Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Karita Kimmel

Korduste arvu hindamine tsentromeeris

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Märt Möls

Tartu 2023

KORDUSTE ARVU HINDAMINE TSENTROMEERIS

Bakalaureusetöö

Karita Kimmel

Lühikokkuvõte

Inimese genoomis on piirkondi, mis sisaldavad korduseid. Üheks palju korduseid sisaldavaks piirkonnaks on tsentromeer - koht, kust saab alguse DNA replikatsioon rakkude jagunemise korral. Bakalaureusetöös keskendutakse tsentromeeri piirkonnas esinevate korduste arvule. Töös uuritakse kolme (erinevatel kromosoomidel paiknevat) korduvat motiivi ja kirjeldatakse nende kolme korduva motiivi korduste arvu jaotust Eestis. Tutvustatakse statistilisi meetodeid korduste arvu hindamiseks; segujaotusel põhinevaid mudeleid korduste arvu jaotuse kirjeldamiseks ja uuritakse kui palju korduste arvu variante võiks Eestis esineda.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: DNA, tsentromeer, k-meer, lugem, kordused, järgmise põlvkonna sekveneerimine, suurima tõepära hinnang, tõepärasuhte test, segujaotus, parameetrid, R(programmeerimiskeel).

ESTIMATION OF THE NUMBER OF REPETITIONS IN THE CENTROMERE

Bachelor thesis

Karita Kimmel

Abstract

There are certain regions in the human genome that contain repeats. One region with many repeats is the centromere - the place where DNA replication begins, when cells divide. The bachelor's thesis focuses on the number of repetitions in the centromere region. The work examines three repetitive motif which are located on different chromosomes. These three recurring motifs will be described for the distribution of the number on repetitions in Estonia. Statistical methods for estimating the number of repetitions are introduced; models based on mixture distribution to describe the distribution of the number of repetitions and how many variants of the number of repetitions could occur in Estonia.

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics.

Keywords: DNA, centromere, k-mer, read, repeats, next generation sequencing, maximum likelihood method, likelihood ratio test, mixture distribution, parameters, R(programming language).

Sisukord

Sissejuhatus

Genoom sisaldab korduvaid DNA nukleotiidide aluspaaride järjestusi. Olgugi, et korduva DNA funktsioon pole veel täielikult selge, on nende tähtsus genoomis ilmne. Sealhulgas tagavad need kromosoomide struktuuri, funktsiooni ja evolutsiooni. Üheks väga palju korduseid sisaldavaks DNA kohaks on kromosoomi heterokromatiin, mis koosneb korduste massiividest. Need massiivid on koondunud tsentromeeridesse ning moodustavad seal suure osa DNAst. Vaatamata sellele, et tsentromeere on palju uuritud, ei osata siiski öelda, kui palju inimese kromosoomis kordusi esineb.

Bakalaureusetöö eesmärk on välja töötada statistilised meetodid, mille abil saaks uurida, sekveneerimisandmeid kasutades, tsentromeerseid korduseid. Kui palju tsentromeerseid korduseid võiks esineda uuritaval inimesel? Kas eri inimestel on tsentromeeri piirkonnas esinevate korduste arvud samasugused või erinevad? Kui selgub, et tsentromeersete korduste arv võib erinevatel inimestel olla erinev, siis uuritakse, kas nähtud varieeruvust saab kirjeldada paari enamlevinud mustri abil - või on pigem igal inimesel temale endale iseloomulik arv korduseid. Kõikide teostatavate testide tulemuste olulisust hinnatakse tõepärasuhte testi abil, olulisuse nivool $\alpha = 0.05$.

Töö esimene peatükk koosneb kahest osast. Esimeses pooles on selgitatud geneetikaga seotud mõisteid töö paremaks mõistmiseks. Teine pool peatükist seab fookuse töös uuritava tentromeeri piirkonna ning sealsete eripärade kirjeldamisele. Teises peatükis on esitatud ülevaade töös kasutatavatest statistilistest meetoditest - selgitatakse, kuidas üldiseid statistikameetodeid antud töös esinevate konkreetsete probleemide lahendamiseks kasutada. Kolmandas peatükis analüüsitakse reaalseid andmeid kasutades eelnevalt kirjeldatud meetodeid. Erinevatel kromosoomidel paiknevate korduvate regioonide

pikkused võiksid teoreetiliselt (ühes kindlas populatsioonis) käituda teineteisest sõltumatult. Kontrollime antud peatükis ka seda, kas antud väide on kooskõlas tegelike vaatlusandmetega.

Bakalaureusetöö on kirjutatud tekstitöötlusprogrammiga LaTeX. Töös olevad statistilised testid, nende analüüsid ja joonised on teostatud statistikapaketi R abil. Kasutatud allikatele on töös viidatud nurksulgude abil.

Käesolevaga tänab autor bakalaureusetöö juhendajat Märt Mölsi suure abi ja toetuse eest, seejuures asjakohaste suunamiste ja paranduste eest töö valmimise eesmärgil. Samuti tänab töö autor molekulaar- ja rakubioloogia instituudi töötajat Tarmo Puurandi vajalike näidisandmete eest.

1 Töö taust

1.1 Kasutatavad geneetika mõisted

Genoom on organismi rakus olev geneetiliste juhiste kogum, mis määrab organismi tunnused ning koosneb veidi üle kolme miljardi DNA aluspaarist [1]. DNA ehk desoksüribonukleiinhape on keemiline ühend, mis kannab elusorganismides geneetilist teavet. DNA koosneb neljast lämmastikalusest: adeniin (A), guaniin (G), tümiin (T) ja tsütosiin (C). [2]

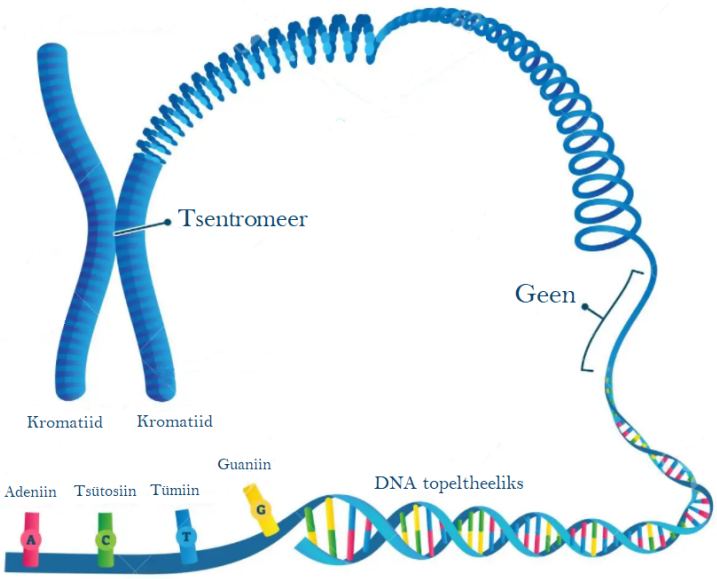
Sekveneerimine on protsess, mille käigus määratakse DNA molekuli moodustavate nelja nukleotiidi (A,T,G,C) täpne järjestus ning mille tulemusena saadakse lugemeid. Lugemid on lühikesed sekveneeritud lõigud genoomist, mis on võetud juhuslike alguspunktidega. [3] Näiteks saab pikast DNA molekuli ahelast sekveneerimise teel määrata juhuslikult valitud kohast 11 nukleotiidi väärtused – saadud 11 aluspaari (nt. ACTAGATAGGA) nimetatakse üheks lugemiks. Saadud järjestus näitab, millist geneetilist teavet teatud DNA segmendis kantakse. Antud töös kasutatud andmestik põhineb järgmise põlvkonna sekveneerimise tehnoloogial, mis võimaldab genereerida miljoneid lugemeid korraga [4].

