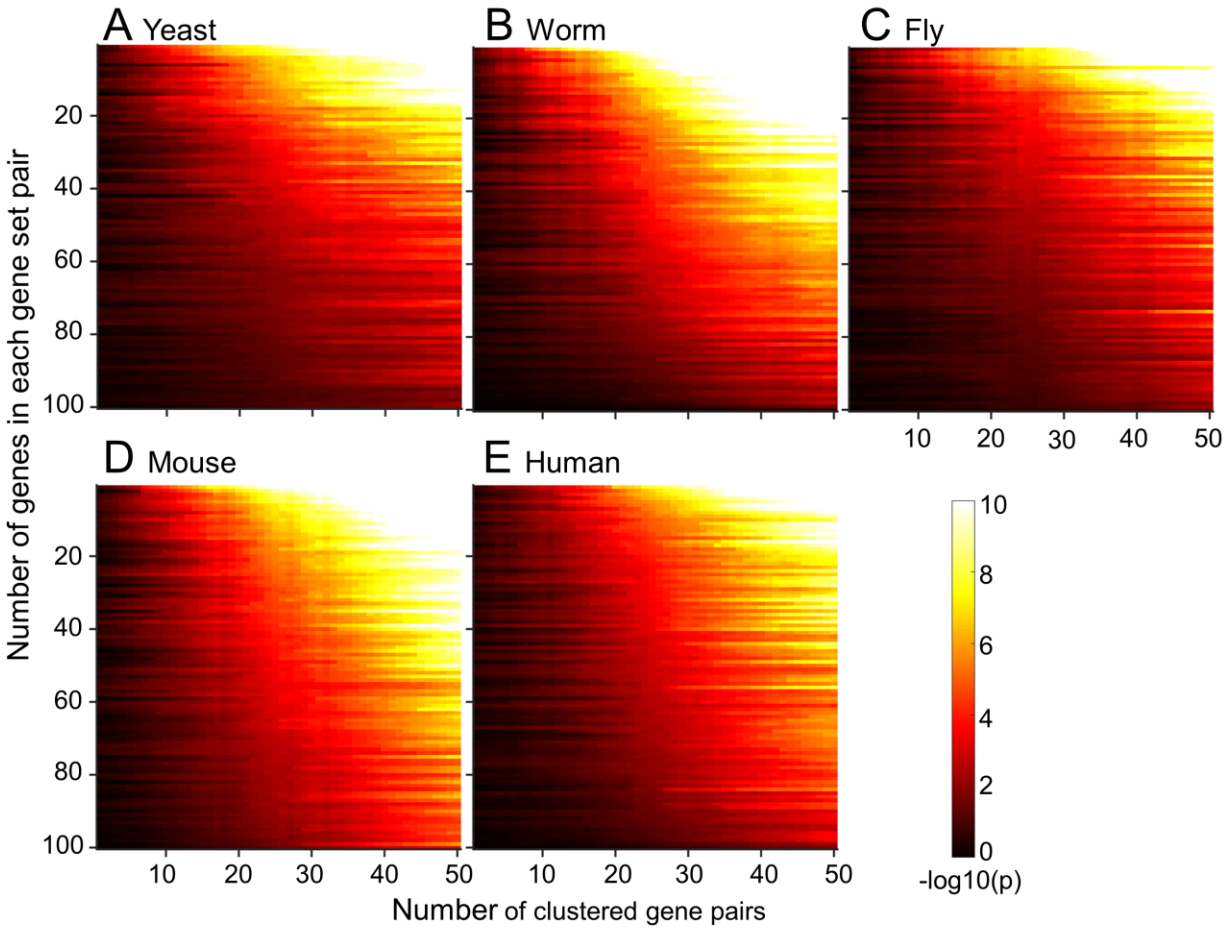


Fig. S5.
The statistical method developed for genome clustering performs as expected. Heat maps depicting the impact on *p*-value as clustered gene pairs are artificially added to randomly generated gene sets. For each indicated organism, *p*-values for clustering of 100 artificially generated pairs of gene sets were calculated. As expected, the artificial addition of clustered gene pairs decreases the *p*-value for clustering in all model organisms. Color is proportional to -log10(*p*).

