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
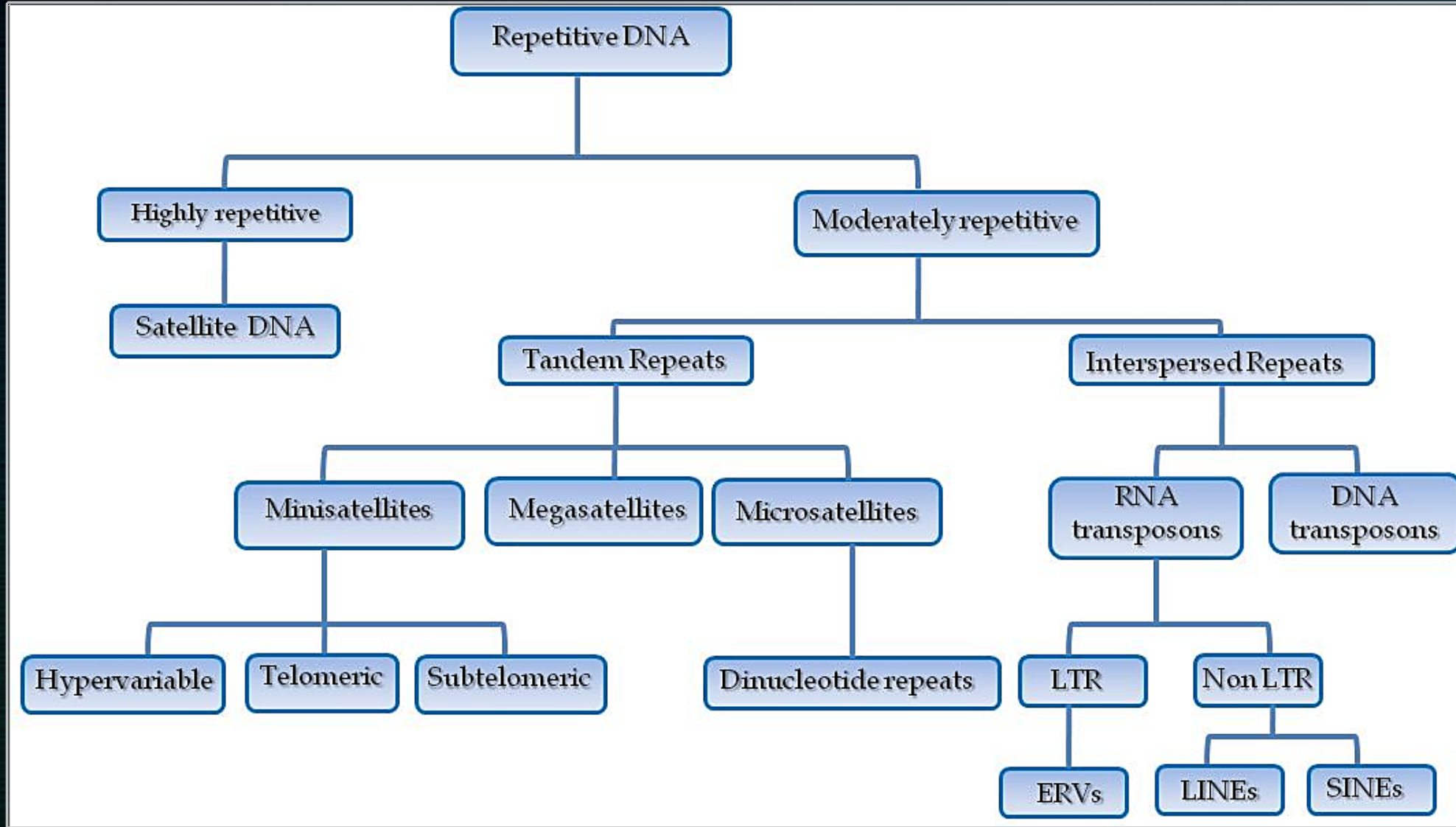
b University College Algebra, Zagreb

c Croatian Academy of Science and Arts

d School of Medicine, University of Zagreb
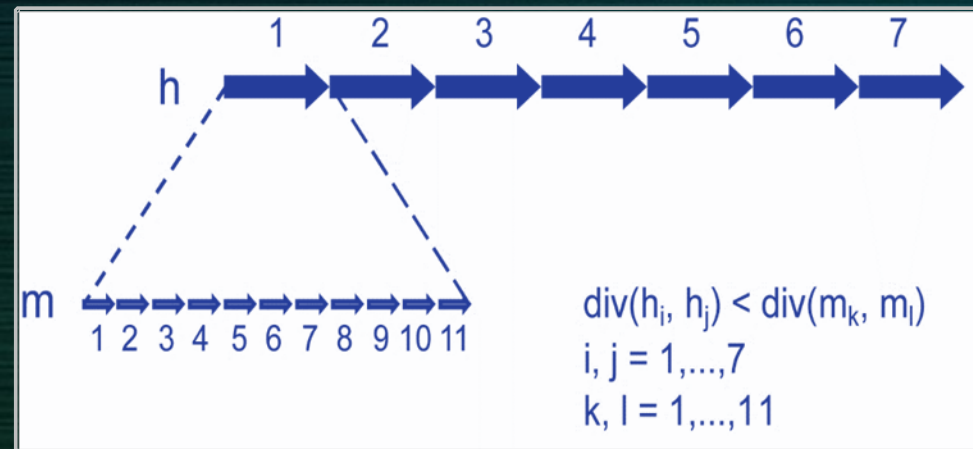
Introduction – types of repetitions

Pathak, D., Ali, S. (2012). Repetitive DNA: A Tool to Explore Animal Genomes/Transcriptomes, Functional Genomics, Dr. Germana Meroni (Ed.), ISBN: 978-953-51-0727-9, InTech, DOI: 10.5772/48259.

Introduction - tandem repeats

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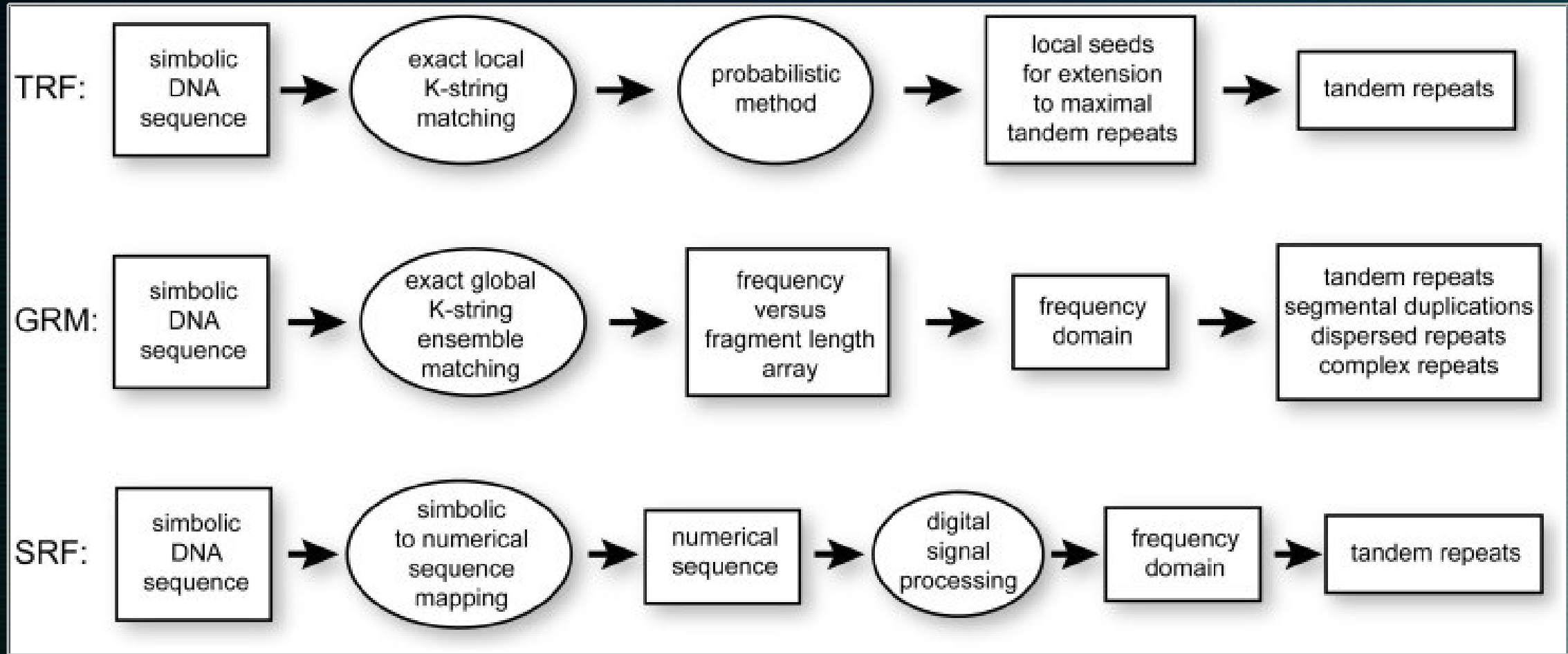

