TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Inimese 15. kromosoomi tsentromeeripuu konstrueerimine inimese pangenoomi versioon I indiviidide põhjal

Bakalaureusetöö

12 EAP

Katariina Ilmoja

Juhendaja MSc Tarmo Puurand

TARTU 2025

INFOLEHT

Inimese 15. kromosoomi tsentromeeripuu konstrueerimine inimese pangenoomi versioon I indiviidide põhjal

Tsentromeerid on kromosoomide funktsioneerimise aluspõhjad, tagades nende korrektse jagunemise rakutsüklis. Tsentromeeride ülesehituse keerukuse tõttu on alles hiljutised tehnoloogilised arengud võimaldanud nende täieliku järjestuse määramist ja analüüsi. Käesolev töö keskendus inimese tsentromeeride järjestusliku mitmekesisuse uurimisele, kasutades selleks T2T-CHM13v2.0 referentsgenoomi ja inimese pangenoomi andmeid. Leiti, et CHM13 referentsgenoomi tsentromeeri sagedaseim k-meer võimaldas eristada indiviide k-meeri mustrite põhjal. Uuringust selgus, et tsentromeeripiirkonnad võivad indiviiditi varieeruda ning valitud k-meeride põhjal on võimalik eristada indiviidigruppe klasteranalüüsi abil. Tulemused näitavad, et kuigi tsentromeerid on korduspiirkonnad, on seal tuvastatavad järjestuspõhised erinevused, mida saab edukalt kasutada genoomse võrdluse aluseks.

Märksõnad: inimese tsentromeer, k-meer, 15. kromosoom, fülogeneetiline puu, CHM13, pangenoom

CERCS kood: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Construction of the human chromosome 15 centromere phylogenetic tree based on individuals from Human Pangenome version I

Centromeres are the foundation for chromosome functioning, ensuring their correct division during cell division. Due to the complex structure of centromeres only recent technological developments have enabled their complete sequence determination and analysis. This work focused on investigating the sequence diversity of human centromeres using the T2T-CHM13v2.0 reference genome and human pangenome data. It was found that the most frequent k-mer of the centromere of the CHM13 reference genome allowed individuals to be distinguished based on k-mer patterns. The study revealed that centromere regions can vary from individual to individual, and it is possible to distinguish groups of individuals based on selected k-mers using cluster analysis. The results show that although centromeres are repetitive regions, sequence-based differences can be detected there, which can be successfully used as the basis for genomic comparison.

Keywords: human centromere, k-mer, chromosome 15, phylogenetic tree, CHM13, pangenome

CERCS code: B110 Bioinformatics, medical informatics, biomathematics, biometrics

SISUKORD

KASUTATUD LÜHENDID 6

SISSEJUHATUS 7

1. KIRJANDUSE ÜLEVAADE 8

1.1. Rakutsükkel 8

1.1.1. Mitoos 8

1.1.2. Meioos 11

1.2. Inimese kromosoomide struktuur 12

1.3. Inimese tsentromeeride ülesehitus 14

1.3.1. Tsentromeeridele omased valgud 16

1.4. Tsentromeeride seos kromosoomi stabiilsuse ja geneetiliste häiretega 17

1.5. Fülogenees 18

1.6. Inimese pangenoom 19

2. EKSPERIMENTAALOSA 22

2.1. Töö eesmärgid 22

2.2. Materjal ja metoodika 23

2.2.1. Referentsgenoom CHM13v2.0 23

2.2.2. 47 indiviidi järjestused inimese pangenoomi andmestikust 23

2.2.3. Töö läbiviimiseks vajaliku kromosoomi valik 24

2.2.4. Töö käik 24

2.2.5. Sagedaseima k-meeri tuvastamine CHM13 referentsgenoomis 25

2.2.6. Tsentromeeripuu loomine 27

2.3. Tulemused 29

2.4. Arutelu 30

KOKKUVÕTE 32

RESÜMEE / SUMMARY 33

TÄNUSÕNAD 35

KIRJANDUSE LOETELU 36

KASUTATUD VEEBIAADRESSID 43

LISAD 45

Lisa 1 45

Lisa 2 46

LIHTLITSENTS 47

KASUTATUD LÜHENDID

aDNA – vana DNA (*ancient DNA*)

CCAN – konstitutiivne tsentromeeriga seotud valguvõrgustik (*Constitutive Centromere Associated Network*)

CENP – tsentromeerivalk (*centromere protein*)

CHM - täielikult hüdatidiformne mool (*complete hydatidiform mole*)

CHM13 – referentsgenoom (*reference genome*)

Chr – kromosoom (*chromosome*)

HOR – kõrgemat järku kordusstruktuurid (*higher-order repeats*)

NGS – järgmise põlvkonna sekveneerimine (*Next-Generation Sequencing*)

SNP – üksiknukleotiidne polümorfism (*single nucleotide polymorphism*)

SNV – üksiknukleotiidne variant (*single nucleotide variant*)

T2T – telomeerist telomeerini (*telomere-to-telomere*)

SISSEJUHATUS

Tsentromeerid on inimese genoomi piirkonnad, mis tagavad enda komponentidega kromosoomide korrektse jagunemise mitoosi ja meioosi ajal, säilitades samal ajal geneetilist stabiilsust. Need koosnevad kordusjärjestustest, millest üks uuritumaid on alfa-satelliit-DNA. Tsentromeeride funktsioon ei ole seotud ainult DNA järjestustest, vaid olulist rolli mängivad ka tsentromeerisisesed komponendid, näiteks CENP-A põhine kintohoor.

Olenemata tsentromeeride olulisusest on nende uurimise keerukus seotud raskesti sekveneeritava struktuuriga. Täispikkade lugemite ning uute bioinformaatiliste tööriistade abil on saanud võimalikuks tsentromeeride järjestamine. Olulist rolli mängib T2T-CHM13v2.0 referentsgenoom, mis esindab esimest järjestuslikult täielikku inimgenoomi. See on pannud aluse tsentromeeride indvidiidevahelisele võrdlusele.

Käesolev bakalaureusetöö keskendub inimese tsentromeeride järjestuse analüüsile, pöörates tähelepanu nende struktuurile, kordustele ja varieeruvusele. Töös antakse teoreetiline ülevaade tsentromeeride ehitusest ja funktsioonist ning viiakse läbi analüüs 15. kromosoomi tsentromeeri põhjal. Eesmärgiks oli määrata T2T-CHM13 referentsgenoomi alusel sagedaseim k-meer ning uurida nende seotussaitide vahel olevaid variatsioonide esinemist genoomide assemblitest. Käesoleva bakalaureusetöö põhieesmärgiks oli saadud tulemuste põhjal koostada fülogeneetiline puu, et hinnata indiviididevahelisi erinevusi ja analüüsida tsentromeeride potentsiaalset evolutsioonilist ja funktsionaalset varieeruvust. Lisaks oli töö üheks eesmärgiks ka GenomeTester4 tarkvara sobivuse hindamine tsentromeeride k-meeridepõhiseks analüüsiks.

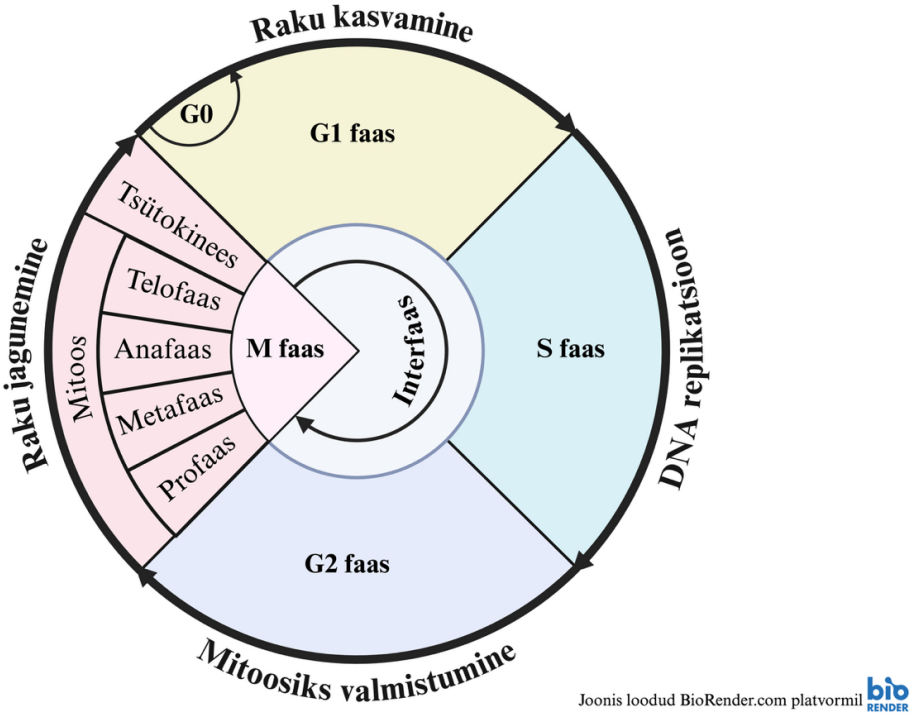
1. KIRJANDUSE ÜLEVAADE

1.1. Rakutsükkel

Rakutsükkel jaguneb kaheks osaks, milleks on mitoos ja meioos. Mõlemad on olulised geneetilise materjali ülekandmisel rakkude jagunemisel. Mitoos toimub somaatilistes rakkudes ning selle tulemusel tekivad kaks geneetiliselt identset tütarrakku (Alberts *et al.*, 2002a). Meioos leiab aset sugurakkudes (*germ cells*) ning vähendab kromosoomide arvu 23-ni, et viljastumisel järglaste rakkudes oleks 46 kromosoomi (Alberts *et al*., 2002c). Kuigi mõlemad protsessid jagavad sarnaseid alamfaase (profaas, metafaas, anafaas, telofaas, tsütokinees) on nende eesmärgid ning tulemused erinevad. Rakutsükli täpne kulg on oluline organismi kasvuks ja arenguks, tagades geneetilise materjali kontrollitud jagunemine (Aly, 2023).

1.1.1. Mitoos

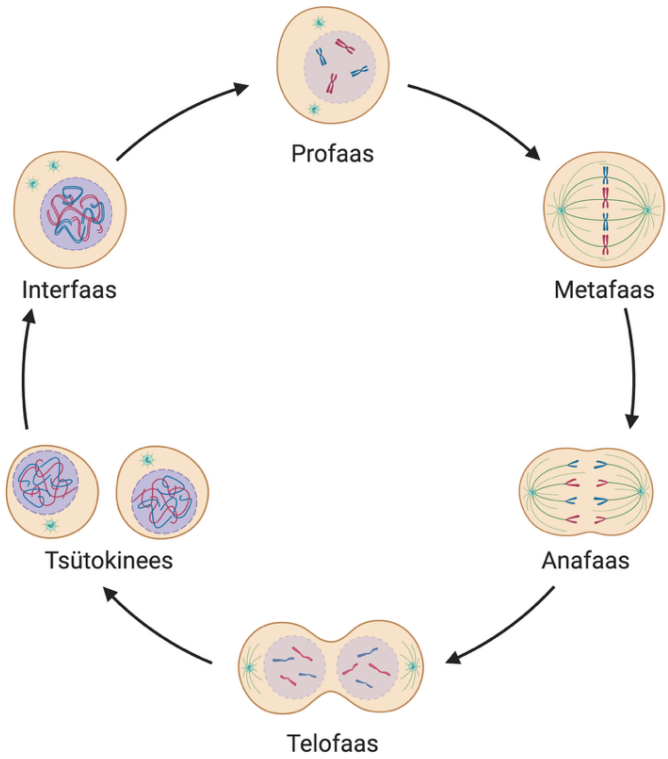
Eukarüootse rakutsükli põhilistest eesmärkidest on DNA replikatsioon, olles aluseks kahe identse tütarraku tekkele. Interfaasi esimeses G1 (vahe ehk *gap*) faasis (Joonis 1) toimub raku pidev kasv sobivate tingimuste saamiseks, sünteesides näiteks valke ja RNA-d. Järgnevas S (süntees ehk *synthesis*) faasis toimub kogu genoomi kahekordistumine ning sünteesitakse histoone ja kromatiini valke DNA stabiilsuse saavutamiseks. G2 faasis jätkub raku kasvamine, valmistades rakku ette mitoosiks. Selles faasis toimub DNA replikatsiooni kontroll ning parandatakse võimalikud tekkinud vead. Samal ajal aktiveeritakse vajalikud valgukinaasid ja regulaatorid M faasi liikumiseks. Järgnev M faas kirjeldab mitoosi, milles toimub kromsoomide lahknemine ja raku tsütoplasma jagunemine ehk tsütokinees (Alberts *et al*., 2002b; Aly, 2023).



Joonis 1. Rakutsükkel. Joonisel on kujutatud faaside kaupa rakutsükkel, sisaldades väljatoodult ka mitoosi erinevaid faase. G0 on rakkude puhkeolek, kus nende aktiivne jagunemine ei toimu, näiteks ebasobivate keskkonnatingimuste tõttu. Selles faasis võivad rakud olla kas ajutiselt või kuni raku/organismi surmani (Alberts *et al*., 2002b). (Loodud BioRender.com saidil võttes aluseks NHGRI, 2025 joonise).

