TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Kristel Luik

Unikaalsete k-meeride keskmise katvuse hindamine ja saadud hinnangu rakendamisnäiteid

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendaja: Märt Möls, PhD

Tartu 2018

Unikaalsete k-meeride keskmise katvuse hindamine ja saadud hinnangu rakendamisnäiteid

Lühikokkuvõte. Magistritöö eesmärgiks on, esiteks hinnata sekveneerimisandmete põhjal, kas teatud plasmiid sisaldub bakteri kromosomaalses DNAs või mitte, ja teiseks hinnata inimese referentsgenoomi andmete abil, kui mitmes korduses esineb erinevatel indiviididel huvipakkuvat piirkonda genoomis. Väljapakutud meetodi sobivust, mida kasutatakse hindamaks, kas plasmiid sisaldub bakteri kromosomaalses DNAs, on esmalt hinnatud genereeritud andmete põhjal. Mõlema probleemi lahendamisel kasutatavad meetodid põhinevad unikaalsete k-meeride keskmise katvuse hinnangutel.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika
Märksõnad: DNA, parameetrid, hindamine, R (programmeerimiskeel)

Estimation of average coverage for unique k-mers and examples of usage of the estimate

Abstract. The goal of this thesis is first of all, to estimate based on sequencing data if a plasmid is integrated or not, and secondly, to estimate how many replications are present in a genome of an individual for an area of interest. The method for estimating if a plasmid is integrated or not, is firstly evaluated based on generated data. The methods for solving each of these problems are both based on estimations of average coverage for unique k-mers.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics
Keywords: DNA, parameters, estimation, R (programming language)

Sisukord

Sissejuhatus

Bakteri genoom sisaldab kromosomaalset DNAd ja plasmiide. Plasmiidid on lühikesed, enamasti rõngjad DNA ahelad, millel on omadus, mis võimaldab neil ühineda bakteri kromosomaalse DNAga. Üldiselt esineb ühte plasmiidi bakteri genoomis mitmes koopias, kuid kromosomaalse DNAga liitunud plasmiidi DNAd esineb genoomis ühes korduses. [1] DNA järjestuse väljaselgitamiseks kasutatakse sekveneerimist. Selle protsessi käigus loetakse DNA järjestusi erinevatest juhuslikest kohtadest alates, saadud tulemusi nimetatakse lugemiteks. Lugemeid võetakse väga palju ja need sobitatakse kattuvate osade abil kokku üheks täielikuks DNA järjestuseks. [2] Sellise protsessi abil on leitud paljudele bakteritele, plasmiididele, inimestele ja teistele organismidele vastavad genoomid, mida saab kasutada referentsgenoomidena uute proovide analüüsimisel.

Magistritöö eesmärgiks on, esiteks hinnata sekveneerimisandmete põhjal, kas teatud plasmiid sisaldub bakteri kromosomaalses DNAs või mitte, ja teiseks hinnata inimese referentsgenoomi andmete abil, kui mitmes korduses esineb erinevatel indiviididel huvipakkuvat piirkonda genoomis. Mõlemad huvipakkuvad küsimused annavad lisainformatsiooni haiguste uurimisel, kuid kummagi probleemi lahenduseks ei ole ühest kindlat meetodit välja kujunenud, mistõttu antud töös keskendutakse probleemide lahendamisele statistiliste meetodite abil.

Töö esimeses peatükis on selgitatud mõningaid bioloogiamõisteid ning kirjeldatud analüüsi raskendavaid andmete eripärasid. Teises peatükis on esitatud lühike ülevaade antud töös kasutatavatest statistilistest meetoditest ning hinnatud simulatsioonide põhjal väljapakutud testi sobivust. Kolmandas peatükis on kirjeldatud andmete genereerimise protsessi ning hinnatud nende põhjal testi võimsust. Neljandas peatükis on kirjeldatud bakteri ja plasmiidi andmestike analüüsi ning saadud tulemusi ja viiendas peatükis on kirjeldatud korduste andmete analüüsi ning vastavaid tulemusi. Neljandas ning viiendas peatükis kasutatud andmed on saadud National Center of Biotechnology Information RefSeq andmebaasist koostöös molekulaar- ja rakubioloogia instituudi õppejõududega.

Analüüsiks ja saadud tulemuste graafiliseks esitamiseks on kasutatud statistikapaketti R. Peamised programmikoodid on esitatud töö lisas. Kogu töö vältel on olulisuse tõenäosuseks valitud $\alpha = 0{,}05$. Magistritöö on kirjutatud tekstitöötlusprogrammiga LaTeX. Kasutatud allikatele on töös viidatud nurksulgude abil.

Käesolevaga tänab autor magistritöö juhendajat Märt Mölsi väärtuslike nõuannete ning arvukate paranduste eest. Samuti tänab autor Märt Roosaaret ning Mikk Puustusmaad töös kasutatud andmete valiku ning eeltöötluse eest.

