JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009
© Mary Ann Liebert, Inc.
Pp. 303–316
DOI: 10.1089/cmb.2008.15TT

Higher-Order Genomic Organization of Cellular Functions in Yeast

TAMIR TULLER,[1,2] UDI RUBINSTEIN,[3] DANI BAR,[3] MICHAEL GUREVITCH,[3]
EYTAN RUPPIN,[1,4] and MARTIN KUPIEC[2]

ABSTRACT

Previous studies have shown that the distribution of genes in prokaryotes and eukaryotic genomes is not random. Using the thousands of cellular functions that appear in the Gene Ontology (GO) project, we exhaustively studied the relation between functionality and genomic localization of genes across 16 organisms with rich GO ontologies (one prokaryote and 15 eukaryotes). Overall, we found that the genomic distribution of cellular functions tends to be more similar in organisms that have higher evolutionary proximity. At the primary level, which measures localization of functionally related genes, the prokaryote *Escherichia coli* exhibits the highest level of organization, as one would expect given its operon-based genomic organization. However, examining a higher level of genomic organization by analyzing the co-localization of pairs of different functional gene groups, we surprisingly find that the eukaryote yeast *Saccharomyces cerevisiae* is markedly more organized than *E. coli*. A network-based analysis further supports this notion and suggests that the eukaryotic genomic architecture is more organized than previously thought. See online Supplementary Material at *www.liebertonline.com*.

Key words: functional organization and co-organization, genomic organization, GO ontologies.

1. INTRODUCTION

In order to function properly, cells must coordinate the expression of a large number of genes. The mechanisms by which cells achieve this feat constitute one of the most intensely studied subjects in modern biology. In recent years, with the acquisition of increasing numbers of genome sequences, it has become possible to compare the distribution of genes in the different genomes, showing that gene location along the genome is not random. In the case of the prokaryotes, it is well known that genes are organized in operons (Miller, 1981). Each operon consists of genes that are located adjacently on the DNA, are expressed in concert, and usually work together (e.g., are part of the same metabolic pathway and/or their products physically interact [Enright et al., 1999; Marcotte et al., 1999]). In contrast, most eukaryotic genomes lack operons and for many years the gene order along their chromosomes was

[1] School of Computer Sciences, Tel Aviv University, Ramat Aviv, Israel.
[2] Department of Molecular Microbiology and Biotechnology, Sheba Medical Center, Tel-Hashomer, Israel.
[3] Multiple Sclerosis Center, Sheba Medical Center, Tel-Hashomer, Israel.
[4] School of Medicine, Tel Aviv University, Ramat Aviv, Israel.

assumed to be largely random. However, this notion has gradually been dispelled in recent years (Eichler and Sankoff, 2003; Hurst et al., 2004; Kosak and Groudine, 2004).

Many studies have investigated different variables that may correlate with genomic organization (i.e., localization across the chromosomes) in the eukaryotic genome. Most of these papers dealt with the connection between co-expression and chromosomal proximity, showing that closely located genes tend to be co-expressed (Cohen et al., 2000; Sémon and Duret, 2006), that clusters of co-expressed genes in mammalian genomes are evolutionarily conserved (Sémon and Duret, 2006; Singer et al., 2005), and that highly expressed genes and housekeeping genes tend to cluster (Caron et al., 2001; Lercher et al., 2002, 2003; Versteeg et al., 2003). This phenomenon can be explained by the way DNA is organized within the nucleus (e.g., chromatin structure and folding), and the sharing of regulatory elements (Bártová and Kozubek, 2006; Branco and Pombo, 2006; Cremer et al., 2006; Sproul et al., 2005). Additional correlations were found between gene organization and other variables such as proximity to replication origins (Huvet et al., 2007) and noise levels during gene expression (Batada and Hurst, 2007).

A number of previous studies explored the genomic distribution of genes belonging to the same biological function or biochemical pathway. It was shown that clustered genes tend to exhibit similar functionality (Hurst et al., 2004; Kosak and Groudine, 2004; Miller et al., 2004; Petkov et al., 2007; Yi et al., 2007), tend to be located in domains with low recombination rate (Pal and Hurst, 2003), encode proteins that tend to interact physically (Poyatos and Hurst, 2006, 2007; Teichmann and Veitia, 2004), and belong to the same metabolic pathway (Lee and Sonnhammer, 2003; Sproul et al., 2005; Wong and Wolfe, 2005). Lee and Sonnhammer (2003) studied the organization level of several metabolic pathways in five eukaryotes. They showed that the degree of gene organization in many pathways exhibits relatively large variation among the studied organisms. Using Linkage Disequilibrium (LD), Petkov et al. (2007) demonstrated that more than 25% of the mice genome contains clusters of functionally related genes. Yi et al. (2007) developed a sophisticated algorithm for finding functionally related genomic clusters. They applied their algorithm to eight organisms and showed that there are species-specific variations in term of the size and functional annotations of gene clusters.

Extending upon this previous work, the aim of the current study is to perform a more comprehensive study of the genomic organization of cellular functions in numerous organisms, focusing on some new research questions. To this end we rely on the functional classification of genes carried out by the Gene Ontology (GO) project (Ashburner et al., 2000). Accordingly, the 16 model organisms studied in this work are the ones with the most abundant GO annotation. As the main focus of this work is the eukaryotes, 15 of the chosen organisms are eukaryotes; one prokaryote (*Escherichia coli*) with abundant and reliable GO annotations, serves as a prokaryotic comparison rod. We define three measures of functional localization that allow us to estimate the organization (or localization, the tendency of genes to appear in proximity to each other) of functional gene groups in the genome, and across chromosomes, and examine the relations between this functional organization and many other variables of interest. We generalize the localization measures and study the co-localization of pairs of GO functions, a new concept that demonstrates a surprisingly higher degree of "second-level" order in a prototypical eukaryote (*Saccharomyces cerevisiae*) compared to a prototypical prokaryote (*E. coli*). This co-localization measure enables us to detect "communities" of co-localized sets of GO functions and uncover a higher-level organization of the eukaryotic genome.

2. RESULTS

2.1. Measures of genomic organization

Throughout this paper, we denominate all the genes tagged with a certain GO-function term a "GO group." We designed three basic and simple measures for estimating the organization level of a GO function in a genome. The first two measures estimate the intra-chromosomal localization of a GO group by comparing the number of gene clusters that map close to each other, to the expected number of clusters in a random background model (Supplementary Fig. 1). (See online Supplementary Material at *www.liebertonline.com*.) In the first measure, *LocN* (Supplementary Fig. 1A), a set of adjacent genes with the same GO group annotation is defined as a cluster. In the second measure, *LocD*, a cluster is a set of genes such that the distance of each one to the closest gene belonging to the same GO group is less than a certain threshold. That is, in contrast to the first measure, the second measure explicitly considers the physical distance (in base pairs) on the chromosome. The third measure, *LocC*, estimates the inter-chromosomal localization of GOs, that is, the tendency of a GO group to reside in fewer chromosomes than expected from a random distribution model. The three measures of localization significantly correlate with each other (e.g., in *S. cerevisiae*, for the Biological Process ontology the Spearman correlation between *LocN* and *LocD* is 0.9259 and between *LocN* and *LocC* is 0.9259, all p-values < 0.001). Therefore, throughout the paper we focus for clarity on the *LocN* measure in the main text, and provide the (qualitatively similar) results for the other measures in the supplementary material.

