

Organizational Heterogeneity of Vertebrate Genomes

Svetlana Frenkel, Valery Kirzhner, Abraham Korol*

Department of Evolutionary and Environmental Biology and Institute of Evolution, University of Haifa, Mount Carmel, Haifa, Israel

Abstract

Genomes of higher eukaryotes are mosaics of segments with various structural, functional, and evolutionary properties. The availability of whole-genome sequences allows the investigation of their structure as “texts” using different statistical and computational methods. One such method, referred to as Compositional Spectra (CS) analysis, is based on scoring the occurrences of fixed-length oligonucleotides (k-mers) in the target DNA sequence. CS analysis allows generating species- or region-specific characteristics of the genome, regardless of their length and the presence of coding DNA. In this study, we consider the heterogeneity of vertebrate genomes as a joint effect of regional variation in sequence organization superimposed on the differences in nucleotide composition. We estimated compositional and organizational heterogeneity of genome and chromosome sequences separately and found that both heterogeneity types vary widely among genomes as well as among chromosomes in all investigated taxonomic groups. The high correspondence of heterogeneity scores obtained on three genome fractions, coding, repetitive, and the remaining part of the noncoding DNA (the genome dark matter - GDM) allows the assumption that CS-heterogeneity may have functional relevance to genome regulation. Of special interest for such interpretation is the fact that natural GDM sequences display the highest deviation from the corresponding reshuffled sequences.

Citation: Frenkel S, Kirzhner V, Korol A (2012) Organizational Heterogeneity of Vertebrate Genomes. PLoS ONE 7(2): e32076. doi:10.1371/journal.pone.0032076

Editor: Vincent Laudet, Ecole Normale Supérieure de Lyon, France

Received: August 15, 2011; **Accepted:** January 23, 2012; **Published:** February 27, 2012

Copyright: © 2012 Frenkel et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Scholarship for Excellence in Converging Technologies Program of Israel Council for Higher Education (http://www.che.org.il/template/default_e.aspx?Pageld=286) and the Scholarship Program for Research Students of the Israeli Ministry of Immigrant Absorption (http://www.moia.gov.il/Moia_en/Scientists/ScholarshipProgram.htm). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: korol@research.haifa.ac.il

Introduction

Unraveling the structural organization of complex eukaryotic genomes is one of the most important problems in current genomics. A plentiful amount of genomes has been sequenced and is available for further analysis. Long DNA sequences, like chromosomes or entire genomes, are known to be heterogeneous in their structural aspects, such as GC content (isochores), CpG distribution, copy number variation, repetitive DNA content, and distribution of indels. Furthermore, they are heterogeneous in their functional and evolution-related features, including dynamics of DNA replication, protein and non-protein-coding DNA content, codon usage, level and tissue-specificity of gene expression, distribution of conserved and ultra-conserved regions, recombination and mutation hot and cold spots, and SNPs and LD-blocks [1–6]. Comparative analysis of mutation rates in mammals indicates that parallel syntenic blocks, rather than entire chromosomes, may represent the units of intragenomic heterogeneity of mutation rates [7]. However, until recently, most of the genome analyses were devoted to the coding space; hence, a major part of sequence organization of eukaryotes remained poorly studied, including intragenomic heterogeneity. One of the main exclusions was, and still remains, intragenomic variation in GC content.

GC content and CpG islands

A simple measure of compositional organization of nucleotide sequences is the molar ratio of G+C in DNA, or GC content. GC content displays wide variation within genomes, chromosomes,

and chromosome segments [8–10]. Long homogeneous regions with certain GC content are called GC isochores [8]; the resolution of isochore maps of the human genome is higher than the resolution of classical Giemsa and Reverse bands [11]. GC content is known to be strongly correlated with biological features of genome organization, such as dynamics of DNA replication [12,13], gene density [14,15], level and tissue-specificity of transcription [16], mutation and recombination rates [17,18]. GC content in non-mammalian vertebrate genomes is less variable than in mammals, but GC-based segmentation of these genomes is still possible [19]. However, as shown by Nekrutenko and Li [20] and supported by our results, heterogeneity of eukaryotic genomes is not confined to the variation of GC isochores. Li [10] showed that some isochores are heterogeneous in their GC content and Costantini and Bernardi [21] found different di- and tri-nucleotide patterns in diverse isochores. Therefore, GC-based characteristics are not sufficient for comprehensive investigations of eukaryotic genome heterogeneity and its relevance to functions and the evolution of genomes. Well known since 1987 [22], CpG islands were found to be associated with gene abundance [23], gene expression [16], local abundance of Alu-sequences and (obviously) with GC content of the region [24]. GC content and CpG islands play a role in tissue-specific differentiation and cancer development [25,26].

Oligonucleotide-based methods

A group of methods for analyzing genome organization based on counting oligonucleotide “word” (or k-mer) occurrences was proposed in the 1980s [27–30]. The main purpose of such

alignment-free analysis is the determination of genomic characteristics (signatures, patterns). It allows differentiation of regions within a genome, comparison of genomes of diverse species, and many other applications [31,32]. Genome-specific characteristics are employed in phylogenetic analyses [33] or in species recognition using their relatively short DNA fragments as training inputs for classification algorithms [34,35]. Region-specific characteristics can be used for the detection of certain elements in DNA sequences such as candidate regulatory elements [36–38], promoter regions [39], and repetitive elements that were not found before [40]. This method proved useful for the detection and determination of the origin of alien DNA segments in studies of horizontal gene transfer [41–43] and duplications of genomic segments [44]. In addition, the oligonucleotide-counting methods are used for preliminary searches of candidates for subsequent gene alignment [45] as well as whole-genome sequence comparisons [30,46–49].

One of the word-counting methods, referred to as Compositional Spectra Analysis (CSA), based on scoring the occurrences of fixed-length words from a predefined set (“vocabulary”) with a chosen level of allowed mismatching, was suggested by Kirzhner et al. [50] and used for genome comparisons of different species. CSA allows generating species- or region-specific characteristics of the genome, regardless of their length and the presence of coding DNA. Here we employ CSA in the investigation of organizational heterogeneity of vertebrate genomes. In addition to the entire DNA sequence, we analyze its coding and two noncoding parts separately: repeated DNA and the rest of the sequence, which we refer to as genome dark matter (see also [51–54] for a similar use of this terminology).

Our results provide new evidence that genome structure is more complicated than just a combination of regions with diverse GC contents; this structure varies at the level of chromosomes, but is shared to a considerable extent between the coding space and the considered two parts of noncoding space. In our CS analysis, we investigated three heterogeneity types: compositional (i.e., variability of nucleotide abundances along the sequence), organizational (i.e., variability of nucleotide order along the sequence), and total CS heterogeneity, which is a result of a complicated interaction of the two former types.

Employed sequences

Chromosomal complete and repeat-masked sequences of 19 vertebrate species and genomic complete and repeat-masked sequences (available as scaffolds) of 26 vertebrate species were downloaded from Ensembl (release 57, <http://mar2010.archive.ensembl.org/index.html>). We analyzed several taxonomic groups of species (Table 1), including five groups of mammals: apes (4) and other primates (5), ungulates (4), rodents (5) and “other mammals” group, consisting of 14 diverse species; and four groups of non-mammalian vertebrates: marsupials (2), birds (3), fishes (5) and “other vertebrates” (anole lizard, frog, and platypus).

A human predicted gene list that includes descriptions and positions of genes and pseudogenes of protein-coding and RNA-coding sequences was imported from Ensembl BioMart Database (<http://www.ensembl.org/biomart/martview>).

Methods

Computation of Compositional Spectrum

Consider a set W (vocabulary) including n different words (oligonucleotides) w_i of length L from standard $\{G, C, A, T\}$ alphabet. Obviously, what follows is $n \leq 4^L$. By moving along a DNA sequence of length K one position per step with a window of

length L , we can collect $M = K - L + 1$ subsequences of length L . Comparing of these subsequences with all words w_i from the set W with a certain number of allowed replacements (mismatches) r allows us to calculate the number of imperfect matches $m_i = m(w_i)$ of each word with the analyzed sequence. Thus, the appearance of a 5-mer CTATG in a sequence of 54 bp length CTTTGA-GTGGCAATAGAGCATTTCAGTAATTGTACCTCTATCC-CTACAAGGAAC with $r = 2$ is $m(\text{CTATG}) = 6$. This word would not be found in the sequence upon $r = 0$ (without mismatches), but with one mismatch it was found twice (CTTTG and CTATC) and with two mismatches four times (CAATA, CTCTA, CTACA, and CAAGG).

The frequency distribution $F(W, S)$ based on frequencies $f_i = m_i / M$ is referred to as Compositional Spectrum (CS) of sequence S relative to a set of words W [50]. In our analysis, sets of words were generated by using an algorithm that randomly chooses the first word of length L in a set, and then examines all L -mers in alphabetical order. Every word, which differs from all previously generated words by a chosen number of letters, is included in the set. When the algorithm reaches the last word in the alphabet order (TTTTTTTTTT), it begins from the first word (AAAAAAAAAA), finishing the search when the first word included in the set is reached. Number n of words in a set does not depend on the first word included. Since this set contains the words with all variations of letters abundances, it is unbiased in its sensitivity to different genomic sequence. However, one may prefer to use specific biased sets of words, e.g., overpopulated by GC- or AT. For $L = 10$ and mismatch $r \leq 3$, we employed sets of 256 words with differences between words by at least six positions, e.g., ATGCTGTCAT, ATGCATAGCA, ATGCCACTGC, ATGCGCGATG, etc. Thus, with $r = 3$ any 10-letter subsequence from the targeted sequence coincides with no more than one word from such set. However, some oligonucleotides in the targeted genome sequence may remain uncovered.

Coverage of targeted sequences

The maximal number of imperfect matches μ for any word of length L in the alphabet with A letters and allowed the mismatching level r per word, can be calculated using the following formula:

$$\mu_i = \binom{L}{r} \cdot (A-1)^r + \binom{L}{r-1} \cdot (A-1)^{r-1} + \dots + \binom{L}{2} \cdot (A-1)^2 + L \cdot (A-1) + 1$$

For $L = 10$, $r = 3$, the maximal number of matches for any word in the standard 4-letter alphabet is 3676, which gives us 941056 matches for a set of 256 words, i.e. 86.7% from all possible 10-letter words. We employed a few sets of words (obtained by the method described above) to analyze the human chromosome sequences using different variations of segment length and found that all sets cover 90–91% of chromosome sequences in all cases. This excess over the expected coverage is probably a result of the fact that the human genome does not use all possible 10-letter words.

Justification of the selected parameter values

The sensitivity of CS to the sequence organizational features increases with the increasing of words length used. However, very long oligonucleotide words could be too specific with vast majority being presented with zero occurrences. The words of 8–15 letters

Table 1. List of genomes used for analyses.