1 Vajalikke bioloogiamõisteid

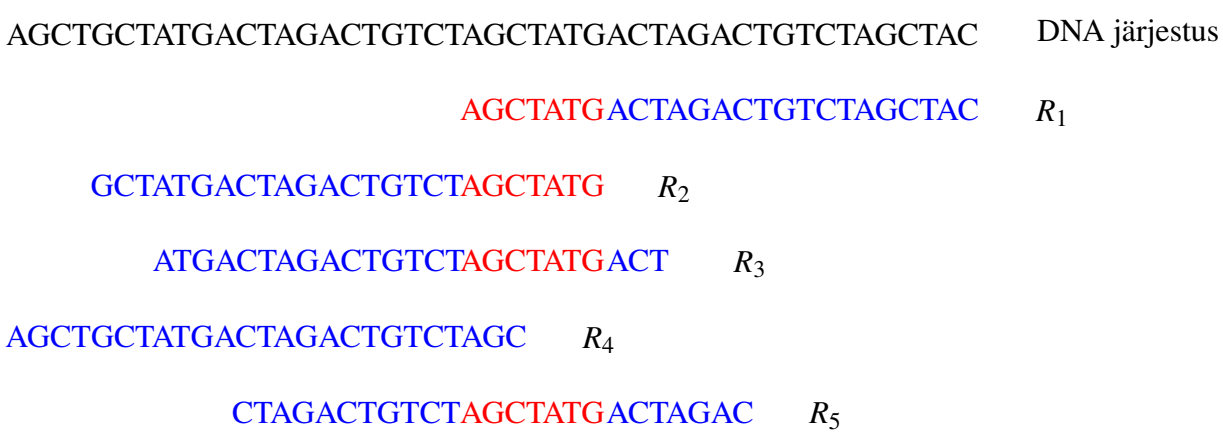
1.1 Sekveneerimine

Genoomide uurimisel on eesmärgiks määrata DNA täielik järjestus. DNA järjestus koosneb nukleotiididest. DNA molekulides nukleotiidide järjestuse väljaselgitamiseks kasutatakse sekveneerimist, mille tulemiks on lugemid. Lugemid on lühikesed sekveneeritud piirkonnad genoomist, mille pikkus sõltub sekveneerimismeetodist. Sekveneerimiskatvus on keskmine lugemite arv, mis katab ühte kindlat nukleotiidi genoomis. Saadud lugemite arv ning sekveneerimiskatvus sõltuvad samuti sekveneerimismeetodist. [2]

Lugemeid võetakse juhuslike alguspunktidega, mistõttu võime vaadelda ühest kohast algavate lugemite arve kui sõltumatuid juhuslikke suuruseid. Teadaolevalt binoomjaotus, mille korral korduste arv n on suur ja tõenäosus p on väga väike, läheneb Poissoni jaotusele [3]. Vaadeldaval juhul nukleotiidide arv genoomis on päris suur, mistõttu tõenäosus, et lugem algab ühest kindlast kohast on väga väike. Samas võetakse suur hulk lugemeid. Seega ühest konkreetsest kohast algavate lugemite arv võiks käituda kui Poissoni jaotusega juhuslik suurus, mille parameetriks on lugemite arv jagatud genoomi pikkusega ehk keskmine lugemite arv positsiooni kohta (λ).

K-meeriks nimetatakse lühikest DNA järjestust pikkusega k, seega nukleotiid on samaväärne 1-meeriga [4]. Ühest genoomist võetakse palju lugemeid, mistõttu esineb ühte ja sama k-meeri lugemites enamasti mitu korda, seega räägitakse ka k-meeride vaatlemisel katvusest. Joonisel 1 on mustaga näidatud mingi osa DNA järjestusest, sinisega on toodud viis sellest järjestusest võetud lugemit pikkusega 25 ning punasega on märgitud üks 7-meer. Jooniselt on näha, et antud 7-meeri katvus on neli, sest lugemis R_4 paikneb ainult kolm vaatluse all oleva 7-meeri nukleotiidi, kuid terve 7-meer sinna ei kuulu, aga ülejäänud lugemid R_1, R_2, R_3 ja R_5 katavad terve vaatluse all oleva 7-meeri. Samuti on näha, et kui ühe k-meeri katvus suureneb uue lugemi võtmisel, siis suureneb ka sellele k-meerile eelnevate või järgnevate k-meeride katvus, mistõttu on k-meeride katvused sõltuvad. Joonisel 1 toodud näite korral

lugemi R_5 lisandumisel suureneb 7-meeri AGCTATG katvus, kuid suurenevad ka näiteks 7-meeride GCTATGA, CTATGAC ja TATGACT katvused.



Joonis 1. K-meeri katvus DNA järjestusest võetud lugemites

Tähistame K_i-ga positsioonilt i algava k-meeri katvuse, A_i –ga i-ndalt positsioonilt algavate lugemite arvu ja L-iga lugemi pikkuse, siis saame arvutada välja k-meeri katvuse kohal i, k, A_i ja L-i kaudu:

$$K_i = A_i + A_{i-1} + A_{i-2} + \ldots + A_{i-L+k}, \tag{1}$$

mistõttu K_{i+1} avaldub järgnevalt:

$$K_{i+1} = A_{i+1} + A_i + A_{i-1} + \ldots + A_{i-L+k+1}. \tag{2}$$

Saadud võrdustest on näha, et lähestikku olevate k-meeride katvused avalduvad osaliselt samade juhuslike suuruste summana. Samuti on nende võrduste põhjal näha, et $i+L-k+1$ on esimene positsioon, mis enam ei sisalda ühtegi samasugust liidetavat nagu k-meeri katvus kohal i, mistõttu k-meeride katvused kohal i ja kohal $i+L-k+1$ on sõltumatud, kõik vahepealsed katvused lõigus $[i, i+L-k]$ on sõltuvad.

Kui i-ndalt positsioonilt algavate lugemite arv A_i on Poissoni jaotusega juhuslik suurus parameetriga λ ning vektori A liikmed on sõltumatud, siis nende juhuslike suuruste summa K_i on samuti Poissoni jaotusega juhuslik suurus, mille keskväärtus on liidetavate juhuslike suurus-

te parameetrite summa [3]. Antud juhul on kõik vektori A liikmed samade keskväärtustega, mistõttu K keskväärtus on avaldatav kui $(L-k+1)*\lambda$.

1.2 Probleemi kirjeldus

Bakterite genoomid sisaldavad bakteri kromosomaalset DNAd ja plasmiidide [1]. Plasmiidid on väikesed enamasti rõngjad DNA ahelad, mis sisaldavad ainult mõnda geeni ning need on kergesti ülekantavad ühelt bakterilt teisele [5]. Lisaks plasmiidide ülekandele, võib plasmiid liituda bakteri kromosomaalse DNAga, sellisel juhul on tegemist integreerunud plasmiidiga. Integreerunud plasmiidi korral ei ole võimalik teistel bakteritel neid enda genoomi üle kanda, kuid bakterid annavad plasmiidid oma järglastele enamasti edasi. Integreerunud plasmiidide omadused antakse edasi alati kromosomaalse DNA kujul, teisel juhul proovitakse paljundada plasmiide piisavalt, et pooldudes satuks juhuse tahtel mõlemasse poolde mõned plasmiididest ja omadused antakse edasi plasmiidide kujul. Seetõttu on ühte plasmiidi bakteri genoomis mitmes korduses, kuid integreerunud plasmiidi korral esineb plasmiidis sisalduvat DNAd ainult ühes korduses bakteri kromosomaalse DNA ahelas. [1] Lisaks plasmiidide ülekandele on bakteritel omane väga kiire DNA muteerumine, mis muudab nende analüüsi keerulisemaks [5].