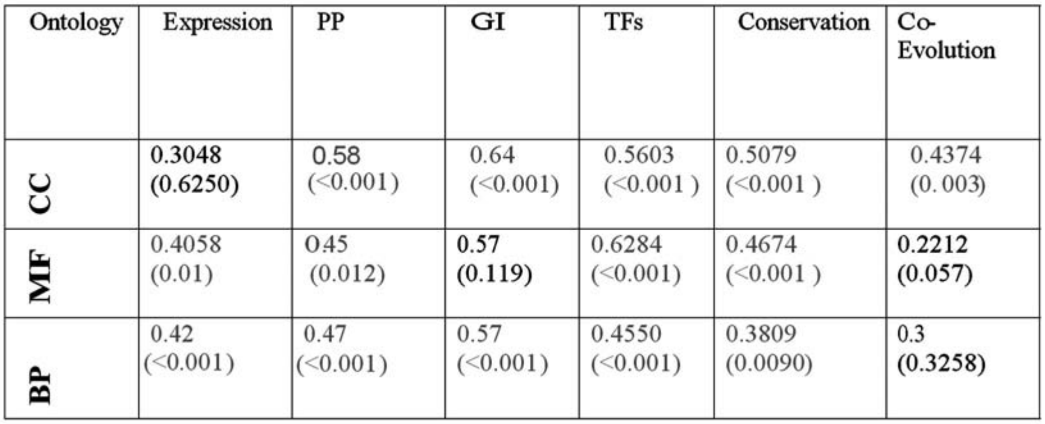
By comparing the cluster distribution of GOs to that expected in a random background model, these measures can detect both weak signals of co-organization (such as the sharing of a common promoter by gene pairs) and stronger signals (such as large operons). The p-value obtained denotes the magnitude of the localization of GO groups of the same order of magnitude, thus allowing a comparison of the organization of functional groups in different organisms in a quantitative manner.

2.2. Functional gene organization in the yeast S. cerevisiae

As a first step, we analyzed the functional genomic architecture of a well-characterized eukaryotic model organism, the yeast *S. cerevisiae*. Supplementary Table 1 includes the localization score of the GO functions that have more than five genes in *S. cerevisiae*, together with their correlations with an array of other variables that we examined, including their enrichment with protein interactions, genetic interactions, the tendency to share common transcription factors (TFs), and their expression coherency. (See online Supplementary Material at *www.liebertonline.com*.) A summary of the correlations is presented in Figure 1A; similar results for other localization measures appear in Supplementary Table 2. (See online Supplementary Material at *www.liebertonline.com*.) Our results are in line with previous findings of a correlation between functional localization and co-expression (Cohen et al., 2000; Sémon and Duret, 2006), as well as physical interactions (Poyatos and Hurst, 2006, 2007; Teichmann and Veitia, 2004), showing a similar trend for all three GO ontologies (Biological Process, Cellular Component, and Molecular Function). We also investigated two additional new relationships, measuring the correlation between gene localization and evolutionary conservation, and its correlation with gene co-evolution. Interestingly, functional groups that are more localized in *S. cerevisiae* tend to be more conserved (see Fig. 1A). We therefore asked whether genomic localization of a GO group correlates with the tendency of its members to co-evolve. Surprisingly, this correlation is significant only for the Cellular Component ontology. This fact can be explained by assuming that the localization of genes grouped under this ontology reflects more stringent physical interactions and thus corresponds to higher levels of co-evolution, whereas co-membership in the other ontologies, although conserved, do not entail a coordinated evolution.

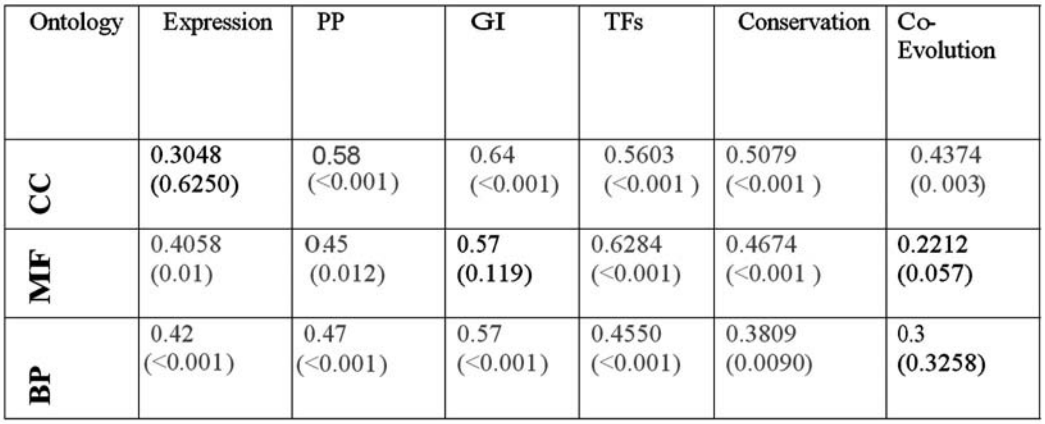
The relation between genome organization and expression coherency is continuously monotonous (not dichotomic; Fig. 1B): in average, an increase in the localization score corresponds to an increase in the co-expression coherency score. Roughly speaking, on average, as the distances between members of a GO group increase, their expression coherency decreases. This supports the notion that the need for coordinated expression is an important evolutionary force behind functional gene organization.

When the three new measures of localization are applied to the GO ontology of *S. cerevisiae*, several biological processes exhibit a high degree of gene clustering (Supplementary Table 1). (See online Supplementary Material at *www.liebertonline.com*.) These include housekeeping functions (e.g., chromatin remodeling, RNA splicing, and processing), as well as cellular processes (cell cycle progression and meiosis) and response to external stress, such as arsenic, copper ion, or osmotic stress. Interestingly, our results also show that genes involved in the metabolism of a small number of compounds (e.g., sulfur, phosphate) are also highly clustered in the yeast genome. A close examination of these GO groups reveals that in many cases they are enriched for pairs of genes that share regulatory regions. For example, the GO group "*chromatin assembly or disassembly*" is composed of 20 genes; out of these, four pairs of genes have common promoters. The GO group "*pyridoxine metabolic process*" includes seven genes; six of these appear as pairs of divergent genes sharing a regulatory sequence. Finally, the GO group "*hexose transport*" (18 genes) contains two clusters of three genes and a pair of genes sharing a common promoter (for a list of clusters in each of the studied organisms, see Supplementary Tables 4 and 5). (See online Supplementary Material at *www.liebertonline.com*.)