- higher order repeats – HOR

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
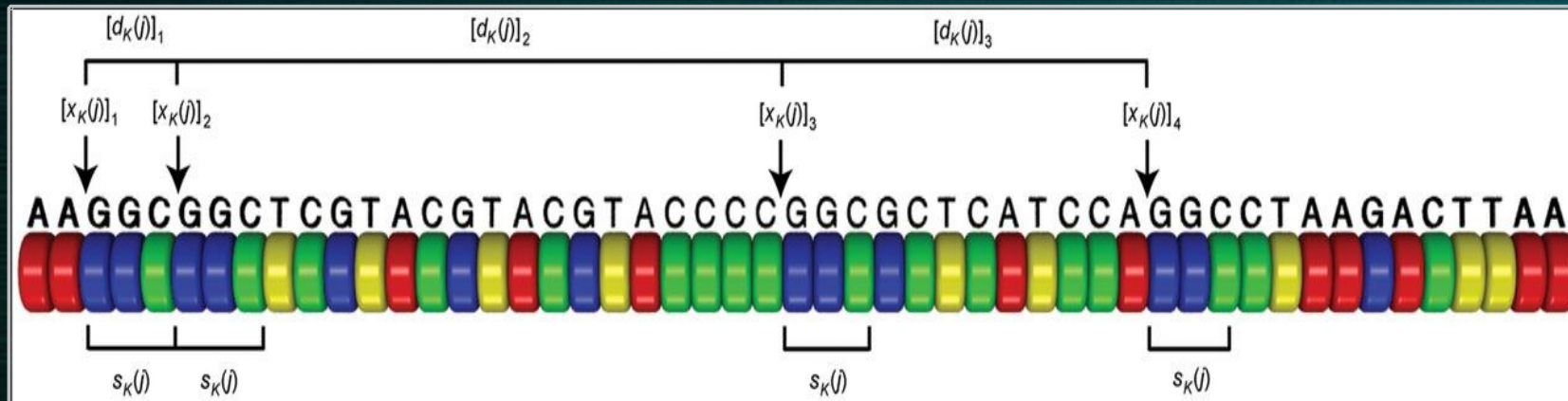
Computational methods for tandem repeats detection

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 **k-word** ensemble (global - local)

- **parameter - free**

- identifies repetitions of **all lengths**

- **robust to copy deviations** from the perfect sample

- identifies higher order repeats **(HOR)**

- consensus lengths and sequences are **simply determined** from results obtained with GRM
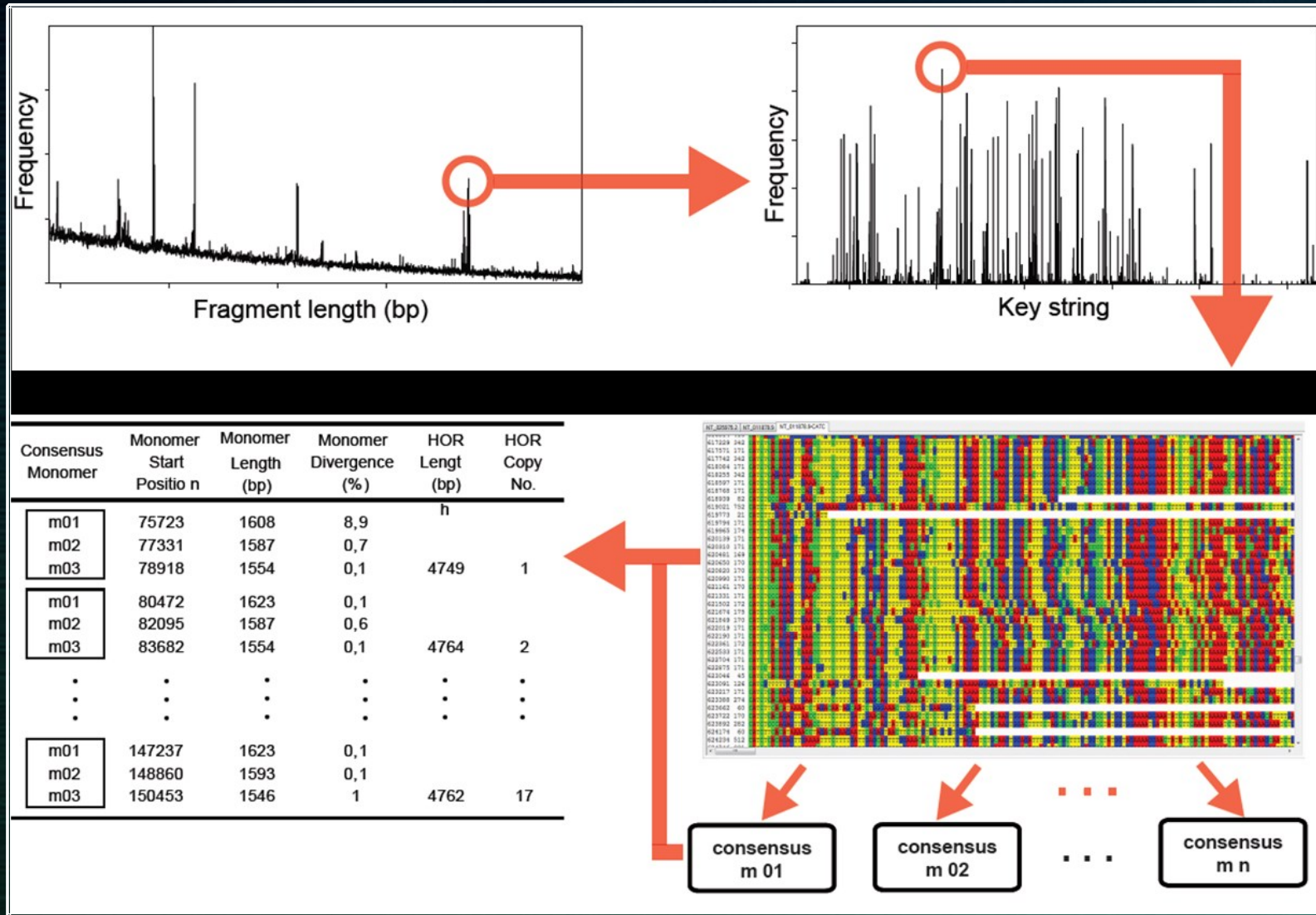
- „good" in combination with **BLAST**