M (mitoos ehk *mitosis*) faas koosneb viiest faasist, millest esimene on profaas (Joonis 2). Selles toimub DNA kondenseerumine tihedalt pakitud kromosoomideks, mida on mikroskoobi all võimalik näha (Aly, 2023; Tram *et al*., 2003). Lisaks kromsosoomile endale on selgelt nähtav ka tsentromeer ja selle ehitus. Profaasis algab kääviniidistiku moodustumine, mille mikrotuubulitest koosnev struktuur liigutab kromosoome. Tsentrosoomid, mis vastutavad mikrotuubulite organiseerituse eest, liiguvad raku erinevatele pooltele. Profaasi lõpus toimub tuumamembraani lagunemine mikrotuubulite ja kromosoomide vaheliseks interaktsiooniks, mida kutsutakse prometafaasiks (Alberts *et al*., 2002a). See etapp on üleminekuperiood profaasi ja metafaasi vahel ning on oluline kromosoomide joondumise ettevalmistuseks. Metafaasi ajal on kromosoomid joondunud raku keskele metafaasiplaadile, et saaks toimuda veatu jagunemine (Tan *et al*., 2015). Iga kromatiid on ühenduses mikrotuubulitega, mis ulatuvad erinevast poolusest. Kontrollina peatub rakk metafaasis seni, kuni on toimunud õige kinnitumine ja joondumine. Anafaas algab tsentromeeri piirkonnas asuva õdekromatiide ühendava struktuuri ehk kohesiini (*cohesin*) (Joonis 7) lagunemisega (Brooker ja Berkowitz, 2014). Selle tulemuseks on õdekromatiidide eraldumine

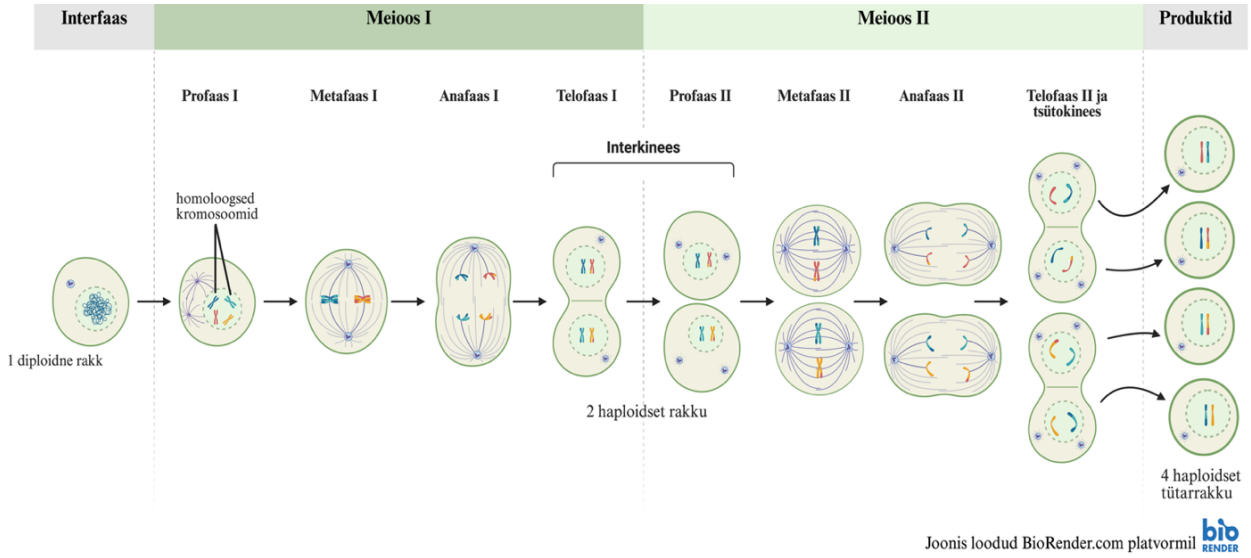
üksteisest ja nende liikumine mööda kääviniite raku vastaspooltele. Nii saavutatakse tütarrakkudele täiskomplekti kromosoomide saamine. Anafaas on kriitilise tähtsusega, sest selles faasis toimunud kromosoomide ebaõnnetusunud jagunemine võib põhjustada kromosoomi aneuploidsust (Brooker ja Berkowitz, 2014). Telofaas on mitoosi viimane faas, mille käigus jagunenud kromosoomide ümber moodustuvad uued tuumamembraanid. Samas etapis toimub kromosoomide dekondenseerumine ja kromatiiniks muutumine. Telofaasile järgneb tsütokinees, mille tulemuseks on kaks iseseisvat tütarrakku, mis mõlemad sisaldavad identset geneetilist materjali. Aktiini- ja müosiinifilamente sisaldav kontraktiilne ring moodustub raku keskjoonele ning kokkutõmbumisel surutakse membraan sissepoole, mis põhjustab raku jagunemist kaheks (Alberts *et al*., 2002a). Iga alam- ja peafaas on täpselt ajastatud ja kontrollitud, et välistada DNA jagunemisega seotud vigade tekkimist. Mitoosi edukas läbimine on kriitilise tähtusega, vastupidisel korral esinevad häired võivad põhjustada genoomi ebastabiilsust ning kasvajate teket inimestes (Bakhoum *et al*., 2014).



Joonis 2. Rakutsükkel täpsustatud M-faasiga. Tsütokineesi faasi lõpuks saavad mõlemad tütarrakud koos kromosoomidega endale ka ühe tsentromsoomi (Alberts *et al*., 2002a). (Koostatud BioRender.com saidil võttes aluseks NHGRI, 2025 joonise).

1.1.2. Meioos

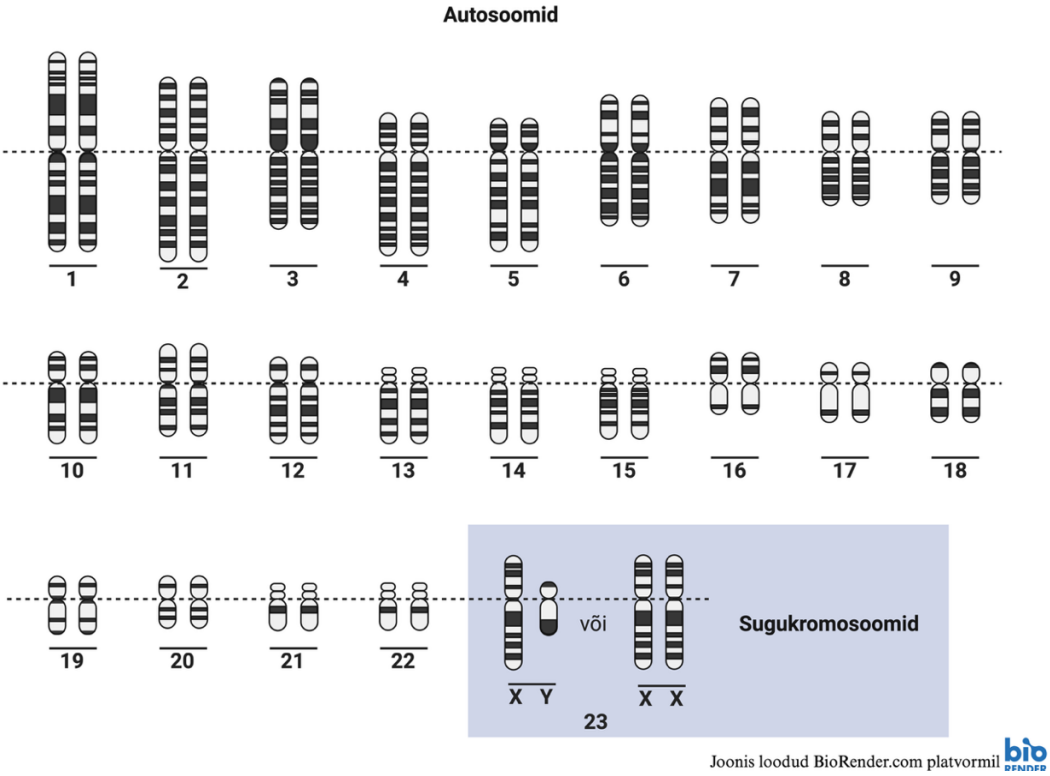
Meioosi kaudu moodustuvad inimese muna- või seemnerakud ehk sugurakud. Erinevalt mitoosist, mis toimub kõikides somaatilistes rakkudes, on meioosi eesmärk kromosoomide arvu vähendamine poole võrra. Selle tulemuseks on 23 kromosoomi, et viljastumisel moodustuks tavapärane diploidne rakk (Alberts *et al*., 2002c). Meioos koosneb kahest osast: meioos I ja meioos II. Esimeses jagunemises eralduvad homoloogsed kromosoomid, teises, sarnaselt mitoosile (Joonis 2), õdekromatiidid. Meioosile eelneb interfaas koos DNA replikatsiooniga. Meioos I algab profaasiga, kus toimub homoloogsete kromosoomide paaritamine ning moodustuvad neljast kromatiidist koosnevad bivalendid (Brooker ja Berkowitz, 2014). Järgnevalt toimub kromosoomide geneetilise materjali vahetamine (*crossing-over*), suurendades geneetilist mitmekesisust. Metafaas I ajal on homoloogsed kromosoomid joondunud samuti raku keskele, kuid ema- ja isaliini (*maternal* ja *paternal*) kromosoomid lahutatakse. Anafaas I käigus tõmmatakse kääviniitide abil kromosoomid raku erinevatele pooltele (Alberts *et al*., 2002c). Selle faasi täpsus sõltub tsentromeeri, täpsemalt CENP-A ja teiste tsentromeerivalkude olemasolust, mis tagavad korrektse kinetohoori moodustumise. Meioos I lõpuks moodustuvad kaks haploidset rakku, kus iga kromosoom koosneb siiski kahest õdekromatiidist. Interkinaasi (Joonis 3) jooksul DNA replikatsiooni ei toimu, kuid rakk võib taastada mõned valgud ja valmistuda teiseks jagunemiseks. Selle etapi ajal säilib haploidne kromosoomistik, mis võimaldab kohe alustada meioosi II-ga (Gottlieb *et al*., 2023). Uue DNA replikatsioonita alustab rakk meioos II, mis kulgneb samade etappidena nagu mitoos (Joonis 1): profaas II, metafaas II, anafaas II ja telofaas II. Eralduvad õdekromatiidid, mis olid meioos I käigus koos püsinud ning moodustuvad neli haploidset tütarrakku. Igaühel on aga enda erinev geneetiline kooslus (Alberts *et al*., 2002c). Kuna meioos määrab, millised kromosoomid satuvad sugurakkudesse, on selle täpne läbimine hädavajalik. Vead kromosoomide joondumises, tsentromeeride funktsioonis või kromosoomide eraldumises võivad viia geneetiliste haiguste, arenguhäirete või viljatuse tekkeni.



Joonis 3. Meioosi kulg etappidena. Joonisel on kujutatud meioos I ja meioos II faasid, mille tulemuseks on 4 geneetiliselt unikaalset haploidset tütarrakku. Homoloogsed kromosoomid vahetavad omavahel DNA-d (*crossing-over*), mille tulemusel moodustuvad uued kombinatsioonid emalt ja isalt pärit kromosoomidest. Meioosi lõppedes on igas sugurakus 23 kromosoomi. (Koostatud BioRender.com saidil võttes aluseks NHGRI, 2025 joonise).

1.2. Inimese kromosoomide struktuur

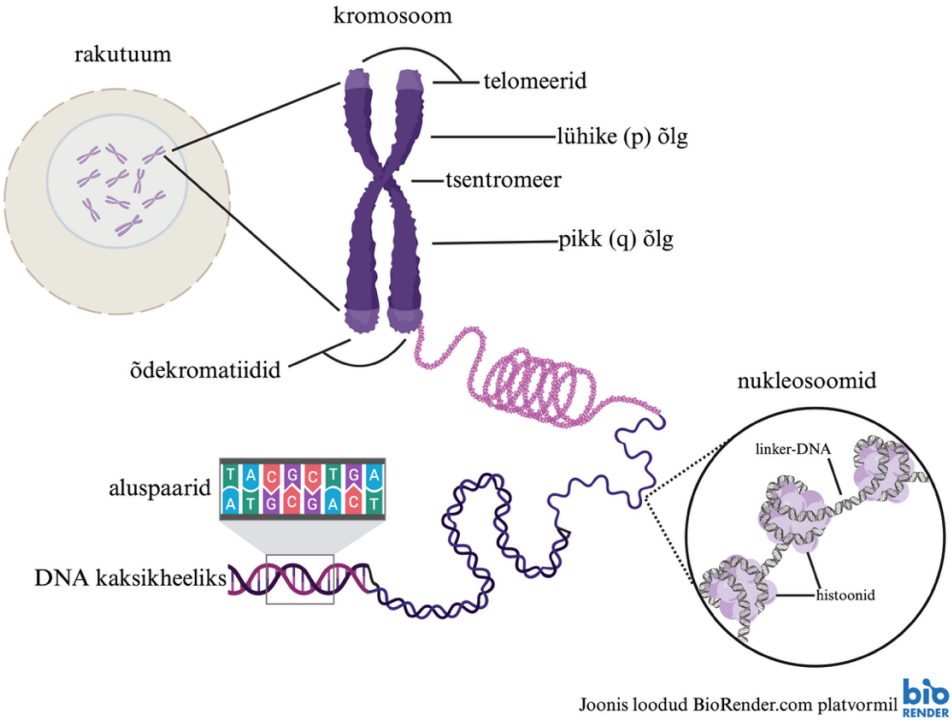
Kromosoomid on rakutuumas asuvad struktuurid, sisaldades organismi pärilikku informatsiooni (NHGRI, 2019). Inimese diploidne genoom, mis sisaldab kahte koopiat igast kromosoomist, koosneb ligikaudu 6,4 miljardist DNA aluspaarist (Goldfeder *et al*., 2017). Aluspaarid on jaotunud 22 autosoomi ja ühte sugukromosoomide paari, kus sugukromosoomide kombinatsioon määrab bioloogilise soo (Brown, 2002). Nendest tuleneb tavapärase inimese 46 kromosoomi ehk 23 paari. Kromosoomid on järjestatud vastavalt pikkusele (Joonis 4), millest suurim kromosoom 1 sisaldab 248,956,422 aluspaari ning väikseim kromosoom 21 sisaldab 46,709,983 aluspaari (Lander *et al*., 2001).