English name	Latin name	Ensembl assembly name	English name	Latin name	Ensembl assembly name
Apes			Hedgehog	<i>Erinaceus europaeus*</i>	HEDGEHOG
Gorilla	<i>Gorilla gorilla</i>	gorGor3	Cat	<i>Felis catus*</i>	CAT
Human	<i>Homo sapiens</i>	GRCh37	Elephant	<i>Loxodonta africana*</i>	loxAfr3
Chimpanzee	<i>Pan troglodytes</i>	CHIMP2.1	Microbat	<i>Myotis lucifugus*</i>	MICROBAT1
Orangutan	<i>Pongo pygmaeus</i>	PPYG2	Pika	<i>Ochotona princeps*</i>	PIKA
Other primates			Hyrax	<i>Procavia capensis*</i>	proCap1
Marmoset	<i>Callithrix jacchus*</i>	caJac3	Megabat	<i>Pteropus vampyrus*</i>	pteVam1
Macaque	<i>Macaca mulatta</i>	MMUL_1	Shrew	<i>Sorex araneus*</i>	COMMON_SHREW1
Mouse lemur	<i>Microcebus murinus*</i>	micMur1	Tree shrew	<i>Tupaia belangeri*</i>	TREESHREW
Bushbaby	<i>Otolemur garnettii*</i>	BUSHBABY1	Dolphin	<i>Tursiops truncatus*</i>	turTru1
Tarsier	<i>Tarsius syrichta*</i>	tarSyr1	Marsupials		
Rodents			Wallaby	<i>Macropus eugenii*</i>	Meug_1.0
Guinea pig	<i>Cavia porcellus*</i>	cavPor3	Opossum	<i>Monodelphis domestica</i>	BROADOS
Kangaroo rat	<i>Dipodomys ordii*</i>	dipOrd1	Birds		
Mouse	<i>Mus musculus</i>	NCBIM37	Chicken	<i>Gallus gallus</i>	WASHUC2
Rat	<i>Rattus norvegicus</i>	RGSC3.4	Turkey	<i>Meleagris gallopavo</i>	UMD2
Squirrel	<i>Spermophilus tridecemlineatus*</i>	SQUIRREL	Zebra finch	<i>Taeniopygia guttata</i>	taeGut3.2.4
Ungulates			Fishes		
Cow	<i>Bos taurus</i>	Btau_4.0	Stickleback	<i>Gasterosteus aculeatus*</i>	BROADS1
Horse	<i>Equus caballus</i>	EquCab2	Zebrafish	<i>Danio rerio</i>	Zv8
Pig	<i>Sus scrofa</i>	Sscrofa9	Medaka	<i>Oryzias latipes</i>	MEDAKA1
Alpaca	<i>Vicugna pacos*</i>	vicPac1	Fugu	<i>Takifugu rubripes*</i>	FUGU4
Other mammals			Tetraodon	<i>Tetraodon nigroviridis</i>	TETRAODON8
Dog	<i>Canis familiaris</i>	BROADD2	Other vertebrates		
Sloth	<i>Choloepus hoffmanni*</i>	choHof1	Anole lizard	<i>Anolis carolinensis*</i>	AnoCar1
Armadillo	<i>Dasypus novemcinctus*</i>	dasNov2	Platypus	<i>Ornithorhynchus anatinus</i>	OANA5
Lesser hedgehog tenrec	<i>Echinops telfairi*</i>	TENREC	Frog	<i>Xenopus tropicalis*</i>	JGI4.1

*incomplete genome sequences.

doi:10.1371/journal.pone.0032076.t001

long are most appropriate for our purposes [50]. Increasing the allowed number of replacements lowers the sensitivity of CS to sequence heterogeneity while decreasing r reduces the number of matches μ and leads to lower coverage of the targeted sequence. To avoid this effect one could use a richer set of words (e.g., all 1,048,576 10-mers at $r=0$ giving 100% coverage without mismatching at all). However, as noted above, decreasing r reduces the observed frequencies of words and leads to a huge prevalence of zero frequencies in CS of short-to-moderate sequences (~100–1000 kb), thereby complicating sequence comparisons. Longer sequences ($\gg 1$ Mb) would relax this problem, but the resolution of such an analysis is not sufficient for investigation of sequence heterogeneity. The segment size of 100 kb length seems to be most suitable for our goals; moreover, this length is comparable with common GC isochore lengths.

Calculating CS-distances between DNA sequences

Let us define a measure of the differences between two sequences, S_1 and S_2 , as the CS-distance d between their spectra $F(W, S_1)$ and $F(W, S_2)$. For this purpose, we use a metric based on the Spearman rank correlation coefficient [50]. Namely, we calculate Spearman rank correlation coefficient r_s between two CS

corresponding to the compared sequences:

$$r_s = 1 - \frac{6 \sum \delta_i^2}{n(n^2 - 1)},$$

where δ_i is the difference between the ranks of the i^{th} word frequency in the compared spectra, and n is the number of words in the set. The obtained correlation coefficient r_s is used to determine CS distance d between the compared sequences S_1 and S_2 :

$$d = (1 - r_s)/2$$

The maximal value $d=1$ corresponds to strictly reverse compositional spectra (i.e., $r_s = -1$), whereas minimal value $d=0$ corresponds to identical spectra (i.e., $r_s = 1$). However, this method does not take into account the possible reverse complement compositional asymmetry, which is related to origins of replication and direction of transcription [55,56]. This compositional asymmetry is a common feature of higher eukaryote genomes, especially in tests with relatively small (less than 1 Mb) windows. A slight modification of our calculations allows the consideration of

this factor. Specifically, for chosen sequences S_1 and S_2 , the calculated difference (say, d_{12}) can be complemented by score d_{12} , using the CS-distance between strand S_1 and the complementary to S_2 strand taken in the reverse direction. In the case of S_2 strand symmetry, d_{12} and d_{12}' are close to each other; their difference will point to S_2 asymmetry. The same calculations are conducted with a complementary to S_1 strand taken in the reverse direction, resulting in $d_{1'2}$ and $d_{1'2}'$. Thus, $d = \min(d_{12}, d_{12}', d_{1'2}, d_{1'2}')$ is employed as CS distance.

Sequence heterogeneity

We consider three types of sequence heterogeneity: compositional, organizational, and total, which is a combination of the first two types. The *compositional heterogeneity* score H_c is a measure of variation in nucleotide composition along the sequence. We calculate H_c as a median difference in GC content between equal-length segments (due to rather strong asymmetry of the distribution of this value). We can disregard “second-order” differences in the within-strand contents of G vs. C and A vs. T due to the rather strict correspondence of strand composition to the second Chargaff parity rule [57,58]. *Organizational heterogeneity* (H_o) reflects the differences in oligonucleotide structures along the sequence, which can be estimated by CS comparisons. Obviously, nucleotide composition of the compared sequences affects the oligonucleotide frequencies found in these sequences. The simplest way to assess the target sequence for H_o not mixed with H_c is to include in the CS analysis only pairs of segments with similar GC content although not all possible pairs of segments will then be included in the calculation. *Total heterogeneity* (H_t) is calculated over all pairs of segments comprising the target sequence and is affected by both compositional and organizational differences between these segments.

To evaluate heterogeneity of a genome, chromosome or any other sequence, we subdivided it into equal lengths not overlapping 100 kb segments (although other sizes can also be

used) and for each such segment calculated its GC content and CS. Median values of GC differences and CS distances calculated over all pairs of segments were used as assessments of H_c and H_t , correspondingly. The median value of CS distances across pairs with zero or small GC differences was used as an estimate of H_o . To evaluate compositional and organizational differences between two sequences (e.g., two chromosomes), referred to as inter-sequence heterogeneity, we calculated the GC differences and CS distances between each segment of one versus each segments of the other sequence.

To determine the acceptable level of inter-segmental GC difference in H_o calculations, we compared the distribution of inter-chromosomal CS distances as a function of GC difference (ΔGC). It appeared that $\Delta GC \leq 0.02$ provides sufficient stability of the distribution, i.e., it does not change considerably with further decreases of ΔGC (Fig. 1). Moreover, using a smaller ΔGC means a reduction of the number of analyzed pairs of segments. Thus, across the human chromosomes, about 26% of segment pairs participate in the analysis of organizational heterogeneity in using $\Delta GC \leq 0.02$ and only 12% and 6% in cases of $\Delta GC \leq 0.01$ and $\Delta GC \leq 0.005$, respectively. On average, for the analyzed vertebrate genomes, using the condition $\Delta GC \leq 0.02$ provides participation of 40% of segment pairs in H_o analyses (range: from 23% for pig *Sus scrofa* to 74% for tarsier *Tarsius syrichta*). Similar results were obtained with 50 kb and 500 kb segments (not shown). Therefore, in all subsequent tests, we employed the condition $\Delta GC \leq 0.02$.

Dependence of intra- and inter-sequence heterogeneity on the set of words

We tested a few different unbiased sets of words and found that intra- and inter-sequence heterogeneity scores of entire genomes, chromosomes, or chromosome segments do not depend on the set used.

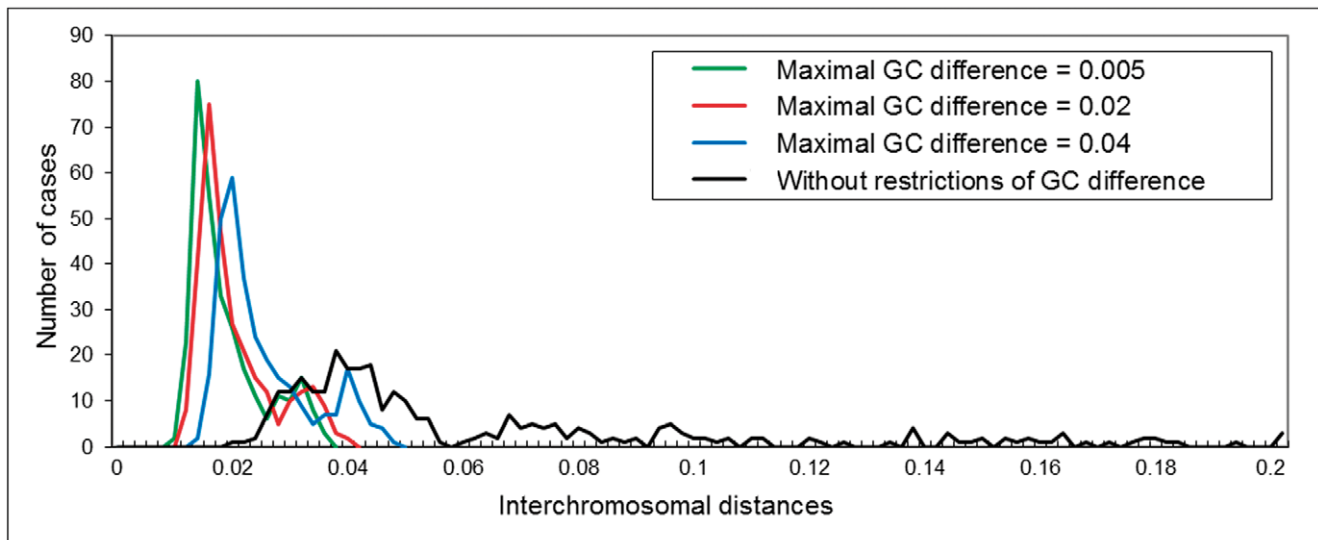


Figure 1. The distribution histogram of human median inter-chromosomal CS-distances obtained with different permitted maximal levels of GC differences between scored 100 kb segments. The inter-chromosomal CS-distances obtained without restriction on inter-segmental GC differences contain both compositional and organizational components of inter-chromosomal heterogeneity (thereby coinciding with H_t). The stricter conditions on ΔGC reduce the influence of H_c on the estimate of H_o . All ΔGC thresholds that are ≤ 0.02 show just slightly different distributions of human inter-chromosomal H_o . Therefore, we believe that $\Delta GC \leq 0.02$ condition permits us to assess organizational inter-chromosomal heterogeneity without the influence of compositional heterogeneity. doi:10.1371/journal.pone.0032076.g001

Results

Intragenomic heterogeneity

First, we analyzed different types of whole-genome heterogeneity: compositional (H_c), organizational (H_o), and total (H_t) for 45 fully sequenced vertebrate genomes from different taxonomic groups. For every genome sequence, subdivided to 100 kb segments, we calculated CS distances and GC differences for all segment pairs and estimated H_c , H_o , and H_t (see section Methods). All three heterogeneity types varied more than 4–5-fold across species (Figure 2). Maximal H_t values were observed within fishes, apes, and rodents; fishes and rodents showed maximal H_o as well, while in apes H_o was relatively small (36–38% of their H_t). Minimal H_t and H_o values were found in marsupials, but H_o of marsupial genomes reached almost 80% of H_t , which is the maximal value among all analyzed genomes. The highest variability of all heterogeneity values was found in fishes, rodents, and “other mammals” group; ape and bird genomes turned out to be very similar to each other. Five species selected for their highest H_t also appeared to have the highest H_o but not H_c (Fig. 3); it is interesting that three of these species were fishes (*Gasterosteus aculeatus*, *Takifugu rubripes*, and *Tetraodon nigroviridis*), two others were mammals: pika (*Ochotona princeps*) and kangaroo rat (*Dipodomys ordii*). A relatively high correlation between H_t and H_c scores was observed in vertebrate genomes (Spearman rank correlation coefficient $\rho = 0.67$) and between H_t and H_o ($\rho = 0.69$), while no correlation was found between H_o and H_c ($\rho = 0.08$). According to linear multiple regression analysis, a combination of both variables (H_c and H_o) explained $R^2 = 0.94$ of variation in H_t between species, while the separate effects of H_o and H_c on H_t were $R^2 = 0.53$ and 0.35 , respectively. Despite considerable dependence of H_o and H_t on genome average GC content ($\rho = 0.63$ and 0.61 , respectively), in the regression analysis the variable “GC content” proved unimportant.

