TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA UURIMISGRUPP

Kromosoomipõhise k-meer sageduse ja tsentromeeri pikkuse vahelise korrelatsiooni leidmine

Bakalaureusetöö
12 EAP

Carmen Beljaev

Juhendaja
MSc Tarmo Puurand

TARTU 2025

Infoleht

Kromosoomipõhise k-meer sageduse ja tsentromeeri pikkuse vahelise korrelatsiooni leidmine

Tsentromeerid on inimese genoomi kriitilised piirkonnad, millel on oluline roll kromosoomide stabiilsuses ja jagunemises. Nende piirkondade uurimist on pikalt takistanud sekveneerimistehnoloogiate piirangud, kuid uued pika ja ülipika lugemiga tehnoloogiad on avanud võimaluse nende struktuuri täpsemaks analüüsiks. Üheks kasulikuks meetodiks on k-meer põhine sagedusanalüüs, mis võimaldab hinnata korduspiirkondade ulatust ning varieeruvust.

Käesoleva töö eesmärk on uurida, kas inimese kromosoomide tsentromeeride pikkuse ja k-meeride sageduse vahel esineb seos. K-meer põhine analüüs võimaldab hinnata korduspiirkondade struktuuri ja potentsiaalselt ka nende pikkust. Töö tulemused aitavad hinnata selle meetodi sobivust tsentromeeri omaduste kaudseks määramiseks.

Märksõnad: inimene, tsentromeeri pikkus, haplotüüp, k-meer

CERCS kood – B110 Bioinformaatika, meditsiiniiformaatika, biomatemaatika ja biomeetrika.

Finding correlation between chromosome-specific k-mer frequency and centromere length

Centromeres are critical regions of the human genome, playing an essential role in chromosome stability and segregation. For a long time, the study on these regions was hindered by limitations in sequencing technologies. However, recent advances in long-read and ultra-long-read sequencing have enabled more precise structural analysis of centromeres. One useful method is k-mer-based frequency analysis, which allows for the assessment of the extent and variability of repetitive regions.

The aim of this study is to investigate whether there is a correlation between the length of human chromosome centromeres and k-mer frequency. K-mer-based analysis enables the evaluation of the structure of repetitive regions and potentially also their length. The results of this work help to assess the suitability of this method for the indirect estimation of centromere characteristics.

Keywords: human, centromere length, haplotype, k-mer

CERCS code – B110 Bioinformatics, medical informatics, biomathematics and biometrics.

SISUKORD

Kasutatud lühendid 4
Sissejuhatus 5
1. Kirjanduse ülevaade 6
1.1. Tsentromeeri struktuur 6
1.1.1. Klassifikatsioon 7
1.1.2. CENP-valgud 8
1.1.3. Rakufaas 9
1.1.4. Tsentromeeri DNA järjestus 11
1.2. Tsentromeeri järjestuse varieeruvus ja tuvastamine 13
1.2.1. Inimese genoomi projekt GRCh38 13
1.2.2. Telomere-To-Telomere konsortsium 14
1.2.3. Pangenoom 15
1.2.4. Pikkade lugemite sekveneerimistehnoloogiad 16
1.2.4.1. PacBio-HiFi 17
1.2.4.2. Oxford Nanopore Technologies 17
1.2.5. K-meerid 18
2. Eksperimentaalosa 21
2.1 Töö eesmärgid 21
2.2 Materjal ja metoodika 21
2.2.1. CHM13, sagedaisema k-meeri leidmine 21
2.2.2. Pangenoomi indiviidid 21
2.2.3. Korrelatsiooni leidmine 22
2.2.4. HG01106 lühikeste lugemite põhise k-mer sageduse leidmine 22
2.3 Tulemused 23
2.4 Arutelu 24
Kokkuvõte 26
Summary 27
Kasutatud kirjandus 28
Kasutatud veebiaadressid 44
Lisad 45
Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks 49

Kasutatud lühendid

1KGP – 1000 Genoomi Projekt (ingl *1000 Genome Project*)

CCAN – püsiv tsentromeerivalkude võrgustik (ingl *Constitutive Centromere Associated Network*)

CenH3^{CENP-A} - tsentromeerispetsiifiline histoon H3 variant, mis asendab klassikalist histooni H3 tsentromeeride nukleosoomides (ingl *centromere-specific histone H3 variant*)

CENP-A – tsentromeerivalk A (ingl *centromere protein A*)

CENP-B – tsentromeerivalk B (ingl *centromere protein B*)

CENP-C – tsentromeerivalk C (ingl *centromere protein C*)

CHM1 – inimese isapoolset päritolu rakuliin, mida on kasutatud referentsgenoomi koostamisel (ingl *a complete hydatidiform mole cell line with a haploid human genome*)

CHM13 – esimene täielik inimese genoomi järjestus (ingl *a complete hydatidiform mole cell line containing a haploid human genome*)

CREST – CREST-sündroom ehk piiratud süsteemse skleroosi vorm, mida iseloomustavad kaltsiumiladestused, Raynaud' fenomen, söögitoruhäired, naha jäikus sõrmedes ja pindmised veresoonte laiendid

GRCh38 – inimese referentsgenoom 38. versioon (ingl *Genome Reference Consortium human build 38*)

HOR – kõrgemat järku kordused (ingl *higher order repeat*)

HPRC – Inimese pangenoomi referentskonsortsium (ingl *Human Pangenome Reference Consortium)*

HSA8 – inimese 8. kromosoom (ingl *human chromosome 8*)

HSAX – inimese X-kromosoom (ingl *human X chromosome*)

ONT – Oxford Nanopoor Tehnoloogia (ingl *Oxford Nanopore Technology*)

PacBio-HiFi – PacBio kõrge täpsusega pikkade lugemite sekveneerimine (ingl *Pacific Biosciences High-Fidelity sequencing*)

SNPs – ühe nukleotiidi polümorfismid (ingl single nucleotide polymorphism)

SV – struktuursed variandid (ingl *structural variation*)

T2T – telomeerist telomeerinini (ingl *Telomere-to-Telomere*)

WGA – terve genoomi amplifikatsioon (ingl *whole genome amblification*)

Sissejuhatus

Tsentromeeride varieeruvuse uurimist on seni takistanud sekveneerimistehnoloogiate piirangud, eelkõige suutmatus tuvastada pikki, asukohaspetsiifilisi mutatsioone, mis on vajalikud haplotüpiseerimiseks ja assambleerimiseks. Kuigi pikkade lugemite tehnoloogia areng algas juba 20 aastat tagasi, toimus oluline läbimurre alles viimase viie aasta jooksul, kui sekveneerimisvigade tase langes 15%-lt 1%-le. See võimaldas genoomide täieliku assambleerimise, mida iseloomustab ka T2T (Telomere-to-Telomere) konsortsiumi nimi.

T2T konsortsiumi töö tulemusena valmis CHM13 – esimene täielikult järjestatud inimese genoom, mis hõlmab ka seni raskesti järjestatavaid piirkondi nagu tsentromeerid ja segmentaalsed duplikatsioonid. Selle kasutamine referentsina parandab järjestusandmete kaardistamise täpsust ja vähendab valepositiivsete leidude hulka (Sergey et al., 2022).

CHM13 ja CHM1 genoomide võrdlus on avanud uusi võimalusi tsentromeeride järjestus- ja struktuurilise varieeruvuse mõistmiseks. Leitud on suuri erinevusi α-satelliitide kordusstruktuurides (HOR-id), nende pikkuses ja järjestuse sarnasuses, mis viitab tsentromeeride geneetilisele ja struktuursele mitmekesisusele (Logsdon et al., 2024). See mitmekesisus võib mõjutada kromosoomide stabiilsust, jagunemist ja epigeneetilist regulatsiooni.

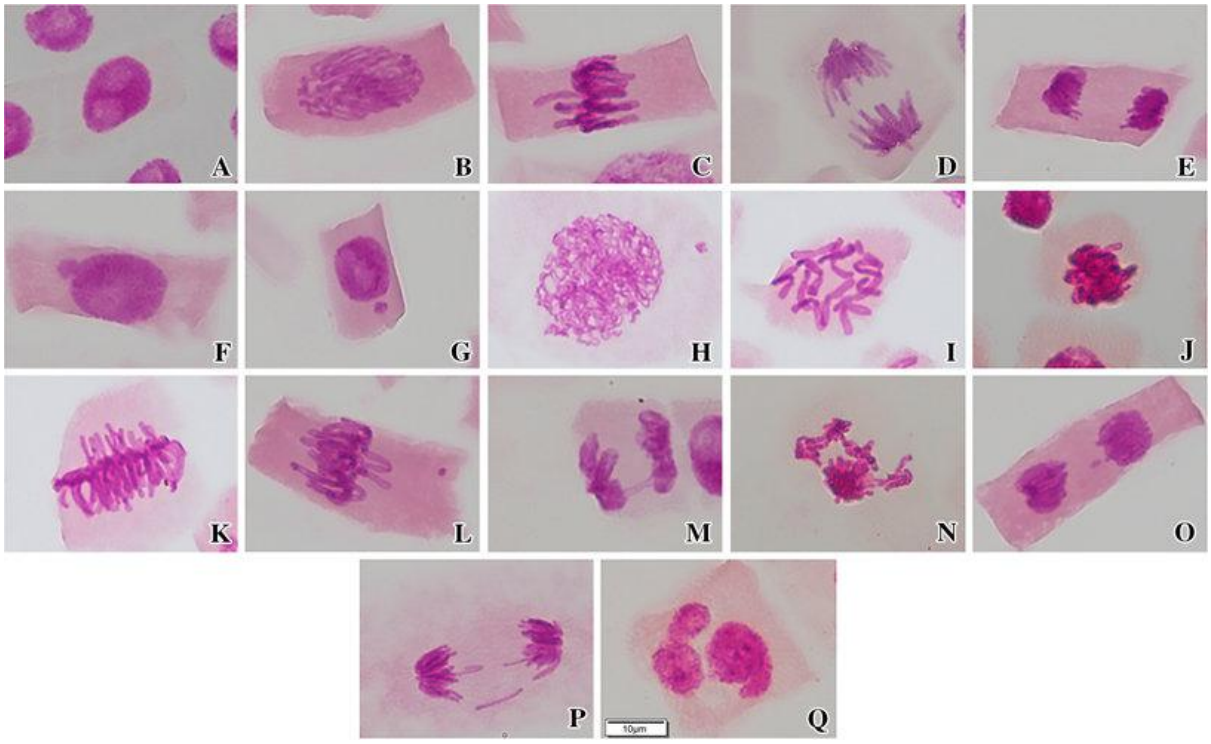
Tsentromeeride struktuuri ja suuruse hindamisel on kasulikuks osutunud k-meer põhine analüüsimeetod, mis kasutab järjestuse lühikeste fragmentide sagedusi korduspiirkondade ulatuse ja keerukuse määramiseks (X. Liao et al., 2024). Lisaks pikkuse hindamisele võimaldab see meetod paremini mõista genoomi kriitiliste piirkondade, nagu tsentromeerid, struktuurset varieeruvust.

Käesolev bakalaureusetöö eesmärgiks on uurida, kas ja millisel määral esineb korrelatsioon k-meeride sageduse ja inimese kromosoomide tsentromeeride pikkuse vahel. Selle kaudu sooviks paremini mõista tsentromeerse DNA struktuurset varieeruvust ning hinnata k-meer põhise analüüsi sobivust tsentromeeri omaduste kaudseks määramiseks.

1. Kirjanduse ülevaade

1.1. Tsentromeeri struktuur

Tsentromeer on kromosoomi piirkond, mis mängib suurt rolli raku jagunemisel. Sinna kinnituvad spindlikiud, mis aitavad kromosoome õigetesse raku pooltesse jaotada (Chadwick, 2025). Inimene tsentromeeride järjestamine ja assambleerimine on keeruline nende suuruse ja korduvate järjestuste tõttu. Selle tulemusel on tsentromeersete variatsioonide kindlaks tegemine keeruline ning ka mudelid, mis kirjeldavad nende arengut ja funktsiooni, on seni puudulikud. Tsentromeerid on just ühed kõige kiiremini muutuvad piirkonnad genoomis (Liao et al., 2023).