Global Repeat Map method - GRM

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

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

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

- $\{f_K(j)\} = [f_K(j)]^1, [f_K(j)]^2, \dots, [f_K(j)]^\nu$

  $\rightarrow \{f_{K(E)}\} = \sum_{j=1}^N [f_K(j)]^1, \sum_{j=1}^N [f_K(j)]^2, \dots, \sum_{j=1}^N [f_K(j)]^\nu, N = 4^K$

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

GRM steps

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

NT_004487.20 KS8
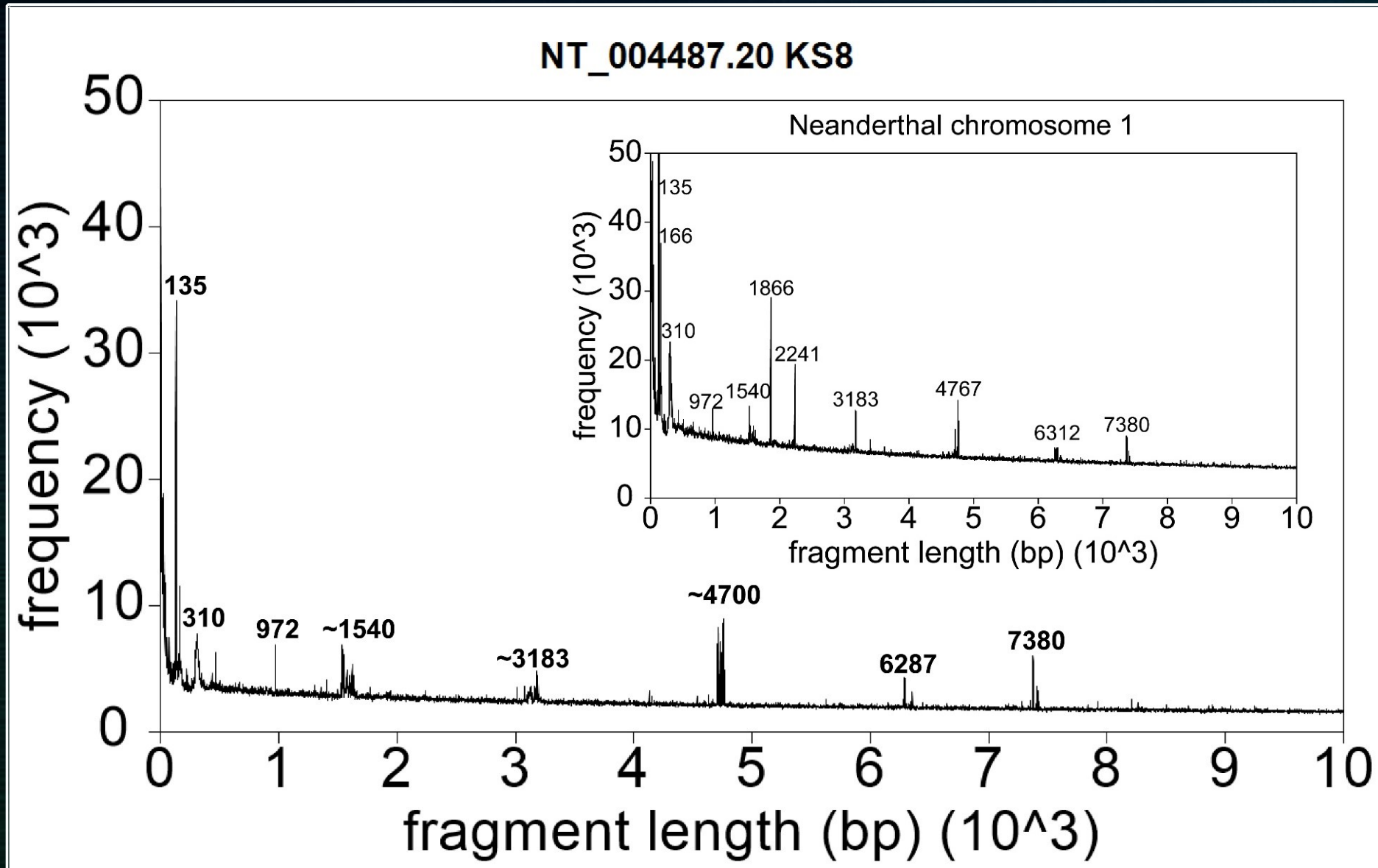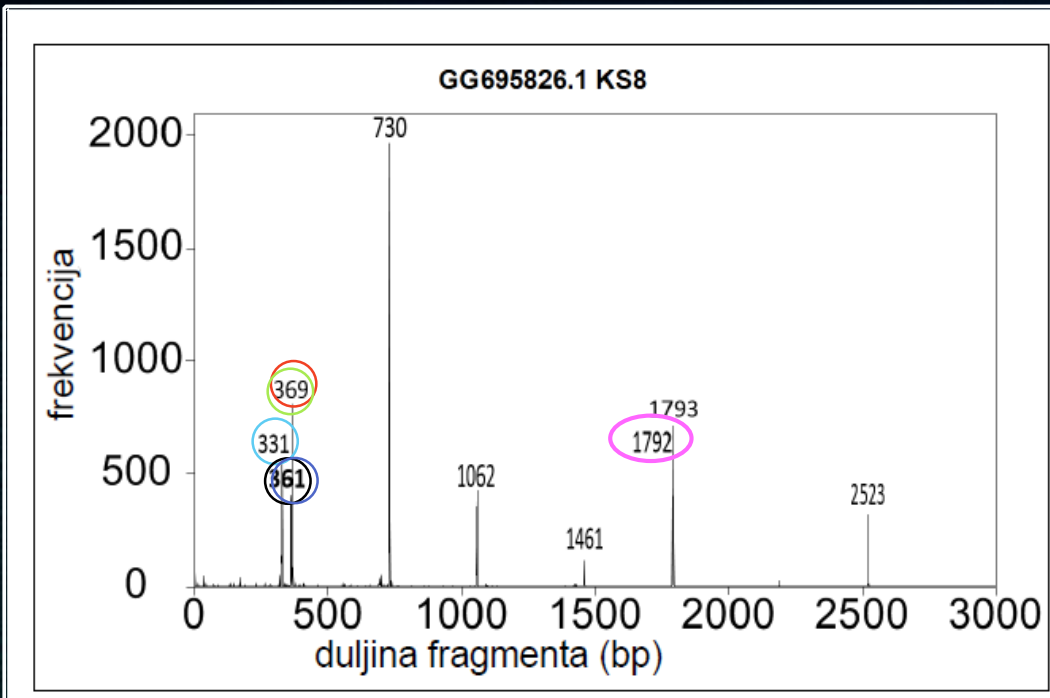
Monomer type	length of consensus(bp)	range of copy length
m1	331	328-332
m2	362	362-363
m3	369	369
m4	361	361-362
m5	369	369