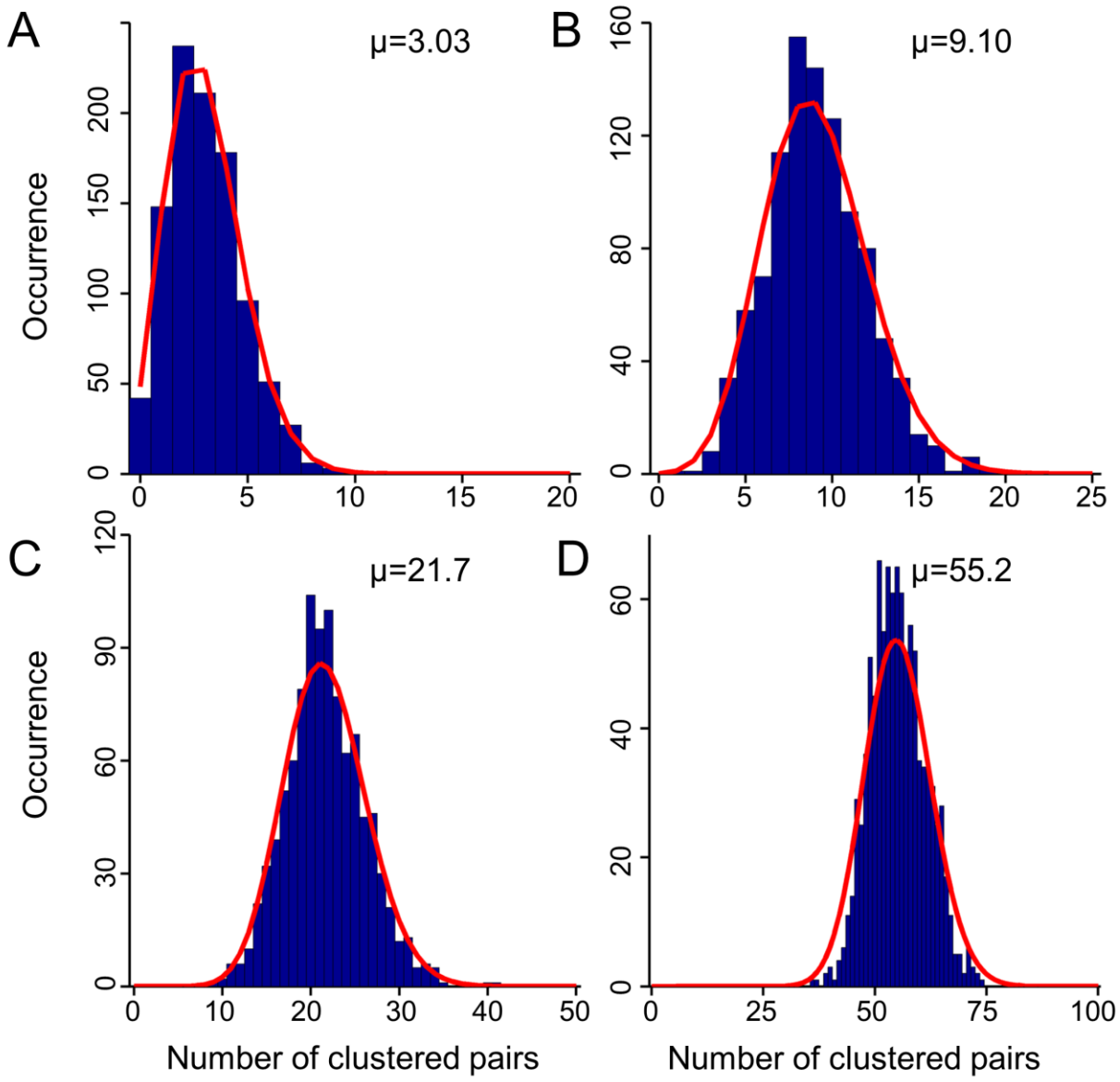


Fig. S6.

Background clustering was quantified by random gene sampling. Graphs depicting the distribution of background gene clustering occurrences (bar graphs) and their Poisson fit (red lines) for four representative sets of BP-BP analyses performed on the human genome. 2,000 random samplings (Materials and Methods) were done to quantify background gene clustering for each BP pair analysis in each organism. μ is the mean number of clustered gene pairs derived from the fitted Poisson distribution.