Lugemist omakorda määratakse k-tähe pikkuseid lõike ehk k-meere (nt. 4-meere: ACTA, CTAG, TAGA jne). Võttes ühe konkreetse k-meeri, saab seda DNA järjestusest saadud lugemitest otsida. Selle tulemusena saadakse teada mitu korda antud tähekombinatsiooni seal keskmiselt esineb. Kuna sekveneerimise käigus tekitatakse palju lugemeid, siis üks ja seesama koht DNAs satub tavaliselt mitmesse lugemisse. Seega nähakse otsitavat k-meeri tavaliselt paljudes lugemites. Kui otsitav k-meer esineb uuritava inimese DNAs vaid ühes kohas, siis kutsutakse vastavat k-meeri unikaalseks. Unikaalsete

k-meeride keskmist arvu sekveneerimisel saadud lugemites kutsutakse sekveneerimiskatvuseks. Seega iseloomustab sekveneerimiskatvus seda, mitu korda keskmiselt ühte ja sedasama DNA-piirkonda sekveneerimise käigus loetakse. [5]

1.2 Korduvad motiivid genoomis

Kromosoomid on raku tuumas olevad pikad DNA ahelad, mis on pakitud geenidest moodustatud keermetaolistesse struktuuridesse (vt. joonis 1) [6]. Selleks, et kaks õdekromatiidi saaksid moodustada kromosoomi, on vajalik kinnituspunkt, mis hoiaks neid koos. Sellist punkti, DNA segmenti ehk lõiku, mis rakkude jagunemise ajal ühendab õdekromatiidide paari, luues X-kujulise struktuuri, nimetatakse tsentromeeriks. Täpsemalt on tsentromeer kromosoomi kitsendatud piirkond (vt. joonis 1), kuhu koguneb kinetokoor, valgustruktuur, mis on kromatiidide vahel siduvaks elemendiks. [7]



Joonis 1: Kromosoomi struktuur ja tsentromeeri paiknemise koht selles.

Tsentromeeri kohta ei ole palju teada, sest kordused raskendavad lühikeste lugemite põhjal DNA järjestuse assambleerimist ehk üheks järjestuseks kokkupanekut. Kuna korduva DNA tükid näevad omavahel väga sarnased välja, võib järjestuse moodustamist iseloomustada kui identsetest tükkidest pusle kokku panemist. Korduvaks piirkonnaks nimetatakse DNA lõiku genoomis, milles üks kindel aluspaaride järjestus kordub üksteise järjestikku mitmeid kordi. Näiteks:

AGTAACTG**ACTAGAACTAGAACTAGA**GGTACGTA,

kus korduv osa, mis on tähistatud paksemaks kirjas, koosneb kolmest ACTAGA motiivist. [8] Siinakohal tasub mainida, et tsentromeerile on iseloomulikud ~ 171 aluspaari pikkused kordused, kust on võimalik leida 25 DNA aluspaari pikkuseid lõike (25-meere), mis võiksid olla iseloomulikud inimese mingi kindla kromosoomi tsentromeerile. Kui oleks teada, mitmes korduses esineb ühe inimese DNAs mingile kromosoomile iseloomulik 25-meer, siis saaks teha oletusi antud kromosoomi tsentromeeri pikkuse (tsentromeersete korduste arvu) kohta.

Siiski on raske öelda, kui palju korduseid nimetatud kromosoomi piirkonnas esineb, sest tsentromeeri piirkonda ei õnnestu enamasti lugemite põhjal korduste tõttu kokku assambleerida. Küll aga on võimalik kokku lugeda k-meeri esinemissagedust sekveneerimise käigus saadud lugemites ja saadud sageduse põhjal hinnata, mitu korda antud k-meer võiks inimese DNAs esineda. [9] Korduvat DNA järjestust peetakse tähtsaks just evolutsioonis. Samuti seostatakse muutuseid korduste arvus mitmete terviseprobleemide suurema esinemissagedusega.

2 Metoodika

Selles peatükis antakse ülevaade statistilistest meetoditest, mida kasutatakse korduste arvu hindamiseks kromosoomi tsentromeeris.

Alapeatükkide 2.1, 2.2 ja 2.3 kirjutamisel toetutakse vastavalt allikate [10], [11], [12] ja [13] materjalidele.

2.1 Korduste arvu hindamine

Käesolevas töös rakendatakse suurima tõepära (ingl. *maximum likelihood*) meetodit, et leida uuritava andmestiku jaotuse tihedusfunktsiooni maksimiseerimise teel hinnatavale parameetrile tõepäraseim väärtus, mille korral realiseerunud valimi nägemise tõenäosus oleks kõige suurem.

Eesmärk on teada saada, kas inimestel on kromosoomi tsentromeeris korduste arvud samad või varieeruvad. Nagu eelpool mainitud, sisaldub tsentromeeri DNA ahelas väga palju kordusi. Teadaolevalt on iga genoomis üks kord esineva k-meeri nägemiskordade jaotus Poissoni jaotus parameetriga, mis on võrdne sekveneerimiskatvusega ehk tulemust võib kirjedada kui juhuslikku suurust $Y_1 \sim Po(\lambda)$. Kui 25-meer esineb t erinevas DNA piirkonnas (t erinevast piirkonnast pärit lugemites) siis antud 25-meeri esinemissagedus sekveneerimisandmetes on $Y_1 + Y_2 \ldots$ ja esinemissageduse jaotuseks on

$$X := Y_1 + Y_2 + \ \ldots \ + Y_t \sim Po(\lambda + \lambda + \ \ldots \ + \lambda)\,.$$

Teadmata, mitmes korduses vaatluse all oleva k-meeri sekveneerimiskatvus esineb, tähistagu indeks t selliste korduste arvu ehk $Po(t\lambda)$. Kui kõigil inimestel oleks tsentromeersete korduste arv sama, siis valimi suurusega n korral kirjeldaks juhusliku suuruse vektori $X = (X_1, X_2, \ldots, X_n)$ element $X_i \sim$

$Po(t\lambda_i)$ i-nda indiividi tulemust, kus t on üldkogumijaotust määrav ühine korduste arv ning seejuures hinnatavaks parameetriks ja vektori $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ element λ_i kirjeldab i-nda indiviidi sekveneerimiskatvust ehk unikaalsete 25-meeride katvust.

