Global Repeat Map (GRM) algorithm as genomic technology for Higher Order Repeat identification

Algoritam globalne karte ponavljanja (Global Repeat map – GRM algorithm) kao genomska tehnologija za identifikaciju periodičnosti višeg reda

Ines Vlahović
Algebra University College
Visoko učilište Algebra
Gradišćanska 24, Zagreb
Faculty of Science, Department of Physics
Prirodoslovno-matematički fakultet, Fizički odsjek
University of Zagreb
Sveučilište u Zagrebu
Bijenička cesta 32, Zagreb
ines.vlahovic@algebra.hr

***Abstract**. In recent years new technologies for DNA sequence widened research of the roles of tandem repeats in eukaryotes, especially ones that are located in centromere region in chromosomes because of their importance for chromosome segregation, microtubule attachment to kinetochores and, as recently showed, their role in some diseases such as cancers. For identification of tandem repeats and higher order repeats (HORs) we use Global Repeat Map (GRM) algorithm and AlphaSub (ASu) algorithm, developed at Department of Physics, Faculty of Science, University of Zagreb. Here we present some illustrative results obtained using this tool. Based on analysis of DNA sequences, we create a database including information on location, length and divergence for alpha satellite arrays and HORs in human, Neanderthal and chimpanzee genomes (at link http://genom.hazu.hr). We summarize analysis of alpha satellite HORs for human Y chromosome, as well as for some other HORs identified in human and other primates' chromosome 1 with novel comparison of ~1.6 kb monomer units in them. We also show preliminary results of GRM analysis of three breast cancer exomes. In this respect, we emphasize a need for employing new sequencing methods to determine complete genome assemblies for identification of more precise HOR structures.*

Keywords. genomes, higher order repeat, tandem repeat, cancer, alpha satellites, NGS, human genome, Y chromosome, NBPF genes, molecular databases

***Sažetak**. Posljednjih godina nove tehnologije za sekvenciranje DNK sekvenci proširile su istraživanja uloga tandemnih repeticija u eukariotima, osobito onih koji se nalaze u području centromere u kromosomima zbog njihove važnosti za segregaciju kromosoma, spajanju mikrotubuli na kinetohore i, kao što je nedavno pokazano, u ulogama u nekim bolestima poput karcinoma. Za identifikaciju tandemnih repeticija i periodičnosti višeg reda (Higher order repeats - HORs) koristimo u našim istraživanjima algoritam globalne karte ponavljanja (Global Repeat Map (GRM) algorithm) te AlphaSub (Asu) algoritam koji su razvijeni na Fizičkom odsjeku, Prirodoslovno-matematičkog fakulteta, Sveučilišta u Zagrebu. U ovom radu predstavljamo dio ilustrativnih rezultata dobivenih pomoću ovih alate. Temeljeno na analizi DNK sekvenci, stvorili smo bazu podataka koja uključuje informacije o pozicijama, duljinama i divergencijama za nizove alfa satelita i HORova za genome čovjeka, neandertalca i čimpanzu (na http://genom.hazu.hr). U ovom radu sumiramo analizu HORova zasnovanih na alfa satelitima za ljudski kromosom Y, kao i za neke druge HORove identificiranje u kromosomu 1 čovjeka te drugih primata s novom usporedbom ~1.6 kb monomerne jedinice u njima. Također, prikazati ćemo preliminarne podatke GRM analize za egzome tri slučaja karcinoma dojke. U tom smislu, naglašavamo potrebu za primjenom novih metoda sekvencioniranja za određivanje potpunih*

1 Introduction

In recent years new sequencing methods have emerged (NGS, nanopore sequencing, single molecule real time sequencing, optical mapping) that enabled collection of huge amounts of data from sequencing projects. With the increasing amount of collected data, repositories were created for storing DNA and protein sequences as well as tools for their exploration and analysis. Recent review paper with list of existing databases, ~1600 databases categorized in 15 categories and 43 subcategories, was published in which their overview was given (Rigden & Fernández, 2022). From collection of databases, NAR online Molecular Biology Database Collection, we use data from some relevant databases for our research, like NCBI, Ensemble, ClinVar, GWAS, TCGA, NCI GDC. Although those data are mostly used in investigation of genetic content, a few are oriented toward repetitive DNA sequences (TRDB, STRBase, Plant Repeat Database, RepBase). For years, the repetitive sequences were considered as “junk” elements, but recently there are more and more evidence that they might have an important role and could be responsible for some diseases such as Huntington, schizophrenia and even cancer [Fondon & Garner, 2004, Glunčić, Vlahović, & Paar, 2019]. Repetitive sequences can emerge in DNA sequence as tandem repeats, dispersed repeats, segmental duplications, complex repeats, and higher order repeats. Some of our main results obtained with GRM algorithm and AlphaSub algorithm belong to the analysis of satellite DNA i.e., research of human alpha satellites of ~171 bp monomer unit length that form high range of HORs – like the one in chromosome 21, (34/36mer HOR) that is largest among human somatic chromosomes, and HORs in human Y chromosome. In addition, our algorithms are very effective with identification of repeats for long monomer units like NBPF monomers with length of ~1.6 kb (classified as mega satellites) as well as smaller monomers (mini and micro satellites in hornerin gene) that are used for DNA fingerprinting. Fig.1 shows a more detailed classification according to ref. (Pathak & Ali, 2012).

In human genome, repetitive elements occupy ~50% of DNA sequence (Glunčić & Paar, 2013). It is shown that tandem repeats have roles such as markers for genotyping a number of pathogens (Flèche, 2001), that copy number variation can influence molecular paths in individuals and that they could also be a reason for phenotype changes in population (Romero, Hosomichi, Nakaoka, Shibata, & Inoue, 2017). One of the most investigated tandem repeats are human centromeric alpha satellites with length of ~171 bp monomer unit that can form higher order repeats (HORs) although their exact schemes with variant copies are not given. Satellites are important for processes of chromosome segregation in cell division which might, if disrupted, cause

genomskih sekvenci za identifikaciju preciznije opisanih HOR struktura.

Ključne riječi. genomi, periodišnosti višeg reda, tandemsko ponavljanje, karcinom, alfa sateliti, NGS, ljudski genom, Y kromosom, NBPF geni, molekularne baze podataka

1 Uvod

Posljednjih godina pojavile su se nove metode sekvenciranja (NGS, sekvenciranje nanopora, sekvenciranje pojedinačne molekule u stvarnom vremenu, optičko mapiranje) koje su omogućile prikupljanje ogromnih količina podataka iz projekata sekvenciranja. S povećanjem količine prikupljenih podataka stvoreni su repozitoriji za pohranjivanje sekvenci DNK i proteina kao i alati za njihovo istraživanje i analizu. Nedavno je objavljen pregledni rad s popisom postojećih baza podataka, ~1600 baza podataka kategoriziranih u 15 kategorija i 43 potkategorija, u kojem je dan njihov pregled (Rigden & Fernández, 2022). Iz zbirke baza podataka, NAR online Molecular Biology Database Collection, koristimo neke relevantne baze podataka u našim istraživanjima, kao što su NCBI, Ensemble, ClinVar, GWAS, TCGA, NCI GDC. Iako se podaci iz tih baza uglavnom koriste u istraživanjima gena, neke od baza su orijentirane na ponavljajuće sekvence DNK (TRDB, STRBase, Plant Repeat Database, RepBase). Godinama su se ponavljajuće sekvence smatrale elementima "smeća", no nedavno je sve više dokaza da bi upravo one mogle imati važnu ulogu te da bi one mogle biti odgovorne za neke bolesti poput Huntingtonove, shizofrenije, pa čak i karcinoma [Fondon & Garner, 2004, Glunčić, Vlahović, & Paar, 2019]. Repeticije sekvenci DNK se mogu pojaviti kao tandemna ponavljanja, disperzirana ponavljanja, segmentne duplikacije, složena ponavljanja te periodičnosti višeg reda (HOR). Neki od naših glavnih rezultata dobivenih s GRM metodom i AlphaSub algoritmom pripadaju analizi satelitske DNA tj. istraživanjima ljudskih alfa satelita s ~171 bp monomernom jedinicom koja može tvoriti veliki raspon HORova – npr. u kromosomu 21 nalazi se najveći među somatskim kromosomima 34/36mer HOR, te HOR strukture u ljudskom Y kromosomu. Dodatno, naši algoritmi su vrlo učinkoviti za repeticije s duljim monomerskim jedinicama poput NBPF monomera duljine od ~1.6 kb (klasificirani kao megasateliti) ali i za manje monomerne jedinice (kao što su mikro i mini sateliti u hornerin genu) koji se koriste u metodama kao što je DNK otisak prsta. Slika 1 prikazuje detaljniju klasifikaciju prema izvoru (Pathak & Ali, 2012).

U ljudskom genomu repeticije zauzimaju ~50% sekvence DNK (Glunčić & Paar, 2013). Pokazalo se da tandemne repeticije mogu poslužiti kao markeri za genotipizaciju brojnih patogena (Flèche, 2001) te missegregation, aneuploidy and cancer [A.K., Mourad, Kaplan, & et al., 2019, Levy-Sakin, Pastor, Mostovoy, & et al., 2019, Black & Giunta, 2018].

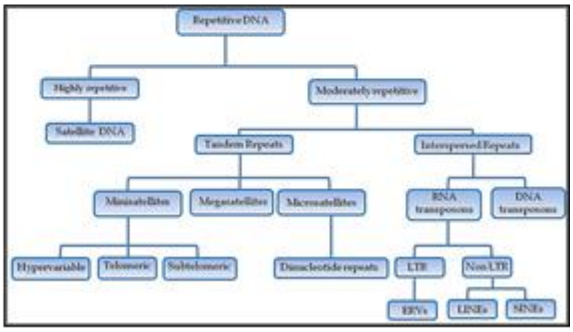
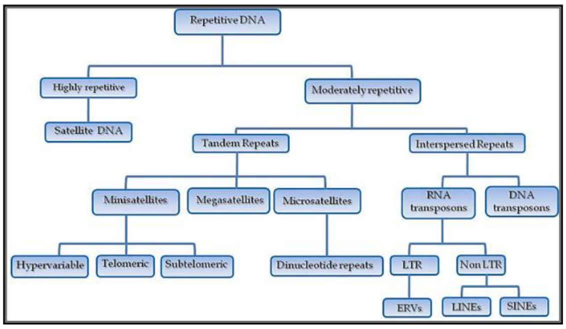


Figure 1. Repetitive elements classification from (Pathak & Ali, 2012). In this paper, we analyze satellite DNA, tandem repeats and HORs in primates for population studies (NBPF copy number in primates, Neanderthal and human genomes and example of hornerin gene in chromosome 1) and breast cancer exomes with our GRM and AlphaSub algorithm.

