Computational Biology and Chemistry 53 (2014) 71–78

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

# DNA clustering and genome complexity

Francisco Dios[a,b], Guillermo Barturen[a,b], Ricardo Lebrón[a,b], Antonio Rueda[c], Michael Hackenberg[a,b], José L. Oliver[a,b,*]

[a] Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain
[b] Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, 18100 Granada, Spain
[c] Plataforma Andaluza de Genómica y Bioinformática (GBPA), Edificio INSUR, Calle Albert Einstein, 41092 Sevilla, Spain

ARTICLE INFO

*Article history:*
Accepted 11 July 2014
Available online 23 August 2014

*Keywords:*
Clustering
Genome elements
Genome complexity
Hierarchical clustering

ABSTRACT

Early global measures of genome complexity (power spectra, the analysis of fluctuations in DNA walks or compositional segmentation) uncovered a high degree of complexity in eukaryotic genome sequences. The main evolutionary mechanisms leading to increases in genome complexity (i.e. gene duplication and transposon proliferation) can all potentially produce increases in DNA clustering. To quantify such clustering and provide a genome-wide description of the formed clusters, we developed *GenomeCluster*, an algorithm able to detect clusters of whatever genome element identified by chromosome coordinates. We obtained a detailed description of clusters for ten categories of human genome elements, including functional (genes, exons, introns), regulatory (CpG islands, TFBSs, enhancers), variant (SNPs) and repeat (Alus, LINE1) elements, as well as DNase hypersensitivity sites. For each category, we located their clusters in the human genome, then quantifying cluster length and composition, and estimated the clustering level as the proportion of clustered genome elements. In average, we found a 27% of elements in clusters, although a considerable variation occurs among different categories. Genes form the lowest number of clusters, but these are the longest ones, both in bp and the average number of components, while the shortest clusters are formed by SNPs. Functional and regulatory elements (genes, CpG islands, TFBSs, enhancers) show the highest clustering level, as compared to DNase sites, repeats (Alus, LINE1) or SNPs. Many of the genome elements we analyzed are known to be composed of clusters of low-level entities. In addition, we found here that the clusters generated by *GenomeCluster* can be in turn clustered into high-level super-clusters. The observation of 'clusters-within-clusters' parallels the 'domains within domains' phenomenon previously detected through global statistical methods in eukaryotic sequences, and reveals a complex human genome landscape dominated by hierarchical clustering.



## 1. Introduction

The increase in genome complexity from prokaryotes to eukaryotes has been mainly driven by a gradual growth in the number of genes, as well as a more sudden growth in the number of introns and transposable genetic elements (Lynch and Conery, 2003). Interestingly, the involved evolutionary mechanisms, such as gene duplication or the increment in the rate of transposon proliferation, can all potentially produce DNA clustering. Gene duplication (Ohno, 1970; Sankoff, 2001) produces at first gene copies clustered in tandem, while at least some mobile elements might initially insert nearly at random but can be later differentially eliminated from some genome regions and accumulated in others (Hackenberg et al., 2005; Jurka et al., 2004). Therefore, a quantitative description of the formed clusters, as well as an estimation of the clustering level shown by distinct categories of genome elements, can increment our understanding of the evolution of genome complexity at the sequence level.

The measurement of genome complexity at the sequence level began as soon as genome sequences of sufficient length were available. Three independent groups (Li and Kaneko, 1992a; Peng et al., 1992; Voss, 1992) applied global statistical methods (i.e. power spectra, analysis of fluctuations in DNA walks) to uncover large-scale genome structure. The emerging view was a complex, patchy genome with long-range, power-law correlations, thus implying that compositional domains should appear at all scales (Bernaola-Galván et al., 1996; Carpena et al., 2007; Li and Kaneko, 1992a,b; Li et al., 1994). Noteworthy, segmenting these complex, long-range correlated sequences leads to the observation of the

* Corresponding author at: Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, 18071-Granada, Spain.
*E-mail address:* oliver@ugr.es (J.L. Oliver).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.011
1476-9271/

'domains-within-domains' phenomenon in eukaryotic, but not in prokaryotic species (Bernaola-Galván et al., 1996; Li and Kaneko, 1992a; Li et al., 1994; Oliver et al., 2004, 2001; Román-Roldán et al., 1998). This phenomenon beautifully mimics the hierarchical nature of biological complexity originated from integrating collections of objects at one level into entities of the next higher hierarchical level: exons clustered within the genes, genes integrated into chromosomes, prokaryotic cells united in an eukaryotic cell, cells combined in multicellular organisms, etc. (Schuster, 1996). At the sequence level, this model works by the integration of nucleotides into sequence motifs (Nussinov et al., 1986; Stormo, 2000) or compositional domains (Bernaola-Galván et al., 1996; Oliver et al., 1999), domains forming isochores (Bernardi et al., 1985), isochores organized into larger chromosome superstructures (Carpena et al., 2011), and so on.

However, with the exception of DNA repeats (Hackenberg et al., 2005; Jurka and Kapitonov, 2007; Jurka and Kohany, 2005; Jurka et al., 2005, 2004; Price et al., 2004; Sellis et al., 2007; Stankiewicz et al., 2004), the chromosome organization of such genome complexity remains largely unexplored. We are interested in characterizing additional, more general based sources of DNA clustering able to contribute to genome complexity. On the basis of the *CpGcluster* algorithm (Hackenberg et al., 2006), we have developed a new program (*GenomeCluster*) able to detect clusters of whatever elements in the genome, provided that chromosome coordinates are available for them. In this way, using the *hg19* assembly of the human genome, we were able to retrieve the clusters not only of repetitive DNA or transposable elements (as Alus or LINE1), but also of functional elements (as genes, exons or introns), regulatory elements (as CpG islands, TFBSs or enhancers), variation sites (as SNPs) or DNase hypersensitivity sites. We explore here the genome-wide clustering of all these elements in the human genome, thus providing a basis to discuss the relation between DNA clustering and genome complexity.