Poissoni jaotusega juhusliku suuruse tõenäosusfunktsioon on:

$$P(X_i = x_i) = \frac{(t\lambda_i)^{x_i}}{x_i!} e^{-t\lambda_i}, \qquad i \in \{1, 2, \ldots, n\},$$

mille väärtuste vektori $x = (x_1, x_2, \ldots, x_n)$ element x_i tähistab mitu korda vaadeldavat k-meeri i. indiviidi DNA lugemitest nähti. Teades, et juhusliku vektori X väärtused X_i on sõltumatud, siis saab avaldada:

$$\begin{aligned} P(X_i = x_i) &= P(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ \ldots \ \cap \{X_n = x_n\}) \\ &= P(\{X_1 = x_1\}) \cdot P(\{X_2 = x_2\}) \cdot \ \ldots \ \cdot P(\{X_n = x_n\}). \end{aligned}$$

Sõltumatute vaatluste korral avaldub tõepärafunktsioon kujul:

$$L(t) = f(x, t) = \prod_{i=1}^{n} \frac{(t\lambda_i)^{x_i}}{x_i!} e^{-t\lambda_i}.$$

Teades, et logaritmfunktsioon on monotoonne, saab lihtsuse mõttes suurima tõepära hinnangut leida ka üleminekul logaritmilisele tõepärafunktsioonile:

$$\begin{aligned} l(t) = \ln L(t) &= \ln \left(\prod_{i=1}^{n} \frac{(t\lambda_i)^{x_i}}{x_i!} e^{-t\lambda_i} \right) \\ &= \sum_{i=1}^{n} (-t\lambda_i + x_i(\ln t + \ln \lambda_i) - \ln(x_i!)) \\ &= -t \sum_{i=1}^{n} \lambda_i + \ln(t) \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i \ln(\lambda_i) - \sum_{i=1}^{n} \ln(x_i!). \end{aligned} \tag{2.1}$$

Edasi maksimiseeritakse tõepära naturaallogaritmi, mille tarvis võetakse ülaltoodud valemist tuletis:

$$\frac{\partial}{\partial t} l(t) = -\sum_{i=1}^{n} \lambda_i + \frac{1}{t} \sum_{i=1}^{n} x_i \,.$$

Võrdsustades tuletise nulliga, saab leida tõepärafunktsiooni maksimumpunkti, mis on ühtlasi parameetri t suurima tõepära hinnanguks:

$$\hat{t} = \arg\max_{t} l(t) = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} \lambda_i} \,. \tag{2.2}$$

Kuigi korduste arv on diskreetne suurus, siis tõepärasuhte testide lihtsamaks kasutamiseks käsitletakse korduste arvu kui ühte pidevat parameetrit. Ühtlasi võimaldab korduste arvu pideva suurusena käsitlemine leevendada ühte teist võimalikku probleemi - sekveneerimiskatvus pole hinnatud töö autori poolt. Kui sekveneerimiskatvuse hinnangud on süstemaatiliselt valed (näiteks 10% väiksemad sellest, mida nad peaksid olema), siis oleksidki mittetäisarvulised korduste arvud ootuspärased.

Siiani arvestati, et kõigil inimestel on tsentromeersete korduste arv sama. Olgu nüüd igal inimesel erinev korduste arv, siis juhusliku suuruse vektori X väärtused X_i kuuluvad jaotusesse $Po\,(t_i\lambda_i)$, kus vektori $t = (t_1, t_2, \ldots, t_n)$ element t_i kirjeldab tõepärasemat korduste arvu i-ndale indiviidile. Sel juhul avaldub tõepärafunktsioon kujul

$$L(t) = \prod_{i=1}^{n} \frac{(t_i\lambda_i)^{x_i}}{x_i!} e^{-t_i\lambda_i} \,,$$

mille log-tõepära avaldub:

$$l(t) = -\sum_{i=1}^{n} t_i\lambda_i + \sum_{i=1}^{n} x_i(\ln t_i + \ln \lambda_i) - \sum_{i=1}^{n} \ln(x_i!)\,. \tag{2.3}$$

Siinkohal hinnatakse valimi iga indiviidi korduste arv eraldi ehk iga vektori t elemendi suurima tõepära hinnang saadakse:

$$\hat{t}_1 = \frac{x_1}{\lambda_1}\,,\ \ \hat{t}_2 = \frac{x_2}{\lambda_2}\,,\ \ldots\ ,\ \hat{t}_n = \frac{x_n}{\lambda_n}\,. \tag{2.4}$$

2.2 Tõepärasuhte test

Tõepärasuhte test (ingl. *likelihood-ratio test*) on statistiline test hüpoteeside kontrollimiseks, mille abil saab lihtsma ja keerulisema mudeli vahel valida, kumb sobib konkreetse andmekogumi kirjeldamiseks paremini. Testi rakendamine on tõhus kui keerulisem mudel erineb lihtsamast mudelist ühe või mitme parameetri võrra.

Olukorras, kus soovitakse välja selgitada, kas valimi jaoks sobib üks ühine korduste arv või tuleb andmestiku kirjeldamiseks võtta rohkem erinevaid väärtusi, jagatakse kogu valimiruum kaheks osaks. Olgu lihtsama mudeli tõepäraks L_0 ja keerulisema mudeli tõepäraks L_1, kus lihtsama mudeli jaoks on STP meetodiga hinnatud üks parameeter t_0 ja keerulisema mudeli korral parameetrite vektor t.

Eeldusel, et keerukam mudel on sobiv, on tõepärasuhte testiks esitatud hüpoteeside paar:

$$H_0 : X_i \sim Po(t_0\lambda_i)\ \ ,$$
$$H_1 : X_i \sim Po(t_i\lambda_i)\,.$$

Kusjuures lihtsamat mudelit nimetatakse pesastatuks keerukamas mudelis, st. olgu kogu parameetriruum Θ, siis $t \in \Theta$, milles $t_0 \in \Theta_0 \subset \Theta$. Tõepärasuhte teststatistik avaldub järgmiselt:

$$\Lambda = \frac{L_0}{L_1}$$

ning selle logaritmiline tõepärasuhe

$$-2 \cdot \ln \Lambda = 2 \cdot \ln \left(\frac{L_1}{L_0} \right) = 2 \cdot (l(t) - l(t_0)) \sim \chi^2_{df} \,, \tag{2.5}$$

kus $L_1 > L_0$. Sõltumatute vaatluste ja suure valimimahu korral on teststatistiku väärtus nullhüpoteesi kehtides χ^2-jaotusega vabadusastmetega $df = h-q$, kus h ja q tähistavad vastavalt keerulisema ja lihtsama mudeli parameetrite arvu. Kui test jääb nullhüpoteesi juurde, siis valitakse lihtsam mudel, kui see aga kummutatakse, siis valitakse keerulisem mudel ning seejuures võib järeldada, et otsus on statistiliselt oluliselt parem.

