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

Inimese genoomi suuruse määramine k-meer metoodikaga

Bakalaureusetöö

Lõputöö maht (12 EAP)

Sylvia Krupp

Juhendaja MSc Tarmo Puurand

TARTU 2018

INFOLEHT

Inimese genoomi suuruse määramine k-meer metoodikaga

Uurimistöö eesmärgiks on inimese genoomi suuruse hindamine k-meer metoodikaga. Teoreetilises pooles tehakse ülevaade genoomide suurusest, seal hulgas inimese genoomist, selle suurusest koos varieeruvust põhjustavate aladega ning suuruse määramise meetoditest. Genoomi suuruse mõõtmine on oluline, et mõõta kromosoomide evolutsiooni ja et edukalt viia läbi laborikatseid, kus DNA hulk määrab katse tulemuse kvaliteedi. Praktilises osas määratakse genoomi suurus 50 mehel ja 50 naisel Tartu Ülikooli Eesti geenivaramu (TÜ EGV) täisgenoomi sekveneeritud indiviidide seast. Mõõdetud genoomide suurus jäi vahemikku 3,0-3,1 Gbp-d meestel ja 3,0-3,3 Gbp-d naistel. Sama protokolli kasutatud indiviidide sekveneerimisandmete põhjal findGSE programmiga mõõdetud genoomi suurused on seda suuremad, mida madalama katvusega sekveneerimistsükkel läbi on viidud.

Märksõnad: genoomi suurus, k-meer metoodika, findGSE, katvus

CERCS kood: B110

Human genome size evaluation with k-mer method

The purpose of the research is to evaluate the size of the genome with k-mer method. The theoretical part will give an overview of the genome sizes, including human genome, its size with variable areas, and size determination methods. Genome size evalution is important, because it measures the evolution of chromosomes and successfully conduct lab tests, where the amount of DNA determines the quality of the test result. In the practical part genome sizes of 50 men and 50 women from Estonian Genome Center at the University of Tartu (EGCUT) whole genome sequenced individuals are determined. The size of the measured genomes varied 3.0-3.1 Gbp in men and 3.0-3.3 Gbp in women. Based on the individuals' sequencing data used in the same protocol, the genome sizes measured by findGSE program are larger when the coverage is low for sequencing cycle.

Keywords: genome size, k-mer method, findGSE, coverage

CERCS code: B110

SISUKORD

KASUTATUD LÜHENDID

bp – aluspaar (*base pair*)

CEU – Põhja- ja Lääne-Euroopa päritoluga Utah elanikud (*Centre d'Etude du Polymorphisme Humain*)

CN – koopiaarv (*Copy-Number*)

CNV – koopiaarvu variatsioon (*Copy-Number Variation*)

LINE – pikk insertsiooniline hajuskorduselement (*Long interspersed nuclear element*)

LTR – DNA kordusjärjestused, mis esinevad retroviiruse DNA mõlemas otsas (*Long terminal repeat*)

mtDNA – mitokondriaalne DNA

qPCR – kvantitatiivne reaalaja PCR (*quantitative real-time Polymerase Chain Reaction*)

rDNA – DNA järjestus, mis kodeerib ribosomaalset RNA-d

SINE – lühike insertsiooniline hajuskorduselement (*Short interspersed nuclear element*)

tRNA – transpordi-RNA

YRI – Yoruba, indiviidid Aafrikast

SISSEJUHATUS

Genoom on organismis sisalduv DNA kogus. Selle suurus võib liigiti palju varieeruda. Genoomi suurust saab määrata nii keemiliste kui arvutuslike meetoditega. Üheks arvutuslikuks meetodiks on k-meeride kasutamine, kus vaadeldakse kõikvõimalikke k-meeride hulka ja nende sageduste jaotust sekveneerimisandmetes. Samuti saab k-meeridega määrata genoomi komponente, mille kohta seni ei osatud täpselt öelda, kuidas nende hulk indiviidide vahel varieerub.

Genoomi suuruse hindamise meetodid on vajalikud, et analüüsida suures hulgas liike, indiviide või kudesid ning uurida muutusi genoomi suuruses fülogeneesil. K-meere kasutades üritan teada saada genoomi kogusuurust ning CNV-de, heterokromatiinide ja geenide hulka genoomis indiviiditi. Selle metoodikaga peaks olema võimalik täpsemalt ja kiiremini määrata inimese tegelikku täisgenoomi suurust, mis hetkel erinevate metoodikate põhjal varieerub vahemikus 2,9-3,7 Gbp.

Tänan oma juhendajat Tarmo Puurand'a, kes oli abiks töö koostamisel. Samuti soovin tänada veel Bioinformaatika õppetooli töötajaid, kes aitasid töö valmimisele kaasa.

1. KIRJANDUSE ÜLEVAADE

1.1 Genoomide suurused

Genoomi suurus on genoomi ühes rakus sisalduv DNA kogus. Seda mõõdetakse aluspaarides või pikogrammides. Üks pikogramm on võrdne 978 mega aluspaariga (Gregory et al., 2007). Diploidses organismis kasutatakse genoomi suurust C-väärtusega (DNA kogus diploidse organismi ühes rakus) vaheldumisi. Iga genoom sisaldab geneetilist informatsiooni, mis juhib kasvu, arengut ja tervist. Seda informatsiooni nimetatakse DNA-ks. DNA koosneb adeniinist (A), guaniinist (G), tsütosiinist (C) ja tümiinist (T). Igal organismil on unikaalne genoom (https://www.yourgenome.org/facts/what-is-a-genome).

Umbes 40 aastat tagasi arvati, et DNA kogus genoomis on vastavuses organismi keerulisusega. Idee seisnes selles, et mida kompleksem on liik, seda rohkem geene ta vajab, mille tõttu ka genoom on suurem. 1960ndatel hakkasid teadlased uurima lähemalt genoomi enda spetsiifilisust. Avastati, et inimesel on unikaalse DNA järjestuse osakaal vaid paar protsenti suure genoomi koguosast, kust sai alguse mõiste „rämps-DNA“ (Mattick, 2004). Tänapäeval me teame, et genoomi suurus ja organismi keerulisus pole omavahel otseses vastavuses (Taft *et al.*, 2007). Samuti on teada, et liigi siseselt genoomi suurus varieerub (Ryan Gregory, 2005). Viirustel, bakteritel ja organismi mitokondril on tavaliselt geneetiline informatsioon genoomis ökonoomselt pakitud, mille tõttu nende suurus on väike. Eukarüootidel, eriti selgroogsetel, sisaldavad genoomid tavaliselt palju korduvaid järjestusi, mis on üks märkimisväärsemaid põhjuseid, miks on genoomid suured (tabel 1).

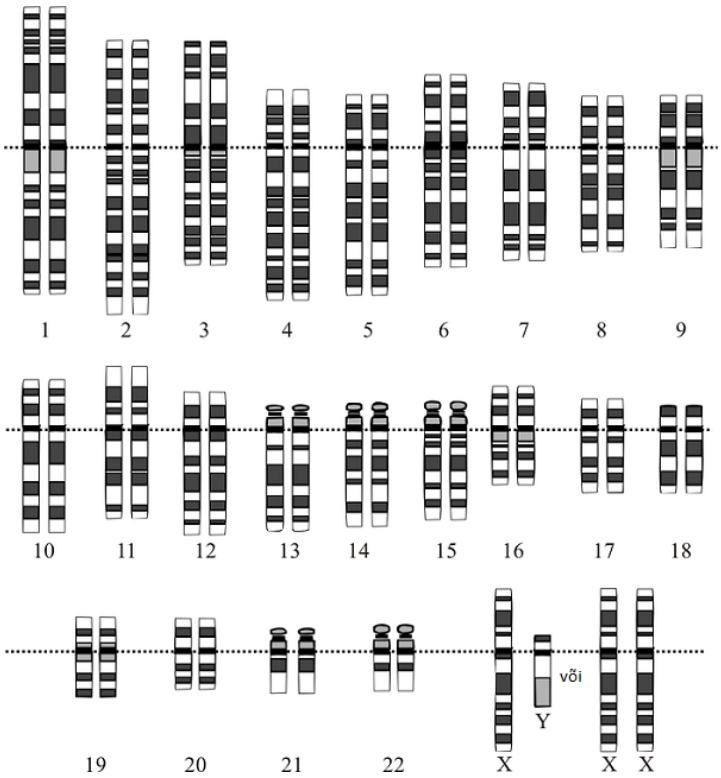
Tabel 1. Genoomi suurused eri organismidel. Inimese genoomi suurus on umbes 3 Gbp, kõige suurem looma genoom on kopskalal 130 Gbp, suurim eukarüootne genoom on taimel *Paris Japonica* – 150 Gbp. Härgkonna genoom on umbes 2 korda suurem kui inimesel. Päevalille genoom on inimese omaga samas suurusklassis.

Organismi tüüp	Ladinakeelne liiginimetus	Eestikeelne liiginimetus	Genoomi suurus (Gbp)	Viide
Bakter	*Staphylococcus aureus*	-	2,8	https://www.ncbi.nlm.nih.gov/genome/?term=staphylococcus+aureus%5Borgn%5D
Bakter	*Escherichia coli*	-	4,6	(Blattner, 1997)
Putukas	*Drosophila melanogaster*	Äädikakärbes	143,9	https://www.ncbi.nlm.nih.gov/genome/?term=drosophila%20melanogaster
Putukas	*Anopheles gambiae*	Sääsk	250,7	https://www.ncbi.nlm.nih.gov/genome/46
Taim	*Solanum lycopersicum*	Tomat	791,9	https://www.ncbi.nlm.nih.gov/genome/?term=Solanum+lycopersicum%5Borgn%5D
Kala	*Oncorhynchus mykiss*	Vikerforell	2,0	https://www.ncbi.nlm.nih.gov/genome/?term=Oncorhynchus+mykiss%5Borgn%5D
Imetaja	*Felis catus*	Kass	2,8	https://www.ncbi.nlm.nih.gov/genome/?term=Felis+catus%5Borgn%5D
Taim	*Helianthus*	Päevalill	3,0	https://www.ncbi.nlm.nih.gov/genome/?term=Helianthus%5Borgn%5D
Imetaja	*Homo sapiens*	Inimene	2,9	https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens%5Borgn%5D
Kahe-paikne	*Rana catesbeiana*	Härgkonn	6,0	https://www.ncbi.nlm.nih.gov/genome/?term=Rana+catesbeiana
Kala	*Protopterus aethiopicus*	Kopskala	131,0	(Leitch, 2007)
Taim	*Paris japonica*	-	151,0	(Pellicer et al., 2010)

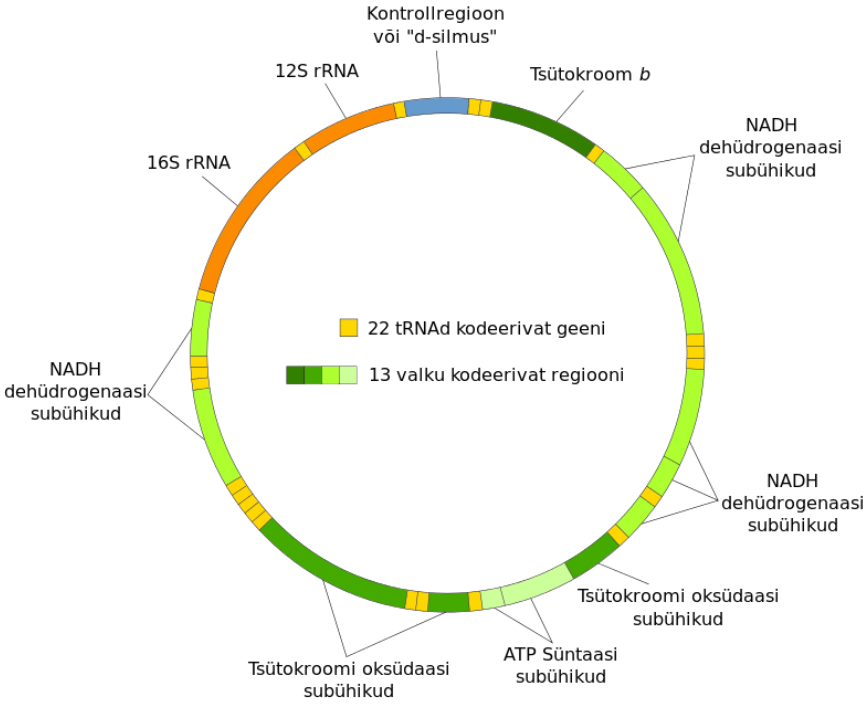
(Blattner, 1997)(Leitch, 2007)(Pellicer *et al.*, 2010)

1.2 Inimese genoom

Genoom on organismi geneetiline materjal. Inimese geneetiline materjal on inimese raku tuumas ja mitokondris. Enamasti on inimeste rakkude tuumades 46 kromosoomi (2n; 22+X, 22+Y) (joonis 1) (Brown, 2002). Ühe raku mitokondrites on sõltuvalt raku tüübist kromosoome arvuliselt väga varieeruvalt (Boyle, 2008). Kui mitokondris olev kromosoom on 16 569 bp pikkune ning moodustab genoomis 37 geeni (joonis 2) (Anderson *et al.*, 1981), siis haploidse raku tuumas olevad kromosoomid sisaldavad üle 3 miljardi bp (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38) jagu erinevaid piirkondi ja vaid 1,5% valke kodeerivaid geene (joonis 3) (Gregory, 2005). Genoom sisaldab nii geene kui ka mittekodeerivat DNA-d (Brosius, 2009), samuti mitokondri geneetilist materjali (Ridley, 2013).

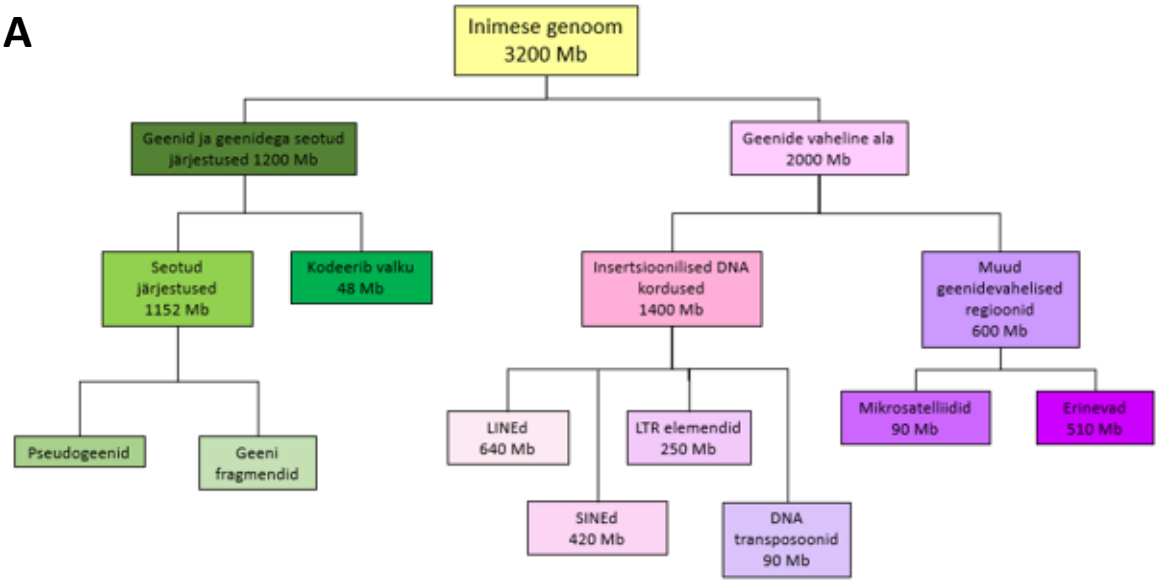


Joonis 1. Inimese kromosoomistiku skeem. Joonisel on 22 paari homoloogilisi kromosoome ja 2 sugukromosoomi. Punktiirjoon tähistab tsentromeere, heterokromatiinid on halli värvi, musta-valge vöödilised on eukromatiinid (http://bio3400.nicerweb.com/Locked/media/ch07/Y_chromosome.html). Eukromatiinid muudab musta-valge vöödiliseks Giemsa värv – alad, mis on madala GC sisaldusega, värvuvad mustaks ning alad, kus on kõrge GC sisaldus, ei värvu, mille tõttu jäävad need kohad kromosoomis valgeks (http://www.garlandscience.com/res/pdf/9780815341499_ch09.pdf). Kromosoomid on nummerdatud pikkuse järjekorras - kromosoom 1 on kõige pikem ning 22 kõige lühem. Lisaks on lõppu pandud ka XY ja XX kromosoomid, mis tähistavad sugu. XY on normaalsel mehel ning XX normaalsel naisel (https://biologydictionary.net/homologous-chromosomes/). Kromosoomides numbritega 13, 14, 15, 21, 22 asub rDNA (Gosden *et al.*, 1981).