We organise this paper in chapters 2 Methods, 3 Results and 4 Conclusion. In the chapter Methods, we describe Global Repeat Map algorithm and AlphaSub algorithm for identification of tandem repeats and alpha satellite sequences. In 2.1. we give links to data used for our research and which databases store those datasets. In chapter 3, Results, we show analysis of human Y chromosome, chromosome 21, primates and human chromosome 1 and novel comparison of NBPF (Neuroblastoma BreakPoint Family) monomer, dimer and trimer repeat units among human, primates and Neanderthal genomes. As novel preliminary result, we show GRM diagram for three cases of whole exome sequences of breast cancer from GDC data portal from CCLE project. One possible way for using GRM analysis could be for diagnostic purposes and classification of cancer exome sequences or whole genomes from various available datasets (from projects such as TCGA, CCLE). Our chapter 4 Conclusion emphasize the need for applying novel sequencing methods in whole genome projects, especially for population studies and studies of different diseases (cancer, autism).

Related work necessary for this research paper included investigation of databases and their usage, as well as some additionally tools and algorithms implemented for DNA analysis like Blast algorithm and tools like UGENE for making consensus sequence in fasta format. The most common database for genomic sequence that we use is NCBI that host a vast number of sequenced species, including viruses, bacteria and eukaryotes that enable us to conduct population studies. Blast algorithm enabled us to gain a novel insight in sequence evolution and their emergence in primate genomes like our novel

da varijacije broja kopija u njima mogu utjecati na određene molekularne putove kod pojedinaca. Također tandemna ponavljanja bi mogla biti uzrok i u promjenama fenotipa u populaciji (Romero, Hosomichi, Nakaoka, Shibata & Inoue, 2017). Jedna grupa najistraživijih tandemskih repeticija su centromerni alfa sateliti duljine ~171 bp koje se mogu dodatno pojaviti i kao periodičnosti višeg reda (HOR) iako do sada njihove egzaktne sheme s varijantnim kopijama nisu dane. Sateliti su važni za procese segregacije kromosoma u staničnoj diobi koji bi, ako se poremete, mogli uzrokovati pogrešnu segregaciju, aneuploidiju i karcinome [A.K., Mourad, Kaplan & sur., 2019, Levy-Sakin, Pastor, Mostovoy & sur., 2019, Black & Giunta, 2018].



Slika 1. Klasifikacija repeticijskih elemenata u DNK (Pathak & Ali, 2012). U ovom radu analiziramo satelitsku DNK, tandemna ponavljanja i HOR-ove kod primata za populacijske studije (broj kopija NBPF-a kod primata, genoma neandertalaca i ljudi te primjer hornerin gena u kromosomu 1) i egzome karcinoma dojke s našim GRM i AlphaSub algoritmom.

Ovaj rad je strukturiran kroz poglavlja 2 Metode, 3 Rezultati i 4 Zaključak. U poglavlju Metode, dajemo opis GRM algoritma i algoritma AlphaSub za identifikaciju tandemnih repeticija te sekvenci alfa satelita. U 2.1. dajemo poveznice na podatke koje koristimo u našem istraživanju i podatke o bazama podataka koji te podatke pohranjuju. U poglavlju 3, Rezultati, prikazujemo analizu ljudskog Y kromosoma, kromosoma 21, kromosoma 1 za čovjeka i primate te novu usporedbu NBPF (Neuroblastoma Break Point Family) monomernih, dimerskih i trimerskih repeticijskih jedinica među genomima čovjeka, neandertalca i primata. Kao novi preliminarni rezultati, također prikazujemo GRM dijagram za tri slučaja karcinoma dojke s GDC podatkovnog portala iz CCLE projekta. Jedan mogući način korištenja GRM algoritma za analize može poslužiti u dijagnostici i klasifikaciji sekvenci egzoma karcinoma ili cijelih genoma iz različitih dostupnih podatkovnih setova (iz projekta TCGA, CCLE). U poglavlju 4 Zaključak naglašavamo potrebu za primjenom novih metoda sekvenciranja u projektima cijelih genoma, posebno u svrhu results for NBPF gene family. Our interpretation of those novel results takes in consideration assemblies used, because of percentage of completeness of genomic sequence and sequencing techniques used. We stress out that combination of different sequencing techniques should fill sequence gaps in the future. Filling the gaps in genomic sequence will enable us to create a more complete picture of tandem repeats and HORs, especially in centromeric region that is responsible for correct chromosome segregation like for human Y, 1 and 21 chromosomes presented in this paper.

Also, the existing databases that store genome and exome data for different diseases (TCGA, CCLE, TARGET, dbGaP, ClinVar, COSMIC, MSSNG, et cetera) could help us to understand progression of diseases and their connection to structural variants and copy number variants. Examples of those kind of diseases are autism, macrocephaly, microcephaly and neuroblastoma that are mentioned earlier and they are connected to NBPF monomer indels (deletions, insertions).

Complete schemes of structural variants and polymorphic copies in HORs could be used as a diagnostic tool or for population studies. In that regard tools such as our GRM and AlphaSub algorithm can be used, but also novel algorithms based on machine learning and AI should be developed that could bring new insight in pattern recognition in studying DNA from large data sets.

2 Methods

For investigation of repeats in DNA sequences of different eukaryotes, we use Global Repeat Map algorithm developed at University of Zagreb, Faculty of Science, Physics department. This method was explained in details in ref. (Gluncić & Paar, 2013), and here we summarize the main points characterizing this tool:

1. It directly maps the DNA symbolic sequence into the frequency domain making „GLOBAL MAP" based on combination of statistical and digital processing methods (Glunčić & Paar, 2013).
2. It uses a complete k-word ensemble (global - local) for creation of global map (Fig. 2).
3. It is parameter – free.
4. It can identify repetitions of all lengths.
5. It is robust to copy deviations from the perfect sample.
6. It identifies higher-order repeats (HOR).
7. We can simply determine consensus lengths and sequences from GRM results.
8. It is „good" for use in combination with BLAST for similar sequence.

This tool is free for download at http://genom.hazu.hr/tools.html. On given link instructions for its usage are given with examples.

populacijskih istraživanja i istraživanja raznih bolesti (karcinomi, autizam).

Povezani rad neophodan za ovo istraživanje uključuje istraživanje baza podataka i njihove upotrebe, kao i neke dodatne alate i algoritme implementirane za analizu DNK kao što je Blast algoritam i alate kao što je UGENE za izradu konsenzusnog sekvence u fasta formatu. Najčešća baza podataka za genomsku sekvencu koju koristimo je NCBI koja sadrži veliki broj sekvenciranih vrsta, uključujući viruse, bakterije i eukariote koji nam omogućuju provođenje populacijskih istraživanja. BLAST algoritam nam je omogućio da steknemo novi uvid u evoluciju sekvenci i njihovu pojavu u genomima primata, poput naših novih rezultata za obitelj gena NBPF. Naše tumačenje tih novih rezultata uzima u obzir korištene sklopove genomskih sekvenci (assembly), zbog postotka potpunosti genomske sekvence i korištenih tehnika sekvenciranja. Naglašavamo da bi kombinacija različitih tehnika sekvenciranja trebala popuniti praznine u sekvencama u budućnosti. Popunjavanje praznina u genomskoj sekvenci omogućit će nam stvaranje cjelovitije slike tandemnih ponavljanja i HORova, posebno u centromernom području koje je odgovorno za ispravnu segregaciju kromosoma kao što pokazuju primjeri za ljudske Y, 1 i 21 kromosome koji su predstavljeni u ovom radu.

Također, postojeće baze podataka koje pohranjuju podatke o genomu i egzomu za različite bolesti (TCGA, CCLE, TARGET, dbGaP, ClinVar, COSMIC, MSSNG, itd.) mogle bi nam pomoći da razumijemo progresiju bolesti i njihovu povezanost sa strukturnim varijantama i varijantama broja kopija.

Primjeri takvih bolesti su autizam, makrocefalija, mikrocefalija i neuroblastom koji su ranije spomenuti i povezani su s NBPF monomernim indelima (delecije, insercije).

Potpune sheme strukturnih varijanti i polimorfnih kopija u HOR-ovima mogu se koristiti kao dijagnostički alat ili za populacijske studije. U tom smislu mogu se koristiti alati kao što su naš algoritam GRM i AlphaSub, ali također bi se trebali razviti novi algoritmi temeljeni na strojnom učenju i umjetnoj inteligenciji koji bi mogli donijeti novi uvid u prepoznavanje uzoraka u proučavanju DNK sekvenci dobivenih iz velikih skupova podataka.

2 Metode

Za istraživanje repeticija u sekvencama DNK različitih eukariota koristimo algoritam globalne karte ponavljanja - Global Repeat Map (GRM) razvijen Fizičkom odsjeku, Prirodoslovno-matematičkog fakulteta, Sveučilišta u Zagrebu. Ova metoda je detaljno objašnjena u izvoru (Glunčić & Paar, 2013), a ovdje sažimamo glavne prednosti koje karakteriziraju ovaj alat:

Example of using GRM method is shown in Fig.2 where we analysed satellites (with monomer unit lengths in range from ~331 bp to ~369 bp) from beetle T.Castaneum genome (Vlahovic, Gluncic, Rosandic, Ugarkovic, & Paar, 2017).

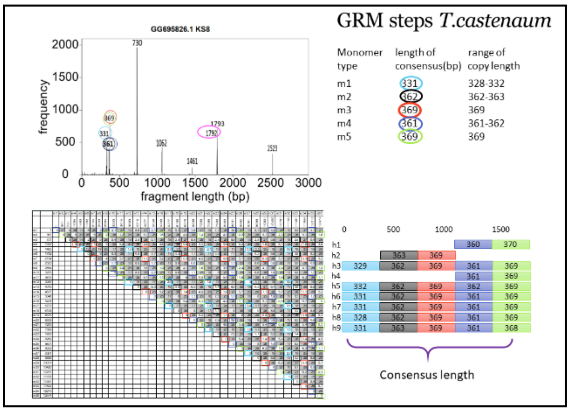


Figure 2. GRM diagram for Tcast satellites show monomer units that are organized in five monomer types as GRM diagram shows (upper left). Needleman–Wunsch algorithm is applied to monomer DNA sequence for their classification into different monomer types according to sequence divergence. Divergence between monomers in array reveals higher order structure of length ~1792 bp i.e., 5mer HOR. Monomers that belong to the same type (with low divergence) have same color in Needleman-Wunsch matrix. This result/example is published in (Vlahovic, Gluncic, Rosandic, Ugarkovic, & Paar, 2017). Currently we are developing novel algorithms for automatization of this process.