2.3 Segumudel

Segujaotus (ingl. *mixture distribution*) on jaotus, mis on formuleeritud lineaarkombinatsioonina kahest või enamast jaotusest. Selle abil on võimalik kirjeldada, millisesse alampopulatsiooni üks või teine juhuslikult valitud väärtus kõige tõenäolisemalt sobib.

Olgu segumudeli jaoks rakendatavaid baasjaotusi mingi arv m ehk $Po(t_1\lambda_i)$, $Po(t_2\lambda_i), \ldots, Po(t_m\lambda_i)$. Tõenäosusfunktsiooni m-komponendilise segumude-

li jaoks, mille parameetriteks on $t = (t_1, t_2, \ldots, t_m)$ avaldub:

$$P(X_i = x_i) = p_1 \cdot P(X = x_i \,|\, \text{korduseid } = t_1) + p_2 \cdot P(X = x_i \,|\, \text{korduseid } = t_2) \\ \ldots + p_m \cdot P(X = x_i \,|\, \text{korduseid } = t_m)\,,$$

kus vektori $p = (p_1, p_2, \ldots, p_m)$ elemendid, mille väärtused kuuluvad vahemikku $0 < p \leqslant 1$ ning kogusumma on võrdne ühega, tähistavad segude kaalusid, teisisõnu, kui suure tõenäosusega juhuslikult valitud vaatlus x_i ühte või teist korduste arvu omavasse gruppi kuulub.

Selleks, et igale segumudeli parameetrile saada tõepäraseim väärtus, kasutatakse ka siinkohal STP meetodit. Tähistagu p_j ja t_j ühe m-komponendilise baasjaotuse parameetreid, siis vastav tõepärafunktsioon on esitatav kujul:

$$L(t) = \prod_{i=1}^{n} \sum_{j=1}^{m} p_j \cdot P(X_i \,|\, \text{korduseid } = t_j)$$

ning selle log-tõepärafunktsioon

$$l(t) = \sum_{i=1}^{n} \ln \left(p_1 \cdot \frac{(t_1\lambda_i)^{x_i}}{x_i!} e^{-t_1\lambda_i} + p_2 \cdot \frac{(t_2\lambda_i)^{x_i}}{x_i!} e^{-t_2\lambda_i} \right. \\ \left. \ldots + p_m \cdot \frac{(t_m\lambda_i)^{x_i}}{x_i!} e^{-t_m\lambda_i} \right). \qquad (2.6)$$

Segumudeli parameetrite hinnangute arvutamiseks pole kerge leida ilmutatud kujul valemit. Autor kasutas parameetrite hindamiseks numbrilisi meetodeid - statistikapaketi R käsku **optim()**. Viimase liikme tõenäosus on avaldatav ka kui $p_m =: 1 - p_1 - p_2 - \ldots - p_{m-1}$, mida kasutati analüüsi läbiviimisel R koodis (vt. lisa 2), et vähendada parameetrite maksimiseerimisel ja tõepära leidmisel tekkivat viga.

3 Testide rakendamine andmetega

3.1 Andmestik

Testide rakendamiseks andmetega on Tartu Ülikooli Eesti geenivaramust saadud reaalsed andmed 29 inimese kohta. Vaatluse all olev andmestik (vt. tabel 1) sisaldab nelja tunnust: sekveneerimiskatvust ja kolme k-meeri esinemissagedust lugemites.

Tabel 1: Andmestiku näidis.

k-meer	**Esinemisi lugemites**					
Sekveneerimiskatvus (Unikaalsed k-meerid keskmiselt)	20	21	...	25	16	29
GAGTGTTTCCAAACCGCTGAATGAA	2323	1859	...	1592	2640	1120
CTTCCTTGTGATATGTGCATTCAAG	438	509	...	630	599	307
AATTTTCAATGCTCTCAAAATATCC	1258	1223	...	1192	880	1795

Tabelis 1 olevad 25-meerid võiksid oletuslikult esineda vaid ühe konkreetse kromosoomi tsentromeeris (iga vaadeldav k-meer on pärit erineva kromosoomi tsentromeerist). Seejärel on loetud kokku mitu korda vastav tähekombinatsioon esines sekveneerimisandmetes. Sekveneerimisandmete pealt on erinevate inimeste jaoks hinnatud ka katvused ehk mitu korda keskmiselt unikaalseid 25-meere lugemites nähakse.

3.2 Mudelite võrdlemise analüüs

Esmalt soovitakse teada, kas inimestel on tsentromeersete korduste arv sama või mitte. Tõepärasuhte testiga kontrollitakse hüpoteeside paari:

$$H_0 : \text{Inimistel on üks ühine korduste arv,}$$
$$H_1 : \text{Igal inimisel on erinev korduste arv.}$$

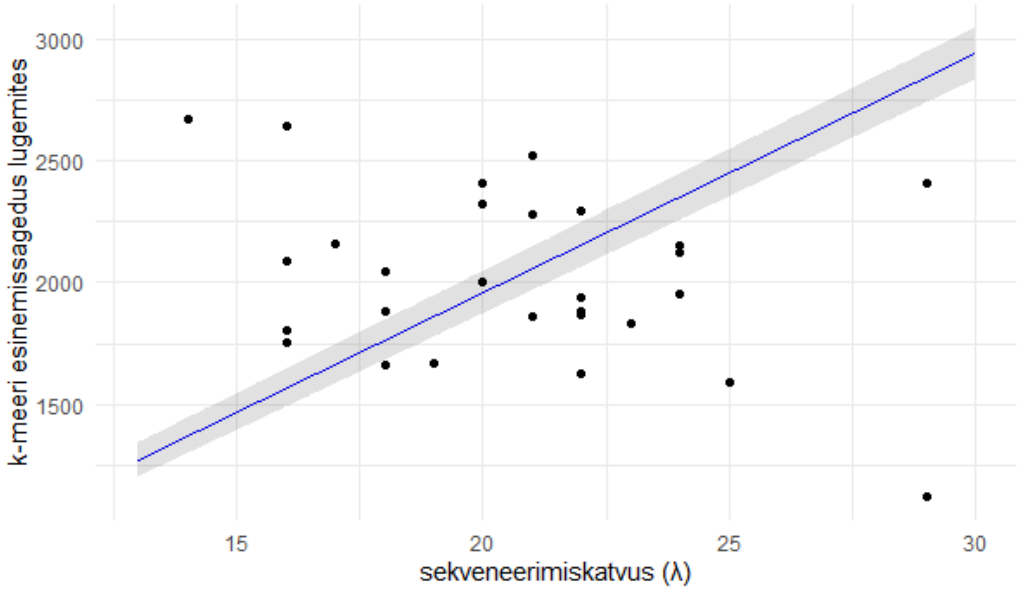
Nullhüpoteesi väide esindab lihtsamat mudelit ja alternatiivne hüpotees keerulisemat mudelit. Parameetrite hinnangud on leitud suurima tõepära meetodil, vastavalt valemite 2.2 ja 2.4 põhjal.

