Article

Visualizing the Echo of Evolution: Tandem Repeats and Sequence Structure of Chromosomes of *Saccharomyces cerevisiae* Shown by Doublet Frequency Distance Maps

Johann Michael Köhler [1,*]

[1] Institute for Micro- and Nanotechnologies, Technische Universitat Ilmenau, 98693 Ilmenau, Germany
[2] Institute for Chemistry and Biotechnology, Department of Physical Chemistry and Microreaction Technology, Technische Universitat Ilmenau, 98693 Ilmenau, Germany;
* Correspondence: michael.koehler@tu-ilmenau.de; Tel.: +49-0-3677-69-3629; Fax.: +69-3179

Abstract: The method of doublet frequency distance (DFD-) maps was applied, here, for visualization of DNA sequence structures of yeast chromosomes. The colour scale "rainbow" of the Octave programming tool is well suited for such visualization. The DFD-maps are generated by comparison of the differences in the number of all base doublets in a shifting frame with all other positions of this frame in a regarded DNA sequence section. This procedure can be applied from DNA-sections of between a few hundred bases (bps) up to a complete chromosome with a million or more bps. An orthogonal DFD-map pattern dominates all chromosomes and its parts. It can be interpreted by a large number of successive mutation events during evolution. In contrast, diagonal patterns indicate duplications of sequence sections and multiple tandem repeats. These periodic structures are found in several chromosomes and are more or less regular or noised. The larger tandem repeat in the subtelomeric region of chromosome 12 presents a characteristic example of a nested multiple-repeat structure. Its pattern in the DFD-map illustrates obviously a temporal order of duplication and mutation events leading to a hierarchical sequence architecture.

Keywords: DNA; Chromosomes; Yeast; Tandem Repeats; Evolution; Genetic Noise; Visualization; Doublet Frequency Maps; Sequence structure

1. Introduction

The sequence of nucleotide in DNA carries the essential information for functional proteins and for regulation of gene expression on the one hand, but contains implicit information of the development of genes, that means it carries an echo of the molecular evolution. Therefore, the DNA sequence can be understood as a superposition of traces reflecting the evolutionary change of phenotype and of traces of undirected fluctuations and modification of the genotype during evolution [1].

It is well known that chromosomes with their millions of base pairs could neither be formed spontaneously in their recent size in natural evolution nor developed by point mutations, alone. From a statistical point of view, the mechanisms of variation and selection have to concern a hierarchy of information storage elements [2]. Despite the fact, that DNA of chromosomes appear as a huge single molecule, it is evident that it is organized in a modular structure. Manfred Eigen gave strong arguments that the first gene in evolution should have a size of about 76 bases and he concluded from different facts that the elementary module in the evolution of genes started with a self-folding molecule with a simple periodic sequence and a structure similar to t-RNA [3].

Simple sequence patterns are obviously not only important for early evolution, but relevant up to now. The presence or absence of small sequence motifs is suited for distinguishing large taxonomical units. Single genera or species are marked by significant over or under representation of certain small signature sequences. Such signatures are very useful for studying evolutionary relationships. Beside the frequency of doubletts [4, 5], also other small motifs as tetranucleotides are applied for taxonomical analyses [6, 7]. Beside taxonomical studies, nucleotide-based techniques have also been introduced for detection of human-pathogenic viruses [8]. Besides the total abundance of certain nucleotide pattern, the distribution of patterns in the genome is of interest, too [9].

Tandem repeats represent a special form of sequence pattern. There are small repeats concern small motifs of a few nucleotides, only, on the one site. But, there are also duplicated large sections with hundreds or thousands of nucleotides. The introduction of such large repeats seem to be an important strategy of evolution for enlarging chromosomes between the level of single nucleotides and small oligonucleotides and the level of large chromosome fragments and tandem repeats are sites of fast mutation and are preferable positions for gen modifying in evolutionary adaption on changing environmental conditions, probably, as shown for yeast [10]. In addition, there are important sites for specific DNA-protein interaction, for example in yeast mitosis [11]. Finally, the coupling of tandem repeats with endonucleases can be used as tool for genome engineering [12].

Here, the example of yeast was selected for demonstrating a simple method for fast evaluation of the structure of chromosomes of a eukaryotic microorganism by the distribution of oligonucleotide motives. The development and testing of alternative methods for data visualization is an urgent need, because the user of sequence data and recent sophisticated numerical procedure need tools for obtaining fast impressions of sequence features and for comparison of sequences [13]. The method of doublet frequency maps was developed and firstly used for visualization of the genome structure of Bacillus subtilis [14]. It is applied, here, for visualizing the structure of sequences in chromosomes of *Saccharomyces cerevisiae*, among them sequence repeats and other characteristic features.

2. Method

2.1. Data

The sequence data of the 16 chromosomes have been downloaded (2021-07-15) from the SGD Saccharomyces Genome Database (http://sgd-archive.yeastgenome.org/sequence/S288C_reference/chromosomes/fasta/; 2019-10-25). This database was also used for comparison of visualized sequence pattern with the functional gene structure (https://browse.yeastgenome.org/).

2.2. Doublet Frequency Distance Maps (DFD-Maps)

16 different doublets can be formed by DNA sequences consisting of the 4 different DNA nucleotides. Each chromosome or sections of it can be described, simply by its nucleotide sequence or by its sequence of overlapping doublets. The number of overlapping doublets Z is given by the number of nucleotides N minus 1:

$$Z = N-1 \qquad (1)$$

For a given sequence (or sequence section), a gliding frame of a smaller number of nucleotides can be defined, which is moved from the start to the end of the regarded sequence. Now, in each frame the total number of all doublets of one of the 16 doublet types is calculated. This results in to a doublet frequency table for each position of the gliding frame. In the sequence “CGTTAATGGCTAG”, for example, following 11 doublets are present:

Doublets: CG GT TT TA AA AT TG GG GT TA AG

This results in to following doublet frequency table:

Doublet:	AA	AT	AC	AG	TA	TT	TC	TG	CA	CT	CC	CG	GA	GT	GC	GG
Number:	1	1	0	1	2	1	0	1	0	0	0	1	0	2	0	1

Exchanges of parts in this sequence would not change the abundance of nucleotides, but would change the doublet frequency table. Therefore, this table represents reduced information of the whole sequence.

In the next step, the differences of the numbers of each doublet type are calculated for all positions of the gliding frame and all other positions of it. The sum of these differences for each position pair supplies the doublet frequency distance table (DFD-table). Its two-dimensional graph represents the doublet frequency distance map (DFD-map), a square map with a diagonal symmetry. Each axis corresponds to the linear extension of the regarded chromosome of a section of it.

For the visualization a colour codes of the octave software have been used. In the "rainbow" colour map, red indicates low and blue a high distance. Yellow and green indicate intermediate doublet frequency distances between the concerned pairs of sequence positions (see, please, supplementary material for more details).