Joonis 4. Inimese karüotüüp. Joonisel on näha 22 paari autosoomse ja 2 sugukromosoomi paari, mis on joonestatud tsentromeeride järgi. Inimese bioloogiline sugu tuleneb meestel tavaliselt XY ning naistel XX paarist (Pathak ja Bordoni, 2023). Joonisel kujutatud kromosoomide must-valge värvus põhineb G-ribade mustril. Tumedad G-positiivsed ribad on tihedamalt pakitud ja AT-rikkad heterokromatiinsed piirkonnad, heledamad G-negatiivsed ribad aga GC-rikkamad ja vastavad pigem aktiivsele eukromatiinile (Niimura ja Gojobori, 2002). (Koostatud BioRender.com saidil võttes aluseks NHGRI, 2025 joonise).

Kromosoomide ehituses (Joonis 5) olevad nukleosoomid koosnevad 147 bp pikkusest DNA lõikudest, mis on keeratud ümber histoonide oktameeri. See on kaheksast valgust koosnev kompleks, sisaldades kahte koopiat H2A, H2B, H3 ja H4 histoonidest (Luger *et al*., 1997). Nukleosoomide vahele jääb linker-DNA, mis ühendab järjestikuseid nukleosoome. Linker-DNA seondumine linker histoon H1-ga aitab hoida kromatiini stabiilsust (Olins ja Olins, 2003). Kromatiin jaotatakse funktsiooni ja tiheduse alusel kaheks põhivormiks: eukromatiiniks ja heterokromatiiniks. Eukromatiin on lõdvemalt pakitud ja transkriptsiooniliselt aktiivne, kuna selles olev DNA on kergemini kättesaadav geenidekspressiooniks. Heterokromatiin seevastu on tihedalt pakitud ja reeglina geneetiliselt passiivne. See esineb sageli genoomi struktuursetes piirkondades, kus geenidekspressioon ei ole vajalik (Misteli, 2007). Tsentromeerid on tüüpiline näide heterokromatiinist, kus tihedalt pakitud DNA aitab kaasa kromosoomide struktuursele stabiilsusele ja segregatsioonile rakutsükli jooksul (Allshire ja Karpen, 2008). Tsentromeeride paiknemine

kromosoomis määrab lisaks ka nende erinevad variandid (metatsentriline, submetatsentriline, akrotsentriline või telotsentriline), olenevalt tsentromeeride kaugusest kromosoomi keskkohast (Bellou *et al*., 2024).

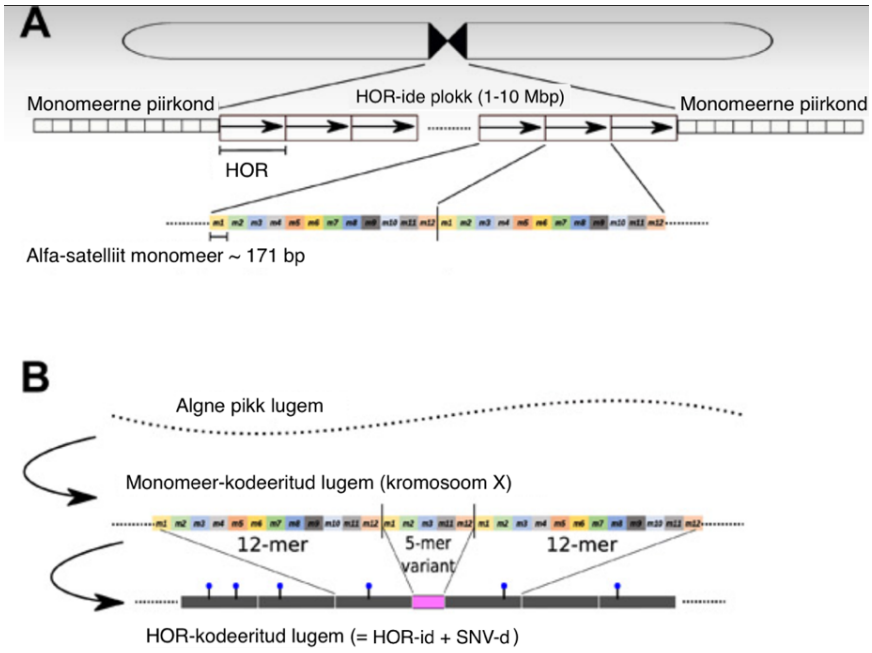


Joonis 5. Kromosoomi ülesehitus. Joonis illustreerib kromosoomi struktuurielemente. Näha on kromosoomi telomeerid ning tsentromeer, mis ühendab õdekromatiide. Telomeerid on kromosoomide otstes paiknevad kordusjärjestused, mis kaitsevad genoomi lagunemise ja valestikinnitumise eest, hoides ära selle, et DNA parandussüsteemid tuvastaksid kromosoomi loomulikke otsi kahjustusena (Blackburn, 2001). Lisaks on välja toodud kromosoomidele omane lühike (p) ja pikk (q) õlg. Allpool on kujutatud DNA järjestuslik pakkimine alates DNA kaksikheeliksist kuni nukleosoomideni, kus DNA on keerdunud histoonvalkude ümber. (Koostatud BioRender.com saidil võttes aluseks NHGRI, 2025 joonise).

1.3. Inimese tsentromeeride ülesehitus

Tsentromeer ja täpsemalt sellele moodustuv valguline kompleks kinetohoor mängivad keskset rolli kromosoomide õigesti jaotumises tütarrakkudesse nii mitoosi kui ka meioosi ajal (Di Tommaso *et al*., 2024). Tsentromeeride suured ja korduvad struktuurid muudavad aga nende järjestamise äärmiselt keeruliseks (Logsdon *et al*., 2024). McNulty *et al*. uuringud näitavad, et tsentromeeri DNA järjestus võib erineda isegi sama kromosoomi eri kohtades, mille tõttu on olnud raskendatud neid ka põhjalikumalt uurida (McNulty *et al*., 2018).

Tsentromeerid koosnevad peamiselt tandeemsetest kordustest, mis on üksteise järel paiknevad DNA järjestused. Tandeemsed kordused on olulised, kuna need loovad tsentromeerse kromatiini (Joonis 7) korrastatud ülesehituse ja kindlad piirkonnad, kuhu tsentromeerivalgud saavad seonduda (McKinley ja Cheeseman, 2015). Korduvad järjestused koos nendele ladestunud valkudega võimaldavad säilitada tsentromeeri funktsiooni läbi rakkude jagunemiste (McNulty ja Sullivan, 2018). Inimese tsentromeerides moodustab tandeemse korduse α-satelliit-DNA, mis koosneb 171 aluspaari pikkustest monomeeridest (Joonis 6A). Monomeerid on pakitud kõrgemat järku kordusstruktuuridesse ehk HOR-idesse (higher-order repeats), mis loovad kindlale kromosoomile omase struktuuri ning varieeruvad monomeeride arvu ja paigutuse poolest (Di Tommaso *et al*., 2024; Sullivan *et al*., 2020; Sullivan *et al*., 2017). Näiteks on inimese X-kromosoomi tsentromeeril 12-monomeeriline HOR, samas kui Y-kromsoomil on 34-monomeeriline HOR, mis rõhutab kordusstruktuuride varieeruvust ja mitmekesisust eri kromosoomidel (Lisa 1) (McNulty *et al*., 2018; Sullivan *et al*., 2020).

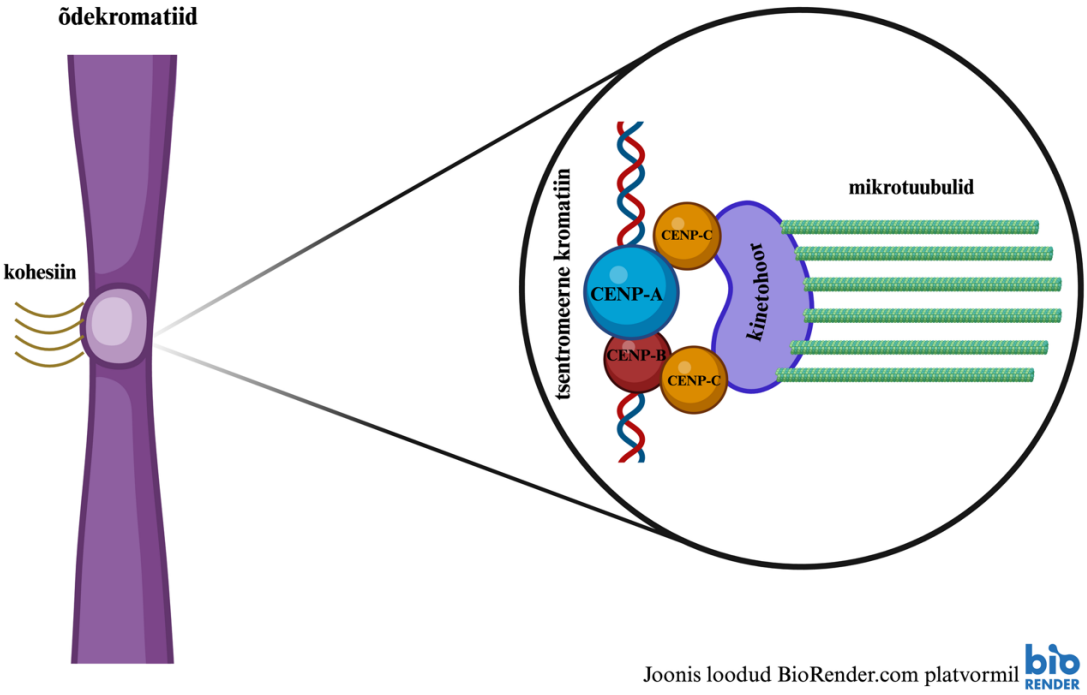


Joonis 6. Tsentromeeride järjestustes esinevate erinevate HOR-variantide põhjalik analüüs. Paneel A visualiseerib skeemilist ülevaadet inimese tsentromeeripiirkondade tüüpilisest DNA-järjestuse struktuurist. Lisaks alfa-satelliit-DNA-st koosnevatele monomeeridele on välja toodud HOR-id. HOR-id korduvad suures ulatuses (1-10 Mbp) ja on spetsiifilised iga kromosoomi jaoks, moodustades HOR-ide ploki tsentromeeri keskmes. Paneel B visualiseerib pikkade järjestuste HOR-kodeerimise etappe. Paneelis B näidatud vahepealne 5-meer näitab HOR-ide variatsiooni võrreldes kõrvalasuva tavalise HOR kordusega, mis võivad olla tingitud üksiknukleotiidsest variandist (SNV) (*single nucleotide variant*) (Suzuki *et al*., 2020). (Modifitseeritud Suzuki *et al*., 2020 järgi).

1.3.1. Tsentromeeridele omased valgud

Üks tsentromeeride unikaalne omadus on spetsiifiline valk CENP-A, mis on histoon H3 asendaja (Di Tommaso *et al*., 2024; Mastrorosa *et al*., 2024). Selle olemasolu eristab tsentromeere teistest kromosoomi piirkondadest. Rakutsükli G1-faasis lisatakse tsentromeerile uus CENP-A, mis aitab hoida selle asukohta stabiilsena. Hilisemas interfaasis võimaldab CENP-A konstitutiivse tsentromeeriga seotud valguvõrgustiku (CCAN) moodustumist, mis on vajalik kinetohoori moodustamiseks (Gamba ja Fachinetti, 2020; Pesenti *et al*, 2022). CENP-A jõuab tsentromeerile tänu selle C-otsa piirkonnas asuvale histooni voltimisdomeenile, mis on sihtimise seisukohalt hädavajalik (Ando *et al.*, 2002). CENP-A olemasoluta ei suudaks rakk kindlaks teha, kuhu kinetohoor moodustada. CENP-A moodustab vundamendi kinetohoorile, mis on mikrotuubulitega ühenduv valkude kogum, mis juhib kromosoomide liikumist raku jagunemise ajal. Täpsemalt on kinetohoori ülesandeks geneetilise informatsiooni stabiilne ülekandumine rakupõlvkondade vahel, aidates kromosoomidel korrektselt jaguneda emarakust tütarrakkudesse, mis on geneetiliselt identsed (Di Tommaso *et al*., 2024). Kui tsentromeeri ei tunta ära õiges kohas kromosoomil, võib kromosoomide eraldumine ebaõnnestuda või need võivad valesti jaguneda.

Lisaks CENP-A-le on tsentromeerile omasteks valkudeks CENP-B ning CENP-C, mis kõik osalevad terve rakutsükli vältel (Ando *et al*., 2002). CENP-B on hetkel ainus inimese tsentromeerivalk, mis seondub otseselt kindla DNA järjestusega nimega CENP-B boks. See on 17 bp pikkune lõik, mis esineb alfa-satelliit-DNA monomeerides (Masumoto *et al*., 1989). Lisaks DNA seondumisele muudab CENP-B DNA ruumilist struktuuri luues silmuseid, mis tekitavad soodsad keskkonnatingimused teiste valkude õigeks positsioneerimiseks (Chardon *et al*., 2022). Kuigi mitte kõik tsentromeerid ei sisalda CENP-B bokse, näiteks Y kromosoom, toimivad nendes vajalikud protsessid ideaalselt (Masumoto *et al*., 1989). See viitab sellele, et CENP-B pole tsentromeeri funktsioneerimiseks hädavajalik, kuid soodustab tsentromeeri organiseeritust, stabiilsust ja funktsionaalset täpsust. CENP-C aga seondub nii CENP-A nukleosoomidega kui ka CENP-B-ga, ning on kinetohoori struktuuri jaoks ülioluline. CENP-B ja CENP-C interaktsiooni kaudu luuakse tsentromeerse DNA ja kinetohoori vahel ühendus isegi madala CENP-A taseme korral (Fukagawa ja Earnshaw, 2014). See mehhanism annab aluse kromosoomide korrektsele segregatsioonile stressitingimustes.