Haiguste uurimisel võetakse pidevalt uusi proove, et tuvastada millise bakteriga on tegu ja milliseid plasmiide vastav bakter võib sisaldada [5]. Uute proovide sekveneerimisel saadud lugemeid võrreldakse referentsgenoomidega. Referentsid jagatakse kindla pikkusega k-meerideks ja seejärel vaadatakse nende k-meeride sisalduvust uuest proovist saadud lugemites, mille tulemusena saadakse igale k-meerile vastavad katvused.[4]

Vaatluse all olevad bakterite ja plasmiidide andmestikud koosnevad kahest tunnusest. Üks on katvus ja teine tunnus näitab vastava katvusega k-meeride arvu. Huvipakkuvaks tunnuseks on keskmine katvus üle genoomis esinevate unikaalsete k-meeride. Unikaalseteks nimetame k-meere, mida esineb vastavas genoomis vaid ühes kohas. Kui unikaalsete k-meeride katvuste keskmine on nullilähedane, siis võib eeldada, et uuritavas proovis ei sisaldu vastavat plasmii-

di või bakterit. Sageli võib juhtuda, et keskmine katvus on nullist suurem, kuid üks mood on siiski null või nullilähedane. Kui vastavat *k*-meeri tegelikult bakteris ei esine, võime siiski vahel näha vastavat *k*-meeri lugemites sekveneerimisvigade tõttu. Muutuste tõttu bakteri DNA ahelas esineb peaaegu alati *k*-meere, mis peaksid referentsgenoomi järgi otsustades bakteris esinema, kuid tegelikult seal ei ole. Lisaks tuleb sageli ette olukordi, kus lisaks ühele moodile nullilähedal, leidub *k*-meeride sagedusjaotusel veel üks, kaks või enam moodi. Vahel esineb DNA lõikude duplitseerumist, mistõttu sama DNA lõik esineb kaks või enam korda bakteri genoomis ja vastavaid *k*-meere sisaldavaid lugemeid leidub mitmes erinevas piirkonnas. Unikaalsete *k*-meeride keskmise katvuse hindamisel jätame sellised *k*-meerid vaatluse alt välja. Kuna katvus on üldjuhul Poissoni jaotusega juhuslik suurus, siis antud kirjeldusele vastab Poissoni jaotuste segu, mille üks segukomponent on nullilähedane, teine on nullist suurem ja järgmised komponendid on teise segukomponendi mingi arvu kordse parameetri väärtusega.

Enamasti on lisaks multimodaalsusele andmete jaotus päris pika sabaga. See tähendab, et vaatluse all on mõned üksikud *k*-meerid, mida on lugemitesse sattunud väga palju kordi. See võib samuti olla põhjustatud *k*-meeride mutatsioonidest ja sekveneerimisvigadest. Potentsiaalselt võib mõningaid üksikuid *k*-meere proovi genoomis esineda peaaegu lõpmatult paljudes kohtades (kordustes), kuid mida suuremaks läheb korduste arv, seda väiksemaks jääb sellise kordusega *k*-meeride arv, mistõttu sellised vaatlused ei paku meile enam huvi ning tuleks määrata vastav piir, millest suuremad vaatlused oleks mõistlik parempoolselt tsenseerida.

Bakteri unikaalsete *k*-meeride keskmise katvuse võrdlemisel plasmiidi unikaalsete *k*-meeride keskmise katvusega, saab hinnata, kas antud plasmiid sisaldub bakteri kromosomaalses DNAs. Kui nii bakteri kui ka plasmiidi keskmised katvused on ligilähedased, siis võib eeldada, et plasmiid sisaldub vaatluse all oleva bakteri kromosomaalses DNAs.

2 Metoodika

Antud peatükis on toodud ülevaade statistilistest meetoditest, mida kasutame unikaalsete *k*-meeride keskmise katvuse hindamiseks ning bakteri ja plasmiidi unikaalsete *k*-meeride keskmiste katvuste võrdlemiseks. Esmalt arutleme selle üle, kuidas võiksime probleemi lahendada sõltumatute vaatluste korral ja seejärel, peatükkides 2.2 ja 2.3, pakume välja lahenduse sõltuvate vaatluste jaoks.

2.1 Tõepärafunktsioon ja tsenseerimine

Olgu $\boldsymbol{x} = (x_1,...,x_n)$ valim pikkusega n, siis sõltumatute vaatluste korral oleks valimi tõepärafunktsioon kujul

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

ja log-tõepärafunktsioon avalduks järgmiselt

$$l(\boldsymbol{\theta};\boldsymbol{x}) = \ln L(\boldsymbol{\theta};\boldsymbol{x}) = \ln\left(\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln f(x_i, \boldsymbol{\theta}),$$

kus $f(x, \boldsymbol{\theta})$ on jaotuse tihedus- või tõenäosusfunktsioon parameetritega $\boldsymbol{\theta}$. [6] Praeguses rakenduses on $f(x_i, \boldsymbol{\theta})$ rollis Poissoni jaotuste segu ja $\boldsymbol{\theta}$ moodustaksid segude osakaale kirjeldavad parameetrid ja parameeter, mis kirjeldab segude paiknemist (katvus, λ).