## 2. Materials and methods

### 2.1. Datasets

Chromosome coordinates for most of the human genome elements analyzed here were obtained from the UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables?hgsid=357122457). The *hg19* assembly for the human genome was used throughout. The *refGene* table (Pruitt et al., 2005) was used to retrieve the coordinates of genes, exons and introns. Only the longest transcript of each gene was considered. The coordinates for TFBSs and DNase hypersensitivity sites of the ENCODE project (Bernstein et al., 2012) were retrieved from the tables *wgEncodeRegTfbsClusteredV3* and *wgEncodeRegDNaseClusteredV2*, respectively. Alu and LINE1 coordinates were retrieved from the *rmsk* table of repeat elements predicted by *RepeatMasker* (Jurka et al., 2005). The *All snp137* set (Sherry et al., 2001) was used to retrieve SNP coordinates. CpG island coordinates were those predicted by *CpGcluster*, using the genome intersection as the distance threshold and a *p*-value cutoff of 1E-5 (Hackenberg et al., 2011). The enhancer dataset was derived by selecting ENCODE TFBSs tagged with the Gene Ontology term GO:0003705 (RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity) and carrying the histone signatures characterizing active enhancers (H3K4me1 and H3K27ac) (see Zentner et al., 2011 for details).

### 2.2. The GenomeCluster algorithm

The algorithm has two main steps. First, based on a distance threshold, the individual genome elements below this threshold are clustered. Second, by means of the negative binomial distribution, a *p*-value is associated to each genome cluster. This *p*-value can then be used as a cutoff: low-significant clusters (i.e. above a given *p*-value) can be filtered out.

#### 2.2.1. The genome intersection point as the distance threshold

The clustering analysis of genome elements presented here is inspired by the level statistics of quantum-disordered systems (Carpena et al., 2009). The spatial distribution of distances ($d$) between adjacent occurrences of a particular genome element can be characterized by the spacing distribution $P(d)$. For systems as the energy levels of quantum disordered systems, the corresponding $P(d)$ follows the *Poisson* distribution. However, the random distribution $P(d)$ is *Poissonian* only for continuous distance distributions, which is valid for the energy levels, but not for the genome elements, where the distances are integers. The discrete counterpart of the *Poisson* distribution is the geometric distribution, which was used to compute the theoretical (expected) distance distribution of genome elements.

As shown previously for CpG dinucleotides, the observed and expected distance distributions show an intersection point separating intra-cluster from inter-cluster distances (see Fig. 1 in Hackenberg et al., 2006). It seems reasonable therefore to use this point as a distance threshold to identify the elements belonging to each cluster. In most human chromosomes, the intersection lies near the median of the observed distance distribution between consecutive CpG dinucleotides. Thus, the default option in the original *CpGcluster* program was to use the median as the distance threshold to identify the clusters. However, we later found notable differences for other chromosomes, concluding that the median is not a good estimator for the intersection (Hackenberg et al., 2011). Therefore, in the *WordCluster* algorithm we added an option to compute the genome intersection as the point showing the maximum difference between observed and expected cumulative density functions (CDFs) of the distances (see Fig. 1 in Hackenberg et al., 2011 and Fig. 1 (bottom) in the present paper). The *GenomeCluster* algorithm inherits this method, and although the script still allows for the use of the median (or some other percentile), the recommended (default) option is taking the genome (or chromosome) intersection point as the distance threshold.

When the data pertain to just one chromosome, the chromosome intersection is defined by the intersection between the

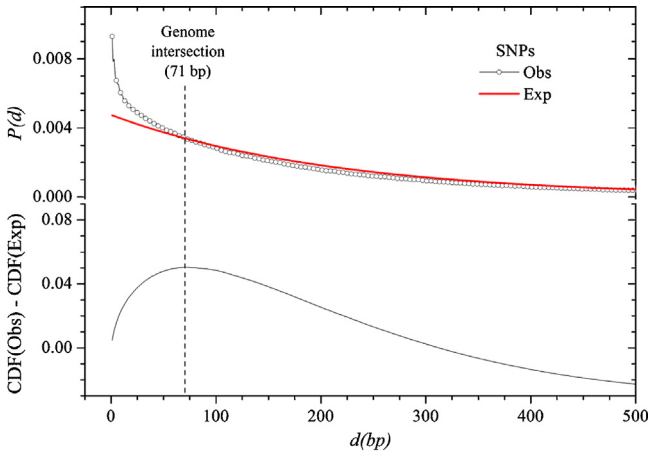


**Fig. 1.** Top: observed and expected distance distributions for the SNPs of *hg19*. Note that short distances are overrepresented and the large ones underrepresented as compared to the expected distances (geometric distribution). The first cross between both curves separates both regimes. Bottom: the intersection between both curves (called the genomic intersection) can be precisely computed (71 bp) as the maximum difference between the observed and expected cumulative density functions (CDFs).

observed and the expected distance distributions in the chromosome. With genome-wide data, the observed distance distributions for all the chromosomes are merged together, then calculating the genome intersection point.

### 2.2.2. Statistical significance of genome clusters

To associate a *p*-value to each of the genome clusters (i.e. the probability of such a cluster appearing by chance in a random sequence) we used the negative binomial distribution (also known as Pascal or Pólya distribution) which can be conveniently tailored to the requirements of genome clusters (Hackenberg et al., 2011). In general, this distribution can be applied to experiments with dichotomous outcomes (either success or failure) and gives the probability of having a certain number of failures when the number of successes was fixed in advance, taking into account that the experiment must always end with a success.

By translating these requirements to a genome context, the successes were equated with the occurrences of the particular genome element being analyzed, while the failures are equated to non-genome elements. When the genome intersection is used to estimate the distance threshold, the success probabilities are not calculated for each chromosome separately, but a genome wide success probability (probability to find the element in the entire genome) is calculated.

### 2.2.3. Implementation

As mentioned above, to detect clusters of genome elements, we used a generalization of the algorithms *CpGcluster* (Hackenberg et al., 2006) and *WordCluster* (Hackenberg et al., 2011) to develop *GenomeCluster* (the corresponding *Perl* script is available at http://bioinfo2.ugr.es/GenomeCluster/software/). Instead of searching the sequence for CpG dinucleotides (as in *CpGcluster*) or *k*-mer occurrences (as in *WordCluster*), the *GenomeCluster* algorithm uses directly the chromosome coordinates of a given genome element retrieved from any annotation table. A requirement is that the table is given in the standard BED format (http://genome.ucsc.edu/FAQ/FAQformat.html#format1). Then, the linear physical distances between consecutive genome elements in the chromosome are directly determined from the coordinates.

