Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome

Annelyse Thévenin[1,3,*], Liat Ein-Dor[2,*], Michal Ozery-Flato[2] and Ron Shamir[3,#]

[1]Genome Informatics, Faculty of Technology and Institute for Bioinformatics, Center for Biotechnology (CeBiTec), Bielefeld University, Germany

[2]IBM Research - Haifa, Mount Carmel, Haifa, 3498825, Israel

[3] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, 69978 Israel

*These authors contributed equally to the paper
[#] corresponding author. email: rshamir@tau.ac.il. Phone:+972-3-6405383. Fax:+972-3-6405384

Keywords: chromosomes organization, gene function, chromosome conformation, spatial organization of the genome

Preliminary version of Nucleic Acids Research doi: 10.1093/nar/gku667 (2014)

Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome

Annelyse Thévenin, Liat Ein-Dor, Michal Ozery-Flato and Ron Shamir

Abstract

Genomes undergo changes in organization as a result of gene duplications, chromosomal rearrangements and local mutations, among other mechanisms. In contrast to prokaryotes, in which genes of a common function are often organized in operons and reside contiguously along the genome, most eukaryotes show much weaker clustering of genes by function, except for few concrete functional groups. We set out to check systematically if there is a relation between gene function and gene organization in the human genome. We test this question for three types of functional groups: pairs of interacting proteins, complexes and pathways. We find a significant concentration of functional groups both in terms of their distance within the same chromosome and in terms of their dispersal over several chromosomes. Moreover, using Hi-C contact map of the tendency of chromosomal segments to appear close in the 3D space of the nucleus, we show that members of the same functional group that reside on distinct chromosomes tend to co-localize in space. The result holds for all three types of functional groups that we tested. Hence, the human genome shows substantial concentration of functional groups within chromosomes and across chromosomes in space.

Introduction

Cellular processes involve multiple types of functional relations between genes, including protein-protein interactions, regulatory relations and co-expression. Substantial research has been carried out regarding the interplay between functionally related genes and their arrangement on the genome. The most dramatic evidence for non-random organization of co-functioning genes is found in prokaryotes, where genes, usually from the same functional family, are often arranged in operons (1, 2) . Genes in an operon reside consecutively along the genome and are governed by a common promoter. In contrast, most studied eukaryotes lack operons, with few exceptions, including nematodes (3) and drosophila, where operons tend to be dicistronic (4) (see (3) for a review).

Various computational studies utilized the availability of whole genome sequences to show that eukaryotic functionally related genes do tend to cluster. Hershberg et al. used network analysis methods to show that adjacent genes are often co-regulated by the same transcription factor (TF) (5). In the same spirit, Janga et al. discovered that the majority of TFs exhibit a strong preference to regulate genes on specific chromosomes (6). Moreover genome-wide studies of expression data in several organisms revealed that genes from the same genomic neighborhood tend to have similar expression (7–9). Tendency of interacting proteins to aggregate on chromosomes was observed in yeast (10, 11). The clustering trend was observed also in pathways, where Lee and Sonnhammer investigated the levels of clustering within pathways in five eukaryotic species, and found that a large fraction of the pathways exhibits significantly higher clustering levels than expected by chance (12). The aforementioned studies along with a handful of others indicate that there is a link between the relative genomic position of genes and their functional relations, though the eukaryotic clusters are usually much less compact than their prokaryotic counterparts (13). This relatively weaker clustering effect may imply that a more complex mechanism underlies gene arrangement in eukaryotes, incorporating a diversity of influences from multiple types of functional relations.

Furthermore, throughout the past decade it has become clear that the spatial arrangement of genes within the nucleus is also non-random (14). Folding and intermingling of chromosomes may result in high

proximity between genes located at distant positions along the genome, including genes from different chromosomes. It was observed that while gene-rich chromosomes in human tend to occupy interior positions in the nucleus, their gene-poor counterparts tend to be peripherally located (15). Several studies have shown that transcription occurs within discrete regions known as transcription factories (16, 17) and nuclear speckles (18) Moreover, evidence for co-expression of spatially proximal genes has been accumulating (19–21).

In this study we develop a general methodology for analyzing the connection between functional gene groups and the linear and spatial arrangement of genes in the human genome. We focus on three types of functional groups: Protein-protein Interactions (PPIs), complexes and pathways. We analyze three different facets of gene arrangement: the tendency of genes from the same group to lie on specific chromosomes, the intra-chromosomal proximity of genes from the same group, and the degree to which genes from the same group tend to lie close to each other in the three dimensional space within the nucleus. Our findings show that functionally related genes tend to co-localize and manifest clustered organization within and across the chromosomes in all three levels.

Materials and Methods

Throughout the paper, we shall use the general term *group* to denote a single functional unit from any type, i.e. a PPI, a complex or a pathway. Note that each type reflects a relation of a different nature: The two members of PPI are in direct physical contact under some conditions, while complex members are simultaneously involved as building blocks in the same physical unit. In contrast, pathways summarize sequences of multiple chemical or signaling reactions, and hence some of their members may not physically interact, co-localize or even simultaneously exist.

Human Data

The Human PPIs, complexes and pathways analyzed in this study were taken from IntAct (22), Corum (23) and KEGG (24) respectively. Basic information about the three types of datasets that were used in the analysis is summarized in Table 1.

Database	Functionality Relation	Number of groups	Group sizes			Total Number of genes involved
			min	median	max	
IntAct	PPIs	27,947	2	2	2	7,669
Corum	Complexes	1,512	2	3	142	2,421
KEGG	Pathways	206	2	49	1,079	4,852

Table 1 Statistics on the group types used.

