Clustering of gene ontology terms in genomes.

Tiirikka, Timo; Siermala, Markku; Vihinen, Mauno

Published in:
Gene

DOI:
10.1016/j.gene.2014.06.060

2014

CLUSTERING OF GENE ONTOLOGY TERMS IN GENOMES

Timo Tiirikka[1,2$], Markku Siermala[1,4] and Mauno Vihinen[1,3]

[1]Institute of Biomedical Technology, FI-33014 University of Tampere, Finland and BioMediTech, Tampere, Finland

[2]Medical Research Center Oulu, Oulu University Hospital and University of Oulu

[3]Department of Experimental Medical Science, Lund University, SE-22 184 Lund, Sweden

[4]Present address: Logica Oy, Hatanpäänkatu 3 F, PL 287, 33900 Tampere, Finland

[$]Corresponding author

E-mail: timo.tiirikka@student.oulu.fi

Tel. +358-29-5317819

Email addresses:

TT: timo.tiirikka@student.oulu.fi

MS: markku.siermala@luukku.com

MV: mauno.vihinen@med.lu.se

Keywords:

Genomics; Bioinformatics; Systems biology; Computational biology

Abstract

Although protein coding genes occupy only a small fraction of genomes in higher species, they are not randomly distributed within or between chromosomes. Clustering of genes with related function(s) and/or characteristics has been evident at several different levels. To study how common the clustering of functionally related genes is and what kind of functions the end products of these genes are involved, we collected gene ontology (GO) terms for complete genomes and developed a method to detect previously undefined gene clustering. Exhaustive analysis was performed for seven widely studied species ranging from human to Escherichia coli. To overcome problems related to varying gene lengths and densities, a novel method was developed and a fixed number of genes were analyzed irrespective of the genome span covered. Statistically very significant GO term clustering was apparent in all the investigated genomes. The analysis window, which ranged from 5 to 50 consecutive genes, revealed extensive GO term clusters for genes with widely varying functions. Here, the most interesting and significant results are discussed and the complete dataset for each analyzed species is available at the GOme database at http://bioinf.uta.fi/GOme. The results indicated that clusters of genes with related functions are very common, not only in bacteria, in which operons are frequent, but also in all the studied species irrespective of how complex they are. There are some differences between species but in all of them GO term clusters are common and of widely differing sizes. The presented method can be applied to analyze any genome or part of a genome for which descriptive features are available, and thus is not restricted to ontology terms. This method can also be applied to investigate gene and protein expression patterns. The results pave a way for further studies of mechanisms that shape genome structure and evolutionary forces related to them.

1 Background

Numerous complete genomes have been sequenced during the last decade. Because only a small fraction of each eukaryotic genome encodes proteins, genes have been thought to be randomly distributed within and between chromosomes. However, the organization of genes within eukaryotic genomes is clearly non-random (Hurst et al., 2004; Kosak and Groudine, 2004; Michalak, 2008). Notably, regions containing the most actively expressed genes have higher gene density (Versteeg et al., 2003; Woo et al., 2010). Consequently, regions of increased gene expression (ridges) are gene dense and have high G + C content (Versteeg et al., 2003). Serial analysis of gene expression (SAGE) and microarray studies have indicated that a large portion of co-expressed genes are clustered in specific areas of genomes (Elizondo et al., 2009; Singer et al., 2005), examples of which come from *Saccharomyces cerevisiae* (Cho et al., 1998; Cohen et al., 2000), *Drosophila melanogaster* (Boutanaev et al., 2002; Spellman and Rubin, 2002), *Homo sapiens* (Caron et al., 2001), *Gallus gallus* (Nie et al., 2010), *Caenorhabditis elegans* (Roy et al., 2002) and *Danio rerio* (Tsai et al., 2009). So called housekeeping or maintenance genes, which are expressed in most tissues, are also clustered (Lercher et al., 2002). In light of these results, genomes seem to be organized to facilitate efficient regulation of specific gene processes relating e.g. tissue formation (Al-Shahrour et al., 2010; Dewey et al., 2010). The clustering of mammalian imprinted genes is a prime example of non-random gene ordering in eukaryotes (Morison et al., 2005).

The analysis of expression data for yeast, fruit fly, worm, rat, mouse and human indicated that neighboring genes are likely co-expressed (Fukuoka et al., 2004). The proximity of a pair of genes has been used to predict gene functions (Raghupathy and Durand, 2009; Yanai et al., 2002). In bacteria, gene essentiality determines chromosome organization (Rocha and Danchin, 2003), and essential genes occur more frequently and are conserved in the leading replicating strand as compared with the expected average frequency for all genes. Protein sequences offer additional information about gene co-expression via network maps using the “betweenness” concept as an indicator of proteins having interrelated functions (Yu et al., 2007).

Analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic and signaling pathways in five eukaryote model species revealed that a high proportion of genes for individual pathways are clustered in each species' genome (Lee and Sonnhammer, 2003). There are differences among the species; however, 30–98% of the genes in the 69 investigated pathways were clustered. Still, only seven of the pathways were clustered for all the eukaryotes studied.

Many functionally related genes are organized in bacteria in operons, and operon-like gene clusters have been identified in many species including e.g. plants, animals, and also human (Osbourn and Field, 2009). Gene duplications generate groups of related genes (for a review see Reams and Neidle, 2004). There are also other mechanisms, especially for clusters of non-homologous genes. As the extent of clustering has not been systematically investigated, we performed genome wide studies for several model organisms based on gene annotations.

Genes and genomes have been annotated in many ways. Gene ontology (GO) terms are rich annotations of function, components, and cellular localization (Ashburner et al., 2000). Previously, GO terms were examined in some of the co-expression clusters in human and yeast (Fukuoka et al., 2004). Certain clusters were identified, but the events were rather rare. In another study, the chromosomal locations of DNA binding proteins encoded on human chromosome 19 strongly correlated with GO annotations (Castresana et al., 2004). Stanley et al. (2006) developed a method to identify statistically significant

GO terms associated with genomic positions. However, the number of genes and GO terms was relatively small in these studies.

The number of sequenced genomes is growing steadily. Although annotations have lagged behind, there are already a number of well annotated genomes with functional information for most genes and proteins. Here, we investigated the genome-wide GO distribution in seven species for which complete genomes are available, namely *H. sapiens*, *Mus musculus*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *Arabidopsis thaliana* and *Escherichia coli*. We determined the GO terms for each gene and gene product and then calculated statistics for the enrichment of GO terms in adjacent genes. The results indicate clear clustering of GO terms and functions within chromosomes and sequence regions in all the investigated species, and certain features of GO term distributions appear to be species specific.

We developed a method to investigate the co-occurrence of ontology annotations in genomes. The method is based on statistical analysis and provides information for a fixed number of consecutive genes, which facilitates analysis independent of gene density. This feature is advantageous because e.g. in human less than 5% of the genome contains protein-coding genes, and gene density varies significantly for different chromosomes and regions in them. Previous genome-wide clustering studies have been restricted to a standard length of a studied genome region (Kano et al., 2003), which provides limited insight into the clustering phenomenon. In another approach, genome positions rather than genes were assumed to be randomly distributed (Stanley et al. 2006). The effect of different gene sizes was avoided; however, they did not perform a genome-wide GO term analysis. The C_Hunter program(Yi et al., 2007) is more similar to our approach; however, it focuses on finding the longest GO term clusters in studied species and does not further analyze its findings on the genome level. In addition, DEFOG, a web based application by Wittkop et al. (2012) uses a resembling way to organize genes in a pathway to functionally related units in order to reduce the complexity of the clustering task. However, our method, which avoids gene length and size bias, was used to analyze GO term distributions in numerous complete genomes.

2. Results and Discussion

Our aim was to reveal how genes with related functions are distributed in genomes. The analysis is based on GO terms: a systematic description of molecular functions, biological processes, and cellular components. GO annotations were retrieved from NCBI Entrez Gene. GO terms are currently incomplete for any species, yet they are very useful and well suited for genome-wide statistical analyses. Some properties of the analyzed genomes are listed in Table 1. The number of genes varies widely, from4279 to 38,699, among the species that we investigated, and the human genome contains the largest number of genes. Among these genomes, there are from 0.88 (S. cerevisiae) to 8.4 (mouse) GO terms per gene on average. The average human gene has 4.4 annotations, which is about half of that for mouse (8.4). The ratios of GO term classes are somewhat different for each species (Fig. 1). Cellular component is the smallest GO term category in all examined cases.