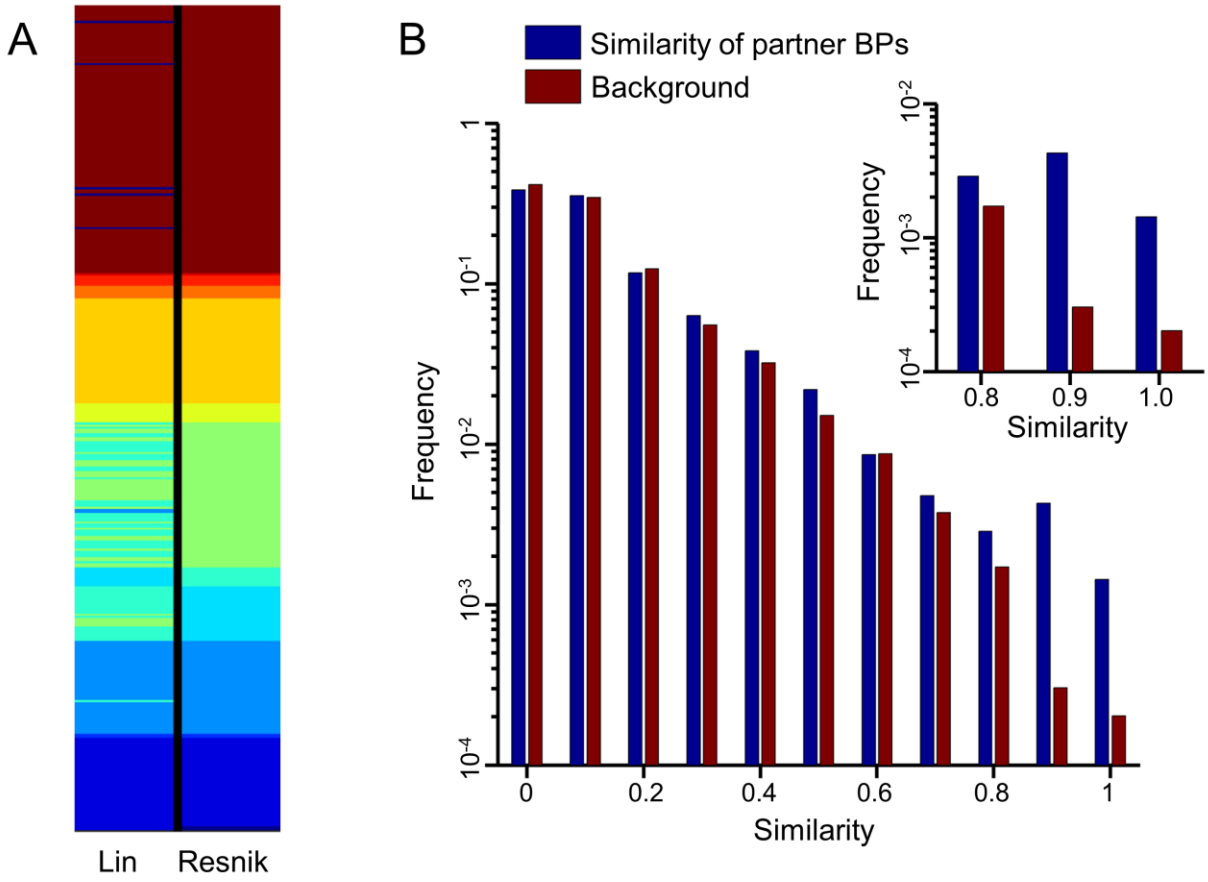


Fig. S7.
Highly similar BP pairs are clustered in eukaryotes. (A) Graph comparing the BP groups generated by hierarchical clustering using Lin's and Resnik's semantic similarity methods. The BPs used in this analysis all have a clustered BP partner in at least two model organisms. (B) Graph depicting the frequency of semantic similarity score of BP terms which share a common clustered partner in at least two organisms. Using the example from the main text, similarity of partner BPs calculates the similarity between BP2 and BP2' which are clustered partners of BP1. Semantic similarity scores of partner BPs were calculated using the Lin method.