Tõepärasuhte teststatistiku leidmisel kasutatakse mudelite logaritmilisi tõepärafunktsioone (valemid 2.1 ja 2.3). Teststatistik mõõdab erinevust nullhüpoteesis väidetu ja andmetest ilmneva vahel - kui erinevus on piisavalt suur, kummutatakse nullhüpotees.

Kahe mudeli tõepära statistiku leidmiseks kasutatakse statistikaprogrammi R, kus vaadeldavate k-meeride jaoks on koostatud vastavad funktsioonid (vt. lisa 2). Funktsioonidele antakse ette vaadeldava k-meeri esinemissagedus lugemites, sekveneerimiskatvused ja vastavalt nendele hinnatud ühine või erinevad korduste arvud. Saadud suhte väärtusest arvutatakse, vabadusastmeid arvestades, olulisustõenäosus p. Selle põhjal saab teha otsuse esitatud hüpoteeside osas, mille juures arvestatakse valitud olulisuse nivood $\alpha = 0.05$. Olulisuse nivoo on maksimaalne lubatud tõenäosus teha I liiki viga. Esimest liiki viga tekib siis, kui võetakse vastu sisukas hüpotees, aga tegelikult on õige nullhüpotees. Seega, kui

- $p < \alpha$, siis võetakse vastu H_1 ehk igal inimesel on erinev korduste arv;
- $p > \alpha$, siis jäädakse H_0 juurde ehk inimestel on üks ühine korduste arv.

Alustuseks võeti tabelis 1 toodud esimese 25-meeri GAG. . . GAA vaatlused. Mudelite võrdlemiseks kasutatava tõepärasuhte testi tulemuseks saadi 4522.46. Vabadusastmete arvuga $df = 28$, saadi nullilähedane olulisustõenäosuse väärtus ehk nullhüpoteesi kehtides ei oleks tohtinud nii ekstreemset teststatistiku väärtust esineda. Kuna testi p-väärtus on väiksem kui määratud olulisuse nivoo, siis kummutati nullhüpotees. Seega võib kindlalt väita, et üks ühine kordus ei kirjelda andmeid sama hästi kui 29 erinevat kordust. Seda väidet kinnitab ka järgnev joonis.



Joonis 2: Sekveneerimiskatvus ja k-meeri esinemissagedus sekveneerimisel saadud lugemites. Joonisele on lisatud ühise korduste arvu hinnangule vastav sirge koos 95%-prognoosiintervalliga.

Joonisel 2 on näha, et vaadeldava 25-meeri korduste arv on inimestel väga erinev. Sinise joonega on illustreeritud vaadeldava 25-meeri ootuspärane keskmine korduste arv juhul, kui kõigil inimestel oleks ühine korduste arv. Sellele on lisatud ka 95%-prognoosiintervall (leitud eeldusel, et kõigi inimeste genoomis esineb antud k-meeri sama arv kordi), millesse satub vaid üks

vaadeldud inimene. Kui nullhüpotees oleks osutunud õigeks, peaks ligikaudu 95% punktidest paiknema graafikul näidatud vahemikus.

Kahe teise vaatluse all oleva k-meeri puhul satub 95%-prognoosiintervalli küll veidi rohkem inimesi (vt. graafikut lisas 1), aga järeldus jääb samaks - nullhüpotees ei kehti. Pole usutav, et kõigi inimeste genoomides esineksid need k-meerid sama arv kordi.

3.3 Erinevad korduste arvude variandid

Kuna eelnevate testide tulemusena selgus, et eri inimestel esineb tsentromeeris eri arv korduseid, siis järgnevalt uuriti, mitu erinevat korduste arvu varianti võiks Eesti populatsioonis olla esindatud.

Selle välja selgitamiseks koostati statistikaprogrammis R vastav funktsioon, millele anti ette hinnatavate parameetrite, k-meeride korduste arvu ja sekveneerimiskatvuse vektorid. Hinnatavateks parameetriteks on siinkohal tsentromeersete korduste arvude variandid ja nendele vastavad segude kaalud.

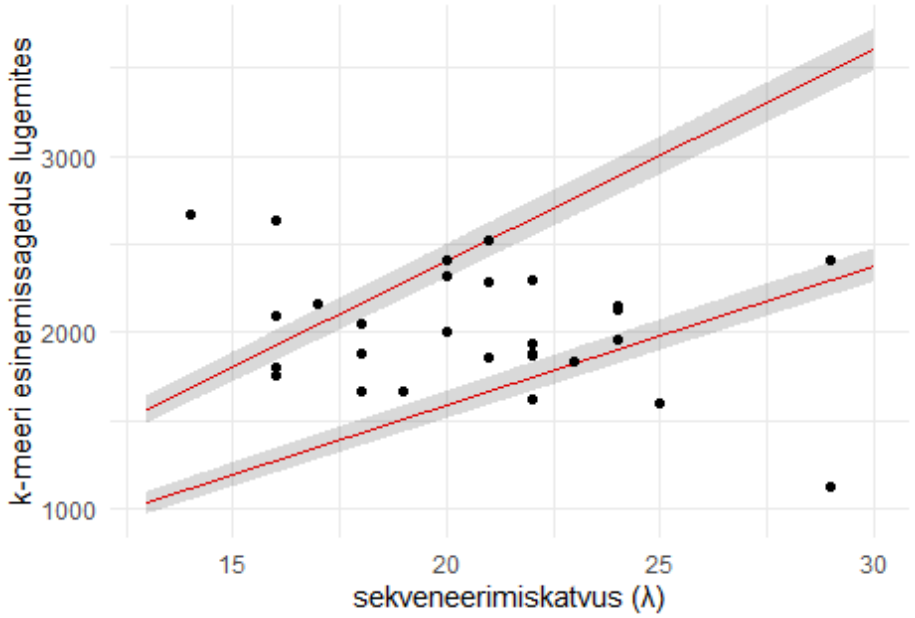
Tabel 2: Valik hinnatud parameetreid lõigu GAG...GAA vaatlustest.

Komponentide arv	Korduste arvu hinnangud	Segude kaalude hinnangud
2	(79, 120)	(0.48, 0.52)
3	(57, 97, 152)	(0.10, 0.76, 0.14)
4	(49, 84, 112, 165)	(0.03, 0.45, 0.47, 0.05)
5	(50, 84, 106, 122, 173)	(0.07, 0.40, 0.27, 0.19, 0.07)
6	(50, 85, 106, 121, 154, 181)	(0.06, 0.45, 0.23, 0.19, 0.03, 0.04)

Vaadeldi kuni 15 erinevat komponenti (populatsioonis on esindatud 15 erinevat korduste arvu) sisaldavaid mudeleid. Esmalt võrreldi kahte komponenti

sisaldavat mudelit mudeliga, mis eeldas, et kõigil inimestel on sama korduste arv. Siis võrreldi kahte komponenti sisaldavat mudelit kolme komponenti sisaldava mudeliga jne.