Ontology	Expression	PP	GI	TFs	Conservation	Co-Evolution
CC	**0.3048 (0.6250)**	0.58 (<0.001)	0.64 (<0.001)	0.5603 (<0.001)	0.5079 (<0.001)	0.4374 (0. 003)
MF	0.4058 (0.01)	0.45 (0.012)	**0.57 (0.119)**	0.6284 (<0.001)	0.4674 (<0.001)	**0.2212 (0.057)**
BP	0.42 (<0.001)	0.47 (<0.001)	0.57 (<0.001)	0.4550 (<0.001)	0.3809 (0.0090)	0.3 **(0.3258)**

(A)



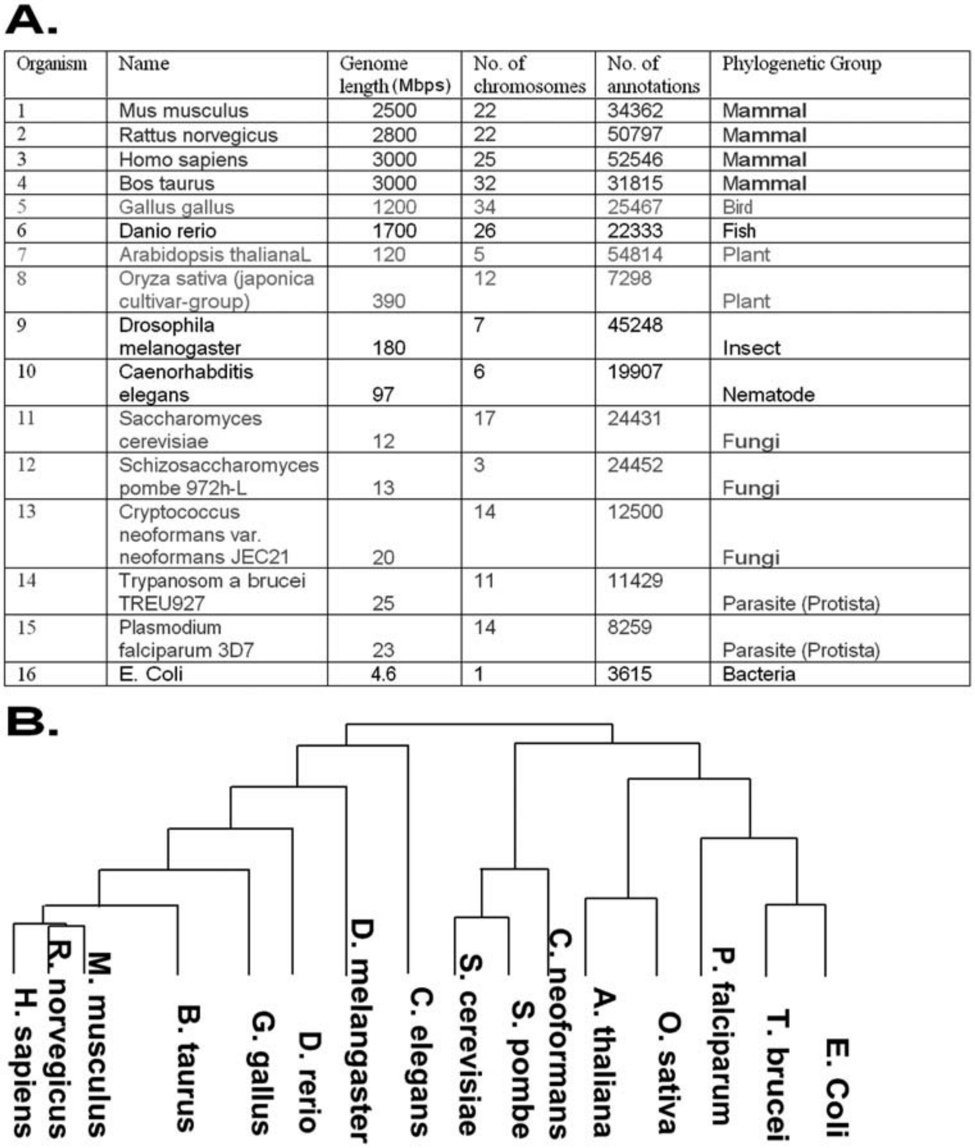
Ontology	Expression	PP	GI	TFs	Conservation	Co-Evolution
CC	**0.3048 (0.6250)**	0.58 (<0.001)	0.64 (<0.001)	0.5603 (<0.001)	0.5079 (<0.001)	0.4374 (0. 003)
MF	0.4058 (0.01)	0.45 (0.012)	**0.57 (0.119)**	0.6284 (<0.001)	0.4674 (<0.001)	**0.2212 (0.057)**
BP	0.42 (<0.001)	0.47 (<0.001)	0.57 (<0.001)	0.4550 (<0.001)	0.3809 (0.0090)	0.3 **(0.3258)**

(B)

FIG. 1. **(A)** Correlation between localization and other properties of GO functions (expression coherency, enrichment for protein-protein-interactions, enrichment for genetic interactions, enrichment for shared TFs, conservation and co-evolution). Correlations are presented for each ontology: Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). Significant correlations appear in gray. Notably, the only ontology whose localization correlates with co-evolution is the Cellular Component ontology. Similar results were obtained when we controlled for the size of the different GO groups (Supplementary Table 3). (See online Supplementary Material at *www.liebertonline.com*.) **(B)** Localization *p*-value versus the fraction of significantly co-expressed GO groups. Roughly speaking, on average, as the distances between members of a GO group increase, their expression coherency decreases.

2.3. Functional genomic organization across species

We turned to analyze the genomic functional organization of the 16 organisms with the most abundant ontologies currently available. Figure 2 provides some descriptors of these organisms and their phylogeny. The localization scores of all the GO groups in all the organisms appear in Supplementary Table 6. (See online Supplementary Material at *www.liebertonline.com*.) The GO groups that are most localized across species belong to the following categories: (1) cellular transport, (2) DNA replication and chromosome packaging, (3) protein metabolism, and (4) immune response. Although these processes involve coordinated expression of many genes, and gene clustering could represent a mechanism that ensures coordinated regulation, it is not clear to us what distinguishes these processes from others.