The analysis of the human genome was performed starting with 38,699 genes, which is higher than the number of current, officially named genes because the automated analysis is based on genome annotations. The GO term coverage, i.e., the percentage of genes for which GO terms were found, was 25.4%. Altogether there were 53,844 molecular function, 51,631 cellular component, and 65,760 biological process ontology terms, totaling 171,235 GO terms (Table 1).

In order to illustrate how syntenic regions, genes and other markers with an evolutionary conserved order localize with GO term distribution we analyzed human and mouse syntenic regions alongside with mouse GO term clusters. The results, chromosome wise, are in Supplementary Figs. 12 to 32. The vast majority of GO clusters and syntenic regions seem to follow each other verifying the biological clustering process. However, there are differences too.

2.1. Analysis method

Hypergeometric distribution was used for statistical tests because it works well even with small datasets. This test has been widely used for GO annotation distribution studies. The uncorrected p-value had to be 10^{-6} or lower for the results to be considered statistically significant. We were interested just on the most significant findings which were obtained with this p-value. In addition, Bonferroni correction for multiple testing and false discovery rate (FDR) were used to overcome statistical problems regarding multiple testing. As the outcome of the analysis was very similar for the two corrections, results are only shown for Bonferroni corrected data (Supplementary Tables 1 and 2). Every GO term and level was considered to be equally important so we did not use scoring methods (see e.g. Alexa et al., 2006) to weigh for more detailed terms. We slid a window of a fixed number of genes for each investigated chromosome. The window was moved in steps of one gene. The width of the window was set to five genes and increased in five gene increments until 50 consecutive genes were included at one time. The reason for the choice of the range of window sizes was that based on published information widely different sizes of clustered genes had been identified. Our goal was to investigate the extent of the phenomenon of functionally related genes in diverse species. Statistics were calculated for the distribution of GO terms within the window. Because the number of genes with ontology term classes varies widely among species, we used expected values for the random occurrence of the GO terms. Results are calculated based on existing annotations and such are affected by any features affecting annotations. Even if the annotations are biased in some way it is likely that the annotations of related genes are affected quite similarly. In order to verify that the phenomenon of the clustering of gene ontologies and thus functions, processes and components is valid, a randomization study was conducted. However, initial hypothesis was not affected by the randomization results; the phenomenon was statistically much stronger than the simulated study. The results are first shown for the human genome, and trends and differences among other genomes are discussed in later chapters.

2.1.1 GOme database

To maintain and distribute vast amounts of data for different species, we created the GOme database. It is a mySQL-based system containing all the available data from our analysis, enabling the user to search and browse through different species, GO terms or genomic positions. The database provides links to additional information using AmiGO service: amigo.geneontology.org (the Gene Ontology project). For example, one can narrow down a search to certain chromosomes of a species and verify the results using the p-value as a qualifier. The results from different species are comparable at the ontology level, which provides a good reference to evolutionary conservation and gene similarity among genomes. Detailed results for each species are available at the GOme web service at http://bioinf.uta.fi/GOme. GOme has a graphical user interface from where the user can choose the organism of interest and then narrow down the search to a chromosome or limit the search by different factors in the database, if wanted. Results can be sorted e.g. by p-value or GO terms.

2.2. GO term distribution

Table 2 lists the 20 most commonly appearing ontology terms in the human genome including biological process (BP; four GO terms), cellular component (CC; eight terms) and molecular function (MF; eight annotations). The majority of these terms are very general such as nucleus, integral to membrane binding, protein binding and signal transduction. However, there are exceptions: the keratin filament (GO:0045095 for 1778 genes) represents a more specific GO term. The level, i.e., the distance (number of steps) from the root of the ontology, is a measure of how detailed a GO term is. The majority of these GO terms range from levels 1 to 8. Many terms are nested and therefore can have different levels in different situations. In Table 2, the highest level for GO terms was used. Different properties have different numbers of levels i.e. molecular function lays on level one whereas regulation of molecular function is on level 3, so it is not possible to define how detailed the data is for a given process, function, or component based solely on the level information. GO is a directed acyclic graph (DAG), and thus levels of annotations do not necessarily illustrate the shortest path to the root term, depending on the details of annotations. It is also possible that there are pairs of related terms (parent–child) on a certain level.

The abundance of GO term annotations does not indicate the frequency of clustered GO terms. The 30 statistically most clustered GO terms from different window sizes of the human genome are shown in Supplementary Table 1. The levels for the GO terms from significantly clustered windows vary greatly from 1 to 12. They are generally on levels 6 to 8 with certain exceptions such as GO:0004872 (level 4), which has a very broad definition: "combining with an extracellular or intracellular messenger to initiate a change in cell activity". For example, GO:0045095, the ontology with the highest distance from the root (level = 12), has a much more specific definition: "A filament composed of acidic and basic keratins (types I and II), typically expressed in epithelial cells. The keratins are the most diverse classes of intermediate filament proteins, with a large number of keratin isoforms being expressed. Each type of epithelium always expresses a characteristic combination of type I and type II keratins."

Some of the enriched GO terms in window 5 differ from those obtained for other window sizes. Otherwise, the GO term enrichment results are similar for the other window sizes with differences in the order of terms. The most abundant term in window 5 is GO:0004872, which is associated with receptor activity. The most common GO term For window size 45 is GO:0004984 (olfactory receptor activity) whereas in the longest window it is GO:0004872 (receptor activity).

Olfactory receptor (OR)-related ontology descriptions are an example of another type of clustering. OR genes form clusters (Glusman et al., 2001; Niimura and Nei, 2003). The genes linked to olfactory stimulus and to the sense of smell in general are postulated to constitute 3% of all known genes (Consortium, I.H.G.S., 2004). Previously, 95 olfactory receptor clusters were identified (Niimura and Nei, 2003). Our analysis indicates that there are 309 individual OR clusters in window 5 and a total of 409 clusters in window 10. The highest amount of clustered OR-related GO terms is found in window 50 (905 genes).

Molecular function (n=12), biological process (n=10) and cellular component (n = 8) constitute the 30 most clustered GO terms in window 5 (Supplementary Table 1). With a window size of 50, the number of molecular function, biological process, and cellular component terms changes to 15, 7 and 8, respectively. Several GO terms are among the significantly most clustered terms for all window sizes. Observations with very low p-values for shorter window sizes are considered significant for longer windows even without additional matching genes. Depending on gene density and gene distribution, the distances between observed clusters vary greatly. The distance is in bp between the end of the first cluster and beginning of the second one. The shortest distance is just one, and the longest distance is 21,565,613 bp between two clusters for GO:0005515, protein binding, in window 5 in chromosome 15.

The window analysis with 50 genes produced general GO terms, however, all the most significant results were obtained for window 5. Shorter windows yield a broader term distribution to different chromosomes when compared in the whole genome context (Supplementary Table 3). Data for window 50 contains 304 unique GO terms, whereas data for window size 5 contains 346 unique terms.

When ranked according to p-values, the most significantly clustered GO terms in window 5 are killing of cells of another organism (GO:0031640) and defense response to fungus (GO:0050832), which are related terms, and epidermis development (GO:0008544) (Supplementary Table 4). However, in window 50 the statistically most clustered terms are sensory perception of taste (GO:0050909), keratinocyte differentiation (GO:0030216) and MHC class II receptor activity (GO:0032395). Immune system related GO terms are highly clustered regardless of the window length.

To further test the significance of the observations we performed a test where genes were randomly picked from the genome and calculated Bonferroni and Šidàk corrected p-values for the observations (Supplementary Table 6). The test was performed for three window lengths, 5, 15 and 25. The test yielded a number of statistically significant GO clusters. For window 5, 974 statistically significant clusters (p-value less than 5%) were observed. The results for windows 15 and 25 contained 3774 and 6464 significance GO clusters, respectively. The p-values for the top findings in window 5 ranged from 1.96 $x10^{-7}$ to 0.00025 (Supplementary Table 6). When we compare these results to those in Supplementary Table 1 it is evident that the results for the human genome have much higher p-values, for example in window 5 for GO:0031640 being 3.16×10^{-11} and for window 15 GO:0031424 below 10^{-30}. Our discussion is based on these most significant observations, which indicate that the GO term clustering has several orders of magnitude more significant results than random sampling observations.