Consensus length

Types of HOR structures
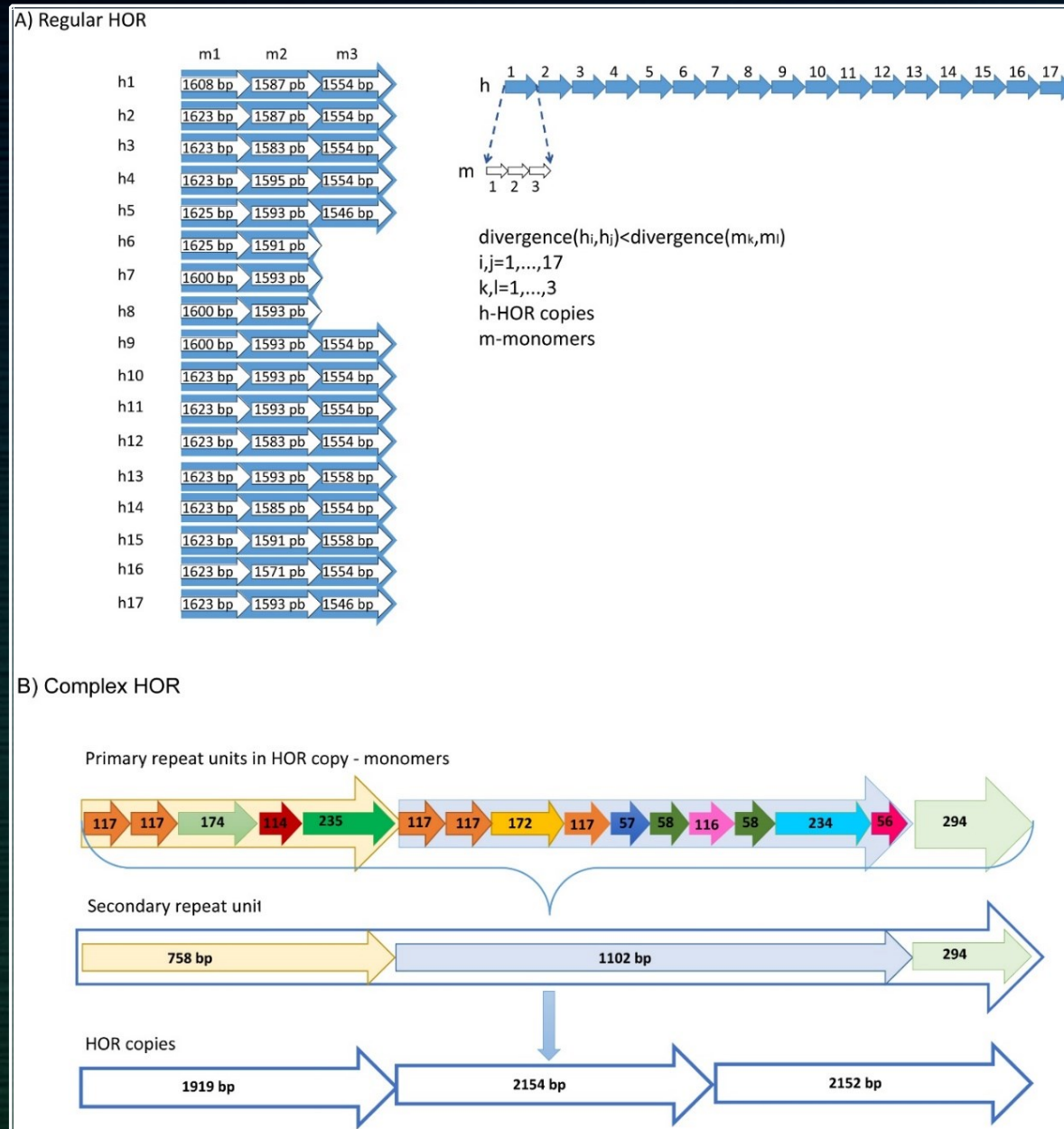
Vlahović I., Glunčić M., Rosandić M.,Ugarković Đ., Paar V.. Regular Higher Order Repeat Structures in Bettle *Tribolium castaneum* Genome. Genome Biol.Evol., 2016. A.) Regular HOR. B) complex HOR.

Results

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

Vladimir Paar,*†[1] Matko Glunčić,†[1] Marija Rosandić,[2] Ivan Basar,[1] and Ines Vlahović[1]

[1] Faculty of Science, University of Zagreb, Zagreb, Croatia
[2] Department of Internal Medicine, University Hospital Rebro, University of Zagreb, Zagreb, Croatia
†These authors contributed equally to this work.
*Corresponding author: E-mail: paar@hazu.hr.
Associate editor: James McInerney

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 intragene 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 11mer 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 hornerin gene, related to regenerating and psoriatic skin. Such an HOR is not found in chimpanzee chromosome 1. We find a remarkable human 3mer HOR organization based on the ~1.6-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 particularly 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 hornerin gene, global repeat map.

V.Paar, M.Glunčić, M.Rosandić, I.Basar, I.Vlahović , **Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees,** *Molecular Biology and Evolution*, Volume 28, Issue 6, June 2011, Pages 1877–1892

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)
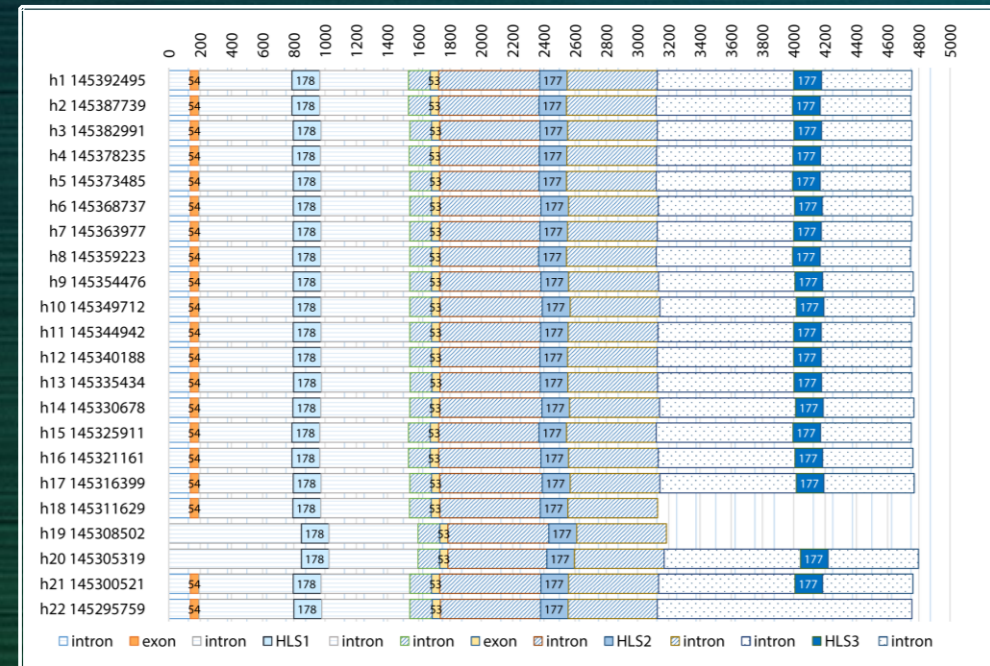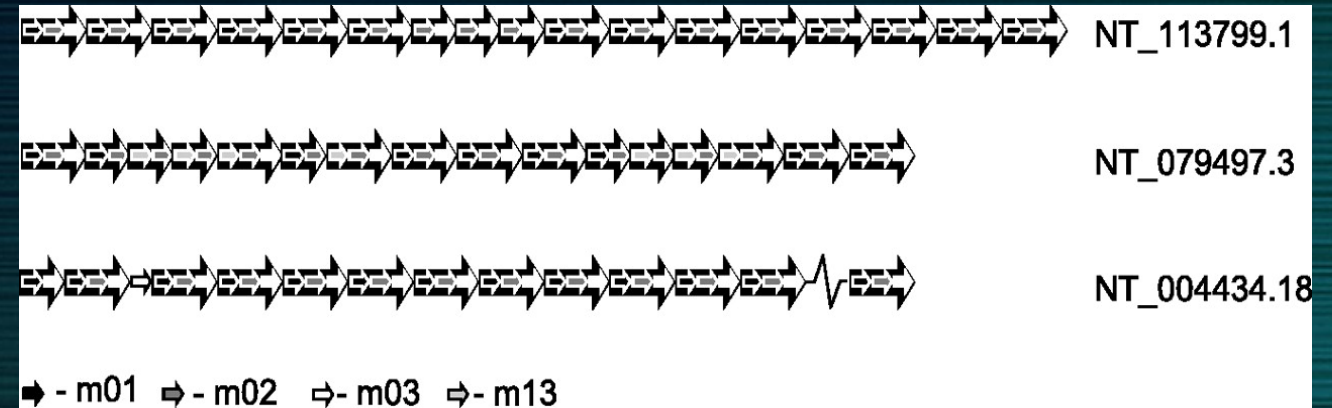
Schematic illustrating three NBPF 3mer HOR copies based on the ~1.6-bp monomers in human chromosome 1. (Results from 2019, Build hg38 assembly)