The previously mentioned characteristics of intragenomic heterogeneity automatically included variation within and between chromosomes. However, the segments with different compositional and organizational characteristics can be more or less proportionally represented in different chromosomes of the same genome. By “CS-genomic states” one, therefore, may consider groups of genomic segments with similar nucleotide

composition and similar organization, i.e., with small H_t , H_o , and H_c within the group. Presumably, each of the five isochore groups in the human genome, in principle, may be further split into sub-groups with small H_o within and high H_o between the sub-groups. It would be natural to refer to these sub-groups as *CS-genomic states* or *organizational isochores*. If these hypothetical CS-genomic states are distributed proportionally between chromosomes, intra-chromosomal and inter-chromosomal heterogeneity (i.e., compositional and organizational differences over all pairs of chromosomes) will be close to the whole-genome heterogeneity. Alternatively, if some genomic state is heavily overrepresented in one or more chromosomes, intra-chromosomal heterogeneity of the corresponding chromosome (chromosomes) should decrease while its average inter-chromosomal distances to the remaining chromosomes should increase.

We have calculated the median inter-chromosomal H_t , H_o and H_c (i.e., total, compositional, and organizational differences for all pairs of chromosomes) and compared them with corresponding whole-genome heterogeneities for the vertebrate species, which were assembled into chromosomes at the moment of analysis (19 species). It was found that average inter-chromosomal H_t and H_o values are greater than the corresponding whole-genome H_t and H_o values for all analyzed genomes, presumably reflecting non-proportional distribution of CS-genomic states between chromosomes (Fig. 4a). On the contrary, average inter-chromosomal H_c was less than whole-genome H_c for all genomes except for birds (Fig. 4b). The greatest differences between average inter-chromosomal and whole-genome H_t and H_o were observed for primate genomes, presumably reflecting a greater chromosome structural specialization reached during the evolution of this group.

Heterogeneity of the coding and non-coding parts of genomic DNA

The genome is a mosaic of sequences with different compositional, organizational, and functional features. The organizational isochores are not available at this stage of our study, but we can investigate the possible relationship of heterogeneity of genomic CS states with its diverse functional features. The simplest separation of genome sequence into functional fractions is by considering coding DNA and repetitive sequences. By “coding” sequences we mean the genes (both exonic and intronic parts),

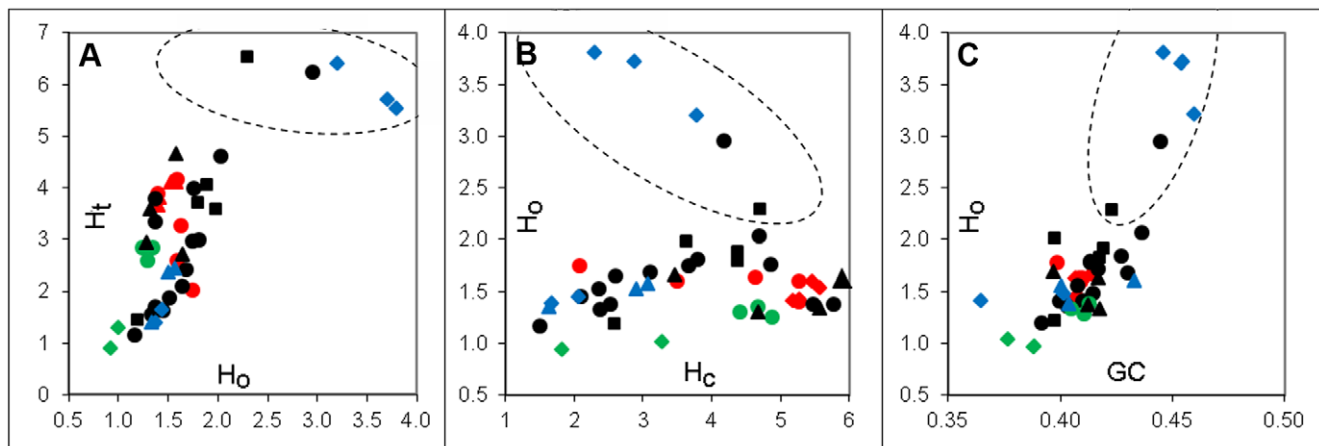
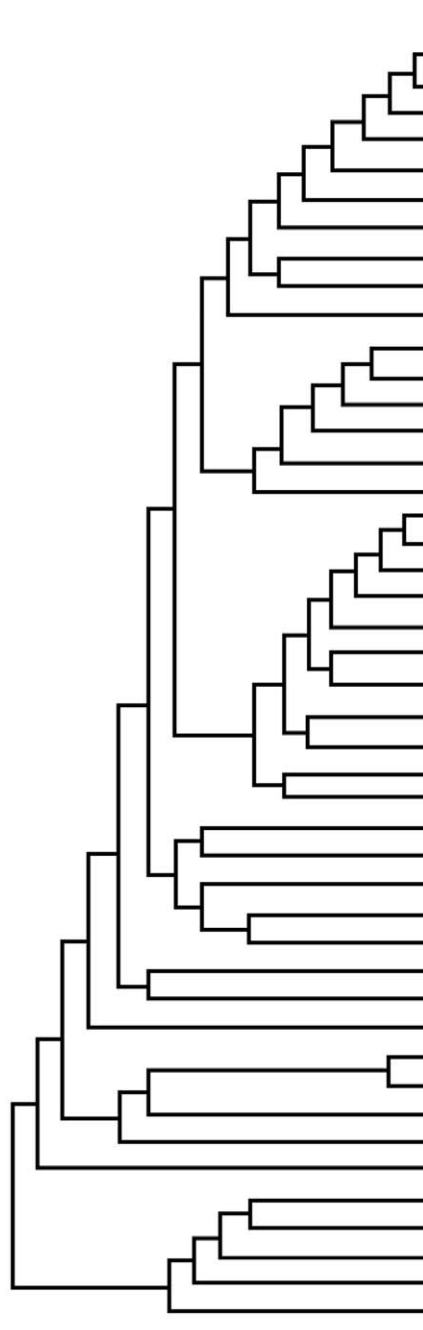


Figure 2. Intragenome heterogeneity and relationships between total and organizational (A), organizational and compositional (B) heterogeneity types, and organizational heterogeneity and genome GC content (C). Every series corresponds to one species group: ◆ apes; ▲ other primates; ■ Rodents; ▲ Ungulates; • other mammals; ◆ Marsupials; ● Birds; ◆ Fishes; ▲ other vertebrates. The species surrounded by a dashed ellipse appeared to be outliers due to their organizational heterogeneity (see explanation in the text). doi:10.1371/journal.pone.0032076.g002



	H_t	H_o	H_c	P^{**}	H_o/H_t
Homo sapiens	4.31	1.56	5.60	0.26	0.36
Pan troglodytes	4.01	1.45	5.30	0.26	0.36
Gorilla gorilla	4.29	1.62	5.48	0.26	0.38
Pongo pygmaeus	3.86	1.43	5.20	0.27	0.37
Macaca mulatta	4.07	1.42	5.30	0.26	0.35
Callithrix jacchus*	4.35	1.62	5.29	0.26	0.37
Tarsius syrichta*	2.22	1.77	2.12	0.74	0.80
Microcebus murinus*	3.45	1.65	4.66	0.29	0.48
Otolemur garnettii*	2.78	1.62	3.53	0.36	0.58
Tupaia belangeri*	3.15	1.77	3.70	0.50	0.56
Mus musculus	3.90	1.82	4.40	0.29	0.47
Rattus norvegicus	4.26	1.91	4.40	0.29	0.45
Cavia porcellus*	1.63	1.21	2.62	0.49	0.74
Spermophilus tridecemlineatus*	3.78	2.00	3.66	0.50	0.53
Dipodomys ordii*	6.52	2.30	4.72	0.27	0.35
Ochotona princeps*	6.22	2.96	4.20	0.31	0.48
Bos taurus	3.13	1.32	4.70	0.31	0.42
Sus scrofa	4.85	1.62	5.90	0.23	0.33
Vicugna pacos*	2.90	1.68	3.50	0.50	0.58
Tursiops truncatus*	4.17	1.78	4.88	0.40	0.43
Equus caballus	3.77	1.36	5.60	0.26	0.36
Canis familiaris	3.98	1.40	5.80	0.24	0.35
Felis catus*	4.79	2.05	4.73	0.29	0.43
Myotis lucifugus*	3.19	1.83	3.82	0.34	0.57
Pteropus vampyrus*	3.54	1.39	5.51	0.24	0.39
Erinaceus europaeus*	1.82	1.47	2.14	0.56	0.81
Sorex araneus*	2.62	1.70	3.14	0.39	0.65
Choloepus hoffmanni*	1.33	1.19	1.54	0.71	0.89
Dasyurus novemcinctus*	1.73	1.35	2.40	0.52	0.78
Echinops telfairi*	2.29	1.67	2.64	0.49	0.73
Loxodonta africana*	2.05	1.54	2.40	0.52	0.75
Procavia capensis*	1.88	1.39	2.57	0.54	0.74
Macropus eugenii*	1.10	0.96	1.85	0.67	0.87
Monodelphis domestica	1.50	1.03	3.30	0.43	0.69
Ornithorhynchus anatinus	2.64	1.60	3.10	0.42	0.61
Gallus gallus	3.03	1.37	4.70	0.29	0.45
Meleagris gallopavo	2.78	1.32	4.45	0.31	0.47
Taeniopygia guttata	3.02	1.27	4.90	0.30	0.42
Anolis carolinensis*	1.59	1.37	1.67	0.68	0.86
Xenopus tropicalis*	2.56	1.54	2.93	0.61	0.60
Takifugu rubripes*	5.70	3.72	2.87	0.43	0.65
Tetraodon nigroviridis	6.38	3.21	3.80	0.34	0.50
Gasterosteus aculeatus*	5.53	3.81	2.31	0.51	0.69
Danio rerio	1.59	1.41	1.70	0.67	0.89
Oryzias latipes	1.85	1.47	2.10	0.59	0.79

Figure 3. Whole-genome compositional, organizational and total heterogeneity for 45 vertebrate species ($\cdot 10^{-2}$). The deviation of the sampled median values is less than 0.0001 in all genomes for each of the three scores. Notes: * – incomplete genome sequences. We evaluate H_o by the same method as H_t , so one can directly compare H_o and H_t in corresponding columns of the table unlike H_c , which is calculated as a median difference in GC content between equal-length segments, so that corresponding H_c values are not directly comparable with H_o and H_t . P^{**} – The part of included segment pairs (i.e., those with very similar GC content) out of the total number of possible segment pairs.
doi:10.1371/journal.pone.0032076.g003

pseudogenes, RNA-coding DNA, etc. From the noncoding part of the genome, we extracted all known repetitive elements (using RepeatMasker as a tool); the remaining part of the noncoding DNA was referred to as *genome dark matter* (GDM). The analysis of every genomic fraction was conducted after isolating of corresponding sequences. Thus, to analyze the heterogeneity of GDM we first masked the repetitive elements and coding DNA in the sequence. The remaining parts of sequence were concatenated and subdivided into 100 kb segments (windows).