Joonis 1. Kromosoomide aberratsioonid, mida täheldati **Allium cepa** (hariliku sibula) meristeemirakkudes: a) normaalne interfaas, b) normaalne profaas, c) normaalne metafaas, d) normaalne anafaas, e) normaalne telofaas, f) interfaas tuumavääntõmbega, g) mikrotuumaga rakk interfaasis, h) mikrotuumaga ja polüploidne rakk profaasis, i) C-metafaas, j) metafaas kromosoomide kleepumisega, k) polüploidne metafaas, l) mikrotuumaga rakk metafaasis, m) anafaasikamber kromosoomisillaga, n) multipolaarne anafaas, o) telofaas kromosoomikatkestusega, p) telofaas kromosoomikaduvusega, q) kahetuumaline ja lobuleerunud (sõlmeline) rakk (Palsikowski et al., 2018).

1.1.1. Klassifikatsioon

Tsentromeerid kromosoomide olulised piirkonnad, kuhu moodustub kinetokoor – valkudest koosnev kompleks, mis võimaldab kromosoomidel kinnituda jagunemiskäävi mikrotuubulitele ja õigesti jaotuda rakkude jagunemisel (Fukagawa & Earnshaw, 2014).
Tsentromeerid ei paikne alati kromosoomide keskel, vaid neid leidub ka kromosoomi otstes (Chadwick, 2025). Tegelikult paiknevad tsentromeerid täpselt keskel vaid metatsentrilistel kromosoomidel. Teiste tüüpide puhul võivad need asuda erinevates kromosoomi osades, mis on iga konkreetse kromosoomi jaoks iseloomulikud. Seetõttu on tsentromeeri asukoht oluline orientiir kromosoomide rühmitamisel karuotüübi alusel ning aitab kaasa geenide paiknemise kirjeldamisel ja kaardistamisel (O'Connor, 2008).

	metatsentriline kromosoom 1, 3, 16, 19, 20	tsentromeer paikneb ligikaudu kromosoomi keskel ning lühike ja pikk haru on peaaegu võrdse pikkusega
	submetatsentriline kromosoom 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 17, 18, X	tsentromeer asub veidi keskkohast eemal, mistõttu on lühike haru selgelt eristatav pikast harust, kuid see ei ole märkimisväärselt lühike
	akrotsentriline kromosoom 13, 14, 15, 21, 22, Y	tsentromeer paikneb kromosoomi otsale väga lähedal ning lühike haru on oluliselt lühem kui pikk haru

Tabel 1. Inimese kromosoomitüübid. Lähtudes kromosoomide pikkusest ja tsentromeeri asukohast jaotatakse kõik kromosoomid metatsentrilisteks, submetatsentrilisteks ja akrotsentrilisteks kromosoomideks (Katagiri & Tamaki, 2021, modifitseeritud).

Tavaliselt leidub igal eukarüootsel kromosoomil vaid üks tsentromeer, mis on vajalik kromosoomide korrektseks jaotumiseks mitoosi käigus. Kui kromosoomil tsentromeer puudub, jaguneb see mitoosis juhuslikult ja võib rakust kaduda (O'Connor, 2008). Vastupidine olukord, kus kromosoomil on mitu tsentromeeri, võib viia kromosoomi katkemiseni, kui erinevad tsentromeerid kinnituvad vastandlikele kääviniitide poolustele kineetokooride vahendusel (O'Connor, 2008).

1.1.2. CENP-valgud

Tsentromeerid määratakse epigeneetiliselt histoon H3 erivariandi ehk tsentromeerivalgu A (CENP-A) kaudu. CENP-A avastati juhuslikult 1985. aastal William Earnshaw'i poolt immunoblottimise ja immunovärvimise eksperimentide käigus. Kui CREST sündroomiga patsientidelt eraldatud seerumit kasutati Western blot'i, millega tuvastati kolm korduvat valgubändi, mis esinesid paljudel patsientidel. Sama seerumiga koekultuurrakkude immunovärvimine näitas, et need valgud paiknevad tsentromeeride piirkonnas. Selle tulemusena nimetati need valgud tsentromeerivalkudeks A, B ja C. Hiljem kinnitasid biokeemilised uuringud, et CENP-A eraldub koos histoonidega ning on tõepoolest osa nukleosoomiosakestest (De Rop et al., 2012).

CENP-B on üks vähestest teadaolevatest tsentromeerivalkudest, mille on selge DNA seondumise spetsiifilisus ning mis on tugevalt konserveerunud imetajate seas (Masumoto et al., 1989). CENP-B sihtjärjestus ehk CENP-B boks koosneb üheksast hädavajalikust nukleotiidist ning selle konsensusjärjestus on 5'-TTCGNNNNAN-3' (Kipling & Warburton, 1997; Gamba & Fachinetti, 2020). Hoolimata selle valgu ja tema seondumiskoha evolutsioonilisest säilimisest, ei ole CENP-B tsentromeeri funktsiooniks siiski hädavajalik. Näiteks ei sisalda inimese neotsentromeerid ja paljude liikide Y-kromosoomid CENP-B seondumiskohti ning seetõttu neid ka ei seo CENP-B (Choo, 2000). Vastupidiselt sellele võivad inaktiivsed pseudo-ditsentrilised kromosoomid siiski sisaldada CENP-B valku, mis viitab, et selle olemasolu ei pruugi olla funktsionaalse tsentromeeri tekkeks piisav (Choo, 2000). Inimese tehislikul kromosoomil on samuti suudetud CENP-A kromatiini siduda kordumatule DNA-järjestusele ilma CENP-B vajaduseta (Logsdon et al., 2019).

CENP-B kõrge konserveeritus ja samaaegne mittevajalikus tekitavad vastuolusid, jättes valgu tegeliku rolli senini vaieldavaks. On pakutud, et CENP-B võib osaleda tsentromeeride aktiivsuse ülesehituses, lahtivõtmises või säilitamises (Dai et al., 2013), ning aitab stabiliseerida CENP-A ja CENP-C olemasolu tsentromeerides, suurendades nende tugevust ja kromosoomide korrektset segregatsiooni (Dumont et al., 2020; Fachinetti et al., 2015; Gamba & Fachinetti, 2020). CENP-B puudumisega Y-kromosoomi on leitud sagedamini mitmetes vähitüüpides, viidates selle kromosoomi suurenenud valejaotuse tõenäosusele (Abdel-Hafiz et al., 2023). On ka arvatud, et CENP-B osaleb peritsentromeerse heterokromatiini kujunemises (McNulty et al., 2017), kuna selle puudumisel kaob H3K9me3-modifikatsioon tsentromeeri ümbrusest, põhjustades heterokromatiini lagunemist ja genoomi ebastabiilsust (Kumon et al., 2021; Morozov et al., 2017). Alternatiivse hüpoteesi järgi võib CENP-B konserveeritus olla

seotud tsentromeeriväliste ülesannetega, nagu transposoonide vaigistamine (Cam et al., 2008; Zaratiegui et al., 2011). Veel on pakutud, et CENP-B aitab tsentromeersetel satelliitidel painutada DNA-d mitte-B konformatsiooni, mis on tsentromeeridele omane (Kasinathan & Henikoff, 2018), ning võib koos CENP-A-ga soodustada kromatiini avatud olekut, võimaldades nukleosoomide DNA osalist lahtikerimist (Nagpal et al., 2023).

CENP-C, üks kõrgemate selgroogsete kinetohoori võtmekomponente, avastati esmakordselt antitsentromeerseid antikehi sisaldavate autoimmuunhaigustega patsientide kaudu (Moroi et al., 1980; Earnshaw & Rothfield, 1985). See valk lokaliseerub sisemistele kinetohooriplaatidele, tsentromeerse DNA vahetus lähedusse (Saitoh et al., 1992) ning on teadaolevalt võimeline seonduma DNA-ga (Yang et al., 1996). Uuringud on näidanud, et CENP-C on rakutsükli kulgemiseks ja rakkude proliferatsiooniks hädavajalik. Kana DT40 rakkudes põhjustas CENP-C inaktiveerimine mitoosiviivitust, kromosoomide väärsegregeerumist ning apoptoosi (Fukagawa & Brown, 1997; Fukagawa et al., 2001). Lisaks on mitootilist peetust täheldatud ka pärast anti-CENP-C antikehade mikroinjektsiooni inimese HeLa rakkudesse (Tomkiel et al., 1994), mis viitab sellele, et CENP-C või sellega seotud valk mängib rolli kinetohoori suuruse määramisel. Kuigi CENP-C homolooge on kirjeldatud mitmetes mudelorganismides, jääb CENP-C täpne roll kõrgemate selgroogsete rakkudes õekromatiidide eraldumise ja eristumise protsessis ebaselgeks (Kwon et al., 2007).

1.1.3. Rakufaas

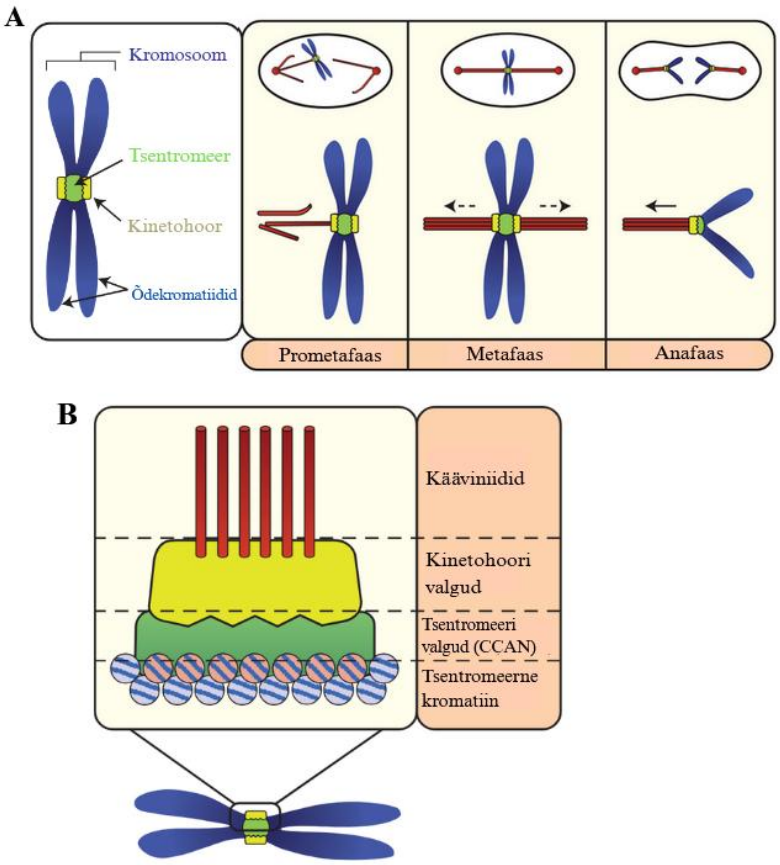
Tsentromeer on suur kromatiini sisaldav valgukompleks, mis toimib mitoosis kinetohoori koostamiskohana. Kinetohoor ise on megadaltoni suurune valgukompleks, mis kinnitub kääviniitidele ning vastutab õdede kromatiidide lahknemise eest anafaasis (Joonis 2) (Westhorpe & Straight, 2015).

Joonisel 2A on toodud enne mitoosi ja selle varajates faasides aktiveeruvad tsentromeerid (roheline ring), mis alustavad kinetohoori valkude (kollased kettad) kogumist. Prometafaasis kujuneb kinetohooriks struktuur, mis moodustab kinnitumiskohad kääviniitidele (punased vardad). Kui mõlema õdekromatiidide paari kinetohoorid on kindlalt ja korrektselt kinnitunud kääviniitidele tõmbavad mikrotuubulite poolt rakendatavad jõud (katkendjooned) kromatiidid metafaasiplaadile. Anafaasis laguneb õdekromatiidide vaheline sidusus ning tsentromeer koos kinetohooriga suunab mikrotuubuliste abil iga kromatiidi rakujagunemisel vastassuundadesse (Westhorpe & Straight, 2015).

Lisaks mikrotuubulite kinnitamisele on kinetohoor ka mitoosikontrollpunkti aktiveerimise kohaks – see mehhanism takistab anafaasi algust, kui mõni kinetohoor on jäänud kääviniitidega ühendamata. Ilma tsentromeerita kinetohoori ei moodustuks ning rakud ei suudaks kromosoome lahutada. Seega on tsentromeer üliioluline kromosoomide segregatsiooni ja mitoosi kontrolli seisukohalt (Westhorpe & Straight, 2015).