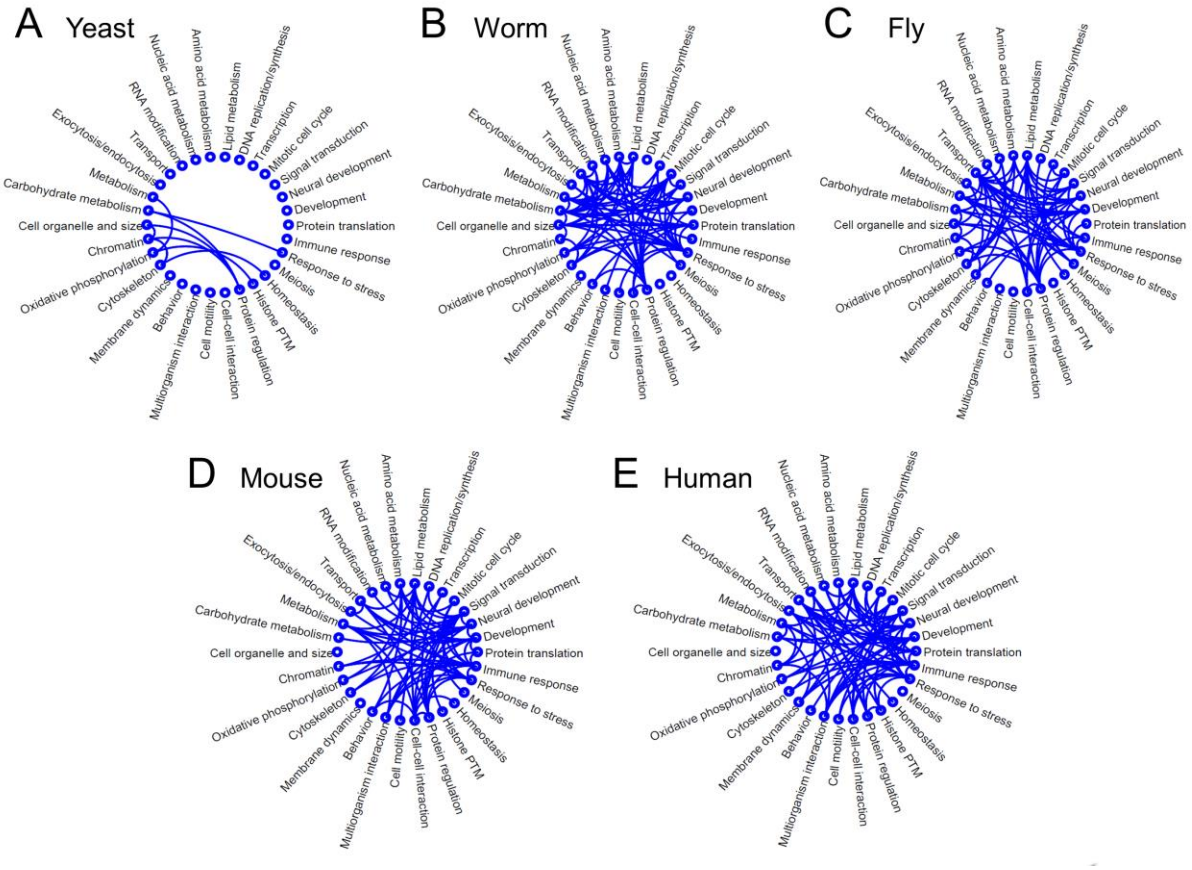


Fig. S8.

Several BP groups cluster in multiple eukaryotic genomes. The lines connect two BP groups which have at least one clustered BP pair between them in the indicated organisms.