Parempoolselt tsenseeritud vaatluste korral ei ole täpselt vaatluse väärtus teada, kuid teatakse mingit piiri, millest vaatluse väärtus on suurem [6]. Vaatluse all oleval juhul on täpsed andmed teada, kuid määrame ise piiri, millest suuremaid vaatlusi vaatleme kui tsenseeritud vaatlusi. Ühte *k*-meeri võib proovis esineda ühes, kahes või enamas kohas, kuid mida suuremaks läheb korduste arv, seda väiksemaks jääb k-meeride arv, mida vastavas korduses proovis esineb. Seetõttu on vaadeldavatel andmetel pikk saba, mis võib mõjutada unikaalsete *k*-meeride keskmise hinnangut, sest me ei taha kaasata lõputult palju unikaalsete *k*-meeride keskmise kordusi oma mudelisse, mistõttu mingist piirist alates tsenseerime vaatlused.

Juhul kui osad vaatlustest on parempoolselt tsenseeritud, siis tõepärafunktsioon avaldub järgnevalt. Olgu $T_1,...,T_n$ tegelikud väärtused ja C tsenseerimispunkt, mis vaatluse all oleval juhul,

on iga vaatluse korral konstante. Parempoolselt tsenseeritud vaatlusteks nimetatakse vaatluste paari (X_i, Δ_i), $i = 1, ..., n$, kus $X_i = min(T_i, C)$ ja $\Delta_i = I(T_i \leq C)$. Kui C on konstantne, siis tõepärafunktsioon on avaldatav kujul

$$L(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{\delta}) = \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i}$$

ja log-tõepära avaldub järgnevalt

$$l(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \Big(\delta_i \cdot \ln(f(x_i)) + (1 - \delta_i) \cdot \ln(S(x_i)) \Big),$$

kus $S(x_i) = \boldsymbol{P}(T_i > x_i)$.[6]

2.2 Tõepärasuhte test

Keskväärtuste võrdlemiseks saab kasutada tõepärasuhte testi, hinnates suurima tõepära meetodil lihtsama mudeli korral ühe keskväärtuse ja keerulisema mudeli korral kaks keskväärtust ning seejärel hinnates, kas keerulisem mudel sobitub oluliselt paremini. Kui tulemuseks saame, et keerulisem mudel on oluliselt parem, siis võime järeldada, et keerulisema mudeli poolt hinnatud keskväärtused on erinevad ning kui keerulisem mudel ei tule statistiliselt oluliselt parem, siis ei saa tõestada, et keskväärtused oleksid erinevad.

Olgu lihtsama mudeli tõepäraks L_0 ja keerulisema mudeli tõepäraks L_1, kusjuures lihtsama mudeli korral on hinnatud p_0 parameetrit ja keerulisema mudeli korral p_1 parameetrit, siis teststatistik

$$\chi^2 := -2 \cdot \ln\left(\frac{L_0}{L_1}\right) = 2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2 \cdot (\ln L_1 - \ln L_0) \tag{3}$$

on sõltumatute vaatluste ja suure valimimahu korral nullhüpoteesi kehtides χ^2 jaotusega vabadusastmete arvuga $p_1 - p_0$. Kui test jääb nullhüpoteesi juurde, siis jääme lihtsama mudeli juurde ning kui test võtab vastu sisuka hüpoteesi, siis on keerulisem mudel statistiliselt oluliselt parem kui lihtsam mudel.[6]

Tõepärasuhte teststatistik on nullhüpoteesi kehtides asümptootiliselt χ^2 jaotusega sõltumatute juhuslike suuruste korral [6]. Modifitseerime tõepärasuhte teststatistikut (vt valem 3) nii,

et see võiks sobida teatut liiki sõltuvate juhuslike suuruste korral. Juhul kui mingi konstantse indeksite vahemiku t järel on juhuslikud suurused sõltumatud, s.t. kui mingi t ja iga i korral juhuslikud suurused X_i ja X_{i+t} on sõltumatud, siis saame grupeerida log-tõepärad ainult üle sõltumatute vaatluste arvutatud summadeks ning arvutada nende korral teststatistikute väärtused $\chi^2(X_1,\, X_{t+1},\, X_{2t+1},\, ...)$, $\chi^2(X_2,\, X_{t+2},\, X_{2t+2},\, ...)$, ..., $\chi^2(X_t,\, X_{2t},\, X_{3t},\, ...)$, mis on nullhüpoteesi kehtides asümptootiliselt χ^2 jaotusega vabadusastmete arvuga $p_1 - p_0$. Valemite 1 ja 2 põhjal on vaadeldaval juhul t väärtuseks $L-k+1$, kus L on lugemi ja k on k-meeri pikkus.

Oletame, et modifitseeritud teststatistiku väärtuseks võiks võtta selliste valiidsete teststatistikute aritmeetilise keskmise. Valiidseid teststatistikuid saame moodustada t tükki, seega avaldub modifitseeritud testatistik järgnevalt

$$\chi^2_{k-meer} := \frac{\chi^2(X_1,...) + \chi^2(X_2,...) + ... + \chi^2(X_t,...)}{t}. \tag{4}$$

Avaldades valemis 4 kõik teststatistikute väärtused valemi 3 kaudu ning võttes arvesse, et igal liidetaval peaks olema ligikaudu sama suur panus log-tõepärasse, saame järgneva tulemuse

$$\begin{aligned} \chi^2_{k-meer} &: = \frac{2[\ln L_1(X_1,..) - \ln L_0(X_1,..) + \ln L_1(X_2,..) - \ln L_0(X_2,..) + .. + \ln L_1(X_t,..) - \ln L_0(X_t,..)]}{t} \\ &= \frac{2}{t} \cdot (\ln L_1 - \ln L_0), \end{aligned} \tag{5}$$

kus viimases avaldises $\ln L_1$ ja $\ln L_0$ on arvutatud üle kõikide vaatluste, käsitledes vaatluseid sõltumatutena.