We developed and used for automatization of alpha satellites identification in chromosomes additional algorithm AlphaSub, described in ref. (Glunčić, Vlahović, & Paar, 2019). This algorithm is based on 28 bp fragment of alpha satellite consensus sequence "TGAGAAACTGCTTTGTGATGTGTGCATT" as ideal "key word" as well as its reverse complement and by using Levenshtein distance algorithm we identify position of those alpha satellites ~171 bp in chromosome DNA sequence. We also used this method for analysis of human chromosome Y as well as whole human, Neanderthal and chimpanzee genomes. Data for those genomes and their alpha satellite arrays can be found at http://genom.hazu.hr. This tool can be downloaded from the link http://genom.hazu.hr/tools.html.

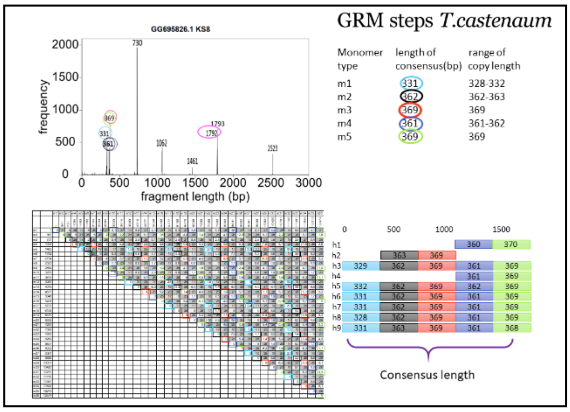
2.1 *Data*

Data used for our research are mostly downloaded from NCBI database ftp site as one of the most used databases in this field of research – NCBI (National Center for Biotechnology Information,

1. Izravno preslikava simboličku sekvencu DNK u frekvencijsku domenu čineći „GLOBALNU MAPU" na temelju kombinacije statističkih i digitalnih metoda obrade signala (Glunčić & Paar, 2013).
2. Koristi kompletan skup k-riječi (globalno - lokalno) za izradu globalne mape (Slika 2).
3. Nema ulaznih dodatnih parametara.
4. Može identificirati repeticije svih duljina.
5. Robustan je za odstupanja repeticija od savršenog uzorka.
6. Identificira periodičnosti višeg reda (HOR).
7. Jednostavno se određuju konsenzusne duljine repeticija i nizovi sekvenci iz GRM rezultata.
8. „Dobar" je za korištenje u kombinaciji s algoritmom BLAST za traženje sličnih nizova.

Ovaj je alat besplatan za preuzimanje na http://genom.hazu.hr/tools.html. Na navedenom linku dane su upute za njegovu uporabu s primjerima.

Primjer korištenja GRM metode prikazan je na Slici 2 gdje smo analizirali satelite (s duljinama monomera u rasponu od ~331 bp do ~369 bp) iz kukca *T.Castaneum* (Vlahovic, Gluncic, Rosandic, Ugarkovic, & Paar, 2017).



Slika 2. GRM dijagram za Tcast satelite koji su organizirani u pet vrsta monomera (primarnih repeticijskih jedinica) kao što prikazuje GRM dijagram (gore lijevo). Needleman-Wunsch algoritam primijenjen je na sekvencu monomerne DNA za njihovu klasifikaciju u različite tipove prema međusobnoj divergenciji monomera. Divergencija između monomera u nizu alfa satelita otkriva strukturu višeg reda duljine ~1792 bp, tj. 5mer HOR. Monomeri koji pripadaju istom tipu (s malom divergencijom) iste su boje u Needleman-Wunsch matrici. Ovaj rezultat je objavljen u radu (Vlahović, Glunčić, Rosandić, Ugarković & Paar, 2017). Trenutno razvijamo nove algoritme za automatizaciju ove procedure.

Za automatizaciju identifikacije alfa satelita u kromosomima razvili smo i koristili dodatni algoritam ALPHAsub, opisan u (Glunčić, Vlahović & Paar, 2019). Ovaj se algoritam temelji na 28 bp fragmentu konsenzusne sekvence alfa satelita https://ftp.ncbi.nlm.nih.gov) for human and primate genomes. We also give links to Nacional Cancer Institute Genomic Data Commons Data Portal – NCI GDC portal (https://portal.gdc.cancer.gov) with cancer databases. NCI GDC portal includes most commonly used cancer databases such as TCGA (The Cancer Genome Atlas), Target (Therapeutically Applicable Research to Generate Effective Therapies) i.e., they host around 72 projects with 828104 files. It also includes Cancer Cell Line Encyclopaedia Project (CCLE) data for breast cancer sequences that we used for our novel preliminary data analysis shown in this paper.

Additional DNA sequence for GRM analysis of chromosome 1 (assembled chromosomes) in *Pan Troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan paniscus, Macaca fascicularis, Macaca mulatta,* are downloaded from NCBI ftp site:

- *Pan troglodytes* - Pan_tro_3.0 (https://www.ncbi.nlm.nih.gov/assembly/ GCF_000001515.7/)
- ftp.ncbi.nih.gov/genomes/Gorilla_gorilla /Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Pongo_abelii/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Papio_anubis/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Pan_paniscus/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Macaca_fascicularis/ Assembled_chromosomes/seq
- ftp.ncbi.nih.gov/genomes/Macaca_mulatta/ Assembled_chromosomes/seq/
- https://www.ncbi.nlm.nih.gov/assembly/ GCF_000001405.26/ (Human hg38)
- Janet Kelso group from Max Planck institute, Department of Evolutionary Genetics in Leipzig, provided DNA sequence for Neanderthal genome. Altai_Neandertal_ERP002097.
- C835.HCC1143_BL.4.bam (https://portal.gdc.cancer.gov/legacy-archive/files/7bf5df2d-0859-43bd-8d2f-b0d584d2bf47).
- C836.Hs_606.T.1.bam (https://portal.gdc.cancer.gov/legacy-archive/files/49a2f622-ee79-43fd-a593-0458a908acc7).
- C836.Hs_343.T.1.bam (https://portal.gdc.cancer.gov/legacy-archive/files/45dd789e-943f-4b25-b4b2-defbdf951ee1).

For our analysis purpose, we first download fasta sequence that we use for GRM analysis for obtaining global repeat map – GRM diagram that show us repeat unit lengths and possible HOR structures (Fig.2). For investigation of alpha satellite sequences, we additionally use AlphaSub algorithm. For breast cancer genomes, we first download raw sequence in BAM format after which we construct

"TGAGAAACTGCTTTGTGATGTGTGCATT" kao idealna "ključna riječ" kao i njen obrnuti komplement te u kombinaciji s Levenshteinovim algoritmom udaljenosti identificiramo položaj alfa satelita ~171 bp u sekvenci DNK kromosoma.

Ovu smo metodu također koristili za analizu ljudskog kromosoma Y kao i cijelih genoma čovjeka, neandertalca i čimpanze. Podaci za te genome i njihove alfa satelitske nizove mogu se pronaći na http://genom.hazu.hr. Ovaj alat može se preuzeti s poveznice http://genom.hazu.hr/tools.html.

2.1 Podatci

Podaci korišteni za naše istraživanje većinom su preuzeti s ftp stranice NCBI baze podataka kao jedne od najkorištenijih baza podataka u ovom području istraživanja – NCBI (National Center for Biotechnology Information, https://ftp.ncbi.nlm.nih.gov) za genome čovjeka i primata. Također, dajemo poveznice na Nacionalni institut za karcinome Genomic Data Commons Data Portal – NCI GDC portal (https://portal.gdc.cancer.gov) s bazama podataka o karcinomima. Portal NCI GDC uključuje najčešće korištene baze podataka o karcinomima kao što su TCGA (The Cancer Genome Atlas), Target (Therapeutically Applicable Research to Generate Effective Therapies) tj. u njima se nalazi oko 72 projekta s 828104 datoteka. Također uključuje podatke Projekta Enciklopedije stanične linije karcinoma (CCLE) za sekvence karcinoma dojke koje smo koristili za našu novu preliminarnu analizu prikazanu u ovom radu.

Dodatne sekvence DNK za GRM analizu kromosoma 1 (sastavljeni kromosomi) u *Pan Troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan paniscus, Macaca fascicularis, Macaca mulatta,* preuzeti su s NCBI ftp stranice:

- *Pan troglodytes* - Pan_tro_3.0 (https://www.ncbi.nlm.nih.gov/assembly/ GCF_000001515.7/)
- ftp.ncbi.nih.gov/genomes/Gorilla_gorilla /Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Pongo_abelii/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Papio_anubis/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Pan_paniscus/ Assembled_chromosomes/seq/
- ftp.ncbi.nih.gov/genomes/Macaca_fascicularis/ Assembled_chromosomes/seq
- ftp.ncbi.nih.gov/genomes/Macaca_mulatta/ Assembled_chromosomes/seq/
- https://www.ncbi.nlm.nih.gov/assembly/ GCF_000001405.26/ (Human hg38)
- Janet Kelso group from Max Planck institute, Department of Evolutionary Genetics in Leipzig, provided DNA sequence for Neanderthal genome. Altai_Neandertal_ERP002097.

fasta consensus sequence using tool UGENE (http://ugene.net) (Okonechnikov, Golosova, Fursov, & Ugene team, 2012) and its algorithms. We afterwards analyse obtained fasta sequences with GRM algorithm for making global repeat map and analyse it further with procedure explained in Fig. 2. For search of NBPF monomers in primate's genomes, we additionally use BLAST algorithm (https://blast.ncbi.nlm.nih.gov /Blast.cgi).

3 Results

Studies of human Y chromosome show that it is replete with different kinds of repeats. The largest alpha satellite higher order repeat unit 34/36mer (with variant units of ~5.7 kb and ~6.0 kb) was identified previously [Tyler-Smith & Brown, 1987, Jain & et al., 2018], but its polymorphic variants have not been studied [Tyler-Smith & Brown, 1987, Jain & et al., 2018, Tyler-Smith & et al., 1993, Rozen & et al., 2003, Perry, Tito, & Verrelli, 2007]. Using GRM and ALPHAsub algorithm, we analysed Y chromosome from hg38 human genome assembly in which most of the centromere gaps have been filled (Uralsky & et al., 2019). We identify 28 alpha satellite arrays ranging from three monomers to 1334 monomers in individual array. Average divergence of monomers in those arrays is ~24%. Those alpha satellite arrays in human Y chromosome occupy 0.71% (407986 bp). Of these alpha satellite arrays, the three arrays constitute HORs located in three domains of Y chromosome (Fig. 3).