Mitoosi käigus on tsentromeeril spetsiifiline ülesehitus. Tsentromeerne kromatiin koosneb erilisest nukleosoomist, mis sisaldavad histoon H3 varianti – tsentromeerspetsiifilist valku CENP-A. See valk värbab omakorda tsentromeerivalgud (roheline), mis moodustavad nn püsiva tsentromeeriga seotud valgulise võrgustiku (CCAN). CCAN toimib platvormina, mille kaudu kinnituvad kinetohoori valgud (kollased) mitoosi ajaks kääviniitidele (Joonis 2B) (Westhorpe & Straight, 2015).

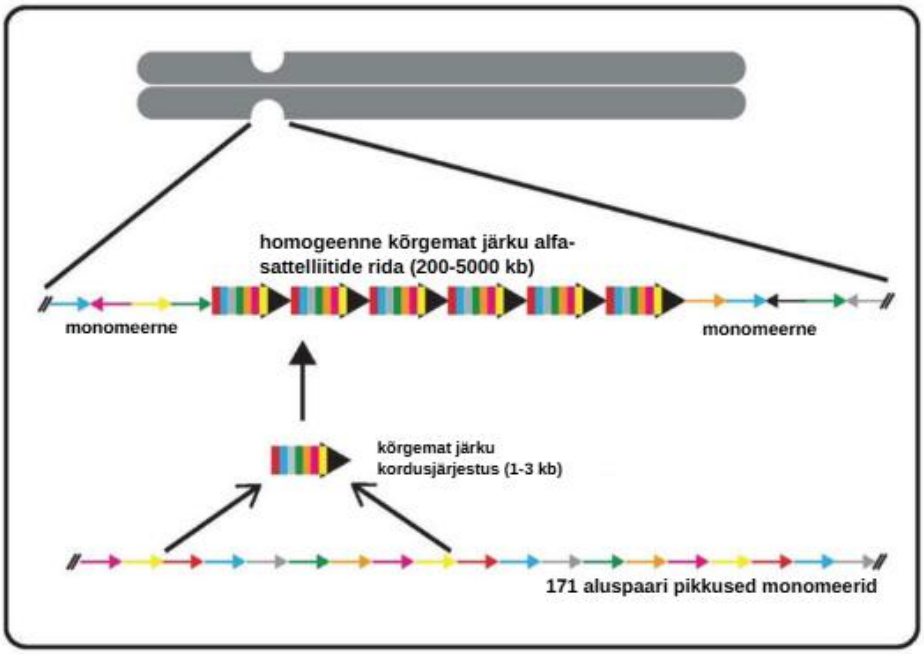
Tsentromeerne kromatiin muutub dünaamiliselt kogu rakutsükli vältel. DNA replikatsiooni käigus S-faasis, pärast replikatsioonikahvli läbimist, lahjendub $CenH3^{CENP-A}$. Seejärel algab sündmuste ahel, kuhu kuulub kromosoomide järkjärguline kondenseerumine ning igale tsentromeerile moodustub üks kintetohoor. $CenH3^{CENP-A}$ ladestumine tsentromeeridesse on rangelt reguleeritud ka toimub inimestel ning mitmetel teistel eukarüootidel telofaasi ja varajase G1-faasi vahel (Jansen et al., 2007). Ka teised tsentromeeri ja kinetohoori komponendid läbivad rakutsükli vältel kindlat vahetusprotsessi erinevatel aegadel. Seetõttu võivad protsessid nagu replikatsioonistress või DNA kahjustused mõjutada tsentromeeri ja kinetohoori toimimist (Westhorpe & Straight, 2015).



Joonis 2. Tsentromeeri funktsioon ja ülesehitus mitoosi ajal (Westhorpe & Straight, 2015, modifitseeritud).

1.1.4. Tsentromeeri DNA järjestus

Inimese tsentromeeri DNA koosneb 171 aluspaari pikkustest A-T-rikastest alfa-satelliitmonomeeridest ehk motiivist, mis on järjestatud tandemkordustena (Harrington et al., 1997; Willard, 1990). Monomeerid on omavahel 50-70% ulatuses identsed nende järjestatud tandemkordustest moodustuvad kõrgemat järku kordused ehk HOR-id. Monomeeride arv ja järjestus HOR-ides määravad krmosoomispetsiifilisuse. Iga inimese kromosoomi tsentromeer on defineeritud unikaalse HOR-struktuuri järgi. Näiteks on inimese X-kromosoomi (HSAX) tsentromeer määratletud 12 monomeerist koosneva HOR-iga, samas kui kromosoomi HSA8 puhul on see 7-monomeeriline kordus (D8Z2) (Miga, 2015).



Joonis 3. Inimese tsentromeeride genoomne korraldus (Aldrup-MacDonald & Sullivan, 2014, modifitseeritud).

Kuigi inimese kromosoomispetsiifilised α-satelliidi HOR-id määratleti juba 1990.aastate keskpaigaks, piirdus olemasolev arusaam tsentromeeride üldehitusest ja evolutsioonist üldiste mudelitega (Smith, 1976). Teadmised nende täpsest järjestuskoostisest ja varieeruvusest inimeste ja populatsioonide lõikes olid puudulikud ning põhinesid peamiselt keemilistel meetoditel või lünklikel järjestusandmetel (Henikoff et al., 2015). Pikkade lugemite sekveneerimistehnoloogiate kasutuselevõtt on avanud võimaluse tsentromeersete satelliit-DNA täielikuks järjestamiseks ja koostamiseks (Altemose et al., 2022; Logsdon et al., 2021; Miga et al., 2020).

Pideva tsentromeeripiirkonna järjestuseta on olnud keeruline seostada α-satelliidi struktuurseid tunnuseid nende funktsionaalse rolliga tsentromeeri kujunemisel. Viimased arvutuslikud lähenemised on siiski võimaldanud luua graafilisi mudeleid inimesi tsentromeeri järjestuste kohta (Hayden, 2012; Miga et al., 2014; Rosenbloom et al., 2015), tähistades olulist sammu lineaarsete tsentromeerikartograafiate suunas. Sellised mudelid on avanud võimaluse uurida α-satelliidi geneetilist sisu ja näidanud nii ühe kromosoomi sisest kui ka kromosoomide vahelist mitmekesisust. Piiranguks on aga see, et need ei kajasta järjestuse täpset järjestust ühegi tsentromeeri piires. Seetõttu jääb α-satelliitide pikkade järjestuste organisatsiooniline ülesehitus valdavalt teadmata (Sullivan et al., 2017).

Senised uuringud on näidanud, et enamikul inimese kromosoomidel on mitu HOR-ide kogumit, see tähendab rohkem kui üks kõrgemat järku α-satelliidi kordus tsentromeerses piirkonnas (Sullivan et al., 2017).

1.2. Tsentromeeri järjestuse varieeruvus ja tuvastamine

1.2.1. Inimese genoomi projekt GRCh38

Viimase 20 aasta jooksul on inimese referentsgenoom GRCh38 kujunenud inimese geneetika ja genoomika keskseks tugistruktuuriks (Hattori, 2005; Lander et al., 2001; Schneider et al., 2017). GRCh38 on kui kõige täpsemalt sekveneeritud inimese genoom maailmas. See koostati paljude doonorite andmete põhjal ning sekveneerimine viid läbi Sangeri sekveneerimise abil, mis võimaldab kui 1000 nukleotiidi pikkuseid ja kümme korda täpsemaid lugemeid kui kõrge läbilaskevõimega lühilugemite sekveneerimine. Võrreldes GRCh37-ga tehti GRCh38-s 8000 nukleotiidimuudatust, parandati mitmeid valesti kokku pandud piirkondi, täideti lünki, lisati järjestusi tsentromeeride jaoks ning suurendati genoomi mitmekesisust (Guo et al., 2017).

GRCh38 üheks olulisemaks kasutusviisiks on alusmaterjali pakkumine kliinilisteks, võrdlevateks ja populatsioonipõhisteks genoomianalüüsideks. Üle miljoni inimese genoomi on sekveneeritud eesmärgiga uurida geneetilist mitmekesisust ja kliinilisi seoseid. Peaaegu kõigi nende analüüside puhul on sekveneeritud andmed joondatud referentsgenoomiga (Karczewski et al., 2020; Schneider et al., 2017; Stephens et al., 2015).

Juhul, kui genoomid koostatakse *de novo* ehk ilma eelneva referentsita, võrreldakse saadud järjestusi tavapäraselt referentsgenoomiga, et tuvastada geneetilisi variatsioone, tuginedes ulatuslikele olemasolevatele annotatsioonikogumikele (Seo et al., 2016; Shafin et al., 2020).

Tänapäeval kasutatakse GRCh38 referentsgenoomi laialdaselt erinevates rakendustes ning selle juurde kuuluvad mitmekesised ressursid, mis võimaldavad järjestuse visualiseerimist ja annotatsiooni eri rakutüüpides ning haigusseisundites (Abascal et al., 2020; Navarro Gonzalez et al., 2021; Schneider et al., 2017; Taliun et al., 2021).

Hoolimata aastakümneid kestnud pingutustest selle täiendamisel ja täpsustamisel, esineb GRCh38-s siiski mitmeid olulisi piiranguid, mis takistavad genoomi terviklikku analüüsi. Näiteks sisaldab GRCh38 üle 100 miljoni nukleotiidi, mis on endiselt täielikult lahendamata. Nende asemele on kirjutatud kas „N“-tähed (nt akrotsentriliste kromosoomide p-õlad) või on asendatud kunstlike mudelitega, milleks on tsentromeersed satelliitjärjestused (Miga et al., 2014). Lisaks sisaldab GRCh38 veel 11,5 miljonit aluspaari paigutamata või lokaliseerimata

järjestusi, mis on esindatud väljaspool peamisi kromosoome (Church et al., 2015; Schneider et al., 2017). Selliste järjestuste uurimine on keeruline ning sageli jäetakse need analüüsidest välja, et vältida valepositiivseid tulemusi variantide või regulatoorsete elementide tuvastamisel (Karczewski et al., 2020).

GRCh38 puhul on täheldatud süsteemseid koostevigu, mis põhjustavad näiteks sisestuste ja kustutuste (indelide) tasakaalustamatust (Audano et al., 2019; Chaisson et al., 2015, 2019). Need vead ja puudused toovad analüüsidesse kallutust, eriti tsentromeeride, satelliitalade ja muude keerukate genoomipiirkondade puhul.

Oluline probleem on ka see, kuidas referentsgenoom mõjutab variatsioonide tuvastamist suurtes populatsiooni- ja kliinilise genoomika andmekogumites. Projektid nagu 1000 Genoomi Projekt (1KGP) (Auton et al., 2015) ja gnomAD (Karczewski et al., 2020) on pakkunud väärtuslikku teavet inimestevahelise geneetilise mitmekesisuse ulatuse kohta. Mendelistlike ja keerukate haiguste analüüsid kasutavad neid andmestikke võimalike haigust põhjustavate variantide järjestamiseks ja prioritiseerimiseks, tuginedes alleelisagedustele ja muudele tõendusmaterjalidele (Gulko et al., 2015; Kircher et al., 2014; Yandell et al., 2011). Selliste andmekogude kasutamisel tuleb aga arvestada referentsgenoomi üldist kvaliteeti, kuna igasugune järjestuse viga või lünk võib varjata geneetilist variatsiooni ning selle mõju inimese fenotüüpidele ja haigustele.

GRCh38 koostati mitme doonori kloonipõhise sekveneerimise alusel, mis tõi kaasa suure hulga kunstlikke haplotüüpstruktuure, mis võivad analüüside tulemusi kallutada (Green et al., 2010; Lander et al., 2001). Kuigi on tehtud jõupingutusi teatud haruldaste alleelide asendamiseks sagedasematega, leidub tänini sadu tuhandeid tehislikke haplotüüpe ja haruldasi alleele (Ballouz et al., 2019; Schneider et al., 2017; Zerbino et al., 2020).

1.2.2. Telomere-To-Telomere konsortsium

Rakuliini CHM13 kasutamine lihtsustas inimese genoomi kokkupanekut, kuna tegemist on ainult ühe inimese haplotüübiga, mis kõrvaldab alleelide varieeruvuse—faktori, mis tavaliselt raskendab keerukate genoomistruktuuride koostamist (Chaisson et al., 2015; Steinberg et al., 2014). Vaatamata olulistele edusammudele on tsentromeeride järjestamine ja kokkupanek endiselt tehniliselt keeruline. Inimgenoomide analüüs Human Pangenome Reference Consortium (HPRC) raames näitas, et seni pole võimalik ühegi teise indiviidi genoomi täielikult järjestada üle tsentromeeride nii, nagu see õnnestus CHM13 puhul (Liao et al., 2023).