2.2.1 Chromosome analysis

GO terms display a highly skewed distribution in chromosomes. Some regions are very rich in ontology terms, whereas others are rather GO term poor. Using our analysis protocol, a fixed number of genes are analyzed irrespective of gene density. If the gene distribution is assumed to be equal within and/or across chromosomes, the average unique GO term count per chromosome would be less than 13. However, this is not always the case; in chromosome 5, for example, there are 27 unique GO terms. The number of terms per gene varies from 0 to 31 (with an average of 3.2) as illustrated in Fig. 2. Typically, the number of GO terms ranges from four to seven, suggesting that GO terms are still quite general because there are only three categories that define GO terms and the classification always starts from the root. There are only a few hundred genes that have 15 or more GO terms.

The largest number of GO terms for a cytoband appears at chromosome 19p13.3, where there are 2129 annotations out of which 53 are unique terms appearing just once (Fig. 3). Whereas in the whole chromosome 19 there are 1285 terms, which is the second highest amount in the human genome (chromosome 1, which is substantially longer, is annotated with 1793 GO terms). Chromosome 19 contains many significant clusters related to DNA binding proteins. Genes in this region have a high G + C content (Castresana et al., 2004). Cytoband 19p13.13 and its surrounding areas (especially 19p13.12–14) contain numerous genes for olfactory receptors (Malnic et al., 2004). Fig. 5 presents the GO term distribution for cytobands in other chromosomes. The smallest number of significant clusters appears in chromosome Y, which also has the fewest number of genes.

The size of the analysis window affects the observations. Narrow windows are sufficient for detecting locally grouped small clusters: for example, small pathways such as γaminobutyric acid signaling pathway (GO:0007124) consist of 10 proteins according to the InterPro database. On the other hand, longer windows can be used to detect more general enrichment or clustering of biological functions. We examined the relationship between window size and commonly clustered GO terms for the major histocompatibility complex (MHC) genes in chromosome 6 (Fig. 4). There are significant GO term clusters for all window sizes. It was previously hypothesized that human genes rarely form clusters (Fukuoka et al., 2004), and in yeast it was hypothesized that clusters contain fewer than 10 genes (Hurst et al., 2004). Based on our genome-wide analysis, there are also many larger clusters in the investigated species. A recent study indicates that the longest statistically significant cluster in humans contains 84 genes (Yi et al., 2007), which is also illustrated by our analysis at the longest window. Our method does not allow identification of the longest window.

A comparison of the cytoband locations and the regions of enriched GO terms, ontologybands, as well as the locations of genes in human chromosomes is shown in Fig. 5. Several of the ontologybands colocalize with cytobands; however, there are cytobands with few or no ontologybands and vice versa. The shorter arms of chromosomes 13, 14, 15, 21 and 22 do not contain any ontologybands. These results differ when using different window sizes; yet numerous ontologybands are conserved over different window sizes. Many ontologybands are identical for window sizes 30 to 50. If there is an ontologyband for longer windows, then there are also bands for shorter windows, but the converse situation does not hold true.

Upon further examination of the distribution of the significant observations, these clusters seldom stretch over the centromere except in chromosomes 4, 5 and 20. This result has nothing to do with gene density, which is typically low in this region, because a fixed number of genes are always used in the analysis. This observation may reflect the effect of centromeres on chromosomal organization in crossing over and duplications.

Analysis of the most statistically significant ontology findings for chromosomes (Supplementary Table 4) reveals that there is a wide variation in the GO terms and the types of them within and between chromosomes. Generally, more detailed terms correlate with shorter windows; however, there are exceptions such as homophilic cell adhesion (GO:0007156) (level 6) in chromosome 5 with window 40 as the most significant finding. GO terms differ among the short and long windows, but for example, bile acid transporter activity (GO:0015125) in chromosome 10 is among the most significant results for all windows.

2.2.2. Comparison of observed clusters to experimental data

The observed GO term clusters are consistent with published information. Chromosome 6 contains several immunology-related genes in clusters, especially those for the MHC complex. MHC genes localize to the short arm of chromosome 6, cytoband 6p21.3, in three blocks (Horton et al., 2004). The class I genes span about 1.9 Mbp with C6orf40 and MICB as border genes, whereas the class II region ranges from C6orf10 to HCG24 covering 0.9 Mbp (Horton et al., 2004). Class III ranges from PPIP9 to NOTCH4 within 0.7 Mbp. Extended MHC areas flank the classical areas (Horton et al., 2004). Altogether, the extended MHC region covers 7.6 Mbp.

There are 374 genes within the MHC area flanked by SCGN and PHF1. Research suggests that there are 461 human genes in the MHC extended region, but only 252 of them are expressed (Horton et al., 2004; MHC sequencing consortium, 1999). Genes in this region have altogether 227 GO terms in 59 topographically separate clusters for different window sizes. On average, there are four terms per gene in the MHC region.

Among significantly enriched GO term clusters, human chromosome 6 contains 65 MHC-related GO terms (the word MHC occurs in the ontology name) from window 5 data. The number of MHC-related terms increases to 452 for window 50. Other windows in numerical order show steady growth: 137, 163, 213, 244, 284, 320, 383, and 440.

In the MHC region, there are 416 GO terms that display significant clustering in window 5; these include the immune response, MHC class II receptor activity, and olfactory receptor activity terms. The number of clustered genes increases concomitantly with an increase in window size - up to 1564 genes for window 50. The number of unique GO terms remains constant for the analyzed window length varying between 32 and 47. The significantly clustered MHC terms are almost exclusively in chromosome 6. Additionally, for window 10, only one significant MHC term cluster is found in chromosome 1, and three are found in chromosome 10. The results for additional MHC GO terms in all the other windows are for genes in chromosome 19.

The mouse MHC region has a very similar structure to that of humans (Kumánovics et al., 2003; Walter et al., 2002), which is also apparent based on the ontology term analysis. The vast majority of mouse MHC genes are located in chromosome 17. In the concatenated overlapping clusters, the occurrences of the same GO term range from24 to 396 (windows 5 and 50, respectively). Although *D. melanogaster* and *C. elegans* have MHC terms, the genes are not significantly clustered and we did not observe significant clusters in these species.

2.3. Other species

2.3.1 Mus musculus

In total, there are 241,226 ontology annotations with an average of 8.4 GO terms per gene (Table 1). The clustering of GO terms is even more apparent than that in the human genome (Fig. 6). Many clusters for longer windows overlap and form very long patterns that cover substantial portions of the genome. Few regions in the mouse genome do not contain clustered genes for any of the analyzed window sizes. Similar to the human genome, certain regions without GO term clusters contain cytobands, for example in chromosomes 1, 2 and X. In window 5 data, 318 of the 15,224 clustered terms are unique. The number of unique terms increases to 412 for window 50.

The most common GO terms in the mouse genome are integral to the membrane (GO:0016021) and nucleus (GO:0005634) (Supplementary Table 2). Receptor activity (GO:0004872) has the highest occurrence rate, 1149, in the clustered data. Next on the list are signal transducer activity (GO:0004871) and G-protein-coupled receptor activity (GO:0004930) for window 5 (Supplementary Table 5a).

Among all chromosomes, chromosome 10 has the greatest number of the most common GO terms (Supplementary Table 3); however, chromosome 10 does not appear in the list of observations with the highest p-values (Supplementary Table 4). Chromosome 2 contains the largest number of GO terms in the p-value-sorted data. The most significant individual term is interleukin-1 receptor binding (GO:0005149) with a p-value of 1.68e^{-12} (Supplementary Table 4). This term also has high information content because it is on level 7. The second most significant term, phospholipase A2 activity (GO:0004623), has a p-value of 1.97e^{-12} at level 9.

2.3.2. D. melanogaster

Drosophila has a total of 64,914 ontology annotations (Table 1) of which 4405 appear in the data for window 5, 333 being unique. The significant clusters are almost evenly distributed throughout the genome (Fig. 7), and no significant clusters are present in chromosome 4. Chromosome 3R contains the highest number of unique GO terms (GO:0001437). The 20 most common terms appear in chromosome 2L (Supplementary Table 2). Nucleus (GO:0005634) is the most frequent term in *Drosophila* followed by proteolysis (GO:0006508) and cellular component (GO:0008372). In window 5, the most abundant clustered GO term is proteolysis (GO:0006508) with 230 occurrences, followed by integral to membrane (GO:0016021) (Supplementary Table 5b). The most statistically significant term in window 5 is phosphatidate phosphatase activity (GO:0008195) with a p-value of 6.30e^{-15} in chromosome 3L (Supplementary Table 4). Genes with this ontology definition have a function in guiding migrating germ cells during *Drosophila* development (Halbleib and Nelson, 2006). Other highly significant terms are puparial adhesion (GO:0007594) and aminoacylase activity (GO:0004046).