The option 'start' was used for all the genome elements analyzed here. This means that the start of each element is used to measure the distances (in bp) between consecutive elements in the chromosome. Therefore, some inter-element distances may need to be interpreted in a slightly different way than usual. For example, gene distances are not the typical intergenic distances that are measured from a gene's end to the following gene's start.

The un-sequenced chromosome stretches (i.e. 'islands of Ns') are previously identified for each chromosome by means of a *Python* script (*N.py*, also available at our website); the distances between genome elements including one or more islands of Ns were discarded before computing the distance threshold and the statistical significance for the clusters.

## 3. Results

We first look for the distance distributions between genome elements. Fig. 1 (top) shows a first example with the observed and expected distance distributions for the SNPs of *hg19*. Note that short distances are overrepresented, while large ones are underrepresented, as compared to the expected distances (geometric distribution). The first crossing point between both curves (the genomic intersection, as distances for all the chromosomes are merged together) separates intra- from inter-cluster distances. It can be accurately computed (71 bp) as the maximum difference between observed and expected cumulative density functions (CDFs, Fig. 1 bottom).

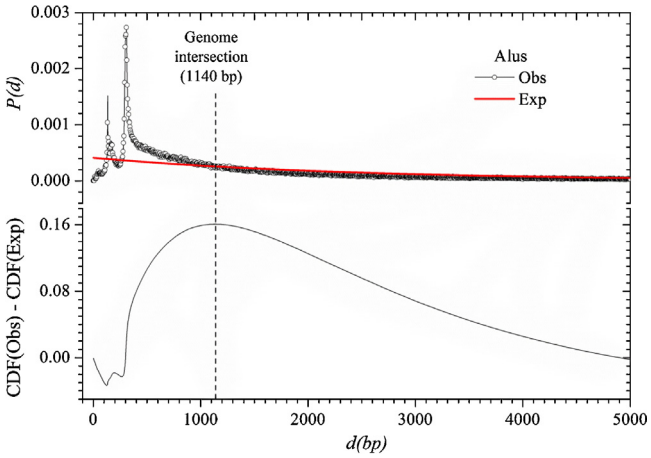


**Fig. 2.** Top: observed and expected distance distributions for the Alus of *hg19*. Bottom: maximum difference between the observed and expected (CDFs) pointing to a genome intersection of 1140 bp. See the legend of Fig. 1 for further details. The two peaks at 135 and 309 bp probably correspond to the lengths of the Alu monomer and of the entire element, respectively.

As a second example, we plot distance distributions for human Alu retroelements (Fig. 2). The genome intersection here is larger (1140 bp); note that it can be easily distinguished from other lower, spurious intersection points by using the maximum difference between observed and expected CDFs (Fig. 2 bottom). The peak at 309 bp surely corresponds to the length of the entire Alu elements, while that at 135 bp may correspond to the length of the Alu monomer, as these repeats are often found fragmented in the genome (Hackenberg et al., 2005; Jurka et al., 2005).

We then used the script *GenomeCluster* with default parameters (i.e. genome intersection as distance threshold, *p*-value $\leq$ 1E−5, clustering method = 'start') to search for clusters of genome elements in the 24 chromosomes (22 autosomes + X + Y) of the *hg19* human genome assembly. We analyzed ten categories of elements (listed in Table 1), including gene-based annotations (as genes, exons or introns), regulatory elements (as CpG islands, TFBSs, enhancers), repeats elements (as Alus or LINE1), variation sites (SNPs) and DNase hypersensitivity sites.

In total, we analyzed more than 64 million elements and found that 27% of them are organized into 424,113 genome clusters. However, a considerable variation occurs between different element categories (see Table 1). A good indicator of clustering level is the percentage of clustered elements (displayed between parentheses in column 4 of Table 1). The highest clustering levels were found for TFBSs and enhancers. The genes and CpG islands show moderate, and strikingly similar, clustering percentages, while exons and introns double the clustering of the genes. The lower percentages of clustered elements were found for repeats (Alus and, above all, LINE1), DNase sites and SNPs. Genes, DNase sites or Alus form he clusters with more elements (Table 1, column 6), while the clusters with a minor number of elements were those for CpG islands or enhancers.

Cluster counts (Table 1, column 3) show that the highest absolute numbers of clusters are formed by SNPs and TFBSs, although their cluster lengths (Table 1, column 5) are the shortest ones. Genes form the lowest number of clusters, but they are the longest ones, both in bp and average number of components (Table 1, column 6). This seems to suggest a relation between cluster and genome-element sizes. However, the clustering level does not follow this rule (e.g. the lowest clustering level was obtained for the SNPs and the largest one for TFBSs, which are only of moderate size). Another

**Table 1**
Clusters of genome elements pertaining to ten different categories in the human genome (hg19).

| Genome entity | Number of elements | Number of clusters | Elements forming clusters (%) | Mean cluster length (bp) ± SD | Mean number of elements by cluster ± SD |
|---|---|---|---|---|---|
| Genes | 19,152 | 206 | 4408 (23%) | 441,645 ± 294,208 | 21 ± 13 |
| Exons | 198,933 | 5089 | 78,425 (39%) | 16,035 ± 10,234 | 15 ± 11 |
| Introns | 179,781 | 5178 | 73,506 (41%) | 14,692 ± 9778 | 14 ± 10 |
| CpG islands | 204,834 | 5563 | 44,408 (22%) | 3384 ± 3143 | 8 ± 6 |
| TFBSs | 4,380,444 | 160,519 | 2,707,380 (62%) | 230 ± 158 | 17 ± 15 |
| Enhancers | 318,454 | 25,944 | 176,925 (56%) | 649 ± 672 | 7 ± 3 |
| DNase sites | 1,281,988 | 5838 | 121,274 (9%) | 10,214 ± 3914 | 21 ± 7 |
| Alus | 1,175,329 | 8020 | 158,121 (13%) | 8140 ± 4593 | 20 ± 10 |
| LINE1 | 1,480,420 | 898 | 13,860 (1%) | 5226 ± 2522 | 15 ± 6 |
| SNPs | 55,448,579 | 206,858 | 2,146,020 (4%) | 25 ± 84 | 10 ± 61 |

interesting observation was that several gene clusters in different chromosomes are over one megabase in length; however, there are chromosomes (as 13, 18 or Y) where the genes do not form clusters. Noteworthy, these three chromosomes are known to have the lowest numbers of both genes and GC-rich isochores (Pavlicek et al., 2002).