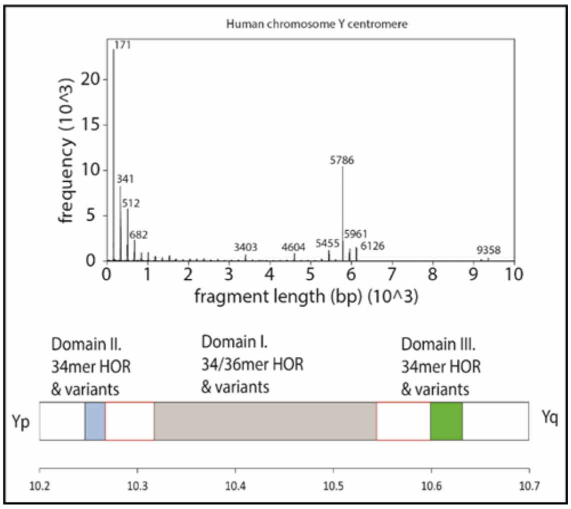


Figure 3. GRM diagram for centromere region and Y ideogram. Ideogram is modified from submitted paper (Vlahović, Glunčić, & Paar).

The novel corresponding computed GRM diagram for centromere region is presented in this paper and details of the variants are presented in (Vlahović, Glunčić, & Paar, (submitted for

- C835.HCC1143_BL.4.bam (https://portal.gdc.cancer.gov/legacy-archive/files/7bf5df2d-0859-43bd-8d2f-b0d584d2bf47).
- C836.Hs_606.T.1.bam (https://portal.gdc.cancer.gov/legacy-archive/files/49a2f622-ee79-43fd-a593-0458a908acc7).
- C836.Hs_343.T.1.bam (https://portal.gdc.cancer.gov/legacy-archive/files/45dd789e-943f-4b25-b4b2-defbdf951ee1).

U svrhu naše analize, prvo preuzimamo fasta sekvencu koju koristimo za dobivanje globalne mape ponavljanja – GRM dijagrama koji nam prikazuje duljine monomernih jedinica repeticija i moguće HOR strukture (Slika 2). Za istraživanje alfa satelitskih sekvenci dodatno koristimo AlphaSub algoritam. Za genome karcinoma dojke prvo preuzimamo "sirovu" sekvencu u BAM formatu nakon čega konstruiramo fasta konsenzusnu sekvencu pomoću alata UGENE (http://ugene.net) (Okonechnikov, Golosova, Fursov i Ugene tim, 2012) i njegovih algoritama. Nakon toga analiziramo dobivene fasta sekvence s GRM algoritmom za izradu mape globalnog ponavljanja i dalje podatke analiziramo postupkom objašnjenim na Slici 2. Za pretraživanje NBPF monomera u genomima primate dodatno koristimo BLAST algoritam (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

3 Rezultati

Studije ljudskog Y kromosoma pokazuju da je prepun različitih vrsta repeticija. Najveća ponovljena jedinica alfa satelita višeg reda 34/36mer (s varijantnim jedinicama od ~5,7 kb i ~6,0 kb) identificirana je ranije [Tyler-Smith & Brown, 1987, Jain & sur., 2018], ali njene polimorfne varijante nisu proučavane [Tyler-Smith & Brown, 1987, Jain & sur., 2018, Tyler-Smith & sur., 1993, Rozen & sur., 2003, Perry, Tito, & Verrelli, 2007]. Koristeći GRM i ALPHAsub algoritam, analizirali smo Y kromosom iz ljudskog genoma hg38 u kojem su većinom popunjene praznine centromera (Uralsky & sur., 2019). Identificirali smo 28 alfa satelitskih nizova u rasponu od 3 monomera do 1334 monomera u pojedinačnim nizovima. Prosječna divergencija monomera u tim nizovima je ~24%. Ti alfa satelitski nizovi u ljudskom Y kromosomu zauzimaju 0,71% (407986 bp). Od tih alfa satelitskih nizova, tri čine HORove smještene u tri domene na Y kromosomu (Slika 3).

publication)). These HORs are located in centromere and near centromere region (domains I., II. and III.). Domain I. constitutes the largest HOR array 34/36mer (positions 10316945-10544039 bp in the centromere). We identified polymorphic variants that have monomer duplications, deletions or insertions, or rearrangements and non-HOR insertions of up to 44-monomer length (details given in (Vlahović, Glunčić, & Paar, (submitted for publication))). These polymorphic variants are evident also from novel Needleman - Wunsch analysis of alpha satellite arrays in these three domains (dot plot in Fig 4. for domain II).

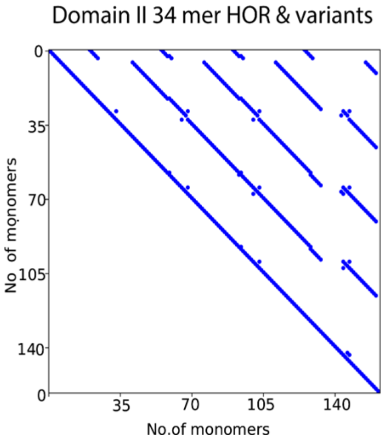
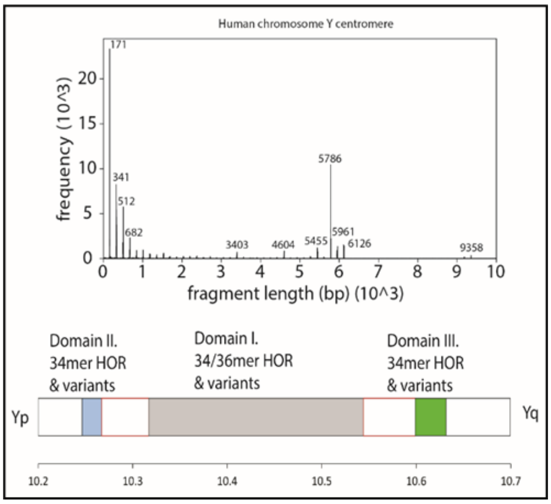


Figure 4. Needleman-Wunsch analysis of monomer divergence. Diagonals represent monomers with divergence less than 7%. Horizontal distances reveal nmer HOR copy that is in this case 34mer and its variants.

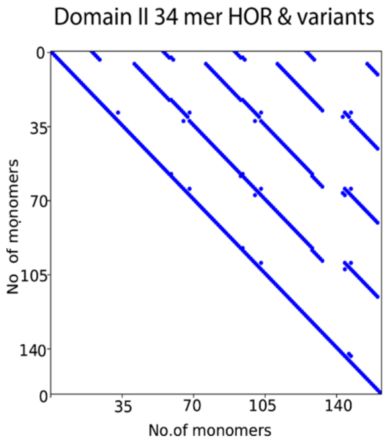
Using GRM and ALPHAsub algorithms, we analyse all alpha satellite arrays in human hg38 assembly as well as in chimpanzee and Neanderthal genomes (data stored in http://genom.hazu.hr database). These data on alpha satellite arrays and their positions within genomes provide a clue about their distribution and evolution between closely related species.

Calculation of percentage of alpha satellite content in genomes show that in chimpanzee (Pan_tro_3.0 assembly) they occupy 0.22%, in Neanderthal (Altai_Neandertal_ERP002097) 0.22% and in human genome (hg38) 2.19% of DNA sequence (Vlahović & et al., 2022). In human chromosomes 17-22 the percentage of alpha satellites is higher, ~5%. In our previous publication (Glunčić, Vlahović, & Paar, 2019) we analysed human chromosome 21 and we discovered the largest alpha satellite HOR unit among all somatic chromosomes - 33mer HOR, and that this chromosome is also replete with numerous other HOR arrays with monomer units composed of eight and more monomer types (Fig. 5). These HORs, based on alpha satellite arrays, could be used as unique chromosome 21 probe in molecular



Slika 3. GRM dijagram za područje centromere i Y ideogram. Ideogram je modificiran prema (Vlahović, Glunčić, & Paar, 2011, predano za objavu).

Novi odgovarajući izračunati GRM dijagram za područje centromera predstavljen je u ovom radu, a detalji varijanti prikazani su u (Vlahović, Glunčić, & Paar, (predano za objavu)). Ovi HOR-ovi se nalaze u području centromere i blizu centromere (domene I., II. i III. Domena I. čini najveći HOR niz 34/36mer (pozicije 10316945-10544039 bp u centromeri). Identificirali smo polimorfne varijante koje imaju duplikacije monomera, delecije ili insercije, ili preraspodjele te insercije koje ne pripadaju repeticijskim elementima HOR-a do duljine od 44 monomera (pojedinosti su dane u (Vlahović, Glunčić, & Paar, (predano za objavu))). Ove polimorfne varijacije također su vidljive iz Needleman-Wunsch analize alfa satelitskih nizova u ove tri domene (dot plot dijagram na slici 4. za domenu II).



Slika 4. Needleman-Wunsch analiza divergencije monomera. Dijagonale predstavljaju monomere s divergencijom manjom od 7%. Horizontalne udaljenosti otkrivanju nmer HOR kopije koje su u ovom slučaju 34 mer HOR te njegove varijante.

cytogenetics for diagnostics purposes (Glunčić, Vlahović, & Paar, 2019).

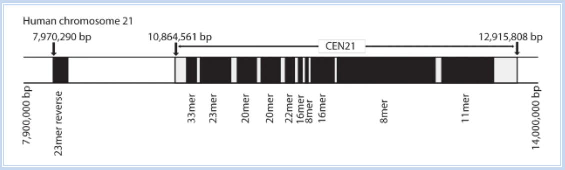


Figure 5. Positions of alpha satellites and nmer HORs in centromere of human chromosome 21 obtained with ALPHAsub algorithm (Glunčić, Vlahović, & Paar, 2019). In centromere region, there are arrays of 33mer, 23mer, 20mer, 20mer, 22mer, 16mer, 8mer, 16mer, 8mer, 11mer HORs. Outside of centromere region in chromosome 21 there is one array of 23mer HOR in its reverse form. HOR arrays of alpha satellites outside of centromere region in human genome are not common.

Global Repeat Map algorithm is well suited for the study of other tandem repeats and HOR identification in genomes. One of our results is identification of novel HORs in human and chimpanzee and other primate's chromosome 1. Of particular interest is our discovery of HORs, which are embedded in Neuroblastoma Breakpoint Family Genes and in hornerin gene (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011). Recently Romero et al. (Romero & al, 2018) confirmed our previous discovery (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011) of 1410-bp quartic HOR repeat in human hornerin gene with phylogenetic analysis based on monomer unit of ~39 bp (Fig. 6.).