Kuna eelnevas alapeatükis selgus, et vaadeldud 25-meer ei ole inimestel kirjeldatav ühe kordusega, siis kontrollitakse enne mudelite rakendamist, kas kaks erinevat kordust (väärtused tabelis 2) oleks sobilik kirjeldama antud vaatlusi.



Joonis 3: Sekveneerimiskatvus ja k-meeri esinemissagedus kahekomponendilise mudeli korral. Ülemine sirge (koos 95%-prognoosiintervalliga) kirjeldab oodatavat k-meeri esinemissagedust erinevate sekveneerimiskatvuste korral juhul kui inimese DNAs esineks 120 kordust. Alumine sirge näitab otsitava k-meeri võimalikke esinemissagedusi 79 korduse korral.

Joonisel 3 on kujutatud kaks joont, mis tähistavad kahte erinevat korduste arvu (79 kordust ja 120 kordust), koos nendele vastavate 95%-prognoosiintervallidega (kui kellegil oleks 120 kordust siis peaks tema vastava k-meeri esinemissagedus jääma ülemist sirget ümbritsevasse prognoosiintervalli tõenäosusega 0.95

ja kui kellegil esineks otsitavat kordust 79 korda siis peaks tema otsitava k-meeri sagedus jääma alumist sirget ümbritsevasse 95%-prognoosiintervalli tõenäosusega 0.95). Joonise põhjal võib järeldada, et kahte segukomponenti sisaldav mudel antud vaatlusandmete jaoks ei sobi.

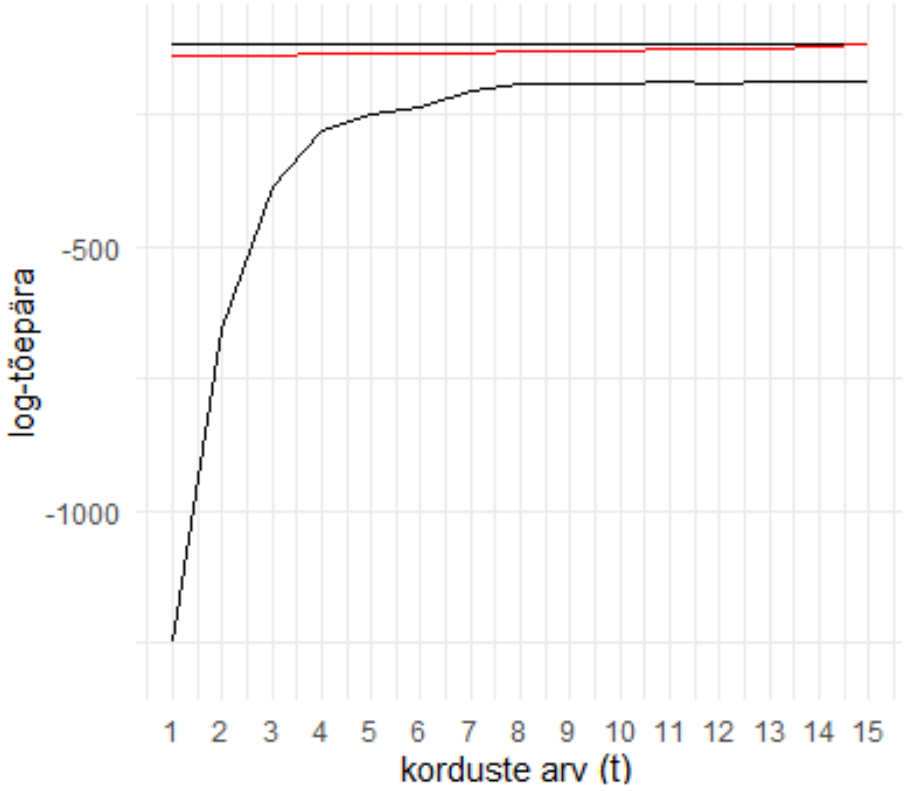
Tabel 3: Mõningaid tulemusi mudelite võrdlemisest.

	Võrreldavad mudelid	Teststatistik	df	p-väärtus
GA...AA	1 vs 2	2459.99	2	0
	4 vs 5	16.32	2	0.00028
	6 vs 7	50.71	2	$9.7 \cdot 10^{-12}$
	7 vs 8	163.24	2	0
CT...AG	4 vs 5	59.90	2	$9.9 \cdot 10^{-14}$
	6 vs 7	66.54	2	$3.6 \cdot 10^{-15}$
	7 vs 8	22.50	2	$1.3 \cdot 10^{-5}$
AA...CC	4 vs 5	264.66	2	0
	6 vs 7	16.11	2	0.00032
	7 vs 8	71.49	2	$3.3 \cdot 10^{-16}$

Kui võrreldi mistahes vaadeldud komponentide arvu (1-15) lubanud mudelit mudeliga, mis lubas igal inimesel omada talle endale iseloomulikku korduste arvu, siis tuli alati vastu võtta alternatiivne hüpotees - igale inimesele erinevat korduste arvu lubanud mudel osutus alati tõestatavalt paremaks.

Joonisel 4 on kujutatud kolm joont. Ülemise musta joonega on kujutatud log-tõepära väärtust kui kõikidel vaadeldud inimestel on erinev korduste arv ning alumine must joon kirjeldab segumudelite log-tõepärasid erinevate korduste arvu ja nendele vastavate segude kaalude korral. Nende vahel olev punane joon näitab, milline oleks olnud segumudeli log-tõepära viimane lubatud väär-

tus, et antud olulisuse nivoo $\alpha = 0.05$ korral oleks öeldud, et mudel sobib. Jooniselt on selgelt näha, et ükski valitud korduste arvu vektor ei kirjelda üldkogumit sama hästi kui 29 erinevat kordust. Küll aga on näha, et erinevate korduste arvu variantide lisamisel läheneb funktsiooni väärtus vähevaahal 95%-kvantiilile.

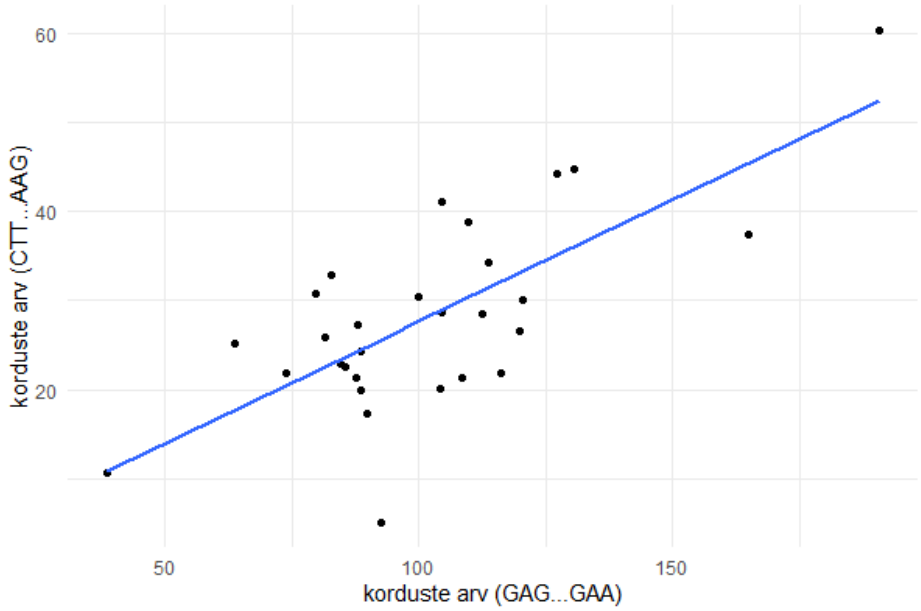


Joonis 4: Tõepärasuhte testi väärtused vastavalt korduste arvule.