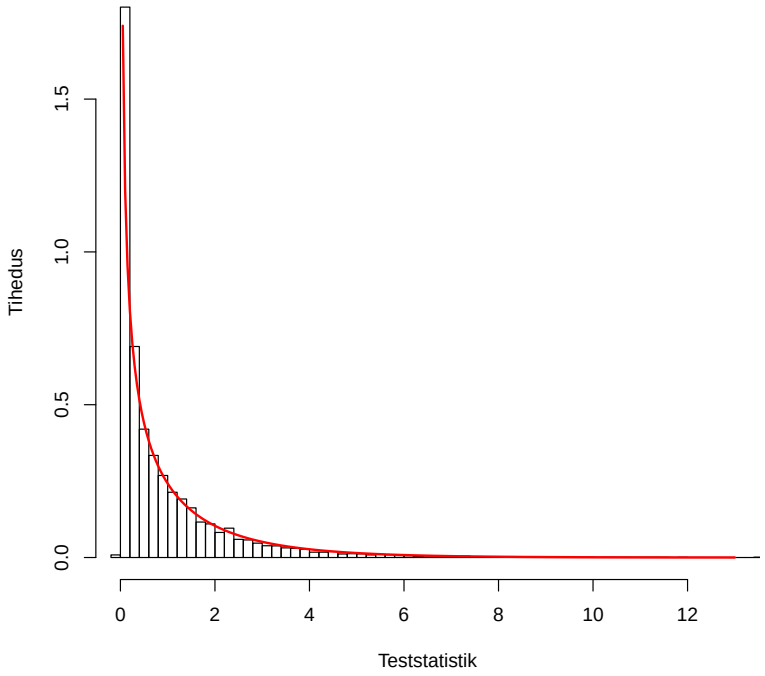
Oletame, et võrreldes sellisel viisil konstrueeritud teststatistikut χ^2 jaotuse kriitilise väärtusega saame täpse või konservatiivse testi vähemalt väikeste olulisusenivoode korral. Teststatistiku konstruktsioon on heuristiline, mistõttu töö käigus kontrollime väljapakutud teststatistiku sobivust simulatsioonide abil.

2.3 Teststatistiku jaotus

Genereerisime reaalsete andmetega sarnasel põhimõttel andmed, et nende põhjal hinnata testi sobivust, kui tõelised parameetrid on teada. Andmete simuleerimise protsessi on kirjeldatud

järgmises peatükis, kuid siinkohal kontrollime eelnevalt püstitatud hüpoteesi õigsust olukorra jaoks, kus $p_1 - p_0 = 1$.

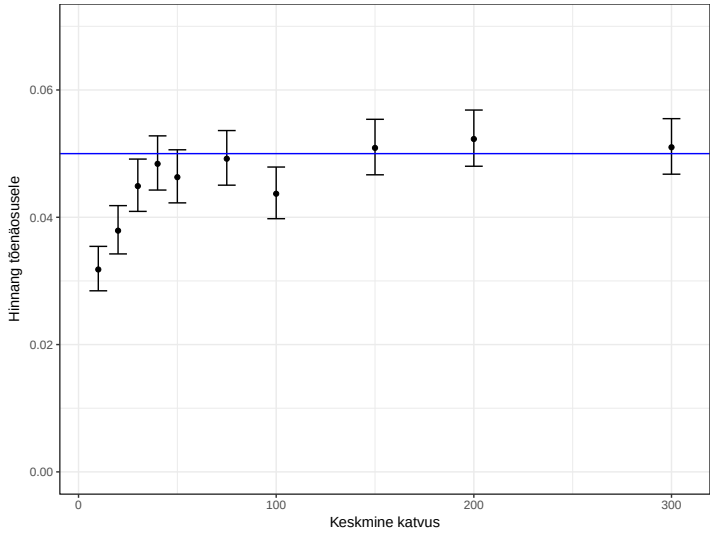
Joonisel 2 on toodud 100 000 simulatsiooni pealt hinnatud teststatistiku jaotus nullhüpoteesi kehtides ning punase joonena on lisatud χ^2 jaotuse tihedusfunktsioon ühe vabadusastme korral. Jooniselt 2 on näha, et võime eeldada nullhüpoteesi kehtides teststatistikult ühe vabadusastmega χ^2 jaotust.



Joonis 2. Teststatistiku jaotus nullhüpoteesi kehtides

Joonisel 3 on 100 000 simulatsiooni ja erinevate unikaalsete 1-meeride keskmiste katvuste korral hinnangud tõenäosusele, et nullhüpoteesi kehtides on teststatistiku väärtus suurem kui ühe vabadusastmega χ^2 jaotuse 0,95-kvantiil ning hinnangute 95%-usaldusintervallid, $x-$teljel on toodud unikaalsete 1-meeride keskmised katvused ja $y-$teljel hinnangud I liiki vea tegemise tõenäosusele. Jooniselt 3 on näha, et I liiki vea tegemise tõenäosuse hinnangud püsivad väikeste keskmiste katvuste korral kindlalt allpool 0,05-te, kuid keskmise katvuse

kasvades hinnangud I liiki vea tegemise tõenäosusele lähenevad 0,05-le.



Joonis 3. Hinnangud I liiki vea tegemise tõenäosusele erinevate katvuste λ korral

3 Testi võimsus

Testi võimsuse arvutamiseks genereerisime reaalsete andmetega sarnasel põhimõttel andmed, kasutades statistikapaketti R. Antud peatükis on kirjeldatud andmete simuleerimise protsessi ning seejärel on toodud simulatsioonide põhjal saadud hinnangud.