3. Results and Discussions

3.1. DFD maps of complete chromosomes

In general, the nucleotide content as well as the doublet content of all chromosomes is similar. The AT/GC ratio is about 1.6 for all regarded chromosomes. Therefore, the A- and T- containing doublets show higher abundancies than the exclusively G- and C- containing doublets. AA and TT represent about 10.7 % of all doublets, CC and GG about 3.4 - 4.1 %. The AT doublet has a little higher abundance (about 8.6 – 9.1 %) than the TA doublet (about 7.2 – 7.4 %)). GC doublets (about 3.6 to 3.9 %) are higher abundant than CG doublets (2.8 – 3.1 %).

This corresponds to the fact, that AC doublets (about 5.2 to 5.4 %) and AG doublets (about 5.7 – 5.9 %) are less abundant then TC (about 6.1 – 6.3 %) and TG (about 6.3 – 6.6 %) doublets. CA (about 6.3 – 6.8 %) and GA doublets (about 6.0 – 6.3 %) are higher abundant than CT (about 5.7 – 5.9 %) and GT doublets (about 5.2 – 5.4 %). These asymmetries in abundances reflect a preferential direction in the sequences (preferential reading arrow).

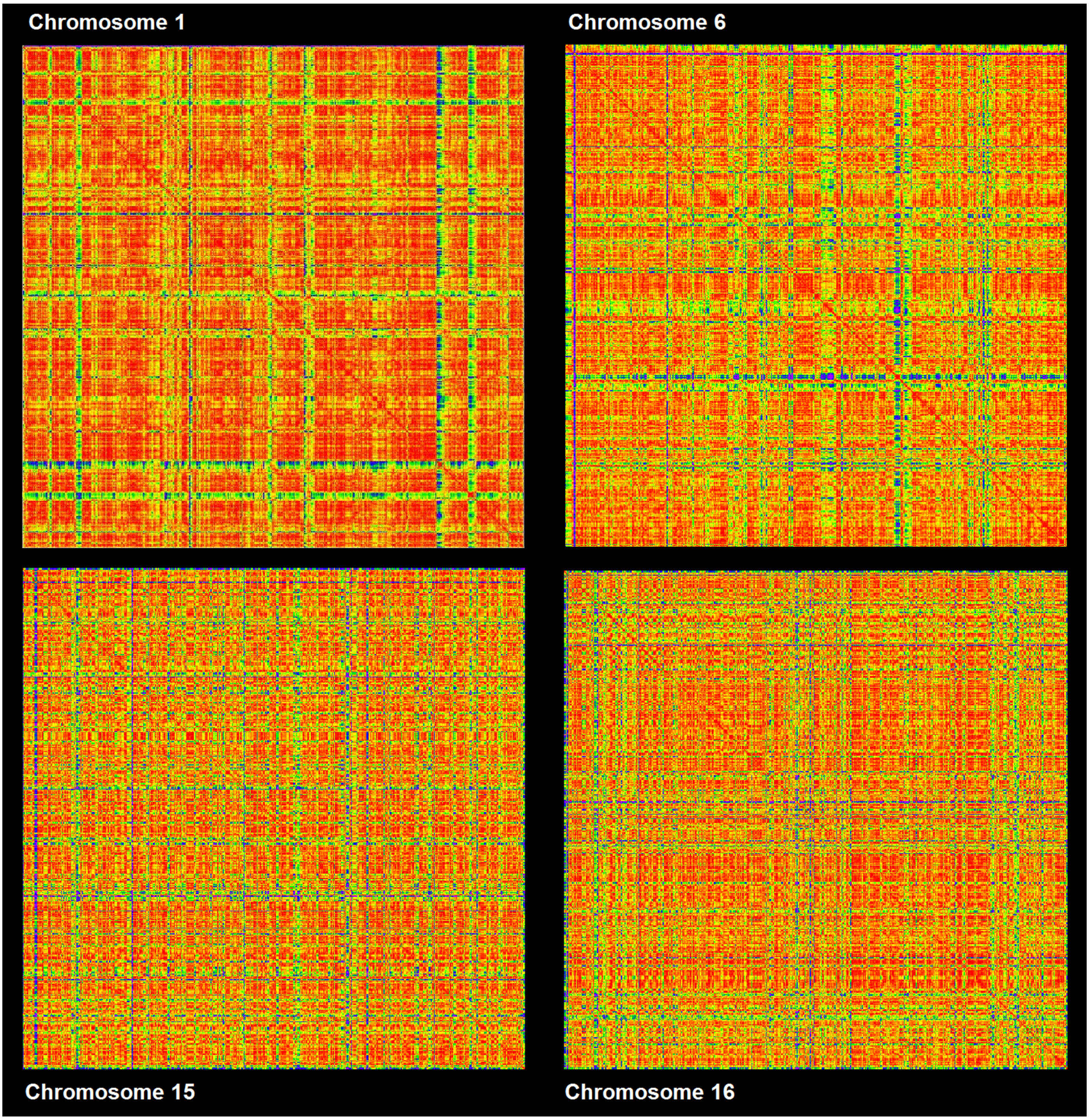


Fig. 1 DFD-maps of the whole chromosomes 1, 8, 15 and 16 of the yeast genome (please, consider that the chromosomes have different sozes)

The dominance of orthogonal structures in the DFD maps results from the mapping principle, which shows the doublet distance of each position against each other position of a regarded sequence. This general feature is observed in all sequences and is also found in DFD maps for whole chromosomes. The dominance of red colour in the "rainbow-scale" maps indicates a comparative high similarity between the most parts of the sequence. But sequence sections of high general similarity are separated by smaller sections with significant different doublet frequencies.

These strange sections are visualized by yellow, blue and green (in the rainbow scale) and are found in the smaller chromsomes as No 1 and 6 as well as in larger chromosomes as No 15 and 16, for example (Fig. 1). The local differences in doublet frequencies are much better visible and distinguishable by the yellow/green/blue strips in the two-dimensional mappings than in linear graphs (for comparison: see supplementary material- Fig. S1).

The whole orthogonal pattern has the character of a formerly more homogeneous red area in which the yellow/green strips are inserted. These strips present obviously a hierarchy of more and less strange regions and larger and smaller strange regions. This general picture can be interpreted by the development of the sequences of the chromosomes. Following this interpretation, the DFD maps reflect numerous insertion events of smaller and mediate sections with significant different doublet distributions in comparison with the formerly average doublet distributions inside the evolving chromosome.