Joonis 2. Inimese mtDNA. Inimesel on mitokondriaalne DNA ringikujuline. mtDNA suurus on ca 16 000 bp, sisaldab 37 geeni, millest 13 kodeerivad valke, 22 kodeerivad tRNA-sid ja 2 kodeerivad rRNA suurt ja väikest alamühikut (http://www.norwaydna.no/wp-content/uploads/2013/10/?C=N;O=A).

A



B

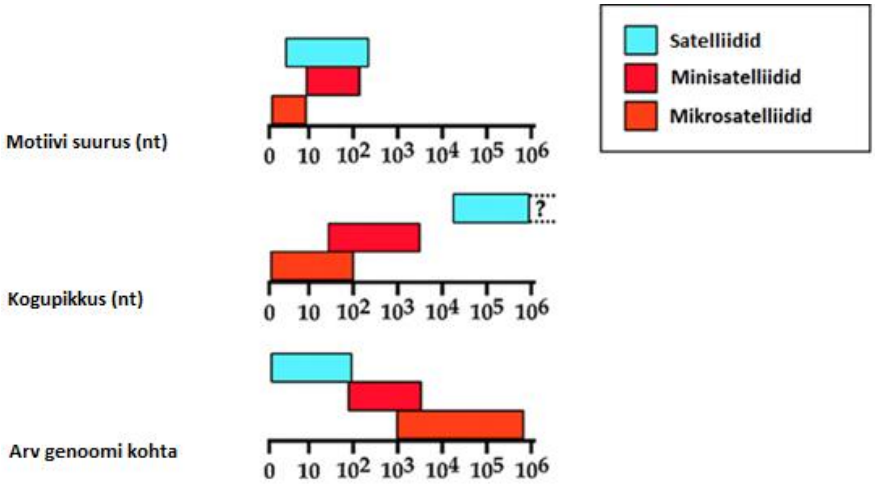
Genoomi komponentide suurused	
Mitokondri genoom	16,6 kb
Tuumagenoom	3,1 Gb
Eukromatiin	2,9 Gb (≈93%)
Valku kodeerivad DNA järjestused	≈35 Mb (≈1,1%)
kõrgkordus-DNA	≈1,6 Gb (≈50%)
Heterokromatiin	≈200 Mb (≈7%)
Transposoonipõhised kordused	≈1,4 Gb (≈45%)
DNA kromosoomi kohta	48 Mb - 249 Mb
Geenide arv	
Tuumagenoom	>26 000
Mitokondri genoom	37
Valku kodeerivad geenid	≈20 000 - 21 000
RNA geenid	>6000
Valku kodeerivate geenidega seotud pseudogeenid	>12 000

Joonis 3. Inimese genoomi komponentide jaotus. Joonise A osas on näidatud, kuidas paljud DNA aluspaarid on jaotunud mitmete erinevate tuvastatavate funktsioonide vahel inimese haploidses genoomis. Ainult väike osa genoomist on otseselt seotud valku kodeerivate piirkondadega (Brown, 2002). Umbes 1,5% genoomist koosneb ligikaudu 20 000 valku kodeerivast järjestusest. Ülejäänud osa koosneb pseudogeenidest, mikrosatelliitidest, transposoonidest ja muudest korduvatest elementidest (Strachan ja Read, 2004). Transposoonide hulka kuuluvad näiteks pikad insertsioonilised hajuskorduselemendid (LINE) ja lühikesed insertsioonilised hajuskorduselemendid (SINE) (Gregory, 2005). „Erinevad“ elementide alla kuuluvad tsentromeerid, telomeerid ja heterokromatiin. Joonise B osas on samuti välja toodud inimese genoomi komponendid. Inimese genoom koosneb mitokondri genoomist ja tuumagenoomist. Geenid sisaldavad korduvat DNA-d, kuid suurem osa kõrgkorduvast DNA-st asub geenidest väljaspool. Sellist tüüpi DNA-d nimetatakse heterokromatiiniks. Umbes 93% genoomist moodustab eukromatiin ja ülejäänud

heterokromatiin. Heterokromatiin hõlmab alasid tsentromeerides ja telomeerides kõigis kromosoomides. Suurem osa Y-kromosoomi ning akrotsentriliste (kromosoomi lühike õlg on väga lühike) kromosoomide lühikesi õlgu koosnevad heterokromatiinist. Teisi sarnaseid DNA järjestusi nimetatakse transposooni kordusteks ja neid on inimese genoomi eri piirkondades ligikaudu 45% (http://www.garlandscience.com/res/pdf/9780815341499_ch09.pdf).

Mitmed töögrupid on erinevatel aegadel hinnanud inimese haploidse genoomi suuruseks 2,9-3,7 Gbp (Venter et al., 2001; Brown, 2002; Dolezel *et al*., 2003; Wilhelm, 2003). Varieeruda saab eukromatiin, kus geenide all olev ala on kordistunud kas korra või rohkem. Samuti on võimalik varieeruda heterokromatiinil, kus tandeemselt paiknev satelliitne järjestus on rekombineerumise käigus lühenenud või pikenenud perioodiliselt korduva motiivi võrra. (http://www.garlandscience.com/res/pdf/9780815341499_ch09.pdf)

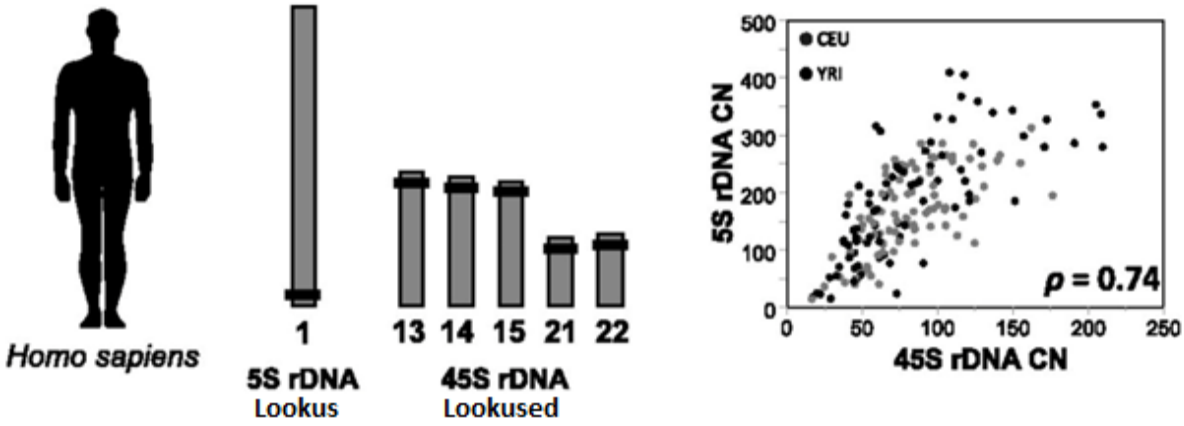
Sagelikorduv DNA asub intronites. Pikkuse ja korduste arvule vastavalt nimetatakse neid kordusi mini-, mikro- või lihtsalt satelliitideks (joonis 4). Satelliidid on pikad, kuni mitmetuhandest kbp-st koosnevad tandeemsed korduvad järjestused. Minisatelliidid on tandeemsed korduvad järjestused, mille pikkus jääb vahemikku 1-15 kbp. Mikrosatelliidid on suuruses 4 kbp või vähem. Satelliitjärjestused koosnevad 171 bp tandemkordusest ning asuvad tsentromeeri ja seda ümbritsevatel aladel kõigis primaatide kromosoomides (Ugarković, 2013). Satelliit-DNA koos minisatelliitide ja mikrosateliit-DNAga moodustavad tandemkordused (Kass ja Batzer, 2001).



Joonis 4. Eukarüootide satelliitjärjestuste suurused, pikkus ja arv genoomis. Igale kategooriale (satelliidid, minisatelliidid, mikrosatelliidid) on esitatud logaritmilise skaala järgi motiivi suurused, kogupikkused ja esinemissagedused eukarüootse genoomi kohta. Satelliit-

DNA võib ulatuda megaaluspaarideni, kuid selle maksimaalne pikkus on teadmata järjestuse andmete puudulikkuse tõttu (punktiirjoon ja küsimärk) (Richard *et al.*, 2008).

Ribosomaalsed RNA-d moodustavad eukarütoosestes rakkudes üle 60% kõigist RNA-dest ja need kodeeritakse ribosomaalse DNA massiivides (rDNA-s). rRNA-d valmistatakse kahest lookuse komplektist: 5S rDNA massiiv paikneb inimese 1. kromosoomis, 45S rDNA paikneb viies inimese akrotsentrilises kromosoomis (Yu ja Lemos, 2016), mis tähendab, et p-õlg on väga lühike (Nussbaum *et al.*, 2016). Eukarüootne ribosoom koosneb umbes 80 valgust ja 4 rRNA molekulist. 80S ribosoom koosneb subühikustest 40S, mis sisaldab 18S rRNA-d ja suurest 60S, mis koosneb 28S, 5.8S ja 5S rRNA-st, millel igaühel on eri funktsioon (Rabl *et al.*, 2011). Inimestel kodeeritakse 18S, 5.8S ja 28S rRNA molekule rDNA geenidelt, mis asuvad 60-800 kbp tandeemsete blokkidena viie kromosoomi otstes (13, 14, 15, 21 ja 22) (https://www.ncbi.nlm.nih.gov/gene/100008588). 5S rRNA-d kodeeritakse umbes 2,2 kbp tandemkorduste klastris kromosoomil 1 (joonis 5) (Sorensen ja Frederiksen, 1991).

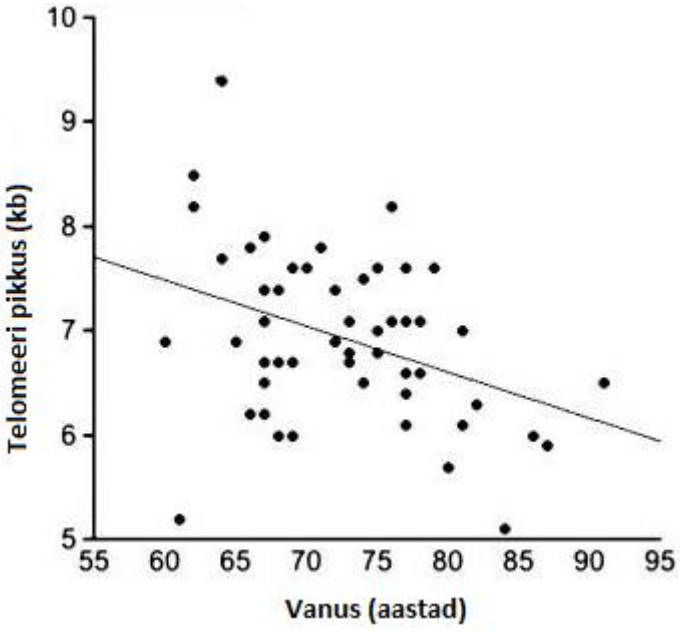


Joonis 5. Inimese 5S ja 45S rDNA koopiaarvu vastavus indiviiditi. 5S rDNA asub esimeses kromosoomis (Yu & Lemos, 2016) ning on ligikaudu 120 nukleotiidi pikk (Pelham & Brown, 1980). 45S rDNA massiiv paikneb viie inimese akrotsentrilise kromosoomi lühikesel õlal. Kui näiteks 5S rDNAd on 100 tükki, siis sellele vastab umbes 50 45S rDNAd. Enamasti muutub 5S rDNA kogus. Hallid ristkülikud tähistavad kromosoome ja mustad ristkülikud nendel 5S ja 45S rDNA lookuste ligikaudset asukohta kromosoomil. X-telg kirjeldab 45S rDNA-d ja y-telg 5S rDNA-d. Nende koopiaarvud (CN) on kõrgelt korrelleerunud. Mustad ja hallid täpid on vastavuses Euroopa (CEU) ja Aafrika (YRI) inimpopulatsioonidega (Gibbons *et al.*, 2015).

Telomeerid koosnevad kuni mitmest tuhandest lühikese korduva struktuuriga järjestusest. Inimese puhul on nendeks TTAGGG järjestused, mis on omavahel seotud

spetsiifiliste valkudega (Riethman, 2008). Telomeerid asuvad kromosoomide mõlemas otsas ning nendel on tähtis roll informatsiooni säilitamisel genoomis. Telomeeri pikkust mõjutab vanus (joonis 6) (Shammas, 2011).

Inimese tsentromeerid paiknevad korduva alfa-satelliidi DNA massiivides, mis moodustavad ligikaudu 5% genoomist (Aldrup-MacDonald ja Sullivan, 2014). Tsentromeerid on spetsiifilised kromosoomi DNA järjestused, mis seovad tütarkromatiide (Alberts *et al*., 2002).



Joonis 6. Telomeeri pikkuse ja vanuse suhe. Telomeeri pikkuse lühenemine on seotud vanusega. Mida vanem on inimene, seda lühem on telomeer (Hochstrasser *et al.*, 2012)**.** Järk-järguline telomeeride lühenemine mõjutab inimese tervist ja eluiga. Lühemaid telomeere on seostatud haiguste sagenemise ja tervise halvenemisega. Telomeeri lühenemise kiirust saab mõjutada elustiiliga. Tervislik toitumine ja aktiivne eluviis takistab telomeeride liigset hõrenemist, mis võimaldab pikemat eluiga (Shammas, 2011).

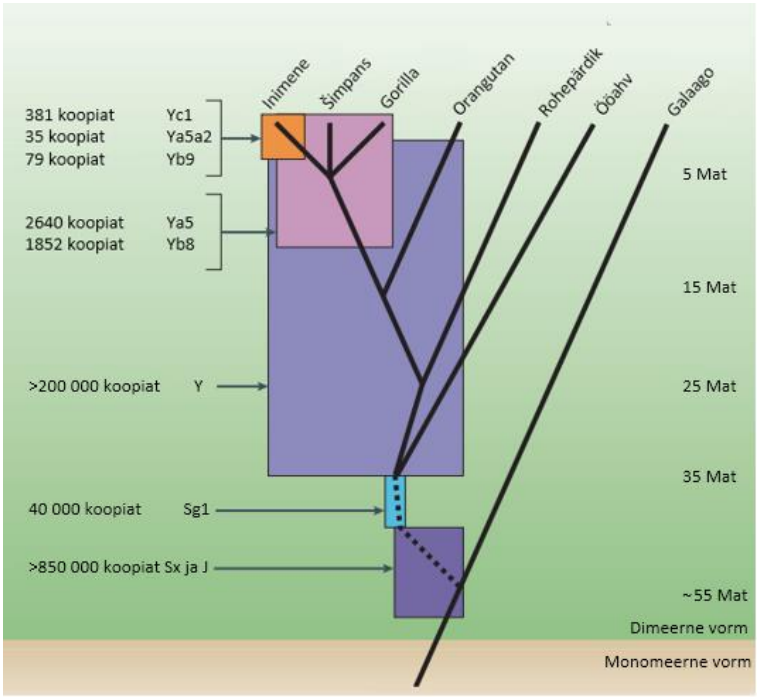
Laialdane DNA sekveneerimine on näidanud, et enamik korduvat DNA-d pärineb transponeeruvatest elementidest – järjestused, mis suudavad liikuda ja replitseeruda genoomis (Wicker et al., 2007), moodustades 45% inimese genoomist (tabel 2) (International Human Genome Sequencing Consortium, 2001).