3.1 Andmete genereerimine

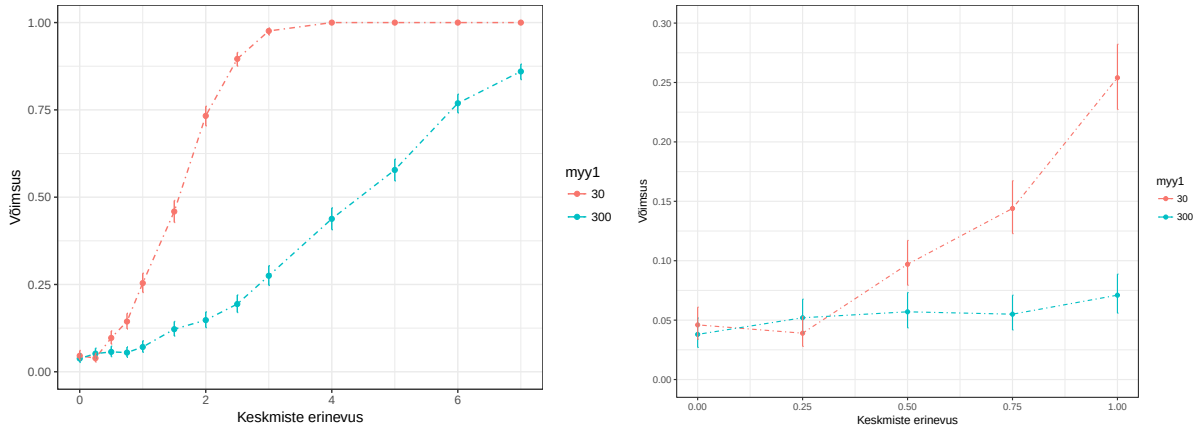
Alustuseks määrasime lugemi pikkuseks $L = 100$ ja ühe simulatsiooni pikkuseks võtsime $N = 10000$. Seejärel genereerisime kaks sama keskväärtusega Poissoni jaotusest juhuslike suuruste vektorit A_1 ja A_2 samuti pikkusega N. $A_1[i]$ ja $A_2[i]$ on i-ndalt positsioonilt algavate lugemite arvud, kusjuures vektorite A_1 ja A_2 kõik liikmed on sõltumatud. Simulatsioonide saamiseks liitsime seejärel kokku vastavad A_1 vektori liikmed, et saada 1-meeride katvused erinevatel positsioonidel i (vt valem 1). Selleks, et simuleerida võimalikke tekkinud kordusi, kus sama 1-meer esineb genoomis mitmes kohas, liitsime väikese osakaalu positsioonide korral juurde A_2 vektori samade indeksitega liikmed, et osade 1-meeride katvused oleksid kahekordistunud. Genereerisime kaks eraldi andmestikku, mis sisaldasid vastavalt bakteri ja plasmiidi 1-meere, erinevate unikaalsete 1-meeride keskmiste katvuste ja duplitseerunud 1-meeride osakaalude korral.

Soovisime simulatsioonide põhjal hinnata, kas paremini sobiks ühe või kahe hinnatud unikaalsete 1-meeride keskmise katvusega mudel. Keerulisema mudeli tõepäraks võtsime kahe simulatsiooni tõepärade summa, kus oli suurima tõepära meetodi abil leitud mõlema simulatsiooni unikaalsete 1-meeride keskmine katvus ja unikaalsete 1-meeride osakaal. Lihtsama mudeli tõepära leidmiseks hindasime kahe simulatsiooni peale kokku ühe unikaalsete 1-meeride keskmise katvuse, kuid mõlema simulatsiooni jaoks eraldi osakaalud. Duplitseerunud 1-meeride keskmiseks katvuseks võtsime mõlemal juhul kahekordse leitud unikaalsete 1-meeride keskmise katvuse hinnangu. Modifitseeritud tõepärasuhte teststatistiku abil hindasime, kas paremini sobib lihtsam või keerulisem mudel ehk kas andmeid kirjeldab paremini üks või kaks hinnatud unikaalsete 1-meeride keskmist katvust.

3.2 Hinnangud simulatsioonidelt

Kordasime eelnevalt kirjeldatud andmete genereerimise protsessi $n = 1000$ korda, valides ühel korral esimese simulatsiooni unikaalsete 1-meeride keskmiseks katvuseks 300 ja teisel korral 30. Nende simulatsioonide pealt hindasime testi võimsuse ning selle usaldusintervalli erinevate unikaalsete 1-meeride keskmiste katvuste erinevuste korral.

Joonisel 4 on kujutatud simulatsioonidelt hinnatud testi võimsus koos 95%-usaldusintervallidega, y-teljel on testi võimsus ning x-teljel on kahe simulatsiooni unikaalsete 1-meeride keskmiste katvuste vahe. Punasega on tähistatud simulatsioonid, mille korral esimese simulatsiooni keskmine katvus oli 30, ning sinisega simulatsioonid, mille esimese simulatsiooni keskmine katvus oli 300. Vasakpoolsel joonisel on toodud testi võimsused kui unikaalsete 1-meeride keskmiste katvuste vahe on nullist seitsme ühikuni ning parempoolsel joonisel on suurendatud vasakpoolse joonise alumist vasakut nurka, kus keskmiste katvuste erinevused jäävad kuni ühe ühiku piiridesse.



Joonis 4. Simulatsioonidelt hinnatud testi võimsus keskmiste katvuste erinevuste lõikes

Vasakpoolselt jooniselt on näha, et kui keskmine katvus on 30 lähedal, siis võimsus suureneb kiiremini võrreldes keskmise katvusega 300 lähistel. Saadud tulemus on loogiline, sest väike keskmiste katvuste erinevus mõjutab väiksemat keskmist katvust tugevamini kui suurt. Mõlema unikaalsete 1-meeride keskmise katvuse korral on testi võimsuse käitumine oodatav. Parempoolse joonise pealt on näha, et väga väikeste keskmiste katvuste erinevuste korral

jääb testi võimsus alla 0,05, mis tagab nullhüpoteesi juurde jäämise, kui kahe simulatsiooni keskmised katvused on ligilähedaselt võrdsed. Kuid kui keskmine katvus on 30 lähedal, siis hakkab testi võimsus päris kiiresti tõusma ja ühe ühiku erinevuse korral on testi võimsus juba 0,25. Kui keskmine katvus on 300 lähistel, siis kasvab võimsus palju aeglasemalt, ühe ühiku erinevuse korral on võimsus 0,075. Üks ühik erinevust 30-kordse katvuse korral tähendab 3,333% suurust muutust, üks ühik erinevust 300-kordse katvuse korral aga kõigest 0,333% suurust muutust, seega viimasel juhul on nähtud väiksem võimsus ootuspärane.