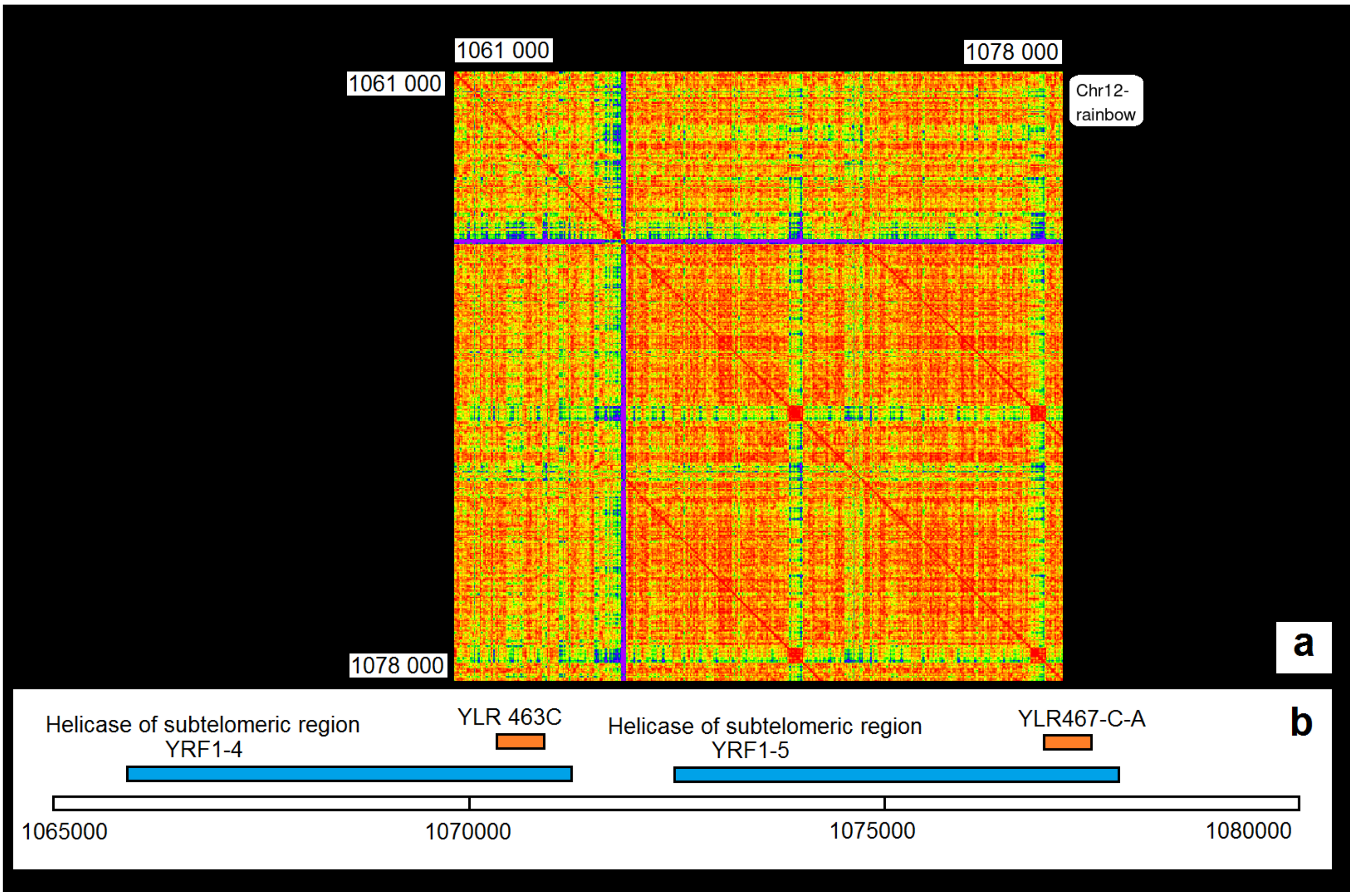


Fig. 2 Subtelomeric region of chromosome 12: a) DFD-map, b) sequences structure (adapted from SGD Saccharomyces Genome Database (http://sgd-archive.yeastgenome.org/sequence/S288C_reference/chromosomes/fasta/)