Tabel 2. Insertsiooniliste korduste koopia-, aluspaaride arv ja fraktsioonide protsent inimese esimese versiooni referentsgenoomi põhjal. Imetajatel jagunevad enamus transpositsioonilistest elementidest nelja põhiklassi ning nende alamklassidesse. Neli põhiklassi moodustavad LINE, SINE, LTR ja DNA elemendid. SINE alamklassi kuuluvad Alu elemendid ning LINE alamklassi LINE1 elemendid. (International Human Genome Sequencing Consortium, 2001). LTR ja DNA elemente siin töös põhjalikumalt ei käsitleta.

	Koopiaarv (x1000)	Aluspaaride arv sekveneeritud genoomis (Mbp)	Fraktsiooni protsent sekveneeritud genoomis (%)
SINE	1558	359,6	13,14
Alu	1090	290,1	10,6
LINE	868	558,8	20,42
LINE1	516	462,1	16,89
LTR elemendid	443	227,0	8,3
DNA elemendid	294	77,6	2,84
Klassifitseerimata	3	3,8	0,14
Insertsiooniliste korduste koguarv		1226,8	44,83

SINEd on umbes 100-400 bp pikkused transposoonid ja ei kodeeri valke. Enamus SINEsid jagavad oma 3’ otsa LINE elemendiga (Okada *et al.*, 1997). SINEd, mis ei jaga 3’ otsa on Alu elemendid, mis on inimese genoomi ainsad transpositsiooniliselt aktiivsed SINEd (Lander *et al*., 2001).

Inimese Alu elemendi amplifikatsiooni kiirus pole olnud ühtlane (Shen *et al.*, 1991). Joonisel 7 on illustreerivalt näidatud Alu perekondade levimise mustrit primaatide genoomides seoses ligikaudse alamperekonna suurusega. Enamus Alu elementide kordusi on duplitseerunud rohkem kui 40 miljonit aastat tagasi. Primaatide evolutsiooni varajases faasis oli ligikaudu üks uus Alu insertsioon iga primaadi sünni kohta. Üks Alu insertsioon on iga 26 sünni kohta (Xing *et al.*, 2013).



Joonis 7. Alu elementide levimine primaatides. Alu alamperekondade (Yc1, Ya5a2, Yb9, Yb8, Sg1, Sx ja J) levimine on välja toodud primaatide evolutsiooni puu joonisel. Erinevate Alu alamperekondade levimine on värviliste kastidega kirjeldatud, et tähistada lahknevuse ajahetk. Umbkaudsed Alu alamperekondade koopiaarvud on samuti ära märgitud. Mat tähendab miljon aastat tagasi (Batzer ja Deininger, 2002).

Alu elemendid on inimgenoomi kõige sagedasemalt esinevad elemendid, ulatudes üle ühe miljoni koopiani ning on kaasatud alternatiivses splassingus, RNA korrektuurides, translatsiooni regulatsioonis (Häsler ja Strub, 2006). Inimese genoomi algne sekveneerimine näitas, et 55% selle nukleotiidijärjestusest koosneb korduvatest elemenditest (International Human Genome Sequencing Consortium, 2001). Alu elemendid moodustavad 11% inimese genoomist (Deininger, 2011) ning kuuluvad SINE perekonda (Quentin, 1992). Praegused Alu elemendid on umbes 280 bp pikad (Deininger *et al.*, 2003).

LINE elemendid moodustavad inimese genoomist umbes 20% ning on umbes 6000 bp pikkused (Martin ja Bushman, 2001). Inimese geoomist on leitud 3 LINE perekonda: LINE1, LINE2 ja LINE3, millest LINE1 on ainsana transpositsiooniliselt aktiivne (Okada *et al.*, 1997).

CNV-d on struktuuri modifikatsiooni tüüp, mis hõlmab spetsiifilistes piirkondades DNA-s muutusi koopiaarvus, mis saavad olla kas deletsioonid või duplikatsioonid. Sellised kromosomaalsed deletsioonid ja duplikatsioonid toimuvad suhteliselt suurtes DNA piirkondades (Thapar ja Cooper, 2013). CNV hulk inimese genoomis on vahemikus 4,8-9,7%

(Zarrei *et al*., 2015) ning suurus kõigub 50 bp ja 3 Mbp vahel. Kui element on väiksem kui 50 bp (MacDonal *et al*., 2014), siis seda loetakse insertsiooniks või deletsiooniks, kui suurem, siis sel juhul kuulub CNV-de hulka. CNV regioonid on genoomis ja kromosoomides ebaühtlaselt jaotunud (Makino *et al.,* 2013; Wong *et al*., 2013). CNV-de täpne arv pole teada, kuna ühegi teadaoleva meetodiga pole hetkel võimalik leida seda variatsiooni täielikult (Levy *et al.*, 2007; Pang *et al*., 2010; Pang *et al.*, 2014).

Pseudogeenid on genoomis üldlevinud ning neid on palju (Tutar, 2012). Need on geenikoopiad, millel on kodeerivate järjestuste puudused nagu raaminihked ja enneaegsed stoppkoodonid, kuid sarnanevad funktsionaalsete geenidega (Pink *et al*., 2011). Pseudogeenide hulka kuuluvad protsessitud pseudogeenid, mis on tekkinud mRNA pöördtranskriptsiooni käigus (Harrison *et al.*, 2005) ning omavad tähtsat rolli geeniregulatsioonis (Sasidharan ja Gerstein, 2008; Salmena *et al.*, 2011). Inimese genoomist on leitud üle 8000 protsessitud pseudogeeni (Zhang ja Gerstein, 2004). Nende keskmine pikkus on 740 aluspaari (Zhang, Harrison, Liu, & Gerstein, 2003) ning on tekkinud umbes 40-50 miljonit aastat tagasi (Ohshima et al., 2003).

1.3 Genoomi suuruse määramise metoodikad

Genoomi kaardistamisel kasutatakse kahte võimalust – geneetiline- ja füüsiline kaardistamine. Geneetilised kaardid põhinevad rekombinatsiooni sagedusel molekulaarmarkerite vahel. Need kaardid on populatsioonispetsiifilised. Geneetilisi kaarte on edukalt kasutatud suhteliselt haruldaste, ühe geeni pärilike häirete, nagu tsüstiline fibroos ja Duchenne'i lihasdüstroofia, eest vastutava geeni leidmiseks. Geneetilised kaardid on kasulikud ka teadlaste suunamisel paljude geenideni, mis arvatakse, et osalevad sagedamate haiguste nagu astma, südamehaiguste, diabeedi, vähi ja psühhiaatriliste seisundite kujunemisel (https://www.genome.gov/10000715/genetic-mapping-fact-sheet/). Füüsilised kaardid on DNA järjestuste joondamine, kus markerite vahelist kaugust on mõõdetud aluspaarides (Dixit *et al.*, 2014).

C-väärtus on haploidses tuumas (nt gameet) sisalduv DNA kogus pikogrammides või pool diploidse eukarüootse organismi somaatilise raku DNA kogusest. Mõningatel juhtudel (eriti diploidsete organismide seas) kasutatakse mõisteid C-väärtus ja genoomi suurus vaheldumisi.

Polüploidides võib C-väärtus esindada kahte või enamat sama tuuma sisalduvat genoomi (Greilhuber *et al.*, 2005).

Genoomi suuruse määramise metoodikaid on mitmeid. Klassikalised meetodid põhinevad suure molekulmassiga genoomse DNA hübridisatsiooni kineetikal. Hilisemad metoodikad kasutavad DNA-spetsiifilisi fluorestsentvärve voolutsütomeetrias, pildianalüüsis või imendumise tsütomeetrias peale Feulgeni värvimist. Üheks levinud meetodiks on veel reaalaja PCR (Wilhelm, 2003).

1.3.1 Hübridisatsiooni kineetika

Hübridisatsiooni kineetika puhul kasutatakse hüdroksüül-apatiit-kromatograafiat, et eraldada üheahelalist ja kaheahelalist DNA-d (dsDNA) (Wilhelm, 2003). Hübridisatsiooni kineetika põhimõtetel on võimalik mõõta kui palju korduvat DNA-d on genoomi DNA proovis (Waring ja Britten, 1966). Seda analüüsi kasutatakse selleks, et uurida genoomi struktuuri ja korraldust ning on kasutatud ka selleks, et lihtsustada genoomide järjestust, mis sisaldavad suures koguses korduvaid järjestusi (Peterson *et al.*, 2002). Meetod põhineb sellel, et uuritav DNA tehakse tükkideks, mis on mõnisada aluspaari pikad ning seotakse seejärel kuumutamisega üheks ahelaks (http://oxfordindex.oup.com/view/10.1093/oi/authority.20110803100407382).

1.3.2 Voolutsütomeetria

Voolutsütomeetrial põhinevate meetoditega analüüsitakse üksikute tuumade DNA sisaldust fluorestsentsmõõtmistega pärast värvimist propiidiumjodiidiga, etiidiumbromiidiga või neeldumise fotomeetriga ja kujutise analüüsiga pärast Feulgeni värvimist (Wilhelm, 2003). Meetod põhineb sellel, et tuhanded rakud läbivad sekundis ükshaaval laserkiirt, mille järel mõõdetakse hajunud valgust ning fluorestseeruvat emissioonivalgust (Picot *et al.*, 2012)**.**

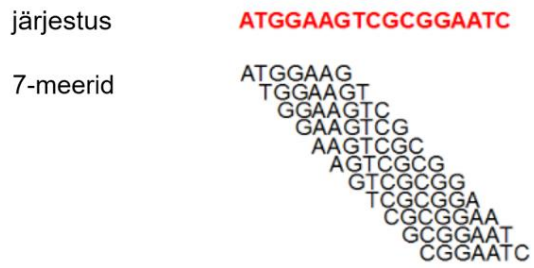
1.3.3 QPCR

QPCR on molekulaarbioloogia tehnika, mis võimaldab sihtmärgiks oleva DNA molekuli paljundamist ja samaaegset kvantifitseerimist. Võrreldes esialgse PCR meetodiga on qPCR arenenud selle poolest, et võimaldab DNA paljundamist jälgida reaalajas (Higuchi *et al.*, 1992). Peale igat tsüklit mõõdetakse DNA kogust fluorestsentsvärviga, mis annab kasvu korral fluorestseeruva signaali ning on otseses seoses PCR-produkti molekulide arvuga

(http://find.thermofisher.com/Global/FileLib/qPCR/2016-Real-Time-qPCR-Handbook-branding.pdf). Selles analüüsis kasutatud proovi kogus (mass) määratakse UV absorptsioon-spektromeetria abil. Proov peab sisaldama puhast DNA-d ilma suurema RNA saasteta, et mõõta kontsentratsiooni. Seejärel saab C-väärtust kergesti arvutada, jagades proovi DNA massi ühe koopia geenide jaoks määratud koopiaarvuga (Wilhelm, 2003).

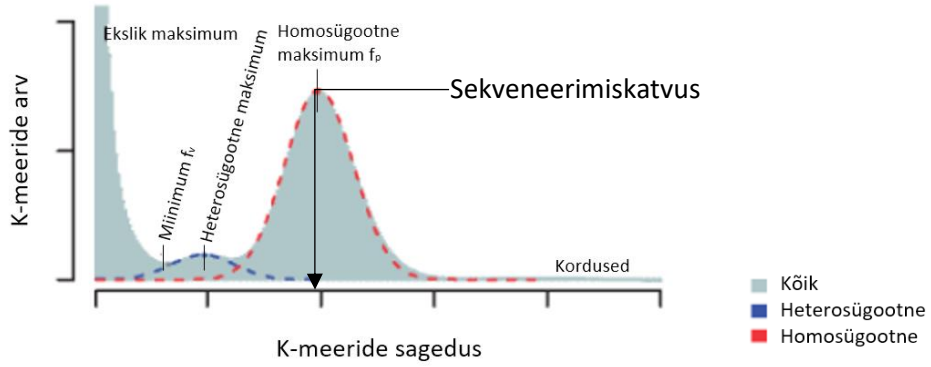
1.3.4 K-meeridel põhinevad meetodid

Viimastel aastatel on k-meeri põhised analüüsid ja võrdlusmeetodid saanud standardseks tööriistaks, mis võimaldab analüüsida suuri DNA järjestusi nagu kromosoomid, terved genoomid või isegi metagenoomid, mis on mikroobide genoomid keskkonnaproovis (Tringe *et al*., 2005). K-meer on k nukleotiidi pikk oligomeer (joonis 8). Näiteks 10-meer tähistab oligomeeri, mis on 10 nukleotiidi pikk (Wood ja Salzberg, 2014).



Joonis 8. K-meeri näidis. Näide 17 nukleotiidi pikkusest DNA järjestusest, millest võetakse 7-meerid (http://www.homolog.us/Tutorials/index.php?p=2.1&s=1).

K-meeride pikkus peab olema piisavalt suur, mis võimaldab määrata k-meeril unikaalset lokalisatsiooni genoomis. Liiga suured k-meerid põhjustavad arvutusressursside liigset kasutamist. FindGSE meetodi põhjal k-meerid pikkusega 21-33 nukleotiidi andsid kõige stabiilsemad (varieeruvus oli väike) genoomide suurused (Sun et al., 2018). Esimeses etapis arvutatakse k-meeride sagedus genoomi katvuse määramiseks sekveneerimise käigus. Selleks saab kasutada tarkvaratööriista, näiteks KMC (https://github.com/refresh-bio/KMC) või Glistquery (https://github.com/bioinfo-ut/GenomeTester4/blob/master/src/glistquery.c), mis aitab leida k-meeride sagedust järjestusprojektides. Kui k-meeride sagedused on arvutatud, siis kuvatakse histogramm (joonis 9) jaotuse visualiseerimiseks ja keskmise katvuse arvutamiseks. Esimene tipp on (ekslik maksimum) peamiselt haruldase ja juhuslike sekveneerimise vigade tõttu lugemites. Graafiku väärtusi saab kärpida kui eemaldada lugemid sekveneerimise vigadega (https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/).



Joonis 9. Diploidse genoomi k-meeri sageduse histogramm. Punktiirjooned tähistavad vastavaid kõveraid, mis näitavad k-meeride hulka heterosügootses (vasakul) või homosügootses (paremal) piirkonnas. Joonis kujutab diploidse genoomi tüüpilist k-meeri jaotust. Vasakpoolne tipp koosneb enamasti k-meeridest, mis koosnevad sekveneerimise vigadest. Need esinevad sageli, kuid madala sagedusega, kuna on vaid ühes või mitmes järjestuses. Teine (heterosügootne) ja kolmas (homosügootne) tipp kajastavad ühes või mitmes kromosoomikomplektis olevaid genoomseid k-meere, mida jagavad kõik järjestuse proovid vastavast lookusest. Jaotuse pikk saba kujutab genoomseid k-meere korduvatest elementidest, mis esinevad kõrgematel sagedustel, kuna jagavad mitut lookust. X-telg väljendab k-meeride sageduste arvu ning y-telg konkreetsete sagedustega k-meeride koguarvu. (Sun *et al.*, 2018). Katvust näitab maksimum, mis on noolega tähistatud (sekveneerimiskatvus ja k-meeride sagedus on võrdväärsed).

