Global Repeat Map algorithm as genomic technology for Higher Order Repeat identification (Case study Human Y Chromosome)

Ines Vlahović a,b, Matko Glunčić a, Marija Rosandić c,d, Vladimir Paar a,c

a Department of Physics Faculty of Science, University of Zagreb
b University College Algebra, Zagreb
c Croatian Academy of Science and Arts
d School of Medicine, University of Zagreb
Introduction – types of repetitions

Introduction - tandem repeats

- tandem repeats – repeating pattern of nucleotide bases in DNA sequence

- higher order repeats – HOR

Glunčić M., Paar V. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research, 2012
What are the roles of tandem repeats?

- gene regulation
- changes in chromatin structure
- protein binding sites
- development of the immune system of cells
- repeat analysis in closely related species
- diseases caused by copy number polymorphism
Computational methods for tandem repeats detection

Glunčić M., Paar V. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research, 2012
Global Repeat Map method

- directly maps the DNA symbolic sequence into the frequency domain- "GLOBAL MAP"
- uses a complete \textit{k-word} ensemble (global - local)
- parameter - free
- identifies repetitions of \textit{all lengths}
- \textbf{robust to copy deviations} from the perfect sample
- identifies higher order repeats (HOR)
- consensus lengths and sequences are \textit{simply determined} from results obtained with GRM
- "good" in combination with \textbf{BLAST}
Global Repeat Map method - GRM

- $S_K(j) = \alpha_1(K,j) \alpha_2(K,j) \alpha_3(K,j) \ldots, j = 1,2, \ldots, 4^K$

- $\{X_K(j)\} = [X_K(j)]_1, [X_K(j)]_2, \ldots, [X_K(j)]_n, [X_K(j)]_{n+1}$

- $[d_K(j)]_n = [X_K(j)]_{n+1} - [X_K(j)]_n \Rightarrow \{d_K(j)\} = [d_K(j)]_1, [d_K(j)]_2, \ldots$

- $\{f_K(j)\} = [f_K(j)]^1, [f_K(j)]^2, \ldots, [f_K(j)]^V$
  $\Rightarrow \{f_{K(E)}\} = \sum_{j=1}^{N}[f_K(j)]^1, \sum_{j=1}^{N}[f_K(j)]^2, \ldots, \sum_{j=1}^{N}[f_K(j)]^V, N = 4^K$

Glunčić M., Paar V. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research, 2012
Glunčić M., Paar V. Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Research, 2012
Examples - GRM diagrams for human and Neanderthal chromosome 1
GRM steps *T. castenaum* example

<table>
<thead>
<tr>
<th>Monomer</th>
<th>length of consensus (bp)</th>
<th>range of copy length</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1</td>
<td>331</td>
<td>328-332</td>
</tr>
<tr>
<td>m2</td>
<td>362</td>
<td>362-363</td>
</tr>
<tr>
<td>m3</td>
<td>369</td>
<td>369</td>
</tr>
<tr>
<td>m4</td>
<td>361</td>
<td>361-362</td>
</tr>
<tr>
<td>m5</td>
<td>369</td>
<td>369</td>
</tr>
</tbody>
</table>

Consensus length
Types of HOR structures

Intragenic Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees

Vladimir Paar, Marko Glunčić, Marija Rosandić, Ivan Basar, and Ines Vlahović

Abstract

Much attention has been devoted to identifying genomic patterns underlying the evolution of the human brain and its emergent advanced cognitive capabilities, which lie at the heart of differences distinguishing humans from chimpanzees, our closest living relatives. Here, we identify two particular intragenic repeat structures of noncoding human DNA, spanning as much as a hundred kilobases, that are present in human genome but are absent from the chimpanzee genome and other nonhuman primates. Using our novel computational method Global Repeat Map, we examine tandem repeat structure in human and chimpanzee chromosome 1. In human chromosome 1, we find three higher order repeats (HORs), two of them novel, not reported previously, whereas in chimpanzee chromosome 1, we find only one HOR, a 2mer aliphod HOR instead of human aliphod 1mer HOR. In human chromosome 1, we identify an HOR based on 20-bp primary repeat unit, with secondary, tertiary, and quartic repeat units, fully embedded in human homerin gene, related to regenerative and psychiatric skin. Such an HOR is not found in chimpanzee chromosome 1. We find a remarkable human 3mer HOR organization based on the ~164-bp primary repeat unit, fully embedded within the neuroblastoma breakpoint family genes, which is related to the function of the human brain. Such HORs are not present in chimpanzees. In general, we find that human–chimpanzee differences are much larger for tandem repeats, in particular for HORs, than for gene sequences. This may be of great significance in light of recent studies that are beginning to reveal the large-scale regulatory architecture of the human genome, in particular the role of noncoding sequences. We hypothesize about the possible importance of human accelerated HOR patterns as components in the gene expression multilayered regulatory network.