Chromosomal locations of genes were extracted from NCBI MapView, where only protein coding genes with a unique position were kept. This preliminary filtering resulted in 19,287 genes.
Spatial distances between genes were based on the Hi-C experimental data of human lymphoblastoid cell line GM06990 (25). The 3D similarity matrices normalized by (26) were used.

Removal of Tandem Duplicate Genes

Duplicate genes are expected to have similar functionality by ancestry. Such genes, if generated by tandem duplication, are often located in physical proximity to one other. To avoid clustering effects

resulting from tandemly duplicated genes, we eliminated them as done in previous studies (8, 9, 27), in the following way. First, to identify proteins belonging to the same gene family, an all-against-all BlastP search was performed on all proteins in the genome, and families were defined using the MCL software (28) with default parameters. We then merged consecutive genes of the same gene family. The location of a merged gene is the interval spanning the consecutive genes that it replaced. The resulting set contained 18,029 genes.

Statistical Methodology

In order to investigate whether genes from the same group tend to preferentially lie in proximity to each other, we created several tests, each examining a different form of co-localization. The p-value calculation for each of the tests is performed as follows:

1. Formulate a test statistic that measures the proximity between functionally related genes.
2. Calculate the value of this statistic, v_0, for the real genome.
3. Estimate the probability to observe this value or higher (alternatively, smaller) for random gene order.
    a. Randomly permute the locations of genes to create a genome with random gene order (functional groups are unchanged).
    b. Calculate the statistic value, v, for the resultant genome.
    c. Repeat steps 3.a. and 3.b. n times
    d. Let k be the number of times $v \geq v_0$.
    e. P-value = (k+1)/(n+1)

The random permutations used to create the null model ensure that the genes in the resultant genomes lie only in loci occupied by genes in the real genome, and that the number of genes in each chromosome remains unchanged. Moreover, the gene composition of the functional groups is unaltered. In this way we exclude from our null hypothesis effects that are not related to the gene order itself. We used the Bonferroni correction whenever multiple tests were performed.

Our tests collect information regarding the distribution of values we are interested in, e.g., how many groups are concentrated in k chromosomes, or what is the distribution of distances along the chromosomes – or in 3D space - between genes from the same group. The most natural test statistics are the moments of the distribution. However, sometimes we need a more sensitive test that focuses on the concentration at the tail of the distribution. In the *distribution tail test*, values are measured and partitioned into bins $b_1, b_2, \ldots, b_k$ where the frequency f_i of values in bin b_i is calculated. The test measures the extent of concentration in the first few bins. We seek the minimal number j for which the cumulative frequency $f_1 + \cdots + f_j$ is significantly higher than expected at random (see more details in the **supplementary information**). The p-value was calculated by comparing the real-genome cumulative histogram as in step 3 above, and Bonferroni corrected via multiplying by j.

Results

We set out to test the tendency of genes with a common function to cluster in the genome using three complementary measures (**Figure 1a**): Inter-chromosomally, by measuring the number of chromosomes co-functioning genes are distributed on; intra-chromosomally, using the genomic distances between co-functioning genes, and in 3D space, by measuring the proximity in the nucleus between co-functioning genes. These three approaches are complementary and each addresses a different aspect of the concentration.

Inter-chromosomal dispersal of genes with a common function

We first investigated the tendency of genes from the same group to concentrate on a small number of chromosomes. Abstractly, each functional type (PPI, complex of pathway) is a collection of groups of

genes, where each group shares a common function. For each group, we defined the number of chromosomes involved in the group as the number of different chromosomes containing genes from that group. Our first test function was defined to be the average of this number over all groups of the same type. We generated 10^6 random genomes by randomly permuting the locations of the genes, and calculated a p-value for each of the group types using the procedure described in Statistical Methodology. The results show that for both PPIs and pathways, the groups tend to concentrate on a small number of chromosomes with p-values 0.001 and $\leq 10^{-6}$ respectively. This test yielded no significant results for complexes (p-value 0.08).

In light of these positive findings, we proceeded with a higher resolution examination of gene arrangement into chromosomes, and applied the distribution tail test. Here f_i is defined as the number of groups involving i chromosomes, in order to measure the extent to which genes from the same group tend to concentrate on few chromosomes. The p-value was calculated by comparing the real-genome frequencies to those in 10^6 random genomes. The results show that there is an enrichment of PPIs and complexes involving a single chromosome (p-value = 0.001 and 0.01, respectively), i.e. the number of PPIs and complexes all of whose genes reside on a single chromosome is significantly higher than randomly expected.

In the case of pathways we reveal a similar trend. The number of pathways that are represented on at most c chromosomes is exhibited in **Figure 1b**. For $c \geq 5$, this number is significantly higher than expected at random (p-value = 0.03 after Bonferroni correction). Hence we observe a tendency of pathways to concentrate on fewer chromosomes than expected by chance.

Intra-chromosomal distances of co-functioning genes

In the previous section, we checked whether genes from the same functional group tend to concentrate on fewer chromosomes than expected by chance. In this section we would like to check whether genes from the same group that belong to the same chromosome tend to be closer than expected. In order to measure this clustering tendency, we calculated for each group i the average distance (in bases), d_i, between pairs of genes in group i that reside on the same chromosome. We defined our test statistic to be the mean value of d_i over all groups in the dataset (see **supplementary information** for more details). We used 10^6 random genomes to calculate the p-values based on our statistical methodology, where this time we randomly permuted the locations of genes within each chromosome separately. In this way we accounted for clustering effects due to concentration of genes from the same group on few chromosomes. The results show that the average intra-chromosomal distance between genes from the same complex and pathway is significantly smaller than expected by chance, obtaining p-values of 0.001 and $\leq 10^{-6}$ respectively. No significant result was obtained for PPIs (p-value = 0.15).