1.4 Illumina sünteesi teel sekveneerimine

Illumina sekveneerimine loob mitmeid miljoneid väga täpseid lugemeid, kuna järjestust analüüsitakse iga lisatud nukleotiidi tagant (https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequencing). Illumina sekveneerimise tehnoloogia põhjal luuakse samaaegselt miljoneid sekveneeritud DNA fragmente klaasist aluse pinnale (joonis 10). DNA fragmendid, mis on ühendatud adapteritega, tehakse üheahelalisteks. Seejärel sünteesitakse neile teine ahel komplementaarselt juurde. Nukleotiide on muudetud selliselt, et igal neist (A, C, G või T) oleks küljes eri fluorestseeruv värv. Seejärel neid pildistatakse ning analüüsitakse arvutis. Tänu värvidele on DNA fragmendi järjestust lihtsam sekveneerida (http://www.historyofnimr.org.uk/mill-hill-essays/essays-yearly-volumes/2010-2/bringing-it-all-back-home-next-generation-sequencing-technology-and-you/). Sekveneeritud DNA fragmendid pannakse referentsgenoomile ning võrreldakse sellega (https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.p

df). DNA hulk peab olema väga hästi paigas kvaliteetsete sekveneerimislugemite saamiseks. Selleks mõõdetakse ära DNA hulk, mida sekveneerimisel kasutatakse.

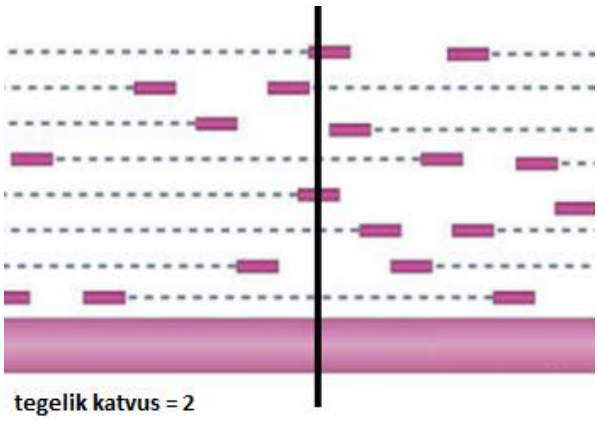
Joonis 10. Illumina HiSeq Flowcell. HiSeq Flowcell jaguneb kaheksaks rajaks, mis võimaldab teha 8 eri katset korraga. Iga rada jaguneb kaheks reaks, mis sisaldavad fikseeritud kohtades nanokannusid. Flowcell lugemi pikkus on kuni 100 bp ning sellega on võimalik tekitab mõnikümmend giga aluspaari päevas. Nende andmete põhjal on võimalik tuvastada üksikuid nukleotiide või insertsioonide ja deletsioonide polümorfisme referentsgenoomiga võrreldes (http://genepool.bio.ed.ac.uk/illumina/index.html; https://genome.duke.edu/cores-and-services/sequencing-and-genomic-technologies/illumina-sequencing).

1.4.1 Katvus

Teoreetilise katvuse arvutamiseks kasutatakse Lander/Watermani võrrandit C=LN/G (Eric S. Lander et al., 1988), kus:

- C - katvus;
- G - haploidse genoomi pikkus;
- L - lugemi pikkus;
- N - lugemite arv.

Kui võtta üks rada (joonis 10) ühe inimese sekveneeritud lugemitest, saame C=(100 bp)*(189x10^6)/(3x10^9 bp) = 6,3, mis ütleb, et iga alus genoomis on sekveneeritud keskmisel 6 kuni 7 korda. See number näitab, et nii palju kordi on eeldatud, et iga nukleotiid on sekveneeritud teatud pikkusega ja arvuga lugemite puhul (Lander *et al.*, 2001). Tegelik katvus tähendab täpset arvu kordi, kui sekveneeritud DNA fragment referentsgenoomiga kattub (joonis 11) (Sims *et al.*, 2014).



Joonis 11. Järjestuse tegelik katvus. Punktiirjooned tähistavad sekveneerimata ala ning punktiirjoonte otsas olevad roosad ristkülikud on sekveneeritud DNA fragmendid. Tegelik katvus tähendab sekveneeritud lugemite arvu, mis katab referentsgenoomi (kõige suurem roosa ristkülik) teatud ala. Must vertikaalne joon on tõmmatud läbi DNA fragmentide, et näidata, kus kohas referentsgenoomis sel hetkel on järjestuse katvus 2 (Meyerson *et al.*, 2010).

Illuminal on Internetis katvuse kalkulaator, mis arvutab reagentide kogust ja kogu sekveneerimistsüklite arvu, mida on tarvis vajaliku katvuse saamiseks (https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf).

2 EKSPERIMENTAALOSA

2.1 Töö eesmärgid

Töö hüpoteesiks on - mida väiksem on katvus seda suurem on genoomi suurus: DNA kogus/katvus = genoomi suurus.

Sellest lähtuvalt on töö eesmärkideks:

a. määrata sekveneerimiskatvused ehk mitu korda keskmiselt on kõik genoomi positsioonid sekveneeritud;
b. määrata erinevate inimeste genoomide suurused teise põlvkonna sekveneerimisandmetest (Illumina);
c. leida inimestevahelise genoomi suuruste erinevust põhjustavad spetsiifilisemad genoomi struktuurielemendid.

2.1.1 Andmestik

Andmestikuks on 100 Tartu Ülikooli Eesti Geenivaramu geenidoonorit (50 meest ja 50 naist), kelle verest eraldatud DNA on sekveneeritud täies mahus MIT Broad Instituudi genotüpiseerimiskeskuses. Siinses töös on kasutatud juba joondatud .bam andmefaile, kus on kõik sekveneerimislugemid säilinud. Lisaks on sekveneeritud indiviidide genotüpiseerimisandmetega kaasas sekveneerimisega seotud tehnilised näitajad, nagu sekveneerimiskatvus, indeks praimerite osakaal, PCR-i duplikaatide osaprotsent, kvaliteetsete lugemite arv ja kogupikkus.

2.1.2 Töövoog

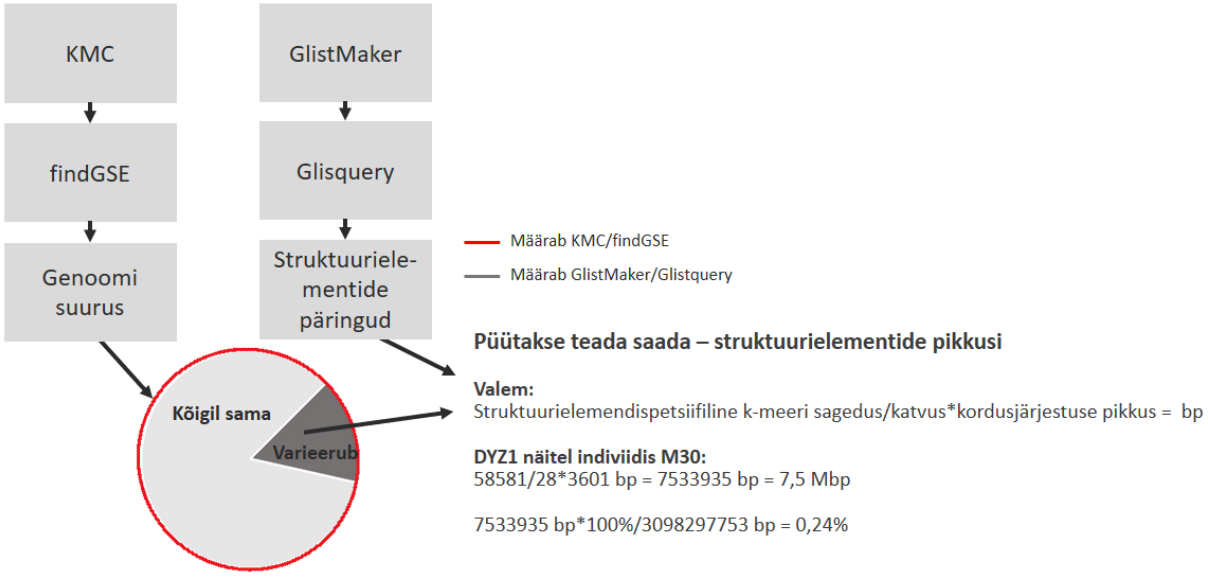
2.1.2.1 Andmed läbi töövoo

Andmestikuks töötlesin 100 TÜ EGV proove. KMC (joonis 12) programmi kasutasin selleks, et saada k-meeri sagedused. See programm on kiire, kuid programmiga ei saa k-meeri sageduse päringuid teha. Seejärel sorteerisin k-meerid sageduste sageduste (k-meeride katvus ja vastava katvusega k-meeride hulk) järgi ning andsin vajalikud andmed sisendiks findGSE programmile.

FindGSE on programm genoomi suuruste määramiseks, mis põhineb k-meeride sageduste sagedustel. FindGSE programm saab genoomi suurusi arvutama hakata peale seda kui sisendiks on antud kahe tulbaga fail, mis sisaldab k-meeride sageduste sagedust. Veel tuleb lisada jaotuskõvera maksimum ning k-meeri pikkus, milleks selles töös oli 25-meer. Nende andmete põhjal arvutab findGSE genoomi suuruse.

KMC ja findGSE abil saadi terve genoomi suurus. Järgmisena püütakse GlistMakeri (https://github.com/bioinfo-ut/GenomeTester4/blob/master/src/glistmaker.c) ja Glistquery abil teada saada struktuurielementide pikkused. Struktuurielemendid on suured varieeruvad piirkonnad, mis varieeruvad nii pikkuses kui ka koopiaarvus organismi genoomis. GlistMaker on oma olemuselt sarnane KMC programmiga. Glistquery on programm, mis on vajalik GlistMakeriga tehtud k-meeri listidest päringute tegemiseks.

Struktuurielementide pikkusi on võimalik leida kasutades valemit, kus struktuurielemendispetsiifiline k-meeri sagedus (tabel 3) jagatakse jaotuspõhise katvusega (lisa 1) ning korrutatakse kordusjärjestuse pikkusega (tabel 3).



Joonis 12. Töökäik. Genoomi suuruse leidmine KMC ja findGSE programmiga ning varieeruvate struktuurielementide pikkuse leidmine GlistMakeri ja Glistquery abil. Joonisel on näitena toodud, kuidas valemit kasutada, kui on soov välja arvutada DYZ1 elemendi pikkust ning seejärel ka uurida, kui suure osa see moodustab indiviidi genoomist, mis on valemi põhjal: struktuurielemendi pikkus*100%/genoomi suurus. Kokkuvõtlikumad andmed nendest arvutustest on tabelis 5.

2.1.2.2 K-meer listide koostamine

Listidega manipuleerimisel on kasutatud kahte k-meer listide tegemise programmi:

a. KMC – on väga kiire, kuid päringute tegemine on kõvaketta ruumi- ja ajamahukas. KMC programmiga sai tehtud listid vaid ajutiselt kuni hetkeni, kui k-meer sageduste sageduste jaotus sai välja arvutatud. KMC programmi oli vaja selleks, et GlistMaker programmi poolt koostatud k-meeri listidest, millest puudusid sagedusega 1 k-meerid.
b. GlistMaker – programm kuulub paketti GenomeTester4, mis koostab sorteeritud listid, milledest tehtud päringud on ülikiired. Selles töös kasutatakse GlistMakeriga varasemalt tehtud k-meeride liste.

2.1.2.3 K-meeri sageduste loendamine

KMC programmi juures kasutati järgnevat tööde käiku:

a. KMC 25-meeri list kirjutati arvuti kõvakettale tervenisti välja;
b. Perli programm liitis kokku kõik sama sagedustega k-meerid;
c. sorteeritud sageduste sageduste järgi määrati sekveneerimise katvus 1-koopia piirkonnale genoomis.

2.1.3 Sekveneerimiskatvuse määramine

Sekveneerimise katvuse määramisel on siin kasutatud kahte võimalust:

a. MIT Broad Instituudi sekveneerimiskeskusest väljastatud number, mis on kogu kvaliteetselt sekveneeritud nukleotiidide arv jagatud referentsgenoomi suurusega;
b. punktis 1.3.4 kirjeldatu (joonis 9);

2.1.4 Genoomi suuruse määramine

Vaadeldud on kolme genoomi suuruse määramise võimalust, kus läbivaks teemaks on kas määrata või hinnata sekveneerimise katvust:

a. genoomi suurus on referentsgenoomi suurus, mis on summaarselt referentsgenoomis olevate nukleotiidide arv. Seda metoodikat on kasutanud MIT Broad Instituudi sekveneerimiskeskus sekveneerimiskatvuse hindamiseks puhtalt sekveneerimise õnnestumise hindamiseks (eesmärk on olnud vähemalt 20-kordne katvus);

b. findGSE on vahend (heterosügootsete diploidsete või homosügootsete) genoomide suuruse hindamiseks, kohandades k-meeri sagedusi normaaljaotuse abil, mis on kirjutatud R-vormingus. FindGSE kasutamiseks peab sisestama k väärtuse ja vastava k-meeri .histo laiendiga faili, mis on loodud lühikeste lugemitega ja sisaldab kahte tabelisse kuuluvat veergu. Esimeses veerus on toodud sagedused, millised k-meerid esinevad järjestuses, samas kui teine veerg loeb selliste eristatavate k-meeride arvu. Kui oleme selle kaheveerulise faili saanud, siis saame genoomi suuruse hindamise jaoks minna R keskkonda.

   Kirjutada vastavad käsud:

   - library("findGSE")
   - findGSE(histo="test_21mer.histo", sizek=25, outdir="hom_test_21mer", exp_hom=21), kus:
     - histo – sorteeritud faili asukoht;
     - sizek – kui pikk soovitud k-meer olema peab;
     - outdir – kataloog, kuhu läheb fail;
     - exp_hom – jaotuskõvera maksimum.

   Kui findGSE on faili läbi jooksutanud, avati valmis tehtud .txt faili, kus on kirjas genoomi suurus, mis lisati Exceli tabelisse, et kõik andmed kokku panna; (https://github.com/schneebergerlab/findGSE/blob/master/R/findGSE_v1.94.R)

c. kõigi sagedustega k-meeride sageduste ja hulga omavaheline korrutis on jagatud sekveneerimiskatvusega, millest on maha lahutatud mitokondri genoomi suurus, indeks praimerite ja sekveneerimise duplikaatide osa.

2.1.5 Järjestusspetsiifiliste k-meer järjestused

a. Siin töös on kasutatud k-meer järjestusi, et hinnata nende osatähtust sekveneerimisandmetes. Andmed on kirjeldatud tabelis 3.

Tabel 3. Järjestusspetsiifiliste k-meeride kirjeldus. Tabelis on kirjeldatud järjestusspetsiifiliste elementide k-meerid, kordujärjestuse pikkused, ID GenBankis ja asukoht järjestuses GenBanki andmebaasis. -mm 1 tähendab seda, et k-meeril on lubatud üks nukleotiid, mis ei ole komplementaarne selle järjestusega.