The closest significant GO terms—for two clusters of transcription factor activity (GO:0003700)—are separated by 537 bp. *Drosophila* has a tight intra-cluster organization: there are 108 genes found with a 10-bp distance from each other. Two pairs of genes have zero distance: Fbgn0040064 and Fbgn0025700 having the cellular component-type GO term, mitochondrion (GO:0005739), in chromosome 2L, and FBgn0037877 and Fbgn0037876 in nucleic acid binding (GO:0003676) in chromosome 3R. In the data for window 5, 5739 genes in clusters occur in the positive strand, and 5176 occur in the negative strand.

Additional results for GO term distribution and clustering in the genomes of *C. elegans, A. thaliana, S. cerevisiae,* and *E. coli* are in Supplement.

3. Conclusions

A new method was developed to study GO term distribution in genomes. Based on the analysis, we conclude that all the investigated genomes are organized such that functionally related genes are often located close to each other. This most likely indicates that because of regulation, expression, chromosome structure or other reasons many genes with related functions have to appear in close proximity. Traditionally, operons have been found from prokaryotes for co-regulated genes. The eukaryote *C. elegans*, however has at least 1000 operons (Blumenthal et al., 2002; Zorio et al., 1994). In eukaryotes, coexpressed genes and their resemblance to operons have been discussed (Osbourn and Field, 2009).

Clusters of paralogous genes have emerged by tandem duplication and amplification and subsequent evolution leading to diverse functions (Reams and Neidle, 2004), whereas clusters of non-homologous but functionally related genes have a different evolutionary origin. The effect of natural selection is still elusive. As explanation has been suggested e.g. the selfish operon model by horizontal gene transfer (Lawrence and Roth, 1996) and persistence model (Fang et al., 2008). These models do not consider direct selection, which however has been proposed as an explanation for operon formation (Price et al., 2005). Our results can be used to test these and other models to explain features related to genomic organization, operon formation, conservation and for example expression of coclustered genes.

GO terms were significantly clustered over a wide range from5 to 50 consecutive genes. The results are statistically significant as analyzed compared to the expected distribution. Some species-specific differences are apparent. For example, the *E. coli* genome is very highly clustered, whereas the *C. elegans* genome contains some relatively large regions without significant ontology enrichment. There is no clear correlation between cytobands and ontologybands. The results are available in a user-friendly GOme database at http://bioinf.uta.fi/GOme.

The analysis approach can be applied to investigate smaller genome regions in relation to gene and protein expression studies. In addition, the statistical analysis approach can be used to study enrichment of any properties and characteristics related to genes and proteins for which data are available in large quantities. The results showed similar trends when compared to not only two genome-wide clustering services GONOME

(Stanley et al., 2006) and C_Hunter (Yi et al., 2007), but also novel clusters. Since the method is different between these approaches our results are not directly comparable.

4. Materials and methods

4.1 Identifying clustered genes and ontology terms

Entrez Gene (release 37.1) was used to retrieve ontology terms for genes derived from the *H. sapiens* genome (Consortium I.H.G.S., 2004). Entrez Gene was chosen because it is manually annotated (being more reliable than automatically annotated systems) and it has GO codes directly linked to corresponding genes. *A. thaliana* (Arabidopsis Genome Initiative, 2000), *D. melanogaster* (Adams et al., 2000), *M. musculus* (Waterston et al., 2002), *C. elegans* (Consortium T.C.e.S., 1998), *S. cerevisiae* (Goffeau et al., 1996) and *E. coli* (Blattner et al., 1997) were also included in the analysis. The chosen taxa represent different types of species including mammals, insects, worms and plants. There were 38,699 human genes obtained from the Entrez Gene compilation. We used full GO annotations for genes available in the www.thegeneontology.org service available in January 2011. These annotations cover the current knowledge of different transcripts and protein products.

GenBank was used to derive GeneIDs because we needed a reliable cross-reference code for species not annotated by NCBI. Genes were localized to contigs, and then GO terms were linked to genes using a set of custom Perl scripts. The BioData database was built to facilitate the processing of ontology terms, GeneIDs and clustering (unpublished data).

To investigate the distribution of GO terms, we utilized a sliding window method where a fixed number of consecutive genes were analyzed within the window, and then the window was moved stepwise by one gene until the entire chromosome was covered. The window size varied from5 to 50 in intervals of 5 genes. The statistical significance of the observed number of GO terms within a window was calculated from the hypergeometric distribution (1), which gives an exact answer even with small sample sizes:

$$p - value = \sum_{x}^{n} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

where *N* is the total number of genes, and *M* is the number of occurrences of a particular GO term. If we observe *x* genes with the particular GO term in a sample of *n* genes, then we can calculate the probability for such an event indicated by the p-value. The window size is the variable *n* having values between 5 to 50 in increments of 5. The minimum number of x is one when in a sample of n genes there are no clustered GO terms. If we have 533 genes (*N*) and there are 74 (*M*) occurrences of certain GO term and we have observed 3 (*x*) genes with the GO term of interest in a sample (or a window) of 50 (n) genes, this produces a p-value of 0.0607256. The formula for the expected value *E(X)* of a hypergeometric distribution is

$$E(X) = \mathrm{n}\,(\frac{b}{C})$$

where n is the window size, *b* is the total number of GO terms used to annotate a chromosome, and C represents the total number of genes in the chromosome. If we consider a window of 50 genes (n) and 10 GO terms (*b*) in a chromosome where there are a total of 500 genes (*C*) the *E(X)* value would be 1. This information can be used to test whether the results are over- or underrepresented compared to the expected random distribution of annotations. To correct for multiple testing, adapted Bonferroni correction was calculated using the number of chromosomes to get the genome-wise p-values. The correction was calculated as follows

$$P_{genome} = 1 - (1 - P_{chr})^{c}$$

where *c* is the number of chromosomes in the studied species (de Koning et al., 1998). P_{chr} is the p-value derived from the hypergeometric test for each chromosome in the studied genome using the number of GO terms in the chromosome as a factor. The corrected P_{genome} value depends on the used window size.

As a further test we conducted a fully randomized ontology enrichment study. With a Perl script a sample of 10,000 human genes with at least one GO term was obtained. The genes were annotated with 96,686 GO terms, which is more than half of that for the genome. Analysis was done using three window sizes — 5, 15 and 25.

We used in here the Šidàk correction (Šidàk, 1967) to overcome the multiple testing problem. In this case we could not use chromosomes as a measurement for GO term distribution; instead we used the number of unique GO terms per window. However, the results for standard Bonferroni correction are included in Supplementary Table 6, where the chromosome number was taken into account.

The DAG nature of gene GO terms did not require special attention as we used annotations from a database and did not infer them. Calculations for all of the window sizes were performed in a Linux cluster of 40 nodes, where the analysis took few days for the human genome. This is mainly due to the computational complexity of the hypergeometric function (4), which is

$$O((\log n)^{2} M(n))$$

The syntenic regions for mouse and human were retrieved from the Cinteny web service (http://cinteny.cchmc.org/) by Sinha and Meller (2007). The number of common markers (genes) was 16,495 out of a total of 32,887 stored in the service. The data was used to construct Supplementary Figs. 12 to 32.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2014.06.060.

Abbreviations

GO: Gene Ontology; SAGE: Serial Analysis of Gene Expression; KEGG: Kyoto Encyclopedia of Genes and Genomes; MHC: Major Histocompatibility Complex; OR: Olfactory Receptor; CC: Cellular Component; MF: Molecular Function; BP: Biological Process;

Author Contributions

TT designed and implemented the utilized programs, and drafted the manuscript with MV. MS participated in the statistical analysis. MV conceived of the study and participated in its design and coordination. All authors read and approved the final manuscript.

Acknowledgements

We gratefully acknowledge financial support from the Competitive Research Funding of the Tampere University Hospital, Biocenter Finland, and Sigrid Juselius Foundation.

References

Adams, MD., et al., 2000. **The genome sequence of *Drosophila melanogaster***. *Science,* **287**: 2185-2195.

Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A., Dopazo, J., 2010. **Selection upon genome architecture: conservation of functional neighborhoods with changing genes.** *PLoS Comput Biol.* Oct 7;**6**(10)

Alexa, A., Rahnenführer, J., Lengauer, T., 2006. **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics*, **22**:1600-1607.

Arabidopsis Genome Initiative, 2000. **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature*, **408**: 796-815.

Ashburner, M., et al., 2000. **Gene ontology: tool for the unification of biology.** The Gene Ontology Consortium. *Nat. Genet.*, **25**: 25-29.

Blattner, FR., et al., 1997. **The complete genome sequence of *Escherichia coli* K-12**. *Science*, **277**: 1453-1474.