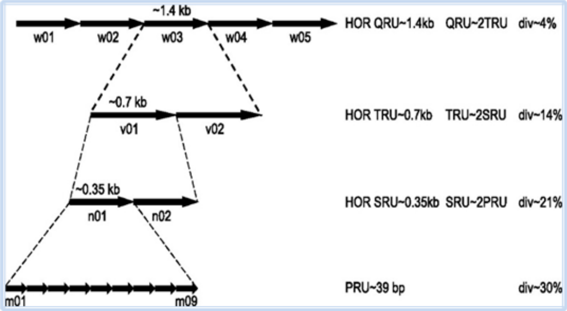
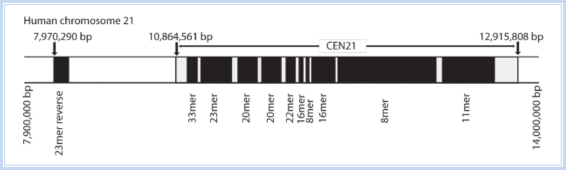

~1.4 kb
w01
w02
w03
w04
w05
HOR QRU~1.4kb QRU~2TRU div~4%
~0.7 kb
v01
v02
HOR TRU~0.7kb TRU~2SRU div~14%
~0.35 kb
n01
n02
HOR SRU~0.35kb SRU~2PRU div~21%
m01
m09
PRU~39 bp
div~30%

Figure 6. Scheme of 1410 bp HOR in human chromosome 1. Figure is taken from (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011).

Hierarchical structure of 1410-bp quartic HOR is composed of nine 39-bp primary repeat units (divergence ~30%) that form a ~0.35-kb secondary repeat unit. The two ~0.35-kb secondary repeat units (divergence ~21%) form a ~0.7-kb tertiary repeat unit, two ~0.7-kb tertiary repeat units (divergence ~14%) form ~1.4-kb quartic HOR repeat unit. HOR copies average divergence is ~4% (Paar, Glunčić,

Koristeći algoritme GRM i ALPHAsub, analizirali smo sve alfa satelitske nizove u ljudskom hg38, kao i u genomima čimpanze i neandertalca (podaci pohranjeni u bazi podataka http://genom.hazu.hr). Ovi podaci o alfa satelitskim nizovima i njihovim položajima unutar genoma daju trag o njihovoj distribuciji i evoluciji između blisko povezanih vrsta.

Izračun postotka sadržaja alfa satelita u genomima pokazuje da kod čimpanze (Pan_tro_3.0) oni zauzimaju 0,22%, kod neandertalca (Altai_Neandertal_ERP002097) 0,22%, a kod genoma čovjeka (hg38) 2,19% sekvence DNK (Vlahović & sur., 2022). U ljudskim kromosomima 17-22 postotak alfa satelita je veći, ~5%. U našoj prethodnoj publikaciji (Glunčić, Vlahović & Paar, 2019) analizirali smo ljudski kromosom 21 i otkrili u njemu najveću alfa satelitsku HOR jedinicu među svim somatskim kromosomima - 33mer HOR, te da je taj kromosom također prepun brojnih HOR jedinica sastavljenih od osam i više monomerskih tipova (Slika 5). Ovi HOR-ovi temeljeni na alfa satelitskim nizovima mogu se koristiti kao jedinstveni identifikator kromosoma 21 u molekularnoj citogenetici za potrebe dijagnostike (Glunčić, Vlahović & Paar, 2019).



Slika 5. Položaji alfa satelita u ljudskom kromosomu 21 dobiveni su algoritmom ALPHAsub (Glunčić, Vlahović & Paar, 2019). U području centromera postoje nizovi 33mer, 23mer, 20mer, 20mer, 22mer, 16mer, 8mer, 16mer, 8mer, 11mer HOR-ova. Izvan područja centromere u kromosomu 21 postoji jedan niz 23mer HOR u svom obrnutom obliku. HOR nizovi alfa satelita izvan područja centromera u ljudskom genomu nisu uobičajeni.

Algoritam Global Repeat Map prikladan je za proučavanje drugih tandemskih ponavljanja (osim alfa satelita) i identifikaciju HORova u genomima. Jedan od naših rezultata je identifikacija novih HORova u kromosomu 1 čovjeka i čimpanze te ostalih primata. Od posebnog je interesa naše otkriće HOR-ova koji su ugrađeni u gene obitelji neuroblastoma (Neuroblastom Breakpoint Family) i u genu hornerin (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011). Nedavno su Romero i sur. (Romero & sur., 2018) potvrdili naše prethodno otkriće (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011) kvartičnog HOR ponavljanja od 1410 bp u ljudskom hornerin genu filogenetskom analizom temeljenom na monomernoj jedinici od ~39 bp (Slika 6.)

Rosandić, Basar, & Vlahović, 2011). Our explanation of the hornerin gene HOR structure was accompanied by our initial discovery of the NBPF HORs (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011). We hypothesized (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011) that the NBPF HORs have important roles in human brain development and cognitive abilities. Later research by other scientists has shown the role of NBPF monomers (also called DUF1220/ Olduvai domain) are related to neurodegenerative diseases schizophrenia, autism, microcephaly, macrocephaly (O'Bleness & et al., 2014) and neuroblastoma (Andries & et al., 2015).

Our novel results for analysed sequences of several primates (*Pan Troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan paniscus, Macaca fascicularis, Macaca mulatta*), as well as for the human and Neanderthal chromosome 1 show peaks for the ~1530 bp NBPF monomer units, ~3180 bp dimer units and for the ~4770 bp 3mer HOR NBPF units in our GRM diagrams. Fig.7 shows the most significant peaks are for human and Neanderthal chromosome 1 and that the corresponding weak signal is observed for Gorilla chromosome 1.

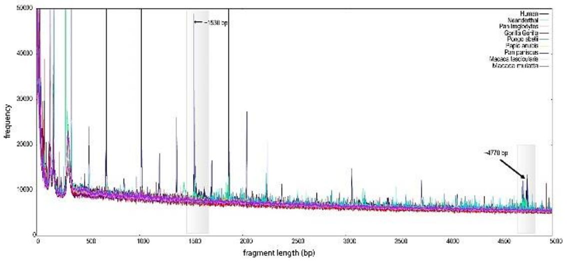
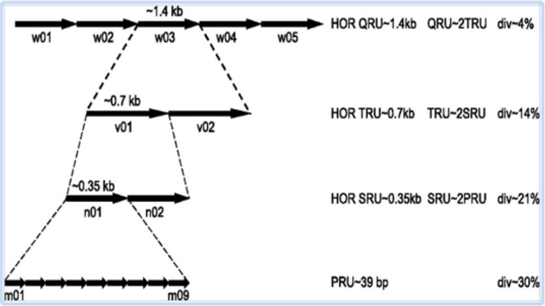


Figure 7. GRM diagrams of chromosome 1 from human, Neanderthal, Pan troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan Paniscus, Macaca fascicularis and Macaca mulatta with pronounced peaks at ~1530 bp (monomer unit) and ~4770 bp (NBPF HOR).

Additionally, using BLAST algorithm we investigated the occurrence of human and Neanderthal monomer units in three GRM length ranges, <2500 bp, 2500 - 4000 bp and >4000 bp, to identify NBPF monomers, dimers and trimers that we identify by GRM analysis in different primates (Table 1). This novel results in Table 1 shows that most copies of those triplets are found in Gorilla genome that represent almost full triplet HOR copies. From Table 1 we can see that for genome assemblies used for this analysis, Gorilla gorilla genome have seven monomer copies with identity of ~82% when compared to human and Neanderthal NBPF consensus monomer sequences. There are 5 dimers when compared to human NBPF dimers but only 2 when compared to Neanderthal NBPF dimer sequence with identity of ~81% for human NBPF dimer and 80% for Neanderthal dimer. Interestingly



~1.4 kb
w01
w02
w03
w04
w05
HOR QRU~1.4kb QRU~2TRU div~4%
~0.7 kb
v01
v02
HOR TRU~0.7kb TRU~2SRU div~14%
~0.35 kb
n01
n02
HOR SRU~0.35kb SRU~2PRU div~21%
m01
m09
PRU~39 bp
div~30%

Slika 6. Shema 1410 bp HOR u ljudskom kromosomu 1. Slika je preuzeta iz (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011).

Hijerarhijska struktura kvartičnog HOR-a od 1410 bp sastoji se od devet primarnih repeticijskih jedinica od 39 bp (divergencija ~30%) koje tvore sekundarnu repeticijsku jedinicu od ~0,35 kb. Dvije sekundarne repeticijske jedinice od ~0,35 kb (divergencija ~21%) tvore tercijarnu repeticijsku jedinicu od ~0,7 kb, dvije tercijarne repeticije od ~0,7 kb (divergencija ~14%) tvore kvartičnu HOR jedinicu od ~1,4 kb. Prosječna divergencija HOR kopija je ~4% (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011). Naše objašnjenje strukture HOR gena hornerina popraćeno je i našim prvim otkrićem NBPF HORova (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011). Postavili smo hipotezu (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011) da NBPF HORovi imaju važnu ulogu u razvoju ljudskog mozga i kognitivnih sposobnosti. Kasnija istraživanja drugih znanstvenika pokazala su da uloga monomera NBPF gena (također nazvanih DUF1220/ Olduvai domene) su povezana s neurodegenerativnim bolestima kao što su shizofrenija, autizam, mikrocefalija, makrocefalija (O'Bleness, & sur., 2014) te neuroblastomom (Andries & sur., 2015).

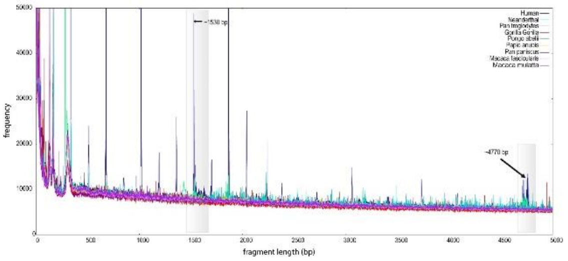
Naši novi rezultati za analizirane sekvence nekoliko primata *(Pan Troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan paniscus, Macaca fascicularis, Macaca mulatta*), kao i za ljudski i neandertalski kromosome 1 pokazuju pikove za ~1530 bp NBPF monomernu jedinicu, ~3180 bp dimer jedinicu i za ~ 4770 bp 3mer HOR NBPF jedinicu na GRM dijagramu. Slika 7 prikazuje najznačajnije vrijednosti pikova za kromosom 1 čovjeka i neandertalaca, te odgovarajući slabi signal u kromosomu 1 gorile.

we identify nine NBPF trimers with ~81% identity for human NBPF 3mer HOR copy and nine trimers for Neanderthal NBPF 3mer HOR copy in *Gorilla gorilla* chromosome 1. Interestingly *Pan Troglodytes* (chimpanzee) have much less identified monomer, dimer and trimer units when compared to human and Neanderthal consensus monomers from NBPF 3mer HORs. In addition, we can see from Table 1 that some primates have greater number of monomers, dimers and trimers when compared to Neanderthal NBPF consensus sequences than to human NBPF monomer, dimer and trimer units. Examples are *Pongo abelii* with six trimer units when compared to Neanderthal vs five trimers compared to human, *Pan Troglodytes* have five dimers compared to Neanderthal vs 4 dimers when compared to human NBPF consensus sequences. *Macaca Fascularis* have two monomers when compared to Neanderthal vs one monomer when compared to human NBPF monomer consensus sequence. We have to emphasize that this big difference can be due to the percentage of completeness of primate genomes assemblies so this result is not finite – new assemblies with more complete genomes will give more detailed picture of NBPF monomer like copy numbers and HOR structures in primates.