Joonis 7. Kinetohoori ja mikrotuubuleid kromosoomiga ühendav kompleks. Õdekromatiidi suurendades on näha tsentromeerse kromatiini piirkond, mis sisaldab tsentromeerivalke. See loob erilise keskkonna, mis tagab kromosoomide korrektse jaotumise rakutsüklis. Lisaks on näha kinetohoori külge seonduvad mikrotuubulid. Välja on toodud õdekromatiide ühendav lüli kohesiin. (Koostatud BioRender.com saidil võttes aluseks French ja Straight, 2013 joonise).

1.4. Tsentromeeride seos kromosoomi stabiilsuse ja geneetiliste häiretega

Inimese genoomis on kromosoomide arv ja ülesehitus väga rangelt reguleeritud, isegi väikseim kõrvalekalle võib põhjustada ulatuslikke arengu- või terviseprobleeme. Sellisteks kõrvalekalleteks on näiteks aneuploidsus ning struktuursed variatsioonid. Tsentromeeri puudumine võib tekitada kromosoomi kaotamise, põhjustades ebanormaalse kromosoomide arvu rakus ehk aneuploidsuse (Sullivan *et al.,* 2020). Tsentromeeri roll kromosoomide stabiilsuses on hädavajalik, eriti CENP-A taseme õige reguleerimise koha pealt. Madal CENP-A tase võib põhjustada kinetohoori nõrka kinnitumist, mis võib viia aneuploiduseni (Santaguida ja Amon, 2015). Dumont *et al*. artiklis toodi välja, et kromosoomispetsiifiline aneuploidsus ei ole juhuslik, vaid on tugevalt mõjutatud tsentromeersetest omadustest. Kromosoomide korrektne jagunemine toimub eelistatult rohkesti DNA-põhiseid tsentromeerikomponente sisaldavate tsentromeeride kaudu. CENP-B ja CENP-C rohkus kromosoomis vähendab selle valesti jaotumise sagedust. Lisaks leiti seos tsentromeeri suuruse ja kromosoomi pikkusega – mida väiksem on tsentromeer ja mida pikem on kromosoom,

seda suurem on valesti jaotumise sagedus (Dumont *et al*., 2020). See peegeldab tsentromeeri vajalikkust kogu raku stabiilsuse tagamisel.

Tsentromeeride olulisust saab uurida lisaks haiguste analüüsimisel. Üks tuntumaid autosoomsete kromosoomidega seotud haigusi on Downi sündroom. Trisoomia 21 ja tsentromeeri vahelist seost on uuritud hiljutises Mastrorosa *et al*. artiklis, kus näidati, et sündroom võib olla seotud ema poolt päritud tsentromeeride asümmeetrilise pikkusega. Leiti, et mida suurem on kahe tsentromeeri pikkuste erinevus, seda suurem on oht, et kromosoomide eraldumine meioosis läheb valesti. Uuringus selgus, et Downi sündroomiga indiviidil olid kolm erinevat 21. kromosoomi tsentromeeri haplotüüpi, millest suurim ja väikseim olid pärit emalt. Suurima tsentromeeri puhul oli CENP-A signaal märkimisväärselt nõrgem, viidates võimalikele probleemidele kinetohoori moodustumises (Mastrorosa *et al*., 2024). 15. kromosoomil põhinev Dup15q sündroom tekib pika õla (q) duplitseerumisel, moodustades kahe tsentromeeriga struktuuri. Kahte koopiat 15q alaosast sisaldav lisakromosoom viib geneetilise materjali neljakordistumiseni, põhjustades erinevaid neuroarengulisi häireid (Butler ja Duis, 2020).

Lisaks aneuploidsusele on tsentromeerile omaseks neotsentromeeride teke. Uus tsentromeer moodustub ebatavalises kromosoomi piirkonnas, mis ei sisalda tavalist alfa-satelliit-DNA-d (Amor ja Choo, 2002). Samuti on uuritud HOR-struktuuride suuruse ja järjestuse varieeruvust, mis mõjutab tsentromeeride funktsiooni ja kromosoomi ennast (McNulty *et al*., 2018). Kõik muutused kromosoomis ja täpsemalt selle tsentromeeris, muudavad inimese vastuvõtlikumaks haigustele, millel on püsivad tagajärjed kogu organismi tasandil. Seetõttu on tsentromeeride molekulaarse ülesehituse mõistmine kriitilise tähtsusega ka meditsiinilises kontekstis.

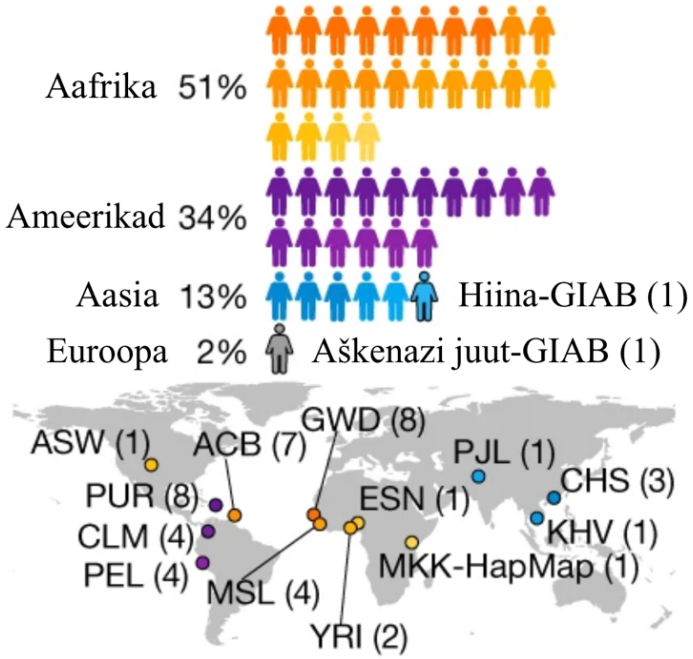
1.5. Fülogenees

Tsentromeeridega seotud geneetilisi muutusi on võimalik analüüsida fülogeneetilisi lähenemisi kasutades. See võimaldab uurida tsentromeeripiirkondade muutusi ja variatsioone, mis võivad olla seotud haigusriskidega. Fülogeneetilised puud põhinevad eeldusel, et geneetilised muutused, näiteks üksiknukleotiidsed polümorfismid (SNP-d), toimuvad harva ja on päritavad. Kui selline mutatsioon tekib ühes kromosoomis, kandub see edasi järglastele ning kujutab endast haruspetsiifilist sündmust. Selliste mutatsioonide kuhjumine võimaldab rekonstruerida kromosoomide arengulist ajalugu ning määrata indiviididevahelisi seoseid (Mallick *et al*., 2016). Fülogeneetilise analüüsi seisukohast on informatiivsed eelkõige need SNP-d, mis esinevad

mitmetel indiviididel, kuid mitte kõigil. Mutatsioonid, mis on kas liiga levinud või esinevad ainult ühes indiviidis, ei anna harude eristamisel lisainfot ning need eemaldatakse analüüsist. Samuti jäetakse välja niinimetatud konfliktseid mutatsioone, mis rikuvad lihtsa hargneva puustruktuuri loogikat, näiteks kui sama muutus on tekkinud iseseisvalt mitmes harus (Green *et al*., 2010). Tsentromeeride puhul on fülogeneetiline lähenemine eriliselt keerukas, sest need piirkonnad koosnevad korduvatest α-satelliitjärjestustest. Siiski võivad nende kordusstruktuuride sees leiduvad üksikmutatsioonid, näiteks SNP-d monomeerides, olla piisavalt unikaalsed ja stabiilsed, et neid kasutada haruspetsiifiliste markeritena. Uuringud on näidanud, et tsentromeerid on genoomi ühed kõige kiiremini evolutsioneeruvad piirkonnad ning sisaldavad märkimisväärset järjestuslikku varieeruvust ka lähisugulastel (Kapli *et al*., 2020; Logsdon *et al*., 2024). Selline mitmekesisus muudab tsentromeerid fülogeneetilise analüüsi sobivaks aluseks, sest mutatsioonid peegeldavad evolutsioonilisi harusid. Fülogeneetiliste puude kaudu saab analüüsida tsentromeersete järjestuste arengut erinevates indiviidides ja tuvastada nende potentsiaalset ühist päritolu.

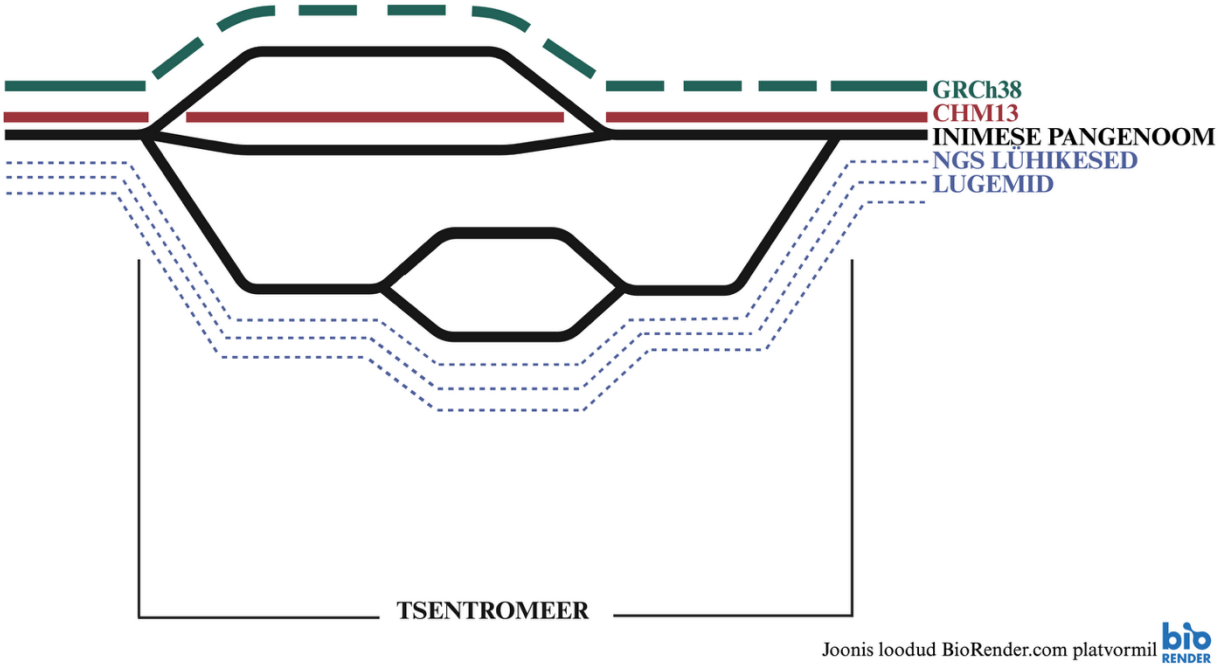
1.6. Inimese pangenoom

Inimese pangenoom viitab mitme indiviidi põhisele genoomi referentsile, mis hõlmab inimeste kogu geneetilist mitmekesisust. Varasem inimgenoomi referents, näiteks GRCh38, kujutab endast ainult üksikute indiviidide järjestusi ning ei sisalda täielikult maailmas esinevaid geenivariante. Selle probleemi lahendas *Human Pangenome Reference Consortium* (HPRC), koostades esimese mustandi inimese pangenoomist, mis sisaldab 47 geneetiliselt mitmekesise indiviidi (Joonis 8) täielikult faasitud diploidseid genoomijärjestusi. Inimese pangenoomi loomiseks järjestati ja analüüsiti indiviidide mõlema vanema (*maternal* ja *paternal*) genoomid. Kasutati täispikki järjestuslugemite tehnoloogiaid, näiteks PacBio HiFi ja Oxford Nanopore, mis võimaldasid lugeda näiteks tsentromeere (Liao *et al*., 2023).



Joonis 8. HPRC 47 indiviidi näidiste valimine. Joonisel kujutatud kolmetähelised lühendid maailmakaardil on määratletud 1000 genoomi projekti poolt. ACB, Kariibi päritolu aafriklased Barbadosel; ASW, Aafrika päritolu inimesed USA edelaosas; CHS, Lõuna-Hiina hanid; CLM, kolumbialased Medellínist (Kolumbia); ESN, esani rahvas Nigeerias; GWD, gambialased Lääne divisjonist; KHV, kinhid Ho Chi Minhi linnast (Vietnam); MKK, maasaid Kinyawa piirkonnas (Keenia); MSL, mende rahvas Sierra Leonest; PEL, peruulased Limast (Peruu); PJL, pandžabi rahvusgrupp Lahore'ist (Pakistan); PUR, puerto ricolased Puerto Ricost; YRI, jorubad Ibadanist (Nigeeria). GIAB (*Genome In A Bottle*) viitab USA-s loodud hästi iseloomustatud ja usaldusväärsetele inimgenoomide referentsaandmestikule (Liao *et al*., 2023). Joonis peegeldab valdavalt Aafrika päritolu selle geneetilise mitmekesisuse tõttu. (Modifitseeritud Liao *et al*., 2023 järgi).

Seejärel koostati iga inimese kohta diploidne, faasitud genoomijärjestus, mille alusel loodi pangenoomne graafistruktuur (Joonis 9), mis esindab kõiki leitud järjestusvariante. See lähenemine võimaldas tuvastada sadu miljoneid uusi aluspaaride lõike, mida varasemad referentsgenoomid ei sisaldanud. Rohkem kui 99% igast genoomist on kaetud ning selle täpsus on enam kui 99% nii struktuuri kui ka aluspaaride tasemel (Liao *et al*., 2023). Sellised täpsed analüüsid aitavad paremini mõista, kui mitmekesine on tegelikult inimeste geneetiline ülesehitus eri maailma piirkondades. See aitab tuvastada järjestusi ja piirkondi, mida varasemad referentsgenoomid teha ei suutnud. Lisaks aitab see ületada piiranguid, mis on seotud ainult ühe võrdlusgenoomi kasutamisega. Pangenoom annab parema aluse tsentromeersete piirkondade varieeruvuse hindamiseks ning kromosoomijärjestuste võrdlemiseks erinevate indiviidide vahel.