### 3.1. Genome-wide maps of genome clusters

Genome-wide maps of the obtained genome clusters can be browsed by means of a track hub – i.e., a public web-accessible directory of genomic data that can be viewed and formatted through the UCSC Genome Browser as they were native tracks. Fig. 3 shows an example map obtained with this facility. Once the data are at the UCSC site, the tables of chromosome coordinates for the different cluster sets can be downloaded by the user simply switching to the Table Browser, choosing the group 'Genome Cluster' and setting the appropriate options to format the output.

### 3.2. 'Circos' maps

Positional information of genome clusters can be also viewed as circular maps for each chromosome generated by means of the software-package *Circos* (Krzywinski et al., 2009). Fig. 4 shows an example for chromosome 19 and images for the remaining chromosomes are available at http://bioinfo2.ugr.es/GenomeCluster/circos-maps/. Intra and inter-chromosome heterogeneity in the number, length and density of clusters for different element categories can be easily compared with the help of these images.

## 4. Discussion

Genes (Ben-Elazar et al., 2013; Durand and Sankoff, 2003; Firneisz et al., 2003; Kendal, 2004; Lercher et al., 2002; Li et al., 2005; Neel, 1961; Thomas, 2002; Wright et al., 2007), CpG dinucleotides (Bird, 1986; Hackenberg et al., 2006), Alu retrotransposons (Hackenberg et al., 2005; Jurka and Kohany, 2005; Jurka et al., 2004, 2002; Pavlícek et al., 2001; Sellis et al., 2007), TFBSs (Berman et al., 2002; Boeva et al., 2007; Murakami et al., 2004), 3D structural motifs in ribosomal RNA (Sargsyan and Lim, 2010), SNPs (Amos, 2010), somatic mutations in cancer (Nik-Zainal et al., 2012), and many DNA *k*-mers (Hackenberg et al., 2012) are all known to occur in clusters. The analysis we carried out here with the help of the *GenomeCluster* algorithm confirms and generalizes these observations. In addition, we were able to detect the clustering level, the chromosome coordinates, the length and the composition of the clusters for a wide variety of functional, regulatory or repeat elements, as well as for variation sites.

### 4.1. Clusters of functional elements

Functional and regulatory elements (genes, CpG islands, TFBSs, enhancers) show a high proportion of clustered elements (40.8% in average, see Table 1), in agreement with recent enrichment/depletion experiments showing that highly clustered words (DNA *k*-mers) are significantly enriched in the functional part of the genome (Hackenberg et al., 2012). Both exons and introns also show high clustering levels (39% and 41%, respectively); the mean numbers of these elements by cluster (15 ± 11 and 14 ± 10 in averages, respectively) agree with the fact that each intron is flanked

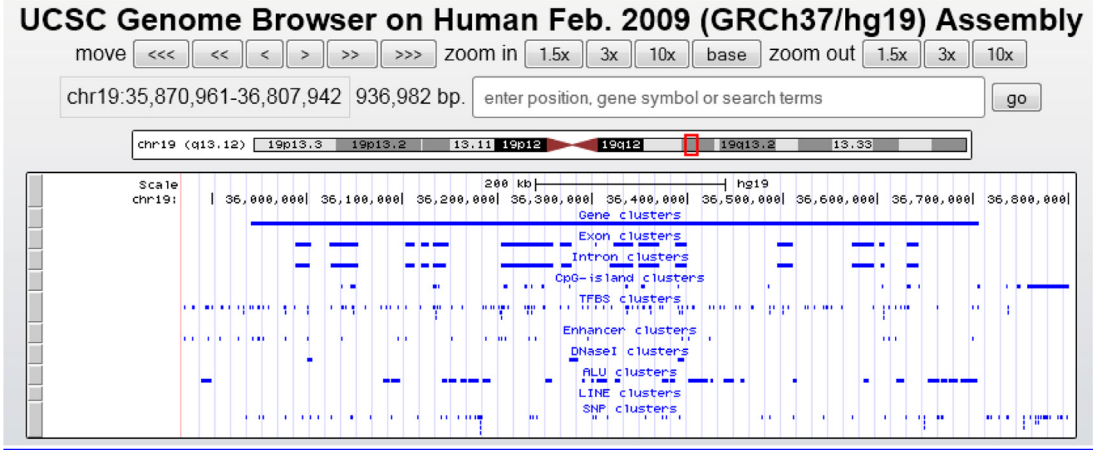


**Fig. 3.** Clusters of genome elements pertaining to ten different categories in a region of 936,982 bp of human chromosome 19. The image was obtained using the UCSC track facility.