Three heterogeneity types: compositional, organizational, and total, of six vertebrate genomes (three mammalian, one avian, and two fish species) were assessed separately in three genome subspaces (genic, repetitive elements, and GDM) and in the whole sequence. Compositional heterogeneity of each of the three genomic fractions in the analyzed species was lower than the whole-genome compositional heterogeneity, with the only exception of the zebrafish genome, where H_c in the repetitive DNA was higher than in the whole genome (Table 2). However, among the

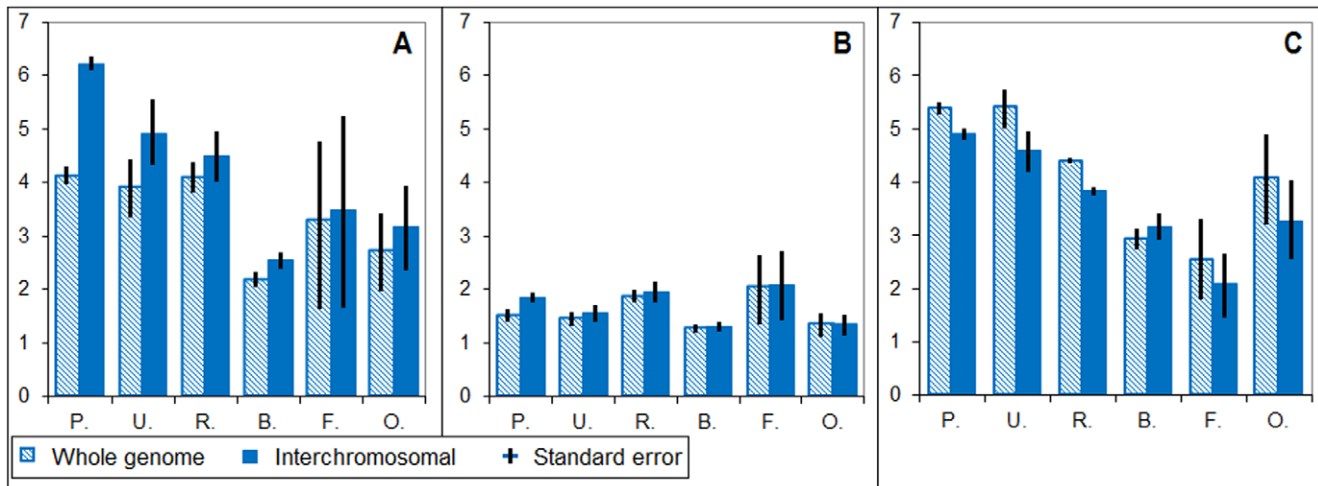


Figure 4. Whole-genome and inter-chromosomal heterogeneity scores H_t , H_o , and H_c for different groups of analyzed vertebrate genomes: (A) total heterogeneity, (B) organizational heterogeneity, (C) compositional heterogeneity; Legends: P. primates, U. ungulates, R. rodents, B. birds, F. fishes, O. other vertebrates; black bars correspond to standard error.
doi:10.1371/journal.pone.0032076.g004

compared six species the GC-poor zebrafish genome showed the lowest values of H_c in all of its functional fractions.

Unlike compositional heterogeneity, the organizational heterogeneity of repetitive and coding DNA in three mammalian genomes was higher than organizational heterogeneity of whole-genome sequences, while their GDM showed minimal H_o among all genome fractions. In the chicken genome, minimal value of H_o was found in coding DNA, whereas its repeats and GDM were more heterogeneous than the whole-genome sequence. Two analyzed fish genomes did not show similar differentiation of H_o among their genomic fractions: in zebrafish, the most heterogeneous genome part was repetitive DNA whereas in pufferfish – it was coding DNA. Interesting that the ratio of H_o/H_t was maximal in the repetitive DNA in all genomes, excluding zebrafish, where it was minimal in repeats and maximal in GDM. However, high H_o/H_t is characteristic of the zebrafish genome and all three of its fractions, caused by combined effect of

relatively high H_o and relatively low H_c of corresponding sequences.

The previously mentioned results on the variation of H_t , H_o , and H_c in whole-genome and inter-chromosomal comparisons (see Fig. 4) point to a possible non-proportional distribution of CS-genomic states between chromosomes. Thus, we may expect to find outlier chromosomes with respect to heterogeneity parameters. Analogously to the whole-genome analysis, we estimated inter-chromosomal H_o and H_c components for the three genomic fractions of the six vertebrate species and complemented these scores by corresponding intra-chromosomal H_o and H_c components. Based on the principal component analysis (PCA), an assessment of between-chromosome variation for each of the six species was conducted using the first two PCA components, which explained 67–90% of the variation in the 12-dimension space (H_o and H_c scores of intra- and inter-chromosomal heterogeneities for the three genomic fractions). One or two outlier chromosomes were

Table 2. Heterogeneity of different genome fractions ($\cdot 10^{-2}$).

	H_t^*	H_o^*	H_c^*	H_o/H_t	H_t^*	H_o^*	H_c^*	H_o/H_t	H_t^*	H_o^*	H_c^*	H_o/H_t
	<i>Homo sapiens</i>				<i>Mus musculus</i>				<i>Canis familiaris</i>			
Full sequence	4.31	1.56	5.60	0.36	3.90	1.82	4.40	0.47	3.98	1.40	5.80	0.35
Coding DNA	5.05	1.76	4.55	0.35	4.78	2.23	3.96	0.47	5.22	1.59	5.10	0.30
Repeats	4.34	1.94	3.19	0.45	3.39	2.11	2.22	0.62	2.72	1.62	2.69	0.60
GDM	3.23	0.83	5.22	0.26	3.13	1.53	3.32	0.49	4.61	1.07	5.59	0.23
	<i>Gallus gallus**</i>				<i>Danio rerio</i>				<i>Tetraodon nigroviridis</i>			
Full sequence	3.03	1.37	4.70	0.45	1.59	1.41	1.70	0.89	6.38	3.21	3.80	0.50
Coding DNA	2.19	1.24	2.90	0.57	1.41	1.29	1.19	0.91	7.09	3.40	3.24	0.48
Repeats	3.47	2.35	2.33	0.68	2.27	1.98	1.79	0.87	4.58	2.88	1.77	0.63
GDM	5.21	1.71	3.62	0.33	0.95	0.89	0.98	0.94	5.20	2.72	2.73	0.52

The deviation of the sampled median values is less than 0.0001 in all genomes for each of the three scores.

* H_c , H_o and H_t – compositional, organizational and total heterogeneity scores. Note that H_c is calculated as median of differences in GC content while H_t and H_o scores are evaluated based on comparison of compositional spectra (see section Methods);

**only large chromosomes 1–8 were taken into account.

doi:10.1371/journal.pone.0032076.t002

found in each of the analyzed mammalian and fish genomes (Fig. 5). Note, that dog outlier chromosomes include large regions syntenic with human outlier chromosomes: gene order of dog chromosome 9 is close to that of human chromosome 17; third of dog chromosome 26 is syntenic with human chromosome 22; and all these chromosomes are relatively GC-rich. In mouse, the outliers appeared to be sex chromosomes. We did not find in the literature any specific features of fish chromosomes that appeared as outliers in our analysis.

Additional details on relationships between heterogeneity components for different genome fractions can be seen in the example of human chromosomes. We were interested in assessing how the total and organizational intra- and inter-chromosomal heterogeneity depend on H_o and H_c scores of the chromosomal sub-sequences (coding, repetitive and GDM) and their GC content. For each human chromosome, intra- and inter-chromosomal H_o and H_c scores for each of the three sub-sequences were calculated together with their GC content, as potential “predictors” for whole-chromosome total and organizational heterogeneity. Using stepwise linear regression analysis, the significant predictors were determined separately for intra- and inter-chromosomal H_i and H_o characteristics (Fig. 6 and Table 3). Combinations of only 2–3 variables out of 9 explained 0.98 of the variation in intra- and inter-chromosomal H_i , and 0.9 and 0.99 of variation in their intra- and inter-chromosomal H_o , correspondingly. An interesting fact is that GC content of GDM proved to be important in explaining both intra- and interchromosomal H_o variation, while in the analysis of

corresponding H_i values, CG contents of all three sub-sequences were not significant. H_c of the repetitive DNA appeared a significant predictor for both characteristics of inter-chromosomal heterogeneity, total and organizational. However, for the organizational heterogeneity, the other score of repetitive DNA, H_o , was even more informative. In turn, H_o of coding DNA was significant for both total and organizational intra-chromosomal heterogeneity. H_c of coding DNA was important only for total intra-chromosomal heterogeneity, but its importance there was much higher than the importance of H_o of coding DNA. The H_o score of GDM was important for intra- and inter-chromosomal whole-sequence total heterogeneity.

Effect of GC content on sequence organization

The previously mentioned comparisons of vertebrate genomes revealed positive correlation between GC content and H_o values (see Fig. 2c). Is this tendency of genome organization also expressed on the intra-genome level (between segments of the same genome)? To address this question, we analyzed 14 vertebrate genomes from diverse taxonomic groups. As before, we subdivided the whole-genome sequence of each species into segments of 100 kb length and classified the segments, regardless of their chromosomal residence, into groups according to GC content (with 1% of GC content width per group). The small groups from the tails of GC distribution were removed from the analysis. In 3 out of 14 tested species the highest H_o was observed in segments with $GC \approx 0.5$ and 6 other species showed the highest

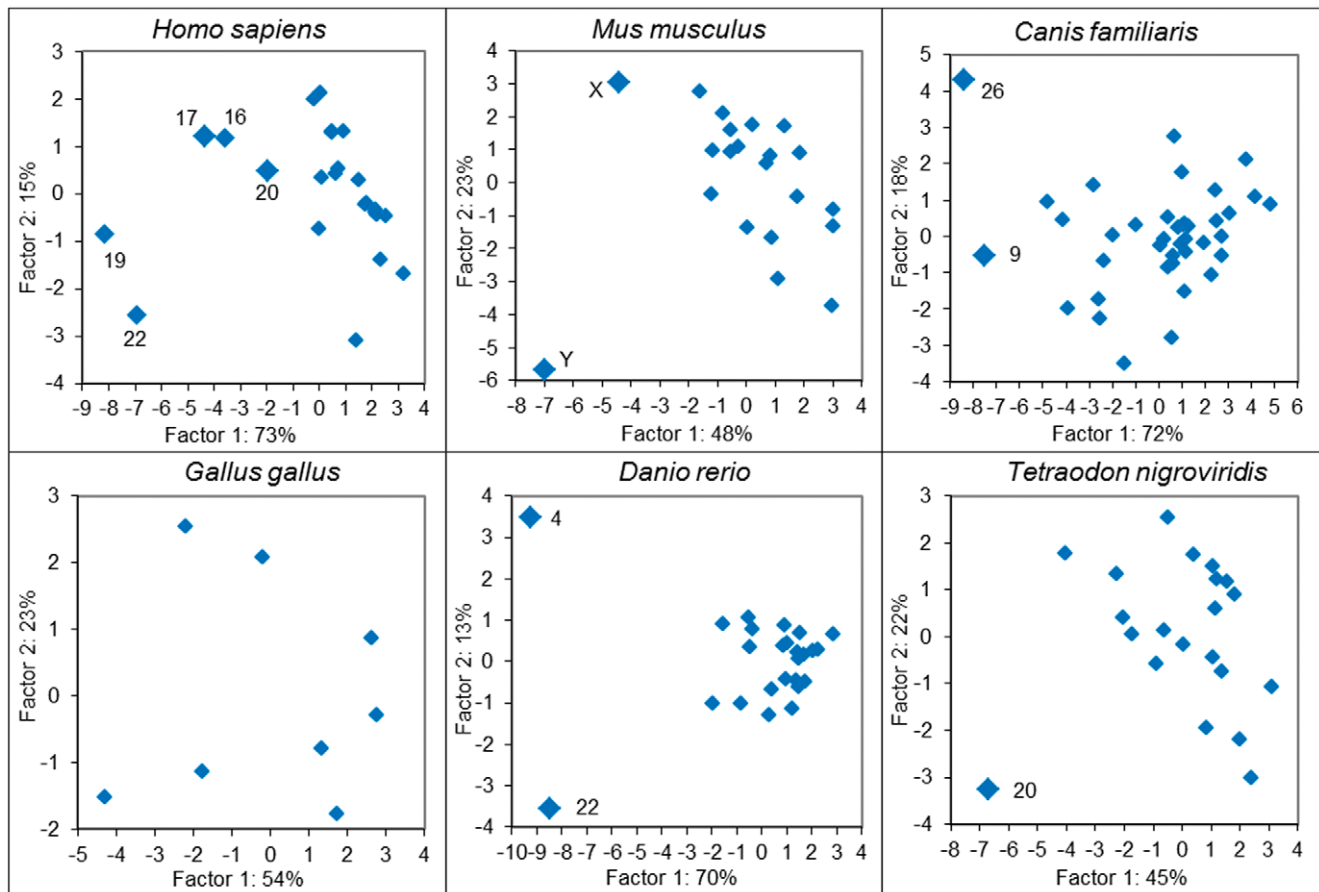


Figure 5. Principal component analysis of vertebrate chromosomes for H_o and H_c scores evaluated in three genome fraction (coding sequences, repetitive elements, and GDM).