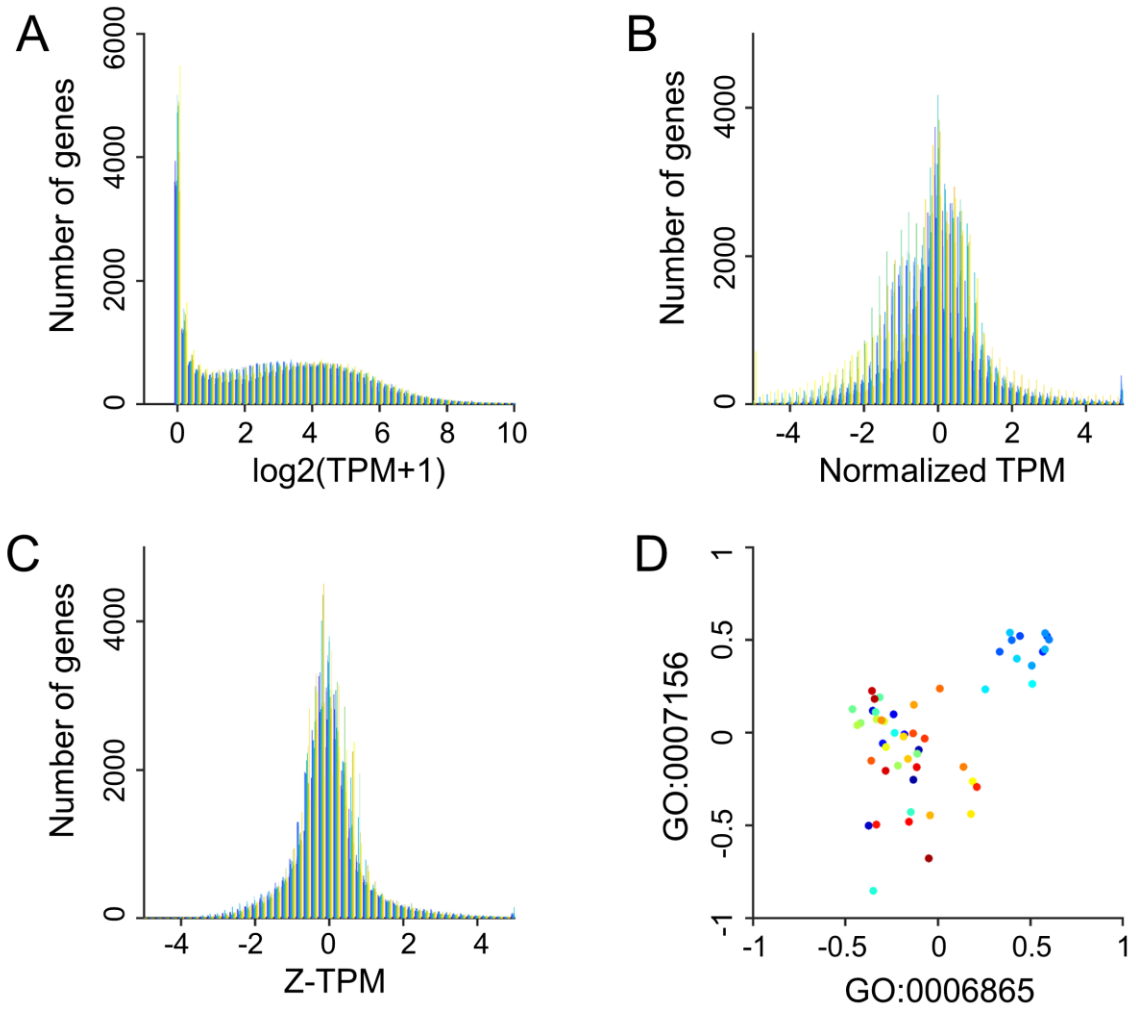


Fig. S9.

GTEx transcriptome data were normalized to quantify transcriptional correlation between BPs. Graphs depict the (A) raw Transcript Per Million (TPM) data (log2(TPM+1)), (B) normalized TPM and (C) Z-score of TPM. Each color corresponds to a different tissue. The Z-scores were used for quantitative comparisons across different tissues. (D) An example of BP-BP transcriptional correlation in human (GO:0006959 (humoral Immune response) and GO:0007162 (negative regulation of cell adhesion)). Each dot represents a different tissue in GTEx.