Table 1. BLAST hit results for comparison of human and Neanderthal HOR consensuses with Pan Troglodytes, Gorilla Gorilla, Pongo abelii, Papio Anubis, Macaca fascicularis, Macaca mulatta and Pan Paniscus.

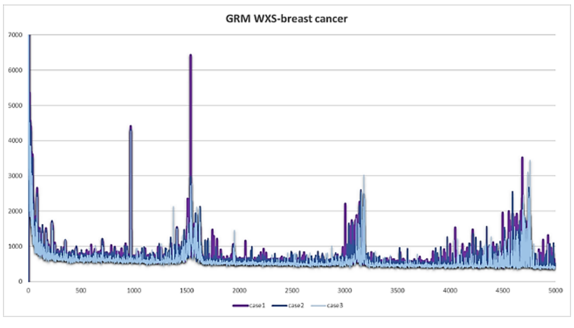
Species	**<2500 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	3	81.57	2	81.26
Pan paniscus	1	81.56	1	79.82
Gorilla gorilla	7	82.29	7	81.14
Pongo abelii	4	81.40	3	80.80
Papio Anubis	2	79.29	2	78.03
Macaca Fascularis	1	79.56	2	79.56
Macaca mulatta	1	79.58	1	78.76
Species	**2500 - 4000 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	4	80.44	5	78.37
Pan paniscus	1	78.93	1	78.60
Gorilla gorilla	5	81.81	2	80.10
Pongo abelii	3	79.84	1	78.92
Papio Anubis	3	80.30	2	79.23
Macaca Fascularis	1	80.85	1	80.79
Macaca mulatta	1	80.51	1	78.60
Species	**>4000 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	3	80.29	3	81.77
Pan paniscus	2	80.25	2	81.22
Gorilla gorilla	9	81.24	9	81.76
Pongo abelii	5	81.54	6	82.07
Papio Anubis	1	77.44	1	77.81
Macaca Fascularis	2	79.21	2	79.59
Macaca mulatta	2	79.19	2	80.40

Another direction of our research is identification of structural variants in cancer genomes related with tandem repeats. Our preliminary novel study of the



Slika 7. GRM dijagrama za kromosom 1 čovjeka, neandertalca, Pan troglodytes, Gorilla gorilla, Pongo abelii, Papio anubis, Pan Paniscus, Macaca fascicularis i Macaca mulatta s naglašenim pikovima na ~1530 bp (monomerna jedinica) te ~4770 bp (NBPF HOR jedinica).

Dodatno, pomoću BLAST algoritma istražili smo pojavu ljudskih i neandertalskih monomernih jedinica za ove vrste u tri GRM raspona duljina, <2500 bp, 2500 - 4000 bp i >4000 bp, kako bismo identificirali NBPF monomere, dimere i trimere koje smo pronašli s GRM metodom za različite primate (Tablica 1). Ova tablica pokazuje naše nove rezultate da se većina kopija tih tripleta nalaze u genomu gorile koji predstavljaju gotovo pune triplet NBPF HOR kopije. Iz Tablice 1 možemo vidjeti da za sklopove genoma sekvenci korištene za ovu analizu, genom Gorilla gorilla ima sedam monomernih kopija s identičnošću od ~82% u usporedbi s ljudskim i neandertalskim NBPF konsenzusnim sekvencama monomera. U njemu postoji 5 dimera u usporedbi s ljudskim NBPF dimerima, ali samo 2 u usporedbi s neandertalskim NBPF dimerskom sekvencom s identičnošću od ~81% za ljudski NBPF dimer i 80% za neandertalski dimer. Zanimljivo je da smo identificirali devet NBPF trimera s ~81% identičnosti za ljudsku NBPF 3mer HOR kopiju i devet trimera za neandertalsku NBPF 3mer HOR kopiju u *Gorila gorila* kromosomu 1. Zanimljivo je da *Pan Troglodytes* (čimpanze) imaju puno manje identificiranih monomernih, dimernih i trimernih jedinica u usporedbi s ljudskim i neandertalskim konsenzusnim monomerima iz NBPF 3mer HORova. Osim toga, iz Tablice 1 možemo vidjeti da neki primati imaju veći broj monomera, dimera i trimera u usporedbi s neandertalskim NBPF konsenzusnim sekvencama u odnosu na ljudske NBPF monomerne, dimerne i trimerne jedinice. Primjeri su *Pongo abelii* sa šest trimernih jedinica u usporedbi s neandertalcem nasuprot pet trimera u usporedbi s čovjekom, *Pan Troglodytes* imaju pet dimera u usporedbi s neandertalcem nasuprot 4 dimera u usporedbi s ljudskim NBPF konsenzusnim sekvencama. *Macaca Fascularis* ima dva monomera u usporedbi s neandertalcem naspram jednog monomera u usporedbi s konsenzusnom sekvencom humanog NBPF monomera. Moramo naglasiti da ova velika razlika može biti posljedica postotka potpunosti sklopova genomske sekvence primata tako da ovaj rezultat nije konačan – novi sklopovi whole exome sequence for three cases of breast cancer from CCLE project indicate that we could, on more extensive data sets, apply those methods for classification of DNA sequences in patients. This usage could be of interest in diagnostic and drug appliances (Fig.8). Fig. 8 shows three breast cancer cases of chromosome 1 and theirs associated GRM diagrams obtained for exome sequences. Novel results show that for even those three cases existence of structural variants i.e., copy number variants (CNVs) in coding part of chromosome 1 is identified. This conclusion comes from peak differences in frequency of individual fragment lengths (that represent monomer, dimer and trimer units) for those cancer cases. Examples of differences are seen, interestingly, in the area of fragment lengths of ~1530 bp, ~3180 bp and ~4770 bp that correspond to NBPF monomer, dimer and trimer units like for NBPF 3mer HOR GRM signals in Fig. 7 (although with lower frequency). The difference between GRM diagrams for human shown on Fig. 7 and GRM diagram for three exome cancer cases (Fig.8) is wider scattering around those fragment lengths that characterize NBPF peaks in cancer cases compared to hg38 human assembly (Fig.7). Precisely that scattering is an indicator of structural variants, partial deletions/insertions or copy number variants, so detailed analysis of that part of the sequence and construction of HOR variant schemes (like on Fig.2. for HOR copy construction) is needed as well as analysis of greater number of cases for obtaining exact structural variants and CNVs for classification purposes. It is not coincidence that NBPF signals for monomer, dimer and trimer units are most pronounced signals on GRM diagram. If we take in consideration that length of NBPF genes (in form of tandem repeats that we identify – NBPF20, NBPF10, NBPF12, NBPF14 and NBPF19 of 351410 bp) occupy 0.14% of chromosome 1 overall length (248956422 bp) while in coding part of chromosome 1 this percentage is 1.36% of overall 10.40% of coding part in whole chromosome 1. For other peaks to distinguish between those cancer cases we will have to use GRM algorithm ability of magnifying glass (zoom in) to smaller ranges of fragment lengths and frequency on diagram (Glunčić & Paar, 2013).



genomskih sekvenci s potpunijim genomima dat će detaljniju sliku NBPF monomera poput broja kopija i HOR struktura u primata.

Tablica 1. Rezultati BLAST pogodaka za usporedbu ljudskih i neandertalskih koncenzusnih HOR sekvenci s primatima Pan Troglodytes, Gorilla Gorilla, Pongo abelii, Papio Anubis, Macaca fascicularis, Macaca mulatta and Pan Paniscus.

Species	**<2500 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	3	81.57	2	81.26
Pan paniscus	1	81.56	1	79.82
Gorilla gorilla	7	82.29	7	81.14
Pongo abelii	4	81.40	3	80.80
Papio Anubis	2	79.29	2	78.03
Macaca Fascularis	1	79.56	2	79.56
Macaca mulatta	1	79.58	1	78.76
Species	**2500 - 4000 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	4	80.44	5	78.37
Pan paniscus	1	78.93	1	78.60
Gorilla gorilla	5	81.81	2	80.10
Pongo abelii	3	79.84	1	78.92
Papio Anubis	3	80.30	2	79.23
Macaca Fascularis	1	80.85	1	80.79
Macaca mulatta	1	80.51	1	78.60
Species	**>4000 bp**			
	Human		**Neanderthal**	
	No.of copies	% identity	No.of copies	%identity
Pan troglodytes	3	80.29	3	81.77
Pan paniscus	2	80.25	2	81.22
Gorilla gorilla	9	81.24	9	81.76
Pongo abelii	5	81.54	6	82.07
Papio Anubis	1	77.44	1	77.81
Macaca Fascularis	2	79.21	2	79.59
Macaca mulatta	2	79.19	2	80.40

Drugi smjer našeg istraživanja je identifikacija strukturnih varijanti u genomima raka povezanih s tandemskim ponavljanjima. Naša preliminarna nova istraživanja cijele sekvence egzoma za tri slučaja karcinoma dojke iz projekta CCLE ukazuje da bismo mogli, na opsežnijim skupovima podataka, primijeniti te metode za klasifikaciju sekvenci DNK kod pacijenata. Ova bi uporaba mogla biti od interesa za dijagnostiku i pametne lijekove (Slika 8). Slika 8 prikazuje tri slučaja raka dojke kromosoma 1 i njihove povezane GRM dijagrame dobivene za sekvence egzoma. Novi rezultati pokazuju da čak i za ta tri slučaja postoji indikacija za strukturne varijante, tj. varijante broja kopija (CNV) u kodirajućem dijelu kromosoma 1. Ovaj zaključak proizlazi iz razlika pikova u frekvenciji individualnih fragmentnih duljina (koje predstavljaju monomerne, dimerne i trimerne jedinice) za ove slučajeve karcinoma. Primjeri ovih razlika se vide, zanimljivo, u području fragmentnih duljina od ~1530 bp, ~3180 bp i ~4770 bp koje odgovaraju signalima NBPF monomernih, dimernih i trimernih jedinica sa Slike 7 (iako nižih frekvencija pošto se radi o kodirajućem dijelu kromosoma 1). Razlike između dijagrama prikazanog za čovjekov kromosom 1 i GRM dijagrama za 3 slučaja raka dojke i njihovog exoma (Slika 8) je šire raspršenje oko fragmentnih duljina

Figure 8. GRM diagram for three cases of cancer exome data for chromosome 1 downloaded from NCI GDC database, CCLE project. From these diagrams, we can conclude that structural variants, CNVs and partial HOR copies exists based on scattering around peaks in GRM diagram for NBPF 3mer HOR, corresponding to peak size in frequency (y-axis) at specific fragment lengths (x-axis).