Seega on tsentromeeride täielikuks järjestamiseks ja kokkupanekuks jätkuvalt vaja täiendavaid metoodilisi lähenemisi.

Inimese tsentromeerne DNA koosneb suures osas umbes 171 aluspaari pikkustest α-satelliitsetest järjestustest, mis esinevad tandeemselt korduvate kõrge orduga kordusüksustena (higher-order repeat, HOR). Tsentromeeride järjestuses esineb indiviidide vahel märkimisväärset varieeruvust, mis tuleneb ebavõrdsest ristumisest, koordineeritud evolutsioonist ja hüppelise paljundamise mehhanismidest (Logsdon & Eichler, 2023; Miga & Alexandrov, 2021). Seetõttu ei ole ükski üksik inimese genoom, sealhulgas CHM13, piisav kogu inimese geneetilise mitmekesisuse esindamiseks.

T2T-CHM13 on esimene täielik ja lünkadeta inimese haploidse genoomi järjestus, mis katab kõik autosoomid ja X-kromosoomi (välja arvatud mõned ribosomaalse DNA ahelad) (Sergey et al., 2022). See genoomireferents võimaldab täpsemaid genoomianalüüse. Näiteks tuvastati 3,7 miljonit täiendavat ühe nukleotiidi polümorfismi (SNP), mis paiknevad piirkondades, mida GRCh38 ei hõlma. Samuti kajastab see täpsemalt 1000 Genoomi Projekti proovide koopiate arvu variatsioone võrreldes GRCh38-g (Aganezov et al., 2022; Auton et al., 2015).

Varasemad uuringud on tuvastanud kümnete megabaaside ulatuses järjestusi, mis esinevad populatsioonis polümorfsete struktuursete variantidena (SV) (Ebert et al., 2021). Kuna need alternatiivsed alleelid puuduvad olemasolevates inimese referentsgenoomides, on enam kui kaks kolmandikku SV-dest jäänud tuvastamata lühilugemite järjestuse ja GRCh38 referentsi kasutavatel meetoditel. On leitud, et üksikud struktuursed variandid mõjutavad sagedamini geenifunktsiooni kui üksikud SNP-d või lühikesed lisamised/kustutamised (Chiang et al., 2017; Sudmant et al., 2015). Referentskallutatuse ületamiseks on kavandatud üleminek pangenoomilisele referentsile (Liao et al., 2023).

1.2.3. Pangenoom

Pangenoom on kogu genoomijärjestuste kogum, mis on saadud mitmelt inimeselt, esitades liigi geneetilist mitmekesisust. Pangenoomi andmestruktuur tugineb sellele, et toodetakse suures mahus kvaliteetset ja faasitud haplotüüpe ehk kromosoomilõike, mis on päritud kas isalt või emalt. Selle tootmise eesmärgiks on täiustada senist inimese võrdlusgenoomi. Andmesüsteem sisaldab koordinaatsüsteemi, millel on lihtne ja intuitiivne raamistik genoomivariantide viitamiseks ning mis säilitab tagurpidi ühilduvuse GRCh38 ja varasemate lineaarsete võrdlusjärjestustega (Wang et al., 2022).

Pangenoomipõhine lähenemine on võimaldanud tuvastada suures mahus geneetilist varieeruvust, mis jäi varasemates uuringutes märkamata just seetõttu, et lühikesed järjestuslugemid joondati peamiselt vaid ühe referentsi suhtes (Ebler et al., 2022; Liao et al., 2023; Rice et al., 2023; Sirén et al., 2021; Wong et al., 2020; Zhou et al., 2022). Lisaks on pangenoomide kasutamine suurendanud tunnuste ja geenide seoste kaardistamise täpsust (Chin et al., 2023; J. M. Song et al., 2020) ning võimaldanud keerukate struktuursete erinevuste usaldusväärsemat tuvastamist (Hickey et al., 2020; B. Song et al., 2024). Pangenoomide abil on ümber hinnatud ka varem dokumenteeritud geneetilised variatsioonid – näiteks on ilmnenud, et osa ühe nukleotiidi polümorfismidest (SNP-id), mida varem peeti tõelisteks geneetilisteks erinevusteks, on tegelikult joondusvigade tagajärg, mis tulenevad struktuursetest erinevustest (Jaegle et al., 2023).

Pangenoomid pakuvad täpsemat ja esinduslikumat pilti geneetilisest mitmekesisusest. Need aitavad vältida referentsipõhist kallutatust – nähtust, kus analüüsi tulemused sõltuvad sellest, millist referentsgenoomi kasutatakse (Chen et al., 2021; Ebler et al., 2022; Gage et al., 2019; Günther & Nettelblad, 2019). Kombineerituna pangenoomide koostamise algoritmide hiljutiste edusammude (Garrison et al., 2023; Hickey et al., 2020) ja järjestusandmete kättesaadavuse kasvuga ka mitte-mudelliikide puhul (Lei et al., 2021), viitab see areng sellele, et pangenoomidest saab tõenäoliselt lähiaastatel populatsioonigeneetika uurimustes valdav tööriist.

Kuigi pangenoomide koostamine ja häälestamine on veel teatud kontekstis keeruline. Kuna pangenoomide koostamine nõuab rohkem järjestusandmeid kui üksiku referentsgenoomi loomine, võib see kujuneda kulukaks eriti suurte genoomide puhul või uuringutes, kus ressursid on piiratud. Pangenoomi loomiseks on võtmetähtsusega kogu genoomi joondamine (WGA), mis nõuab täpset parameetrite häälestamist ning mille keerukus võib oluliselt mõjutada lõpptulemust. Kuigi tänapäevased WGA algoritmid on tehniliselt väga arenenud ning neid rakendatakse edukalt paljude liikide puhul (Garrison et al., 2023), on need enamasti loodud inimese ja teiste mudelliikide genoomide jaoks. Seetõttu ei pruugi need hästi sobituda genoomidele, mis on suured, polüploidsed, kõrge kordusastmega, geneetiliselt väga mitmekesised või sisaldavad ulatuslikke struktuurseid variatsioone (Song et al., 2024).

1.2.4. Pikkade lugemite sekveneerimistehnoloogiad

Pika lugemispikkusega tehnoloogiad on võimaldanud kokku panna täielikke kromosoomide järjestusi (Logsdon et al., 2021; Miga et al., 2020; Sergey et al., 2021). Tehnoloogia on

näidanud võimekust analüüsida suuri ja keerukaid inimese struktuurseid variante (Ebert et al., 2021). PacBio võimaldab väga täpset konsensuslugemeid ja lugemite mõõdukas pikkus (10-20 kb), suudab järjekindlalt lahendada pikki tandemkorduseid, satelliitstruktuure ja suuri segmentduplikatsioone. Oxford Nanopore Technologies pakub samal ajal nanopooripõhist pika lugemispikkusega andmeid, mis katavad sadade kilobaaside pikkused lugemeid ehk katab ära ülipikad järjestused (Wang et al., 2022).

1.2.4.1. PacBio-HiFi

PacBio-HiFi järjestamismeetod võimaldab toota väga täpseid, pikkade lugemitega järjestusandmekogumeid, mille keskmine lugemipikkus on 10–25 kilobaasi ja täpsus ületab 99,5%. Sellised kvaliteetsed pikad lugemid on eriti väärtuslikud keerukates rakendustes, sealhulgas üksiknukleotiidsete ja struktuursete variantide tuvastamisel, genoomi de novo kokkupanekul, raskesti järjestatavate polüploidsete või väga korduvate genoomide analüüsil ning metagenoomide kokkupanekul. Käesoleval hetkel on suur vajadus avalikult kättesaadavate proovipõhiste andmekogumite järele, mis võimaldaksid hinnata nende täpsete ja pikkade lugemite eeliseid ning toetaksid bioinformaatiliste tööriistade, näiteks genoomi kokkupanekuprogrammide, variantide tuvastamise tööriistade ja haplotüpiseerimisalgoritmide arendamist (Hon, 2020).

PacBio mitmepassiline ringikujuline konsensusjärjestamine, mille käigus järjestatakse üksikuid pikki (kuni ~25 kb) DNA molekule, saavutades väga kõrge täpsuse — need ongi HiFi-lugemid (Wenger et al., 2019). Saavutati 28-kordne inimese genoomi katvus, keskmise lugemipikkusega 13,5 kb ja keskmise täpsusega 99,8%. Tulemused näitasid paremat kokkupanekut ja haplotüüpide määramist kui traditsioonilised mürarikkad pikad või lühikesed lugemid. Lisaks tuvastati üksiknukleotiidsed variandid täpsuse ja tundlikkuse poolest samaväärselt Illumina NovaSeq™ andmetega.

Pärast seda on HiFi-järjestusel põhinevates inimgenoomi projektides täheldatud märkimisväärseid edusamme genoomide kokkupanekus (Porubsky et al., 2019; Shumate et al., 2020; Nurk et al., 2020).

1.2.4.2. Oxford Nanopore Technologies

Viimastel aastatel on nanopoorjärjestamise tehnoloogia areng toonud kaasa märkimisväärseid edusamme DNA ja RNA üksikute pikkade molekulide järjestamisel, sealhulgas täpsuse,

lugemite pikkuse ning andmeläbilaskevõime olulise paranemise. Et neid tehnoloogilisi läbimurdeid tõhusalt rakendada, on olnud vaja ulatuslikke täiustusi nii katselistes kui ka bioinformaatilistes meetodites. Need täiustused on võimaldanud nanopooripõhist järjestamist kasutada keerukates genoomide, transkriptoomide, epigenoomide ja epitranskriptoomide analüüsides (Wang et al., 2021).

Nanopoorjärjestamise tehnoloogiat kasutatakse laialdaselt mitmesugustes rakendustes, sealhulgas de novo genoomide kokkupanekus, täispikkade transkriptide tuvastamises ja DNA või RNA alusmodifikatsioonide määramises. Samuti on see osutunud väärtuslikuks kiiretes kliinilistes diagnoosides ja nakkushaiguste puhangute seires. Kuigi nanopoorjärjestamine on kiiresti arenenud, on siiski alles mitmeid arenguvõimalusi, eelkõige andmekvaliteedi ja analüütiliste töövoogude osas. Edasine areng hõlmab uute nanopooride loomist, täpsemate alusetuvastusalgoritmide arendamist ning spetsiifilistele rakendustele kohandatud katseliste protokollide väljatöötamist.

Nanopoorjärjestamise tehnoloogia ja selle rakendused nii alus- kui ka rakendusuuringutes on teinud olulise läbimurde alates 2014. aastast, mil Oxford Nanopore Technologies (ONT) tõi turule esimese nanopoorjärjestusseadme MinIONi (Deamer et al., 2016; Jain et al., 2016). See tehnoloogia põhineb nanoskaalas valgulisel pooril ehk „nanopooril“, mis toimib biosensorina ja on paigutatud elektrit mittejuhtivasse polümeermembraani (Deamer et al., 2016; van Dijk et al., 2018). Elektroodlahuses rakendatakse nanopoori kaudu püsivat pinget, mille tulemusel liigub ioonvool läbi poorikanali. Negatiivselt laetud üksikahelalised DNA või RNA molekulid suunatakse selle voolu abil negatiivselt laetud „cis“-küljelt positiivselt laetud „trans“-küljele. Molekulide liikumist läbi nanopoori juhib mootorvalk, mis reguleerib nukleiinhappe liikumiskiirust sammhaaval. Selle tulemusena võimaldab nanopoorijärjestus reaalajas jälgida nukleotiidide järjestust, kuna molekuli läbimisel toimuvad muutused ioonvoolus, mida dekodeeritakse arvutuslike algoritmide abil. Lisaks kiiruse reguleerimisele omab mootorvalk ka helikaasi aktiivsust, mis võimaldab DNA või RNA–DNA kaksikahelate lahtikeeramist üksikahelalisteks molekulideks, mis läbivad nanopoori (Wang et al., 2021).

1.2.5. K-meerid

K-meeriteks nimetatakse lühikesi, k pikkuseid DNA järjestusi, mida saab kasutada genoomijärjestuste analüüsiks. Inimese geenikomplektil on miljardeid k-meere (Pavlichin et al., 2022). K-meeride põhised meetodid on äratanud suuremat huvi genoomi analüüsimisel, kuna ei nõua genoomi kokkupanekut ja neid saab kasutada otse sekveneerimise lugemisel.