Results

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

Vladimir Paar,*†[1] Matko Glunčić,†[1] Marija Rosandić,[2] Ivan Basar,[1] and Ines Vlahović[1]

[1]Faculty of Science, University of Zagreb, Zagreb, Croatia

[2]Department of Internal Medicine, University Hospital Rebro, University of Zagreb, Zagreb, Croatia

†These authors contributed equally to this work.

*Corresponding author: E-mail: paar@hazu.hr.

Associate editor: James McInerney

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 intragene 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 11mer 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 hornerin gene, related to regenerating and psoriatric skin. Such an HOR is not found in chimpanzee chromosome 1. We find a remarkable human 3mer HOR organization based on the ~1.6-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 particularly 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 hornerin gene, global repeat map.

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

V.Paar, M.Glunčić, M.Rosandić, I.Basar, I.Vlahović, **Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees,** *Molecular Biology and Evolution*, Volume 28, Issue 6, June 2011, Pages 1877–1892
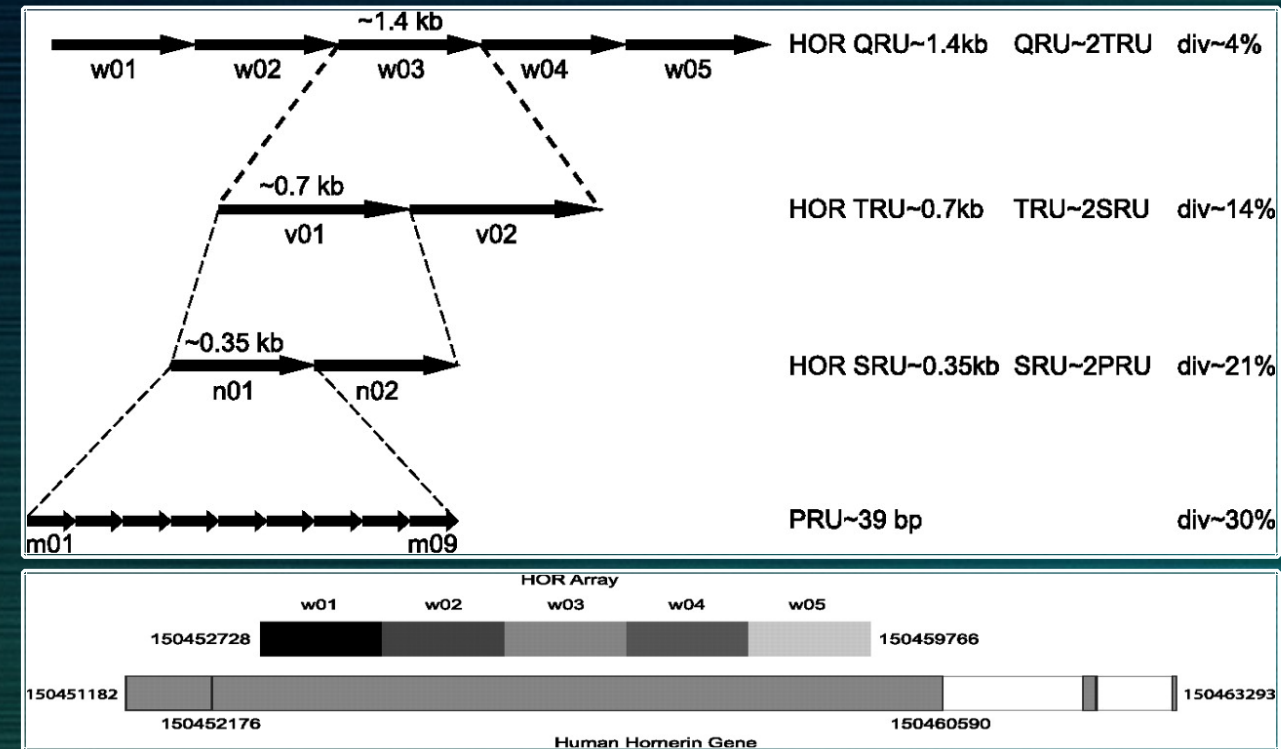
Confirmed by V.Romero et al. **High Order Formation and Evolution of Hornerin in Primates.** *Genome Biology and Evolution*, Volume 10, Issue 12, December 2018, Pages 3167–3175.