Joonis 9. Referentsjärjestuste (GRCh38 ja CHM13) ja k-meeride joondus pangenoomile. Skeem illustreerib, kuidas erinevad genoomi referentsjärjestused ja lühikesed lugemid paiknevad inimese pangenoomi kontekstis tsentromeeri piirkonnas. Roheline katkestatud joon tähistab GRCh38 referentsgenoomi, mille katkestatud jooned näitavad, et joonduses ollakse ebakindlad. Näiteks chr15 sagedaseim k-meer paikneb GRCh38 assemblis küll tandeemselt, kuid vastas ahela põhiselt võrreldes CHM13-ga. NGS viitab järgmise põlvkonna sekveneerimisele (*Next-Generation Sequencing*). (Koostatud BioRender.com saidil võttes aluseks Hickey *et al*., 2023 joonise).

2. EKSPERIMENTAALOSA

2.1. Töö eesmärgid

Käesoleva töö põhieesmärkideks oli:

1. analüüsida CHM13 referentsgenoomi alusel 15. kromosoomi tsentromeere ja tuvastada selle alusel sagedaseim k-meer;
2. koostada 15. kromosoomi tsentromeeride põhjal fülogeneetiline puu, kasutades unikaalse lokalisatsiooniga SNP alleele, mis asuvad sagedaseima k-meeri asukohtade vahel kõigis 47 indiviidis.

Selleks, et täita töö eesmärke:

- eraldati CHM13 v2.0 referentsgenoomist annotatsioonifaili põhjal tsentromeeripiirkond ja sellest genereeriti k-meeride glistid, kasutades GenomeTester4 tööriista *glistmaker*;
- tuvastati genoomi indekseeritud kujul esinemissageduste põhjal kõige sagedasem k-meer 15. kromosoomi tsentromeeris;
- kontrolliti valitud k-meeride esinemist 47 pangenoomi indiviidi kahelt haplotüübilt pärit järjestustes, kasutades *glistcompare* ja *glistquery* tööriistu;
- loodi tsentromeeriga seotud SNP-de põhjal sagedusmaatriks, mille alusel koostati R-programmis klasterdendrogramm (funktsiooniga *hclust*), mis kujutas indiviididevahelist järjestuslikku sarnasust.
- koostati fülogeneetiline puu k-meeride sagedusandmete põhjal, kaasates CHM13 võrdlusgenoomi.

2.2. Materjal ja metoodika

2.2.1. Referentsgenoom CHM13v2.0

Tsentromeeripiirkonna järjestused pärinesid telomeerist-telomeerini koostatud inimese referentsgenoomist CHM13v2.0, mis on esimene täielik ja katkematu lineaarne inimese genoom (Joonis 9). See referents on eriline, kuna sisaldab piirkondi, mis varasemates versioonides olid kas täielikult puudu või ainult osaliselt järjestatud, näiteks tsentromeerid, telomeerid ja muud kordusjärjestused. CHM13v2.0 genoom põhineb indiviidi haploidsel rakuliinil (CHM13hTERT), mis võimaldas ühtlast ning täpset järjestamist ilma homoloogsete kromosoomide segunemise riskita (Nurk *et al*., 2022). CHM ehk täielikult hüdatidiformne mool (*complete hydatidiform mole*) tähistab haruldast inimese embrüot, mille kogu genoom on päritud isalt, jättes välja emapoolse geneetilise materjali (Liu ja Hao, 2023). Selline viis aitab lihtsustada järjestuste joondamist (*alignment*) ja vähendada analüüsivigu emapoolsete alleelide puudumise tõttu (Nurk *et al*., 2022). CHM13 genoomi kasutatakse referentsina, kuna see pakub väga puhast geneetilist sisu võrreldes tavapärase „müraga", mis tuleneb kahe vanema DNA segunemisest. See teeb CHM13 genoomist eriti kasuliku tööriista genoomi struktuuri ja varieeruvuse uurimisel, võimaldades paremat ligipääsu kordusjärjestustele ja struktuursetele variantidele ning vältides tõlgendamisprobleeme, mis on seotud heterosügootsusega. See referents on oluliselt parandanud genoomi kvaliteeti, eriti keeruliste piirkondade, nagu tsentromeerid, uurimisel, kus varasemates genoomiversioonides esines lünki või vigu kordusstruktuuride tõttu. Siinses töös kasutati CHM13v2.0 referentsgenoomi (GCF_009914755.1), mis on T2T konsortsiumi ametlik versioon ja allalaaditud NCBI andmekataloogist.

2.2.2. 47 indiviidi järjestused inimese pangenoomi andmestikust

Inimese pangenoomi esimese etapi andmestikus kasutati 47 indiviidi genoomi, millest igaühe kohta olid kättesaadavad mõlemalt vanemalt (*maternal* ja *paternal*) päritud assemblid https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/. Vastavad failid asusid kataloogis kujul: assemblies/year1_freeze_assembly_v2/.maternal.f1_assembly_v2.fa.gz või .paternal.f1_assembly_v2.fa.gz. Ühelt vanemalt päritud kromosoomide põhjal moodustatud assemblite kogupikkus oli ligikaudu 3 miljardit aluspaari.

2.2.3. Töö läbiviimiseks vajaliku kromosoomi valik

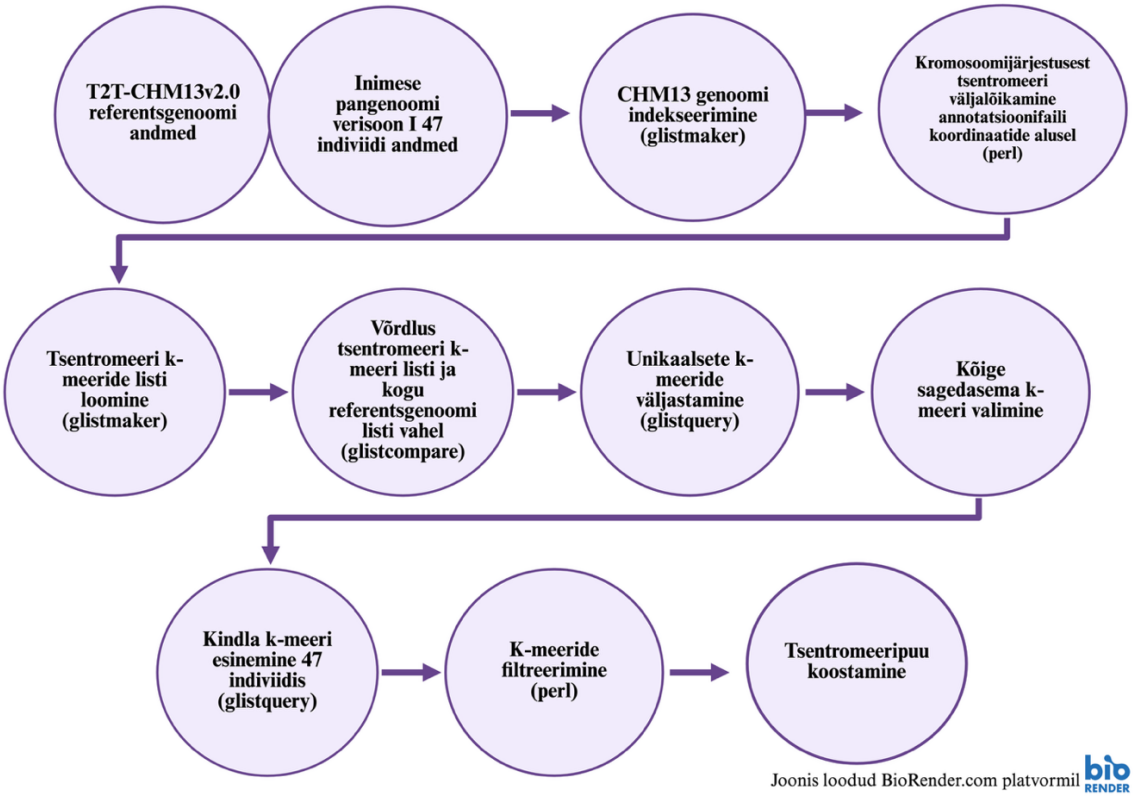
Töö viidi läbi autori poolt valitud 15. kromosoomi põhiselt. Valik tulenes teadmisest, et 15. kromosoomiga on seotud mitmed geneetilised häired nagu Prader-Willi sündroom ja Angelmani sündroom (Lisa 1) (National Center for Biotechnology Information, 1998). Need haigused on geneetiliselt omavahel seotud, kuid fenotüübiliselt väga erinevad, mistõttu pakub kromosoomi struktuurne varieeruvus huvitavat uurimismaterjali.

2.2.4. Töö käik

Kõik käsureal käivitatavad skriptid ja analüüsid viidi läbi MacOS terminali kasutades. Töös kasutati GenomeTester4 tarkvarapaketti, et hallata ja analüüsida k-meeride esinemist genoomijärjestustes (Kaplinski *et al*., 2015). K-meerid on lühikesed DNA järjestused, mille nukleotiididest koosnevat pikkust (k) saab valida vastavalt vajadusele (Jenike *et al*., 2025). Nende kohandamise võimaluse tõttu on need kasulikud näiteks tsentromeeride uurimisel, kus järjestused on korduvad. T2T-CHM13v2.0 referentsgenoomist eraldati satelliitjärjestuste annotatsioonifaili koordinaatide alusel tsentromeerid. Annotatsioonifail sisaldas tsentromeeripiirkondade täpseid asukohti, mille põhjal isoleeriti analüüsi aluseks olevad järjestused (Joonis 10).

Eraldatud järjestuste põhjal loodi GenomeTester4 *glistmaker*-funktsiooniga binaarsed k-meeride loendid (glistid), mis sisaldavad kompaktseid k-meeride kogumeid tsentromeeripiirkonnast. Seejärel indekseeriti CHM13 põhjal loodud glistid, et määrata iga k-meeri esinemissagedus ja asukoht referentsgenoomis. Sellest tulenevalt oli võimalik hinnata, millised k-meerid on tsentromeerile iseloomulikud ning uurida nende esinemissagedust.

Tsentromeeri analüüsi aluseks valiti kõige sagedasem k-meer, mis esines CHM13 referentsgenoomis. Selle põhjal võrreldi individuaalsete järjestuste gliste referentsist pärineva valitud k-meeriga, kasutades glistcompare tööriista. Lisaks võimaldas glistquery-funktsioon tuvastada, millistes indiviidide põhistes assemblites antud k-meer esineb ja millistes mitte. Kuna kõik k-meerid ei pruugi referentsgenoomis üldse esineda, oli vajalik hinnata nende olemasolu otse individuaalsetes järjestustes.



Joonis 10. Tsentromeeripuu koostamise etapid. Analüüs algas CHM13v2.0 genoomi ja pangenoomi individuaalsete järjestuste allalaadimise ning referentsgenoomi 25-meeride kaupa indekseerimisega (*glistmaker*). CHM13 annotatsioonifaili põhjal eraldati 15. kromosoomi tsentromeeripiirkond, millest loodi samuti k-meeride list. Seda listi võrreldi kogu genoomi indeksiga (*glistcompare*), et identifitseerida tsentromeerile iseloomulikke ja mujal genoomis mitteesinevaid k-meere. Leitud unikaalsed k-meerid väljastati tekstifailina (*glistquery*), mille järel valiti välja suurima esinemissagedusega k-meer, mida kasutati markerina edasises analüüsis. Seejärel kontrolliti selle k-meeri esinemist kõigis indiviidis ning andmed filtreeriti Perli skriptiga. Lõpuks kasutati skaleeritud andmeid R-keskkonnas tsentromeeripõhise fülogeneetilise puu koostamiseks. (Koostatud BioRender.com saidil autori poolt).

2.2.5. Sagedaseima k-meeri tuvastamine CHM13 referentsgenoomis

Esmalt kasutati indekseerimiseks T2T-CHM13v2.0 genoomijärjestust (Homo Sapiens Genome Assembly T2T-CHM13v2.0, n.d.), mis oli kirjeldatud kui chm13.draft_v2.0.fasta. K-meeride pikkuseks valiti 25 aluspaari. Indekseerimiseks kasutati glistmaker-funktsiooni järgmise käsuga:

glistmaker chm13.draft_v2.0.fasta -w 25 --index -o chm13.draft_v2.0

Indeksfaili loomisel salvestati iga k-meeri esinemiskoht ning sagedus referentsgenoomis, mis võimaldas hiljem nende põhjalikku võrdlust ja selektsiooni. Indeks koos kromosoomipõhiste listidega paiknes kataloogis t2t-chm13-v2.0/INDEX_25. Selles on kirjas k-meeride asukohad.