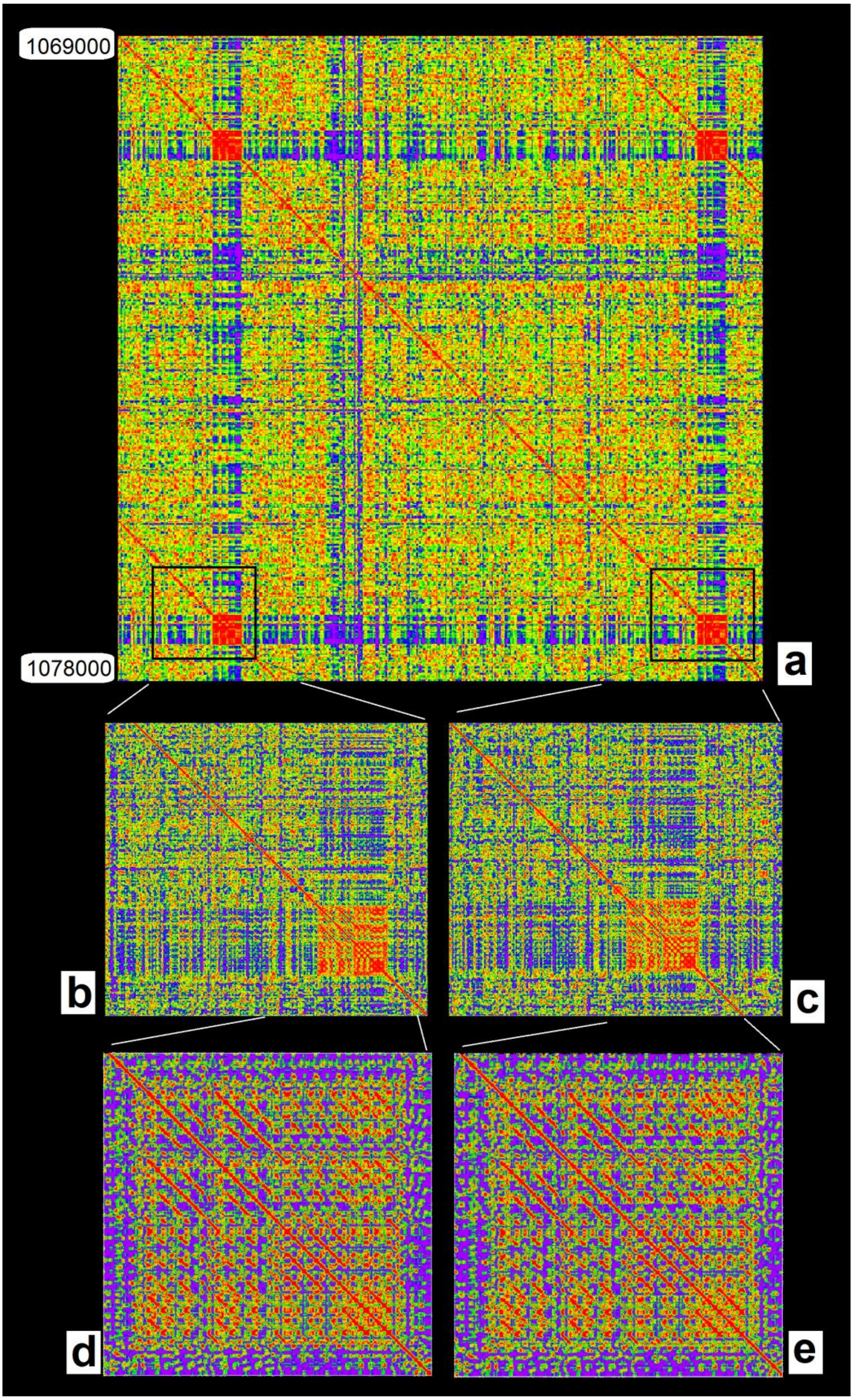


Fig. 3 Subtelomeric region of yeast chromosome 12: a) tandem repeat section (cutout of 9 kb); b) two regions of high internal similarity and large doublet frequency difference distances to the environment, c) DFD-map showing a noised periodic pattern of high similarity in the both strange regions (size of cutout: about 0.5 bp)

3.2. Segment duplication

Segment duplications appear as diagonal lines (side diagonals) parallel to the main diagonal in the DFD-maps. An important example is found in the subtelomeric region around position 1,070,000 on chromosome 12 (Fig. 2a). The whole subtelomeric region is separated by small strange section (high DFD to other parts of the chromosome) around position 1,065 760. This separation is clearly expressed in the DFD map by the blues strip.

The whole duplicate region (between about 1,065700 and 1,078,000) appears as a slightly enhanced ratio of red and yellow/green areas. This indicates a comparatively high similarity inside the duplicate region. This similarity is interrupted by two small sections with stronger deviation of their local DFDs to the DFDs of the other parts of duplicate region (yellow/green strips around position 1,070,350 and 1,077,150). A very high doublet frequency similarity appears inside these both small strange sections, indicated by the small red squares on the main diagobal. The appearance of the same red squares on the side diagonals corresponds to the duplication of the larger sequence section.

The duplicated sequence section has a length of about 5 kb. Both corresponding sequence sections are separated by a section with a length of about 0.9 kb which is not distinguished by a significant difference in the DFDs from the duplicated region.

The data from the SGD archive teach that the duplicate region is formed by the genes YRF1-4 and YLR1-5 and YLR463C and YLR467-C-A (Fig. 2b). The first mentioned genes are related to helicase encoded by the subtelomeric Y-structure, the last mentioned relate to a smaller "dubious" open reading frame (https://browse.yeastgenome.org/).

3.3. Multiple repeats

The duplicate structure in the subtelomeric region of chromosome 12 allows also a good insight into the reason for strange sequence sections. Therefore, it is helpful to zoom into the both small strange sections around 1,070,350 and 1,077,150 (small red squares). It is clearly visible that each of the both small sections is strange to its environment, but possesses a comparatively high similarity in internal doublet frequencies. Therefore, the both regions appear red in y yellow/green environment in the DFD maps (Fig. 3a). At higher resolution, the internal structure of the both region becomes visible. It reflects the analogous sequence of both sections (Fig. 3 b, c) At still higher magnification, a nearly periodic structure with the character of a noised translational symmetry becomes visible (Fig. 3d, e). It represents a disturbed multiple repeat over a length of about 450 bp. It consists of 11 small sequence sections of about 40 bp.

Such multiple repeats exist in other chromosomes of the yeast genome, too. In general, they are marked by high similarities in internal nucleotide and doublet frequencies and large differences in the doublet frequencies to most other regions of the concerned chromosome. Typically, a crossed strips arise at the position of the multiple repeat region in the DFD map A related map is given for region around position 1300000 of chromosome 4, for example (Fig. 4a). The zoom shows a red square, what mean that the doublet pattern of the multiple repeats possess a high internal similarity. The blue strips prove the high dissimilarity between the multiple-repeat region and the environment (Fig. 4b). The multiplicity of elementary gene structure becomes visible in a DFD map at further enhanced resolution (Fig. 4c). The repeat region consists of 11 elements of a length of 85 bp.