Viimastel aastatel on kasutusele võetud oligomeeride sageduste analüüs, mille analüüsi viivad läbi k-meerid. See meetod on kasulikuks osutunud, kuna on kiire ja vähem veaohtlik. Selle meetodiga on võimalik leida ka sekveneerimisvigu ja kattuvate lugemite avastamine (Kaplinski et al., 2015).

K-meerid pakuvad suuremat potentsiaali populatsioonigeneetikas kui neid seni on rakendatud. K-meeriks nimetatakse järjestuse alamsõnet pikkusega *k*. Näiteks järjestuses „ATGCA" esinevad unikaalsed 2-meerid on: AT, TG, GC ja CA. Üheks k-meeride oluliseks eeliseks on see, et nende analüüs ei nõua eelnevalt joondust. Selle asemel loendatakse kõik unikaalsed k-meerid antud lugemite kogumist – pöördkomplementaarid loetakse üldjuhul identseks ja käsitletakse ühiselt (Kokot et al., 2017) ning saadud loenditabeli alusel viiakse läbi edasine analüüs. Seetõttu säilitavad analüüsi võimaluse ka need järjestusosad, mis on referentsgenoomi suhtes tugevalt lahknenud või sellele sootuks puuduvad, võimaldades realistlikumat ülevaadet pangenoomilisest varieeruvusest (Roberts et al., 2024).

Üks esmaseid küsimusi k-meer-analüüsides on, millist k väärtust tuleks kasutada. Optimaalse k valik sõltub eelkõige kompromissist k pikkuse ja järjestusvigade mõju vahel: pikemad k-meerid võimaldavad paremat eristust unikaalsete genoomijärjestuste osas, kuid on samas tundlikumad järjestusvigade suhtes (Rahman et al., 2018). Praktikas jäävad kasutatavad k väärtused enamasti vahemikku 20–40 nukleotiidi (Ponsero et al., 2023), kuna selles vahemikus on k-meerid hästi järjestatavad lühilugemite põhistel andmestikel ning kipuvad joonduma unikaalselt lähtegenoomi suhtes (Wu et al., 1991; Becher et al., 2022). Näiteks on k = 32 puhul võimalik katta ligikaudu 85,7% inimese genoomi unikaalsetest järjestustest (Shajii et al., 2016), samal ajal kui k = 21 võimaldab eristada erinevaid eukarüootseid, bakteriaalseid ja arheelseid liike (Bussi et al., 2021).

Üheks lähenemiseks tulemuste tundlikkuse hindamiseks k suhtes on sama analüüsi kordamine mitmete erinevate k väärtuste puhul. See „toore jõu" meetod on eriti levinud genoomi koostamise algoritmides (Chikhi & Medvedev, 2014; Durai & Schulz, 2016), kuid on teoreetiliselt rakendatav ka muudes analüüsides. Meetodi peamiseks piiranguks on aga märkimisväärne arvutuslik koormus, mis kaasneb korduvate analüüsidega. Lisaks puuduvad sageli objektiivsed kriteeriumid, mille põhjal hinnata, kas mingi konkreetne k väärtus annab täpsemaid tulemusi, mis omakorda suurendab vajadust süsteemsete valikupõhimõtete järele (Roberts et al., 2025).

Pikemad k-meerid suurendavad üldiselt võimekust tuvastada genoomis unikaalseid piirkondi, ent samal ajal suurendavad need ka tõenäosust, et iga k-mer sisaldab vähemalt ühte järjestusviga (Chikhi & Medvedev, 2014; Rahman et al., 2018). Kuna üksainus järjestusviga

võib põhjustada kuni k vigast k-meeri, on soovitatav leida k väärtus, mis maksimeerib veavabade ja unikaalsete k-meride arvu. Selline lähenemine sobib eriti hästi genoomi kokkupanekuks (Chikhi & Medvedev, 2014).

Optimaalse k leidmiseks analüüsitakse sageli k-meride sagedusjaotust ja sobitatakse sellele statistiline mudel, et hinnata vigade põhjustatud k-meride osakaalu (Chikhi & Medvedev, 2014). Vigased k-meerid kipuvad esinema madalate sagedustega, kuna järjestusvead on haruldased ja ei kordu samal viisil lähtegenoomis (Kelley et al., 2010).

Kuigi sama põhimõtet võiks rakendada ka populatsioonigeneetilistes analüüsides, ei ole selge, kas sellisel viisil tuvastatud optimaalne k on stabiilne ühe liigi genoomide lõikes. Tõenäoliselt varieerub optimaalne k sõltuvalt kordusjärjestuste osakaalust, genoomi suurusest (Schmuths jt, 2004) ja ploidsusest (Kolář et al., 2017). Edasistes populatsioonigeneetilistes uuringutes võiks keskenduda sellele, kuidas valida k, mis maksimeerib veavabade k-meride arvu kogu populatsiooni ulatuses (Haberer et al., 2020).

2. Eksperimentaalosa

2.1 Töö eesmärgid

Käesoleva töö eesmärkideks on:

1. CHM13 põhise k-meeri abil Inimese Pangenoomi v.1 indiviidide tsentromeeride pikkuse ja k-meeri sageduse vahelise korrelatsiooni leidmine.
2. Indiviid HG01106 kromosoom 9 assemblipõhise pikkuse võrdlemine WGS põhise k-meer sagedusega

2.2 Materjal ja metoodika

2.2.1. CHM13, sagedaisema k-meeri leidmine

Töös kasutatakse T2T-CHM13v2.0 referentsgenoomi, mille andmed saadi NCBI Assembly database. T2T-CHM13v2.0 referentsgenoomi (GCF_009914755.1, T2T Consortium, 2022) põhjal eraldati tsentromeerina määratletud järjestused, kasutades satelliitsete järjestuste annatatsioonfaili nimega *chm13v2.0_censat_v2.1.bed* (Altemose et al., 2022), mis sisaldab koordinaatide tsentromeeripiirkondade määramiseks. Eristatud tsentromeeripiirkondadest loodi GenomeTester4 paketi glistmaker funktsiooniga binaarne k-meer list. Seejärel kasutati glistcompare funktsiooni, et moodustada uus list, mis sisaldas ainult vastava kromosoomi tsentromeeris paiknevaid k-meere. Nende seast valiti T2T-CHM13v2.0 referentsgenoomi põhjal sagedasim k-meer. Vastavad andmed on sain juhendajalt.

2.2.2. Pangenoomi indiviidid

Inimese pangenoomi (Human Pangenome Reference Consortium) I etapi 47 indiviidi puhul kasutati isalt ja emalt päritud assembleid, mis paiknevad kataloogis *working/* (tabelis tulp 2) / *indiviid* (tabelis tulp 1) / *assemblies/year1_freeze_assembly_v2/*, kus faili nimedeks on *[indiviid].maternal.f1_assembly_v2.fa.gz* ja *[indiviid].paternal.f1_assembly_v2.fa.gz*. Assemblite kogupikkus ühelt vanemalt päritud kromosoomide põhiselt on umbes 3 miljardit aluspaari. Neid andmeid kasutasin kromosoomipõhise tsentromeeri pikkuse hindamisel.

2.2.3. Korrelatsiooni leidmine

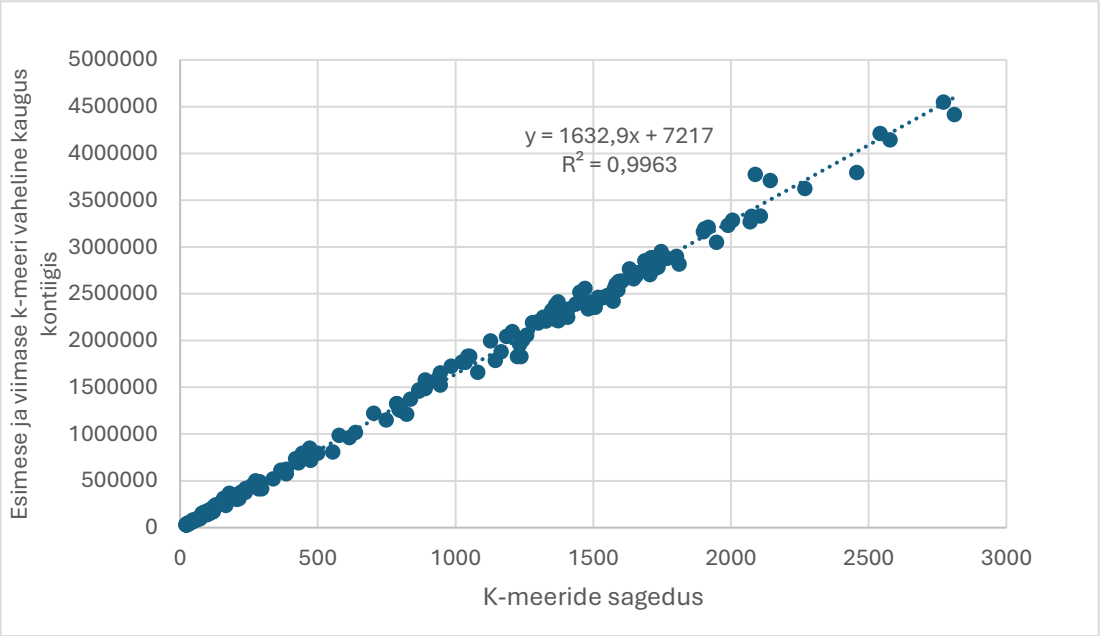
Korrelatsioonianalüüsi läbiviimiseks kasutasin andmestikku, mis paikneb lokaalses kataloogis */path/pangenome/samples/*. Iga kromosoomi kohta oli koostatud vastav *logi.txt* fail, mis sisaldas infot indiviidi identifikaatori, assembli päritolu (emalt või isalt), k-meeride järjestust, nende arvukuse, kromosoomisisese algus- ja lõppkoordinaadi ning tsentromeeri pikkuse kohta. Failide sisu vaatasin käsurea käsu *more* abil. Seejärel kopeerisin failis olevad andmed Microsoft Exceli töölehele edasiseks analüüsiks.
Iga kromosoomi kohta arvutasin korrelatsioonikordaja k-meeride koguarvu ja tsentromeeri pikkuse vahel, eesmärgiga tuvastada kromosoomid, millel on tugevam seos nende kahe tunnuse vahel. Korrelatsiooni leidmiseks kasutasin Pearsoni korrelatsioonikordajat. Panin ka kirja iga kromosoomi andmete hulga ehk ridade arvu.

2.2.4. HG01106 lühikeste lugemite põhise k-mer sageduse leidmine

Kõikide kromosoomide kohta on tehtud fülogeneesipuud, kus mina valisin indiviidi HG01106, kuna see on homosügoot. Selleks, et uurida indiviidi HG01106 kohta laadisin alla Human Pangenome Reference Consortium (HPRC) avalikust andmekogust PacBio HiFi järjestusandmed käsuga *wget https://human-pangenomics.s3.amazonaws.com/working/HPRC/HG01106/raw_data/PacBio_HiFi/m64043_200625_174853.ccs.bam*. Järgnevaks töötlemiseks Bioinformaatika õppetooli arvutusserverit, kus konverteerisin *.bam* formaadist järjestusfaili FASTQ vormingusse, kasutades *samtools* versiooni 1.12. Kuna fail oli mahukas ja protsess aeganõudev, siis käivitasin käsu nohup perl pacbio.pl, kuhu olin lisanud käsu *samtools fastq m64043_200625_174853.ccs.bam > m64043_200625_174853.ccs.fastq*. Edasi lõin järjestuse põhjal k-meer loendi, kus k-meeri pikkuseks seati 25 nukleotiidi. Selleks kasutasin GenomeTester4 tööriista glistmaker järgmiselt:
glistmaker m64043_200625_174853.ccs.fastq -w 25 -o m64043_200625_174853.ccs. K-meeride jaotuse ja sekveneerimiskatvuse hindamiskes kasutasin glistquery tööriista koos –distribution parameetriga:
glistquery m64043_200625_174853.ccs_25.list --distribution 100 > m64043_jaotus.txt. Tsentromeeridele iseloomuliku k-meeri TAAAAAGAAAGGTTCATCTCTGTTA absoluutse esinemissageduse määramiseks kasutasin käsku:
glistquery m64043_200625_174853.ccs_25.list -q TAAAAAGAAAGGTTCATCTCTGTTA.