doi:10.1371/journal.pone.0032076.g005

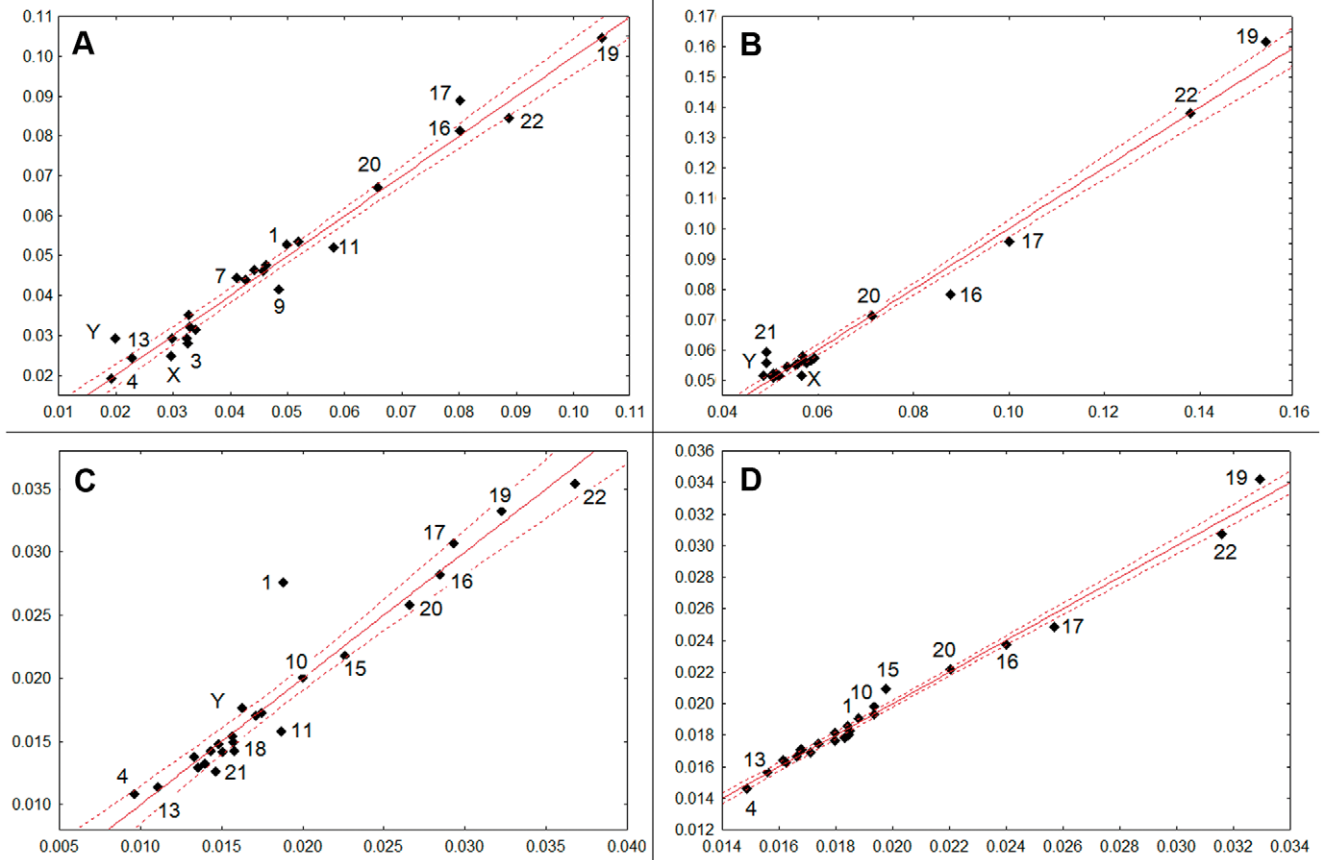


Figure 6. Predicted and observed values of human intra- and inter-chromosomal heterogeneity estimates. Axes X: predicted values, axes Y: observed values. (A) intra-chromosomal H_t ; (B) average inter-chromosomal H_t ; (C) intra-chromosomal H_o ; (D) average inter-chromosomal H_o . doi:10.1371/journal.pone.0032076.g006

H_o at GC = 0.48–0.50 or 0.50–0.52 (Fig. 7). The pika genome had an “extreme” point at CG = 0.45 and the wallaby, at CG = 0.46. Three other species did not have any “extreme” point and showed the highest H_o in their GC-richest segments. Organizational heterogeneity of GC-rich segments in fish genomes was higher than in corresponding groups of mammalian, birds, and marsupial genomes.

In previous studies, we employed a reshuffling test to assess the significance of genome organization compared to random

sequences with the same nucleotide abundances [59]. This test was applied here (with some modifications) to vertebrate species in order to assess the nonrandomness of genome organization in terms of compositional spectra and its dependence on GC content. To conduct reshuffling, each letter of segment sequence, starting from the first position, was swapped with a letter from a randomly selected position within the segment. This procedure returns a new sequence of the segment with unchanged letter abundances but a random letter order. Therefore, in addition to the initial (natural)

Table 3. The most significant predictors of intra- and inter-chromosomal heterogeneity scores H_t and H_o .

Intrachromosomal H_t	t(20)	p-value	Interchromosomal H_t	t(21)	p-value
Coding DNA H_c	11.34	$<5 \times 10^{-7}$	Repetitive DNA H_c	12.35	$<5 \times 10^{-7}$
GDM H_o	6.78	$<5 \times 10^{-6}$	GDM H_o	5.19	$<4 \times 10^{-5}$
Coding DNA H_o	3.78	0.001			
Adjusted R^2	0.98		Adjusted R^2	0.98	
Intrachromosomal H_o	t(21)	p-value	Interchromosomal H_o	t(20)	p-value
GC content of GDM	3.77	0.001	Repetitive DNA H_o	8.95	$<5 \times 10^{-7}$
Coding DNA H_o	3.05	0.006	Repetitive DNA H_c	4.65	$<2 \times 10^{-4}$
			GC content of GDM	2.15	0.043
Adjusted R^2	0.91		Adjusted R^2	0.99	

doi:10.1371/journal.pone.0032076.t003

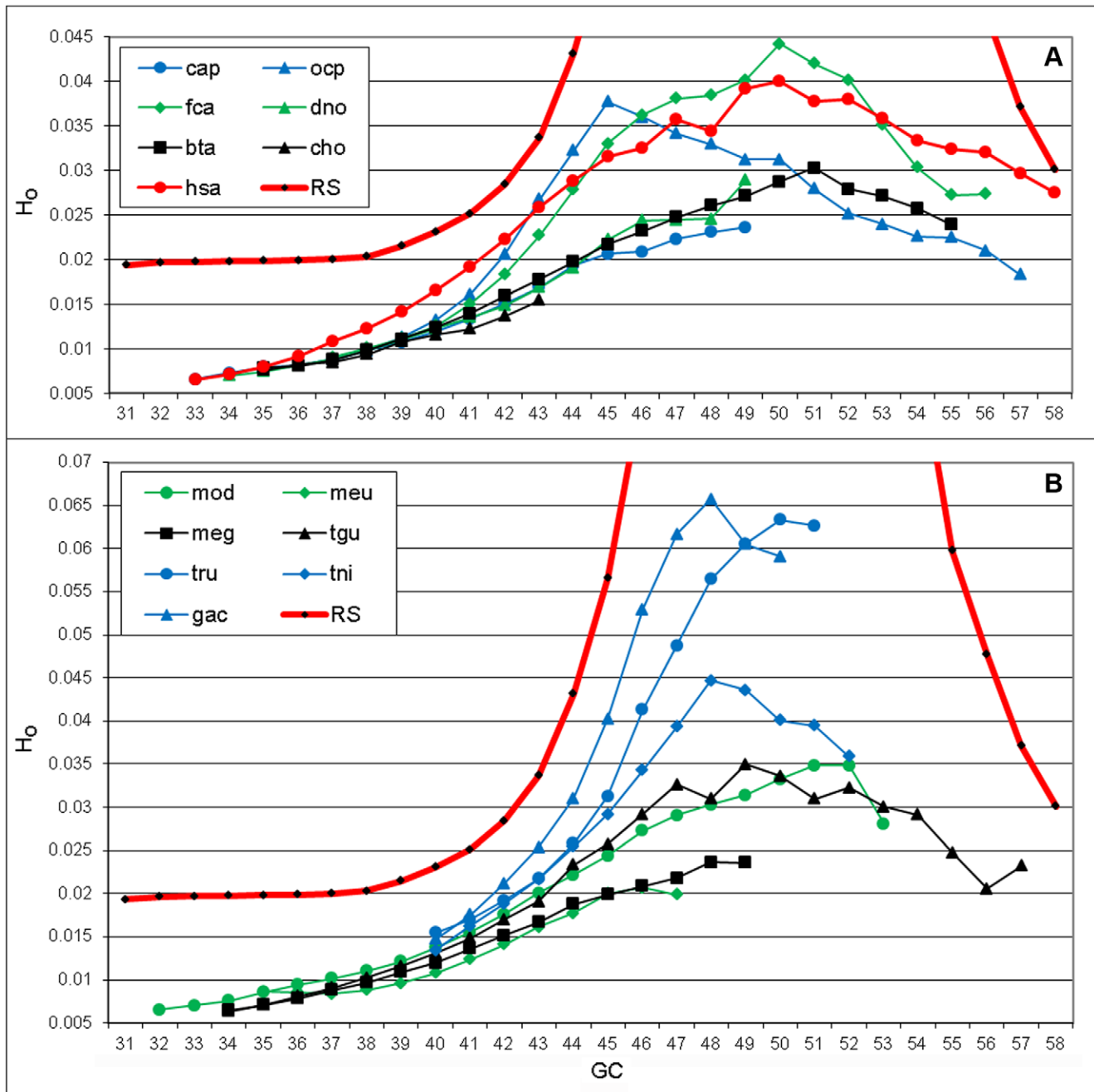


Figure 7. Organizational heterogeneity of natural sequences of different vertebrate genomes. (A) mammals, (B) non-mammalian species. Legend: **cap** guinea pig (*Cavia porcellus*), **dno** armadillo (*Dasypus novemcinctus*), **gac** stickleback (*Gasterosteus aculeatus*), **meu** wallaby (*Macropus eugenii*), **mod** opossum (*Monodelphis domestica*), **ocp** American pika (*Ochotona princeps*), **tru** fugu (*Takifugu rubripes*), **tni** pufferfish (*Tetraodon nigroviridis*), **hsa** human (*Homo sapiens*), **bta** cow (*Bos taurus*), **cho** cloth (*Choloepus hoffmanni*), **fca** cat (*Felis catus*), **meg** turkey (*Meleagris gallopavo*), **tgu** zebra finch (*Taeniopygia guttata*), **RS** reshuffled sequence. doi:10.1371/journal.pone.0032076.g007

sequences, we calculated the heterogeneity within each group of segments using the reshuffled sequences.