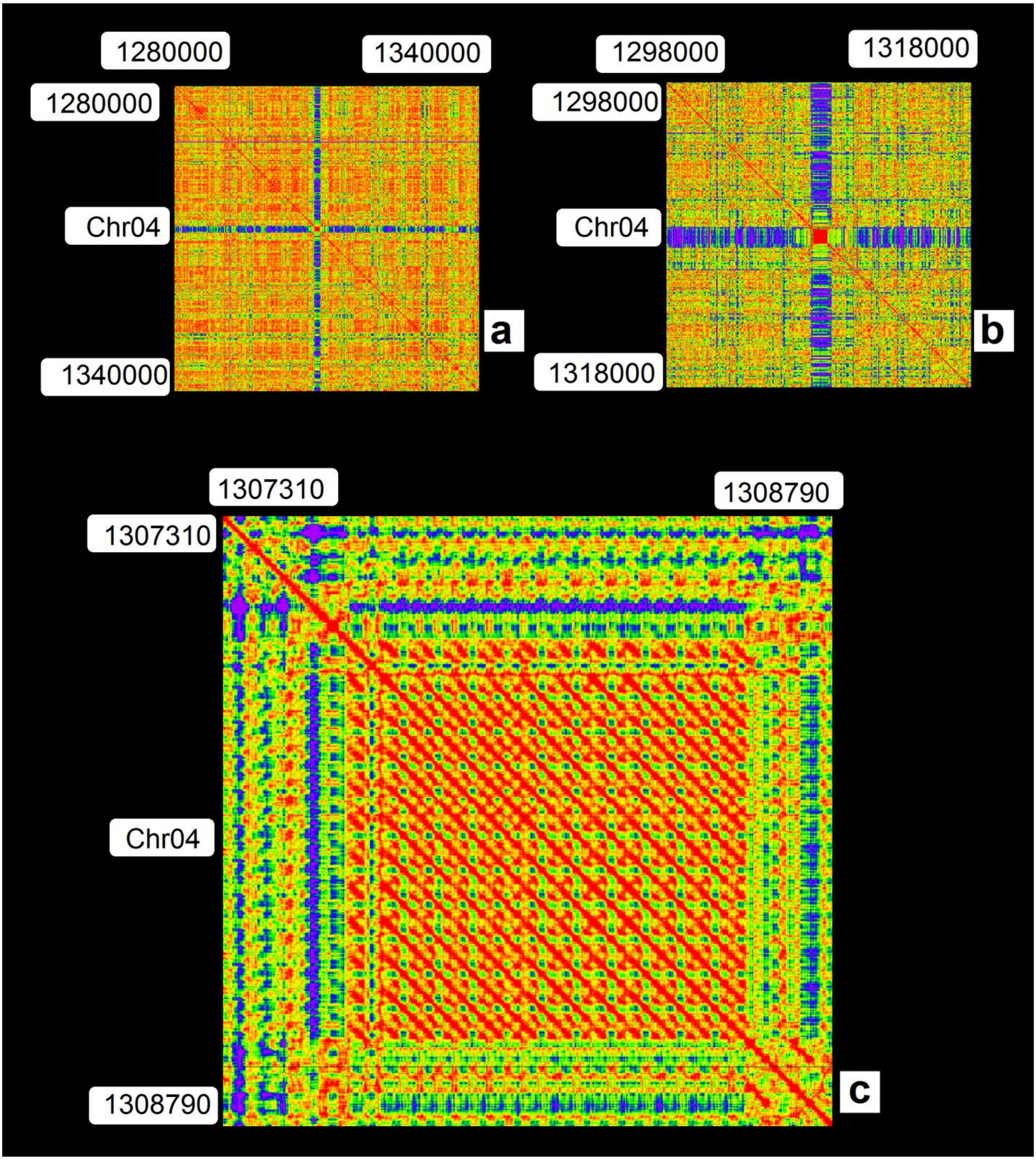


Fig. 4 Multiple-repeat region of chromsome 4: a) presentation of the multiple-repeat region as small strange line in the DFD-map of a section of a size of section of 60 kb, b) banded strip pattern reflecting the periodic sequence structure inside the multiple-repeat region in for a map of 20 kb, c) DFD-map zooming into the multi-repreat region showing the periodic DNA sequence pattern by the parallel arranged side diagonals

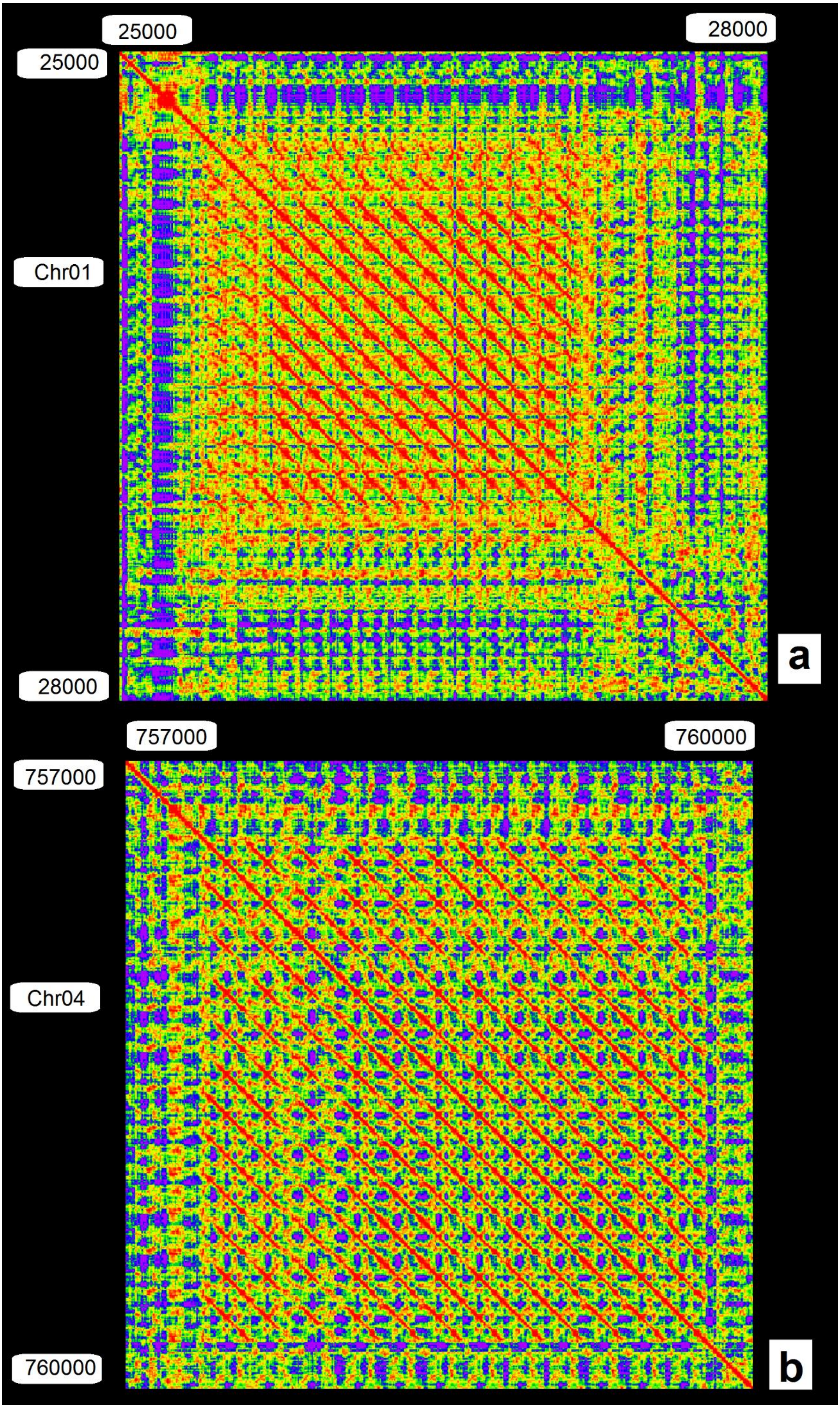


Fig. 5 Multiple-repeat regions of chromosome 1 with high regularity: a) sequence section of 3 kb around positions 26,500, b) sequence section of 3 kb around positions 75,850

Two other repeat structures of chromosomes 1 and 4 are shown in Fig 5. The DFD maps make clear that the symmetry between the single gene sections is not perfect, but marked by small deviations. These deviations cause a modified spot structure in the DFDs. A shift in the side diagonals by few bp indicates an insertion event, obviously (Fig. 5b). It can be speculated, here, that the strange sequence section of the multiple repeat regions corresponds to a late insertion or multiplication in the gene evolution. But, the small offset in the side diagonals is probably caused by a still later insertion event.