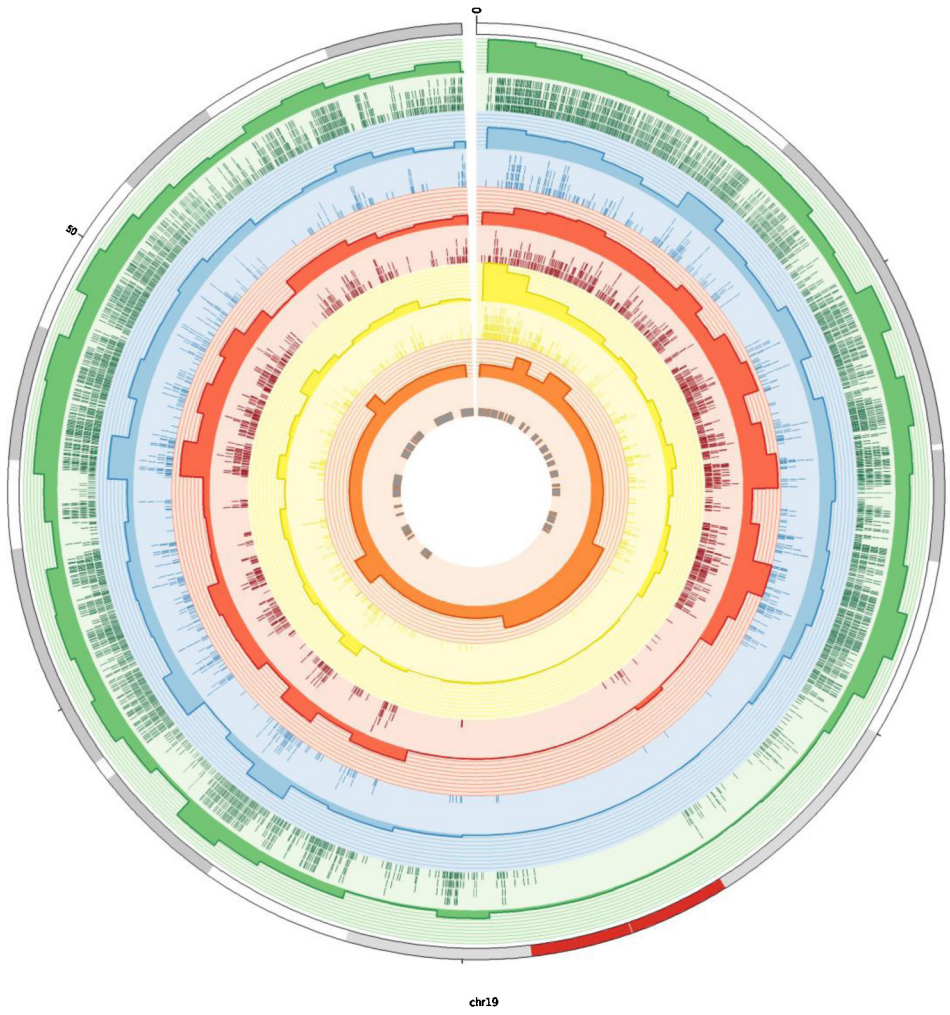


**Fig. 4.** Clusters of genome elements pertaining to different categories in the human chromosome 19 represented as a series of concentric circles by means of the program *Circos* (Krzywinski et al., 2009). Histograms of densities and tiles of genome clusters are shown for each genome category. Five representative categories are drawn from outside to inside: TFBSs, enhancers, Alus, CpG islands and genes.

by exons at both sides. The small difference between these two proportions may be due to single-exon genes. Some of the introns in the human genome are hundred thousand bp in length (Sakharkar et al., 2004), and therefore some genes could contain two or more exon clusters. In fact, we found 813 (or 4.2%) of such genes in *hg19* (see an example in Fig. 5).

Since TFBSs are often located within gene bodies (Wittkopp and Kalay, 2011), one would expect similar levels of clustering for these two types of elements. However, Table 1 shows that TFBSs show a far higher proportion (62%) of clustered elements than genes (23%). The fact that a plethora of transcription factors binds within protein-coding regions, in addition to nearby noncoding regions

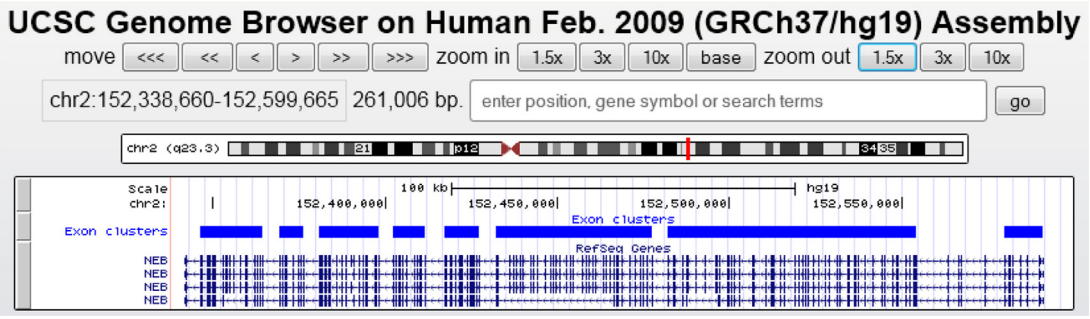


**Fig. 5.** The 183 exons of the gene NEB (chr2: 152,341,853–152,591,001) are grouped by our algorithm into 8 identifiable clusters. This gene encodes nebulin, a giant modular protein component of the cytoskeletal matrix within the sarcomeres of skeletal muscle. The encoded protein contains approximately 30-amino acid long modules that can be classified into 7 types and other repeated modules. Of the 183 exons in the nebulin gene, at least 43 are alternatively spliced. The figure also shows four of the several thousand transcript variants predicted for nebulin by the RefSeq Project.

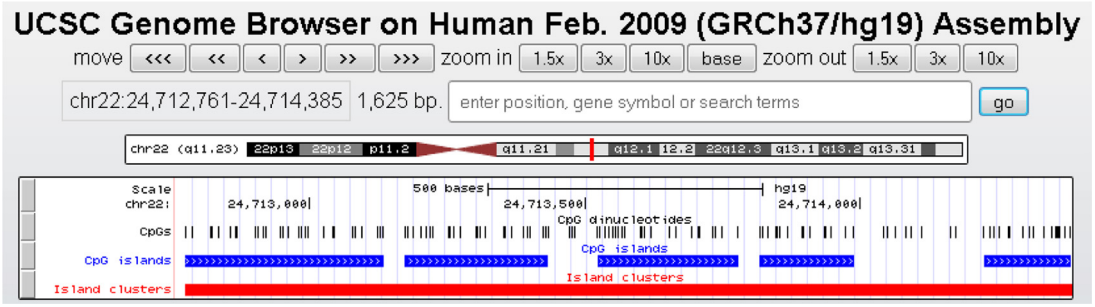


**Fig. 6.** Hierarchical clustering for CpG dinucleotides in a region of 1625 bp of human chromosome 22 (24,712,761–24,714,385). The dinucleotides are clustered within CpG islands, which in turn are clustered within larger CpG-island clusters. Both CpG islands and genome clusters for CpG islands are statistically significant ($p$-value $\leq 0.00001$).