Next, we conducted the more sensitive distribution tail test, focusing on groups with short average distances between members. We partitioned the gene groups into 20 bins based on the distances d_i defined above, as follows. The distances were sorted, and thresholds $0 = t_0 < t_1 < \cdots < t_{20}$ were set such that 5% of the distances were between t_{i-1} and t_i (see **supplementary information** for more details). We then used these thresholds to bin distances for each of 10^5 random genomes defined as above. We used a lexicographic order of bin frequencies to refine the results (see **supplementary information**). We discovered that for all group types, a statistically significant number of groups tend to cluster within smaller distances than expected at random (all three with p-values $\leq 10^{-5}$, Bonferroni corrected). This tendency is illustrated in **Figure 2**. The cumulative distributions of the true genome are plotted along with their random counterparts obtained by averaging over the 10^5 random histograms. The figure shows the enrichment of short distances for each of the three functional group types.

Spatial arrangement of co-functioning genes

In this section we analyze the spatial proximity of functionally related genes in the nucleus. The spatial distances that we used were based on contact map data generated by Lieberman-Aiden et al. using the Hi-C technology (25) and renormalized by Yaffe and Tanay (26) (see **supplementary information** for details). The contact map gives the frequency of observing each two genomic segments next to each other in the experiment. Segment sizes were 1Mb. As in (25), correlation between the frequency vectors of segments was used to measure proximity, and we use 1-correlation as a measure akin to 3D distance.

We first applied the distribution tail test as follows. For each group we computed the average distance between pairs of genes in the group, irrespectively of whether they reside on the same chromosome or on different chromosomes. The results show that for all three types of groups, functional gene groups exhibit more spatial concentration than expected at random (p-value $= 10^{-4}$, calculated from 10^{-5} simulations). However, the high correlation between linear intra-chromosomal distances (as measured in base pairs) and the corresponding 3D distances (see **Figure S1** in **supplementary information**) raises the question whether the apparent 3D concentration is merely a result of the linear intra-chromosomal concentration observed earlier. In order to test for spatial concentration effects that are not related to linear gene proximity, we *considered only inter-chromosomal gene pairs*, excluding all intra-chromosomal distances. To respect the chromosomal organization of each group, we again randomly permuted the locations of genes within each chromosome separately. So, for each group the number of pairs of genes along different chromosomes stays the same in simulated genomes. The results show that the average inter-chromosomal 3D distances between genes from the same pathway are significantly smaller than expected by chance (p-value = 0.009; complexes p-value = 0.09, PPIs p-value= 0.7). Applying the distribution tail test with 20 bins resulted in significant over-population of the first bin for PPIs only (p-value = 0.004). For complexes and pathways, p-values were 0.06 and 0.33 respectively.

The distribution tail test used above checks whether the *average* inter-chromosomal distances between genes pairs within a group is significantly smaller than expected at random. However, if proximity tendency exists only between specific genes within a group, it may be undetected after averaging all pairs in the group. This could explain the fact that the test was significant for PPIs but not for the other types, which have larger groups. To examine whether such tendency exists, we applied the distribution tail test again, but this time we did not average over the distances in each group, but used the individual distances between gene pairs. We computed the distance between each pair of genes from the same group that reside on different chromosomes, binned the values obtained from the entire set of such pairs into 20 bins, and tested the concentration at the distribution tail. We found that for all three group types, namely pathways, PPIs and complexes, gene pairs from the same group tend to cluster within a small spatial region even when they lie on different chromosomes. For the three resultant distributions, the first bin in the real genome (5% of pairs with highest spatial inter-chromosomal proximity) was significantly more populated than the same bin in the random genomes, with p-values 0.004, 10^{-4} and 0.02 for PPIs, complexes and pathways respectively. This result reflects the clustering tendency of genes from the same groups. For PPIs and complexes, the cumulative distribution of the histogram tail remained statistically significant also beyond the first bin. For PPIs, more than 25% of the pairs displayed strong clustering tendency (e.g., for the sum of frequencies in bins 1-6, the obtained p-value was ≤ 0.03). An even more pronounced effect was found for complexes, where about 90% of the pairs, populating bins 1-18 in the cumulative histogram, had all p-values below 0.02. These results, normalized by dividing by the real genome values for the sake of better visibility, are illustrated in **Figure 3.**

All the tests that we conducted aimed to deduct concentration by looking together at the signal from all groups of co-functioning genes of the same types. As such, they provide answers regarding the general phenomenon of concentration. In addition, our analysis can also be applied to study the concentration of individual groups, which may be of independent interest. **Supplementary file 1** contains the full results of each PPI, complex and pathway in each of our three tests. We also tested functional categories of complexes for concentration but obtained no significant results (see **supplementary information**).

Discussion

We have observed that co-functioning genes manifest significant concentration in terms of their organization in the human genome. This holds separately for three types of sets of co-functioning genes (**Table 1**): gene pairs corresponding to interacting proteins, genes whose proteins belong to the same complex, and genes whose proteins take part in the same pathway. The concentration of co-functioning genes is established in three independent ways (**Figure 1a**). Co-functioning genes tend to reside on fewer chromosomes than expected by chance. When they are on the same chromosome, they are positioned more closely to each other than randomly selected genes. Moreover, co-functioning genes on different chromosomes tend to be closer to each other in the three-dimensional nuclear space, based on chromosome conformation capture (3C (29) or Hi-C) data.