4 Bakterite ja plasmiidide andmed

Rakendame simulatsioonide põhjal toimivat testi bakteri ja plasmiidi k-meeride esinemissagedusi sisaldavate andmestike analüüsiks. Peatüki alguses on kirjeldatud täpsemalt analüüsi käiku ning seejärel on esitatud sekveneerimissimulaatori MetaSim abil genereeritud andmete ning reaalsete bakterite ja plasmiidide andmete analüüsi tulemused.

4.1 Analüüsi käik

Nagu varasemalt kirjeldatud, lähendame k-meeride esinemissageduste jaotuse Poissoni jaotuste segule, kus üks segukomponent on null või nullilähedane, teine on nullist suurem ning kolmas komponent on kahekordse teise komponendi parameetri väärtusega. Kuigi k-meeride katvuste jaotuse täielikuks kirjeldamiseks läheks vaja rohkem komponente, sest jaotusel on pikk parempoolne saba, siis antud juhul huvitab meid ainult teise komponendi parameetri hinnang ehk unikaalsete k-meeride keskmise katvuse hinnang, mistõttu, mudeli lihtsuse huvides, kaasame veel ainult kolmanda komponendi ning teatud piirist alates tsenseerime vaatlused.

Esiteks võtame vaatluse alla mõlemad andmestikud eraldi ja hindame mõlema andmestiku jaoks nullilähedase katvusega ehk puuduvate k-meeride osakaalud, vastavalt $\hat{\pi}_{01}$ ja $\hat{\pi}_{02}$, unikaalsete k-meeride keskmised katvused $\hat{\mu}_1$ ja $\hat{\mu}_2$, ning unikaalsete k-meeride osakaalud $\hat{\pi}_1$ ja $\hat{\pi}_2$. Puuduvate k-meeride keskmiseks katvuseks fikseerime väärtuse 0,1, sest mõned k-meerid, mida tegelikult proovis ei esine, võivad andmetes esineda väikese katvusega sekveneerimisvigade tõttu. Duplitseerunud k-meeride keskmiste katvustete hinnanguteks võtame kahekordsed saadud unikaalsete k-meeride keskmiste katvuste hinnangud $2 \cdot \hat{\mu}_1$ ja $2 \cdot \hat{\mu}_2$ ning duplitseerunud k-meeride oskaalude hinnanguteks on $1 - \hat{\pi}_{01} - \hat{\pi}_1$ ja $1 - \hat{\pi}_{02} - \hat{\pi}_2$. Mõlema andmestiku korral leiame hinnangud pärast vaatluste tsenseerimist, kusjuures tsenseerimispunkti määrame bakteri andmestiku põhjal.

Tsenseerimispunkti määramisel kasutame kahte erinevat lähenemist, mis sõltuvad bakteri andmestiku kvantiilidest. Eesmärgiks on tsenseerida vaatlused nii, et väheneks pika parempoolse saba mõju unikaalsete k-meeride keskmise katvuse hinnangule. Kuna hindame ühe

keskmise katvuse ja teiseks keskmiseks katvuseks võtame kahekordse saadud hinnangu, kuid alates kolmekordsest keskmisest katvusest enam mudelisse ei kaasa, siis see oleks loogiline koht, kust alates vaatlused tsenseerida. Ligikaudu võiks selle tsenseerimispunkti määrata kahekordne mediaan. Teiseks võimaluseks oleks võtta mingi katvus, millest suuremate katvustega k-meere leidub andmestikus väike hulk, näiteks kuni 5% kõikidest vaatlustest. Mõlema lähenemise korral leiame mitmete erinevate punktide korral hinnangud, et näidata kui suurel määral sõltuvad hinnangud tsenseerimispunktist ja otsustada, milline oleks parim tsenseerimispunkt.

Pärast mõlema andmestiku põhjal eraldi kesmiste katvuste hindamist, hindame mõlemale andmestikule ühise unikaalsete k-meeride keskmise katvuse $\hat{\mu}_3$, kuid mõlema andmestiku jaoks eraldi unikaalsete k-meeride osakaalud $\hat{\pi}_3$ ja $\hat{\pi}_4$ ning puuduvate k-meeride osakaalud $\hat{\pi}_{03}$ ja $\hat{\pi}_{04}$. Ka sellel juhul võtame duplitseerunud k-meeride keskmise katvuse hinnanguks kahekordse saadud unikaalsete k-meeride keskmise katvuse hinnangu $2 \cdot \hat{\mu}_3$ ning duplitseerunud on esimeses andmestikus hinnanguliselt osakaaluga $1 - \hat{\pi}_{03} - \hat{\pi}_3$ ja teises andmestikus $1 - \hat{\pi}_{04} - \hat{\pi}_4$ k-meeridest.