Blumenthal, T., et al., 2002. **A global analysis of *Caenorhabditis elegans* operons**. *Nature*, **417**: 851-854.

Boutanaev, AM., Kalmykova, AI., Shevelyov, YY., Nurminsky DI., 2002. **Large clusters of co-expressed genes in the Drosophila genome**. *Nature*, **420**: 666-669.

Caron, H., et al., 2001. **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science,* **291**: 1289-1292.

Castresana, J., Guigo, R., Alba, MM., 2004. **Clustering of genes coding for DNA binding proteins in a region of atypical evolution of the human genome.** *J Mol Evol*, **59**: 72-79.

Cho, RJ et al., 1998. **A genome-wide transcriptional analysis of the mitotic cell cycle**. *Mol Cell*, **2**: 65-73.

Cohen, BA., Mitra, RD., Hughes, JD., Church, GM., 2000. **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet,* **26**: 183-186.

Consortium, I.H.G.S, 2004. **Finishing the euchromatic sequence of the human genome**. *Nature*, **431**: 931-945.

Consortium, T.C.e.S, 1998. Genome **sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science,* **282**: 2012-2018.

de Koning, DJ., Visscher, PM., Knott, S., Haley, CS., 1998. **A strategy for detection of QTL in half-sib populations.** *Anim. Sci.*, **67**: 257-268.

Dewey, FE., Perez, MV., Wheeler, MT., Watt, C., Spin, J., Langfelder, P., Horvath, S., Hannenhalli, S., Cappola, TP., Ashley, EA., 2010. **Gene coexpression network topology of cardiac development, hypertrophy, and failure.** *Circ Cardiovasc Genet.*, **4**:26-35.

Elizondo, LI., Jafar-Nejad, P., Clewing, JM., Boerkoel, CF., 2009. **Gene clusters, molecular evolution and disease: a speculation.** *Curr Genomics*;10:64-75.

Fang, G., Rocha, E.P.C., Danchin, A., 2008. **Persistence drives gene clustering in bacterial genomes.** *BMC Genomics*, **9**: 4

Fukuoka, Y., Inaoka, H., Kohane, IS., 2004. **Inter-species differences of co-expression of neighboring genes in eukaryotic genomes.** *BMC Genomics*, **5**: 4.

Glusman, G., Yanai, I., Rubin, I., Lancet, D., 2001. **The complete human olfactory subgenome**. *Genome Res*, **11**: 685-702.

Goffeau, A., et al., 1996. **Life with 6000 genes**. *Science*, **274**: 546, 563-547.

Halbleib, JM., Nelson, WJ., 2006. **Cadherins in development: cell adhesion, sorting, and tissue morphogenesis.** *Genes Dev*, **20**: 3199-3214.

Horton, R., et al, 2004. **Gene map of the extended human MHC**. *Nat Rev Genet,* **5**:889-899

Hurst, L., Pal, C., Lercher, MJ., 2004. **The evolutionary dynamics of eukaryotic gene order**. *Nat Rev Genet*, **5**: 299-310.

Kano, M., Nishimura, K., Ishikawa, S., Tsutsumi, S., Hirota, K., Hirose, M., Aburatani, H., 2003. **Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions.** *Physiol Genomics*, **13**: 31-46.

Kosak, ST., Groudine M., 2004. **Gene order and dynamic domains**. *Science*, **306**: 644-647.

Kumánovics, A., Takada, T., Lindahl, KF., 2003. **Genomic organization of the mammalian MH**C. *Annu Rev Immunol*, **21**:629-657.

Lawrence, J.G., Roth, J.R., 1996. **Selfish operons: horizontal transfer may drive the evolution of gene clusters**. *Genetics Department of Biology, University of Utah, Salt Lake City 84112, USA. lawrence@biology.utah.edu*, **143,**: 1843-1860

Lee , JM., Sonnhammer, EL,. 2003. **Genomic gene clustering analysis of pathways in eukaryotes**. *Genome Res*, **13**: 875-882.

Lercher, MJ., Urrutia, AO., Hurst, LD., 2002. **Clustering of housekeeping genes provides a unified model of gene order in the human genome**. *Nat Genet,* **31**: 180-183.

Malnic, B., Godfrey, PA., Buck, LB., 2004. **The human olfactory receptor gene family**. *Proc Natl Acad Sci U S A*, **8**:2584-2589.

MHC sequencing consortium, 1999. **Complete sequence and gene map of a human major histocompatibility complex.** The MHC sequencing consortium. *Nature*, **401**:921-923.

Michalak, P., 2008. **Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes**. *Genomics*, **3:** 243-248.

Morison, IM., Ramsay, JP., Spencer, HG., 2005. **A census of mammalian imprinting**. *Trends Genet*, **21**: 457-465.

Nie, H., Crooijmans, RP., Bastiaansen, JW., Megens, HJ., Groenen, MA., 2010. **Regional regulation of transcription in the chicken genome**. *BMC Genomics*, 11:28.

Niimura, Y., Nei, M., 2003. **Evolution of olfactory receptor genes in the human genome**. *Proc Natl Acad Sci U S A*, **100**: 12235-12240.

Osbourn, AE., Field, B., 2009. **Operons**. *Cell. Mol. Life Sci.*, 66:3755–3775.

Price, M.N., Huang, K.H., Arkin, A.P., Alm, E.J., 2005. **Operon formation is driven by co-regulation and not by horizontal gene transfer**. *Genome Res*, **15**: 809-819.

Raghupathy, N., Durand, D., 2009. **Gene cluster statistics with gene families**. *Mol Biol Evol.*, **5**:957-968.

Reams, A.B., Neidle, E. L., 2004. **Gene amplification involves site-specific short homology-independent illegitimate recombination in *Acinetobacter sp.* strain ADP1**. *J Mol Biol*, **338**: 643-656

Rocha, EP., Danchin, A., 2003. **Gene essentiality determines chromosome organisation in bacteria.** *Nucleic Acids Res*, **31**: 6570-6577.

Roy, PJ., Stuart, JM., Lund, J., Kim, SK., 2002. **Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*.** *Nature*, **418**: 975-979.

Šidàk, Z., 1967. Rectangular confidence region for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62:** 626–633.

Singer, GA., Lloyd, AT., Huminiecki, LB., Wolfe, KH., 2005. **Clusters of co-expressed genes in mammalian genomes are conserved by natural selection.** *Mol Biol Evol*, **3**: 767-775.

Spellman, PT., Rubin GM., 2002. **Evidence for large domains of similarly expressed genes in the Drosophila genome.** *J Biol*, **1**: 5.

Stanley, SM., Bailey, TL., Mattick, JS., 2006. **GONOME: measuring correlations between GO terms and genomic positions.** *BMC Bioinformatics*, **7**: 94.

Starz-Gaiano, M., Cho, NK., Forbes, A., Lehmann, R., 2001. **Spatially restricted activity of a Drosophila lipid phosphatase guides migrating germ cells.** *Development*, **6**:983-991.

The Gene Ontology project (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi)

The InterPro Database (http://www.ebi.ac.uk/interpro/)

Tsai, HK., Huang, PY., Kao, CY., Wang, D., 2009. **Co-expression of neighboring genes in the zebrafish (*Danio rerio*) genome**. *Int J Mol Sci.,*10:3658-3670.

Versteeg, R., van Schaik, BD., van Batenburg, MF., Roos, M., Monajemi, R., Caron, H., Bussemaker, HJ., van Kampen, AH., 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.

Walter, L., Hurt, P., Himmelbauer, H., Sudbrak, R., Günther, E., 2002. **Physical mapping of the major histocompatibility complex class II and class III regions of the rat**. *Immunogenetics,* **4**:268-275.

Waterston, RH., et al., 2002. Initial **sequencing and comparative analysis of the mouse genome.** *Nature*, **420**: 520-562.

Woo, YH., Walker, M., Churchill. GA., 2010. **Coordinated expression domains in mammalian genomes**. *PLoS One*;5(**8**):e12158.

Wu, Q., Maniatis, T., 1999. **A striking organization of a large family of human neural cadherin-like cell adhesion genes.** *Cell*, **97**:779-790.

Yanai, I., Mellor, JC., DeLisi, C., 2002. **Identifying functional links between genes using conserved chromosomal proximity.** *Trends Genet*, **18**: 176-179.

Yi, G., Sze, SH., Thon, MR., 2007. **Identifying clusters of functionally related genes in genomes**. *Bioinformatics,* **23**: 1053-1060.