These tendencies are statistically significant, based on a cumulative signal collected from many groups of co-functioning genes. The distribution of the test statistic (e.g., the spatial distance or the number of involved chromosomes) in the known functional groups is compared to randomly permuted genes (within and/or across chromosomes, where appropriate). In some cases a simple test statistic, like the distribution mean, suffices to determine significance. In others, we tailored a test statistic to focus on the tail of the distribution, corresponding to the closest pairs.

Why are co-functioning genes concentrated? The broadly accepted explanation has to do with co-transcription. Co-location of genes of common function can facilitate direct cis-regulation of several genes simultaneously, at the level of transcription factors and co-factors, and on the nucleosome and other epigenetic levels. Caron et al. observed clustering of highly expressed genes on intervals along the chromosomes in a variety of human tissues (30). Lercher et al. later observed the same phenomenon for housekeeping genes, and in fact argued that the findings of Caron et al. are due primarily to housekeeping genes (9). In lower eukaryotes, by focusing on two or at most three consecutive genes, the co-regulation of adjacent divergent transcriptional units was shown to be prevalent in yeast (5). Taking an evolutionary perspective, Veron et al. analyzed chromosomal rearrangements between mouse and human and showed significant correlation between intra-chromosomal 3D proximity in the human genome and breakpoint pairs, suggesting the functional relevance of the structure (31). Another evolutionary analysis was recently provided by Dai et al., using gene orders in 17 yeast species (32). The authors showed that gene pairs that are adjacent in other yeast species but reside on different chromosomes in *S. cerevisiae* tend to show stronger nuclear co-localization, as measured in (33). Moreover, these co-localized pairs tend to be regulated by the same transcription factors and by the same histone modifications. Hence co-localization is correlated with co-regulation even after separation due to recombination.

The connection between spatial organization within chromosomes and gene expression was attributed to active chromatin hubs (34), nuclear speckles (18) and more generally to transcription factories. These are discrete nuclear regions in which multiple RNA polymerases are active (35). However, evidence for and against the existence of transcription factories is still debated (36). Li et al. recently studied extensively chromatin interactions in human cell lines and observed promoter-promoter interactions, to the extent that they proposed a chromatin-based operon-like mechanism ("chroperon") for gene regulation in eukaryotic cells (37). Co-expression and proximity in space were shown to be associated both in studies focusing on a few genes using FISH and microscopy, and, more recently, in genome-wide studies of promoter-enhancer associations (38, 39). One novel perspective that we add to this area is the inter-chromosomal dispersion: We show that genes of a common complex or pathway tend to be dispersed on fewer chromosomes than expected by chance. With such clustering, co-regulation of the co-functioning is conceivably better than when the dispersal is completely random.

Is it possible that we see co-clustering of members of the same functional gene groups only because they have similar expression levels? Put differently, is the primary phenomenon co-expression of genes of the same functional group, and the co-localization is only its secondary effect? While it is hard to say which effect is primary, co-expression and common function clearly both affect co-localization. There

are, however, several advantages to analysis based on common function over analysis based on co-expression: (a) Sharing the same functional group is a "cleaner" and more universal property than co-expression, which is measured on condition-dependent datasets. (b) Co-expression quantification is based on measurements of expression, which are noisy. (In fact, co-expression may even be the result of biological noise, cf. (40)). Moreover, gene transcription, the main source of co-expression measurements today, is only moderately correlated with protein transcription, where the function is manifest (41, 42). Moreover, gene transcription, the main source of co-expression measurements today, is only moderately correlated with protein transcription, where the function is manifest (41, 42) (c) Different pathways or complexes may show co-expression under some conditions even if they have completely different functions. (d) Since many pathways summarize a temporal sequence of events and interactions, different segments of the pathway may be active at different times and in that case will not show co-expression, even though they belong to the same functional unit. (e) The definition of a group of co-expressed genes can vary depending on the correlation function, the correlation threshold, the normalization methods etc. On the other hand, functional groups are defined based on a holistic understanding of the underlying biology. (f) Our analysis enables us to examine different types of co-functioning groups, and discern differences among them, which is impossible using co-expression. Further study is required to show which effect is more primary, or that perhaps both co-expression and co-localization are artifacts of yet another more basic, global effect.

The study of nuclear organization has undergone a revolution over the last decade, with the combined contribution of microscopy techniques, chromosome conformation and epigenomics. Seminal studies have established chromatin proximity maps in human (25), baker's yeast (33), fission yeast (43), drosophila (44) and mouse (45), among others. Very recently, an interesting paper by Ben-Elazar et al. (46) studied localization of co-regulated genes in *S. cerevisiae* using 4C data (33). Focusing on the set of targets of each transcription factor, the authors showed that for about half the transcription factors, the concentration of these targets in space exceeds their linear clustering along the chromosomes. This adds support to the transcription factories paradigm. Note, however, that the statistical test for 3D concentration does not distinguish between targets that are on the same chromosome and those on different chromosomes. In fact, the intra-chromosome contact level observed in 3C maps exceeds the inter-chromosomal level by orders of magnitude (25, 33), and therefore it is highly likely that the effect observed by Ben-Elazar et al. is based overwhelmingly on intra-chromosomal contacts. In contrast, our test (**Figure 3**) separates completely the inter- and intra-chromosomal signals, and shows directly inter-chromosomal spatial proximity among co-functioning genes. Another recent paper (47) analyzed the same yeast contact map data (33) together with gene expression data. Using a large panel of expression profiles, and focusing on inter-chromosomal gene pairs only, Homouz and Kudlicki showed that the measured expression levels of nearby genes are significantly correlated. Moreover, they showed that many of the high level gene ontology groups (GO-slim groups) significantly show more 3D contacts between gene pairs than random gene groups of the same sizes. This test is similar in spirit to the one we have performed here. Note, however, that we took care to create gene sets by randomly permuting genes within each chromosome independently, thereby avoiding possible bias due to uneven distribution of group genes across chromosomes. Another salient difference between our study and these two reports is that our test was conducted on human DNA, for which the contact map resolution is much lower than for yeast, and hence detecting the concentration signal is more challenging. To the best of our knowledge, this is the first study of this kind on the human chromosomal conformation data.