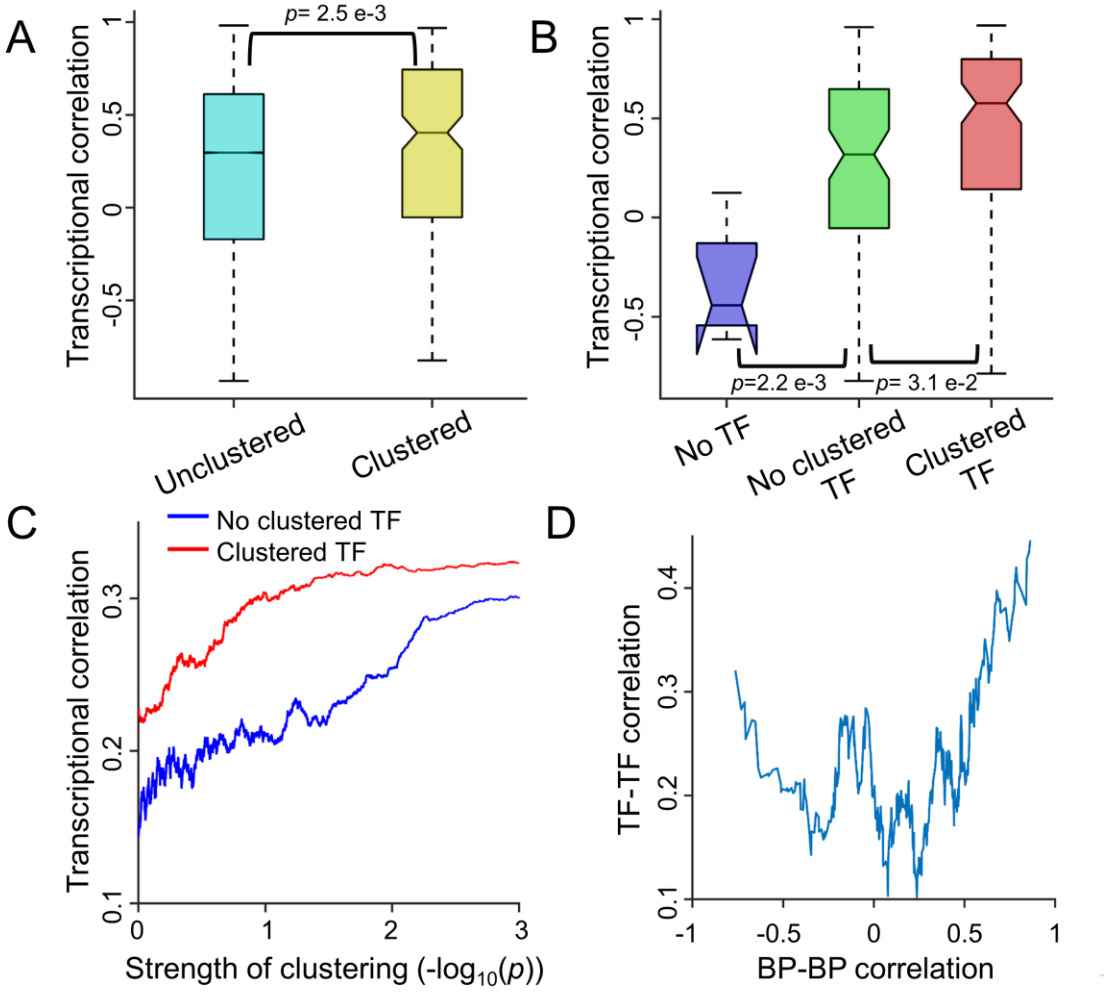


Fig. S10.

Similar (Lin score > 0.1) clustered BP pairs, especially those with clustered transcription factors (TFs), are transcriptionally correlated in human cells. (A) Box plots depicting transcriptional correlation of similar clustered (N=188) and unclustered (N=186,330) BP pairs. (B) Box plots depicting transcriptional correlation of similar clustered BP pairs assigned to three groups based on the presence and clustering of TFs. 'No TF' refers to clustered BP pairs in which one or both of the two BP terms is missing a TF (N=7); 'No clustered TF' refers to clustered BP pairs in which both BP terms have a TF, but their TFs are not clustered (N=78); and 'Clustered TF' refers to clustered BP pairs in which both BP pairs have a TF and those TFs are clustered (N=103). *p*-values were calculated by the two-sample t-test. (C) Graph depicting the transcriptional correlation of all similar (Lin score > 0.1) human BP pairs as a function of the strength of their genomic clustering. Red and blue lines represent transcriptional correlation of BP pairs with (N=27,898) or without (N=158,016) clustered TFs, respectively. Each BP pair was sorted by the strength of clustering and the correlation represents the moving average of 5,000 points. (D) Graph depicting TF-TF transcriptional correlation as a function of the transcriptional correlation of the associated BP pairs. TF-TF pairs were sorted by BP-BP correlation, and the moving averages of 50 points are shown.

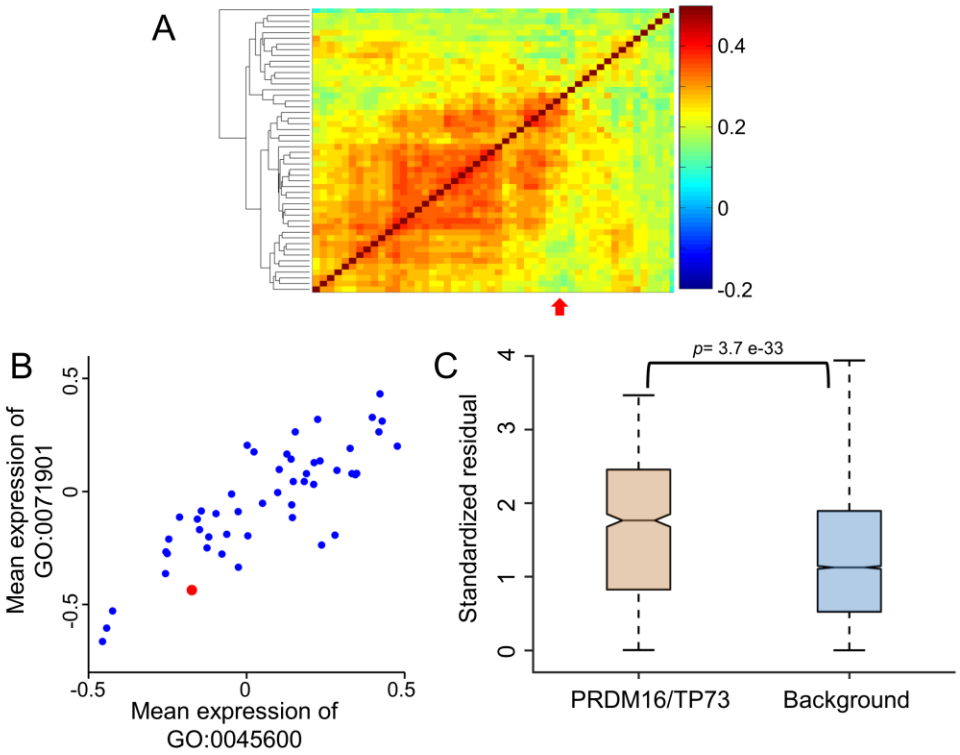


Fig. S11.