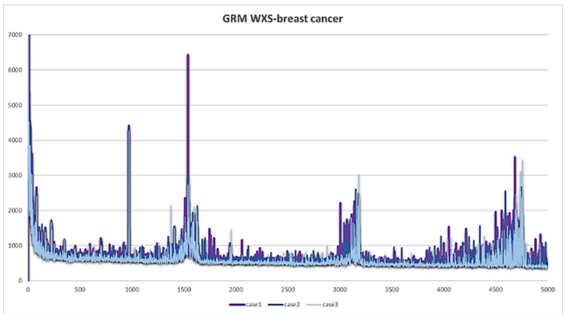
4 Conclusion

A combination of novel methods for more precise genome assemblies involving combination of short-read, long-read and optical mapping (Weissensteiner, Pang, AWC, Bunikis, & et al., 2017) can cover large-scale tandem repeat arrays to avoid assembly errors [Tørresen, et al., 2019, Levy-Sakin & et al., 2019] especially in complex chromosome regions such as centromere. That could help in diagnostics as well as in search of new drugs and therapies, in combination with growing field of AI and machine learning methods for pattern recognition in bioinformatics.

Our GRM tool gives an advantage and opportunity to identify HORs in genomes of various species. With new sequencing technologies, our bioinformatics algorithms can provide even greater insight into tandem and HOR structure variants and CNVs in more precisely assembled genomes. Such a case is seen in hornerin gene HOR in human chromosome 1 for which we proposed a structure in ref. (Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011), and later Romero found similar structure in other primates, using novel genome assembly (Romero & et al., 2018). Analysing human hg37 and hg38 genomes and chimpanzee genomes, we obtained novel results for HOR variants in human Y chromosome (Vlahović, Glunčić, & Paar, (submitted for publication)) with respect to earlier results for previous incomplete genomic assemblies (Paar & et al., 2011). Of particular interest for evolution of cognitive abilities and of neurodegenerative diseases are our discoveries of NBPF 3mer HORs in human and Neanderthal genomes and their absence/low occurrence in other primates [Paar, Glunčić, Rosandić, Basar, & Vlahović, 2011, Vlahović, Glunčić, Rosandić, & Paar, submitted for publication] probably due to incomplete primate genomes used for research. Example of assembly difference is for human hg37 that have in its chromosome 1 assembly 165 monomers of NBPF monomer unit length, organized in 57 HOR copies, while in hg38 there are 177 monomer units organized in 65 HOR copies. In hg38 chromosome Y sequence there are 51 HOR copies based on alpha satellite monomer unit with 1616 monomers, while in hg37 there are 11 HOR copies with 329 monomers, but bigger difference with newer assemblies of sequenced genome is changed HOR scheme for this chromosome. In hg37 it was 45mer HOR form of the

koje karakteriziraju NBPF pikove kod slučaja karcinoma u usporedbi s hg38 ljudskim sklopom DNK sekvence kromosoma 1 (Slika 7).

Upravo je to raspršenje indikator za postojanje strukturalnih varijanti, djelomičnih delecija/insercija ili varijanti broja kopija, stoga je detaljna analiza tog dijela sekvence potreba i konstrukcija shema HOR varijanti (kao što je prikazano na Slici 2 za konstrukciju HOR kopija) te je potrebna analiza većeg broja slučajeva za dobivanje egzaktnih strukturalnih varijanti i CNV-ova za potrebe klasifikacije. Nije slučajno da su NBPF signali za monomerne, dimerne i trimerne jedinice najizraženiji signali na GRM dijagramu. Ako uzmemo u obzir da duljina NBPF gena (u obliku tandemskih ponavljanja koje smo identificirali – NBPF20, NBPF10, NBPF12, NBPF14 i NBPF19 od 351410 bp) zauzima 0.14% ukupne duljine kromosoma 1 (248956422 bp) te u kodirajućem dijelu kromosoma 1 ovaj postotak iznosi 1.36% od ukupnih 10.40% kodirajućeg dijela cijelog kromosoma 1. Za ostale pikove na GRM dijagramu, da bismo ih razlikovali u ovim slučajevima karcinoma moramo upotrijebiti sposobnost GRM algoritma - povećalo (zoom in) za uvećanje dijagrama na manje raspone fragmentnih duljina i frekvencije na dijagramu (Glunčić & Paar, 2013).



Slika 8. GRM dijagram za tri slučaja egzoma karcinoma za kromosom 1 preuzet iz NCI GDC baze podataka, s CCLE projekta. Iz ovih dijagrama možemo zaključiti da strukturne varijante, CNV i djelomične HOR kopije postoje na temelju raspršenja oko pikova za NBPF 3mer HOR, što odgovara veličini pika frekvencije (y-os) na specifičnim fragmentnim duljinama (x-os).

4 Zaključak

Kombinacija novih metoda za preciznije slaganje genoma koje uključuju kombinacije kratkog čitanja, dugog čitanja te optičkog mapiranja (Weissensteiner, Pang, AWC, Bunikis & sur., 2017) mogu pokriti nizove tandemnih repeticija velikih razmjera kako bi se izbjegle pogreške u slaganju genoma [Tørresen, & sur., 2019, Levy-Sakin, & sur., 2019], posebno u područjima kromosoma s velikim nizovima repeticija kao što je centromera. To bi moglo pomoći scheme for Y chromosome, while within hg38 the HOR scheme is in 34/36mer HOR form.

Our future work in this field will be oriented to analysis of complex tandem repeats and HORs in eukaryotes, both plants and mammals, as well as in study of human genomes related to cancer (breast cancer, colorectal cancer) and neurodegenerative diseases (autism spectrum disorder) for which genomic data are available. In addition, we will use our tools to investigate population genomics/genetics of closely relates species such as humans and primates and theirs corresponding structural variants, CNVs in HOR structures. For those studies, we will make exact variant schemes in HOR structures that could be used for comparison and classification of available large data sets collected through different sequencing projects with machine learning and AI algorithms.

Acknowledgments

This work was supported by IP-2019-04-2757 from Croatian Science Foundation.

References

A.K., S., Mourad, M., Kaplan, M., & al., e. (2019). The Genomic Landscape of Centromeres in Cancers. *9*, str. 11259. doi: https://doi.org/10.1038/s41598-019-47757-6

Andries, V., Vandepoele, K., Staes, K., Berx, G., Bogaert, P., Isterdael, G. V., . . . Roy, F. v. (2015). NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer, 10*. doi:doi: 10.1186/s12885-015-1408-5

Black, E., & Giunta, S. (2018). Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes, 9*, str. 615. doi:doi:10.3390/genes9120615

Flèche, P. (2001). tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *MC Microbiol, 1*. doi:doi:10.1186/1471-2180-1-2.

Fondon, J., & Garner, H. (2004). Molecular origins of rapid and continuous morphological evolution. *PNAS, 101*, str. 18058–63.

Glunčić, M., & Paar, V. (2013). Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acid Research, 41*, str. e17. doi:doi:10.1093/nar/gks721

Glunčić, M., Vlahović, I., & Paar, V. (2019). Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit

u dijagnostici, kao i u potrazi za novim lijekovima i terapijama, u kombinaciji s rastućim područjem umjetne inteligencije i metoda strojnog učenja za prepoznavanje uzoraka u području bioinformatike.

Naš GRM algoritam daje prednost i priliku za identifikaciju HORova u genomima različitih vrsta. S novim tehnologijama sekvenciranja, naši bioinformatički algoritmi mogu pružiti još bolji uvid u tandemne varijacije i HOR strukture u preciznije sastavljenim genomima. Takav se slučaj vidi u HORu hornerin gena u ljudskom kromosomu 1 za koji smo predložili strukturu u ref. (Paar, Glunčić, Rosandić, Basar & Vlahović, 2011), a kasnije su i Romero & sur. pronašli sličnu strukturu u drugih primata, koristeći nove sekvence genoma (Romero et al, 2018). Analizirajući ljudske hg37 i hg38 sklopove genomskih sekvenci i genom čimpanze, dobili smo nove rezultate za HOR varijante u ljudskom Y kromosomu (Vlahović, Glunčić i Paar, predano za objavu) u odnosu na ranije rezultate za prethodne nepotpune sklopove genomskih sekvenci (Paar & sur., 2011). Od posebnog interesa za evoluciju kognitivnih sposobnosti i neurodegenerativnih bolesti su naša otkrića NBPF 3mer HORa u genomima čovjeka i neandertalca te njihova odsutnost/mala pojavnost kod drugih primata [Paar, Glunčić, Rosandić, Basar & Vlahović, 2011, Vlahović, Glunčić, Rosandić & Paar, predano na objavu] vjerojatno zbog nepotpunih genoma primata korištenih za istraživanje. Primjer razlike sklopa genomske sekvence je za ljudski hg37 koji u svom sklopu DNK sekvence kromosoma 1 ima 165 monomera duljine NBPF monomerne jedinice, organiziranih u 57 HOR kopija, dok u hg38 postoji 177 monomernih jedinica organiziranih u 65 HOR kopija. U sekvenci kromosoma Y iz hg38 postoji 51 HOR kopija bazirana na alfa satelitskoj monomernoj jedinici sa 1616 monomera, dok u hg37 postoji 11 HOR kopija sa 329 monomera, ali je veća razlika u odnosu na novije sklopove sekvenciranog genoma promijenjena HOR shema za ovaj kromosom. U hg37 to je bio 45mer-ni HOR oblik sheme za Y kromosom, dok je unutar hg38 HOR shema u 34/36mer HOR obliku.