13

HRZZ project – Human-Neanderthal – Chimpanzee -Results
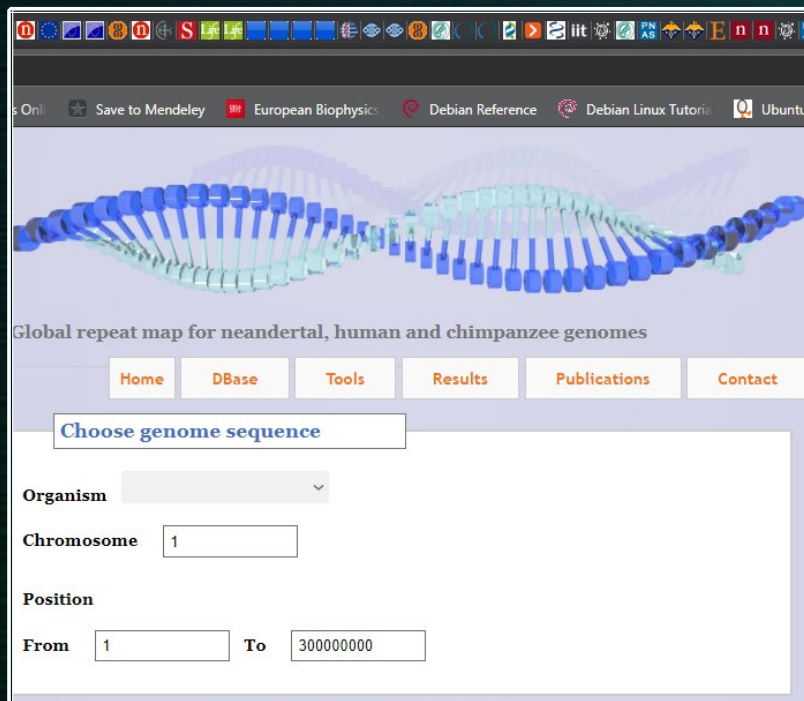
- Goal of the project:
  - Creation of the repetition database for human, Neanderthal and chimpanzee genomes (http://genom.hazu.hr)
- Over 2000 records
- **ALPHAsub** algorithm -extension
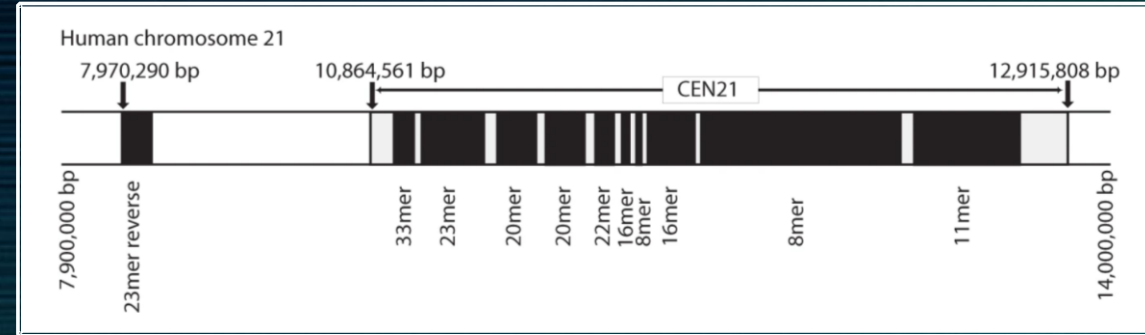
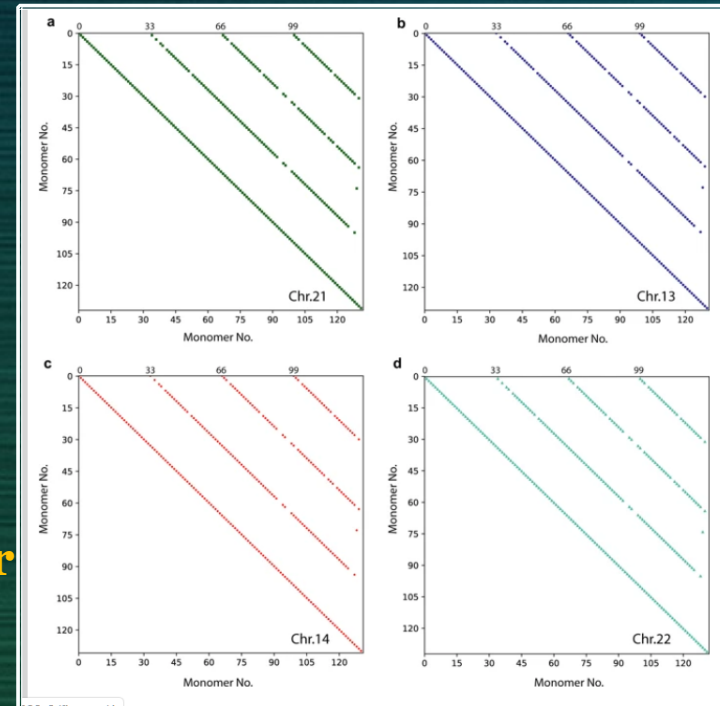# Results chromosome 21

SCIENTIFIC REPORTS
natureresearch

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

Matko Glunčić[1], Ines Vlahović[1,2] & Vladimir Paar[1,3]

The centromere is important for segregation of chromosomes during cell division in eukaryotes. Its destabilization results in chromosomal missegregation, aneuploidy, hallmarks of cancers and birth defects. In primate genomes centromeres contain tandem repeats of ~171 bp alpha satellite DNA, commonly organized into higher order repeats (HORs). In spite of crucial importance, satellites have been understudied because of gaps in sequencing - genomic "black holes". Bioinformatical studies of genomic sequences open possibilities to revolutionize understanding of repetitive DNA datasets. Here, using robust (Global Repeat Map) algorithm we identified in hg38 sequence of human chromosome 21 complete ensemble of alpha satellite HORs with six long repeat units (≥20 mers), five of them novel. Novel 33mer HOR has the longest HOR unit identified so far among all somatic chromosomes and novel 23mer reverse HOR is distant far from the centromere. Also, we discovered that for hg38 assembly the 33mer sequences in chromosomes 21, 13, 14, and 22 are 100% identical but nearby gaps are present; that seems to require an additional more precise sequencing. Chromosome 21 is of significant interest for deciphering the molecular base of Down syndrome and of aneuploidies in general. Since the chromosome identifier probes are largely based on the detection of higher order alpha satellite repeats, distinctions between alpha satellite HORs in chromosomes 21 and 13 here identified might lead to

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)