(Stergachis et al., 2013; Weatheritt and Babu, 2013) may explain these results.

### 4.2. Repeat and variant clusters

Given its higher density in GC-rich regions (Lander et al., 2001; Pavlícek et al., 2001), Alu retrotransposons are thought to be one of the most clustered elements in the genome (Hackenberg et al., 2005; Jurka and Kohany, 2005; Jurka et al., 2004, 2002; Pavlícek et al., 2001; Sellis et al., 2007). However, we found that, in comparison with other genome elements (Table 1), Alus show only a moderate number of clusters. In addition, the percentage of clustered elements is low (only the 13% of Alus are forming clusters), although the mean number of elements by cluster is one of the higher ones ($20 \pm 10$ in average).

A low clustering level (4%) was found for variation sites (table *All SNPs137* at UCSC). When common (table *Common SNPs137*) and Not-common (the difference between *All* and *Common* tables) variants were separately analyzed, 2% and 5% of clustered elements were found, respectively. Clinically associated SNPs (tables SNPFlagged and ClinVar of the UCSC Table Browser) show far higher clustering levels (74% and 79%, respectively). However, these data should be taken with caution, since sample sizes of flagged SNPs are much smaller. An additional caution is that clinically associated SNPs co-localize with gene bodies, and therefore the clustering of genes, exons and introns are probably conditioning the clustering levels we found for this class of SNPs.

The non-random distribution of SNPs in the human genome has been explained by mutational non-independence (Amos, 2010), and also by the action of natural selection, with purifying selection eliminating SNPs from functional regions and balancing selection promoting the clustering of SNPs (Varela and Amos, 2010; Zhao et al., 2003).

### 4.3. Clusters within clusters

It is well known that clusters of low-level entities in fact compose many of the genome elements analyzed here. For example, CpG islands are due to the clustering of CpG dinucleotides, TFBSs were obtained by clustering peaks of transcription factor occupancy, DNase sites are clusters of peaks of DNase hypersensitivity, and so on. We investigated therefore if the clusters obtained with *GenomeCluster* could be in turn grouped forming structures of a higher rank. We found that all the genome elements analyzed in this work (Table 1) were able to form 'super-clusters' when they were taken as units to feed again the *GenomeCluster* script. Further work is needed, however, to properly determine the statistical and biological significance of these higher-order superstructures.

The emerging view was, therefore, a genome landscape dominated by hierarchical clustering, thus supporting previous observations by global statistical methods of 'domains-within-domains' in eukaryotic genomes (Bernaola-Galván et al., 1996; Li and Kaneko, 1992a; Li et al., 1994; Oliver et al., 2004, 2001; Román-Roldán et al., 1998). Examples of hierarchical clustering are shown in Fig. 5 (the exons form clusters within a gene coding for a modular protein) and 6 (the dinucleotide CpG is clustered within CpG islands, which are in turn clustered within larger clusters of CpG-islands). To date, the 'domains-within-domains' phenomenon has been uniquely observed in the complex, long-range correlated sequences of eukaryotic genomes, but not in bacterial genomes (see Fig. 3 in Bernaola-Galván et al., 1996); thus, it seems to be related only to complex genomes.

### 4.4. Perspectives

The present approach to describe the organization and evaluate the role of genome clusters in genome complexity has the limitation that we only searched for clusters of elements belonging to a same category, i.e. only homoclusters were detected. Most probably, as envisaged by the *circos* maps (Fig. 4), the situation in the genome is more complex, with homoclusters of a same element type intermixed or co-clustered with homoclusters of other element types, a process that would lead to heteroclusters of disparate genome elements. The next step therefore would be to develop computational tools able to reliably detect such heteroclusters, and investigate how these are organized, leading to the hugely complex genome structure anticipated by global statistical measures of genome complexity (Fig. 6).

Another, more general, limitation of our study is that it only consider genome clustering along the one-dimensional chromosome sequence, thus ignoring the spatial clustering that may result from the 3D organization of the chromosomes within the nucleus, which can put together genome elements actually far in the chromosome or even located on different chromosomes (Trieu and Cheng, 2014). The 3D vicinity of genetic elements is surely most relevant to gene function (Pennisi, 2011), but unfortunately the sequence data we used here do not allow to address this interesting problem. The chromosomal contact data generated by Hi-C chromosome conformation capturing techniques (Lieberman-Aiden et al., 2009) should allow to approach this problem in the near future.

## Web supplementary data

Supplementary data associated with this article can be found, in the online version, at http://bioinfo2.ugr.es/GenomeCluster/.

## Acknowledgements

We are most grateful to Wentian Li by testing the software and solve some bugs. Helpful comments from two reviewers are also greatly acknowledged. This work was carried out by using the facilities of the Computational Genomics and Bioinformatics Group,

Dept. of Genetics, Inst. of Biotechnology, University of Granada (Spain).

## References


Amos, W., 2010. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc. R. Soc. B 277, 1443–1449.

Ben-Elazar, S., Yakhini, Z., Yanai, I., 2013. Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. Nucleic Acids Res. 41, 2191–2201.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., Eisen, M.B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. U.S.A. 99, 757–762.