As expected, after reshuffling, the groups of segments with the same GC content showed very similar values of H_0 , no matter to which genome they belonged. Likewise, they showed a strong correlation between their H_0 and deviation of GC content from 0.5. The following straightforward argument explains this fact. For simplicity, we assume that $\%A = \%T$ and $\%C = \%G$ in single-strand DNA, in accordance to the Chargaff's second parity rule

[56,57]. Violation of this rule was found in some regions related to replication origin, gene location and transcriptional hot spots [60], and the calculations below can be extended to such cases with qualitatively similar results.

Let the frequencies of letters A and T in the sequence be equal to $n_1/2 = p_1$, and letters C and G to $n_2/2 = p_2$ (we assume that n_1 and n_2 are even); the length of the entire sequence is $n_1 + n_2 = n$. Thus, $p_1 + p_2 = n/2$. It is easy to calculate the number N of all possible different sequences with these parameters:

$$N(n_1, n_2) = \binom{n}{n_1} \binom{n_1}{p_1} \binom{n_2}{p_2} = \frac{n!}{(p_1!)^2 (p_2!)^2}$$

For symmetry, we write the value N as a function of two variables, although it is clear that n_1 , for example, can be directly calculated from the values n and n_2 . It is easy to show that this expression reaches its greatest value at $n_1 = n_2$. Using the known asymptotic approximations for the binomial coefficients we can calculate

$$N(n_1, n_2) \sim \frac{2^n n^n \sqrt{n}}{n_1^{n_1+1} n_2^{n_2+1}} \text{ or } N(x) \sim \frac{2^n n^n \sqrt{n}}{(xn)^{xn+1} [(1-x)n]^{(1-x)n+1}},$$

where α is GC content of segments.

As follows from these calculations, sequences with GC content close to 0.5 have a much higher potential of variation compared to sequences with other GC contents. In accordance with the used method of CS-distance calculation, heterogeneity (i.e., median CS distance) of large enough group of random segments with strictly identical proportions of all four nucleotides (25% of each) should be equal to 0.5 (i.e., rank correlation coefficient between compositional spectra of such segments should be equal or close to zero). All random segments with deviations from identical proportions of all four nucleotides, such as different CG content of analyzed segment groups, should result in spectra with $d < 0.5$ (i.e., with positive rank correlation) because of overrepresentation of certain nucleotide(s). In fact, H_o of the reshuffled sequences for the analyzed groups with GC ≈ 0.5 ranged in different genomes from 0.33 to 0.42, presumably due to small-scale variability of GC in these groups (from 0.495 to 0.505) and possible violations of the Chargaff's second parity rule for either G vs. C, or A vs. T, or both.

In general, the natural sequences of the tested species proved to be much less heterogeneous than corresponding random sequences with the same GC content, reflecting the high level of genome organization. Moreover, the difference between the reshuffled and natural sequences was found to be maximal for the groups of sequences with GC in the range 0.49–0.51 and decreased with the increase in |GC-0.5|. This may indicate that anti-entropic evolutionary forces shaping the genome organization have imposed stronger impact on regions with equal abundances of all four nucleotides compared to other regions.

In addition to above analysis of deviation of the full sequence of human genome from corresponding random sequences for each GC window, the same analysis was conducted for separate genome fractions: coding, repetitive sequences, and GDM. Due to different compositional limitations we obtained different numbers of GC-groups: 23 for the coding DNA (CG range from 0.33 to 0.55), 15 for repetitive DNA (CG range from 0.36 to 0.50), and 20 for the GDM (CG range from 0.31 to 0.50). As before, natural sequences were considerably less heterogeneous than the reshuffled ones; the largest difference was observed in GDM, implying that GDM is *the most organized DNA fraction in the human genome* (Fig. 8a). The observed differences between H_o values of reshuffled sequences of separated genome fractions (Fig. 8b) can presumably be explained by varied deviations from Chargaff's parity rule in these fractions.

Discussion

The problem of genome heterogeneity has been widely discussed in the last decade [1,8,20,61–65]. A considerable part of the studies is devoted to GC heterogeneity and its relationship

with the functional and evolutionary heterogeneity. GC content displays a wide interspecific variation as well as high within-genome heterogeneity, especially in mammals, where it is presented in the form of GC isochores [66]. GC content is strongly correlated with biological features of the genome organization such as distribution of various classes of repeated elements, gene density, level and tissue-specificity of transcription, and mutation rate. The available data point to a strong correlation between GC content and recombination. This connection may be caused by increased recombination in GC-rich regions. An alternative assumption is that meiotic recombination machinery is responsible for the evolution of GC distribution due to the bias of mismatch repair within the gene conversion tracts in favor of GC, which preferentially converts A/T into G/C at sites that are heterozygous for AT and GC - the biased gene conversion hypothesis [67,68]. A possibility that another factor, affecting both the evolution of intragenomic heterogeneity in GC content and recombination, is responsible for their positive association cannot be excluded.

We used the average difference of GC content between genome segments as an indicator of its compositional heterogeneity, which can be approximately compared with GC-profiles discussed in previous publications. Our results on interspecific variation of compositional heterogeneity, as expressed by the differences in H_c , in general correspond well to previous studies based on assessment of isochore structure of vertebrate genomes. In particular, closely related primate genomes displayed similar and relatively high values of compositional heterogeneity H_c corroborating with similar isochore patterns [69]. Generally, genomes of cold-blooded vertebrates are less heterogeneous than genomes of warm-blooded vertebrates in terms of H_c (our results) and isochore organization [69]. Although bird genomes have wider GC profiles due to the presence of very GC-rich H4 isochores [70], in terms of H_c they are less heterogeneous than primate genomes. The last discrepancy can be explained by a small proportion of GC-rich segments in bird genomes and their relatively weak contribution to the whole-genome H_c estimate.

Organizational heterogeneity of the studied genomes also varied among analyzed species: relatively high values were observed in 3 out of 5 fish genomes and in 2 out of 32 studied mammal genomes, while the lowest values were found in marsupials, sloth, and guinea pig (see Table 1). It is noteworthy that similar values of H_o , H_c , and H_t are displayed by closely related species in two groups of organisms (birds and apes), while in two other groups (rodents and fishes) all heterogeneity types were more variable. In lower (non-human) primates, ungulates, and marsupials, low intra-group variation of H_o values was observed together with considerable variation of H_c and H_t , presumably reflecting higher functional importance of organizational patterns in genome structure and evolution. Compositional and organizational similarity of different segments of the same genome indicates the existence of genome structural patterns.

Similar to H_c values, the absolute H_o values are relatively small, especially when compared to corresponding reshuffled sequences. However, their variability among genomes, chromosomes or chromosome segments may be an important source of information for the investigations of genome structure, functions and evolution. We found that intra-genome variability of the compositional and organizational patterns may contain sufficient information for revealing conserved parallelism upon a modified dot-plot genome comparison based on compositional spectra analysis [48]. Remarkably, GDM by itself allows to get a rather good coverage in such dot-plot comparisons between humans and apes [Kirzhner et al. unpublished results], implying evolutionary conservation in

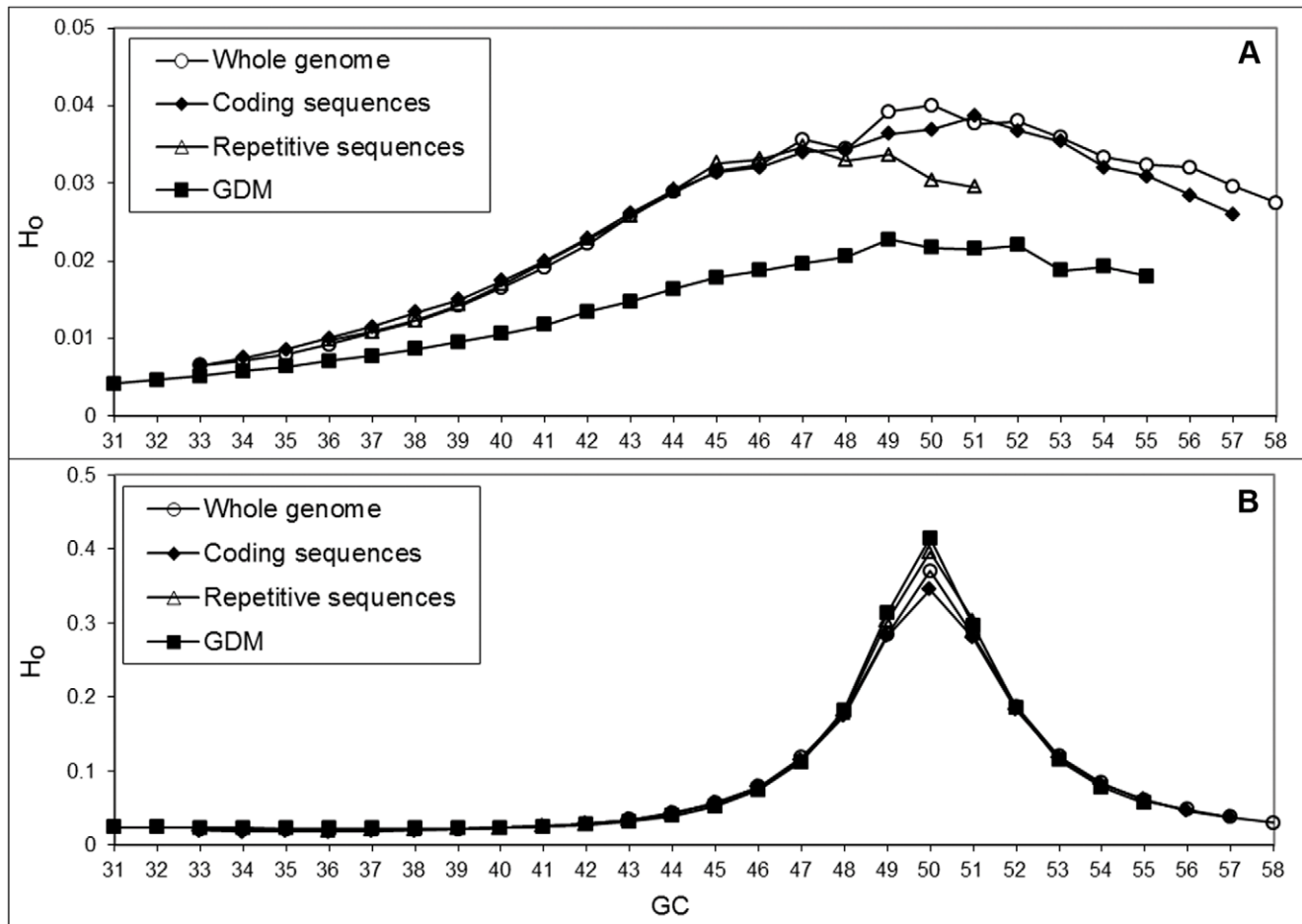


Figure 8. Organizational heterogeneity of natural sequences of different genome fractions and corresponding reshuffled sequences with varied GC content: (A) natural sequences, (B) reshuffled sequences.

doi:10.1371/journal.pone.0032076.g008

intragenomic variation of compositional and organizational patterns that extends beyond the genic space. Evolutionary conservation of noncoding DNA among vertebrates was intensively studied during the last decade [52,71–73]. It was found that some noncoding elements are even more conserved than genes. Although the location of a part of these elements in the intergenic regions does not necessarily point to their role as gene regulatory elements, many examples are known in which this is indeed the case. Some estimates indicate that the vertebrate-specific ultra-conserved noncoding regions may be under selection comparable to or stronger than selection on protein-coding regions [74]. The total length of evolutionary conserved noncoding DNA is about 2.5–5% of the mammalian genome length, whereas our definition of GDM leads to several-fold larger estimates, implying that evolutionary less conserved sequences may be the majority in GDM.

The results of multiple regression analysis of the whole-genome heterogeneity permits the conclusion that inter-specific variation of total genome heterogeneity characteristics as measured by CS analysis depends both on genome composition and organization. A similar effect was observed in the analysis of inter- and intra-chromosomal heterogeneities of different genomes, including the analysis of whole-chromosome sequences and their separate subsequences (coding, repetitive, and GDM sequences). This can be explained by assuming the existence of “organizational isochores”,

i.e., segments with different organization within the compositional (GC) isochores. Similar ideas of organizational heterogeneity on a variable range of scales lower than those of GC isochores were proposed earlier based on the concept of long-range correlations in DNA sequences [75]. The segments with different composition obviously have different spectra, but their organizations could not be fully compared by the proposed method. A CS comparison of sequences with the same (or similar) nucleotide composition allows detecting groups of segments with similar CS organization, which may be related to different structural, functional, and evolutionary features of genomic sequences.