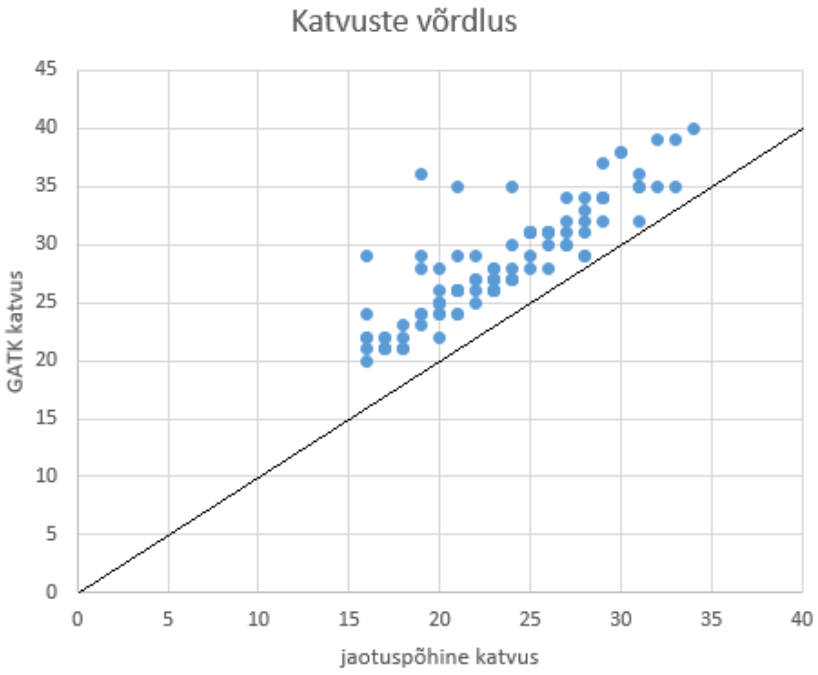
Tabeli kirjeldus	K-meer	Kordusjärjestuse pikkus	Genbank ID	Asukoht Genbank järjestuses
polüA	AAAAAAAAAAAAAAAAAAAAAAAAA	1	-	-
polüC	CCCCCCCCCCCCCCCCCCCCCCCCC	1	-	-
DYZ1	CCAGTCCATTTAATTCAAGGGCATT	3601	KF941192	2190-2215
telomeerid	GGGTTAGGGTTAGGGTTAGGGTTAG	6	-	-
45S rRNA	GCCCTTAGATGTCCGGGGCTGCACG	42999	U13369	5156-5180
5S rRNA	GGGCGGAGGACCGGAGGGCGTCCCA	2230	AL713899	6-30; 2247-2271; 4488-4512; 6729-6753; 8970-8994; 11211-11235; 13426-13450; 15646-15670; 17888-17912; 20127-20151; 22369-22393; 24610-24634; 26850-26874; 29075-29099; 31316-31340; 33547-33571; 35788-35812; 38019-38043
satelliit-DNA	TCTTTTTGTAGAATCTGCAAGTGGA	5	-	-
TTCCA	TTCCATTCCATTCCATTCCATTCCA -mm 1	5	-	-
tsentromeerid	TGTGGAATTTGCAAGTGGAGATTTC -mm 1	171	-	-
Alu	GGCCTCTCAAAGTGCTGGGATTACA -mm 1	300	-	24-48
LINE	AAGTTCATATGGAACCAAAAAAGAG -mm 1	6017	U93571	4519-4543
CCTT	AGGAAGGAAGGAAGGAAGGAAGGAA -mm 1	4	-	-

b. Kõigile k-meeridele, mis on kasutatud, vastab konkreetne korduva elemendi pikkus, mis on läbi korrutatud ühe indiviidi vastava k-meeri sageduse ja sekveneerimiskatvuse jagatisega.

2.2 Tulemused

Tulemused on jagatud kolmeks:

a. katvused;
b. genoomi suurused meestel ja naistel;
c. genoomi suurust mõjutavad järjestused meestel ja naistel.



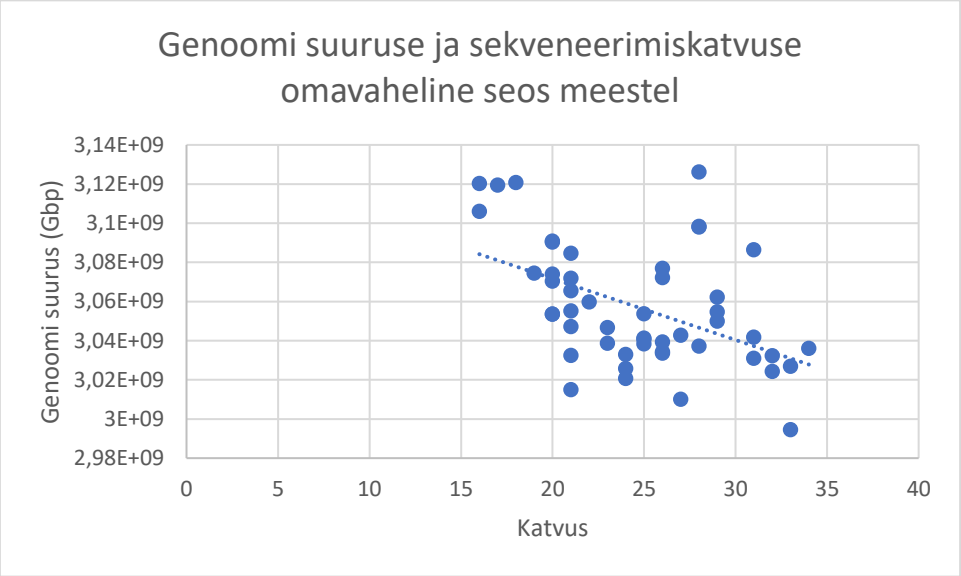
Joonis 13. Katvuse võrdlus findGSE ja GATK vahel. X-teljel on kujutatud jaotuspõhine katvus ning y-teljel GATK katvus. Katvuste erinevuse põhjustavad nende programmide meetodid. FindGSE võtab katvuse jaotuskõvera põhiselt, GATK programmil on referentsgenoomi pikkused fikseeritud. Must diagonaalne joon tähistab katvusi, mis oleks vastavuses üksteisega, selle järgi on aru saada, et GATK katvus on suurem kui jaotuspõhine katvus. Katvusi on võrreldud 100 indiviidi puhul.

Kuna Y-kromosoom on oluliselt lühem kui X-kromosoom, siis naiste ja meeste genoomi suuruse mõõtmisi vaadeldi eraldi.

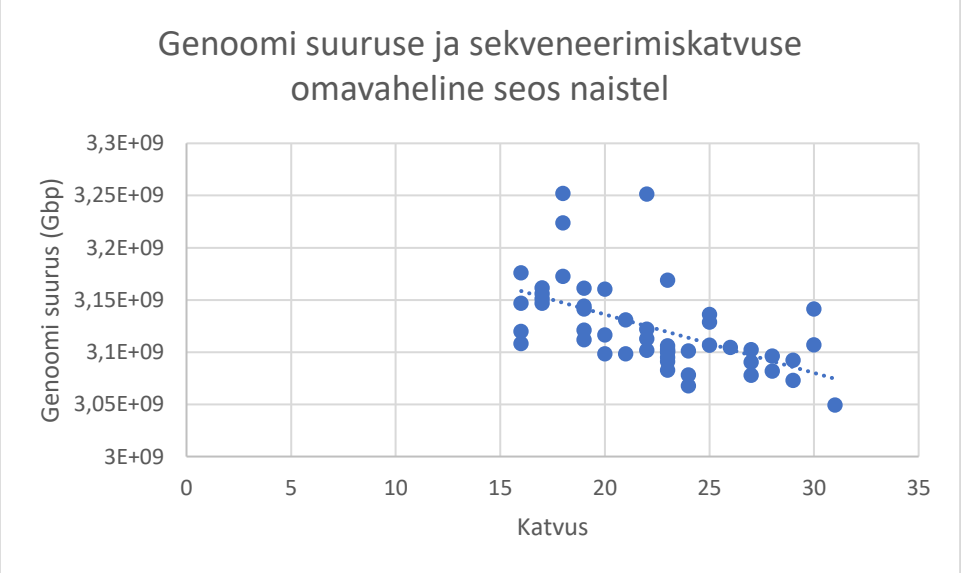
i. Sekveneerimiskatvuste võrdluses on findGSE ja GATK andmed erinevad, kuna findGSE põhineb k-meeri sagedustel, kuid GATK programmil on referentsgenoomi pikkused fikseeritud. Tulemused on välja toodud joonisel 13.
ii. Genoomi suurused on vahemikus 2,99-3,13 Gbp meestel ja 3,05-3,25 Gbp naistel, kasutades findGSE programmi. Genoomi suurused on vahemikus 3,26-4,23 Gbp meestel ja 3,22-4,22 Gbp naistel, kasutades lihtsat pindala arvutust. Genoomi suurused on vahemikus 2,65-4,00 Gbp meestel ja 2,83-4,03 Gbp naistel, kasutades pindala arvutust ja mahaarvutusi (mitokondri genoom, praimeri ja PCR duplikaatide osahulka) (tabel 4). Vastavad tulemused on kokkuvõtvalt lisa 1 tabelites 6 ja 7.
iii. Genoomi suuruse ja sekveneerimiskatvuste omavahelise sõltuvuste graafikud on meestel ja naistel välja toodud joonistel 14 ja 15 (Sun et al., 2018).

Tabel 4. FindGSE, pindala ja mahaarvutustega pindala keskmised genoomi suurused.

Sugu	FindGSE (Gbp)	Pindala (Gbp)	Mahaarvutustega pindala (Gbp)
mehed	3,06	3,57	3,27
naised	3,12	3,59	3,30



Joonis 14. Genoomi suuruse ja sekveneerimiskatvuse omavaheline seos meestel. Joonis on koostatud 50 mehe tulemustest. Katvus on võetud jaotuskõvera maksimumi põhjal. Sinine lineaarne punktiirjoon näitab - mida suurem on genoomi suurus, seda väiksem on katvus. X-telg väljendab katvust ning y-telg genoomi suurust.



Joonis 15. Genoomi suuruse ja sekveneerimiskatvuse omavaheline seos naistel. Joonis on koostatud 50 naise tulemustest. Mida suurem on genoomi suurus, seda väiksem on katvus (näitab sinine lineaarne punktiirjoon).

Toodud tulemustest on näha seost väiksemal genoomil ja suuremal katvusel ning vastupidi.

iv. Struktuurielementide pikkused on samuti soospetsiifikast sõltuvad. Mõõdetud k-meeri sagedused ja vastavad struktuurielementide pikkused on lisades 2 (tabel 8 ja 9) ja 3 (tabel 10 ja 11) ning kokkuvõtvad tulemused pikkuste vahemikust ja osakaalust genoomis tabelis 5. Siin töös kasutatud DYZ1 regioon on ainult mehel Y-kromosoomis. Samuti on Alu ja LINE elementide hulk muutuv.

Tabel 5. Meeste ja naiste struktuurielementide pikkused ja osakaal genoomis. Tabelis on tehtud kokkuvõtted tabelitest 8 ja 9. Keskmised genoomi suurused on saadud findGSE järgi – meestel 3,06 Gbp ja naistel 3,12 Gbp.

Tabeli kirjeldus	Pikkuste vahemik (Mbp)		Osakaal keskmisest genoomist (%)	
	Mehed	Naised	Mehed	Naised
DYZ1	4,79-9,53	0-0,07	0,228	0,0007
Telomeerid	0,05-1,38	0,06-1,62	0,003	0,003
45S rRNA	6,36-14,76	5,81-14,46	0,316	0,297
5S rRNA	0,07-0,30	0,08-0,26	0,005	0,005
Satelliit-DNA	0,08-0,12	0,08-011	0,003	0,003
TTCCA	4,03-8,40	4,84-8,94	0,205	0,205
Tsentromeerid	7,19-10,30	7,82-11,77	0,297	0,295
Alu	48,64-60,11	50,05-63,52	1,726	1,712
LINE	88,80-100,00	93,66-101,14	3,067	3,140
CCTT	0,19-0,22	0,19-0,23	0,007	0,007

Arutelu

DNA sekveneerimisel üritatakse sekveneerida sama kogus DNA-d. Selleks mõõdetakse DNA kontsentratsioon ära ja pannakse eeldatav kogus DNA-d reaktsiooni. 2284-st TÜ EGV proovist võeti 100 indiviidi (50 meest ja 50 naist). 100 indiviidi andmete töötlemine võttis aega umbes 4 nädalat. Kõigi andmete läbitöötamine oleks võtnud liiga kaua aega. Paari indiviidi puhul ei saanud algselt täpseid andmeid genoomi suuruse kohta sellepärast, et KMC failid ei olnud saanud lõpuni andmeid faili kirjutada, kuna kettaruum sai otsa.

Käesolevas töös püstitatud hüpotees, mida väiksem on katvus, seda suurem on genoomi suurus, leidis kinnitust 50 mehe ja 50 naise kogugenoomi Illumina sekveneerimisandmete põhjal.

Siinses töös kasutatud genoomi suuruse hindamise meetod eeldab sekveneerimiskatvuse mõõtmist. Sekveneerimiskatvuse tõlgendamine on lihtsamini mõistetav sellisel juhul, kui genoomilõik on genoomis esindatud vaid ühes kohas. Kui on genoomilõik kas deleteerunud või duplitseerunud või vaadatakse hoopis mitokondriaalset DNA-d, on sekveneerimiskatvuse määramine keerulisem. Lisaks on eukarüootses rakus kaks õdekromosoomi ja katvust väljendatakse haploides genoomi kohta. Keerulisest olukorrast saadakse üle niimoodi, et määratakse visuaalselt k-meer sageduste jaotust nende k-meeride osas, mis on esindatud vaid ühe korra mõlemal õdekromosoomil. Sekveneerimiskatvuste vahemik jäi k-meer sageduste jaotuse põhjal 16 ja 34 vahele. Illumina platvormi juures kasutatav sekveveerimiskatvus, kõikide sekveneerimisjugemite kogupikkus läbijagatuna referentsgenoomi pikkusega, jäi vahemikku 22 kuni 40. Katvused varieerusid kahe programmi puhul märgatavalt, mis on tingitud meetodite erinevusest. FindGSE programm on k-meeri sagedustel põhinev, GATK programm on regerentsgenoomi pikkusel põhinev.

Kuna sekveneerimiskatvuste väärtus ise koheselt genoomi suurust ei määra, sai genoomi suurus mõõdetud kolmel erineval moel:

- findGSE programmiga;
- k-meeri katvuse ja vastava katvusega k-meeride hulga korrutisega;
- sama, mis eelnevas punktis, kuid teatud mahaarvamistega.

Kaks viimast meetodit on ise välja mõeldud, findGSE on juba tsiteerimistleidnud meetod. Töö praktilises osas ongi selle metoodikaga rohkem arvutusi tehtud. Lisaks on analoogilisi meetodeid veel, kuid neil puudub viitamisvõimalus. Aastal 2011. on k-meeri põhine genoomisuuruse määrmise meetodile võetud lausa patent US2014/188397 A1.

Eri aegadel on kasutatud erinevaid meetodeid genoomi suuruse hindamiseks. Kuna meetodeid on mitmeid, varieerub ka inimese genoomi suurus selle tõttu 2,9 Gbp ja 3,7 Gbp vahel. Töös kasutatud meetodite puhul sai täheldada samuti genoomi suuruste varieeruvust. FindGSE keskmine genoomi suurus meestel on 3,06 Gbp, naistel 3,12 Gbp, pindala järgi vastavalt 3,57 Gbp ja 3,59 Gbp, mahaarvutustega pindala järgi 3,27 Gbp ja 3,30 Gbp. Genoomi suurus on naistel suurem, kuna sisaldab X-kromosoomi, mis on tunduvalt suurem kui meestel Y-kromosoom. FindGSE, pindala ja mahaarvutustega pindalale vastavalt varieerus genoomi suurus keskmiselt 340 Mbp, mis on umbes 10% piires kogu genoomist (arvutatud kolme meetodi keskmise genoomi suuruse järgi). Kuna arvutused põhinesid findGSE katvust arvestades, siis seda võib pidada päris suureks erinevuseks.