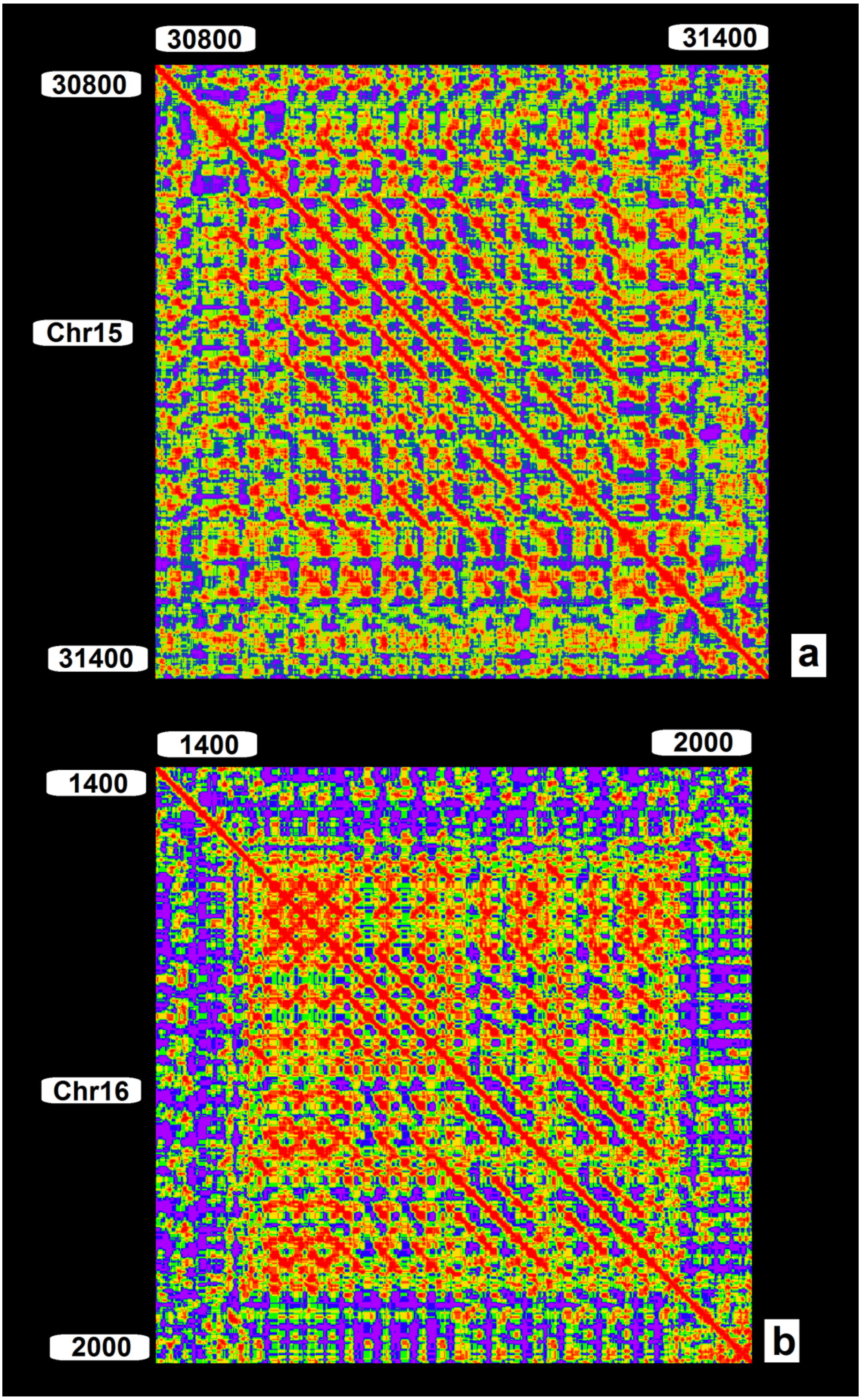
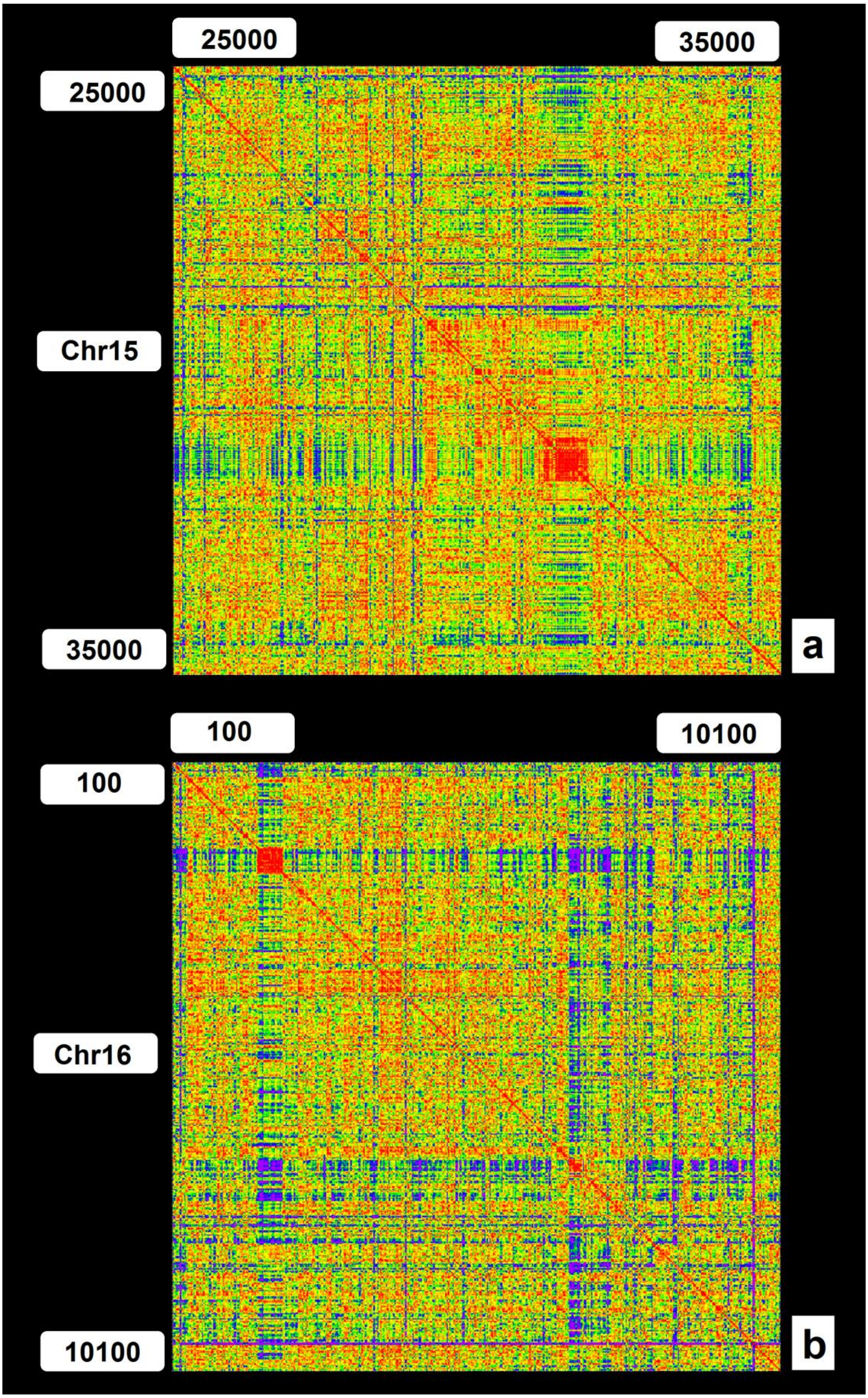