Alpha satellite HOR ideogram for linear positioning of alpha satellite HOR arrays with long repeat units (n ≥ 8).
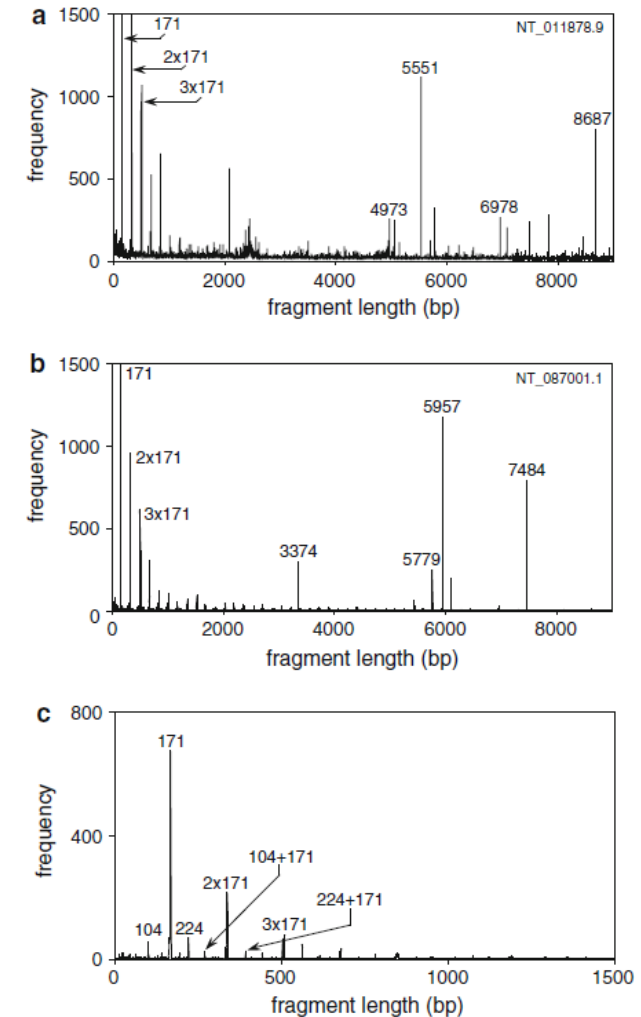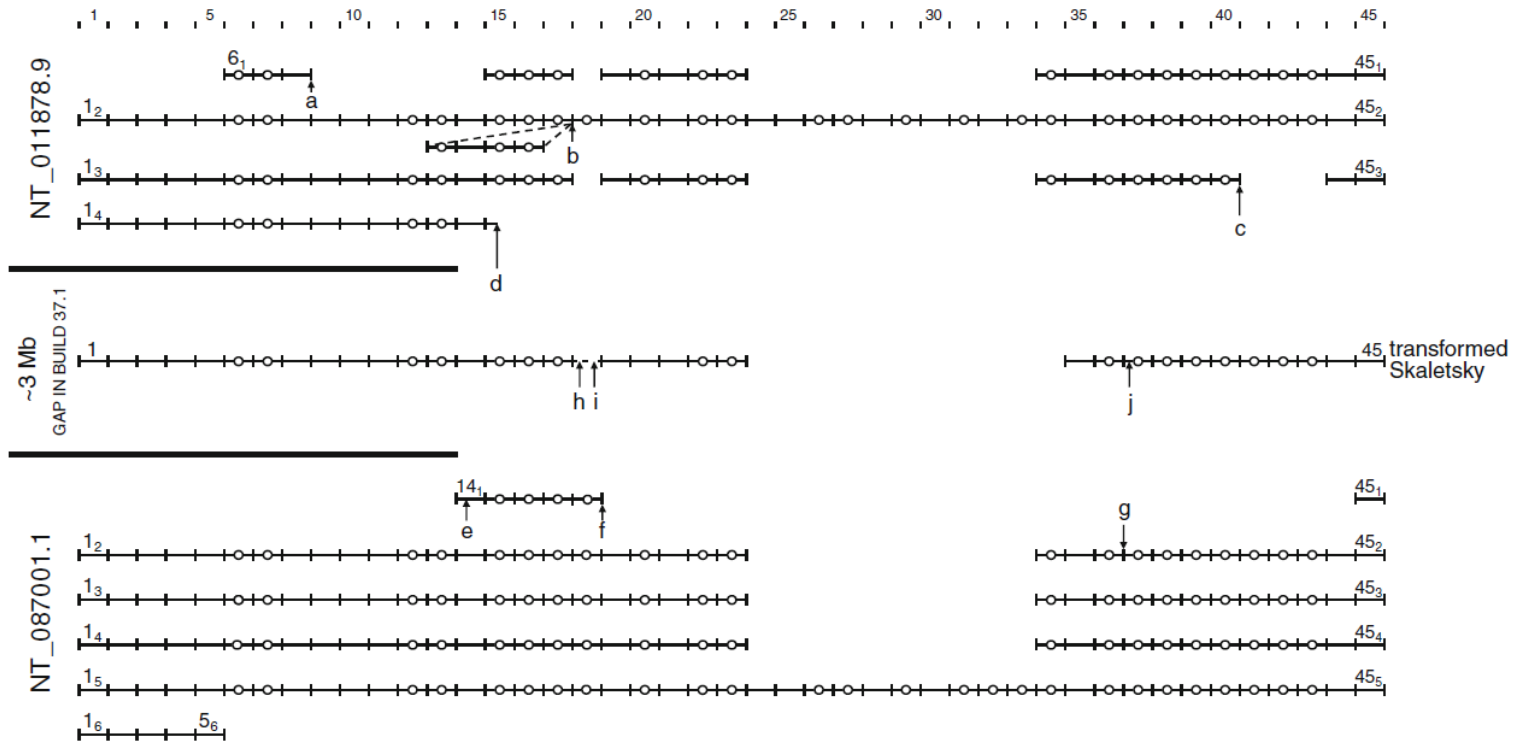

Dot-matrix plots of 33mer HORs in four acrocentric chromosomes. (**a**) chromosome 21; (**b**) chromosome 13; (**c**) chromosome 14; (**d**) chromosome 22.

15

Results chromosome Y

Schematic presentation of aligned monomer structure of 45mer alphoid HOR (consensus length 7662 bp) in human chromosome Y (Build 37.1). V. Paar , M. Glunčić, I. Basar, M.Rosandic, P. Paar, M. Cvitković. **Large Tandem, Higher Order Repeats and Regularly Dispersed Repeat Units Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes**. 2011, Journal of Molecular Evolution 72(1):34-55
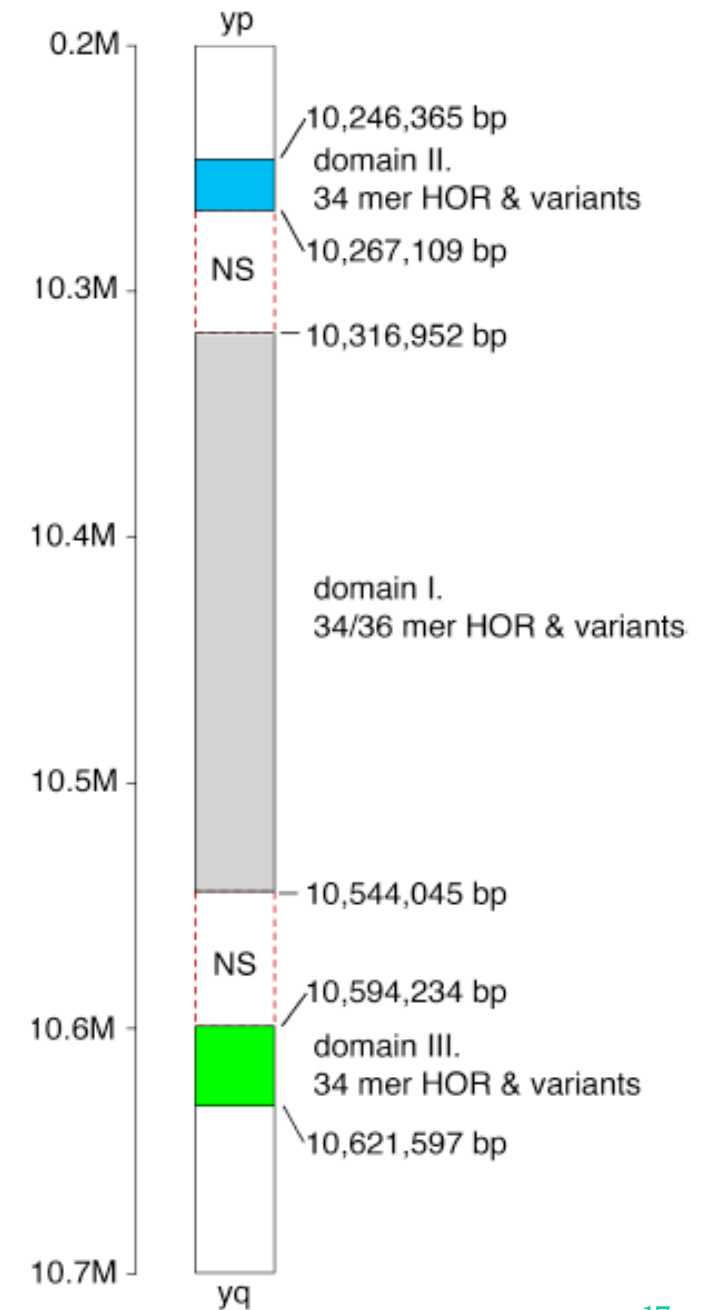
Results chromosome Y