The comparison of intragenome heterogeneity values with and without taking into account intrachromosomal heterogeneities suggests disproportional distributions of CS-genomic states between chromosomes. Moreover, PCA and multiple regression analysis reveals chromosomes dissimilarities of the same genome by their H_0 and H_c values calculated on diverse genome fractions (see Fig. 5). Notably, for human genome, dissimilarities of the same chromosomes (16, 17, 19, 20 and 22) were observed by several authors. It was found that these chromosomes are GC-, CpG-, and gene-rich [76–78] and differ from other human chromosomes by enrichment with various repeats [77–80]. According to numerous publications, these chromosomes also differ by their structural and evolutionary features. Malcolm et al. [7] clustered human chromosomes into families according to genic mutation rates

and noted a significant distinction of chromosomes 19 and 21 from other chromosomes. Surprisingly, our results on human inter-chromosomal organizational CS-heterogeneity (see Fig. 6) corroborate with inter-chromosomal heterogeneity with respect to the proportion of sequences under selection [81]. Authors marked chromosomes 19 and 22 as outliers due to relatively low proportions of both total and nonprotein-coding sequences under selection in spite of very high proportions of coding DNA in these chromosomes. Buschiazzi et al. [63] explain significant differences between human chromosomes related to the fraction of DNA alignable with other vertebrate genome sequences, which is lowest in chromosomes X, 19 and 22 and highest in chromosomes 13 and 18; this also implies dissimilarities in the rates of chromosome evolution. Correlation between chromosomal and regional variation in recombination rates and corresponding compositional features (GC isochores) of human chromosomes [17] fits well the biased gene conversion hypothesis [67].

Most aspects of intra- and inter-chromosomal heterogeneity correlate with chromosome GC content. For clarification of the influence of GC content on the sequence organizational heterogeneity, we compared heterogeneity of genome segments with different GC contents. It was found that organizational heterogeneity of segments increases with their GC content up to some “extreme point” and then decreases. This could be related to the expected relationship between the observed and potential variability of sequences as a function of GC content: higher variation is expected when GC is close to 0.5. However, in several cases, the greatest H_0 values were found in segment groups with $GC < 0.5$. An alternative explanation of greater organizational variability of CG-rich segments may be based on the fact that classical CG-isochores are not really compositionally homogeneous (i.e., GC content within isochores is variable). Moreover, GC-rich isochores are more compositionally heterogeneous than those that are GC-poor [10]. However, our “GC-slices” are much thinner than classical GC-isochores, and the variability of GC-content within every group of segments does not exceed 0.01; thus, different compositional heterogeneity of segments cannot be considered a cause of the observed differences in their organizational heterogeneity. Therefore, isochores of each GC range can be further classified into sub-sets, with similar CS patterns within the sub-sets and those that are dissimilar between sub-sets. Such sub-sets superimposed on the classical GC isochores can be referred to as “organizational isochores”.

High gene and SINE density is characteristic of GC-rich regions in most vertebrate genomes [81–82]. One may assume that a high level of organization of GC-rich regions (expressed as strong deviation from reshuffled sequences) is predetermined by the presence of well-organized coding DNA or highly repetitive sequences. However, the analysis of diverse fractions of the human genome (coding, repetitive sequences, and GDM) shows the same heterogeneity distribution pattern in each of these fractions (see Fig. 8a). Moreover, coding DNA appears to be the most heterogeneous fraction, whereas GDM, which does not have known genes and repeats (otherwise they would be masked by our masking pretreatment), shows the highest degree of organizational homogeneity, hence the highest deviation from the reshuffled (random) sequence. This may be partially explained by the presence of a fraction of highly conserved non-protein-coding DNA that in our analysis should be a part of GDM. We, therefore, speculate that the high conservation of numerous short intergenic sequences may be just the tip of the iceberg, namely, a more massive, albeit more fuzzy, organizational conservation (see also [73]).

The results of our analysis fit the concepts of nonrandom organization of genetic material within and between chromosomes and within the nucleus, represented in terms of chromosome fields, genomic neighborhoods, and expression domains, which are related to the three-dimensional architecture of the eukaryotic genome [83–85] and changes in nuclear compartmentalization during transcriptional activation and in the course of evolution [86,87]. Besides GC content, the organization peculiarities of genomic neighborhoods may be affected by different families of repeated elements, chromosomal position relative to centromere or telomere, and the distance to heterochromatin blocks. Organizational heterogeneity revealed by CS analysis may also be related to this phenomenon. The high correspondence of the CS-heterogeneity scores between the three types of DNA (coding, repetitive, and GDM) permits us to assume that CS-intragenomic heterogeneity may have functional relevance to genome regulation rather than only to reflect different composition and organization of gene-rich and gene-poor regions or local GC enrichment caused by biased gene conversion. Of special interest for such interpretation is that natural GDM sequences display the highest deviation from the corresponding reshuffled sequences, which may hint to a potential role of their organization in the formation of genomic neighborhoods and three-dimensional genome nuclear architecture, hence the regulation of genome dynamics and transcription.

The similarity of effects observed in diverse DNA spaces (coding, repetitive, and GDM) provides evidence of the inter-relatedness of different structural and functional genomic elements. Earlier we showed high positive correlation between interspecific distances based on compositional spectra analysis and corresponding distances based on orthologous genes encoding for information-processing enzymes involved in replication, recombination, DNA repair, and transcription [88]. We hypothesized that high structuring of genome sequences may be associated with intracellular mechanisms where interactions between template DNA and corresponding information processing enzymatic machinery play a leading role. These interactions may derive from several mechanisms: (a) DNA polymerase may have a key role in the evolution of its product (DNA sequence); (b) the structure of repair-recombination enzymes may be evolutionarily more sensitive and “responsive” to changes in its predominant organizational pattern (referred to as “genome dialect”) than proteins involved in structural and metabolic processes; (c) changes in DNA sequences (caused by mutation, recombination, and transposable elements) displayed as changes in genome dialect may improve the mutual correspondence of genome organization and its information processing machinery. These results were obtained on genomes of 20 species of proteobacteria (i.e., with a predominant proportion of protein coding DNA in the genome sequence). The remarkable similarity of the genome heterogeneity patterns across the three considered spaces of vertebrate genomes (coding DNA, repetitive DNA and genome dark matter) found in the current study indicates that co-evolution between information processing enzymes and genome dialect may be a reasonable hypothesis for higher eukaryotes. However, the “genome dialect” concept does not imply unification of organizational variation along the noncoding part of the genome. Indeed, we recently found that compositional and organizational variability patterns in GDM contain sufficient information to reveal conserved parallelism upon a modified dot-plot genome comparison between humans and apes [Kirzhner et al. unpublished results].

The obtained results suggest that the heterogeneity of genomic sequences is a product of a complex interplay between organizational and compositional heterogeneities. The influence of compositional heterogeneity is obvious because any two

sequences with diverse nucleotide composition necessarily differ in their compositional spectra. However, according to our results, any two sequences with a very similar composition may still have very different compositional spectra, which probably would be better named “organizational spectra” because they reflect organizational rather than compositional differences of the sequences. Simultaneously, organizational heterogeneity of human genome segments with the same nucleotide composition displays a strong dependence on the deviation of GC content from 50%. Permutation tests indicate that natural genomic sequences do not utilize the whole potential of sequence variation “offered” by their composition: their organizational heterogeneity proved much lower than the heterogeneity of corresponding reshuffled sequences with the same composition. The analysis of inter- and intra-chromosomal heterogeneities forces us to assume the existence of some basic (predominant) organizational patterns in each genome. The degree of deviations of the organizational patterns (genomic states) presented in each chromosome from the predominant patterns determines the level of intra- and, therefore, inter-chromosomal heterogeneities.

We further speculate that in parallel to compositional (GC) isochores, genome sequences deviating from the basic organizational pattern(s) also form a mosaic structure of “organizational” isochores calling for corresponding genome segmentation analysis. An example of such an approach based on the analysis of abundances of tri-nucleotide words was provided by Bingham and

co-authors [89]. Chromosomal segments with organizational patterns that are similar in their coding, repetitive and GDM subsequences, may represent the aforementioned neighborhoods involved in spatial organization of the nucleus [87,90]. A new analysis is underway in our lab aimed at testing for possible association between genome-wise distribution of organizational “isochores” and various evolutionary and functional features, such as rate of gene duplications, indels and SNPs, hotspots of transcription, rate of mutation and recombination, and distribution of linkage disequilibrium blocks.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions, Dr. Zeev Frenkel for useful discussions and Alexander Frenkel for help in improving the scripts. This work was supported by the Israeli Ministry of Absorption (S.F. and V.K.). S.F. was also supported by a fellowship for excellence in Converging Technologies Program of The Council for Higher Education.

Author Contributions

Conceived and designed the experiments: VK AK. Performed the experiments: SF. Analyzed the data: SF VK. Contributed reagents/materials/analysis tools: VK SF AK. Wrote the paper: SF AK. Designed the software used in analysis: SF VK.

References

- Karlin S, Ladunga I, Blaisdell BE (1994) Heterogeneity of genomes: measures and values. *Proceedings of the National Academy of Sciences of the USA* 91: 12837–41.
- Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD (2003) A unification of mosaic structures in the human genome. *Human Molecular Genetics* 12: 2411–5. doi:10.1093/hmg/ddg251.
- Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Research* 15: 1468–76. doi:10.1101/gr.4398405.
- Schmegner C, Hameister H, Vogel W, Assum G (2007) Isochores and replication time zones: a perfect match. *Cytogenetic and Genome Research* 116: 167–72. doi:10.1159/000098182.
- 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. *Molecular Biology and Evolution* 24: 2385–99. doi:10.1093/molbev/msm181.
- Eory L, Halligan DL, Keightley PD (2010) Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Molecular Biology and Evolution* 27: 177–92. doi:10.1093/molbev/msp219.
- Malcom CM, Wyckoff GJ, Lahn BT (2003) Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Molecular Biology and Evolution* 20: 1633–41. doi:10.1093/molbev/msg178.
- Bernardi G (1989) The isochore organization of the human genome. *Annual Review of Genetics* 23: 637–61.
- Li W (2001) Delineating relative homogeneous G+C domains in DNA sequences. *Gene* 276: 57–72.
- Li W (2002) Are isochore sequences homogeneous? *Gene* 300: 129–39.
- Costantini M, Clay O, Federico C, Saccone S, Auletta F, et al. (2007) Human chromosomal bands: nested structure, high-definition map and molecular basis. *Chromosoma* 116: 29–40. doi:10.1007/s00412-006-0078-0.
- Costantini M, Bernardi G (2008) Replication timing, chromosomal bands, and isochores. *Proceedings of the National Academy of Sciences of the USA* 105: 3433–7. doi:10.1073/pnas.0710587105.
- Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, et al. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research* 20: 761–70. doi:10.1101/gr.099655.109.
- Sémon M, Mouchiroud D, Duret L (2005) Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Human Molecular Genetics* 14: 421.
- Versteeg R, Schaik BDC van, Batenburg MF van, Roos M, Monajemi R, et al. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Research* 13: 1998–2004. doi:10.1101/gr.1649303.
- Vinogradov AE (2005) Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends in Genetics* 21: 69–43.
- Jensen-Seaman MI, Furey TS, Payseur Ba, Lu Y, Roskin KM, et al. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research* 14: 528–38. doi:10.1101/gr.1970304.
- Schmegner C, Hoegel J, Vogel W, Assum G (2007) The rate, not the spectrum, of base pair substitutions changes at a GC-content transition in the human NF1 gene region: implications for the evolution of the mammalian genome structure. *Genetics* 175: 421–8. doi:10.1534/genetics.106.064386.
- Melodelima C, Gautier C (2008) The GC-heterogeneity of teleost fishes. *BMC Genomics* 9: 632. doi:10.1186/1471-2164-9-632.
- Nekrutenko A, Li W-H (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research* 10: 1986–1995. doi:10.1101/gr.10.12.1986.
- Costantini M, Bernardi G (2008) The short-sequence designs of isochores from the human genome. *Proceedings of the National Academy of Sciences of the USA* 105: 13971–6. doi:10.1073/pnas.0803916105.
- Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. *Journal of Molecular Biology* 196: 261–82.
- Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. *Genomics* 13: 1095–107.
- Jabbari K, Bernardi G (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* 224: 123–7.
- Vinogradov AE (2003) Isochores and tissue-specificity. *Nucleic Acids Research* 31(17): 5212–20. doi: 10.1093/nar/gkg69.
- Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP (2010) Redefining CpG islands using hidden Markov models. *Biostatistics (Oxford, England)* 11: 499–514. doi:10.1093/biostatistics/kxq005.
- Karlin S, Cardon LR (1994) Computational DNA sequence analysis. *Annual Review of Microbiology* 48: 619–54. doi:10.1146/annurev.mi.48.100194.003155.
- Nussinov R (1980) Some rules in the ordering of nucleotides in the DNA. *Nucleic Acids Research* 378: doi:10.1093/nar/gkn942.
- Petrokovski S, Hirshon J, Trifonov E (1990) Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *Journal of Biomolecular Structure and Dynamics* 7: 1251–1268.
- Trifonov EN, Brendel V (1986) *Gnomic: A dictionary of genetic codes*. Rehovot: Balaban. 272 p.
- Sims GE, Jun S-R, Wu Ga, Kim S-H (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the USA* 106: 2677–82. doi:10.1073/pnas.0813249106.
- Vinga S, Almeida J (2003) Alignment-free sequence comparison — a review. *Bioinformatics* 19: 513–523. doi:10.1093/bioinformatics/btg005.
- Hedges SB (2002) The origin and evolution of model organisms. *Nature Reviews. Genetics* 3: 838–49. doi:10.1038/nrg929.
- Rosen G, Garbarine E, Caseiro D, Polikar R, Sokhansanj B (2008) Metagenome fragment classification using n-mer frequency profiles. *Advances in Bioinformatics* 2008. doi:10.1155/2008/205969.