Bernaola-Galván, P., Román-Roldán, R., Oliver, J., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev. E: Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top. 53, 5181–5189.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F., 1985. The mosaic genome of warm-blooded vertebrates. Science 228, 953–958.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.-K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J.M., Song, L., Trinklein, N.D., Altshuler, R.C., Brown, J.B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T.S., Gerstein, M., Giardine, B., Greven, M., Hardison, R.C., Harris, R.S., Herrero, J., Hoffman, M.M., Iyer, S., Kelllis, M., Kheradpour, P., Lassman, T., Li, Q., Lin, X., Marinov, G.K., Merkel, A., Mortazavi, A., Parker, S.C.J.S.L., Reddy, T.E., Rozowsky, J., Schlesinger, F., Thurman, R.E., Wang, J., Ward, L.D., Whitfield, T.W., Wilder, S.P., Wu, W., Xi, H.S., Yip, K.Y., Zhuang, J., Pazin, M.J., Lowdon, R.F., Dillon, L.A.L., Adams, L.B., Kelly, C.J., Zhang, J., Wexler, J.R., Good, P.J., Feingold, E.A., Crawford, G.E., Dekker, J., Elinitski, L., Farnham, P.J., Giddings, M.C., Gingeras, T.R., Guigó, R., Hubbard, T.J., Kellis, M., Kent, W.J., Lieb, J.D., Margulies, E.H., Myers, R.M., Starnatoyannopoulos, J.A., Tennebaum, S.A., Weng, Z., White, K.P., Wold, B., Yu, Y., Wrobel, J., Risk, B.A., Gunawardena, H.P., Kuiper, H.C., Maier, C.W., Xie, L., Chen, X., Mikkelsen, T.S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M.J., Durham, T., Ku, M., Truong, T., Eaton, M.L., Dobin, A., Lassmann, T., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B.A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O.J., Park, E., Preall, J.B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandu, K.S., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H.H., Hayashizaki, Y., Reymond, A., Antonarakis, S.E., Hannon, G.J., Ruan, Y., Carninci, P., Sloan, C.A., Learned, K., Malladi, V.S., Wong, M.C., Barber, G.P., Cline, M.S., Dreszer, T.R., Heitner, S.G., Karolchik, D., Kirkup, V.M., Meyer, L.R., Long, J.C., Maddren, M., Raney, B.J., Grasfeder, L.L., Giresi, P.G., Battenhouse, A., Sheffield, N.C., Showers, K.A., London, D., Bhinge, A.A., Shestak, C., Schaner, M.R., Kim, S.K., Zhang, Z.Z.Z., Mieczkowski, P.A., Mieczkowska, J.O., Liu, Z., McDaniell, R.M., Ni, Y., Rashid, N.U., Kim, M.J., Adar, S., Wang, T., Winter, D., Keefe, D., Iyer, V.R., Sandhu, K.S., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E.C., Varley, K.E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K.M., Anaya, M., Cross, M.K., Muratet, M.A., Newberry, K.M., McCue, K., Nesmith, A.S., Fisher-Aylor, K.I., Pusey, B., DeSalvo, G., Balasubramanian, S.S., Davis, N.S., Meadows, S.K., Eggleston, T., Newberry, J.S., Levy, S.E., Absher, D.M., Wong, W.H., Blow, M.J., Visel, A., Pennachio, L.A., Elnitski, L., Petrykowska, H.M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J.M., Griffiths, E., Harte, R., Hendrix, D.A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M.F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J.M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M.L., van Baren, M.J., Washieti, S., Wilming, L., Zadissa, A., Zhengdong, Z., Brent, M., Haussler, D., Valencia, A., Raymond, A., Addleman, N., Alexander, R.P., Auerbach, R.K., Bettinger, K., Bhardwaj, N., Boyle, A.P., Cao, A.R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J.D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyenger, S., Jin, V.X., Karczewski, K.J., Kasowski, M., Lacroute, P., Lam, H., Larnarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X.J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Struhl, K., Weissman, S.M., Tenebaum, S.A., Penalva, L.O., Karmakar, S., Bhanvadia, R.R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centarin, L., Eichenlaub, M., Gruhl, F., Heerman, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D.L., Byron, R., Canfield, T.K., Diegel, M.J., Dunn, D., Ebersol, A.K., Frum, T., Garg, K., Gist, E., Hansen, R.S., Boatman, L., Haugen, E., Humbert, R., Johnson, A.K., Johnson, E.M., Kutyavin, T.M., Lee, K., Lotakis, D., Maurano, M.T., Neph, S.J., Neri, F.V., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Rynes, E., Sanchez, M.E., Sandstrom, R.S., Shafer, A.O., Stergachis, A.B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Weaver, M.A., Yan, Y., Zhang, M., Akey, J.A., Bender, M., Dorschner, M.O., Groudine, M., MacCoss, M.J., Navas, P., Stamatoyannopoulos, G., Stamatoyannopoulos, J.A., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N.M., Sobral, D., Vaquerizas, J.M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M.W., Schaub, M.A., Miller, W., Bickel, P.J., Banfai, B., Boley, N.P., Huang, H., Li, J.J., Noble, W.S., Bilmes, J.A., Buske, O.J., Sahu, A.O., Kharchenko, P.V., Park, P.J., Baker, D., Taylor, J., Lochovsky, L., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Bird, A., 1986. CpG-rich islands and the function of DNA methylation. Nature 321, 209–213.

Boeva, V., Clément, J., Régnier, M., Roytberg, M.A., Makeev, V.J., 2007. Exact *p*-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. Algorithms Mol. Biol. 2, 13.

Carpena, P., Bernaola-Galván, P., Coronado, A., Hackenberg, M., Oliver, J., 2007. Identifying characteristic scales in the human genome. Phys. Rev. E 75, 032903.

Carpena, P., Bernaola-Galván, P., Hackenberg, M., Coronado, A., Oliver, J., 2009. Level statistics of words: finding keywords in literary texts and symbolic sequences. Phys. Rev. E 79, 035102.

Carpena, P., Oliver, J.L., Hackenberg, M., Coronado, A.V., Barturen, G., Bernaola-Galván, P., 2011. High-level organization of isochores into gigantic superstructures in the human genome. Phys. Rev. E 83, 031908.

Durand, D., Sankoff, D., 2003. Tests for gene clustering. J. Comput. Biol. 10, 453–482.

Firneisz, G., Zehavi, I., Vermes, C., Hanyecz, A., Frieman, J.A., Glant, T.T., 2003. Identification and quantification of disease-related gene clusters. Bioinformatics (Oxford, Engl.) 19, 1781–1786.

Hackenberg, M., Bernaola-Galván, P., Carpena, P., Oliver, J., 2005. The biased distribution of Alus in human isochores might be driven by recombination. J. Mol. Evol. 60, 365–377.

Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, A.M., Oliver, J.L., 2011. WordCluster: detecting clusters of DNA words and genomic elements. Algorithms Mol. Biol. 6, 2.

Hackenberg, M., Previti, C., Luque-Escamilla, P., Carpena, P., Martínez-Aroza, J., Oliver, J., 2006. CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinform. 7, 446.