Fig. 6 Strongerly noised multiple-repeat regions of a length of 0.6 kb: a) chromosome 15, region around position 31,500; b) chromosome 16, region around position 1,700

3.4. Strongly noised repeats

In some cases, tandem repeats and multiple repeats are less sharp recognizable. The gene pattern is disrupted and noised by many disturbances. Two typical examples of strongly noised multiple repeats are shown in Fig. 6. It concerns regions of about 0.4 kb in both cases. In chromosome 15, a period of about



35 bp appears (Fig. 6a). In chromosome 16 periods of about 25-32 bp are visible in the DFD map. The regarded region in chromosome 16 is stronger noised than the example of chromosome 15. Despite the considerable loss of regularity, both regions are well distinguishable from their environment. The strange regions appear as red square spots on the main diagonal (Fig. 7). Even very strongly noised multiple repeats are visible by their strange colour in the DFD maps as, for example a relict sequence section with strongly noised periodicity in chromosome 13 (Fig. 8).

Fig. 7 Embedding DNA sections (10 kb) of noised multiple-repeat regions in two chromosomes (Fig. 6): a) chromosome 15, aroud position 20,000 b) chromosome 16, subtelomeric region

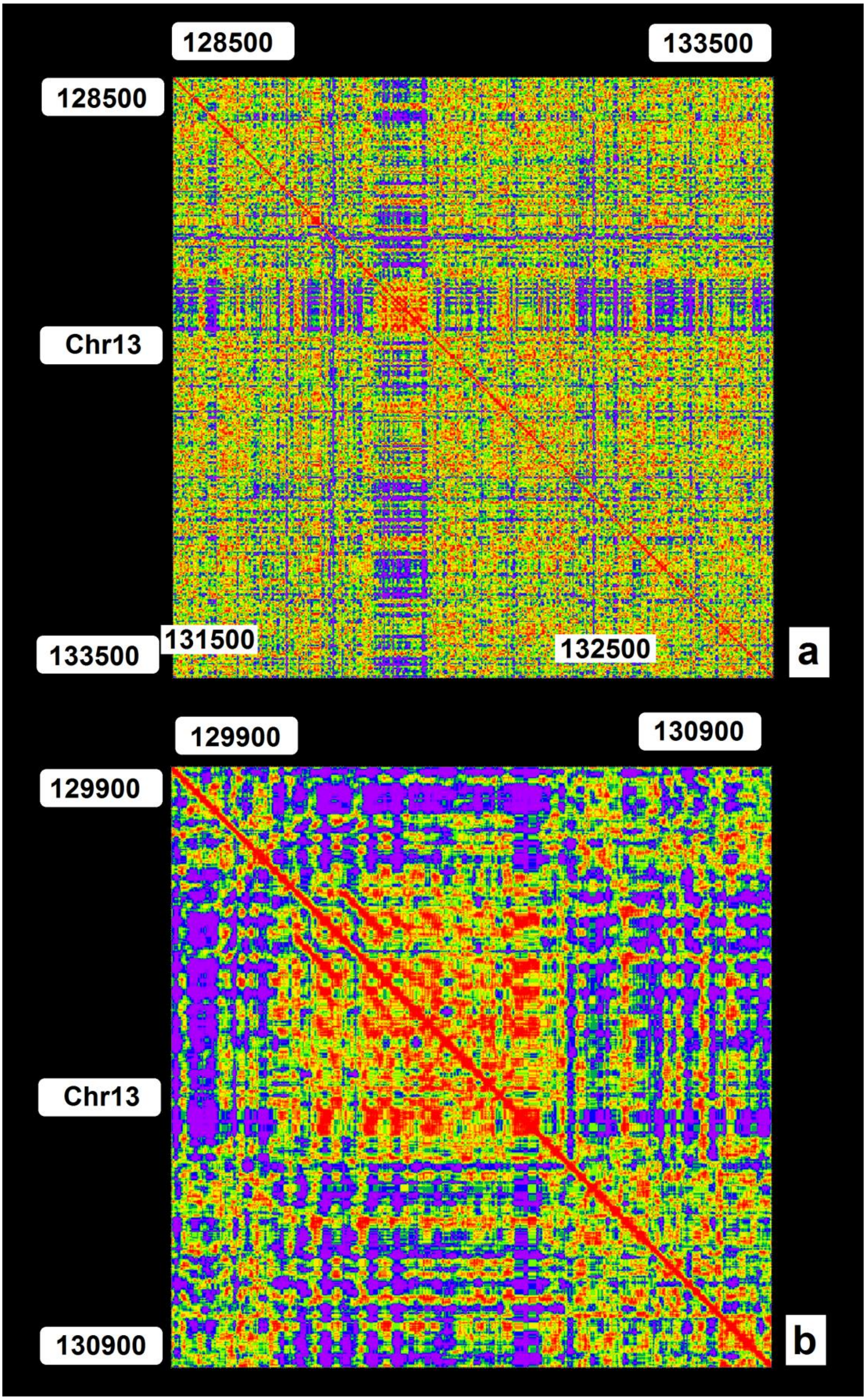


Fig. 8 Strongly noised repeat region of chromosome 13: a) embedding sequence section of 5 kb, b) the region (size about: 0.5 kb) around position 132,000

The four red squares in Fig. 3a are also indicating two regions of noised repeats. These noised repeats are parts of the large duplicated sequence section (Fig. 2). The DFD picture allows to identifying a simple relative chronology of events corresponding to following interpretation: At first, a multiplication of a short sequence section containing (here, about 40 bp) occurs (schematically in Fig. 9a). In a second step, the resulting multiple-repeat pattern was noised by additional mutations (Fig. 9b). Finally, the multiple-repeat pattern was duplicated together with a larger embedding sequence region (Fig. 9c).