Seejärel tuli eraldada 15. kromosoomi tsentromeeripiirkond, mille koordinaadid saadi vastava annotatsioonifaili põhjal https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_censat_v2.1.bed. Kasutades Perl-skripti, lõigati välja soovitud järjestusfail:

perl loika_fasta_failist.pl t2t-chm13-v2.0/chrs/chr15.fna 16678794 17694466 > tsentromeer_chr15_katariina.fa

Tsentromeerist eraldatud järjestusest genereeriti 25-mreride glist järgmise käsuga:

GenomeTester4/src/glistmaker tsentromeer_chr15.fa -w 25 -o tsentromeer_chr15

Seejärel võrreldi tsentromeeripiirkonnast genereeritud k-meeride listi kogu CHM13 referentsgenoomi indeksiga glistcompare-funktsiooni abil. Võrdlus viidi läbi, et identifitseerida need k-meerid, mis esinevad ainult konkreetse kromosoomi tsentromeeris, mitte mujal genoomis:

GenomeTester4/src/glistcompare tsentromeer_chr15_25.list t2t-chm13-v2.0/chm13.draft_v2.0_25.index -du -o tsentromeer_chr15

Pärast võrdlust kasutati glistquery-funktsiooni, et väljastada tekstifail kujul .txt, mis sisaldas ainult neid k-meere, mis olid unikaalsed kromosoom 15 tsentromeeripiirkonnale:

GenomeTester4/src/glistquery tsentromeer_chr15_25_0_diff1.list > ainult_tsentrosoom_chr15_k-merid_katariina.txt

Tulemuseks saadud tekstifail sisaldas kahte tulpa: esimene k-meeri järjestus ning teine selle esinemissagedus antud piirkonnas. Faili nimetuses esinev tsentrosoom ei viita otseselt sellele, vaid on autori poolt valitud failide lihtsama eristamise eesmärgil. Kõige sagedasema k-meeri määramiseks sorteeriti tulemused sageduse järgi kahanevas järjekorras ja valiti välja kõrgeima väärtusega rida, kasutades järgmise käsureakäsu kombinatsiooni:

sort -k 2 -n -r ainult_tsentrosoom_chr15_k-merid_katariina.txt | head -1

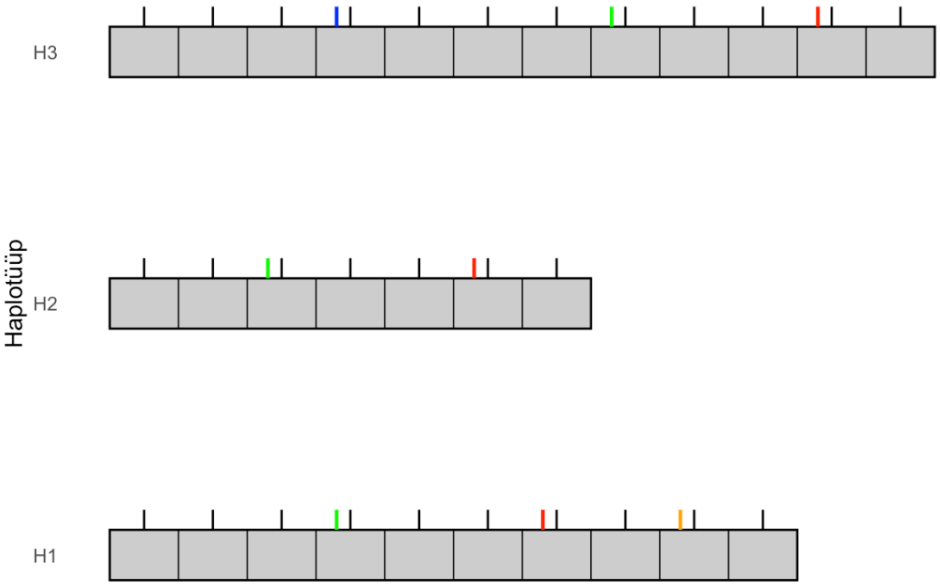
Selle käsu tulemusel saadud k-meer valiti edasise analüüsi markeriks ning kasutati teiste individuaalsete genoomide võrdluses ja fülogeneetilise puu konstrueerimisel.

2.2.6. Tsentromeeripuu loomine

Tsentromeeripuu loomiseks vajaliku sisendfaili koostamisel kasutati Perli skripti proov_01.pl, mille eesmärk oli filtreerida välja ainult need k-meerid, mis esinesid vähem kui neljal indiviidil. Skript loeb sisendina nimed.txt faili, mis sisaldab FASTA järjestuste nimesid, ning loob väljundfaili nimega 15_proov_01_4.txt. Faili nimetuses „15" viitab kasutatud kromosoominumbrile, „proov_01" viitab kasutatud skriptiversioonile ning „4" määrab filtriläve – k-meerid, mis esinesid rohkem kui neljal indiviidil, jäeti välja. Filtreerimine on oluline, kuna liiga sagedased k-meerid, mis esinevad paljudes isendites, võivad tuleneda kordustest ja põhjustada puus moonutusi, näiteks duplikeerunud signaale ja klastrite liitumist. Ilma sellise eelfiltreerimiseta muutuks puu raskesti loetavaks ja vähem informatiivseks, kuna tulemuseks oleks ebatäpselt grupeeritud või liiga „laialivalguv" dendrogramm.

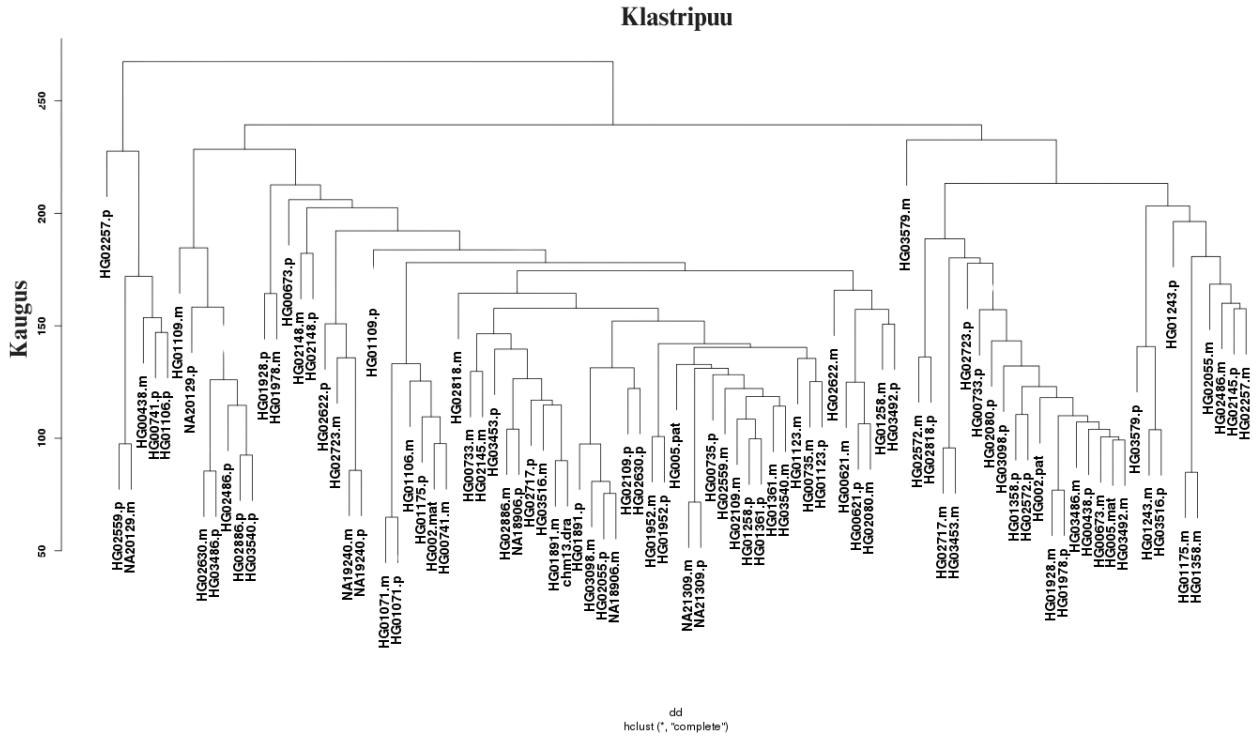
Skript käivitati käsurealt järgmise käsuga: perl proov_01.pl chr_15, kus 15 tähistab analüüsitavat kromosoomi. Skript loeb järjestused ja koostab nende põhjal sagedusmaatriksi, millest genereeritakse k-meeride loend edasiseks kasutamiseks R-skriptis. Selline eeltöötlus tagas, et R-i kaudu tehtud puu põhines ainult informatiivsetel ja eristavatel k-meeridel, mitte kogu andmestikul.

GenomeTester4 abil eraldatud ja indekseeritud k-meeride loendid andsid aluse tsentromeeripuu koostamiseks. Puu tegemiseks valmistati R-skript nimega "proov.R". Valitud kõige sagedasem k-meer toimis analüüsi põhikriteeriumina, mille abil arvutati indiviididevahelised kaugused ja koostati klasterdendrogramm. Töös loeti esmalt sisse andmed failist "15_proov_01_4.txt", kus esimeseks veeruks määrati tsentromeeripiirkonna näidete identifikaatorid. Andmed skaleeriti funktsiooniga scale() ning nende põhjal arvutati indiviididevaheline kaugusmaatriks funktsiooniga dist(). Selle maatriksi alusel rakendati hclust(), mis grupeeris indiviidide tsentromeeripõhised k-meeride esinemismustrid hierarhiliseks dendrogrammiks. Joonistamiseks kasutati graafikaseadet png(), mille tulemusena salvestati tsentromeeripuu failina "15_puu.png". Lisaks on oluline välja tuua, et analüüsi käigus oli eesmärgiks leida k-meerid, mida esineb kogu tsentromeeri ulatuses ainult üks kord. Selleks rakendati täiendavat filtreerimismeetodit, kus loendist eemaldati kõik need k-meerid, mille esinemissagedus ületas ühe ning säilitati vaid unikaalsed k-meerid (Joonis 11). See lähenemine võimaldas keskenduda tsentromeerile iseloomulikele mustritele, toetades edasist SNP-de analüüsi ja võimaldades täpsemalt hinnata indiviididevahelisi genoomilisi erinevusi.



Joonis 11. Lihtsustatud skeem SNP-de paiknemisest tsentromeeri tandeemsete korduste sees. Halli kasti pikkus on 171 bp, viidates tüüpilisele tandmeese korduse pikkusele tsentromeeris. Mustaga on märgistatud enimlevinud k-meer. Punase, oranži, sinise ja rohelisega on tähistatud unikaalse lokalisatsiooniga SNP-d. (Koostatud RStudio programmis autori poolt).

2.3. Tulemused



Joonis 12. 15. kromosoomi fülogeneetiline puu. Joonis väljendab 47 pangenoomi indiviidi ja referentsgenoomi CHM13 kõige sagedasema k-meeri esinemist. Koodide lõpus märgistavad p isapoolset (*paternal*) ja m emapoolset (*maternal*) assemblit. Dendrogrammi Y-telg ehk kaugus (*height*) näitab kahe haru vahelise distantsi suurust, kus pikemad harud tähendavad suuremat k-meeride sageduste erinevust (rohkem unikaalseid k-meere), ehk harude omavaheline kaugus kasvas. Mida madalamal (madalam *height* väärtus) kaks haru puus kokku tulevad, seda rohkem ühiseid k-meere neil on. Puult otsis autor ka homosügootsete indiviidide olemasolu, kelleks olid HG02148, NA19240, HG01071, HG01952, NA21309. (Koostatud R-skriptiga MacOS terminalis autori poolt).

Käesoleva töö põhieesmärgiks oli tsentromeeripuu loomine. Töö käigus eraldati CHM13 referentsgenoomi tsentromeerist kõige sagedasem 25 nukleotiidipikkune k-meer järjestusega AATTTTCAATGCTCTCAAAATATCC, mida esines kokku 731 korda. Kuigi leiti ka mitmeid teisi sagedasi k-meere (Lisa 2), valiti just see juhendajaga kokkuleppeliselt edasiseks analüüsiks, kuna see sobis lühemate järjestuslugemite töötluseks ning tagas piisava sekveneerimiskattuvuse. Seega võib järeldada, et esimene eesmärk saavutati edukalt. Pikemate k-meeride kasutamisel oleks töötlus olnud mahukam ning oluliselt kauem aega võtnud. Teise eesmärgina seati 15. kromosoomi tsentromeeride põhjal fülogeneetilise puu koostamine, kasutades unikaalse lokalisatsiooniga SNP-alleele, mis paiknesid sagedaseima k-meeri asukohtade vahel 47 pangenoomi indiviidis. Analüüsis keskenduti k-meeridele, mis paiknesid ainult üks kord kogu tsentromeeripiirkonna peale ning,

mille vahelised SNP-d võimaldasid koostada informatiivse puustruktuuri. Puu harude pikkus (*height*) vastab eristuvate k-meeride hulgale ehk mida pikem lüli, seda rohkem unikaalseid k-meere. Fülogeneetilises puus kõrvuti paiknevad sama koodiga harud viitavad homosügootsusele (Joonis 12). *Maternal* ja *paternal* haplotüüpide kõrvutiasetus dendrogrammis kinnitab, et ühe indiviidi kaks homoloogset kromosoomi jagavad omavahel peaaegu identset k-meeride profiili. See on ootuspärane, sest mõlemad alleelid pärinevad sama indiviidi 15. kromosoomilt. Puult on vaja need leida aDNA (vana ehk *ancient*) mudeli tegemiseks kui valida on tuhandete inimeste sekveneerimisandmete seast. Mõne indiviidi puhul võis haru paikneda kõrgemal seetõttu, et vastavas piirkonnas võis esineda deleteerunud järjestusi, põhjustades k-meeride puudumise või muutumise. Kui tsentromeer on deleteerunud, puuduvad sealt kõik markerid, mis aitaks fülogeneetilise puu tegemisel. Töö käigus kasutati tsentromeeripuu loomiseks käsureapõhiseid tööriistu (terminal) ja Rscripti, kuna andmestik sisaldas täpselt 15 107 k-meeri, mille maht ületab tavapäraste graafiliste programmide töötlusvõimekuse. See lähenemine tagas protsessi kiiruse, mälusäästlikkuse ning võimaldas tsentromeeripuu koostamise piisava täpsusega. Lisaks võimaldas käsurea kasutamine paremat kontrolli kõikide etappide üle. Teine eesmärk täideti edukalt, sest loodi 15. kromosoomi tsentromeeri põhjal informatiivne ja visuaalselt uuritav fülogeneetiline puu, mis näitas indiviididevahelisi erinevusi tsentromeeripiirkonnas.