2.3 Tulemused

1. Töös leidsin CHM13 põhise k-meeri abil Inimese Pangenoomi v.1 indiviidide tsentromeeride pikkuse ja k-meeri sageduse vahelise korrelatsiooni. Joonis 3 kirjeldab kromosoom 9 tsentromeeri pikkuse ja k-meeride sageduste vahelist korrelatsiooni, mis näitab seda, et kahe tunnuse vahel on kindel korrelatsioon ning see on tõusev. Teiste kromosoomidega tegin samamoodi ning nende korrelatsioonide kokkuvõttev tabel on leitav lisades (Lisa 2.).



Joonis 4. Kromosoom 9 tsentromeeride pikkuse ja k-meeri sageduste vaheline korrelatsioon.

2. Indiviid HG01106 kromosoom 9 assemblipõhise pikkuse võrdlemine WGS põhise k-meer sagedusega. Kasutades Bioinformaatika õppetooli arvutusserverit ja eelnevalt kirjeldatud käske leidsin, et indiviid HG01106 PacBio HiFi järjestusfailist saadud andmete põhjal esines tsentromeeridele iseloomulik k-meer TAAAAAGAAAGGTTCATCTCTGTTA sagedusega 33620. Selle esinemissageduse jagamisel väärtusega 11 (mis vastas jaotusfailis kõige kõrgema katvuseklassiga väärtusele) saadi hinnanguliselt 3056 k-meeri.

   See tulemus on ligilähedane Excelis saadud väärtusele, kus sama k-meeri üldarvuks oli 3216 (sh 1510 isalt ja 1706 emalt). Saadud andmed kinnitavad, et GenomeTester4 tööriistade abil loodud k-meeride jaotus ning sageduspõhine lähenemine

tsentromeeride analüüsiks on kooskõlas eelnevate arvutustega ning sobib hästi korduvate järjestuste kvantifitseerimiseks inimese genoomis.

2.4 Arutelu

Tsentromeerid on kromosomaalsed struktuurid, mis koosnevad suurel määral kordusjärjestustest ja mille pikkus võib varieeruda indiviiditi. Kuna teatud k-meerid ehk fikseeritud pikkusega nukleotiidjärjestused esinevad eelistatult tsentromeeride piirkonnas, siis võib nende sagedus genoomis peegeldada vastava kromosoomi tsentromeeri pikkust. Käes olevas töös varieerusid korrelatsioonid kromosoomid lõikes. Kui enamuses oli korrelatsioon kõrge, mis näitab seda, et k-meer võib olla bioloogiliselt seotud tsentromeeri kordustega ja seega olla hea marker tsentromeeri pikkuse ennustamiseks.

Kuid esines ka kromosoome, mille oli väga madal korrelatsioonikordaja k-meeri sageduse ja tsentromeeri pikkuse vahel. Näiteks kromosoom 13 korrelatsioonikordaja oli kõigest 0,51. Madal korrelatsioonikordaja tuleneb sellest, et see võib olla tingitud kromosoomi 13 α-satelliitsete korduste (HOR) suurest järjestuslikust varieeruvusest ning struktuurilisest mitmekesisusest CHM1 ja CHM13 genoomides (Logsdon et al., 2024). Lisaks sisaldab kromosoom 13 uusi evolutsioonilisi kihte ja seni kirjeldamata kordusetüüpe, mis võivad mõjutada k-meeride jaotust ja sagedust (Logsdon et al., 2024).

Trisoomia 21 juhtumiuuring pakub hea näite sellest, et tsentromeeri pikkus ja k-meeride sagedus ei pruugi alati korreleeruda. Uuritud lapsel, kellel oli kolm kromosoomi 21 – kaks emalt ja üks isalt. Selles uuringus erinesid tsentromeeride pikkused kuni 11-kordselt. Just pikim tsentromeer (H1) oli funktsionaalselt kõige nõrgem. Sellel tsentromeeril puudus selge tööpiirkond (CDR) ning CENP-A valgu seondumine oli väiksem kui lühematel alleelidel (Mastrorosa et al., 2024).

See uuring toob esile olulisi seoseid tsentromeeri pikkuse ja funktsionaalsuse vahel. Otseselt ei käsitleta korrelatsiooni k-meeride sageduse ja tsentromeeri pikkuse vahel. Kuna tegemist on üksikjuhtumiga, rõhutab see vajadust teha edasisi uuringuid, et uurida, kuidas k-meeride sagedus võib peegeldada tsentromeeri struktuurilisi ja funktsionaalseid omadusi.

Hiljuti avaldatud inimese pangenoomi teine versioon (Human Pangenome Release 2) tähistab olulist sammu genoomi mitmekesisuse paremal kaardistamisel. Selle asemel, et tugineda ühele referentsgenoomile, koondab uus pangenoom üle 200 indiviidi andmed ja enam kui 400 haplotüüpi, pakkudes oluliselt mitmekesisemat ja täpsemat vaadet inimese genoomile.

Oluline edasiminek on see, et ligikaudu poole kromosoomid on indiviiditi täielikult assambleeritud, ulatudes telomeerist telomeerini. See loob esmakordselt võimaluse analüüsida tsentromeere täies pikkuses ja võrrelda nende järjestuslikku struktuuri indiviiditi. Samuti sisaldab andmestik suure täpsusega HiFi transkriptsiooniandmeid, mis võimaldavad paremat funktsionaalset annotatsiooni.

Kuigi see areng loob tugeva aluse tsentromeeri omaduste ja k-meeride sageduse seoste uurimiseks, ei ole käesolevas töös nende uute andmetega veel süvitsi tegeletud. Edasised analüüsid pangenoomi põhjal on kindlasti vajalikud, et neid seoseid paremini mõista.

Kuna k-meeride sagedust kasutatakse sageli tsentromeeri pikkuse kaudseks hinnanguks, sõltub tulemuste usaldusväärsus suuresti sellest, kui täpne ja järjepidev on kasutatud genoomikoostis. HG01106 puhul võib erinevates genoomiversioonides esineda erinevusi just kromosoom 9 tsentromeeripiirkonnas, mis on teadaolevalt keerulise ja korduva järjestusega. Sellised piirkonnad on tundlikud erinevatele analüüsi-pipeline'ide ja montaažitöötluse (assembly pipeline) variatsioonidele.

Kui ühes pipeline'is on tsentromeeri piirkond assambleeritud osaliselt (või ebatäpselt), võib see alahinnata teatud k-meeride sagedust või jätta need täielikult tuvastamata. Teises pipeline'is, kus sama piirkond on täpsemalt monteeritud (nt kasutades T2T või HiFi lugemisi), võivad samad k-meerid esineda suurema sagedusega ja tsentromeeri hinnanguline pikkus olla suurem.

Sellised erinevused mõjutavad k-meeride ja tsentromeeri pikkuse vahelise korrelatsiooni tugevust ning täpsust. Kui andmeallikad ei ole homogeensete meetoditega loodud, võib see moonutada korrelatsioonianalüüsi tulemusi või tekitada näivaid erandeid.

See töö on osa suuremast tööst, kus püüame määrata tsentromeeri puu haru põhist pikkust, mis diploidses sekveneerimisandmetes on.

Kokkuvõte

Tsentromeeride järjestuslik ja struktuurne varieeruvus on olnud keeruline uurimisvaldkond, eelkõige kordusjärjestuste rohkuse ja piiratud sekveneerimis täpsuse tõttu. Käesolevas bakalaureusetöös üritasin välja selgitada, kas k-meeride sagedus inimese genoomis võiks kajastada tsentromeeri pikkust ja võimaldada nende varieeruvuse kvantitatiivset hindamist. Analüüs tugines T2T-CHM13 referentsgenoomile ning Inimese Pangenoomi konsortsiumi esimese versiooni indiviidide WGS-andmetele.

Esmalt sai määratud CHM13 põhjal igale kromosoomile iseloomulik sagedasim tsentromeerispetsiifiline k-meer ning analüüsisin selle sageduse ja tsentromeeri pikkuse vahelist korrelatsiooni 47 indiviidi puhul, hõlmates kõiki kromosoome. Enamiku kromosoomide puhul ilmnes tugev positiivne korrelatsioon, mis viitab, et k-meeride sagedus peegeldab hästi tsentromeeride pikkust. Erandiks oli näiteks kromosoom 13, millel madala korrelatsioon võib viidata suuremale järjestuslikule mitmekesisusele või piirkonna keerukusele.

Lisaks võrdlesin indiviidi HG01106 näitel WGS-andmetes leitud k-meeride sagedusi tsentromeeride pikkustega, mis põhinesid genoomi kokkupandud järjestustel. Tulemused toetasid hüpoteesi, et k-meeride sagedused võivad olla heaks indikaatoriks tsentromeeri pikkuse leidmiseks.

Töö tulemused kinnitavad, et k-meeride sageduspõhine lähenemine on tõhus vahend tsentromeeride pikkuse ja mitmekesisuse uurimisel, avades uusi võimalusi nende keerukate genoomipiirkondade analüüsimiseks. Tuleviku uuringud, mis kasutavad täpsemaid pangenoomiversioone ja ulatuslikumaid andmestikke, võivad aidata veelgi paremini mõista tsentromeeride rolli inimese genoomi evolutsioonis ja toimimises.

Finding correlation between chromosome-specific k-mer frequency and centromere length

Carmen Beljaev

Summary

Centromeres are essential chromosomal regions responsible for proper segregation during cell division and are primarily composed of repetitive DNA sequences. Due to their highly repetitive nature, centromeres have long posed challenges for sequencing and analysis. However, recent advances in long-read sequencing technologies have enabled complete centromere assembly. This bachelor's thesis investigates whether the length of human centromeres correlates with the frequency of specific k-mers and to what extent k-mer-based methods can be used to assess structural variability in centromeres.

Frequent centromeric k-mers identified from the T2T-CHM13v2.0 reference genome were used to analyze their frequency across the centromeres of 47 individuals from the Human Pangenome Project. The analysis revealed that the total frequency of centromeric k-mers in pangenome assemblies shows a significant positive correlation with centromere length. This suggests that k-mer frequency is a potential proxy for estimating centromere size. The same trend was confirmed using whole-genome sequencing (WGS) data, where a single individual's (HG01106) k-mer frequencies were compared to centromere lengths determined from their assembly.

The results demonstrate that a k-mer frequency-based approach is a powerful tool for investigating centromere length and diversity, offering new possibilities for studying these complex genomic regions. Further research using improved versions of the pangenome and broader datasets could deepen our understanding of centromere evolution and function in the human genome.

Kasutatud kirjandus

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abdel-Hafiz, H. A., Schafer, J. M., Chen, X., Xiao, T., Gauntner, T. D., Li, Z., & Theodorescu, D. (2023). Y chromosome loss in cancer drives growth by evasion of adaptive immunity. *Nature*, *619*(7970), 624–631. https://doi.org/10.1038/s41586-023-06234-x

Aldrup-Macdonald, M. E., & Sullivan, B. A. (2014). The past, present, and future of human centromere genomics. *Genes*, *5*(1), 33–50. https://doi.org/10.3390/genes5010033

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science (New York, N.Y.)*, *376*(6588), eabl4178. https://doi.org/10.1126/science.abl4178

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., … Wilson, R. K., & Eichler, E. E. (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, *176*(3), 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019

Ballouz, S., Dobin, A., & Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, *20*(1), 159. https://doi.org/10.1186/s13059-019-1774-4

Becher, H., Sampson, J., & Twyford, A. D. (2022). Measuring the invisible: The sequences causal of genome size differences in eyebrights (Euphrasia) revealed by k-mers. *Frontiers in Plant Science*, *13*, 818410. https://doi.org/10.3389/fpls.2022.818410

Bussi, Y., Kapon, R., & Reich, Z. (2021). Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PloS One*, *16*(10), e0258693. https://doi.org/10.1371/journal.pone.0258693

Cam, H. P., Noma, K.-I., Ebina, H., Levin, H. L., & Grewal, S. I. S. (2008). Host genome surveillance for retrotransposons by transposon-derived proteins. *Nature*, *451*(7177), 431–436. https://doi.org/10.1038/nature06499

Chadwick, L. H. (2025). *Centromere*. National Human Genome Research Institute. https://www.genome.gov/genetics-glossary/Centromere

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Stamatoyannopoulos, J. A., Hunkapiller, M. W., … Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E. J., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Chen, N.-C., Solomon, B., Mun, T., Iyer, S., & Langmead, B. (2021). Reference flow: reducing reference bias using multiple population genomes. *Genome Biology*, *22*(1), 8. https://doi.org/10.1186/s13059-020-02229-3