Töös püüti leida genoomsed regioonid, mis võiks mõjutada kõige rohkem genoomi suuruse erinevusi. Selleks valiti tuntud, oma koopiaarvult varieeruvad genoomilõigud nagu 45S, 5S, DYZ1 mehel, Alu ja Line elemendid, lihtsad kordused ja heterokromatiini kordus TTCCA. Iga elemendi k-meeri sageduste põhjal hinnati osahulka genoomis aluspaarides. Selles osas jäid tulemused tagasihoidlikuks, kuna tulemusi saab võrrelda vaid visuaalselt taustavärvi põhjal. FindGSE artiklis seostatakse genoomi suuruse erinevust eelkõige LINE elementidega, mida siinsete mõõtmise juures on täheldatav vaid pindala ja mahaarvutustega pindala mõõtmiste variandi juures, kuid findGSE puhul mitte. Tulemused tabelis 2 ei ühtinud tabelis 5 saadud tulemustega. Sellest võib järeldada, et välja valitud k-meerid polnud piisavalt spetsiifilised, et leida otsitud järjestusi ning seda meetodid tuleks veel arendada, et oleks rohkem k-meere, millega otsida koopiaarvult varieeruvaid genoomilõike. DYZ1 elemente on meestel umbes 300 korda rohkem kui naistel, kuna DYZ1 on meessoo-spetsiifiline. Siiski leidus seda järjestust ka naissoost indiviididel, mille põhjuseks võivad olla seveneerimisvead või siis mõnel naisterahval on selline järjestus olemas.

Telomeeride pikkus mõjutab samuti genoomi suurust, sest mida vanem on inimene, seda lühem on telomeer, mille tõttu ka genoomi suurus on selle võrra väiksem. Nii meestel kui ka

naistel oli indiviiditi võrreldes kõigi järjestuste varieeruvus, mida on käsitletud selles töös, umbes 3%. Genoomi struktuurielementide osatähtsuses genoomi suuruse varieeruvuse kohta midagi olulist hetke järeldada ei saa.

Minu arvates saab findGSE meetodit pidada usaldusväärseks sel juhul, kui on võetud piisavalt optimaalse suurusega k-meerid, mille põhjal saab stabiilsed andmed genoomi suuruse kohta. Kuna kasutasin 25-meere, siis saadud genoomi suurusi võib pidada päris usaldusväärseks tulemuseks.

Geenivaramu 2284 täissekveneeritud genoomi suuruse hindamine vajab kiiremat katvuse määramise metoodikat. Siis on võimalik genoomi suurust kasutada fenotüübilise tunnusena ülegenoomse assotsiatsiooni analüüsi tegemisel.

Kokkuvõte

Käesoleva töö hüpoteesiks oli seatud väide, et mida väiksem on katvus, seda suurem on genoomi suurus. Töö eesmärkideks oli määrata sekveneerimiskatvused, erinevate inimeste genoomide suurused teise põlvkonna sekveneerimisandmetest ja leida inimestevahelise genoomi suuruste erinevusi põhjustavad spetsiifilisemad genoomi struktuurielemendid.

Inimese genoomi suurus varieerub mitmete elementide tõttu, mis võivad olla nii soo ja vanuse-spetsiifilised või sõltuda üldse mõnest muust tegurist. Peale selle varieeruvad genoomi suurused ka mõõtmiste meetoditest sõltuvalt. FindGSE on kiire ja mugav abivahend inimese genoomi suuruse hindamiseks k-meere kasutades. Selle jaoks on vaja eelnevalt sorteeritud k-meeride sageduste sageduste faili ning etteantud jaotuskõvera maksimumi.

Genoomi suurusi sekveneerimiskatvustega võrreldes tuli välja, et mida väiksem on katvus, seda suurem on genoom, mille tõttu saab väita, et hüpotees pidas paika. Katvus varieerus FindGSE meetodiga 16-34 vahel. FindGSE meetodiga saadi keskmiseks genoomi suuruseks meestel 3,06 Gbp ja naistel 3,12 Gbp. Suurem genoomi suurus naistel tuleneb sellest, et X-kromosoom on suurem kui meestel Y-kromosoom. Genoomi struktuurielementide puhul oli ainuke suurem erinevus tingitud DYZ1 järjestusest, mida oli meestel 300 korda rohkem kui naistel, kuna on meessoo-spetsiifiline. Käesolevas töös valitud k-meeride meetodiga ei saa struktuurielementide osatähtsuses genoomi suuruse varieeruvuse kohta midagi olulist järeldada.

Human genome size evaluation with k-mer method

Sylvia Krupp

Summary

Genome size is the amount of DNA in one cell. Genome size varies by species. Genome size of the human varies from 2,9 Gbp to 3,7 Gbp. The different estimations come from which method was used. Copy number variations are the reason why the genome size varies in human. Genome size depends on the sex of the individual therefore X-chromosome is larger than Y-chromosome. This means that women have larger genome than men.

In this study the author estimates 100 (50 men and 50 women) individuals' genome sizes with findGSE and varying areas. FindGSE method is based on k-mers. To get the correct results with findGSE, it is needed to give input a k-mer length, coverage, and histo file that consists k-mer fequencies which every individual has and k-mer counts. After the results were gathered from 50 women and 50 men, the estimation of the genome size for men was 3,06 Gbp and women 3,12 Gbp. The author also estimated varying ares like DYZ1, telomeres, 45S RNA, 5S RNA, satellite-DNA, TTCCA, centromeres, Alu, LINE and CCTT sequences. They were measured with k-mers and the variation of each was approximately 3% throughout men and women. DYZ1 element is male-spesific – the element was represented 300 times more in men individuals than in women.

The hypothesis that claimed the smaller the coverage, the greater the genome was proven to be correct.

KASUTATUD KIRJANDUS

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell, Fourth Edition*. *Molecular Biology*. https://doi.org/citeulike-article-id:691434

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

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Anderson, S., Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG., B. A. T., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., … Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*. https://doi.org/10.1038/290457a0

Batzer, M. A., & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg798

Blattner, F. R. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, *277*(5331), 1453–1462. https://doi.org/10.1126/science.277.5331.1453

Boyle, J. (2008). Molecular biology of the cell, 5th edition by B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Biochemistry and Molecular Biology Education*, *36*(4), 317–318. https://doi.org/10.1002/bmb.20192

Brosius, J. (2009). The fragmented gene. In *Annals of the New York Academy of Sciences* (Vol. 1178, pp. 186–193). https://doi.org/10.1111/j.1749-6632.2009.05004.x

Brown, T. A. (2002). *Genomes. 2nd*. *UK: Wiley-Liss Manchester*. https://doi.org/NBK21128 [bookaccession]

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. *Biology Direct*. https://doi.org/10.1186/1745-6150-8-3

Deininger, P. (2011). Alu elements: know the SINEs. *Genome Biol*, *12*(12), 236. https://doi.org/gb-2011-12-12-236 [pii]\r10.1186/gb-2011-12-12-236

Deininger, P. L., Moran, J. V., Batzer, M. A., & Kazazian, H. H. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2003.10.013

Dixit, R., Rai, D., Agarwal, R., & Pundhir, A. (2014). PHYSICAL MAPPING OF GENOME AND GENES. *J. Biol. Engg. Res. & Rev*, *1*(1), 6–11.

Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S., & Lemos, B. (2015). Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *Proceedings of the National Academy of Sciences*, *112*(8), 2485–2490. https://doi.org/10.1073/pnas.1416878112

Gosden, J. R., Lawrie, S. S., & Gosden, C. M. (1981). Satellite DNA sequences in the human acrocentric chromosomes: information from translocations and heteromorphisms. *American Journal of Human Genetics*, *33*(2), 243–251.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg1674

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., … Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(SUPPL. 1). https://doi.org/10.1093/nar/gkl828

Greilhuber, J., Doležel, J., Lysák, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms “genome size” and “C-value” to describe nuclear DNA contents. In *Annals of Botany* (Vol. 95, pp. 255–260). https://doi.org/10.1093/aob/mci019

Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005). Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Research*, *33*(8), 2374–2383. https://doi.org/10.1093/nar/gki531

Häsler, J., & Strub, K. (2006). Alu elements as regulators of gene expression. *Nucleic Acids Research*, *34*(19), 5491–5497. https://doi.org/10.1093/nar/gkl706

Higuchi, R., Dollinger, G., Walsh, P. S., & Griffith, R. (1992). Simultaneous amplification and detection of specific DNA sequences. *Biotechnology*, *10*(4), 413–417. https://doi.org/10.1038/nbt0492-413

Hochstrasser, T., Marksteiner, J., & Humpel, C. (2012). Telomere length is age-dependent and reduced in monocytes of Alzheimer patients. *Experimental Gerontology*, *47*(2), 160–163. https://doi.org/10.1016/j.exger.2011.11.012

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

Kass, D. H., & Batzer, M. A. (2001). Genome Organization: Human. In *Encyclopedia of Life Sciences*. https://doi.org/10.1038/npg.els.0001889

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lander, E. S., Waterman, M. S., Gu, H., Gnirke, A., Meissner, A., Lowe, C., … Feinberg, A. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, *2*(3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Leitch, I. J. (2007). Genome sizes through the ages. *Heredity*. https://doi.org/10.1038/sj.hdy.6800981

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, *5*(10), 2113–2144. https://doi.org/10.1371/journal.pbio.0050254

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt958

Makino, T., McLysaght, A., & Kawata, M. (2013). Genome-wide deserts for copy number variation in vertebrates. *Nature Communications*, *4*. https://doi.org/10.1038/ncomms3283

Martin, S. L., & Bushman, F. D. (2001). Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. *Molecular and Cellular Biology*, *21*(2), 467–475. https://doi.org/10.1128/MCB.21.2.467-475.2001

Mattick, J. S. (2004). The hidden genetic program of complex organisms. *Scientific American*. https://doi.org/10.1038/scientificamerican1004-60

Meyerson, M., Gabriel, S., & Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2841

Nussbaum, Robert L; McInnes, Roderick R; Huntington, F. W. (2016). *Thompson & Thompson Genetics in Medicine*. *Elsevier*. https://doi.org/10.1001/jama.1992.03480150121052

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., & Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biology*, *4*(11). https://doi.org/10.1186/gb-2003-4-11-r74

Okada, N., Hamada, M., Ogiwara, I., & Ohshima, K. (1997). SINEs and LINEs share common 3' sequences: A review. In *Gene* (Vol. 205, pp. 229–243). https://doi.org/10.1016/S0378-1119(97)00409-5

Pang, A. W. C., MacDonald, J. R., Yuen, R. K. C., Hayes, V. M., & Scherer, S. W. (2014). Performance of High-Throughput Sequencing for the Discovery of Genetic Variation Across the Complete Size Spectrum. *G3: Genes|Genomes|Genetics*, *4*(1), 63–65. https://doi.org/10.1534/g3.113.008797

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., … Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, *11*(5). https://doi.org/10.1186/gb-2010-11-5-r52

Pelham, H. R., & Brown, D. D. (1980). A specific transcription factor that can bind either the 5S RNA gene or 5S RNA. *Proceedings of the National Academy of Sciences*, *77*(7), 4170–4174. https://doi.org/10.1073/pnas.77.7.4170

Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, *164*(1), 10–15. https://doi.org/10.1111/j.1095-8339.2010.01072.x

Peterson, D. G., Wessler, S. R., & Paterson, A. H. (2002). Efficient capture of unique sequences from eukaryotic genomes. *Trends in Genetics*. https://doi.org/10.1016/S0168-9525(02)02764-6

Picot, J., Guerin, C. L., Le Van Kim, C., & Boulanger, C. M. (2012). Flow cytometry: Retrospective, fundamentals and recent instrumentation. *Cytotechnology*. https://doi.org/10.1007/s10616-011-9415-0

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., & Francisco Carter, D. R. (2011). Pseudogenes: Pseudo-functional or key regulators in health and disease? *RNA*, *17*(5), 792–798. https://doi.org/10.1261/rna.2658311

Quentin, Y. (1992). Origin of the alu family: A family of alu-like monomers gave birth to the left and the right arms of the alu elements. *Nucleic Acids Research*, *20*(13), 3397–3401. https://doi.org/10.1093/nar/20.13.3397

Rabl, J., Leibundgut, M., Ataide, S. F., Haag, A., & Ban, N. (2011). Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. *Science*, *331*(6018), 730–736. https://doi.org/10.1126/science.1198308

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews : MMBR*, *72*(4), 686–727. https://doi.org/10.1128/MMBR.00011-08

Richard Shen, M., Batzer, M. A., & Deininger, P. L. (1991). Evolution of the master Alu gene(s). *Journal of Molecular Evolution*, *33*(4), 311–320. https://doi.org/10.1007/BF02102862

Ridley, M. (2013). *Genome : the autobiography of a species in 23 chapters*. *The Autobiography of a Species in 23 Chapters*. https://doi.org/10.1176/appi.ps.51.11.1457

Riethman, H. (2008). Human Telomere Structure and Biology. *Annual Review of Genomics and Human Genetics*, *9*(1), 1–19. https://doi.org/10.1146/annurev.genom.8.021506.172017

Ryan Gregory, T. (2005). Genome Size Evolution in Animals. *The Evolution of the Genome*, 3–87. https://doi.org/10.1016/B978-012301463-4/50003-6

Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: The rosetta stone of a hidden RNA language? *Cell*, *146*(3), 353–358. https://doi.org/10.1016/j.cell.2011.07.014

Sasidharan, R., & Gerstein, M. (2008). Genomics: Protein fossils live on as RNA. *Nature*. https://doi.org/10.1038/453729a

Shammas, M. A. (2011). Telomeres, lifestyle, cancer, and aging. *Current Opinion in Clinical Nutrition and Metabolic Care*, *14*(1), 28–34. https://doi.org/10.1097/MCO.0b013e32834121b1

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3642

Sorensen, P. D., & Frederiksen, S. (1991). Characterization of human 5S rRNA genes. *Nucleic Acids Res.*, *19*(15), 4147–4151. https://doi.org/10.1093/nar/19.15.4147

Strachan, T., & Read, a P. (2004). Chapter 9: Organization of the human genome. *Human Molecular Genetics 3*. https://doi.org/10.1007/BF00711355

Sun, H., Ding, J., Piednoël, M., & Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics*, *34*(4), 550–557. https://doi.org/10.1093/bioinformatics/btx637

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and

eukaryotic complexity. *BioEssays*. https://doi.org/10.1002/bies.20544

Thapar, A., & Cooper, M. (2013). Copy number variation: What is it and what has it told us about child psychiatric disorders? *Journal of the American Academy of Child and Adolescent Psychiatry*. https://doi.org/10.1016/j.jaac.2013.05.013

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., … Rubin, E. M. (2005). Comparative metagenomics of microbial communities. *Science*, *308*(5721), 554–557. https://doi.org/10.1126/science.1107851

Tutar, Y. (2012). Pseudogenes. *Comparative and Functional Genomics*. https://doi.org/10.1155/2012/424526

Ugarković, Đ. (2013). Evolution of Alpha-Satellite DNA. In *eLS*. https://doi.org/10.1002/9780470015902.a0020829.pub2

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Koonin, E. V. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Waring, M., & Britten, R. J. (1966). Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science (New York, N.Y.)*, *154*(3750), 791–794. https://doi.org/10.1126/science.154.3750.791

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2165

Wilhelm, J. (2003). Real-time PCR-based method for the estimation of genome sizes. *Nucleic Acids Research*, *31*(10), 56e–56. https://doi.org/10.1093/nar/gng056

Wong, L. P., Ong, R. T. H., Poh, W. T., Liu, X., Chen, P., Li, R., … Teo, Y. Y. (2013). Deep whole-genome sequencing of 100 southeast Asian malays. *American Journal of Human Genetics*, *92*(1), 52–66. https://doi.org/10.1016/j.ajhg.2012.12.005

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3). https://doi.org/10.1186/gb-2014-15-3-r46