Transcription factor (TF) clustering drives the coregulation of discrete biological processes.

(A) Hierarchical clustering of transcriptomes of transverse colon cancer cell lines (N=50). Red arrow indicates SNU1033, the cell line carrying a deletion in PRDM16. The color indicates to the overall pairwise correlation between two cancer cell lines. (B) Graph depicting the mean expression of two PRDM16- and TP73-associated BPs (GO:0045600 positive regulation of fat cell differentiation and GO:0071901 negative regulation of protein serine/threonine kinase activity, respectively) among transverse colon cancer cell lines. Red dot represents SNU1033. (C) Boxplot depicting the distribution of standardized residuals of BP pairs in SNU1033 versus other transverse colon cancer cell lines. The left and right bars depict this distribution for the PRDM16- and TP73-associated BP pairs and random BP pairs, respectively.

A

Cell lines in which TADs were identified
A549
AdrenalGland_Donor-AD2
Aorta_STL002
Bladder_Donor-BL1
Bowel_Small_Donor-SB2
Caki2
Cortex_DLPFC_Donor-CO
G401
GM12878
H1-ESC
H1-MES
H1-MSC
H1-NPC
H1-TRO
HMEC
HUVEC
IMR90
K562
KBM7
LNCaP
Liver_STL011
Lung_Donor-LG1
Muscle_Psoas_Donor-PO1
NCIH460
NHEK
PANC1
Pancreas_Donor-PA2
RPMI7951
SJCRH30
SKMEL5
SKNDZ
SKNMC
Spleen_Donor-PX1
T470
Thymus_STL001
VentricleLeft_STL003
VentricleRight_Donor-RV3

B

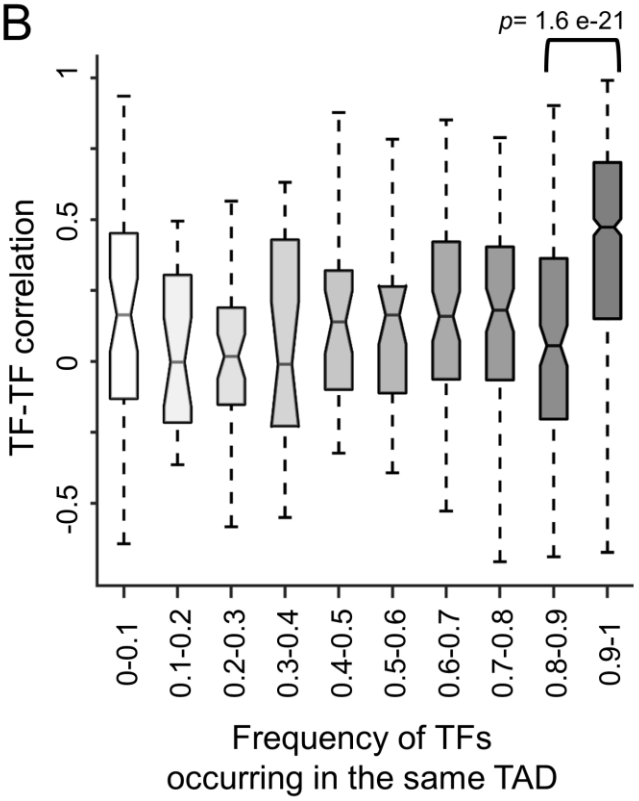


Fig. S12.
Clustered TFs found frequently in the same TAD display high TF-TF transcriptional correlation. (A) The list of 37 cell lines in which TADs were identified. This was used to calculate the frequency by which a clustered TF pairs falls in the same TAD based on these studies. (B) Graph depicting TF-TF correlation and the frequency which clustered TFs are found in the same TAD. *p*-values were calculated by the two-sample t-test.