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., GTEx Consortium, Montgomery, S. B., Battle, A., Conrad, D. F., & Hall, I. M. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, *49*(5), 692–699. https://doi.org/10.1038/ng.3834

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics (Oxford, England)*, *30*(1), 31–37. https://doi.org/10.1093/bioinformatics/btt310

Chin, C.-S., Behera, S., Khalak, A., Sedlazeck, F. J., Sudmant, P. H., Wagner, J., & Zook, J. M. (2023). Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nature Methods*, *20*(8), 1213–1221. https://doi.org/10.1038/s41592-023-01914-y

Choo, K. H. (2000). Centromerization. *Trends in Cell Biology*, *10*(5), 182–188. https://doi.org/10.1016/s0962-8924(00)01739-6

Church, D. M., Schneider, V. A., Steinberg, K. M., Schatz, M. C., Quinlan, A. R., Chin, C.-S., Kitts, P. A., Aken, B., Marth, G. T., Hoffman, M. M., Herrero, J., Mendoza, M. L. Z., Durbin, R., & Flicek, P. (2015). Extending reference assembly models. *Genome Biology*, *16*(1), 13. https://doi.org/10.1186/s13059-015-0587-3

Dai, X., Otake, K., You, C., Cai, Q., Wang, Z., Masumoto, H., & Wang, Y. (2013). Identification of novel α-n-methylation of CENP-B that regulates its binding to the centromeric DNA. *Journal of Proteome Research*, *12*(9), 4167–4175. https://doi.org/10.1021/pr400498y

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524. https://doi.org/10.1038/nbt.3423

De Rop, V., Padeganeh, A., & Maddox, P. S. (2012). CENP-A: the key player behind centromere identity, propagation, and kinetochore assembly. *Chromosoma*, *121*(6), 527–538. https://doi.org/10.1007/s00412-012-0386-5

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

Durai, D. A., & Schulz, M. H. (2016). Informed kmer selection for de novo transcriptome assembly. *Bioinformatics (Oxford, England)*, *32*(11), 1670–1677. https://doi.org/10.1093/bioinformatics/btw217

Earnshaw, W. C., & Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. *Chromosoma*, *91*(3–4), 313–321. https://doi.org/10.1007/bf00328227

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (New York, N.Y.)*, *372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, *54*(4), 518–525. https://doi.org/10.1038/s41588-022-01043-w

ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Mackiewicz, M., Pauli-Behn, F., … Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

Fachinetti, D., Han, J. S., McMahon, M. A., Ly, P., Abdullah, A., Wong, A. J., & Cleveland, D. W. (2015). DNA sequence-specific binding of CENP-B enhances the fidelity of human centromere function. *Developmental Cell*, *33*(3), 314–327. https://doi.org/10.1016/j.devcel.2015.03.020

Fukagawa, T., Mikami, Y., Nishihashi, A., Regnier, V., Haraguchi, T., Hiraoka, Y., Sugata, N., Todokoro, K., Brown, W., & Ikemura, T. (2001). CENP-H, a constitutive centromere component, is required for centromere targeting of CENP-C in vertebrate cells. *The EMBO Journal*, *20*(16), 4603–4617. https://doi.org/10.1093/emboj/20.16.4603

Fukagawa, T., & Brown, W. R. (1997). Efficient conditional mutation of the vertebrate CENP-C gene. *Human Molecular Genetics*, *6*(13), 2301–2308. https://doi.org/10.1093/hmg/6.13.2301

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

Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., Lipzen, A., Tracy, W. F., Mikel, M. A., Kaeppler, S. M., Buell, C. R., & de Leon, N. (2019). Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *The Plant Genome*, *12*(2), 180069. https://doi.org/10.3835/plantgenome2018.09.0069

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

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., Hagmann, J., Vorbrugg, S., Marco-Sola, S., Kubica, C., Ashbrook, … Prins, P. (2024). Building pangenome graphs. *Nature Methods*, *21*(11), 2008–2012. https://doi.org/10.1038/s41592-024-02430-3

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Gulko, B., Hubisz, M. J., Gronau, I., & Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics*, *47*(3), 276–283. https://doi.org/10.1038/ng.3196

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Haberer, G., Kamal, N., Bauer, E., Gundlach, H., Fischer, I., Seidel, M. A., Spannagl, M., Marcon, C., Ruban, A., Urbany, C., Nemri, A., Hochholdinger, F., Ouzunova, M., Houben, A., Schön, C.-C., & Mayer, K. F. X. (2020b). European maize genomes highlight intraspecies variation in repeat and gene content. *Nature Genetics*, *52*(9), 950–957. https://doi.org/10.1038/s41588-020-0671-9

Harrington, J. J., Van Bokkelen, G., Mays, R. W., Gustashaw, K., & Willard, H. F. (1997). Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nature Genetics*, *15*(4), 345–355. https://doi.org/10.1038/ng0497-345

Hattori M. (2005). Finishing the euchromatic sequence of the human genome. *Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme*, *50*(2), 162–168.

Hayden, K. E. (2012). Human centromere genomics: now it's personal. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *20*(5), 621–633. https://doi.org/10.1007/s10577-012-9295-y

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, *21*(1), 35. https://doi.org/10.1186/s13059-020-1941-7

Henikoff, J. G., Thakur, J., Kasinathan, S., & Henikoff, S. (2015). A unique chromatin complex occupies young α-satellite arrays of human centromeres. *Science Advances*, *1*(1), e1400234. https://doi.org/10.1126/sciadv.1400234

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., Knapp, S. J., Ware, D., Shapiro, B., Peluso, P., & Rank, D.

R. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, *7*(1), 399. https://doi.org/10.1038/s41597-020-00743-4

Jaegle, B., Pisupati, R., Soto-Jiménez, L. M., Burns, R., Rabanal, F. A., & Nordborg, M. (2023). Extensive sequence duplication in Arabidopsis revealed by pseudo-heterozygosity. *Genome Biology*, *24*(1), 44. https://doi.org/10.1186/s13059-023-02875-3

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-1122-x

Jansen, L. E. T., Black, B. E., Foltz, D. R., & Cleveland, D. W. (2007). Propagation of centromeric chromatin requires exit from mitosis. *The Journal of Cell Biology*, *176*(6), 795–805. https://doi.org/10.1083/jcb.200701066

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

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kasinathan, S., & Henikoff, S. (2018). Non-B-form DNA is enriched at centromeres. *Molecular Biology and Evolution*, *35*(4), 949–962. https://doi.org/10.1093/molbev/msy010

Katagiri, Y., & Tamaki, Y. (2021). Genetic counseling prior to assisted reproductive technology. *Reproductive Medicine and Biology*, *20*(2), 133–143. https://doi.org/10.1002/rmb2.12361

Kelley, D. R., Schatz, M. C., & Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, *11*(11), R116. https://doi.org/10.1186/gb-2010-11-11-r116

Kipling, D., & Warburton, P. E. (1997). Centromeres, CENP-B and Tigger too. *Trends in Genetics: TIG*, *13*(4), 141–145. https://doi.org/10.1016/s0168-9525(97)01098-6

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kokot, M., Dlugosz, M., & Deorowicz, S. (2017). KMC 3: counting and manipulating k-mer statistics. *Bioinformatics (Oxford, England)*, *33*(17), 2759–2761. https://doi.org/10.1093/bioinformatics/btx304

Kolář, F., Čertner, M., Suda, J., Schönswetter, P., & Husband, B. C. (2017). Mixed-ploidy species: Progress and opportunities in polyploid research. *Trends in Plant Science*, *22*(12), 1041–1055. https://doi.org/10.1016/j.tplants.2017.09.011

Kumon, T., Ma, J., Akins, R. B., Stefanik, D., Nordgren, C. E., Kim, J., Levine, M. T., & Lampson, M. A. (2021). Parallel pathways for recruiting effector proteins determine centromere drive and suppression. *Cell*, *184*(19), 4904-4918.e11. https://doi.org/10.1016/j.cell.2021.07.037

Kwon, M.-S., Hori, T., Okada, M., & Fukagawa, T. (2007). CENP-C is involved in chromosome segregation, mitotic checkpoint function, and kinetochore assembly. *Molecular Biology of the Cell*, *18*(6), 2155–2168. https://doi.org/10.1091/mbc.E07-01-0045

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., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lei, L., Goltsman, E., Goodstein, D., Wu, G. A., Rokhsar, D. S., Vogel, J. P. (2021). Plant pan-genomics comes of age. *Annual Review of Plant Biology, 72*(1), 411-435. https://doi.org/10.1146/arplant.2021.72.issue-1

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., 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

Liao, X., Zhu, W., & Liu, C. (2024). A high-precision genome size estimator based on the k-mer histogram correction. *Frontiers in Genetics*, *15*, 1451730. https://doi.org/10.3389/fgene.2024.1451730

Logsdon, G. A., & Eichler, E. E. (2023). The dynamic structure and rapid evolution of human centromeric satellite DNA. *Genes*, *14*(1), 92. https://doi.org/10.3390/genes14010092

Logsdon, G. A., Gambogi, C. W., Liskovykh, M. A., Barrey, E. J., Larionov, V., Miga, K. H., Heun, P., & Black, B. E. (2019). Human Artificial Chromosomes that Bypass Centromeric DNA. *Cell*, *178*(3), 624-639.e19. https://doi.org/10.1016/j.cell.2019.06.006

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., … Eichler, E. E. (2024). The variation and evolution of complete human centromeres. *Nature*, *629*(8010), 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., … 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

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

McNulty, S. M., Sullivan, L. L., & Sullivan, B. A. (2017). Human centromeres produce chromosome-specific and array-specific alpha satellite transcripts that are complexed with

CENP-A and CENP-C. *Developmental Cell*, *42*(3), 226-240.e6. https://doi.org/10.1016/j.devcel.2017.07.001

Miga, K. H. (2015). Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *23*(3), 421–426. https://doi.org/10.1007/s10577-015-9488-2

Miga, K. H., & Alexandrov, I. A. (2021). Variation and evolution of human centromeres: A field guide and perspective. *Annual Review of Genetics*, *55*(1), 583–602. https://doi.org/10.1146/annurev-genet-071719-020519

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., Schneider, … Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, *585*(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Miga, K. H., Newton, Y., Jain, M., Altemose, N., Willard, H. F., & Kent, W. J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Research*, *24*(4), 697–707. https://doi.org/10.1101/gr.159624.113

Moroi, Y., Peebles, C., Fritzler, M. J., Steigerwald, J., & Tan, E. M. (1980). Autoantibody to centromere (kinetochore) in scleroderma sera. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(3), 1627–1631. https://doi.org/10.1073/pnas.77.3.1627

Morozov, V. M., Giovinazzi, S., & Ishov, A. M. (2017). CENP-B protects centromere chromatin integrity by facilitating histone deposition via the H3.3-specific chaperone Daxx. *Epigenetics & Chromatin*, *10*(1), 63. https://doi.org/10.1186/s13072-017-0164-y

Nagpal, H., Ali-Ahmad, A., Hirano, Y., Cai, W., Halic, M., Fukagawa, T., Sekulić, N., & Fierz, B. (2023). CENP-A and CENP-B collaborate to create an open centromeric chromatin state. *Nature Communications*, *14*(1), 8227. https://doi.org/10.1038/s41467-023-43739-5

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., Powell, C. C., Nassar, L. R., Maulding, N. D., Lee, C. M., Barber, G. P., … Kent, W. J. (2021). The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*, *49*(D1), D1046–D1057. https://doi.org/10.1093/nar/gkaa1070

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., … Phillippy, A. M. (2021). The complete sequence of a human genome. In *bioRxiv*. https://doi.org/10.1101/2021.05.26.445798

Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., Miga, K. H., Eichler, E. E., Phillippy, A. M., & Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, *30*(9), 1291–1305. https://doi.org/10.1101/gr.263566.120

O'Connor, C. (2008). Chromosome Segregation in Mitosis: The Role of Centromeres. *Nature Education, 1*(1), 28. https://www.nature.com/scitable/topicpage/chromosome-segregation-in-mitosis-the-role-of-242/

Palsikowski, P. A., Roberto, M. M., Sommaggio, L. R. D., Souza, P. M. S., Morales, A. R., & Marin-Morales, M. A. (2018). Ecotoxicity evaluation of the biodegradable polymers PLA, PBAT and its blends using Allium cepa as test organism. *Journal of Polymers and the Environment*, *26*(3), 938–945. https://doi.org/10.1007/s10924-017-0990-9

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

Ponsero, A. J., Miller, M., & Hurwitz, B. L. (2023). Comparison of k-mer-based de novo comparative metagenomic tools and approaches. *Microbiome Research Reports*, *2*(4), 27. https://doi.org/10.20517/mrr.2023.26

Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Munson, K. M., Sorensen, M., Sulovari, A., Devine, S. E., Sanders, A. D., … Marschall, T., & Human Genome Structural

Variation Consortium. (2019). A fully phased accurate assembly of an individual human genome. In *bioRxiv*.