2.4. Arutelu

Tsentromeeride analüüs põhineb eeldusel, et need genoomipiirkonnad ei läbi rekombinatsiooni, mistõttu on neis toimuvad mutatsioonid pärilikult stabiilsed ja sobivad hästi evolutsiooniliste suhete jälgimiseks. Varasemad uuringud on näidanud, et tsentromeerid, olles suuresti üles ehitatud tandeemsetest kordustest (McNulty *et al*., 2018, Sullivan *et al*., 2020), on evolutsiooniliselt ebastabiilsed ja raskesti järjestatavad piirkonnad, mille varieeruvus indiviidide vahel on märkimisväärne (Logsdon *et al*., 2024). Just seetõttu on fülogeneetilisi puid koostatud peamiselt tsentromeeri lähialadest (peritsentromeersed regioonid), kus järjestuse stabiilsus on suurem. Samuti on teada, et kordusaladel esineb sagedamini mutatsioone, mis muudavad need potentsiaalselt informatiivseks evolutsioonilise seose hindamisel. Käesoleva töö raames püüti seda lähenemist laiendada, kasutades puu koostamisel esmakordselt andmeid otse tsentromeeripiirkonnast, mis muudab töö olulisel määral võrreldavaks uuema tsentromeerianalüüsi suunaga, mida võimaldab täielikult järjestatud CHM13v2.0 genoom. On aga alust eeldada, et kromosoomide vahelise rekombinatsiooni vähesuse tõttu on tsentromeeri sees ja selle lähialadel

paiknev järjestus sarnase fülogeneetilise mustriga. Tulevikus võiks teha võrdluse nii tsentromeeri kui ka selle ümbritseva piirkonna sarnasustest.

Puust ei saa hetkel vanuselisi järeldusi teha, kuid selle struktuur võimaldas visuaalselt uurida isendite grupeerumist ja sarnasust tsentromeeripiirkonna järjestuse põhjal. 47 pangenoomi indiviidi analüüsi 15 107 k-meeri ei ole fikseeritud arv ehk uuema versioon II alusel, mis sisaldab 232 indiviidi faasitud genoome (Human Pangenome Reference Consortium, n.d.), tuleks koostada sarnane 15. kromosoomi tsentromeeripuu. Suurema andmestikuga on suurenenud geneetiline mitmekesisus ning, mida rohkem on andmeid, seda statistiliselt usaldusväärsemaks muutub analüüs. Tulevikus võiks tsentromeeripuu analüüsi täiendada harupõhiste k-meeride määramisega, mis võimaldaks tuvastada igale harus iseloomulikud järjestused. Selle kaudu saaks tsentromeere rühmitada nii kromosoomi, indiviidi kui ka järjestuslike sarnasuste põhjal. See oleks abiks tsentromeeride evolutsiooni uurimisel ning erinevate indiviidide või populatsioonide võrdlemisel. Lisaks võiks tulevikus täiendada fülogeneetilist puud selgete harunimetuste, numbrite või haplogruppidega, et igal harus oleks kordumatu tähistus. See aitaks kirjeldada, millised isendid moodustavad omavahelise grupi, ning lihtsustaks tulemuste võrdlemist teiste kromosoomide, populatsioonide või võimalike haigusmarkeritega, eriti suuremate andmekogumite analüüsimisel. Hetkel puudub andmestik, millega saaks käesoleva töö tulemusi kinnitada või täpsustada. Selleks võiks kasutada NGS (PacBio või Oxford Nanopore) tehnoloogiate abil saadud pikki lugemeid, mis võimaldaks meil täpsemalt hinnata tsentromeeri kordusjärjestuste, SNP-de asukohtade ja deletsioonipiirkondade korraldust. Need analüüsid võimaldaksid hinnata, kui täpne on kirjeldatud tsentromeeripuu võrreldes tegeliku järjestusliku varieeruvusega. Edasine uurimine saab ka kinnitada, kas samad SNP-d esinevad ainult 15. kromosoomi tsentromeeris või leidub neid ka teistel kromosoomidel. See kinnitaks, et kasutatud markerid on ainulaadsed 15. kromosoomi tsentromeerile ega pärine juhuslikult mõnest muust kromosoomipiirkonnast.

KOKKUVÕTE

Käesoleva bakalaureusetöö eesmärgiks oli analüüsida inimese 15. kromosoomi tsentromeeripiirkonda, leida selle alusel kõige sagedasem k-meer ning kasutada unikaalse lokalisatsiooniga SNP-de vahelisi piirkondi fülogeneetilise puu koostamiseks. Töö teoreetilises osas anti ülevaade inimese kromosoomide ja tsentromeeride struktuurist, nende ülesannetest ning rollist kromosoomide jagunemisel.

Kirjanduse põhjal selgus, et kuigi tsentromeerid koosnevad peamiselt kordusjärjestustest (α-satelliit-DNA), leidub nende monomeerides üksiknukleotiidseid polümorfisme, mida saab kasutada järjestuslike erinevuste tuvastamiseks indiviidide vahel (Logsdon *et al*., 2024; McNulty *et al*., 2018; Sullivan *et al*., 2020). Fülogeneetilise puu konstrueerimise aluseks on oletus, et tsentromeerid ei läbi rekombinatsiooni, mistõttu on seal toimuvad mutatsioonid päritavad ning sobivad hästi evolutsiooniliste suhete uurimiseks (Kapli *et al*., 2020).

Praktilises osas tuvastati CHM13v2.0 referentsgenoomi põhjal kõige sagedasem 25-meer, mis esines ainult tsentromeeris. Selle k-meeri esinemise ja sellevaheliste SNP-de põhjal koostati 47 pangenoomi indiviidi fülogeneetiline puu. Tulemused näitasid, et indiviidide haplotüübid grupeerusid loogiliselt ning osad sama isiku ema- ja isapoolsed harud paiknesid lähestikku, viidates homosügootsusele. Mõned indiviidid erinesid teistest deleteerunud järjestuste tõttu, mis vähendas k-meeride esinemist.

Töö kinnitas, et tsentromeeride piires leiduvad järjestuslikud erinevused võimaldavad koostada informatiivseid fülogeneetilisi puid. CHM13 genoomi kasutamine referentsina osutus sobivaks, kuna see välistas heterosügootsusest tingitud määramatuse ja võimaldas keskenduda kordusjärjestuste täpsele hindamisele. Töö tulemused vastasid püstitatud eesmärkidele ning annavad aluse edasisteks uurimusteks tsentromeeride evolutsiooni ja funktsionaalse varieeruvuse vallas. Tulevikus võiks sarnase analüüsi viia läbi laiendatud pangenoomi (nt 232 indiviidi) põhjal ning täiendada fülogeneetilist puud selgemate harutähiste ja haplogruppide määramisega.

RESÜMEE / SUMMARY

Construction of the human chromosome 15 centromere phylogenetic tree based on individuals from Human Pangenome version I

Katariina Ilmoja

Summary

Centromeres are highly specialized chromosomal regions essential for proper segregation of chromosomes upon cell division. Centromeres are composed predominantly of repetitive DNA repeats such as alpha-satellite DNA in higher-order repeat (HOR) arrays. Improved long-read sequencing and availability of completed human genome assemblies such as CHM13 of the T2T (Telomere-to-Telomere) Consortium have made it possible for scientists to study the sequence-level variability and structure organization of centromeres with unprecedented precision.

This thesis dealt with the k-mer sequence analysis of human chromosome 15 centromere and the construction of a phylogenetic tree based on k-mer profiles and SNP variants. The primary objective was to identify the most frequent 25-mer in the CHM13 reference genome centromere and check if SNPs in these characteristic k-mer sites were able to track lineage relationships between 47 phased genomes within the Human Pangenome Reference Consortium.

Theoretical section of the thesis provided an overview of structure, function, and evolution of centromeres with the intricacies involved due to their repetitive nature. From the literature, it was resolved that even though they are not subject to recombination along with fast evolution, they still accumulate informative mutations such as SNPs in satellite repeats which are utilized to reconstruct phylogenetic relationships (Kapli *et al*., 2020; Logsdon *et al*., 2024). The concept that centromeric areas are inherited without recombination makes them suitable for evolutionary analysis.

The most frequent centromeric 25-mer for chromosome 15 was found by the analysis and used as an anchor in constructing the tree of phylogeny. The tree was constructed using a custom k-mer-based workflow in R and at the command line, with haplotypes for the same individual (paternal and maternal) clustering together, showing high sequence similarity and shared homozygosity in the region. Some had the evidence of erased centromeric regions, influencing clustering and completeness of their k-mer profiles.

The results demonstrate that even within the most repetitive and most structured regions of the genome such as centromeres, certain sequence variants can provide useful phylogenetic signals. The work confirms once again that the human centromere—especially with the assistance of whole-genome references such as CHM13—is amenable to high-resolution analysis. Future studies might further enhance the knowledge regarding centromeric evolution and population-level diversity using more pangenomic datasets, more centromeres, and improved tree annotation (e.g., haplogroup assignment).

TÄNUSÕNAD

Suurimad tänud minu juhendajale Tarmo Puurandile asjalike nõuannete, rohkete teadmiste jagamise ning järjekindla toe eest kogu töö vältel. Samuti soovin tänada Tartu Ülikooli bioinformaatika õppetooli meeldiva ja toetava keskkonna eest.

KIRJANDUSE LOETELU

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002a). *An overview of M phase.* Molecular Biology of the Cell - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK26931/

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002b). *An overview of the cell cycle.* Molecular Biology of the Cell - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK26869/

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002c). *Meiosis*. Molecular Biology of the Cell - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK26840/

Allshire, R. C., & Karpen, G. H. (2008). Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nature Reviews Genetics*, *9*(12), 923–937. https://doi.org/10.1038/nrg2466

Aly, H. (2023). Understanding cell Cycle: Regulation and Significance. *Longdom*. https://doi.org/10.35248/2332-0737.23.11.047

Amor, D. J., & Choo, K. A. (2002). Neocentromeres: Role in Human Disease, Evolution, and centromere study. *The American Journal of Human Genetics, 71*(4), 695–714. https://doi.org/10.1086/342730

Ando, S., Yang, H., Nozaki, N., Okazaki, T., & Yoda, K. (2002). CENP-A, -B, and -C chromatin complex that contains the I-type alpha-satellite array constitutes the prekinetochore in HeLa cells. *Molecular and cellular biology, 22*(7), 2229–2241. https://doi.org/10.1128/MCB.22.7.2229-2241.2002

Bakhoum, S. F., Kabeche, L., Murnane, J. P., Zaki, B. I., & Compton, D. A. (2014). DNA-damage response during mitosis induces whole-chromosome missegregation. *Cancer discovery, 4*(11), 1281–1289. https://doi.org/10.1158/2159-8290.CD-14-0403

Bellou, E., Zielinska, A. P., Mönnich, E. U., Schweizer, N., Politi, A. Z., Wellecke, A., Sibold, C., Tandler-Schneider, A., & Schuh, M. (2024). Chromosome architecture and low cohesion bias acrocentric chromosomes towards aneuploidy during mammalian meiosis. *Nature Communications, 15*(1). https://doi.org/10.1038/s41467-024-54659-3

Blackburn E. H. (2001). Switching and signaling at the telomere. *Cell, 106*(6), 661–673. https://doi.org/10.1016/s0092-8674(01)00492-5

Brooker, A. S., & Berkowitz, K. M. (2014). The roles of cohesins in mitosis, meiosis, and human health and disease. *Methods in Molecular Biology*, 229–266. https://doi.org/10.1007/978-1-4939-0888-2_11

Brown, T. A. (2002). Genomes, 2nd edition. *National Center for Biotechnology Information*. https://www.ncbi.nlm.nih.gov/books/NBK21134/

Butler, M. G., & Duis, J. (2020). Chromosome 15 Imprinting Disorders: Genetic Laboratory Methodology and Approaches. *Frontiers in Pediatrics, 8*. https://doi.org/10.3389/fped.2020.00154

Chardon, F., Japaridze, A., Witt, H., Velikovsky, L., Chakraborty, C., Wilhelm, T., Dumont, M., Yang, W., Kikuti, C., Gangnard, S., Mace, A., Wuite, G., Dekker, C., & Fachinetti, D. (2022). CENP-B-mediated DNA loops regulate activity and stability of human centromeres. *Molecular Cell, 82*(9), 1751-1767.e8. https://doi.org/10.1016/j.molcel.2022.02.032

Dumont, M., Gamba, R., Gestraud, P., Klaasen, S., Worrall, J. T., De Vries, S. G., Boudreau, V., Salinas-Luypaert, C., Maddox, P. S., Lens, S. M., Kops, G. J., McClelland, S. E., Miga, K. H., & Fachinetti, D. (2020). Human chromosome-specific aneuploidy is influenced by DNA-dependent centromeric features. *The EMBO journal, 39*(2), e102924. https://doi.org/10.15252/embj.2019102924

Di Tommaso, E., & Giunta, S. (2024). Dynamic interplay between human alpha-satellite DNA structure and centromere functions. *Seminars in Cell & Developmental Biology*, *156*, 130–140. https://doi-org.ezproxy.utlib.ut.ee/10.1016/j.semcdb.2023.10.002

French, B. T., & Straight, A. F. (2013). Swapping CENP-A at the centromere. *Nature Cell Biology, 15*(9), 1028–1030. https://doi.org/10.1038/ncb2833

Fukagawa, T., & Earnshaw, W. C. (2014). The centromere: chromatin foundation for the kinetochore machinery. *Developmental cell, 30*(5), 496–508. https://doi.org/10.1016/j.devcel.2014.08.016