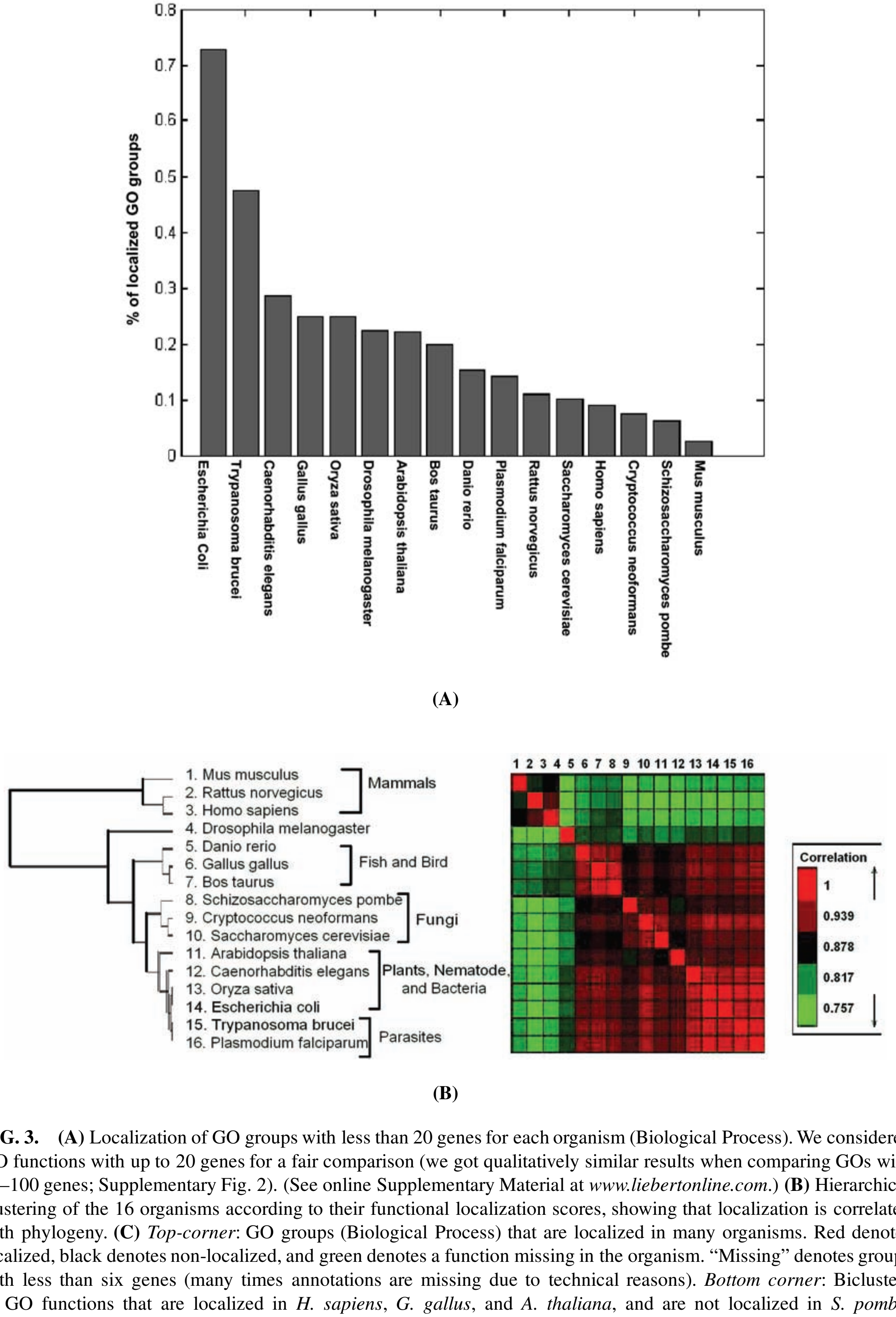
A.

Organism	Name	Genome length (Mbps)	No. of chromosomes	No. of annotations	Phylogenetic Group
1	Mus musculus	2500	22	34362	Mammal
2	Rattus norvegicus	2800	22	50797	Mammal
3	Homo sapiens	3000	25	52546	Mammal
4	Bos taurus	3000	32	31815	Mammal
5	Gallus gallus	1200	34	25467	Bird
6	Danio rerio	1700	26	22333	Fish
7	Arabidopsis thalianaL	120	5	54814	Plant
8	Oryza sativa (japonica cultivar-group)	390	12	7298	Plant
9	Drosophila melanogaster	180	7	45248	Insect
10	Caenorhabditis elegans	97	6	19907	Nematode
11	Saccharomyces cerevisiae	12	17	24431	Fungi
12	Schizosaccharomyces pombe 972h-L	13	3	24452	Fungi
13	Cryptococcus neoformans var. neoformans JEC21	20	14	12500	Fungi
14	Trypanosom a brucei TREU927	25	11	11429	Parasite (Protista)
15	Plasmodium falciparum 3D7	23	14	8259	Parasite (Protista)
16	E. Coli	4.6	1	3615	Bacteria

FIG. 2. **(A)** Name, genome size, number of annotations, and phylogenetic group for each organism studied here. **(B)** Their phylogeny, based on NCBI taxonomy (*www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/*) and ITOL (*http://itol.embl.de/*); the lengths of the branches are arbitrary.

To study the functional organization level of the different organisms in a comparative fashion, we computed the fraction of localized GO groups in each organism (i.e., the fraction of GO groups with a localization score p-value of <0.05; Fig. 3).

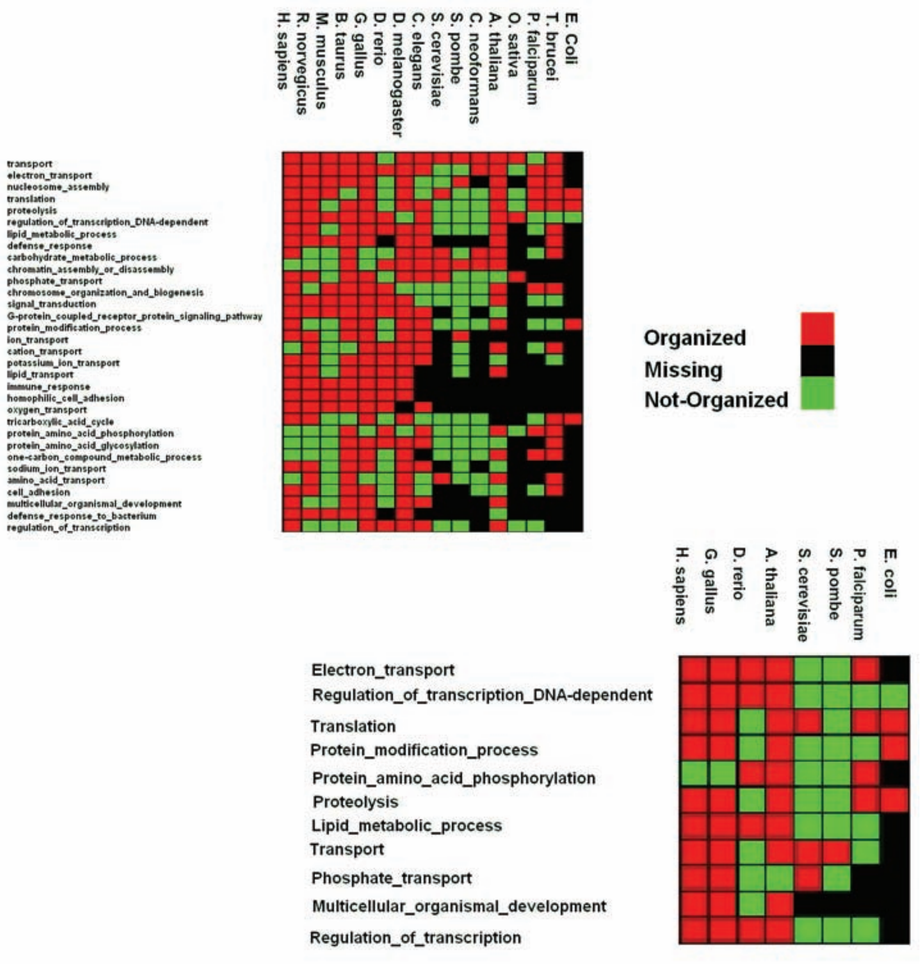
As expected, *E. coli* shows the highest level of organization (Fig. 3A), with 75% of the GO functions significantly localized, due to its operon-based genomic organization. Yet, some eukaryotes also have a fairly large fraction of organized functions, with *Trypanosoma brucei* (48% are localized) and *Caenorhabditis elegans* (28%) showing the highest organization level. In *C. elegans*, the high level of organization is at least probably partially due to the operon-like organization of ~15% of the worm genes (Blumenthal et al., 2002). The *Trypanosoma* results, however, are surprising: although it is known that its genome is transcribed in long polycistronic units, previous work failed to see any order in the way genes are distributed among these units (Ivens et al., 2005).