Rahman, A., Hallgrímsdóttir, I., Eisen, M., & Pachter, L. (2018). Association mapping from sequencing reads using k-mers. *eLife*, *7*. https://doi.org/10.7554/eLife.32920

Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A., & Friedman, J. M. (2024). Sequence composition changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical implications. *Nature Reviews. Genetics*, *25*(7), 476–499. https://doi.org/10.1038/s41576-024-00696-z

Rice, E. S., Alberdi, A., Alfieri, J., Athrey, G., Balacco, J. R., Bardou, P., Blackmon, H., Charles, M., Cheng, H. H., Klopp, C., Marcos, S., Mason, A. S., … Warren, W. C. (2023). A pangenome graph reference of 30 chicken genomes allows genotyping of large and complex structural variants. *BMC Biology*, *21*(1), 267. https://doi.org/10.1186/s12915-023-01758-0

Roberts, M. D., Davis, O., Josephs, E. B., Williamson, R. J. (2024). K-mer-based Approaches to Bridging Pangenomics and Population Genetics. *Molecular Biology and Evolution, 42,* 1-15. https://doi.org/10.1093/molbev/msaf047

Rosenbloom, K. R., Armstrong, J., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Lee, B. T., Li, C. H., … Kent, W. J. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research*, *43*(Database issue), D670-81. https://doi.org/10.1093/nar/gku1177

Saitoh, H., Tomkiel, J., Cooke, C. A., Ratrie, H., 3rd, Maurer, M., Rothfield, N. F., & Earnshaw, W. C. (1992). CENP-C, an autoantigen in scleroderma, is a component of the human inner kinetochore plate. *Cell*, *70*(1), 115–125. https://doi.org/10.1016/0092-8674(92)90538-n

Schmuths, H., Meister, A., Horres, R., & Bachmann, K. (2004). Genome size variation among accessions of Arabidopsis thaliana. *Annals of Botany*, *93*(3), 317–321. https://doi.org/10.1093/aob/mch037

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, Chow, W., Collins, J., … Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., Kuk, J., Park, G. H., Hunkapiller, M. W., Korlach, J., … Kim, C. (2016). De novo assembly and phasing of a Korean human genome. *Nature*, *538*(7624), 243–247. https://doi.org/10.1038/nature20098

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), 1044–1053. https://doi.org/10.1038/s41587-020-0503-6

Shajii, A., Yorukoglu, D., William Yu, Y., & Berger, B. (2016). Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics (Oxford, England)*, *32*(17), i538–i544. https://doi.org/10.1093/bioinformatics/btw460

Shumate, A., Zimin, A. V., Sherman, R. M., Puiu, D., Wagner, J. M., Olson, N. D., Pertea, M., Salit, M. L., Zook, J. M., & Salzberg, S. L. (2020). Assembly and annotation of an Ashkenazi human reference genome. *Genome Biology*, *21*(1), 129. https://doi.org/10.1186/s13059-020-02047-7

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., J. I., Haussler, D., … Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science (New York, N.Y.)*, *374*(6574), abg8871. https://doi.org/10.1126/science.abg8871

Smith, G. P. (1976). Evolution of repeated DNA sequences by unequal crossover: DNA whose sequence is not maintained by selection will develop periodicities as a result of random crossover. *Science (New York, N.Y.)*, *191*(4227), 528–535. https://doi.org/10.1126/science.1251186

Song, B., Buckler, E. S., & Stitzer, M. C. (2024). New whole-genome alignment tools are needed for tapping into plant diversity. *Trends in Plant Science*, *29*(3), 355–369. https://doi.org/10.1016/j.tplants.2023.08.013

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S., Zhou, R., Xie, W.-Z., Cheng, Y., Zhang, Y., Liu, K., Yang, Q.-Y., Chen, L.-L., & Guo, L. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, *6*(1), 34–45. https://doi.org/10.1038/s41477-019-0577-7

Steinberg, K. M., Schneider, V. A., Graves-Lindsay, T. A., Fulton, R. S., Agarwala, R., Huddleston, J., Shiryev, S. A., Morgulis, A., Surti, U., Warren, W. C., Church, D. M., Eichler, E. E., & Wilson, R. K. (2014). Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Research*, *24*(12), 2066–2076. https://doi.org/10.1101/gr.180893.114

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or genomical? *PLoS Biology*, *13*(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

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

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tanudisastro, H. A., Deveson, I. W., Dashnow, H., & MacArthur, D. G. (2024). Sequencing and characterizing short tandem repeats in the human genome. *Nature Reviews. Genetics*, *25*(7), 460–475. https://doi.org/10.1038/s41576-024-00692-3

Tomkiel, J., Cooke, C. A., Saitoh, H., Bernat, R. L., & Earnshaw, W. C. (1994). CENP-C is required for maintaining proper kinetochore size and for a timely transition to anaphase. *The Journal of Cell Biology*, *125*(3), 531–545. https://doi.org/10.1083/jcb.125.3.531

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics: TIG*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Human Pangenome Reference Consortium. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Westhorpe, F. G., & Straight, A. F. (2015). The centromere: epigenetic control of chromosome segregation during mitosis. *Cold Spring Harbor Perspectives in Biology*, *7*(1), a015818. https://doi.org/10.1101/cshperspect.a015818

Willard, H. F. (1990). Centromeres of mammalian chromosomes. *Trends in Genetics: TIG*, *6*, 410–416. https://doi.org/10.1016/0168-9525(90)90302-m

Wong, K. H. Y., Ma, W., Wei, C.-Y., Yeh, E.-C., Lin, W.-J., Wang, E. H. F., Su, J.-P., Hsieh, F.-J., Kao, H.-J., Chen, H.-H., … Kwok, P.-Y. (2020). Towards a reference genome that captures global genetic diversity. *Nature Communications*, *11*(1), 5482. https://doi.org/10.1038/s41467-020-19311-w

Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J., & Wallace, R. B. (1991). The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA and Cell Biology*, *10*(3), 233–238. https://doi.org/10.1089/dna.1991.10.233

Xie, N. (2024). Building a catalogue of short tandem repeats in diverse populations. *Nature Reviews. Genetics*, *25*(7), 457–457. https://doi.org/10.1038/s41576-024-00726-w

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L. B., & Reese, M. G. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, *21*(9), 1529–1542. https://doi.org/10.1101/gr.123158.111

Yang, F., Moss, L. G., & Phillips, G. N., Jr. (1996). The molecular structure of green fluorescent protein. *Nature Biotechnology*, *14*(10), 1246–1251. https://doi.org/10.1038/nbt1096-1246

Zaratiegui, M., Vaughn, M. W., Irvine, D. V., Goto, D., Watt, S., Bähler, J., Arcangioli, B., & Martienssen, R. A. (2011). CENP-B preserves genome integrity at replication forks paused by retrotransposon LTR. *Nature*, *469*(7328), 112–115. https://doi.org/10.1038/nature09608

Zerbino, D. R., Frankish, A., & Flicek, P. (2020). Progress, challenges, and surprises in annotating the human genome. *Annual Review of Genomics and Human Genetics*, *21*(1), 55–79. https://doi.org/10.1146/annurev-genom-121119-083418

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., Zhang, J., Lyu, H., Lin, T., Gao, Q., Saha, S., Mueller, L., Fei, Z., Städler, T., Xu, S., … Huang, S. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, *606*(7914), 527–534. https://doi.org/10.1038/s41586-022-04808-9

Kasutatud veebiaadressid

Homo sapiens genome assembly T2T-CHM13v2.0. (n.d.). NCBI. (21.05.2025) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/

HPRC data release 2. (n.d.). Human Pangenome Reference Consortium. (25.05.2025) https://humanpangenome.org/hprc-data-release-2/

Human Pangenome Reference Consortium (21.05.2025) https://humanpangenome.org/

Lisad

Lisa 1. CHM13 assembli põhised sagedaseimad kromosoomi tsentromeeri spetsiifilised k-meerid.

Kromosoom	k-mer
1	GAGTGTTTCCAAACCGCTGAATGAA
2	ATTTGTGATGTGCGCCCTCAACTAA
3	GTGAGTTGAATACACACACACAGAA
4	AAATATCTTCTCATAAAAACCAGAA
5	AGAGTTGAACGCACACATCGCAGAG
6	AACCCAGACAGAAGAATTCTCAGAG
7	AAAGGGACGTTCCACTCTGTGAGTT
8	CACTGTTAGTTGAGTACCCACATCA
9	TAAAAAGAAAGGTTCATCTCTGTTA
10	AAAAACTACACAGAATCATTCTCAA
11	AATGCTTCCGTTTGGTTTTTAGATG
12	AGAATTCTCAGTAACTTATTTGTGG
13	AGATTCTTCCAAAAGAGTGTTTGAA
14	CTTCCTTGTGATATGTGCATTCAAG
15	AATTTTCAATGCTCTCAAAATATCC
16	CTAAAGGAAGGTTCAACTCTGTCAG
17	AGAATGCTTCTGTTTAGTTCTGTGC
18	ACCTGCTCTACCAAAGGGAATGTTC
19	AAAGCCTCAAGGATGTCTGAATATC
20	AGAATGCTTCTGTCTAGAGTTTATC
21	TTCTTTTGATGCAGCAATTTGGAAA
22	CGTTTCAAAGAGCAGCTTTGAGGCA
X	AACTGCTCTGTGATGATTGCATTCA
Y	TATCGTTTGAGAGAGCATTTCGAAA

Lisa 2. Kromosoomi põhised tsentromeeri k-meere sisaldavate summarne kontiigide arv kõikide pangenoomi v1 indiviididelt ning korrelatsioonikoefitsiendid k-meeri seostumisarvude ja tsentromeeri pikkuste vahel.

Kromosoom	**Tsentromeeri k-meri sisaldavate kontiigide arv**	**Korrelatsioon**
1	143	0,94380013

2	118	0,92164314
3	669	0,63521213
4	423	0,87044479
5	297	0,99303239
6	605	0,99899088
7	395	0,99719016
8	262	0,99164353
9	181	0,99812378
10	213	0,98263648
11	164	0,99839595
12	238	0,99580518
13	93	0,50759841
14	183	0,90101624
15	299	0,98693562
16	290	0,96421167
17	612	0,96192911
18	837	0,99077402
19	168	0,8652015
20	328	0,99615288
21	208	0,98990474
22	45	0,92100677
X	167	0,99799393
Y	14	0,97729724

Lisa 3. Pangenoomi haplotüpiseeritud kontiigide arv, mis sisaldab chr9 tsentromeerseid järjestusi. HG01106 nii isalt (1) ja emalt (2) parit chr9 tsentromeer on samalt harult evolutsioonipuult (andmed juhendajalt).

Indiviidid	**Kontiigid isalt**	**Kontiigid emalt**
HG002	4	3
HG00438	1	1
HG005	2	1
HG00621	2	1
HG00673	1	2
HG00733	4	2

HG00735	1	1
HG00741	2	1
HG01071	1	1
HG01106	1	1
HG01109	2	2
HG01123	2	1
HG01175	2	1
HG01243	2	3
HG01258	2	5
HG01358	1	1
HG01361	2	1
HG01891	6	2
HG01928	3	1
HG01952	2	1
HG01978	3	2
HG02055	1	3
HG02080	1	3
HG02109	2	1
HG02145	2	2
HG02148	3	1
HG02257	3	1
HG02486	1	5
HG02559	1	1
HG02572	1	2
HG02622	3	1
HG02630	1	3
HG02717	2	1
HG02723	3	2
HG02818	3	1
HG02886	1	4
HG03098	1	4
HG03453	3	1
HG03486	1	1
HG03492	1	1
HG03516	1	1

HG03540	3	2
HG03579	2	2
NA18906	2	3
NA19240	1	2
NA20129	5	1
NA21309	1	3

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina Carmen Beljaev,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Kromosoomipõhise k-meer sageduse ja tsentromeeri pikkuse vahelise korrelatsiooni leidmine,

mille juhendaja on Tarmo Puurand,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi 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.

Carmen Beljaev
26.05.2025