Xing, J., Witherspoon, D. J., & Jorde, L. B. (2013). Mobile element biology: New possibilities with high-throughput sequencing. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2012.12.002

Yu, S., & Lemos, B. (2016). A Portrait of Ribosomal DNA Contacts with Hi-C Reveals 5S and 45S rDNA Anchoring Points in the Folded Human Genome. *Genome Biology and Evolution*, *8*(11), 3545–3558. https://doi.org/10.1093/gbe/evw257

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3871

Zhang, Z., & Gerstein, M. (2004). Large-scale analysis of pseudogenes in the human genome. *Current Opinion in Genetics and Development*. https://doi.org/10.1016/j.gde.2004.06.003

Zhang, Z., Harrison, P. M., Liu, Y., & Gerstein, M. (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Research*, *13*(12), 2541–2558. https://doi.org/10.1101/gr.1429003

KASUTATUD VEEBRIAADRESSID

https://www.yourgenome.org/facts/what-is-a-genome

http://www.garlandscience.com/res/pdf/9780815341499_ch09.pdf

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38

http://www.ncbi.nlm.nih.gov/genome

https://biologydictionary.net/homologous-chromosomes/

http://bio3400.nicerweb.com/Locked/media/ch07/Y_chromosome.html

http://www.norwaydna.no/wp-content/uploads/2013/10/?C=N;O=A

https://www.ncbi.nlm.nih.gov/gene/100008588

https://www.genome.gov/10000715/genetic-mapping-fact-sheet/

http://oxfordindex.oup.com/view/10.1093/oi/authority.20110803100407382

http://find.thermofisher.com/Global/FileLib/qPCR/2016-Real-Time-qPCR-Handbook-branding.pdf

http://www.homolog.us/Tutorials/index.php?p=2.1&s=1

https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/

https://www.yourgenome.org/facts/what-is-the-illumina-method-of-dna-sequencing

http://www.historyofnimr.org.uk/mill-hill-essays/essays-yearly-volumes/2010-2/bringing-it-all-back-home-next-generation-sequencing-technology-and-you/

https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

http://genepool.bio.ed.ac.uk/illumina/index.html

https://genome.duke.edu/cores-and-services/sequencing-and-genomic-technologies/illumina-sequencing

https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf

https://github.com/schneebergerlab/findGSE/blob/master/R/findGSE_v1.94.R

https://github.com/refresh-bio/KMC

https://github.com/bioinfo-ut/GenomeTester4/blob/master/src/glistquery.c

LISA 1

Tabel 6. Meeste genoomi suurused findGSE, pindala ja mahaarvutustega pindala järgi ning katvus. Mida punasem on tulbas real oleva numbri lahter, seda suurem on genoom. FindGSE keskmine meestel on 3,06 Gbp, pindala 3,91 Gbp ja pindala mahaarvutuste järgi 3,59 Gbp. Genoomi suurused varieeruvad meetodite puhul seetõttu, et meetodid on erinevad. FindGSE arvutab genoomi suuruse k-meeride sageduste sageduste järgi. Pindala on arvutatud k-meeri katvuse ja vastava katvusega k-meerida hulga korrutise järgi. Pindala mahaarvutus koosneb pindalast, millest on maha arvutatud adapterid, PCR duplikaadid ja mitokondri pikkus.

	Jaotuspõhine katvus	GATK katvus	FindGSE (bp)	Pindala (bp)	Mahaarvutustega pindala (bp)
M1	27	30	3 010 020 204	3 362 193 579	3 058 736 193
M2	28	34	3 037 164 921	3 617 757 350	3 352 449 066
M3	31	32	3 086 425 867	3 377 806 921	2 652 802 728
M4	21	35	3 047 220 637	3 408 657 412	3 068 063 834
M5	19	24	3 074 467 701	3 636 356 651	3 395 719 865
M6	20	25	3 090 334 930	3 739 722 926	3 563 138 406
M7	29	34	3 054 756 907	3 536 403 044	3 132 005 516
M8	23	27	3 046 747 292	3 577 596 126	3 308 023 296
M9	28	29	3 126 233 235	3 351 141 185	2 721 786 475
M10	29	37	3 050 052 523	3 924 359 998	3 621 935 357
M11	33	35	2 994 555 477	3 263 154 984	2 896 849 678
M12	34	40	3 036 004 681	3 450 702 091	3 122 862 435
M13	31	36	3 041 722 262	3 264 388 877	3 042 314 334
M14	31	35	3 030 974 080	3 369 412 388	3 166 973 997
M15	25	28	3 053 795 090	3 571 825 810	3 097 412 795
M16	24	28	3 032 909 794	3 503 868 310	3 217 832 850
M17	23	26	3 038 641 287	3 712 551 706	3 229 925 330
M18	29	32	3 062 215 371	3 340 266 881	2 790 226 780
M19	21	24	3 014 899 392	3 514 274 210	3 237 819 841
M20	21	26	3 055 135 966	3 705 553 299	3 493 935 275
M21	20	24	3 053 570 017	3 459 735 607	3 258 485 300
M22	21	26	3 071 900 809	3 606 300 455	3 459 250 079
M23	20	25	3 090 911 675	3 551 765 424	3 362 203 369
M24	17	22	3 119 496 888	3 670 522 279	3 486 262 219
M25	25	31	3 038 410 882	3 670 803 377	3 465 298 227
M26	16	22	3 120 364 070	4 228 133 229	3 987 760 578
M27	21	26	3 065 487 904	3 615 752 447	3 415 974 651
M28	20	22	3 070 365 365	3 370 874 331	3 056 685 840
M29	16	24	3 106 082 602	4 032 503 947	3 784 275 547
M30	28	32	3 098 297 753	3 617 541 345	3 158 543 341
M31	26	31	3 076 912 730	3 563 415 476	3 373 748 982
M32	18	23	3 120 760 873	3 806 617 328	3 614 863 050
M33	21	24	3 032 454 107	3 328 064 585	3 148 394 514
M34	20	25	3 073 964 761	4 000 159 292	3 791 121 713
M35	26	31	3 034 341 256	3 550 975 740	3 341 566 547
M36	21	29	3 084 574 471	3 903 266 574	3 637 931 307
M37	22	27	3 059 844 186	3 672 847 500	3 477 854 148
M38	27	32	3 042 827 446	3 386 783 140	3 218 708 342
M39	26	31	3 033 676 096	3 484 115 731	3 261 461 716
M40	25	31	3 041 322 445	3 546 874 429	3 320 208 811
M41	24	27	3 025 828 563	3 426 031 342	3 055 272 990
M42	24	27	3 020 782 339	3 375 709 416	3 021 711 554
M43	25	31	3 040 722 170	3 728 702 690	3 561 853 154
M44	32	39	3 032 274 970	3 941 739 858	3 570 673 673
M45	32	35	3 024 342 495	3 482 582 327	3 009 077 132
M46	26	31	3 072 208 601	3 473 622 064	3 228 427 366
M47	20	28	3 053 627 877	3 411 502 507	3 220 438 595
M48	28	29	3 098 038 479	3 377 415 176	2 736 720 287
M49	26	30	3 039 329 898	3 437 894 592	3 201 131 738
M50	33	39	3 026 844 617	3 356 106 708	3 170 215 627

Tabel 7. Naiste genoomi suurused findGSE, pindala ja pindalast mahaarvutuste järgi ning katvused. Mida sinisem on tulpade järgi reas oleva number lahter, seda väiksem on genoomi suurus. FindGSE järgi genoomi keskmine suurus naistel on 3,12 Gbp, pindala 3,59 Gbp ja pindala mahaarvutuste järgi 3,61 Gbp.

Indiviid	Jaotuspõhine katvus	GATK katvus	FindGSE (bp)	Pindala (bp)	Mahaarvutustega pindala (bp)
N1	23	28	3 100 834 217	3 363 881 017	3 194 069 070
N2	23	26	3 094 970 651	3 367 281 826	2 980 753 185
N3	17	21	3 155 921 635	3 573 865 889	3 352 828 005
N4	18	22	3 172 423 338	3 834 795 078	3 444 835 497
N5	22	27	3 121 949 645	3 549 595 932	3 221 386 467
N6	20	26	3 098 469 686	3 630 683 394	3 444 018 628
N7	23	26	3 102 725 620	3 493 197 839	3 095 302 564
N8	17	22	3 151 103 793	3 749 846 904	3 546 907 027
N9	28	31	3 096 351 248	3 324 439 544	2 826 089 347
N10	29	34	3 092 352 917	3 502 707 437	3 204 136 171
N11	24	30	3 078 154 394	3 572 417 105	3 398 024 655
N12	29	34	3 072 774 005	3 437 177 910	3 108 329 121
N13	31	35	3 049 311 013	3 359 065 582	2 981 107 010
N14	16	29	3 146 753 309	3 821 204 595	3 593 086 542
N15	19	29	3 121 130 074	3 751 385 249	3 555 939 737
N16	22	25	3 101 789 295	3 355 875 846	3 041 440 933
N17	16	22	3 108 105 312	3 874 141 558	3 742 353 063
N18	26	28	3 104 397 594	3 356 936 738	2 961 863 818
N19	17	21	3 161 617 193	3 883 853 329	3 510 907 371
N20	19	24	3 111 828 120	3 668 680 027	3 522 324 479
N21	28	33	3 081 726 226	3 468 847 694	3 177 599 665
N22	23	28	3 168 895 193	3 876 316 145	3 555 347 183
N23	18	21	3 223 657 580	3 485 165 970	2 972 127 276
N24	24	35	3 101 231 874	3 580 796 073	3 382 774 739
N25	19	23	3 143 939 026	3 594 507 131	3 238 423 001
N26	16	21	3 176 053 234	4 218 740 159	4 034 978 510
N27	22	26	3 112 724 603	3 453 197 111	3 126 709 289
N28	20	24	3 116 506 218	3 469 786 020	3 154 799 693
N29	30	38	3 141 346 733	3 217 268 539	2 915 930 362
N30	16	20	3 119 911 618	3 849 602 653	3 676 292 039
N31	23	27	3 099 779 453	3 562 083 799	3 228 771 781
N32	18	21	3 252 070 668	3 575 310 331	3 178 674 582
N33	22	29	3 251 280 527	3 894 483 371	3 657 840 658
N34	17	21	3 146 894 503	3 641 455 559	3 413 001 843
N35	19	36	3 141 276 120	3 866 583 742	3 618 158 384
N36	25	31	3 128 629 213	3 553 871 563	3 316 811 688
N37	20	25	3 160 157 665	3 857 903 576	3 586 073 553
N38	25	29	3 136 219 605	3 543 183 293	3 164 461 624
N39	30	38	3 106 933 125	3 380 677 109	3 202 560 684
N40	21	26	3 098 518 744	3 627 701 577	3 354 560 702
N41	27	31	3 077 696 145	3 424 170 760	3 109 958 149
N42	25	31	3 106 691 188	3 627 055 426	3 493 153 946
N43	27	30	3 090 529 588	3 393 454 685	2 984 909 221
N44	23	27	3 082 604 921	3 514 743 315	3 155 873 245
N45	24	27	3 067 747 053	3 382 093 088	3 044 409 416
N46	23	26	3 091 035 213	3 417 435 827	3 077 097 089
N47	19	28	3 161 371 551	3 773 624 603	3 540 598 121
N48	21	26	3 130 881 603	3 642 495 328	3 307 669 099
N49	23	27	3 105 928 609	3 481 361 070	3 116 758 334
N50	27	34	3 102 425 380	3 662 635 227	3 389 312 073

LISA 2

Tabel 8. Meeste struktuurielementide pikkused genoomis. DYZ1 regiooni vahemik on 31,8-69,2 kbp. Kõigi genoomis olevate struktuurielementide pikkused varieerusid keskmisel 3% 50 indiviidi vahel. Struktuurielementide pikkused on saadud struktuurielementide k-meeride arvu (tabel 3 veerg nimega „K-meer“) ja jaotuspõhise katvuse jagatise ning struktuurielementide pikkuse korrutise järgi (tabel 3 veerg nimega „Kordusjärjestuse pikkus“).

Indiviid	DYZ1	Telomeer	45S rRNA	5S rRNA	Satelliit-DNA	TTCCA	Tsentro-meer	Alu	LINE	CCTT
M1	6807757	54843	8335436	101259	89941	5366930	8589938	51955089	90988628	189699
M2	5687780	98057	8149846	138260	90763	6540487	9752619	53764768	94086969	202095
M3	6686825	97679	7243251	128261	98994	4738679	9263462	59227965	94514260	202750
M4	6844644	62453	7424494	122438	88094	4733749	8946077	51953729	91258120	186632
M5	7395696	90895	8862320	133683	101211	7469402	9230751	50600289	95301046	199744
M6	7885290	99645	11637679	137257	98118	6846376	9916906	52241730	97026532	203111
M7	6625095	73732	9866046	193549	90350	6098383	7812129	53525752	95352229	195678
M8	7370777	78689	14761744	155033	96569	6918336	8792441	51602843	94460883	198620
M9	6366825	115589	11680371	256052	94086	6525820	9695334	58735629	92951046	201496
M10	7003572	118615	8699143	205006	96923	7719153	8821359	53281324	98061539	211864
M11	6090273	74761	9867619	89741	75065	6147258	9652706	53916909	91227384	194731
M12	7329836	97784	7503326	131176	95061	7372075	9741895	53037238	92309806	195487
M13	6580886	123305	8713539	271485	84978	6726091	9428504	58964100	91361158	198335
M14	6832839	98320	6490075	118765	87987	6913940	8520610	49074000	88789375	191705
M15	6098942	67163	12574628	139509	83552	5100683	9311853	52105752	95641659	200775
M16	5951853	72742	7680696	139468	109359	5253183	9638023	52781600	96516691	191721
M17	6487906	72971	6519022	259456	115664	6466840	10295828	51644491	96279063	206100
M18	7002703	107958	10470998	237072	90076	6248740	9188773	56357979	93205820	209328
M19	5930847	95396	6875745	113624	93837	4147839	9798162	48641800	93621941	189584
M20	5453457	124783	6361804	149835	88627	7347261	9752659	48689043	93592429	194607
M21	7279782	86924	10373509	110608	91267	6078503	9051543	52113195	93566757	190547
M22	7687106	131277	11863629	139640	92940	6443251	8796207	50530371	92276999	194933
M23	7372687	91170	8960992	203488	97133	7113722	9060597	49911180	94349569	187773
M24	6754841	117275	9004496	201356	87617	6011609	9510437	51300953	93612840	189769
M25	7388100	99904	6958958	155565	101956	8398898	8559241	50471604	94461364	200506
M26	8784414	136050	8911543	164463	104234	7714803	9485477	50644631	99981104	210585
M27	7774388	135484	7459303	187957	98075	5769325	9807762	52166714	95110146	205960
M28	9414454	101069	11704328	174721	98538	4025323	8996909	52546170	90319984	197730
M29	9529371	109842	7721008	142581	104233	7424226	9992011	50900775	96211830	204987
M30	7533935	130117	9300069	128384	96178	5521287	9463726	54267825	91760754	207153
M31	6217127	111723	8563416	142548	90348	6918039	8092055	49975050	92154983	200366
M32	7753753	113168	12570041	129216	104635	6602278	9846209	52953117	95961456	207556
M33	6910490	82302	8173905	138154	89276	5655126	8334353	50771286	90534074	198481
M34	7427783	138101	12347163	70468	92293	5167132	8739049	49273995	97809043	211748
M35	7094940	90880	10554601	115703	94056	6231429	9777504	50285931	92992272	200108
M36	8382099	129209	11230929	166294	91587	5861917	8574510	51267571	96866537	205341
M37	7475349	125825	11398644	204856	96629	6998812	9520775	49441077	93704656	200924
M38	5070208	92246	13033475	222752	85893	5362349	8507959	59815122	89786343	219370
M39	7721514	113866	10144456	212879	100361	7135979	8745867	50229462	92425748	200510
M40	7501027	111965	10923466	155030	99195	5111782	8392865	51533028	93791311	207918
M41	7232458	63187	10919954	190758	93584	6026030	8499135	55292150	92603385	202612
M42	4792481	82100	12009262	154242	93228	7268668	7596739	51407763	93567609	216597
M43	7435345	122163	11109222	73144	89062	6245999	9515056	50812320	94527311	206737
M44	5211322	80784	9013665	81953	94387	5829136	9663739	51739931	98151936	197750
M45	7431451	63836	9037852	122511	87628	5948666	8518354	52478897	94144991	191071
M46	6702846	102376	11560116	191351	87038	6607182	9027110	59091946	92361876	205263
M47	6066425	89772	10248812	139041	81741	7281663	8618468	53142420	90668970	198468
M48	6737600	122642	9596455	296829	91656	5946775	9611281	60034050	93369657	201831
M49	6722929	61370	8482380	180887	89518	5881160	8804244	52344692	92836293	188490
M50	6679091	79946	10555603	177386	75463	6260663	7193270	60109309	90936380	211028

Tabel 9. Naiste struktuurielementide pikkused genoomis. DYZ1 regiooni vahemik on 0-0,4 kbp.