Yu, H., Kim, PM., Sprecher, E., Trifonov, V., Gerstein, M., 2007. **The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.** *PLoS Comput Biol,* **3**: e59.

Zorio, DA., Cheng, NN., Blumenthal, T., Spieth, J., 1994. **Operons as a common form of chromosomal organization in C. elegans.** *Nature*, **372**: 270-272.

Figure legends

Figure 1.
Number of GO terms for each investigated species. GO terms for molecular function (black), biological process (gray) and cellular component (white) are indicated.

Figure 2.
Distribution of the GO terms per gene for *Homo sapiens*.

Figure 3.
Number of GO terms as a function of cytogenetic locus in human chromosome 19.

Figure 4.
Representation of statistically significant GO clusters. Clustered MHC-related GO terms in chromosome 6 and the effect of analysis window size. The chromosome 6 ideogram is shown on the left. The significant GO term clusters are illustrated as black blocks on the right. The blocks are in 10 columns from5 to 50 in intervals of 5 genes. When clusters overlap, they form longer concatenated GO term superclusters. The blocks are drawn from the beginning of the first gene to the end of the last gene for each cluster.

Figure 5.
Comparison of cytobands versus ontologybands in the human genome. The cytobands are shown on the left, and ontologybands in 10 columns for the statistically significant GO terms are indicated on the right for each chromosome. Centromeres are indicated (gray): the darker gray illustrates the exact centromere region, whereas the lighter gray shows the neighboring areas low on genes.

Figure 6.
GO term distribution for mouse genome. For details see Fig. 5 and for the syntenic regions of mouse relative to human with GO clusters see Supplementary Figs. 12 to 32.

Figure 7.
GO term distribution for the *Drosophila melanogaster* genome. For details see Figure 5.

Tables

Table 1. The properties of investigated genomes.

Organism	Genome		Number of GO term annotations				
	No. of gene identifiers	Size (Mbp)	Biological process	Cellular component	Molecular function	Total	Coverage (GO terms/gene)
Homo sapiens	38,699	3070	65,760	51,631	53,844	171,235	4,4
Mus musculus	28,642	2634	96,168	68,559	76,499	241,226	8,4
Arabidopsis thaliana	27,820	135	39,268	33,933	45,045	118,246	4,2
Drosophila melanogaster	14,838	120	29,255	14,012	21,647	64,914	4,4
Caenorhabditis elegans	20,922	100	31,304	13,382	19,092	63,778	3,0
Saccharomyces cerevisiae	6,189	12	2,322	1,798	1,332	5,452	0,9
Escherichia coli K-12	4,279	5	9,335	5,298	15,304	29,937	7,0

Table 2. The 20 most commonly appearing GO terms in the human genome

Ontology	Occurrence	Definition	GO term type[a]	Level
GO:0004872	6,792	Receptor activity	MF	5
GO:0004984	6,369	Olfactory receptor activity	MF	8
GO:0050896	6,322	Response to stimulus	BP	3
GO:0007608	6,261	Sensory perception of smell	BP	8
GO:0005886	5,943	Plasma membrane	CC	6
GO:0016021	4,763	Integral to membrane	CC	8
GO:0006355	4,649	Regulation of transcription, DNA-dependent	BP	10
GO:0008270	4,026	Zinc ion binding	MF	8
GO:0003677	3,767	DNA binding	MF	5
GO:0005634	3,763	Nucleus	CC	9
GO:0005622	3,737	Intracellular	CC	5
GO:0005576	3,571	Extracellular region	CC	3
GO:0046872	3,180	Metal ion binding	MF	6
GO:0006955	2,342	Immune response	BP	4
GO:0005615	2,042	Extracellular space	CC	5
GO:0004252	1,990	Serine-type endopeptidase activity	MF	8
GO:0005509	1,953	Calcium ion binding	MF	7
GO:0045095	1,778	Keratin filament	CC	12
GO:0005792	1,767	Microsome	CC	9
GO:0003700	1,759	Sequence-specific DNA binding transcription factor activity	MF	4
GO:0007186	671	G-protein coupled receptor protein signaling pathway	BP	6

[a]CC, cellular component; BP, biological process and; MF, molecular function

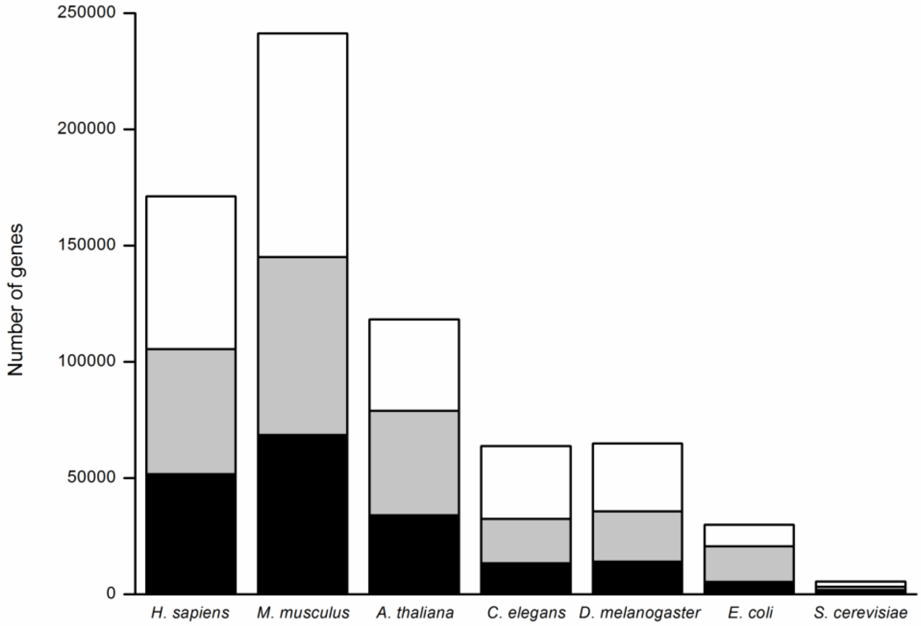
Number of genes
250000
200000
150000
100000
50000
0
H. sapiens
M. musculus
A. thaliana
C. elegans
D. melanogaster
E. coli
S. cerevisiae

Fig 1.

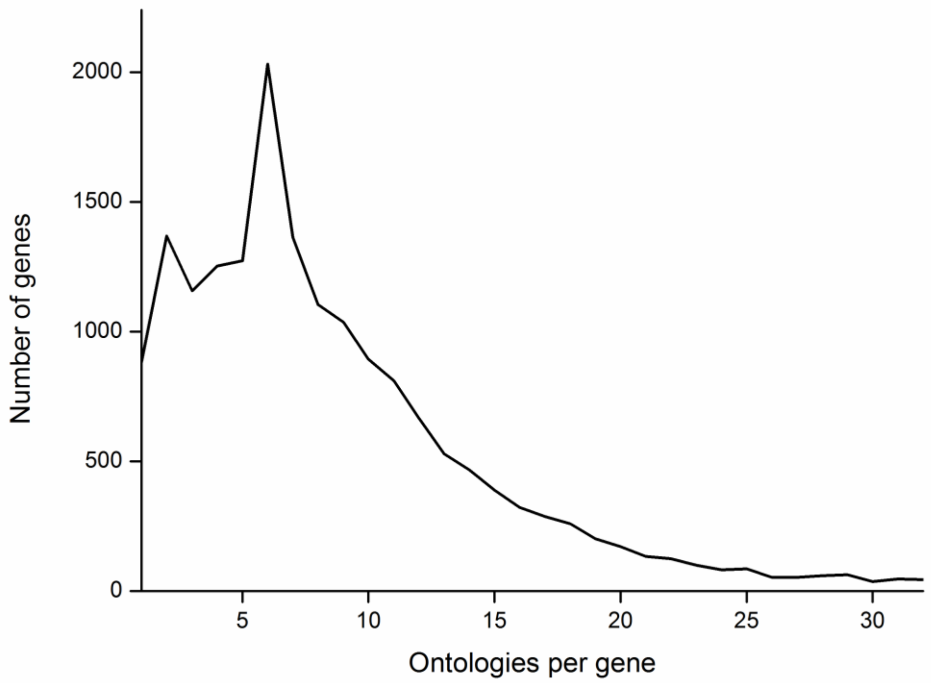
Number of genes
2000
1500
1000
500
0
5
10
15
20
25
30
Ontologies per gene

Fig. 2.