4 Lineaarne regressioonimudel

Eelnevate analüüside käigus selgus, et k-meeride esinemissagedusi sekveneerimisandmetes ei olnud võimalik selgitada väikese arvu tsentromeeri variantide abil. Seega võib arvata, et igal inimesel on talle iseloomulik arv tsentromeerseid korduseid. Kas korduste arvu hinnnangud võiksid olla omavahel seotud? Kui inimese ühe kromosoomi tsentromeeris esineb palju korduseid, kas siis võiks ka tema teise kromosoomi tsentromeeris esineda tavapärasest rohkem korduseid?

Näiteks võib võrrelda 25-meeride "GAG...GAA"ja "CTT...AAG" suurima tõepära meetodil hinnatud tsentromeersete korduste arve. Korduste arvu vahelist seost on iseloomustatud joonisel 5. Seos korduste arvu hinnanguте vahel on suhteliselt tugev (Pearsoni korrelatsioonikordaja: 0.7255496 ; p-väärtus: 0.0000084).



Joonis 5: Kahe kromosoomi tsentromeeride hinnatud korduste arvude hajuvusgraafik.

Kokkuvõte

Käesolevas bakalaureusetöös hinnati sekveneerimisandmete põhjal, kas erinevad inimesed omavad tsentromeeri piirkonnas sama korduste arvu või mitte. Selle teada saamiseks kasutati tõepärasuhte testi, mille tarvis hinnati suurima tõepära meetodil mudelite jaoks korduste arvude parameetreid. Töö praktilises pooles hinnati kasutatava andmestiku abil kolme unikaalse 25-meeri korduste arvu ja sekveneerimiskatvuse kaudu korduste arvud erinevate inimeste jaoks.

Andemete analüüsi tulemusena selgus, et inimestel ei ole tsentromeersete korduste arv ühesugune. Seejärel võrreldi segumudeleid, mille parameetrite STP hinnangute leidmiseks kasutati statistikapaketi R optimeerimise käsku. Segumudelite kaudu sooviti näha, kas korduste arvu varieeruvus on kirjeldatav paari variandiga. Võrdluste tulemusena selgus, et korduste arvud ei ole grupeeritavad ehk igal indiviidil on erinev arv korduseid.

Täiendava analüüsina kontrolliti lineaarse regressioonimudeliga, kas erinevate kromosoomide tsentromeeride korduste arvud on kuidagi omavahel seotud. Selle tulemusena selgus, et seos hinnangute vahel on suhteliselt tugev.

Kasutatud allikad

[1] *A Brief Guide to Genomics* (2020). National Human Genome Research Institute. URL: https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics (vaadatud 30.03.2022).

[2] Austin, C.P. *Deoxyribonucleic Acid (DNA).* National Human Genome Research Institute. URL: https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid (vaadatud 30.03.2022).

[3] Green, E.D. *DNA Sequencing.* National Human Genome Research Institute. URL: https://www.genome.gov/genetics-glossary/DNA-Sequencing (vaadatud 30.03.2022).

[4] Kchouk, M., Gibrat, J.-F. ja Elloumi, M. (2017). *Generations of Sequencing Technologies: From First to Next Generation.* Biology ja Medicine, köide 9, nr. 3, lk 3. DOI: 10.4172/0974-8369.1000395. (Vaadatud 05.04.2022).

[5] *The Importance of Sequencing Coverage and Throughput* (2020). Thermo Fisher Scientific, lk 4. URL: https://www.thermofisher.com/ee/en/home/life-science/sequencing/sequencing-learning-center/next-generation-sequencing-information/ngs-basics/importance-coverage-throughput.html (vaadatud 01.05.2022).

[6] *An Introduction to DNA, RNA, Genes and Chromosomes* (2020). Centre for Genetics Education, lk 4. URL: https://www.genetics.edu.au/PDF/DNA_RNA_genes_and_chromosomes_fact_sheet-CGE.pdf (vaadatud 01.04.2022).

[7] *Centromere* (2021). Biology Online. URL: https://www.biologyonline.com/dictionary/centromere (vaadatud 04.04.2022).

[8] Remm, M. (2021). *Kordusjärjestuse defineerimisest*. Kogumiku "Bioloogia ja matemaatika" artikkel. (Vaadatud 06.05.2022).

[9] Valich, L. (2019). *Researchers sequence the genome's elusive centromere*. University of Rochester. URL: https://www.rochester.edu/newscenter/centromere-genome-sequencing-research-fruit-flies-381262/ (vaadatud 06.04.2022).

[10] Lepik, N. (2017). *Kursuse "Tõenäosusteooria ja statistika II" loengukonspekt*. Tartu Ülikooli matemaatika statistika instituut. URL: http://www-1.ms.ut.ee/mart/TS2/tntstat2_17.pdf (vaadatud 08.04.2022).

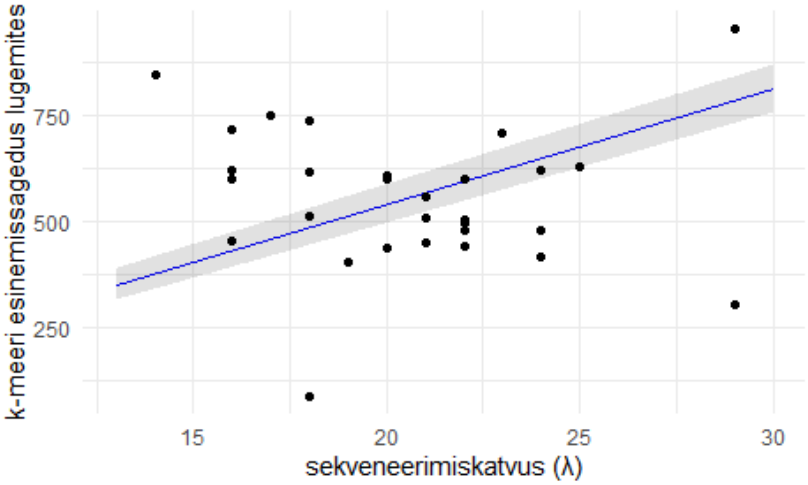
[11] *Likelihood Ratio Test*. Evolution ja Genomics. URL: https://evomics.org/resources/likelihood-ratio-test/ (vaadatud 24.04.2022).