Indiviid	DYZ1	Telomeer	45S rRNA	5S rRNA	Satellit-DNA	TTCCA	Tsentro-meer	Alu	LINE	CCTT
N1	35697	107865	11946244	210105	89734	6335197	7815719	63515974	94473440	206879
N2	27399	128515	7984727	199246	98292	5833836	9308913	53274730	95864675	211551
N3	14192	60437	7201068	188632	89212	6878319	10150409	54565500	98453339	194616
N4	15404	92578	5814420	130579	102611	5499739	9282184	52020867	101100977	204515
N5	16696	106033	9717774	181238	92491	6122326	8269140	50730791	96452237	188655
N6	12423	91613	9984368	125661	99603	6374646	8427248	50905215	95777102	188139
N7	11899	86519	7880034	98993	88144	6035409	9180128	52912657	95969842	197024
N8	13557	106511	12242068	76214	98776	7576768	9265021	52990941	98694373	204740
N9	27136	105801	11285702	153472	85834	5508979	9497975	54430725	94833292	205857
N10	19247	79314	6285268	208851	87584	5985536	9316151	61503176	93657510	204414
N11	9753	87524	9010082	134729	95227	7632137	9349539	54312750	96111797	203583
N12	20488	78868	7446241	172171	98909	6223031	8251428	54216062	95976960	197292
N13	34268	62579	8207261	160416	105395	5838182	8313369	55624345	96521414	197260
N14	11253	123191	8559488	107737	89180	7417169	8177188	52815525	100209750	202471
N15	11751	126563	9400939	200817	91399	6453517	8999973	54199737	98757021	204627
N16	23570	127933	10032449	146775	90129	5734708	9591071	52238059	94822997	207367
N17	10128	88686	8172497	184672	113351	6060487	9442673	53100750	101144266	197698
N18	31301	101442	10288338	114845	78683	5918763	8427104	57196258	96394191	204705
N19	67360	131060	9753185	145475	85648	6465545	11768361	50272782	100676798	214454
N20	25586	121979	8140390	198587	87595	7224920	9310374	50901221	96827464	208561
N21	60831	62540	6016789	156498	85707	6010236	8642542	52236514	96153164	191436
N22	14404	67961	13352124	172583	102265	5500181	8203331	51690104	98868466	197342
N23	33609	122944	14000952	183479	95478	4843818	9536214	57666950	98145293	215973
N24	11553	118597	10459507	119584	100404	6080261	9223797	52802063	96229881	199726
N25	14783	81531	7461458	164551	101153	6641380	9198351	54126932	100588089	200379
N26	6527	143285	6699782	94636	104939	4855507	9735821	50652131	100569266	225907
N27	41739	89931	5990543	95282	96660	6169117	8479082	53962527	98609331	204074
N28	0	80460	8232159	112281	84976	5799168	8767965	53953065	99455294	205537
N29	25567	161929	7779952	130901	91779	8434999	8832236	50277270	99830454	198995
N30	8102	150962	8503052	83625	96936	8938378	9704816	50051269	100606872	209672
N31	11116	98476	8046422	170837	98291	6808487	10569175	52597030	99348780	197043
N32	35610	140652	8762241	206771	93833	6763803	9513205	52206767	99996523	221337
N33	27171	103420	8154174	148092	102882	7244472	9396839	51647114	101131001	204744
N34	11862	107489	10008650	226673	100837	6578279	9599769	50822718	96511972	192776
N35	19332	91966	7085783	174409	92133	6248899	10220616	52320063	100520952	210169
N36	16853	112856	12758663	234328	100401	5826206	9596493	61532988	95990645	212796
N37	20526	126528	12669655	207725	104687	6156389	8960631	54350730	99419493	217806
N38	29960	85654	11669929	115960	96710	6605995	9391429	53459424	98189738	194723
N39	14164	124960	12410945	159519	99871	6323907	10042226	53097030	100460634	206563
N40	20063	98901	12391902	148454	101788	5493853	8905704	52561014	99858419	194274
N41	37877	88453	8542468	176583	99869	6029461	8469915	54494967	97590837	207587
N42	10227	138862	12089599	163504	83635	7417145	9849812	51181332	96922077	208423
N43	39344	72005	8974051	178152	99116	6483653	9467168	54656367	97365088	194926
N44	28182	96370	6438633	86873	98094	5566755	8514945	53894883	97861796	197791
N45	36310	96019	8721631	255149	100697	4944846	9019801	51653725	95200473	199192
N46	0	92806	7829557	154937	88339	6260349	9005314	52988100	96168926	201001
N47	22743	116697	9520884	131805	88549	7166398	8218917	50494200	100398079	200600
N48	13890	105581	7012932	136136	102450	7706076	9672354	54836586	100413702	193010
N49	18162	96455	14455142	192750	94249	6691742	8928854	53750830	97867813	213907
N50	10803	149525	9142861	153870	96393	6844207	10163391	52188189	100037528	211965

LISA 3

Tabel 10. Meeste struktuurielementide arv. Struktuurielementide arvud on saadud struktuurielementide k-meeride arvu (tabel 3 veerg nimega „K-meer“) ja jaotuspõhise katvuse jagatise järgi.

Indiviid	DYZ1	Telomeer	45S rRNA	5S rRNA	Satelliit-DNA	TTCCA	Tsentro-meer	Alu	LINE	CCTT
M1	51044	246792	5234	1226	485680	28981420	1356306	4675958	408292	1280468
M2	44226	457600	5307	1736	508273	36626725	1596920	5018045	437832	1414664
M3	57565	504677	5222	1783	613765	29379810	1679341	6120223	486944	1571309
M4	39916	218585	3626	1153	369995	19881747	1098641	3636761	318501	979816
M5	39022	287834	3916	1139	384600	28383727	1025639	3204685	300934	948782
M6	43795	332150	5413	1231	392473	27385503	1159872	3482782	322508	1015556
M7	53354	356372	6654	2517	524028	35370620	1324864	5174156	459567	1418662
M8	47078	301641	7896	1599	444218	31824346	1182609	3956218	361077	1142065
M9	49506	539416	7606	3215	526884	36544591	1587540	5481992	432546	1410475
M10	56402	573307	5867	2666	562156	44771089	1496020	5150528	472625	1536017
M11	55812	411185	7573	1328	495432	40571900	1862803	5930860	500333	1606534
M12	69207	554110	5933	2000	646415	50130113	1936985	6010887	521611	1661638
M13	56653	637077	6282	3774	526861	41701766	1709261	6092957	470699	1537097
M14	58822	507987	4679	1651	545521	42866428	1544672	5070980	457449	1485712
M15	42342	279847	7311	1564	417761	25503416	1361382	4342146	397381	1254845
M16	39668	290967	4287	1501	524922	25215279	1352705	4222528	384976	1150325
M17	41439	279721	3487	2676	532055	29747464	1384819	3959411	368027	1185076
M18	56395	521796	7062	3083	522443	36242691	1558330	5447938	449222	1517625
M19	34587	333887	3358	1070	394114	17420923	1203283	3404926	326751	995315
M20	31803	436739	3107	1411	372234	30858496	1197695	3408233	326648	1021687
M21	40432	289745	4825	992	365067	24314013	1058660	3474213	311008	952735
M22	44829	459470	5794	1315	390347	27061655	1080236	3537126	322057	1023396
M23	40948	303899	4168	1825	388530	28454886	1059719	3327412	313610	938864
M24	31889	332280	3560	1535	297898	20439471	945482	2907054	264487	806520
M25	51292	416267	4046	1744	509778	41994488	1251351	4205967	392477	1253163
M26	39031	362801	3316	1180	333548	24687368	887530	2701047	265863	842338
M27	45338	474193	3643	1770	411914	24231163	1204462	3651670	331945	1081291
M28	52288	336898	5444	1567	394151	16101291	1052270	3503078	300216	988649
M29	42341	292913	2873	1023	333545	23757523	934925	2714708	255840	819947
M30	58581	607213	6056	1612	538596	30919207	1549616	5064997	427007	1450069
M31	44889	484132	5178	1662	469807	35973804	1230371	4331171	398210	1302378
M32	38758	339504	5262	1043	376687	23768199	1036443	3177187	287071	934002
M33	40300	288056	3992	1301	374961	23751529	1023517	3553990	315974	1042024
M34	41254	460337	5743	632	369170	20668529	1022111	3284933	325109	1058742
M35	51227	393812	6382	1349	489089	32403430	1486638	4358114	401828	1300701
M36	48882	452232	5485	1566	384667	24620051	1053010	3588730	338075	1078040
M37	45670	461358	5832	2021	425166	30794772	1224895	3625679	342613	1105081
M38	38016	415109	8184	2697	463823	28956683	1343362	5383361	402897	1480746
M39	55751	493418	6134	2482	521875	37107090	1329781	4353220	399380	1303312
M40	52076	466519	6351	1738	495974	25558910	1227027	4294419	389693	1299485
M41	48203	252746	6095	2053	449205	28924944	1192861	4423372	369367	1215670
M42	31941	328399	6703	1660	447492	34889606	1066209	4112621	373213	1299580
M43	51620	509011	6459	820	445312	31229997	1391090	4234360	392751	1292106
M44	46310	430847	6708	1176	604078	37306468	1808419	5518926	521998	1581996
M45	66039	340457	6726	1758	560818	38071464	1594078	5597749	500688	1528568
M46	48396	443629	6990	2231	452595	34357344	1372543	5121302	399104	1334207
M47	33693	299241	4767	1247	326963	29126653	1008008	3542828	301376	992338
M48	52389	572328	6249	3727	513275	33301942	1573777	5603178	434494	1412819
M49	48541	265936	5129	2109	465492	30582033	1338657	4536540	401154	1225188
M50	61208	439701	8101	2625	498053	41320373	1388175	6612024	498737	1740981

Tabel 11. Naiste struktuurielementide arv genoomis.

Indiviid	DYZ1	Telomeer	45S rRNA	5S rRNA	Satellit-DNA	TTCCA	Tsentro-meer	Alu	LINE	CCTT
N1	228	413481	6390	2167	412778	29141906	1051237	4869558	361125	1189554
N2	175	492642	4271	2055	452142	26835645	1252076	4084396	366443	1216416
N3	67	171237	2847	1438	303320	23386283	1009105	3092045	278163	827118
N4	77	277733	2434	1054	369400	19799060	977072	3121252	302446	920318
N5	102	388788	4972	1788	406959	26938235	1063866	3720258	352659	1037601
N6	69	305376	4644	1127	398412	25498584	985643	3393681	318355	940694
N7	76	331655	4215	1021	405463	27762883	1234754	4056637	366845	1132886
N8	64	301782	4840	581	335839	25761011	921084	3002820	278844	870146
N9	211	493739	7349	1927	480673	30850284	1555224	5080201	441305	1441001
N10	155	383353	4239	2716	507989	34716109	1579932	5945307	451399	1481998
N11	65	350094	5029	1450	457091	36634257	1312216	4345020	383361	1221495
N12	165	381194	5022	2239	573672	36093578	1399365	5240886	462578	1430369
N13	295	323324	5917	2230	653449	36196730	1507102	5747849	497285	1528767
N14	50	328510	3185	773	285377	23734940	765117	2816828	266471	809885
N15	62	400782	4154	1711	347315	24523365	999997	3432650	311847	971979
N16	144	469087	5133	1448	396569	25232714	1233939	3830791	346702	1140517
N17	45	236495	3041	1325	362724	19393557	883525	2832040	268956	790792
N18	226	439584	6221	1339	409150	30777566	1281314	4957009	416528	1330581
N19	318	371336	3856	1109	291204	21982852	1169954	2848791	284445	911428
N20	135	386267	3597	1692	332861	27454695	1034486	3223744	305754	990665
N21	473	291851	3918	1965	479958	33657324	1415153	4875408	447447	1340050
N22	92	260518	7142	1780	470421	25300832	1103372	3962908	377925	1134719
N23	168	368831	5861	1481	343721	17437746	1003812	3460017	293604	971877
N24	77	474388	5838	1287	481938	29185252	1294568	4224165	383832	1198353
N25	78	258180	3297	1402	384380	25237243	1022039	3428039	317629	951801
N26	29	382094	2493	679	335804	15537622	910954	2701447	267427	903627
N27	255	329746	3065	940	425303	27144115	1090876	3957252	360546	1122407
N28	0	268201	3829	1007	339904	23196673	1025493	3596871	330581	1027686
N29	213	809644	5428	1761	550673	50609993	1549515	5027727	497742	1492460
N30	36	402566	3164	600	310196	28602810	908053	2669401	267527	838689
N31	71	377490	4304	1762	452137	31319040	1421585	4032439	379761	1132997
N32	178	421957	3668	1669	337798	24349690	1001390	3132406	299142	996018
N33	166	379206	4172	1461	452682	31875677	1208950	3787455	369766	1126094
N34	56	304552	3957	1728	342845	22366147	954363	2879954	272678	819299
N35	102	291225	3131	1486	350107	23745815	1135624	3313604	317417	998304
N36	117	470233	7418	2627	502005	29131028	1402996	5127749	398831	1329977
N37	114	421759	5893	1863	418746	24625555	1048027	3623382	330462	1089028
N38	208	356890	6785	1300	483549	33029977	1373016	4454952	407968	1217021
N39	118	624798	8659	2146	599224	37943444	1761794	5309703	500884	1549225
N40	117	346153	6052	1398	427510	23074184	1093683	3679271	348517	1019939
N41	284	398037	5364	2138	539293	32559087	1337355	4904547	437918	1401210
N42	71	578592	7029	1833	418174	37085727	1440031	4265111	402701	1302641
N43	295	324024	5635	2157	535226	35011724	1494816	4919073	436905	1315753
N44	180	369420	3444	896	451231	25607074	1145285	4131941	374077	1137300
N45	242	384075	4868	2746	483346	23735261	1265937	4132298	379726	1195149
N46	0	355755	4188	1598	406358	28797607	1211241	4062421	367606	1155756
N47	120	369539	4207	1123	336486	27232312	913213	3197966	317029	952849
N48	81	369532	3425	1282	430290	32365520	1187833	3838561	350455	1013303
N49	116	369745	7732	1988	433545	30782015	1200957	4120897	374100	1229966
N50	81	672864	5741	1863	520524	36958718	1604746	4696937	448897	1430767

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Sylvia Krupp (23.10.1995)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Inimese genoomi suuruse määramine k-meer metoodikaga,

mille juhendaja on Tarmo Puurand,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 28.05.2018