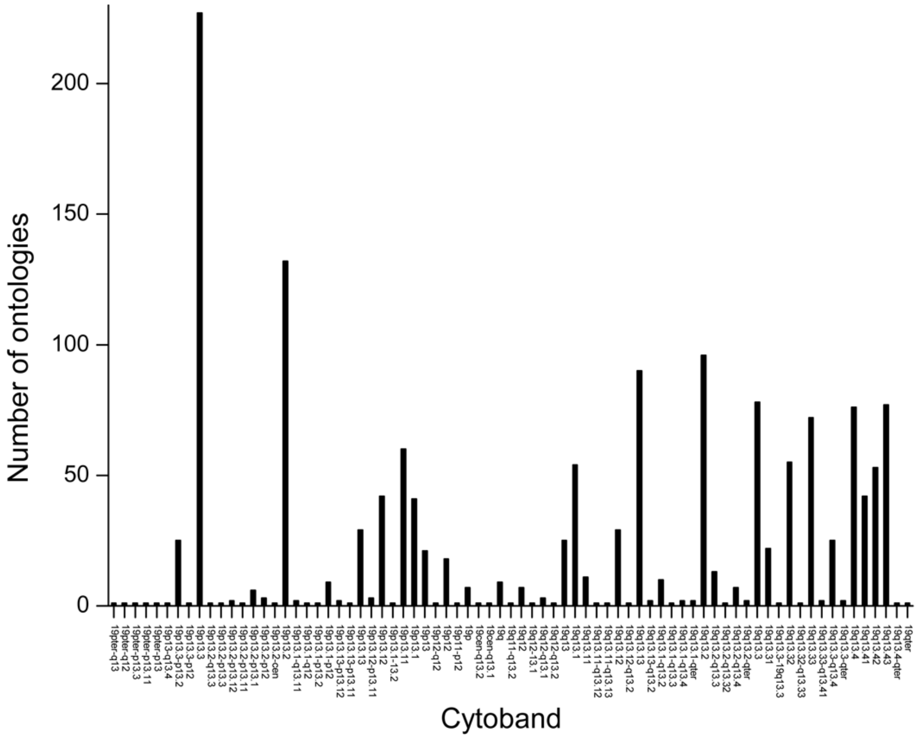
Number of ontologies
200
150
100
50
0
Cytoband
19pter-q13
19pter-q12
19pter-p13.3
19pter-p13.11
19pter-p13
19p13-q13.4
19p13.3-p13.2
19p13.3-p12
19p13.3
19p13.2-q13.3
19p13.2-p13.3
19p13.2-p13.12
19p13.2-p13.11
19p13.2-p13.1
19p13.2-p12
19p13.2-cen
19p13.2
19p13.1-q13.11
19p13.1-q12
19p13.1-p13.2
19p13.1-p12
19p13.13-p13.12
19p13.13-p13.11
19p13.13
19p13.12-p13.11
19p13.12
19p13.1-13.2
19p13.11
19p13.1
19p13
19p12-q12
19p12
19p11-p12
19p
19cen-q13.2
19cen-q13.1
19q
19q11-q13.2
19q12
19q12-13.1
19q12-q13.1
19q12-q13.2
19q13
19q13.1
19q13.11
19q13.11-q13.12
19q13.11-q13.13
19q13.12
19q13.12-q13.2
19q13.13
19q13.13-q13.2
19q13.1-q13.2
19q13.1-q13.3
19q13.1-q13.4
19q13.1-qter
19q13.2
19q13.2-q13.3
19q13.2-q13.32
19q13.2-q13.4
19q13.2-qter
19q13.3
19q13.31
19q13.3-19q13.3
19q13.32
19q13.32-q13.33
19q13.33
19q13.33-q13.41
19q13.3-q13.4
19q13.3-qter
19q13.4
19q13.41
19q13.42
19q13.43
19q13.4-qter
19qter

Fig. 3.