In **Figure 3**, the normalized cumulative distributions for randomized genomes remain below the real genome plot for PPIs and complexes, showing spatial concentration of co-functioning genes. For short distances this is the case for pathways as well, but for longer distances the real genome apparently has fewer pairs than the randomized genomes. A possible tempting explanation may be due to the different nature of the functional groups. PPIs and complexes consist of proteins that are simultaneously interacting, and thus their co-location (and consequently co-expression) may be an advantage. In contrast, for pathways a time dimension is involved (e.g. when a sequence of metabolic or signaling reactions is performed), and therefore not all the building blocks of the pathway may be active simultaneously. In a large pathway it may be favorable for subunits that act together to be closer in space, and subgroups that act at different times to be well separated in space. Since many of the

pathways that we have analyzed are rather large (median size 49), they may contain such non-simultaneous blocks that may give rise to their distinct distribution. This hypothesis requires further analysis and testing.

Gene clusters and tandem gene duplications can affect our intra-chromosomal statistics. In order to remove such effects, we removed known gene clusters and also filtered tandem duplicated genes as done previously (8, 9, 27). It is possible though that part of the effect that we observe is a result of unknown clusters or remaining duplicated genes that do not appear in tandem. However, this would not explain the inter-chromosomal spatial concentration.

Finally, the test methodology that we developed here can be useful for studying other questions as well. It provides a uniform approach for comparison of true and random distributions for a broad variety of test statistics.

Funding

This study was supported in part by an IBM Research Open Collaborative Research grant on clinical genome analysis. RS was supported in part by the Israel Science Foundation [grant 317/13]; the Raymond and Beverly Sackler chair in bioinformatics; and I-CORE Program of the Planning and Budgeting Committee and The Israel Science Foundation [grant No 41/11]. AT was supported in part by a post-doctoral fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and by a fellowship from the Alexander von Humboldt Foundation at Bielefeld University.

Acknowledgements

We thank Shai Izraeli, Omer Schwartzman, Manolo Gouy and Osnat Penn for helpful discussions on the manuscript. We thank Laurence Hurst for important comments and references.

References

1. JACOB, F. and MONOD, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
2. Malke, H. (1981) J. H. Miller and W. S. Reznikoff (Editors), The Operon (2nd Edition). VII, 469 S., 128 Abb., 36 Tab. Cold Spring Harbor 1980. Cold Spring Harbor Laboratory. *Zeitschrift für allgemeine Mikrobiologie*, **21**, 697–697.
3. Blumenthal, T., Evans, D., Link, C.D., Guffanti, A., Lawson, D., Thierry-Mieg, J., Thierry-Mieg, D., Chiu, W.L., Duke, K., Kiraly, M., et al. (2002) A global analysis of Caenorhabditis elegans operons. *Nature*, **417**, 851–854.
4. Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. (2002) Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol*, **3**, research0083.1–83.22.
5. Hershberg, R., Yeger-Lotem, E. and Margalit, H. (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet.*, **21**, 138–142.
6. Janga, S.C., Collado-Vides, J. and Babu, M.M. (2008) Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 15761–15766.
7. Woo, Y.H., Walker, M. and Churchill, G.A. (2010) Coordinated Expression Domains in

Mammalian Genomes. *PLoS ONE*, **5**, e12158.
8. Williams, E.J.B. and Bowles, D.J. (2004) Coexpression of Neighboring Genes in the Genome of Arabidopsis thaliana. *Genome Res*, **14**, 1060–1067.
9. 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.
10. 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.
11. Poyatos, J.F. and Hurst, L.D. (2006) Is optimal gene order impossible? *Trends Genet.*, **22**, 420–423.
12. Lee, J.M. and Sonnhammer, E.L.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.*, **13**, 875–882.
13. Hurst, L.D., Pál, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, **5**, 299–310.
14. Rajapakse, I. and Groudine, M. (2011) On emerging nuclear order. *J Cell Biol*, **192**, 711–721.
15. Reddy, T.E., Pauli, F., Sprouse, R.O., Neff, N.F., Newberry, K.M., Garabedian, M.J. and Myers, R.M. (2009) Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.*, **19**, 2163–2171.
16. Wansink, D.G., Schul, W., van der Kraan, I., van Steensel, B., van Driel, R. and de Jong, L. (1993) Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J. Cell Biol.*, **122**, 283–293.
17. Jackson, D.A., Hassan, A.B., Errington, R.J. and Cook, P.R. (1993) Visualization of focal sites of transcription within human nuclei. *EMBO J*, **12**, 1059–1065.
18. Spector, D.L. and Lamond, A.I. (2011) Nuclear speckles. *Cold Spring Harb Perspect Biol*, **3**.
19. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., et al. (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.*, **42**, 53–61.
20. Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M.S. and Mhlanga, M.M. (2013) Chromosomal Contact Permits Transcription between Coregulated Genes. *Cell*, **155**, 606–620.
21. Rieder, D., Ploner, C., Krogsdam, A.M., Stocker, G., Fischer, M., Scheideler, M., Dani, C., Amri, E.-Z., Müller, W.G., McNally, J.G., et al. (2014) Co-expressed genes prepositioned in spatial neighborhoods stochastically associate with SC35 speckles and RNA polymerase II factories. *Cell. Mol. Life Sci.*, **71**, 1741–1759.
22. Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., et al. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–531.
23. Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C. and Mewes, H.-W. (2010) CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.*, **38**, D497–501.
24. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.*, **28**, 27–30.
25. Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–293.
26. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates

systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
27. Singer, G.A.C., Lloyd, A.T., Huminiecki, L.B. and Wolfe, K.H. (2005) Clusters of Co-expressed Genes in Mammalian Genomes Are Conserved by Natural Selection. *Mol Biol Evol*, **22**, 767–775.
28. Van Dongen, S. (2000) Graph Clustering by Flow Simulation.
29. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
30. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voûte, P.A., et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
31. Véron, A.S., Lemaitre, C., Gautier, C., Lacroix, V. and Sagot, M.-F. (2011) Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. *BMC Genomics*, **12**, 303.
32. Dai, Z., Xiong, Y. and Dai, X. (2014) Neighboring genes show inter-chromosomal colocalization after their separation. *Mol Biol Evol*, 10.1093/molbev/msu065.
33. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A. and Noble, W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
34. De Laat, W. and Grosveld, F. (2003) Spatial organization of gene expression: the active chromatin hub. *Chromosome Res.*, **11**, 447–459.
35. Iborra, F.J., Pombo, A., Jackson, D.A. and Cook, P.R. (1996) Active RNA polymerases are localized within discrete transcription 'factories' in human nuclei. *J Cell Sci*, **109**, 1427–1436.
36. Sutherland, H. and Bickmore, W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457–466.
37. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012) Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, **148**, 84–98.
38. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. and Zhao, K. (2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.*, **22**, 490–503.
39. Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, **504**, 306–310.
40. Wang, G.-Z., Lercher, M.J. and Hurst, L.D. (2011) Transcriptional Coupling of Neighboring Genes and Gene Expression Noise: Evidence that Gene Orientation and Noncoding Transcripts Are Modulators of Noise. *Genome Biol Evol*, **3**, 320–331.
41. De Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst*, **5**, 1512–1526.
42. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
43. Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J.R., Wickramasinghe, P., Lee, M., Fu, Z. and Noma, K. (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.*, **38**, 8164–8177.
44. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-Dimensional Folding and Functional Organization

Principles of the Drosophila Genome. *Cell*, **148**, 458–472.
45. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. and Fraser, P. (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, 59–64.
46. Ben-Elazar, S., Yakhini, Z. and Yanai, I. (2013) Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic Acids Res.*, 10.1093/nar/gks1360.
47. Homouz, D. and Kudlicki, A.S. (2013) The 3D Organization of the Yeast Genome Correlates with Co-Expression and Reflects Functional Relations between Genes. *PLoS ONE*, **8**, e54699.

Figure legends

Figure 1 Distribution criteria and pathway concentration. (a) Criteria for chromosome concentration of co-functioning genes. Circles correspond to genes and the shaded circles are all the members of a group with a common function. i: concentration within a chromosome (intra-chromosomal). Here clustering/concentration is gauged based on pairwise linear distance (in base pairs or in the number of intervening genes) between co-functioning genes. ii: dispersal across chromosomes (inter-chromosomal). Out of the 23 pairs of chromosomes, 18 do not contain the group's genes, so this group is concentrated in few chromosomes. This measure takes into account only the chromosomes on which the co-functioning genes reside and ignores the relative locations within each chromosome. iii: concentration in the 3D space. Curved lines show positions of chromosomal segments in space, with the genes on them. The group is concentrated in space. By identifying the chromosome each segment belongs to, one can distinguish between proximity of inter- and intra-chromosomal gene pairs, and analyze them separately. **(b) Pathway concentration in few chromosomes**. For each number j of chromosomes, the plots show the number of pathways whose genes reside in at most j chromosomes. Plots for the real genome (red curve) and for an average over 10^6 random genomes (blue curve) are shown. The shaded area around the blue curve shows ±1 standard deviation. Inset: Zoom in on the region of a small number of chromosomes.

Figure 2 Intra chromosomal distances. The plot shows the average intra-chromosomal distance between genes from the same group in the real (red) and randomized (blue) genomes. **(a)** PPIs; **(b)** Complexes; **(c)** Pathways. Bins were selected so that the occupancy of pairs from the real genome is uniform (hence the straight red line). The clustering effect is reflected by the larger cumulative fraction in the real genome histograms compared to the random model in the smaller distance bins. The light blue shaded region around the blue curve stands for ±1 standard deviation.

Figure 3 Spatial proximity between inter-chromosomal gene pairs from the same functional group. The cumulative distribution function (cdf) of inter-chromosomal proximity between genes from the same group was computed for real and randomized genomes. The plot shows each cdf divided by the cdf of the real genome. As a result the real genome curve has a constant y-value of 1. Red: real genome, blue: average over 10^5 random genomes. The light blue bands show ±1 standard deviation. **(a)** PPIs; **(b)** Complexes; **(c)** Pathways. The x-axis units are 1 minus the correlation between the normalized Hi-C contact profiles of the regions containing the gene pairs, so that smaller values reflect higher correlation and shorter distances.