Keywords: human brain evolution, chromosome 1, higher order repeats, NBPF genes, human homerin gene, global repeat map.

Schematic illustrating three NBPF 3mer HOR copies based on the ~1.6-bp monomers in human chromosome 1. (Results from 2011 - Build 36.3 assembly)
Intragenic Higher Order Repeats in Neuroblastoma Breakpoint Family Genes Distinguish Humans from Chimpanzees

Vladmir Paar,†; Držko Glučić,†; Marija Rosandić,†; Ivan Basar,†; and Ines Vlahović†
†Faculty of Science, University of Zagreb, Zagreb, Croatia
‡Department of Internal Medicine, University Hospital Rogač, University of Zagreb, Zagreb, Croatia
†These authors contributed equally to this work.
‡Corresponding author: E-mail: paar@hzuc.hr.
Associate editor: James Mckinney

Abstract

Much attention has been devoted to identifying genomic patterns underlying the evolution of the human brain and its emergent advanced cognitive capabilities, which lie at the heart of differences distinguishing humans from chimpanzees, our closest living relatives. Here, we identify two particular intragenic repeat structures of noncoding human DNA, spanning as much as a hundred kilobases, that are present in human genome but are absent from the chimpanzee genome and other nonhuman primates. Using our novel computational method Global Repeat Map, we examine tandem repeat structure in human and chimpanzee chromosome 1. In human chromosome 1, we find three higher order repeats (HORs), two of them novel, not reported previously, whereas in chimpanzee chromosome 1, we find only one HOR, a 2mer alphoid HOR instead of human alphoid 1mer HOR. In human chromosome 1, we identify an HOR based on 39-bp primary repeat unit, with secondary, tertiary, and quartic repeat units, fully embedded in human homerin gene, related to regenerating and psoriatic skin. Such an HOR is not found in chimpanzee chromosome 1. We find a remarkable human 1mer HOR organization based on the ~16-kb primary repeat unit, fully embedded within the neuroblastoma breakpoint family genes, which is related to the function of the human brain. Such HORs are not present in chimpanzees. In general, we find that human–chimpanzee differences are much larger for tandem repeats, in particular for HORs, than for gene sequences. This may be of great significance in light of recent studies that are beginning to reveal the large-scale regulatory architecture of the human genome, in particular the role of noncoding sequences. We hypothesize about the possible importance of human accelerated HOR patterns as components in the gene expression multilayered regulatory network.

Key words: human brain evolution, chromosome 1, higher order repeats, NBPF genes, human homerin gene, global repeat map.

Schematic illustrating hierarchical structure of 1,410-bp quartic HOR.

Goal of the project:

- Creation of the repetition database for human, Neanderthal and chimpanzee genomes (http://genom.hazu.hr)
- Over 2000 records
- ALPHAsub algorithm -extension
RESULTS

Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes

M. Glunčić, I. Vlahović, V. Paar. Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes. Scientific Reports volume 9, Article number: 12629 (2019)
Results chromosome Y

GRM diagram and ideogram for human chromosome Y (Build hg38).

What are disadvantages of GRM algorithm?

• main: it depends only on DNA sequences, so variation in schemas are **due to different assemblies of genomes** because of tandem repeats which are very hard to assemble

• solution → **new sequencing technologies** able to sequence complex region of genomes

Future work

DNA analysis of some diseases:

- autism genomes: [https://research.mss.ng/](https://research.mss.ng/)

population genomics/genetics:

- structural variants in tandem and HOR repeats

differences in repeats between close related species (in animals and plants)
Group members

doc. dr.sc. Matko Glunčić
prof. dr.sc. Marija Rosandić
dr. sc. Ines Vlahović
akademik, prof. dr.sc. Vladimir Paar