(A)

(B)

FIG. 3. **(A)** Localization of GO groups with less than 20 genes for each organism (Biological Process). We considered GO functions with up to 20 genes for a fair comparison (we got qualitatively similar results when comparing GOs with 20–100 genes; Supplementary Fig. 2). (See online Supplementary Material at *www.liebertonline.com*.) **(B)** Hierarchical clustering of the 16 organisms according to their functional localization scores, showing that localization is correlated with phylogeny. **(C)** *Top-corner*: GO groups (Biological Process) that are localized in many organisms. Red denotes localized, black denotes non-localized, and green denotes a function missing in the organism. "Missing" denotes groups with less than six genes (many times annotations are missing due to technical reasons). *Bottom corner*: Biclusters of GO functions that are localized in *H. sapiens*, *G. gallus*, and *A. thaliana*, and are not localized in *S. pombe*, *S. cerevisiae*, *D. rerio*, and *P. falsiparum*.



(C)

FIG. 3. (*Continued*).

Using the functional localization patterns computed for each organism (using all three ontologies), we performed a hierarchical clustering of the species surveyed. Figure 3B shows that the pattern obtained is significantly similar to the phylogeny of these organisms [the correlation between the pair-wise distances that are induced by the two trees is 0.68 (p-values of $<10^{-16}$)]. This similarity is maintained whether one considers each ontology separately or only the 52 GO functions that appear in all the organisms.

Further, the results remain significant ($r = 0.67$, p-values $= 10^{-4}$) even when comparing the corresponding trees after contracting each group of organisms to a single leaf. This additional test was performed to reject the possibility that these trees are similar mainly due to the fact that annotations are made many times by sequence similarity with closely related better-studied species. These results demonstrate that the functional genomic architecture tends to be more similar in organisms that have higher evolutionary proximity.

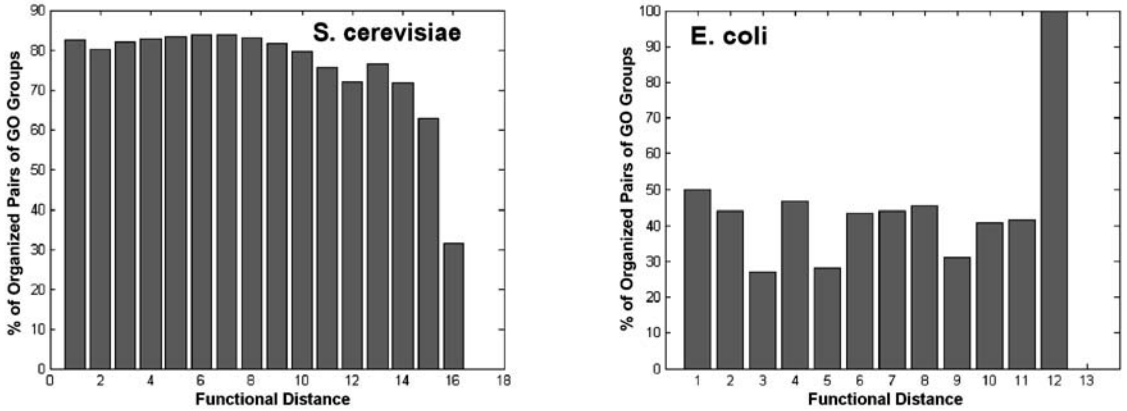
The top corner of Figure 3C includes the GO groups (Biological Process ontology) that are localized in many organisms. We performed a bi-clustering analysis [using the SAMBA algorithm (Shamir et al., 2005)] identifying subsets of GO groups that have similar localization levels in a subset of organisms. An example of a bi-cluster appears at the bottom corner of Figure 3C; more bi-clusters appear in Supplementary Table 7). (See online Supplementary Material at *www.liebertonline.com*.) This bi-cluster includes GO groups that exhibit an elevated level of localization in multi-cellular organisms (e.g., *H. sapiens*, *G. gallus*, and *A. thaliana*) compared to unicellular organisms (e.g., *S. pombe*, *S. cerevisiae*, and *P. falsiparum*). A plausible explanation for the existence of such a bi-cluster is that the ontologies it contains are related to the development of multi-cellular organisms, a process that requires accurate coordination between many signal transduction pathways and the correct communication between cells. Indeed, one of the GO groups in this cluster is *Multicellular organismal development*. Accuracy in gene expression may be achieved by increasing the level of organization of genes that affect the flow of information within the cell at all possible levels (e.g., *Regulation of transcription, Translation, Protein modification process, Protein aminoacid phosphorylation, Proteolysis*), all of which appear in the bi-cluster.

2.4. Co-localization of pairs of GO functions: S. cerevisiae *versus* E. coli

Next, we asked whether a second level of order can be detected, in which GO groups tend to be localized with respect to other GO groups. We defined *CoLoc*, a measure of co-localization for pairs of GO functions, that expresses the tendency of genes in two GO groups to be clustered together (see Supplementary Fig. 1C and more details in the Methods section, with emphasis on controlling for the individual organization level of each function on its own). (See online Supplementary Material at *www.liebertonline.com*.)