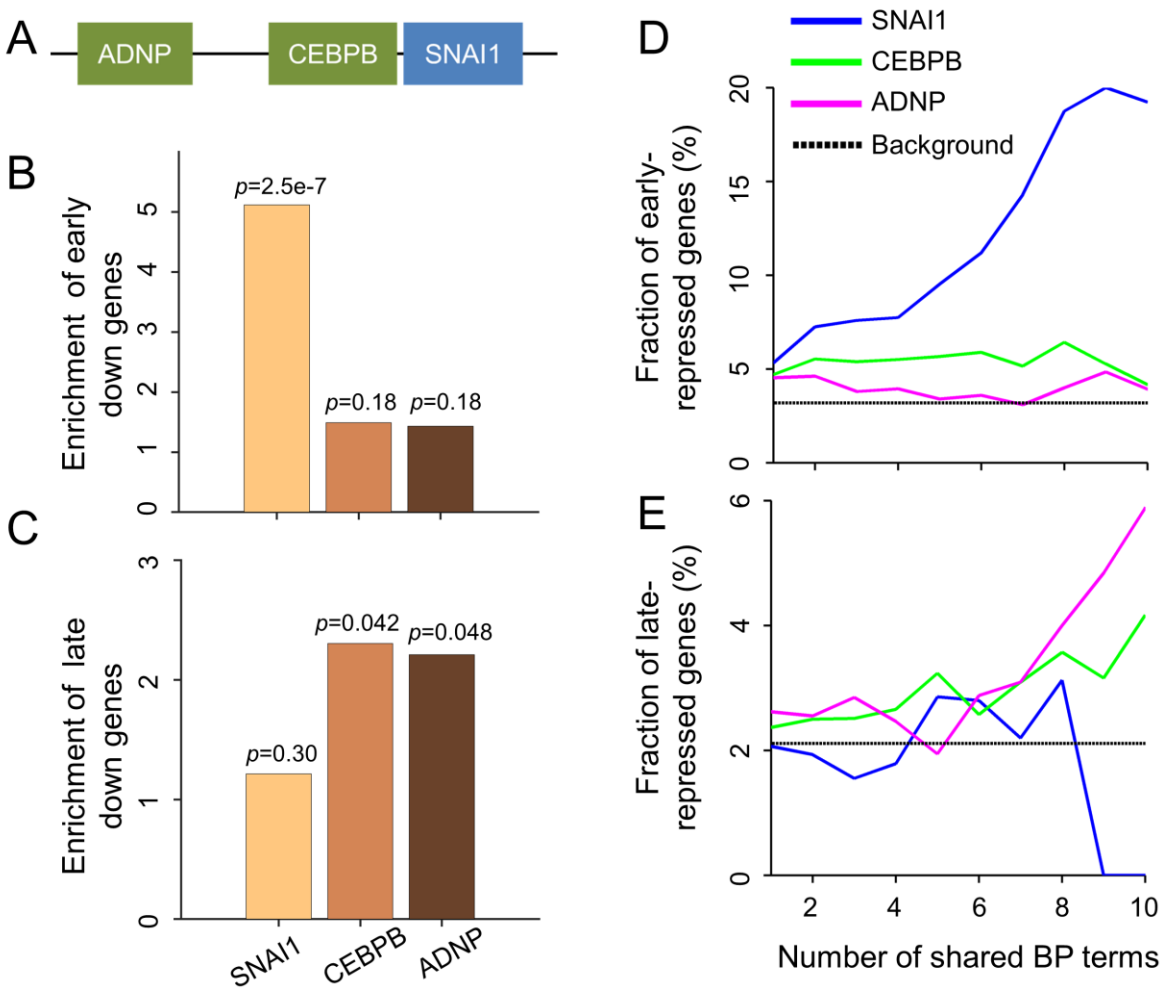


Fig. S13.

SNAI1 overexpression drives the temporal coregulation of ADNP- and CEBP-associated BPs. (A) Schematic representation of the ADNP-CEPBP-SNAI1 clustered TFs in the human genome. (B-C) Enrichment of genes regulated by the indicated TFs that are also found in the (E) early- and (F) late-repressed genes over background. Background is the expected number of indicated genes by random gene sampling. The early and late repressed gene sets were identified in a time-course assay after the inducible overexpression of SNAI1 in MCF10A human breast epithelial cells (Javaid et al 2013) [1]. *p*-values are from one-tailed binomial test. Fractions of SNAI1-, CEBP-, and ADNP-associated genes found in the (D) early- and (E) late-repressed transcriptomes after the inducible overexpression of SNAI1 in MCF10A epithelial human breast cells. The x-axis is the number of times that a given gene shares the same BP term with SNAI1, CEBPB and ADNP. This measure was used as a surrogate for core genes regulated by these TFs.