Hackenberg, M., Rueda, A., Carpena, P., Bernaola-Galván, P., Barturen, G., Oliver, J.L., 2012. Clustering of DNA words and biological function: a proof of principle. J. Theor. Biol. 297, 127–136.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110, 462–467.

Jurka, J., Kapitonov, V., 2007. Repetitive sequences in complex genomes: structure and evolution. Annu. Rev. Genomics Hum. Genet. 8, 241–259.

Jurka, J., Kohany, O., 2005. Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet. Genome Res. 110, 117–123.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V., Jurka, M.V., 2004. Duplication, coclustering, and selection of human Alu retrotransposons. Proc. Natl. Acad. Sci. U.S.A. 101, 1268–1272.

Jurka, J., Krnjajic, M., Kapitonov, V.V., Stenger, J.E., Kokhanyy, O., 2002. Active Alu elements are passed primarily through paternal germlines. Theor. Popul. Biol. 61, 519–530.

Kendal, W., 2004. A scale invariant clustering of genes on human chromosome 7. BMC Evol. Biol., http://dx.doi.org/10.1186/1471-2148-4-3.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang,

W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

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

Li, Q., Lee, B.T.K., Zhang, L., 2005. Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics 6, 7.

Li, W., Kaneko, K., 1992a. Long-range correlations and partial 1/*f* spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 555–660.

Li, W., Kaneko, K., 1992b. DNA correlations. Nature 360, 635–636.

Li, W., Marr, T., Kaneko, K., 1994. Understanding long-range correlations in DNA sequences. Phys. D: Nonlinear Phenom. Vol. 75, 392–416.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J., 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, N.Y.) 326, 289–293.

Lynch, M., Conery, J.S., 2003. The origins of genome complexity. Science (New York, N.Y.) 302, 1401–1404.

Murakami, K., Kojima, T., Sakaki, Y., 2004. Assessment of clusters of transcription factor binding sites in relationship to human promoter, CpG islands and gene expression. BMC Genomics 5, 16.

Neel, J.V., 1961. The hemoglobin genes: a remarkable example of the clustering of related genetic functions on a single mammalian chromosome. Blood 18, 769–777.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K.W., Mudie, L.J., Varela, I., McBride, D.J., Bignell, G.R., Cooke, S.L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P.S., Davies, H.R., Papaemmanuil, E., Stephens, P.J., McLaren, S., Butler, A.P., Teague, J.W., Jönsson, G., Garber, J.E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J.W.M., Aparicio, S.A.J.R., Borg, Å., Salomon, A.V., Thomas, G., Børresen-Dale, A.-L., Richardson, A.L., Neuberger, M.S., Futreal, P.A., Campbell, P.J., Stratton, M.R., 2012. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979–993.

Nussinov, R., Owens, J., Maizel, J.V., 1986. Sequence signals in eukaryotic upstream regions. Biochim. Biophys. Acta 866, 109–119.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag.

Oliver, J., Román-Roldán, R., Pérez, J.B.-G.P., 1999. SEGMENT: identifying compositional domains in DNA sequences. Bioinformatics 15, 974–979.

Oliver, J.L., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276, 47–56.

Oliver, J.L., Carpena, P., Hackenberg, M., Bernaola-Galván, P., 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. 32, W287–W292.

Pavlícek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., Bernardi, G., 2001. Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39–45.

Pavlicek, A., Paces, J., Clay, O., Bernardi, G., 2002. A compact view of isochores in the draft human genome sequence. FEBS Lett. 511, 165–169.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–170.

Pennisi, E., 2011. Mysteries of the cell. Does a gene's location in the nucleus matter? Science (New York, N.Y.) 334, 1050–1051.

Price, A.L., Eskin, E., Pevzner, P.A., 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Res. 14, 2245–2252.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33, D501–D504.

Román-Roldán, R., Bernaola-Galván, P., Oliver, J., 1998. Sequence compositional complexity of DNA through an entropic segmentation method. Phys. Rev. Lett. 80, 1344–1347.

Sakharkar, M.K., Chow, V.T.K., Kangueane, P., 2004. Distributions of exons and introns in the human genome. In Silico Biol. 4, 387–393.

Sankoff, D., 2001. Gene and genome duplication. Curr. Opin. Genet. Dev. 11, 681–684.

Sargsyan, K., Lim, C., 2010. Arrangement of 3D structural motifs in ribosomal RNA. Nucleic Acids Res. 38, 3512–3522.

Schuster, P., 1996. How does complexity arise in evolution. Complexity 2, 22–30.

Sellis, D., Provata, A., Almirantis, Y., 2007. Alu and LINE1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws. Mol. Biol. Evol. 24, 2385–2399.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311.

Stankiewicz, P., Shaw, C.J., Withers, M., Inoue, K., Lupski, J.R., 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res. 14, 2209–2220.

Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M., Stamatoyannopoulos, J.A., 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. Science 342, 1367–1372.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. Bioinformatics 16, 16–23.

Thomas, J., 2002. White House Concerns Block Doubling Bill Jumbled DNA Separates Chimps and Humans Koski Steps Down After Bumpy Ride. The First Director of a Federal Office Created., pp. 64–65.

Trieu, T., Cheng, J., 2014. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. Nucleic Acids Res. 42, e52.

Varela, M.A., Amos, W., 2010. Heterogeneous distribution of SNPs in the human genome: microsatellites as predictors of nucleotide diversity and divergence. Genomics 95, 151–159.

Voss, R., 1992. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Weatheritt, R.J., Babu, M.M., 2013. The hidden codes that shape protein evolution. Science 342, 1325–1326.

Wittkopp, P., Kalay, G., 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 13, 59–69.

Wright, M.A., Kharchenko, P., Church, G.M., Segrè, D., 2007. Chromosomal periodicity of evolutionarily conserved gene pairs. Proc. Natl. Acad. Sci. U.S.A. 104, 10559–10564.

Zentner, G.E., Tesar, P.J., Scacheri, P.C., 2011. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. 21, 1273–1283.

Zhao, Z., Fu, Y.-X., Hewett-Emmett, D., Boerwinkle, E., 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312, 207–213.