35. Rosen GL, Reichenberger ER, Rosenfeld AM (2011) NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27: 127–9. doi:10.1093/bioinformatics/btq619.
36. Csurös M, Noé L, Kucherov G (2007) Reconsidering the significance of genomic word frequencies. *Trends in Genetics: TIG* 23: 543–6. doi:10.1016/j.tig.2007.07.008.
37. Sivaraman K, Seshasayee ASN, Swaminathan K, Muthukumar G, Pennathur G (2005) Promoter addresses: revelations from oligonucleotide profiling applied to the *Escherichia coli* genome. *Theoretical Biology & Medical Modelling* 2: 20. doi: 10.1186/1742-4682-2-20.
38. Helden J van, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281: 827–42. doi:10.1006/jmbi.1998.1947.
39. Mariño-Ramírez L, Spouge JL, Kanga GC, Landsman D (2004) Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Research* 32: 949–58. doi:10.1093/nar/gkh246.
40. Healy J, Thomas EE, Schwartz JT, Wigler M (2003) Annotating large genomes with exact word matches. *Genome Research* 13: 2306. doi:10.1101/gr.1350803.
41. Chapus C, Dufraigne C, Edwards S, Giron A, Fertl B, et al. (2005) Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology* 5: 63. doi:10.1186/1471-2148-5-63.
42. Dufraigne C, Fertl B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Research* 33: e6. doi:10.1093/nar/gmi004.
43. Karlin S (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends in Microbiology* 9: 335–43.
44. Li W, Stolovitzky G, Bernaola-Galvan P, Oliver JL (1998) Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Research*. pp 916–928. doi:10.1101/gr.8.9.916.
45. Kent WJ (2002) BLAT — The BLAST-like alignment tool. *Genome Research* 12: 656. doi:10.1101/gr.229202.
46. Höhl M, Ragan MA (2007) Is multiple-sequence alignment required for accurate inference of phylogeny? *Systematic Biology* 56: 206–221. doi:10.1080/10635150701294741.
47. Kirzhner V, Paz A, Volkovich Z, Nevo E, Korol A (2007) Different clustering of genomes across life using the A-T-C-G and degenerate R-Y alphabets: early and late signaling on genome evolution? *Journal of Molecular Evolution* 64: 448–56. doi:10.1007/s00239-006-0178-8.
48. Kirzhner V, Frenkel S, Korol a (2011) Minimal-dot plot: “Old tale in new skin” about sequence comparison. *Information Sciences* 181: 1454–1462. doi:10.1016/j.ins.2010.12.009.
49. Liao B-Y, Chang Y-J, Ho J-M, Hwang M-J (2004) The UniMarker (UM) method for synteny mapping of large genomes. *Bioinformatics* 20(17): 3156–65. doi:10.1093/bioinformatics/bth380.
50. Kirzhner V, Korol A, Bolshoy A, Nevo E (2002) Compositional spectrum—revealing patterns for genomic sequence characterization and comparison. *Physica A: Statistical Mechanics and its Applications* 312: 447–457. doi:10.1016/S0378-4371(02)00843-9.
51. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. *Science* 304: 1321–5. doi:10.1126/science.1098119.
52. Ponting CP, Lunter G (2006) Signatures of adaptive evolution within human non-coding sequence. *Human Molecular Genetics* 15 Spec No 2: R170–5. doi:10.1093/hmg/ddl182.
53. Woolfe A, Elgar G (2008) Organization of conserved elements near key developmental regulators in vertebrate genomes. *Advances in Genetics* 61: 307–38. doi:10.1016/S0065-2660(07)00012-0.
54. Yamada K, Lim J, Dale JM, Chen H, Shinn P, et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302(5646): 842–6. doi:10.1126/science.1088305.
55. Chen L, Zhao H (2005) Negative correlation between compositional symmetries and local recombination rates. *Bioinformatics* 21(21): 3951–8. doi:10.1093/bioinformatics/bti651.
56. Bell SJ, Forsdyke DR (1999) Deviation from Chargaff's second parity rule Correlate with direction of transcription. *Journal of Theoretical Biology* 197: 63–76. doi:10.1006/jtbi.1998.0858.
57. Bell SJ, Forsdyke DR (1999) Accounting units in DNA. *Journal of Theoretical Biology* 197: 51–61. doi:10.1006/jtbi.1999.0857.
58. Deng B (2007) Mismatch repair error implies Chargaff's Second Parity Rule. *Arxiv preprint arXiv:0704.2191*: 1–8.
59. Kirzhner V, Bolshoy A, Volkovich Z, Korol A, Nevo E (2005) Large-scale genome clustering across life based on a linguistic approach. *Bio Systems* 81: 208–22. doi:10.1016/j.biosystems.2005.04.003.
60. Touchon M, Nicolay S, Audit B, Brodie E, Brodie E-B, Aubenton-Carafa Yd, et al. (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proceedings of the National Academy of Sciences of the USA* 102: 9836–41. doi:10.1073/pnas.0500577102.
61. Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, et al. (2003) Informatics for unveiling hidden genome signatures. *Genome Research* 13: 693–702. doi:10.1101/gr.634603.
62. Azad RK, Rao JS, Li W, Ramaswamy R (2002) Simplifying the mosaic description of DNA sequences. *Physical Review E* 66: 031913.
63. Buschiazio E, Gemell NJ (2010) Conservation of human microsatellites across 450 million years of evolution. *Genome Biology and Evolution* 2: 153–65. doi:10.1093/gbe/evq007.
64. Porceddu A, Camiolo S (2011) Spatial analyses of mono, di and trinucleotide trends in plant genes. *PLoS ONE* 6: e22855. doi:10.1371/journal.pone.0022855.
65. Smith AV, Thomas DJ, Munro HM, Abecasis GR (2005) Sequence features in regions of weak and strong linkage disequilibrium. *Genome Research* 15: 1519–34. doi:10.1101/gr.4421405.
66. Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* 241: 3–17.
67. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics* 10: 285–311. doi:10.1146/annurev-genom-082908-150001.
68. Katzman S, Capra JA, Haussler D, Pollard KS (2011) GBE Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hotspots. *Genome*. pp 1–17. doi:10.1093/gbe/.
69. Costantini M, Cammarano R, Bernardi G (2009) The evolution of isochore patterns in vertebrate genomes. *BMC Genomics* 10: 146. doi:10.1186/1471-2164-10-146.
70. Costantini M, Filippo MDi, Auletta F, Bernardi G (2007) Isochore pattern and gene distribution in the chicken genome. *Gene* 400: 9–15. doi:10.1016/j.gene.2007.05.025.
71. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, et al. (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420: 578–82. doi:10.1038/nature01251.
72. Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences — an unexpected feature of mammalian genomes. *Nature Reviews. Genetics* 6: 151–7.
73. McLean C, Bejerano G (2008) Dispensability of mammalian DNA. *Genome Research* 18: 1743–51. doi:10.1101/gr.080184.108.
74. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. (2007) Human genome ultraconserved elements are ultrasected. *Science* 317: 915. doi:10.1126/science.1142430.
75. Carpena P, Bernaola-Galván P, Coronado M, Hackenberg M, Oliver J (2007) Identifying characteristic scales in the human genome. *Physical Review E* 75: 2–5. doi:10.1103/PhysRevE.75.032903.
76. Dunham I, Shimizu N, Roe B, Chissoe S, Hunt A, et al. (1999) The DNA sequence of human chromosome 22. *Nature* 402: 489–495.
77. Grimwood J, Gordon La, Olsen A, Terry A, Schmutz J, et al. (2004) The DNA sequence and biology of human chromosome 19. *Nature* 428: 529–35. doi:10.1038/nature02399.
78. Hillier LW, Graves Ta, Fulton RS, Fulton La, Pepin KH, et al. (2005) Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* 434: 724–31. doi:10.1038/nature03466.
79. Hattori M, Fujiyama A, Taylor T, Watanabe H, Yada T, et al. (2000) The DNA sequence of human chromosome 21. *American Journal of Ophthalmology* 130: 383. doi:10.1016/S0002-9394(00)00695-4.
80. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, et al. (2006) DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature* 440: 1045–9. doi:10.1038/nature04689.
81. Nusbaum C, Zody MC, Borowsky ML, Kamal M, Kodira CD, et al. (2005) DNA sequence and analysis of human chromosome 18. *Nature* 437: 551–5. doi:10.1038/nature03983.
82. Jurka J, Kohany O, Pavlicek A, Kapitonov VV, Jurka MV (2004) Duplication, coclustering, and selection of human Alu retrotransposons. *Proceedings of the National Academy of Sciences of the USA* 101: 1268–72. doi:10.1073/pnas.0308084100.
83. Caron H, Schaik B van, Mee M van der, Baas F, Riggins G, et al. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* 291: 1289–92. doi:10.1126/science.1056794.
84. Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews. Genetics* 2: 292–301. doi:10.1038/35066075.
85. Meaburn KJ (2007) Chromosome territories. *Nature* 445: 379–381.
86. Chubb JR, Bickmore WA (2003) Compartmentalization in the light of nuclear dynamics. *Cell* 112: 403–406.
87. De S, Babu MM (2010) Genomic neighbourhood and the regulation of gene expression. *Current Opinion in Cell Biology* 22: 326–33. doi:10.1016/j.ccb.2010.04.004.
88. Paz A, Kirzhner V, Nevo E, Korol A (2006) Coevolution of DNA-interacting proteins and genome “dialect”. *Molecular Biology and Evolution* 23: 56–64. doi:10.1093/molbev/msj007.
89. Bingham E, Gionis A, Haiminen N, Hiisila H, Mannila H, et al. (2006) Segmentation and dimensionality reduction. In: *Proceedings of the Sixth SIAM International Conference on Data Mining*. Society for Industrial Mathematics, Vol. 124. 372 p.
90. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragozy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289–93. doi:10.1126/science.1181369.