[12] *Mixture distribution*. Analytic Visionary Modelling. URL: http://wiki.analytica.com/Mixture_distribution (vaadatud 27.04.2022).

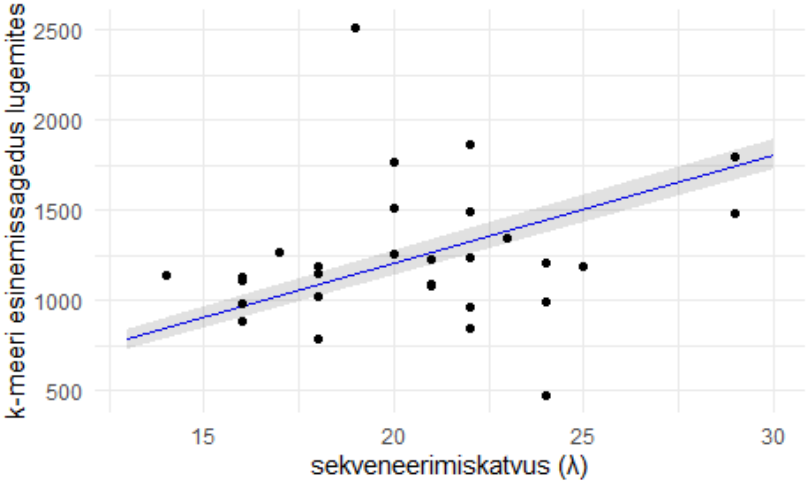
[13] *Coverage depth recommendations*. Illumina. URL: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html (vaadatud 06.04.2022).

Lisad

Lisa 1. Andmestiku põhjal koostatud joonised



Joonis 6: 25-meeri CTT. . . AAG sekveneerimiskatvus ja esinemissagedus sekveneerimisel saadud lugemites. Joonisele on lisatud ühise korduste arvu hinnangule vastav sirge koos 95%-prognoosiintervalliga.



Joonis 7: 25-meeri AAT. . . TCC sekveneerimiskatvus ja esinemissagedus sekveneerimisel saadud lugemites. Joonisele on lisatud ühise korduste arvu hinnangule vastav sirge koos 95%-prognoosiintervalliga.

Lisa 2. R kood

```
# yhise korduste arvu leidmine
kordus <- function(xi, lambda){
  return(sum(xi)/sum(lambda))
}

# yhise korduste arvu log-toepara
l0 <- function(kordus, lambda, xi){
  return(-kordus(xi,lambda)*sum(lambda)
         +log(kordus(xi, lambda))*sum(xi)+sum(log(lambda)*xi)
         -sum(xi*log(xi)-xi+log(sqrt(2*pi*xi))))
}

# igale erineva korduse arvu leidmine
kordus1 <- function(xi, lambda){
  return(xi/lambda)
}

# erinevate korduste arvudega mudel
l1 <- function(n, kordus1, lambda, xi){
  tulemus = rep(NA,n)
  esimene = rep(NA,n)
  teine = rep(NA,n)
  kolmas = rep(NA,n)
  esimene[1] = -kordus1(xi,lambda)[1]*lambda[1]
  teine[1] = xi[1]*(log(kordus1(xi,lambda)[1])+log(lambda[1]))
  kolmas[1] = -(xi[1]*log(xi[1])-xi[1]+log(sqrt(2*pi*xi[1])))
  tulemus[1] = esimene[1]+teine[1]+kolmas[1]
```

```
  for (i in 2:n){
    esimene[i] = -kordus1(xi,lambda)[i]*lambda[i]
    teine[i] = xi[i]*(log(kordus1(xi,lambda)[i])+log(lambda[i]))
    kolmas[i] = -(xi[i]*log(xi[i])-xi[i]+log(sqrt(2*pi*xi[i])))
    tulemus[i] = esimene[i]+teine[i]+kolmas[i]
  }
  return(sum(tulemus))
}

#toeparasuhte test
toepara_test <- function(lihtsam, keerukam){
  return(2*(keerukam-lihtsam))
}

# kas 1 erinevat kordust on parem kui 29?
suhe <- toepara_test(lihtsam=l0(kordus, lambda, xi),
                     keerukam=l1(n=length(xi), kordus1, lambda, xi))

#hii-ruut testi p-vaartus
hii_ruut <- 1-pchisq(suhe, df=28)

#naiteid segumudeli leidmisest ja rakendamisest
segumudel4 <- function(param, lambda, xi){
  segu1 = param[1]*dpois(xi, lambda=param[4]*lambda)
  segu2 = param[2]*dpois(xi, lambda=param[5]*lambda)
  segu3 = param[3]*dpois(xi, lambda=param[6]*lambda)
  segu4 = (1-param[1]-param[2]-param[3])*
    dpois(xi, lambda=param[7]*lambda)
  tulem = log(segu1+segu2+segu3+segu4)
```

```
  return(sum(tulem))
}

C <- optim(par=c(0.1, 0.4, 0.3, 15, 23, 33, 43),
           fn=segumudel4, n=length(xi),lambda=lambda, xi=xi,
           control=list(fnscale=-1, maxit=10000))
sum(C$par[1:3]) #kontrollin kas tulemus on ootusparane

segumudel5 <- function(param, lambda, xi){
  segu1 = param[1]*dpois(xi, lambda=param[5]*lambda)
  segu2 = param[2]*dpois(xi, lambda=param[6]*lambda)
  segu3 = param[3]*dpois(xi, lambda=param[7]*lambda)
  segu4 = param[4]*dpois(xi, lambda=param[8]*lambda)
  segu5 = (1-param[1]-param[2]-param[3]-param[4])*
    dpois(xi, lambda=param[9]*lambda)
  tulem = log(segu1+segu2+segu3+segu4+segu5)
  if(sum(param < 0) > 0){
    return(-Inf)
  } else return(sum(tulem))
}

D <- optim(par=c(0.1, 0.4, 0.25, 0.1, 15, 23, 33, 43, 55),
           fn=segumudel5, n=length(xi),lambda=lambda, xi=xi,
           control=list(fnscale=-1, maxit=10000))
sum(D$par[1:4]) #kontrollin kas tulemus on ootusparane

# kas 5 erinevat kordust on parem kui 4?
suhe51 <- toepara_test(lihtsam=C$value,keerukam=D$value)
```

```
1-pchisq(suhe51, df=2)

# kas 5 erinevat kordust on parem kui 29?
suhe52 <- toepara_test(lihtsam=D$value,
                       keerukam=l1(n=length(xi), kordus1, lambda, xi))

1-pchisq(suhe52, df=20)

# lineaarne regressioon
cor.test(kordus1(xi, lambda), kordus1(xi2, lambda))

kordused.lm = lm(kordus1(xi2, lambda) ~ kordus1(xi, lambda))
summary(kordused.lm)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karita Kimmel,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Korduste arvu hindamine tsentromeeris", mille juhendaja on Märt Möls, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalar-hiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Karita Kimmel
09.05.2023