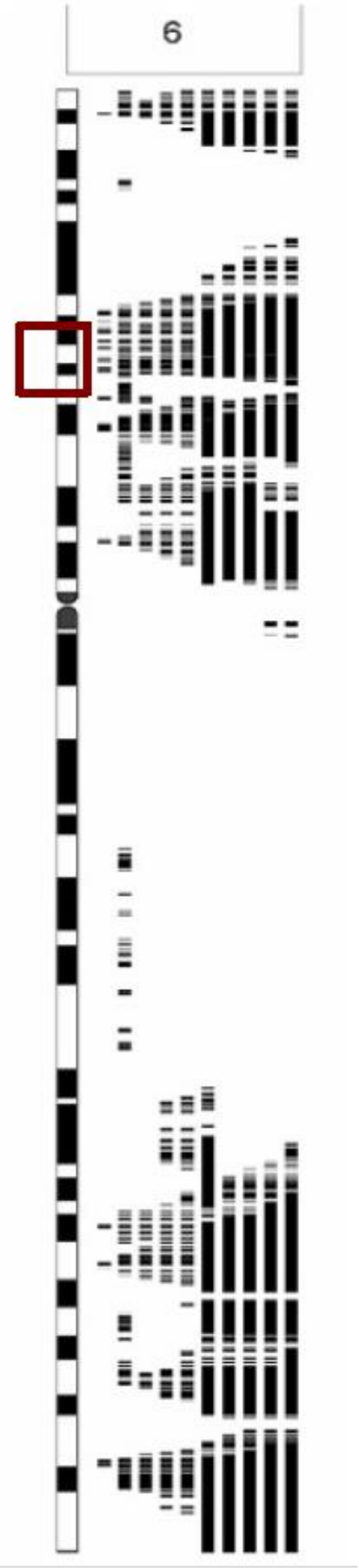

Fig. 4

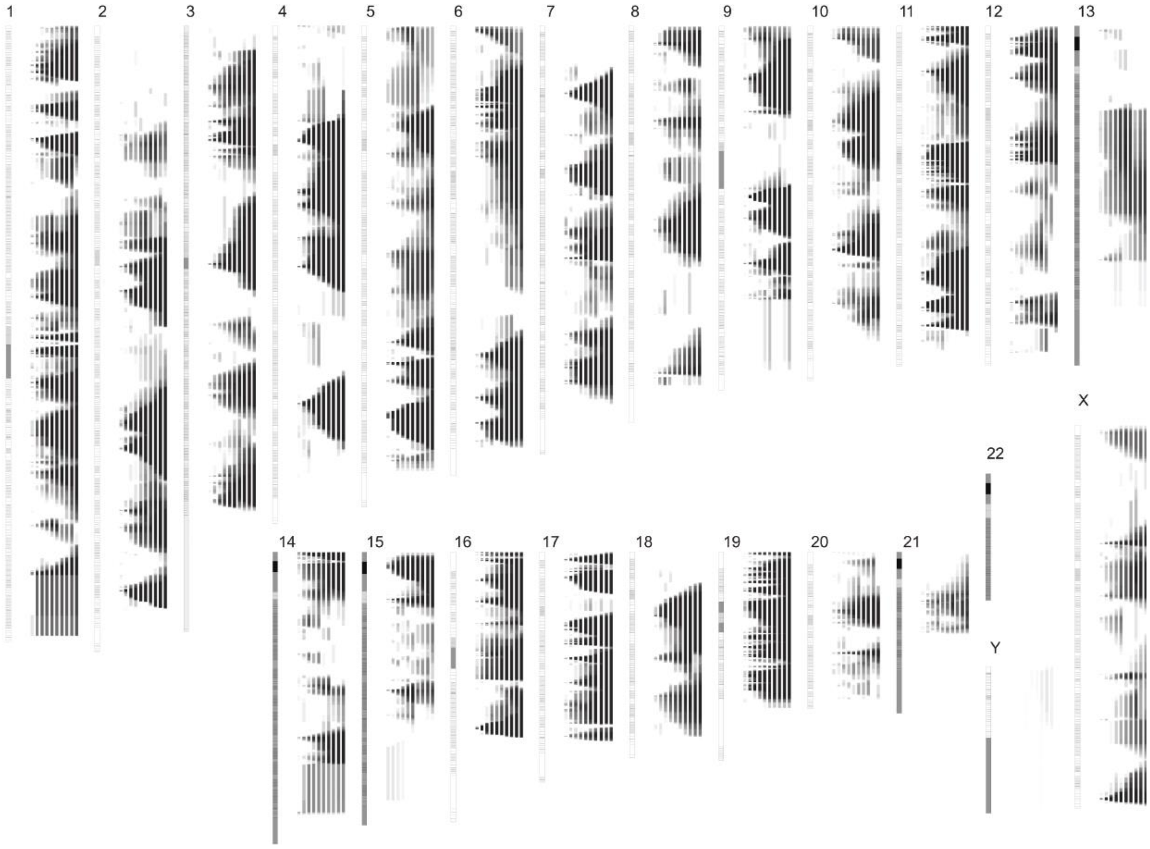

Fig. 5

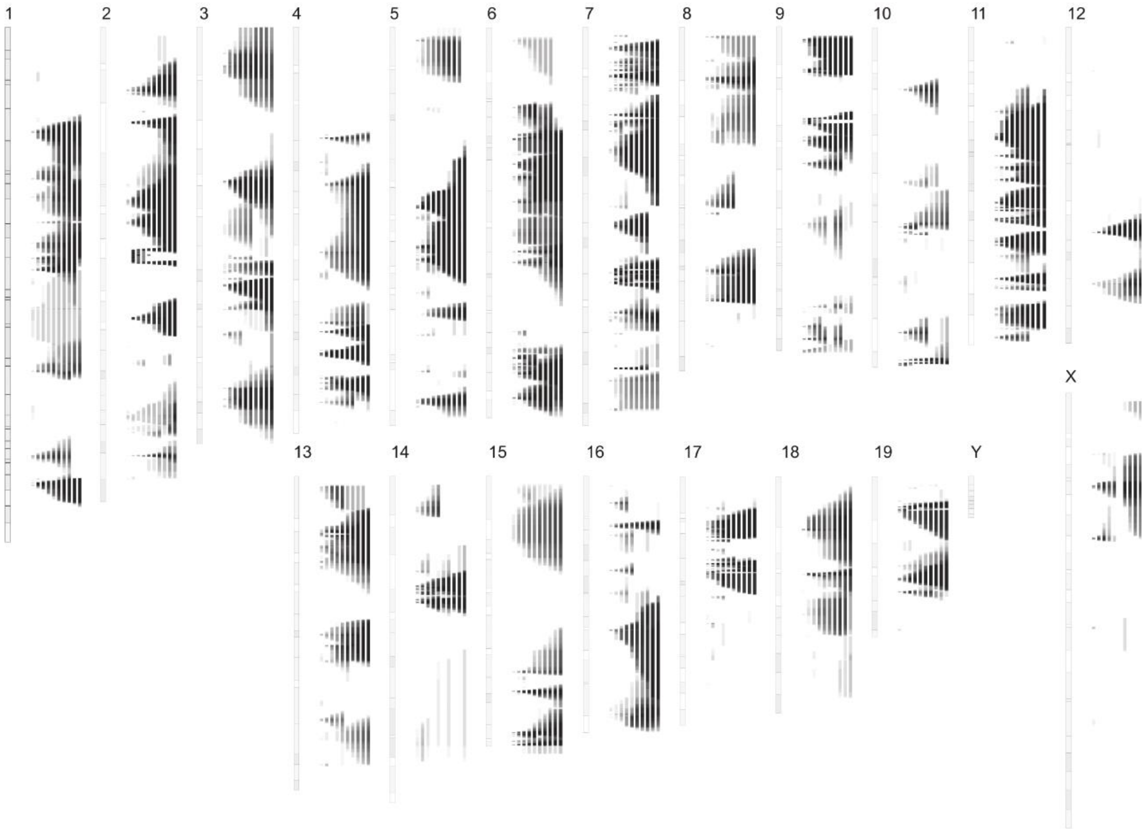

Fig. 6

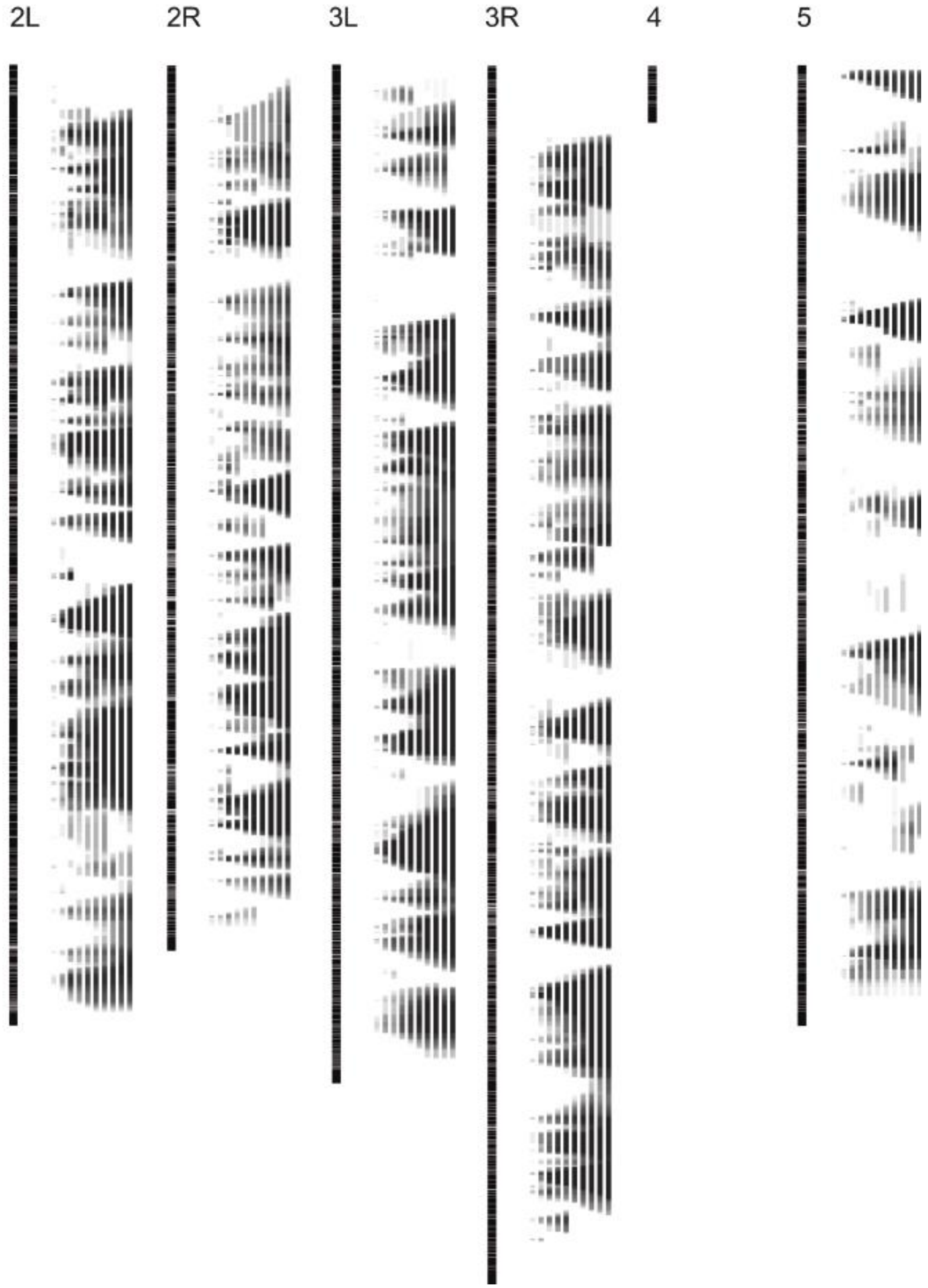

Fig. 7.

Additional files

There are altogether 15 additional files included in this article. Supplementary Results and Discussion contains information for four model species.

Supplementary table 1 shows the 30 most clustered GO terms per window in *Homo sapiens.* The most clustered GO terms mean terms that form the majority of the clusters.

Supplementary table 2 lists the 30 most commonly appearing GO terms in the studied species. Most common GO terms mean terms that have the highest abundance in the genome.

Supplementary table 3 describes the 30 most statistically significant clustered GO terms according to their p-values in *Homo sapiens.*

Other investigated species and their top 30 p-value sorted gene clusters are presented in Supplementary table 4.

The Supplementary tables 5a – 5f show the most clustered GO terms in *Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Saccharomyces cerevisiae* and in *Escherichia coli,* respectively.

Supplementary table 6 describes the results from randomized study using windows 5, 15 and 25 in *Homo sapiens*. The results are sorted by p-value or by frequency.

Supplementary figure 8 shows the GO term distribution for the *Caenorhabditis elegans* genome. For details see Figure 5.

Supplementary figure 9 shows the GO term distribution for the *Arabidopsis thaliana* genome. For details see Figure 5.

Supplementary figure 10 shows the GO term distribution for the *Saccharomyces cerevisiae* genome. For details see Figure 5.

Supplementary figure 11 shows the GO term distribution for the *Escherichia coli* genome. For details see Figure 5.