Naš budući rad u ovom području bit će usmjeren na analizu složenih tandemskih ponavljanja i HORova u eukariota, kako biljaka tako i sisavaca, kao i na proučavanje ljudskih genoma povezanih s karcinomima (karcinom dojke, kolorektalni karcinom) i neurodegenerativnim bolestima (poremećaj iz autističnog spektra) za koje su dostupni genomski podaci. Osim toga, koristit ćemo naše alate za istraživanje populacijske genomike/genetike blisko srodnih vrsta kao što su ljudi i primati i njihovih odgovarajućih strukturnih varijanti, CNVova u HOR strukturama. Za te ćemo studije izraditi točne sheme varijanti HOR struktura koje bi se mogle koristiti za usporedbu i klasifikaciju dostupnih velikih skupova podataka prikupljenih

among all human somatic chromosomes. *Sci Rep*. doi:doi:10.1038/s41598-019-49022-2

Jain, M., Olsen, H., Turner, D., Stoddart, D., Bulazel, K., Paten, B., & et al. (2018). Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol., 4*, str. 321-3. doi:doi: 10.1038/nbt.4109

Levy-Sakin, M., & al., e. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications, 10*, str. 1025. doi:https://doi.org/10.1038/s41467-019-08992-7

Levy-Sakin, M., Pastor, S., Mostovoy, Y., & al, e. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat Commun, 10*, str. 1025. doi:https://doi.org/10.1038/s41467-019-08992-7

Okonechnikov K, Golosova O, Fursov M & the GENE team. Unipro UGENE: a unified bioinformatics toolkit . Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091

O'Bleness, M., Searles, V. B., Dickens, C. M., Astling, D., Albracht, D., Mak, A. C., . . . Sikela, J. M. (2014). Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics, 387*. doi:https://doi.org/10.1186/1471-2164-15-387

Paar, V., Glunčić, M., Rosandić, M., Basar, I., & Vlahović, I. (2011). Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees. *Molecular Biology and Evolution, 28*, str. 1877–1892. doi:https://doi.org/10.1093/molbev/msr009

Paar, V., x Glunčić, V., Basar, I., Rosandic, M., Paar, P., & Cvitković, M. (2011). Large Tandem, Higher Order Repeats and Regularly Dispersed Repeat Units Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes. *Journal of Molecular Evolution, 1*, str. 34-55.

Pathak, D., & Ali, S. (2012). Repetitive DNA: A Tool to Explore Animal Genomes/Transcriptomes, Functional Genomics. (D. G. Meroni, Ur.) *InTech*. doi:DOI: 10.5772/48259

Perry, G., Tito, R., & Verrelli, B. (2007). The evolutionary history of human and chimpanzee Y-chromosome gene loss. *Mol Biol Evol., 3*, str. 853-9. doi:doi: 10.1093/molbev/msm002

Rigden, D. J., & Fernández, X. M. (2022). The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic acids research, 50(D1), D1–D10. https://doi.org/10.1093/nar/gkab1195.

Romero, V., & al, e. (2018). High Order Formation and Evolution of Hornerin in Primates. *Genome Biology and Evolution, 10*, str. 3167–3175.

kroz različite projekte sekvenciranja s algoritmima strojnog učenja i umjetne inteligencije.

Zahvale

Ovaj rad je izrađen pod projektom IP-2019-04-2757 od Hrvatske zaklade za znanost.

Reference

A.K., S., Mourad, M., Kaplan, M., & al., e. (2019). The Genomic Landscape of Centromeres in Cancers. *9*, str. 11259. doi: https://doi.org/10.1038/s41598-019-47757-6

Andries, V., Vandepoele, K., Staes, K., Berx, G., Bogaert, P., Isterdael, G. V., . . . Roy, F. v. (2015). NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. *BMC Cancer, 10*. doi:doi: 10.1186/s12885-015-1408-5

Black, E., & Giunta, S. (2018). Repetitive Fragile Sites: Centromere Satellite DNA As a Source of Genome Instability in Human Diseases. *Genes, 9*, str. 615. doi:doi:10.3390/genes9120615

Flèche, P. (2001). tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *MC Microbiol, 1*. doi:doi:10.1186/1471-2180-1-2.

Fondon, J., & Garner, H. (2004). Molecular origins of rapid and continuous morphological evolution. *PNAS, 101*, str. 18058–63.

Glunčić, M., & Paar, V. (2013). Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. *Nucleic Acid Research, 41*, str. e17. doi:doi:10.1093/nar/gks721

Glunčić, M., Vlahović, I., & Paar, V. (2019). Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes. *Sci Rep*. doi:doi:10.1038/s41598-019-49022-2

Jain, M., Olsen, H., Turner, D., Stoddart, D., Bulazel, K., Paten, B., & et al. (2018). Linear assembly of a human centromere on the Y chromosome. *Nat Biotechnol., 4*, str. 321-3. doi:doi: 10.1038/nbt.4109

Levy-Sakin, M., & al., e. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications, 10*, str. 1025. doi:https://doi.org/10.1038/s41467-019-08992-7

Levy-Sakin, M., Pastor, S., Mostovoy, Y., & al, e. (2019). Genome maps across 26 human

Romero, V., Hosomichi, K., Nakaoka, H., Shibata, H., & Inoue, I. (2017). Structure and evolution of the filaggrin gene repeated region in primates. *BMC Evol Biol, 17*. doi:doi:10.1186/s12862-016-0851-5

Rozen, S., Skaletsky, H., Marszalek, J., Minx, P., Cordum, H., Waterston, R., & al, e. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature, 6942*, str. 873-6. doi:doi: 10.1038/nature01723

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., . . . Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research, 47*(21), str. 10994–11006 . doi:doi: 10.1093/nar/gkz841.

Tyler-Smith, C., & Brown, W. (1987). Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol, 3*, str. 457-70.

Tyler-Smith, C., Oakey, R., Larin, Z., Fisher, R., Crocker, M., Affara, N., & al, e. (1993). Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat Genet., 4*, str. 368-75. doi:doi: 10.1038/ng1293-368

Uralsky, L., Shepelev, V., Alexandrov, A., Yurov, Y., Rogaev, E., & Alexandrov, I. (2019). Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data Brief., 24*, str. 103708. doi:doi: 10.1016/j.dib.2019.103708

Vlahovic, I., Gluncic, M., Rosandic, M., Ugarkovic, Đ., & Paar, V. (2017). Regular Higher Order Repeat Structures in Beetle Tribolium castaneum Genome. *Genome Biol Evol, 9*, str. 2668–2680. doi:doi:10.1093/gbe/evw174

Vlahović, I., Glunčić, M., & Paar, V. (n.d.). Rich polymorphic variants of alpha satellite 34mer higher order repeats in hg38 assembly of human chromosome Y (submitted for publication).

Vlahović, I., Glunčić, M., Dekanić, K., Mršić, L., Jerković, H., Martinjak, I., & Paar, V. (2022). Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, str. 237-242. doi:doi: 10.23919/MIPRO48935.2020.9245278.

Vlahović, I., Glunčić, M., Rosandić, M., & Paar, V. (n.d.). Higher Order Repeats (HORs) in Neuroblastoma BreakPoint Family (NBPF) genes distinguish Neanderthal and Human populations reveal population-specific patterns of structural variation. *Nat Commun, 10*, str. 1025. doi:https://doi.org/10.1038/s41467-019-08992-7

Okonechnikov K, Golosova O, Fursov M & the GENE team. Unipro UGENE: a unified bioinformatics toolkit . Bioinformatics 2012 28: 1166-1167. doi:10.1093/bioinformatics/bts091

O'Bleness, M., Searles, V. B., Dickens, C. M., Astling, D., Albracht, D., Mak, A. C., . . . Sikela, J. M. (2014). Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. *BMC Genomics, 387*. doi:https://doi.org/10.1186/1471-2164-15-387

Paar, V., Glunčić, M., Rosandić, M., Basar, I., & Vlahović, I. (2011). Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees. *Molecular Biology and Evolution, 28,* str. 1877–1892. doi:https://doi.org/10.1093/molbev/msr009

Paar, V., x Glunčić, V., Basar, I., Rosandic, M., Paar, P., & Cvitković, M. (2011). Large Tandem, Higher Order Repeats and Regularly Dispersed Repeat Units Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes. *Journal of Molecular Evolution, 1*, str. 34-55.

Pathak, D., & Ali, S. (2012). Repetitive DNA: A Tool to Explore Animal Genomes/Transcriptomes, Functional Genomics. (D. G. Meroni, Ur.) *InTech*. doi:DOI: 10.5772/48259

Perry, G., Tito, R., & Verrelli, B. (2007). The evolutionary history of human and chimpanzee Y-chromosome gene loss. *Mol Biol Evol., 3*, str. 853-9. doi:doi: 10.1093/molbev/msm002

Rigden, D. J., & Fernández, X. M. (2022). The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. Nucleic acids research, 50(D1), D1–D10. https://doi.org/10.1093/nar/gkab1195.

Romero, V., & al, e. (2018). High Order Formation and Evolution of Hornerin in Primates. *Genome Biology and Evolution, 10*, str. 3167–3175.

Romero, V., Hosomichi, K., Nakaoka, H., Shibata, H., & Inoue, I. (2017). Structure and evolution of the filaggrin gene repeated region in primates. *BMC Evol Biol, 17*. doi:doi:10.1186/s12862-016-0851-5

Rozen, S., Skaletsky, H., Marszalek, J., Minx, P., Cordum, H., Waterston, R., & al, e. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature, 6942*, str. 873-6. doi:doi: 10.1038/nature01723

Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., . . . Linke, D. (2019). Tandem repeats lead to

lineage with respect to cognitive capabilities (submitted for publication).

Weissensteiner, M., Pang, AWC, Bunikis, I., & al, e. (2017). Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res., 5*, str. 697–708. doi:doi:10.1101/gr.215095.116.

sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research, 47*(21), str. 10994–11006 . doi:doi: 10.1093/nar/gkz841.

Tyler-Smith, C., & Brown, W. (1987). Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J Mol Biol, 3*, str. 457-70.

Tyler-Smith, C., Oakey, R., Larin, Z., Fisher, R., Crocker, M., Affara, N., & al, e. (1993). Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat Genet., 4*, str. 368-75. doi:doi: 10.1038/ng1293-368

Uralsky, L., Shepelev, V., Alexandrov, A., Yurov, Y., Rogaev, E., & Alexandrov, I. (2019). Classification and monomer-by-monomer annotation dataset of suprachromosomal family 1 alpha satellite higher-order repeats in hg38 human genome assembly. *Data Brief., 24*, str. 103708. doi:doi: 10.1016/j.dib.2019.103708

Vlahovic, I., Gluncic, M., Rosandic, M., Ugarkovic, Đ., & Paar, V. (2017). Regular Higher Order Repeat Structures in Beetle Tribolium castaneum Genome. *Genome Biol Evol, 9*, str. 2668–2680. doi:doi:10.1093/gbe/evw174

Vlahović, I., Glunčić, M., & Paar, M. (n.d.). Rich polymorphic variants of alpha satellite 34mer higher order repeats in hg38 assembly of human chromosome Y (poslano za objavu).

Vlahović, I., Glunčić, M., Dekanić, K., Mršić, L., Jerković, H., Martinjak, I., & Paar, V. (2022). Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, str. 237-242. doi:doi: 10.23919/MIPRO48935.2020.9245278.

Vlahović, I., Glunčić, M., Rosandić, M., & Paar, V. (n.d.). Higher Order Repeats (HORs) in Neuroblastoma BreakPoint Family (NBPF) genes distinguish Neanderthal and Human lineage with respect to cognitive capabilities (poslano za objavu).

Weissensteiner, M., Pang, AWC, Bunikis, I., & al, e. (2017). Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res., 5*, str. 697–708. doi:doi:10.1101/gr.215095.116.