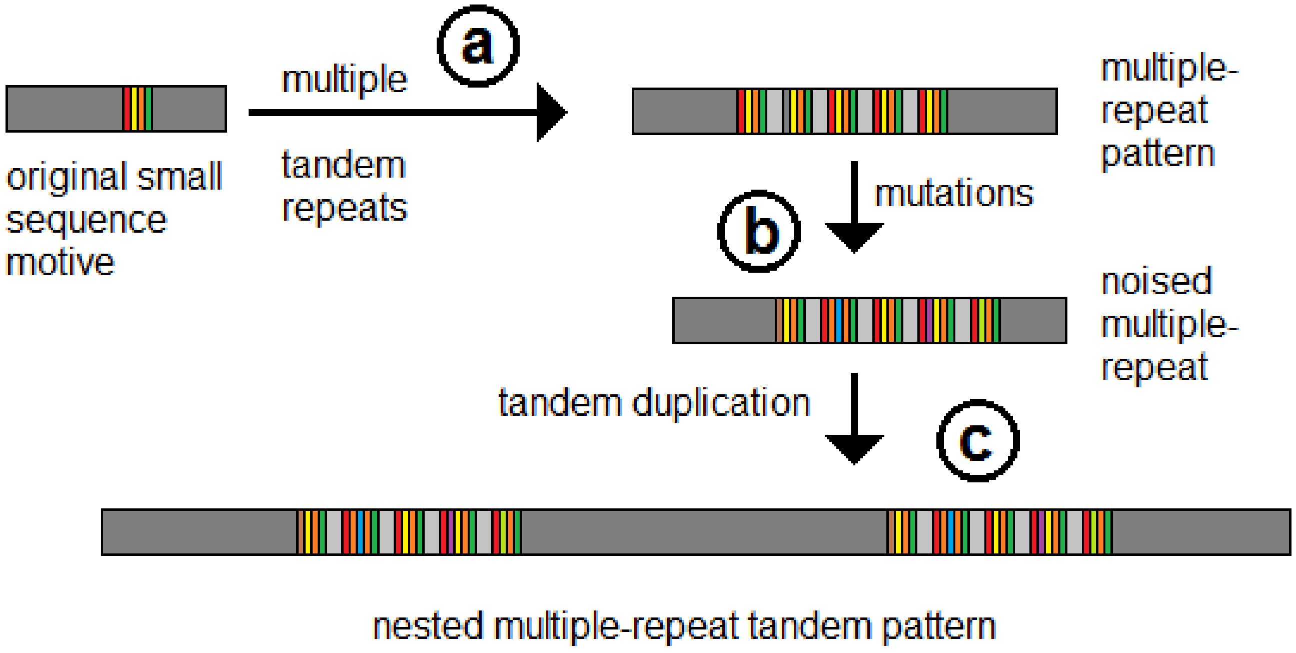


Fig. 9 Formation of a nested repeat region as reflected in DFD-maps (schematically)

In result, a nested tandem repeat region was formed as presented in the example of Fig. 3. The high similarity between the both small multiple-repeat sections (Fig. 3 d and e) speaks for a comparatively young age of the outside larger duplication, on the one hand. This similarity and the stronger noise in parts of the multiple-repeat regions indicates, obviously, a higher age of the multiple tandem duplication of the small original sequence motive.

4. Conclusions

Sequence patterns of chromosomes can be visualized easily by doublet frequence distance maps (DFD maps). The example of yeast shows the general dominance of orthogonal pattern. They can be interpreted by a lot of insertion or other mutation events. Such changes in an original shorter DNA sequence are much better visible in a two-dimensional than in a one-dimensional plot. The superposition of large numbers of strange-colour strips in a more homogeneous matrix can be interpreted by a succession of mutation events during the evolution of the chromosomal DNA sequence. The dominance of this orthogonal structure is evident in all chromosomes of *Saccharomyces cerevisiae*. It is suggested here that the strip pattern of DFD-maps, in general, reflects the evolution of hierarchical structures of chromosomes by a succession of insertions, tandem repeats and smaller mutation events.

Particular significant strips in the DFD-maps corresponding to strange sections of DNA sequences are caused by tandem duplications and multiple tandem repeats. Such patterns are found in several chromosomes of the yeast genome. Some of the multiple repeats are rather regular, other are more or less noised. An example of a nested tandem duplication pattern exists in the subtelomeric region of chromosome 12. The hierarchical structure of this sequence pattern is detailed pictured by DFD-maps with different zoom factors.

Funding: This research received no external funding.

Conflicts of Interest: The author declare no conflict of interest.

References

1. [1] Sharp, P.M.; Averof, M.; Llooyd, A.T., Matassi, G.; Peden, J.F. DNA-sequence evolution – the sound of silence. *Philos. Transact. Royal Soc. B – Biol. Sci.* **1995,** *349,* 241-247.
2. [2] Köhler, J.M.; Saluz, H.P. Molecular building principles in nature. Microsystem technology: a powerful tool for biomolecular studies, *Biomethods* **1999,** *10,* 1-15.
3. [3] Eigen, M. Das Urgen. *Nova Acta Leopoldina* **1976**, *52/243*, 5-40.
4. [4] Campbell, A.; Mrazek, J.; Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *PNAS* **1999,** *96*, 9184-9789.
5. [5] VanPassel, M.W.J.; Kuramae, E.E.; Luyf, A.C.M.; Bart, A.; Boekhout, T. The reach of the genome signature in prokaryotes. *BMC Evol. Biol.* **2006,** *6,* 84.
6. [6] Terzian, C.; Laprevotte, I.; Brouillet, S.; Henaut, A. Genomic signatures: tracing the origin of retroelements at the nucleotide level. *Genetica* **1997,** *100,* 271-279.
7. [7] Alsop, E.B. Resolving prokaryotic taxonomy without rRNA: longer oligonucleotide word lengths improve genome and metagenome taxonomic classification. *PLOS ONE* **2013,** *8*, e67337.
8. [8] Cha, T.A.; Kolberg, J.; Irvine, B. et al. Use of signature nucleotide-sequence of hepatitis-C virus for detection of viral-RNA in human serum and plasma. *J. Clinic. Microbiol.* **1991,** *29*, 2528-2534.
9. [9] Rocha, E.; Viari, A.; Danchin, A. Oligonucleotide bias in Bacillus subtilis: general trends and taxonomical comparisons. *Nucleic Acid Res. 1998,* **26,** 2971-2980.
10. [10] Wilkins, M. Biological roles of protein-coding tandem repeats in the yeast Candida albicans. J. Funghi 2018, 4, 78.
11. [11] Fuch, J,; Lorenz, A.; Loidl, J. Chromosome associations in budding yeast caused by integrated tandem repeated transgenes. *J. Cell. Sci.* **2002,** *115,* 1213-1220.
12. [12] Noskov, V.N.; Segall-Shapiro, T.H.; Chuang, R.Y. Tandem repeat coupled with endonuclease cleavage (TRCE): a seamless modification tool for genome engineering in yeast. *Nucl. Acid Res.* **2010,** *38,* 2570-2576. Batley
13. [13] Batley, J; Edwards, D. Genome sequence data: management, storage, and visualization*. Biotechniques* **2009,** *46,* 333.
14. [14] Koehler, J.M. Symmetries in the genome structure of Bacillus subtilis shown by doublet frequency maps. *Proc. SPIE* **2002,** *4623*, 59-66.