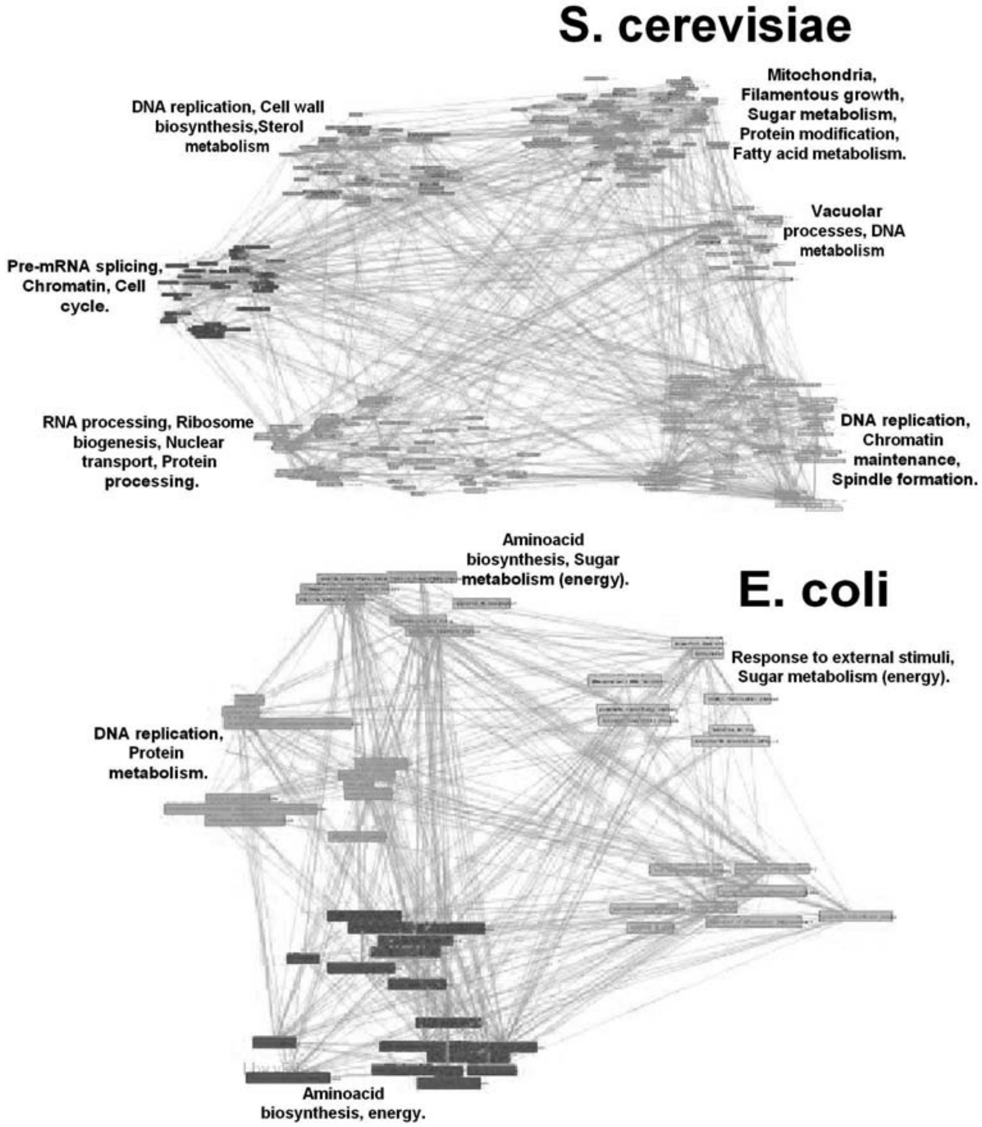
Figure 4A depicts the frequency of co-localized pairs of GO groups with less than 20 genes (y-axis) as a function of the functional distance between the GO functions (defined by the distance in the GO hierarchy, see, for example, *www.geneontology.org*; x-axis) for *S. cerevisiae* and for *E. coli*. Suprisingly, *S. cerevisiae* has a higher level of co-localization (∼80% in *S. cerevisiae* versus ∼40% in *E. coli*). Furthermore, *S. cerevisiae* has a higher correlation between the level of co-localization and the functional distance between the pair of co-localized functions than that observed in *E. coli* (*S. cerevisiae*: Spearman correlation -0.79, p-value $= 1.4 * 10^{-4}$; *E. coli*: Spearman correlation -0.2088, p-value $=$ 0.49) (Fig. 4B). Thus, although at the first-order level, individual functions are more organized in *E. coli* than in *S. cerevisiae*, at a second level, that of co-localization of pairs of functions, *S. cerevisiae* is markedly more organized than *E. coli*. Note that this figure depicts the *frequency* of co-localized pairs of GO groups. Thus, the fact that there are less GO groups in *E. coli* than in *S. cerevisiae* can not explain this figure (a sampling of the *S. cerevisiae* GO annotation to get similar number of GO annotations as in *E. coli* results in similar results).

We generated two GO co-localization networks for *S. cerevisiae* and *E. coli*, where each GO category is a node and pairs of GO groups are connected by an edge if their CoLoc p-value is below 0.05 (Fig. 4B). Supplementary Table 8 includes the *S. cerevisiae* pair wise co-localization p-values for the three ontologies and for various measures of co-localization; results for *E. coli* appear in Supplementary Table 9. (See online Supplementary Material at *www.liebertonline.com*.)



(A)

FIG. 4. **(A)** The frequency of co-localized pairs of GO groups with less than 20 genes (y-axis) as a function of the distance in the GO hierarchy, for both *S. cerevisiae* and *E. coli*. The inverse relation between these two variables is stronger in *S. cerevisiae*. **(B)** Co-localization networks for *S. cerevisiae* and *E. coli*. Each node denotes a GO BP annotation; edges connect two nodes that have co-localization p-values < 0.05. Supplementary Figures 3 and 4 are enlarged versions of this figure. (See online Supplementary Material at *www.liebertonline.com*.)


S. cerevisiae
DNA replication, Cell wall biosynthesis,Sterol metabolism
Mitochondria, Filamentous growth, Sugar metabolism, Protein modification, Fatty acid metabolism.
Vacuolar processes, DNA metabolism
Pre-mRNA splicing, Chromatin, Cell cycle.
RNA processing, Ribosome biogenesis, Nuclear transport, Protein processing.
DNA replication, Chromatin maintenance, Spindle formation.
Aminoacid biosynthesis, Sugar metabolism (energy).
E. coli
Response to external stimuli, Sugar metabolism (energy).
DNA replication, Protein metabolism.
Aminoacid biosynthesis, energy.

(B)

FIG. 4. (*Continued*).

The two networks show a significant connection between functionality (distance in the GO graph) and co-localization (distance in the co-localization graph). For *E. coli*, the Spearman correlation is 0.1165 (p-value = 0.05), while for *S. cerevisiae* the correlation is both higher and more significant: 0.14 (p-value < 0.001), again this may indicate that at this level of functional co-localization, *S. cerevisiae* is more organized than *E. coli*. These relatively low correlations suggest the existence of additional constraints, different from functionality, which determine the genomic location of functional gene groups.

We additionally used a clustering algorithm (Girvan and Newman, 2002) to partition these two co-localization graphs to clusters of functionally related GO groups, aiming to uncover their "meta-localization" architecture (Fig. 4B). In both organisms, we can detect the clustering of related functions. In *E. coli*, five "neighborhoods" are apparent. Two of them are enriched with GO groups involved in amino-acid biosynthesis and in sugar metabolism. A third group includes additional sugar metabolism pathways; a fourth clusters protein metabolism and DNA replication, whereas the fifth neighborhood contains apparently unrelated GO groups. *S. cerevisiae* shows an even more refined meta-localization structure (Fig. 4B): With the exception of a few GO groups, such as DNA replication, RNA processing, and chromatin, most GO functions appear restricted to a single neighborhood. The clusters formed are functionally coherent and genes within related GO categories map close to each other. These findings further strengthen the notion that the eukaryotic yeast has a more marked higher-level functional genomic organization than the prokaryotic *E. coli*.

3. CONCLUSION

One of the interesting results reported in this work is that, when considering co-localization of GO group pairs, *E. coli* seems less organized than *S. cerevisiae*. This result is surprising as in general prokaryotes are considered to be more organized than eukaryotes. Indeed, Figure 3A shows that most GO groups in *E. coli* exhibit a high level of order, reflecting the existence of operons in prokaryotic genomes. We term this level of gene organization the *primary level*. However, at a *secondary level* of organization (in which the relative organization between GO groups is measured), *E. coli* exhibits only low level of order (Fig. 4A). In contrast, *S. cerevisiae* has no operons and a much lower primary level of functional organization; however, at the secondary level, the yeast genome demonstrates a high degree of co-localization between related GO groups. The differences in functional gene architecture can be clearly seen in the density and distribution of closely located, functionally related GO groups (Fig. 4B). Our results thus uncover, at this secondary level, a high degree of organization of the eukaryotic genome that has not been explored to date.