GRM diagram and ideogram for human chromosome Y (Build hg38).
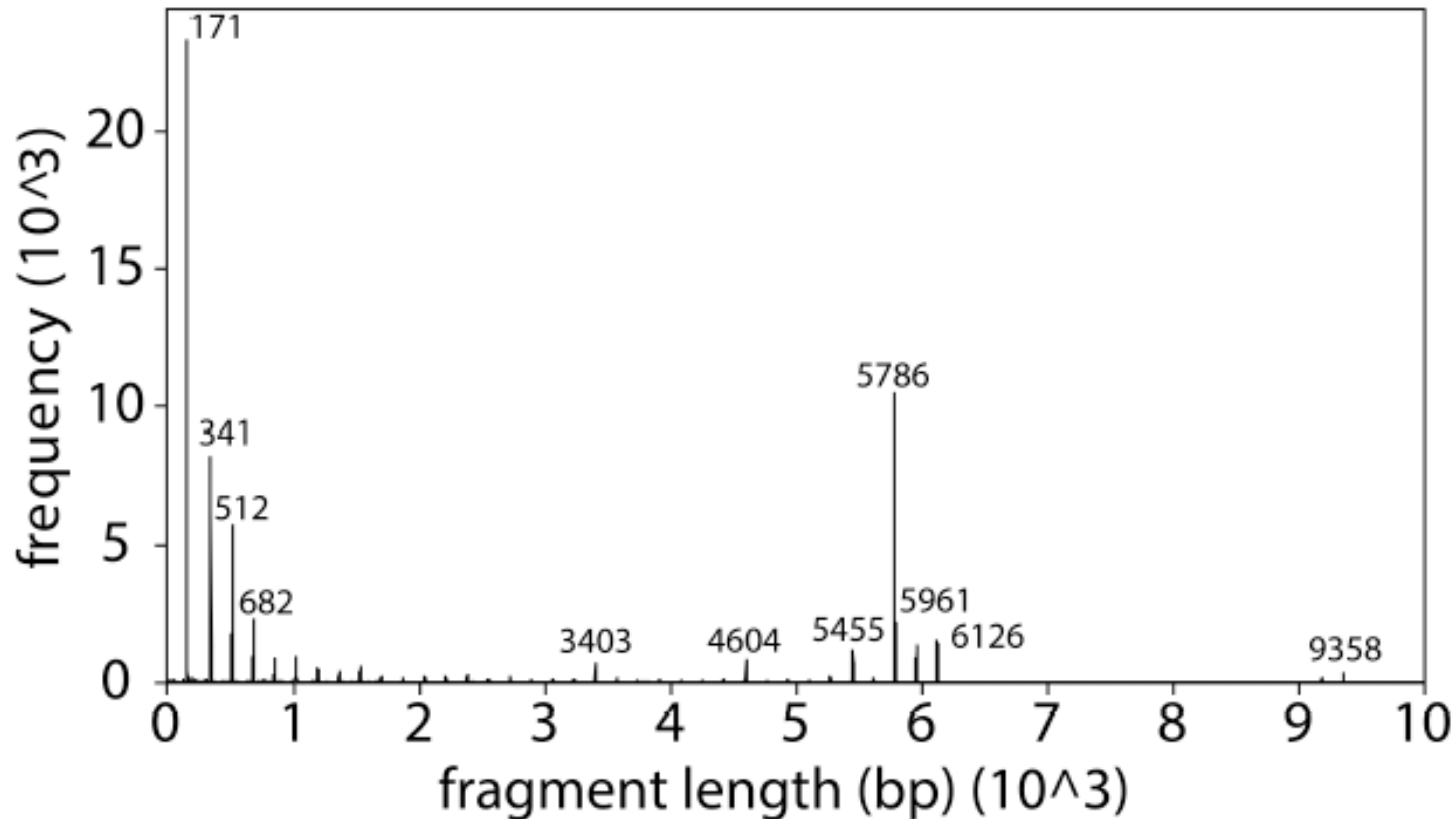
I. Vlahović, M. Glunčić, V. Paar. **Rich polymorphic variants of alpha satellite 34mer higher order repeats in hg38 assembly of human chromosome Y.** (submitted paper)

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
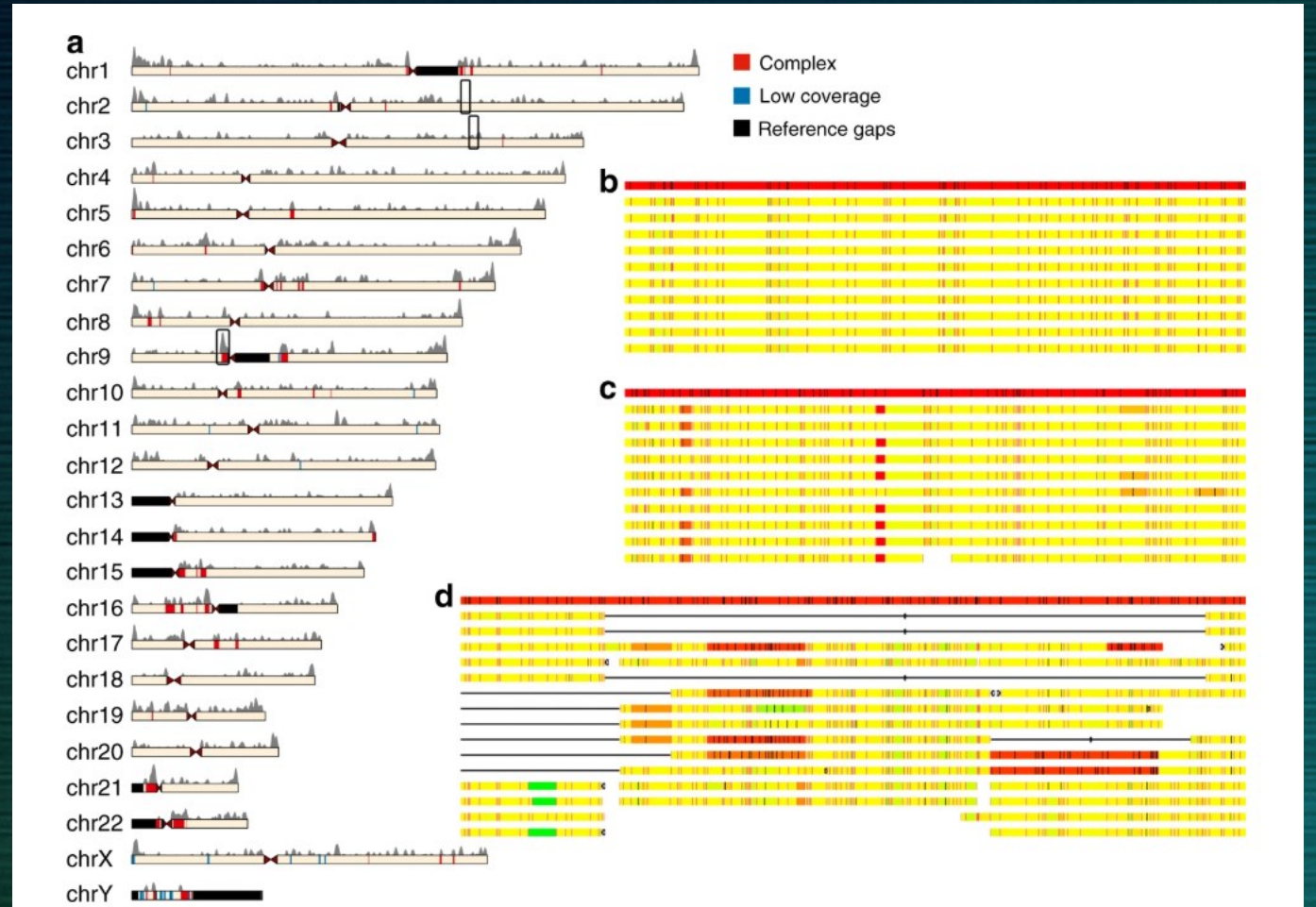
Michal Levy-Sakin et al. **Genome maps across 26 human populations reveal population-specific patterns of structural variation.** Nature Communications volume 10, Article number: 1025 (2019).

Future work

DNA analysis of some **diseases:**

- autism genomes
  https://research.mss.ng/

- cancer genomes
  https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

- **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