"Supplementary Data are available at NAR online: Supplementary table S1, Supplementary figure S1, Supplementary methods and Supplementary references."

Thévenin et al. Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome

Supplementary Information

Data

Human genome

For gene location in the Human genome, we used the July 2011 versions of NCBI [1] files Homo_sapiens.gene_info.txt[2] for the curated genes, and seq_gene.md[3] for the genes' base position. Among the 42,158 entries of Homo_sapiens.gene_info.txt, we selected the set $\mathcal{G}_{GI}$ of 20,329 genes annotated as protein coding and present in a homologous chromosome. Among the 1,534,264 entries of seq_gene.md, we selected the set $\mathcal{G}_{MV}$ of 36,429 annotated as protein coding genes, and according to the GRCh37.p2-Primary assembly, present in a homologous chromosome with only one known base position. Then, we defined the set $\mathcal{G}' \subseteq \mathcal{G}_{GI} \cup \mathcal{G}_{MV}$ of 19,287 genes whose chromosome was the same in $\mathcal{G}_{GI}$ and $\mathcal{G}_{MV}$. In this way, all genes in $\mathcal{G}'$ had a single gene ID in Entrez. Genes from chromosome Y were excluded from the analysis.

Finally, from $\mathcal{G}'$ we obtained a set $\mathcal{G}$ of size 18,029 by merging tandem duplicated genes into a single gene as described in the Methods section.

Co-functioning genes

For PPIs we used the IntAct database (July 2011 version). For pathways, we used KEGG (30 June 2011 version). For complexes, we used CORUM (September 2009 version). In order for a functional gene group to be included in our analysis the following conditions were set:

- The group is unique,
- It has at least two different genes in $\mathcal{G}$,
- At least 95% of the genes in the group are in $\mathcal{G}$,
- The group has at most one gene from each known gene cluster (Hox genes, olfactory receptor genes, Human Leukocyte Antigen (HLA) genes and Hemoglobin genes).

The last constraint removes the influence of large known gene clusters from our analysis.

In IntAct, the genes in each group were given by their UniProt[4] identifier which we converted to gene IDs using the file HUMAN_9606_idmapping.dat[5]. PPIs with at least one protein that had either no geneID match or several matches were ignored.

3D Human genome

Lieberman-Aiden et al. used the Hi-C method to study the three-dimensional architecture of a whole genome by coupling proximity-based ligation with massively parallel sequencing (1). They constructed spatial proximity matrix of the human lymphoblastoid cell line GM06990 genome, based on contact probability between genomic regions at a resolution of 1Mb. In order to reduce biases, we adopted the normalization method proposed by Yaffe and Tanay (2) , which was shown to outperform the original one of Lieberman-Aiden et al.. A later method of Imakaev et al. (3) was reported to yield essentially the same matrix of biases. Because the extremities of chromosomes are not included in the proximity

[1] http://www.ncbi.nlm.nih.gov/
[2] ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/
[3] ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.37.2/initial_release/
[4] http://www.uniprot.org/
[5] ftp://ftp.uniprot.org/pub/databases/uniprot/previous_release/releases-2011/knowledgebase/idmapping/by_organism/

matrix, some genes do not have spatial measurements. As a result, groups containing less than 2 genes with spatial measurements are ignored during our the 3D proximity tests. This excluded 426 PPIs (among 27,947) and 5 complexes (among 1,512).

Methods

Inter-chromosomal dispersal of genes with a common function - the distribution tail test

Let N_i^G be the number of groups involving exactly i chromosomes in genome G, and let X^G be a vector representing the total number of groups involving at most i chromosomes, namely, the i-th component of X^G is given by $X_i^G = \sum_{j=1}^{i} N_j^G$. Denote the number of chromosomes by N. To calculate the relative fraction of random distributions that are at least as concentrated as the real one, we use the following definition: Given two frequency vectors X^G and $X^{G\prime}$, we define X^G to be more concentrated than $X^{G\prime}$ starting from bin i, iff X^G is lexicographically larger than $X^{G\prime}$ starting from bin i, i.e., there exists m, $i \leq m \leq N$, such that $X_m^G > X_m^{G\prime}$ and $\forall\, i \leq k \leq m\ X_k^G = X_k^{G\prime}$. Clearly, the two histograms are equally concentrated if and only if $X_k^G = X_k^{G\prime}\ \forall\, 1 \leq k \leq N$.

Testing intra-chromosomal distances between co-functioning genes

Define the distance between two genes on the same chromosome as the number of base pairs between the last base of the first gene and the first base of the second gene. Let D_ω be the average distance between all pairs of genes from the same chromosome involved in group ω. (Note that the average can include different pairs from different chromosomes).

- **The average test:** The average test function is defined to be the average of D_ω over all groups ω of the same type.
- **The distribution tail test:** The distances D_ω for the real genome were sorted, and thresholds $0 = t_0 < t_1 < \cdots < t_{20}$ were set such that 5% of the distances were between t_{i-1} and t_i, i.e., in bin i. More precisely, when the number of distances N_G did not allow the bins to be evenly populated, the first 19 bins were populated with $\lfloor N_G/20 \rfloor$ groups each, and the last bin with the remaining groups. Moreover, cases where multiple groups had the same average distance as the bin threshold often resulted in slight differences between the bin populations. For genome G, let Y_i^G be the number of distances that fall in bin i. By generating many randomized genomes and recording their Y_i^G vectors, the significance of the concentration of the values within the first few bins in the real genome can be evaluated. Starting from the first bin we perform a sequence of tests with the test functions Y_i^G to find the first i for which a significant p-value (after Bonferroni correction) is obtained.