Our observations imply an important distinction between prokaryotes and eukaryotic cells: in the first, gene regulation occurs mainly at the operon level, with each operon acting as an independent unit. Operons that affect related processes do not tend to map in the vicinity of each other. In contrast, yeast cells, which lack polycistronic units, show a low level of gene clustering, but gene distribution tends to be such that related functions co-localize. Potentially, such co-localized clusters may share particular chromosomal or nuclear domains, facilitating coordinated expression of many genes acting in related functions.

Interestingly, the human genome has a primary level of organization that is quite similar in its magnitude to that of the yeast (Fig. 3A). Characterizing the secondary level of organization of human and other available organisms is an interesting challenge for future related research. This may be used to answer questions about the systemic links between different levels of organization, and for understanding the properties that correlate with co-localization. Further, a natural generalization of our approach is to recursively define and study even higher levels of organization (third level, fourth level, and so on).

In summary, eukaryote genomes are more organized than was previously believed, and more careful studies tracking down their higher-order organization are called for.

4. METHODS

4.1. Localization of GO groups: definition of clusters and p-values

We used the number of gene clusters corresponding to a GO group as a measure of localization (i.e., lower number of clusters corresponds to higher level of localization). We examined two definitions of clusters:

Distance-based clustering, *LocD*: Let d_g denote the average distance between adjacent genes in a genome, g. A cluster of genes from a GO annotation, G, is either a set of genes such that the distance between pairs of consecutive genes in this set is $\leq d_g$, or it is a single gene. By this definition, two genes in a cluster can be non-adjacent (see Supplementary Fig. 1A, Clustering_methods.doc A). (See online Supplementary Material at *www.liebertonline.com.*)

Adjacency-based clustering, *LocN*: A cluster of genes from a GO group, G, is either a set of genes such that any pair of consecutive genes in the set are adjacent, or it is a single gene (see Supplementary Fig. 1B). (See online Supplementary Material at *www.liebertonline.com.*)

As the locations of genes along genomes is not random (Hurst et al., 2004), we computed a localization p-value for a GO group by comparing the number of clusters corresponding to the group to the number of clusters after performing 1000 random permutations of gene locations (a similar idea was described in

Hurst et al. [2004]). The p-value was defined as the fraction of random permutations with fewer or equal number of clusters than in the original case. Our measures are similar to the measure described in (Hurst et al. 2004). However, in Hurst et al. (2004), the score is the frequency of genes from the same group (in our case, the same GO annotations) that have another gene from the group as an immediate neighbor. In our work, we counted clusters, as we think it is a measure that provides a more intuitive understanding of the problem at hand. For example, according to Hurst et al. (2004), a case of three clusters, each with a pair of genes, will have higher score than a case of only two clusters, one with four and one with two genes. By our measures, the second case, which is clearly more organized, will have a higher score.

4.2. p-Value for chromosomal localization of a GO group

This p-value, *LocC*, corresponded to the distribution of the gene of a GO group among the organism chromosomes. Cellular function with more significant p-value are distributed less uniformly then expected among the different chromosomes (i.e., a larger fraction of the GO genes appear in a smaller fraction of the chromosomes). The p-value was based on the entropy of this chromosomal distribution; it was computed as follows:

Let n_i denote the number of genes from the GO group that are mapped to chromosome i. Let $p_i = \frac{n_i}{\sum_i n_i}$ denote the fraction of genes in chromosome i. The corresponding entropy of the GO group is $-\sum_i p_i \cdot \log(p_i)$, and it is an accepted measure for non-uniformity. An empirical p-value was computed as we mentioned above for *LocD* and *LocN.*

Note that this measure can not be implemented on organisms with only one chromosome (e.g., *E. coli*).

4.3. Localization p-value across few organisms

Let p_s denote a threshold (we used $p_s = 0.05$) that reflect a significant p-value. We used the following upper bound as a p-value reflecting localization in n out of the 16 studied organisms:

$$\sum_{x=n}^{16} \binom{16}{n} \cdot p_s^n \cdot (1 - p_s)^{16-n}.$$

p-Values were filtered by false discovery rate (FDR) to correct for multiple testing (Benjamini and Hochberg, 1995).

4.4. Co-localization (CoLoc) score for GO pairs

In this case, did not consider pairs of GO groups where one of the GO groups was a subset of the second GO group, or GO groups with at least six genes. We checked if a pair of GO groups, G_1 and G_2, tend to be cluster together by computing the two p-values in the following way: (1) Let G_U denote the union of G_1 and G_2. Compute the number of clusters of genes from G_U as described in the previous subsection. (2) Perform 1000 random shifts of all the genes from G_1 while maintaining the distances between all the pairs of genes in G_1 (see Supplementary Fig. 1C). (See online Supplementary Material at *www.liebertonline.com.*) For each shift, we computed the new number of clusters in G_U.

The first p-value is the fraction of cases that the number of clusters in a random shift was less or equal to the number of clusters in the original case. The second p-value was computed in a similar way by shifting G_2instead of G_1.

The goal of this process is to maintain the internal clustering structure in each GO group while testing the co-clustering of the two GO groups. Pair of GO groups are co-localized if the average of the above two p-values < 0.05.

4.5. Functional distance

The distance between GO groups on the GO hierarchy was computed by replacing each directed edge in the original graph with an undirected one, and computing the length of the minimal path between the two GO groups.

The GO annotation hierarchy is based on the designers of GO and as such it is not clear what is the accurate distance metric that should be used when comparing GO annotations. However, it is clear that roughly distances along the GO hierarchy graph correspond to distances between pairs of GO annotations. Thus, we used in the relevant parts of the paper, the non-parametric Spearman correlation (and the corresponding p-values) when we compared functional distances.

4.6. *Enrichment with genes from the same cohort in a GO function*

Chip-chip information of 203 TFs was downloaded from the work of Harbison et al. (2004; *http://web.wi.mit.edu/young/regulatory_code/*). We considered only interactions with p-value of ≤ 0.001. We computed for each GO group the number of genes from the GO group that are regulated by each TF. A p-value was computed by comparing the average genes regulated per TF to this average in random groups of genes with identical size.

4.7. *Various sources of data*

Information about the GO annotation and gene-order of the analyzed organisms was downloaded from NCBI. The GO ontology hierarchy was downloaded from OBO Foundry Ontologies (*http://obofoundry.org/*).

We used the *S. cerevisiae* protein interaction network from Sharan et al. (2005) and the genetic interaction network data from Tong et al. (2004). Gene expression data was taken from the Stanford MicroArray Database (Sherlock et al., 2001). The *S. cerevisiae* gene evolutionary rates were downloaded from Wall et al. (2005).

4.8. *Coherency enrichment and conservation scores*

In this work, we computed two coherency scores (co-expression, co-evolution), and two enrichment scores (enrichment with pp-interactions, and enrichment with genetic interactions). The co-expression score of a group of genes is the frequency of gene groups with similar size that have lower average pair wise correlation (computed empirically by sampling 100 groups with size identical to the size of the original one). The co-evolution score for a pair of genes was defined as in the work of Chen and Dokholyan (2006), as the absolute value of the difference between their evolutionary rates. Co-evolution score of a group of genes is the KS p-value when comparing the distribution of all the pairwise co-evolution scores in the group to the distribution of all pairwise co-evolution scores in the dataset. We used hypergeometric p-values for evaluating enrichments with pp-interaction and genetic interaction.