Gamba, R., & Fachinetti, D. (2020). From evolution to function: Two sides of the same CENP-B coin?. *Experimental cell research, 390*(2), 111959. https://doi.org/10.1016/j.yexcr.2020.111959

Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. A., & Ashley, E. A. (2017). Human Genome Sequencing at the Population scale: A primer on High-Throughput DNA Sequencing and Analysis. *American Journal of Epidemiology, 186*(8), 1000–1009. https://doi.org/10.1093/aje/kww224

Gottlieb, S. F., Gulani, A., & Tegay, D. H. (2023, August 14). *Genetics, meiosis*. StatPearls - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK482462/#:~:text=During%20telophase%20I%2C%20the%20cytoplasm,cell%20and%20another%20polar%20body.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., . . . Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science, 328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Harris, P., Boyd, E., Young, B. D., & Ferguson-Smith, M. A. (1986). Determination of the DNA content of human chromosomes by flow cytometry. *Cytogenetics and cell genetics*, *41*(1), 14–21. https://doi.org/10.1159/000132190

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Buonaiuto, S., Carroll, A., Chaisson, M. J. P., Chang, P., Chang, X. H., . . . Paten, B. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology, 42*(4), 663–673. https://doi.org/10.1038/s41587-023-01793-w

Human Pangenome Reference Consortium. (n.d.). *HPRC Data Release 2*. https://humanpangenome.org/hprc-data-release-2/

Jenike, K. M., Campos-Domínguez, L., Boddé, M., Cerca, J., Hodson, C. N., Schatz, M. C., & Jaron, K. S. (2025). k-mer approaches for biodiversity genomics. *Genome Research*. https://doi.org/10.1101/gr.279452.124

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics, 21*(7), 428–444. https://doi.org/10.1038/s41576-020-0233-0

Kaplinski, L., Lepamets, M., & Remm, M. (2015). GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience, 4*(1). https://doi.org/10.1186/s13742-015-0097-y

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., . . . Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860–921. https://doi.org/10.1038/35057062

Liao, W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., . . . Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Liu, Y., & Hao, P. (2023). Complete hydatidiform mole. *New England Journal of Medicine, 389*(18), 1708. https://doi.org/10.1056/nejmicm2304005

Logsdon, G. A., Rozanski, A. N., Ryabov, F., Potapova, T., Shepelev, V. A., Catacchio, C. R., Porubsky, D., Mao, Y., Yoo, D., Rautiainen, M., Koren, S., Nurk, S., Lucas, J. K., Hoekzema, K., Munson, K. M., Gerton, J. L., Phillippy, A. M., Ventura, M., Alexandrov, I. A., & Eichler, E. E. (2024). The variation and evolution of complete human centromeres. *Nature*, *629*, 136–145. https://doi.org/10.1038/s41586-024-07278-3

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S., Mercuri, L., Dishuck, P. C., Rhie, A., De Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., . .

. Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107. https://doi.org/10.1038/s41586-021-03420-7

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature, 389*(6648), 251–260. https://doi.org/10.1038/38444

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., . . . Reich, D. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature, 538*(7624), 201–206. https://doi.org/10.1038/nature18964

Mastrorosa, F. K., Rozanski, A. N., Harvey, W. T., Knuth, J., Garcia, G., Munson, K. M., Hoekzema, K., Logsdon, G. A., & Eichler, E. E. (2024). Complete chromosome 21 centromere sequences from a Down syndrome family reveal size asymmetry and differences in kinetochore attachment. *bioRxiv: the preprint server for biology*, 2024.02.25.581464. https://doi.org/10.1101/2024.02.25.581464

Masumoto, H., Masukata, H., Muro, Y., Nozaki, N., & Okazaki, T. (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *The Journal of cell biology, 109*(5), 1963–1973. https://doi.org/10.1083/jcb.109.5.1963

McKinley, K. L., & Cheeseman, I. M. (2015). The molecular basis for centromere identity and function. *Nature Reviews Molecular Cell Biology, 17*(1), 16–29. https://doi.org/10.1038/nrm.2015.5

McNulty, S. M., & Sullivan, B. A. (2018). Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome research: an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *26*(3), 115–138. https://doi.org/10.1007/s10577-018-9582-3

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell*, *128*(4), 787–800. https://doi.org/10.1016/j.cell.2007.01.028

Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G. A., & Georgakopoulos-Soares, I. (2024). A survey of K-Mer Methods and Applications in bioinformatics. *Computational and Structural Biotechnology Journal, 23*, 2289–2303. https://doi.org/10.1016/j.csbj.2024.05.025

National Center for Biotechnology Information (US). (1998). *Chromosome map*. Genes and Disease - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK22266/

National Human Genome Research Institute (NHGRI). (2019, March 9). *Chromosomes fact sheet*. Genome.gov. https://www.genome.gov/about-genomics/fact-sheets/Chromosomes-Fact-Sheet#:~:text=is%20a%20chromosome%3F-,Chromosomes%20are%20thread%2Dlike%20structures%20located%20inside%20the%20nucleus%20of,type%20of%20living%20creature%20unique.

Niimura, Y., & Gojobori, T. (2002). In silico chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence. *Proceedings of the National Academy of Sciences, 99*(2), 797–802. https://doi.org/10.1073/pnas.022437999

Olins, D. E., & Olins, A. L. (2003). Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology, 4*(10), 809–814. https://doi.org/10.1038/nrm1225

Pathak, I., & Bordoni, B. (2023). *Genetics, Chromosomes*. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK557784/

Pavlichin, D. S., Lee, H., Greer, S. U., Grimes, S. M., Weissman, T., & Ji, H. P. (2022). KmerKeys: a web resource for searching indexed genome assemblies and variants. *Nucleic Acids Research, 50*(W1), W448–W453. https://doi.org/10.1093/nar/gkac266

Pesenti, M. E., Raisch, T., Conti, D., Walstein, K., Hoffmann, I., Vogt, D., Prumbaum, D., Vetter, I. R., Raunser, S., & Musacchio, A. (2022). Structure of the human inner kinetochore CCAN complex and its significance for human centromere organization. *Molecular Cell, 82*(11), 2113-2131.e8. https://doi.org/10.1016/j.molcel.2022.04.027

Santaguida, S., & Amon, A. (2015). Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat Rev Mol Cell Biol, 16*, 473–485. https://doi.org/10.1038/nrm4025

Sullivan, L. L., Chew, K., & Sullivan, B. A. (2017). α satellite DNA variation and function of the human centromere. *Nucleus (1949-1034)*, *8*(4), 331–339. https://doi-org.ezproxy.utlib.ut.ee/10.1080/19491034.2017.1308989

Sullivan, L. L., & Sullivan, B. A. (2020). Genomic and functional variation of human centromeres. *Experimental Cell Research, 389*(2), N.PAG. https://doi-org.ezproxy.utlib.ut.ee/10.1016/j.yexcr.2020.111896

Suzuki, Y., Myers, E. W., & Morishita, S. (2020). Rapid and ongoing evolution of repetitive sequence structures in human centromeres. *Science Advances, 6*(50). https://doi.org/10.1126/sciadv.abd9230

Tan, C. H., Gasic, I., Huber-Reggi, S. P., Dudka, D., Barisic, M., Maiato, H., & Meraldi, P. (2015). The equatorial position of the metaphase plate ensures symmetric cell divisions. *eLife, 4*. https://doi.org/10.7554/elife.05124

Tram, U., Ferree, P. M., & Sullivan, W. (2003). Identification of Wolbachia--host interacting factors through cytological analysis. *Microbes and infection, 5*(11), 999–1011. https://doi.org/10.1016/s1286-4579(03)00192-8

KASUTATUD VEEBIAADRESSID

GenomeTester4. (n.d.). GenomeTester4 software by bioinfo-ut. Vaadatud 21.03.2025, https://github.com/bioinfo-ut/GenomeTester4

Human Pangenome Reference Consortium. (n.d.). *Centromere annotation file for T2T-CHM13v2.0 assembly (censat_v2.1.bed)*. Vaadatud 21.03.2025, https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_censat_v2.1.bed.

Human Pangenome Reference Consortium. (n.d.). *Human Pangenome sequencing data – working directory*. Vaadatud 21.03.2025, https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/

Kaplinski, L., Lepamets, M., & Remm, M. (2015). *GenomeTester4* [Tarkvarapakett]. Tartu Ülikooli bioinformaatika töörühm. Vaadatud 21.03.2025, https://github.com/bioinfo-ut/GenomeTester4

National Center for Biotechnology Information (US). (1998). *Chromosome map*. Genes and Disease - NCBI Bookshelf. Vaadatud 19.03.2025, https://www.ncbi.nlm.nih.gov/books/NBK22266/

National Human Genome Research Institute (NHGRI). (2025). *Autosome*. Genome.gov. Vaadatud 12.05.2025, https://www.genome.gov/genetics-glossary/Autosome

National Human Genome Research Institute (NHGRI). (2025). *Cell cycle*. Genome.gov. Vaadatud 13.05.2025, https://www.genome.gov/genetics-glossary/Cell-Cycle

National Human Genome Research Institute (NHGRI). (2025). *Chromosome*. Genome.gov. Vaadatud 13.05.2025, https://www.genome.gov/genetics-glossary/Chromosome

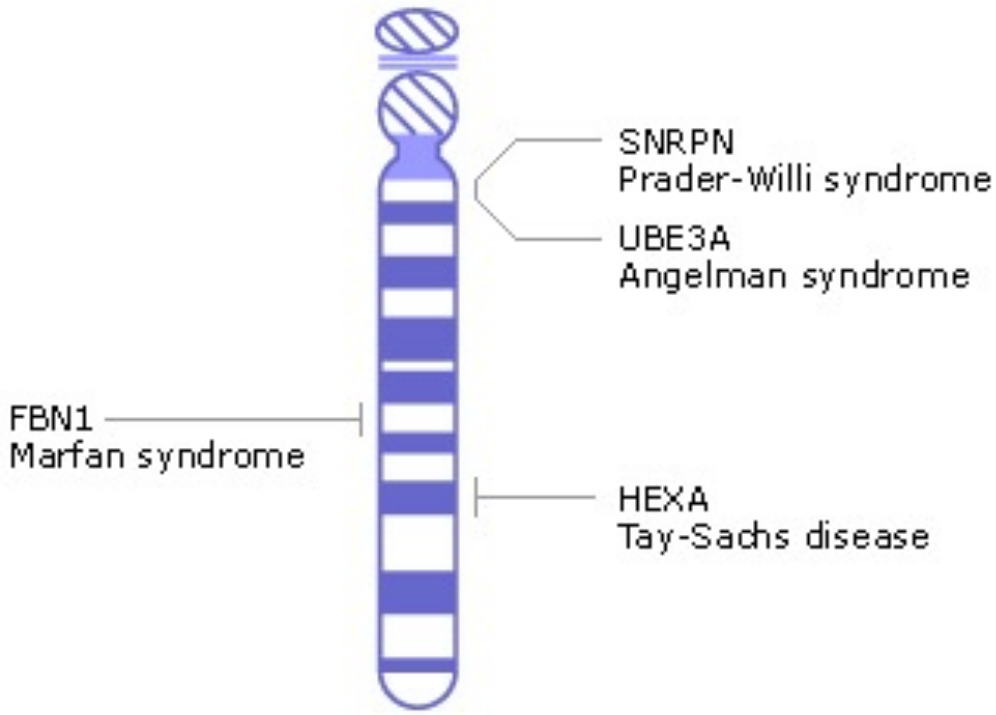
National Human Genome Research Institute (NHGRI). (2025). *Meiosis*. Genome.gov. Vaadatud 03.05.2025, https://www.genome.gov/genetics-glossary/Meiosis

National Library of Medicine (n.d.). *Homo sapiens genome assembly T2T-CHM13v2.0*. NCBI. Vaadatud 21.03.2025, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/

PubMed Central (2020). *Chromosome structure and function*. Vaadatud 14.05.2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC7732198/

LISAD

Lisa 1



Joonis 14. Kromosoom 15 ja sellega seotud geneetilised haigused. Joonisel on kujutatud nelja erineva geneetilise seisundiga seotud geene chr15-s: SNRPN (Prader-Willi sündroom), UBE3A (Angelmani sündroom), FBN1 (Marfani sündroom) ja HEXA (Tay-Sachsi tõbi). Need haigused on seotud erinevate mutatsioonide või geenide ekspressioonihäiretega kromosoomi eri piirkondades (Joonis võetud allikast National Center for Biotechnology Information (US), 1998).

Lisa 2

Tabel 1. Esimesed 10 kõige sagedasemat 25 nukleotiidist koosnev k-meeri (25-meer). Tabelis on toodud välja k-meeri järjestus ning paremal pool selle esinemissagedus CHM13 referentsgenoomi tsentromeerides.

K-meeri järjestus	K-meeride hulk
AATTTTCAATGCTCTCAAAATATCC	731
AAATTTTCAATGCTCTCAAAATATC	731
ATATTTTGAGAGCATTGAAAATTTC	730
TAAGGTGAGAAAGGAAATGTCTTCA	728
CTAAGGTGAGAAAGGAAATGTCTTC	728
AGGTGAGAAAGGAAATGTCTTCAAA	728
AGAAAGGAAATGTCTTCAAATAAGA	728
AAGGTGAGAAAGGAAATGTCTTCAA	728
CTTATTTGAAGACATTTCCTTTCTC	727
GAAAGGAAATGTCTTCAAATAAGAA	726

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Katariina Ilmoja,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Inimese 15. kromosoomi tsentromeeripuu konstrueerimine inimese pangenoomi versioon I indiviidide põhjal “, mille juhendaja on MSc Tarmo Puurand, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Katariina Ilmoja

26.05.2025