Intra-chromosomal distances along the genome and in space

We wanted to test the correlation between spatial and linear intra-chromosomal distances. For the spatial normalized matrix, we computed the Pearson correlation of each pair of rows. This computation yielded, for each pair of 1Mb chromosomal regions r_1 and r_2, a value $m(r_1, r_2)$ between -1 and 1. For every pair of genes, g_1 and g_2, residing on regions r_1 and r_2 respectively, we defined the distance between them to be $d(g_1, g_2) = 1 - m(r_1, r_2)$ (if a gene is spread over several different regions, we take the weighted average of m according the number of bases present in each region). The 265 genes located on regions that were not included in the spatial proximity matrix from (1) were removed from our analysis. **Figure S1** shows the high correlation between the spatial and the linear measures, at short intra-chromosomal distances**.**

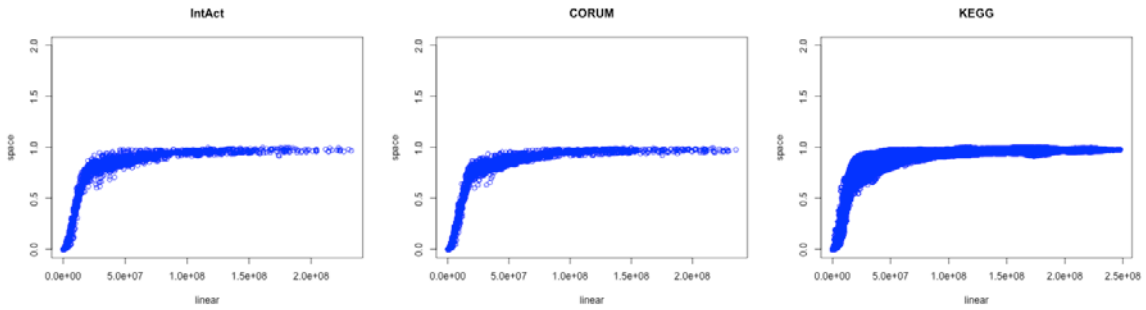

Figure S1 Relation between intra-chromosomal linear distance (number of bases) between pairs of genes and the spatial proximity between them (represented by 1-correlation), as measured in (1, 2).

Testing the association between functional categories of complexes and inter-chromosomal 3D distances

We conducted the following additional test to see if certain functional categories of complexes are significantly concentrated in 3D.

1. Filter out any gene that is not covered by the Hi-C data. Filter out complexes whose number of genes is zero or one after genes filtering. [5 complexes were filtered out due to missing 3D data.]

2. Order all inter-chromosomal gene-pairs that are found within complexes by their proximity in 3D (overall there were 31648 such pairs)

3. For a tested functional category of complexes, derive the set of inter-chromosomal gene pairs appearing in its complexes. Call this set S.

4. Test whether S is enriched with low/high 3D distances by a GSEA test (4, 5) and obtain an enrichment score ES(S). To compute its significance, apply the following procedure:
    i. Randomly permute all genes within their chromosomes
    ii. Rank inter-chromosomal gene pairs by the 3D-proximity values corresponding to their new locations
    iii. Recompute the enrichment score of S on the resulting set.

Report the empirical p-value of ES(S) obtained by 1000 runs of the procedure

We used MIPS FunCat for functional annotation scheme of CORUM complexes, covering 1512 complexes. We applied the test described above to the 23 classes in the highest level in this annotation hierarchy. Table S1 below presents the results of the test. None of the tested functional categories was found to be significantly enriched with lower 3D values after Bonferroni correction for multiple testing.

FunCat group no.	FunCat desc	no. complexes	no. interchr. pairs	GSEA pval
1	METABOLISM	65	534	0,847
2	ENERGY	4	148	0,108
10	CELL CYCLE AND DNA PROCESSING	404	5202	0,275
11	TRANSCRIPTION	474	17849	0,574
12	PROTEIN SYNTHESIS	27	9745	0,670
14	PROTEIN FATE (folding, modification, destination)	352	4638	0,311
16	PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)	436	16337	0,314
18	REGULATION OF METABOLISM AND PROTEIN FUNCTION	153	2066	0,261
20	CELLULAR TRANSPORT, TRANSPORT FACILITATION AND TRANSPORT ROUTES	136	776	0,385
30	CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	349	2718	0,277
32	CELL RESCUE, DEFENSE AND VIRULENCE	117	1109	0,162
34	INTERACTION WITH THE ENVIRONMENT	118	335	0,210
36	SYSTEMIC INTERACTION WITH THE ENVIRONMENT	33	174	0,520
40	CELL FATE	101	1153	0,833
41	DEVELOPMENT (Systemic)	60	255	0,513
42	BIOGENESIS OF CELLULAR COMPONENTS	230	3235	0,853
43	CELL TYPE DIFFERENTIATION	29	237	0,416
45	TISSUE DIFFERENTIATION	24	113	0,008
47	ORGAN DIFFERENTIATION	31	136	0,041
70	SUBCELLULAR LOCALIZATION	843	30662	0,224
73	CELL TYPE LOCALIZATION	47	256	0,452
75	TISSUE LOCALIZATION	45	270	0,135
77	ORGAN LOCALIZATION	45	255	0,284

Table S1: Significance results for enrichment of MIPS functional categories in inter-chromosomal 3D distances

1. Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–293.

2. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

3. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J. and Mirny, L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

4. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267–273.

5. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, **102**, 15545–15550.