To test for evolutionary conservation, for each GO function (containing more than five genes), we counted the number of organisms in which it appears (for the list of organisms used, see Fig. 2).

4.9. *p-Values for Spearman correlations*

Some of the correlations reported in this work were between vectors of p-values. Many times, these vectors have a non-uniform distribution (i.e., they included many "1" and many "0" values). In such cases, the standard Spearman correlation p-values are biased (they are more significant than they should be). Thus, we computed empirical p-values by comparing the correlation to the correlations after permuting the vectors.

4.10. *Hierarchical clustering and p-values for comparing trees*

The distance matrix for the hierarchical clustering was based on the Spearman correlations between the localization p-values of the GO groups that are common to pairs of organisms.

For comparing two trees, we computed the Spearman correlation and a corresponding p-value between all the pair wise distances that were induced by each of the trees (i.e., the number of tree branches between each pair of organisms/leaf).

The correlation between the organism tree and the tree that was based on hierarchical clustering according to the GO groups from the three ontologies was 0.68 (p-values $< 10^{-16}$). When the hierarchical clustering was based on the GO groups that are common to all the organisms, the correlation was $r = 0.51$ (p-values $= 3 * 10^{-7}$). When we considered each ontology separately, the resulting hierarchical clusterings remained significantly similar to the organism tree (Biological Process: $r = 0.52$, p-values $= 8 * 10^{-10}$; Cellular Component: $r = 0.41$, p-values $= 3 * 10^{-6}$; Molecular Function: $r = 0.65$, p-values $= 2 * 10^{-15}$).

In this work, we used all GO annotations, regardless of evidence codes. Thus, it is possible that the trees reconstructed by the hierarchical clustering are similar to the real phylogenetic tree mainly due to this fact. Therefore, in order to check if the two trees are significantly similar even after controlling for this effect we performed the following additional experiment:

First, we contracted each group of organisms (mammals, fish, birds, Fungi, plants, bacteria, parasites, insects), clustered in the analyzed trees to a single leaf. The result of this stage was pairs of trees with eight leaves (corresponding to these groups). Next, we compared such pairs of trees, as described. As sequence similarity annotations are made by using a closely related organism that should be from the same organism group, the tree corresponding to these groups should not be affected by the annotation process. The result remain significant even when we compared these new trees ($r = 0.67$, p-values $= 10^{-4}$).

ACKNOWLEDGMENTS

T.T. was supported by the Edmond J. Safra Bioinformatics program at Tel Aviv University and the Yeshaya Horowitz Association through the Center for Complexity Science.

DISCLOSURE STATEMENT

No conflicting financial interests exist.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.

Bártová, E., and Kozubek, S. 2006. Nuclear architecture in the light of gene expression and cell differentiation studies. *Biol. Cell.* 98, 323–336.

Batada, N., and Hurst, L.D. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* 39, 945–949.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Mat.* 57, 289–300.

Blumenthal, T., Evans, D., Link, C.D., et al. 2002. A global analysis of *Caenorhabditis elegans* operons. *Nature* 417, 851–854.

Branco, M.R., and Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* 4, e138.

Caron, H., van Schaik, B., van der Mee, M., et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291, 1289–1292.

Chen, Y., and Dokholyan, N.V. 2006. The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends Genet.* 22, 416–419.

Cohen, B.A., Mitra, R.D., Hughes, J.D., et al. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186.

Cremer, T., Cremer, M., Dietzel, S., et al. 2006. Chromosome territories—a functional nuclear landscape. *Curr. Opin. Cell. Biol.* 18, 307–316.

Eichler, E.E., and Sankoff, D. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* 301, 793–797.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., et al. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90.

Girvan, M., and Newman, M.E. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 7821–7826.

Harbison, C.T., Gordon, D.B., Lee, T.I., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.

Hurst, L.D., Pál, C. and Lercher, M.J. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5, 299–331.

Huvet, M., Nicolay, S., Touchon, M., et al. 2007. Human gene organization driven by the coordination of replication and transcription. *Genome. Res.* 17, 1278–1285.

Ivens, A.C., Peacock, C.S., Worthey, E.A., et al. 2005. The genome of the kinetoplastid parasite, *Leishmania major*. *Science* 309, 436–442.

Kosak, S.T., and Groudine, M. 2004. Gene order and dynamic domains. *Science* 306, 644–647.

Lee, J.M., and Sonnhammer, E.L. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13, 875–882.

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* 31, 180–183.

Lercher, M.J., Urrutia, A.O., Pavlíek, A., et al. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* 12, 2411–2415.

Marcotte, E.M., Pellegrini, M., Ng, H.L., et al. 1999. Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753.

Miller, J. 1981. *The Operon*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Miller, M.A., Cutter, A.D., Yamamoto, I., et al. 2004. Clustered organization of reproductive genes in the *C. elegans* genome. *Curr. Biol.* 14, 1284–1290.

Pal, C., and Hurst, L.D. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat. Genet.* 33, 392–395.

Petkov, P.M., Graber, J.H., Churchill, G.A., et al. 2007. Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS. Biol.* 5, e127.

Poyatos, J.F., and Hurst, L.D. 2006. Is optimal gene order impossible? *Trends Genet.* 22, 420–423.

Poyatos, J.F., and Hurst, L.D. 2007. The determinants of gene order conservation in yeasts. *Genome Biol.* 8, R233.

Sémon, M., and Duret, L. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* 23, 1715–1723.

Shamir, R., Maron-Katz, A., Tanay, A., et al., 2005. Expander—an integrative program suite for microarray data analysis. *BMC Bioinform.* 6, 1–12.

Sharan, R., Suthram, S., Kelley, R.M., et al. 2005. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* 102, 1974–1979.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., et al. 2001. The Stanford Microarray Database. *Nucleic Acids Res.* 29, 152–155.

Singer, G.A., Lloyd, A.T., Huminiecki, L.B., et al. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* 22, 767–775.

Sproul, D., Gilbert, N., and Bickmore, W.A. 2005. The role of chromatin structure in regulating the expression of clustered genes. *Nat. Rev. Genet.* 6, 775–781.

Teichmann, S.A., and Veitia, R.A. 2004. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics* 167, 2121–2125.

Tong, A.H., Lesage, G., Bader, G.D., et al. 2004. Global mapping of the yeast genetic interaction network. *Science* 303, 808–813.

Versteeg, R., van Schaik, B.D., van Batenburg, M.F., et al. 2003. 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.* 13, 1998–2004.

Wall, D.P., Hirsh, A.E., Fraser, H.B., et al. 2005. Functional genomic analysis of the rate of protein evolution. *Proc. Natl. Acad. Sci. USA* 102, 5483–5488.

Wong, S., and Wolfe, K.H. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat. Genet.* 37, 777–782.

Yi, G., Sze, S.H., and Thon, M.R. 2007. Identifying clusters of functionally related genes in genomes. *Bioinformatics* 23, 1053–1060.

Address reprint requests to:
Dr. Tamir Tuller
School of Computer Sciences
Tel Aviv University
Tel Aviv 69978, Israel

E-mail: tamirtul@post.tau.ac.il