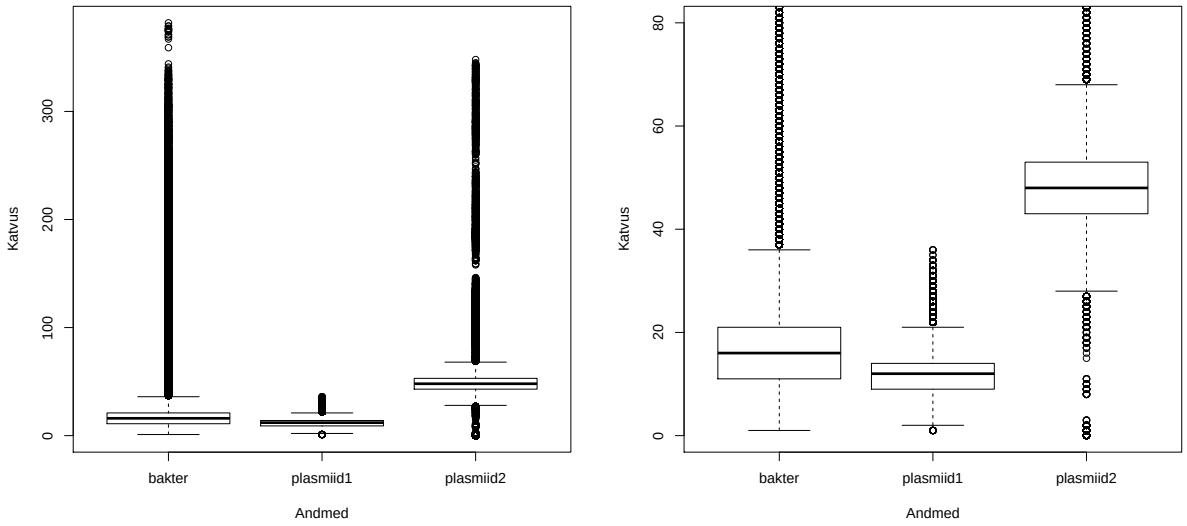
Mõlema mudeli korral arvutame log-tõepärad ning nende põhjal modifitseeritud teststatistiku väärtuse (vt valem 4). Lihtsama mudeli korral hindame 5 parameetrit ($\hat{\mu}_3$, $\hat{\pi}_{03}$, $\hat{\pi}_{04}$, $\hat{\pi}_3$, $\hat{\pi}_4$) ja keerulisema mudeli korral 6 parameetrit ($\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\pi}_{01}$, $\hat{\pi}_{02}$, $\hat{\pi}_1$, $\hat{\pi}_2$), mistõttu vabadusastmete arv tuleb 1. Vaatluse all olevate andmete korral on lugemi pikkus 100 ja k-meeri pikkus 25, mistõttu t väärtuseks võtame 76.

4.2 Andmete kirjeldus

Vaatame lähemalt kolme simuleeritud ning kolme reaalset andmestikku, mõlemal juhul ühe bakteri ning kahe plasmiidi andmeid. Simuleeritud andmestikud on genereeritud sekveneerimissimulaatori MetaSim abil. Selle simulatsiooni käigus luuakse suhteliselt realistlikud sekveneerimisandmed, mille korral on teada, kas antud juhul oli plasmiidi DNA integreeritud bakteri kromosomaalsesse DNAsse või mitte. Reaalsete andmete korral ei ole kunagi teada,

kas tegelikult plasmiid on integreerunud või mitte, mistõttu on hea alguses analüüsida simuleeritud andmeid, kus tõde on teada, kuid hiljem analüüsime ka reaalseid andmeid.

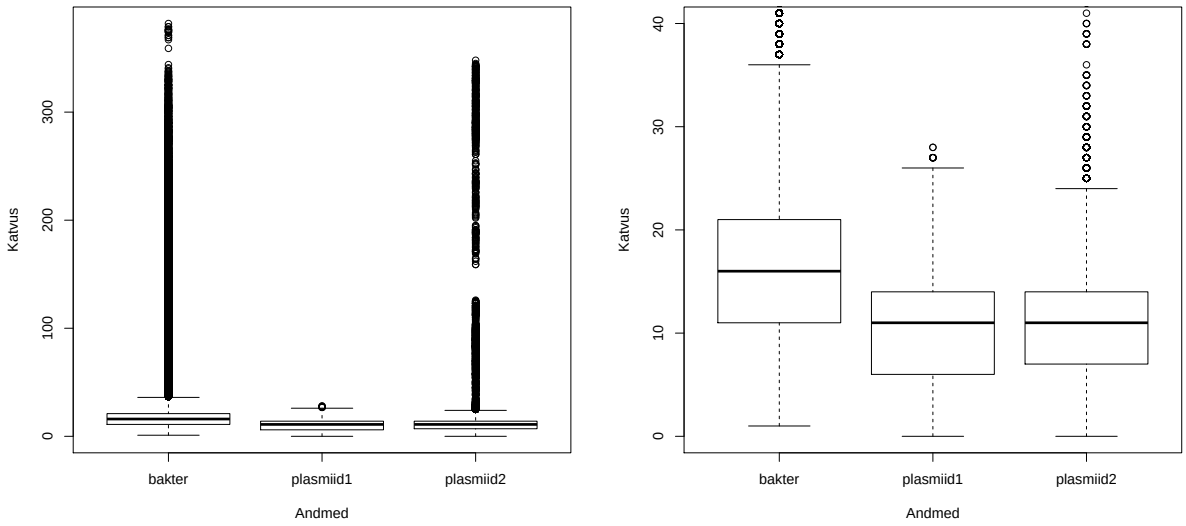
Joonisel 5 on kolme simuleeritud andmestiku karpdiagrammid, kus x-teljel on andmestik ning y-teljel k-meeri katvus. Vasakpoolsel joonisel on kujutatud kogu andmestikku ning parempoolsel joonisel on y-telge kujutatud kuni katvuseni 80, et oleks paremini näha erinevad karakteristikud. Maksimaalne katvus on bakteri andmestiku korral 382, ühe plasmiidi andmestiku korral 36 ning teise korral 348. Teise plasmiidi ja bakteri andmestiku korral on näha, et andmetel on päris pikk parempoolne saba. Mediaankatvus on bakteri andmete jaoks 16, esimese plasmiidi andmete jaoks 12 ja teise jaoks 48. Keskmised katvused on kõikide andmestike korral natukene mediaankatvustest suuremad, kuid mitte suurel määral. Karpdiagrammide põhjal tundub, et esimese plasmiidi ja bakteri k-meeride keskmised katvused võivad olla võrdsed, kuid teise plasmiidi ja bakteri korral need arvatavasti pole võrdsed. Kuna tegemist on simuleeritud andmetega, siis antud juhul on teada, et esimese plasmiidi DNA on integreeritud bakteri kromosomaalsesse DNAsse ning teise plasmiidi DNA ei ole.



Joonis 5. Simuleeritud andmete ülevaade

Joonisel 6 on toodud ühe bakteri ja kahe plasmiidi reaalsete andmete karpdiagrammid. Vasak-

poolsel joonisel on kujutatud kogu andmestikku ning parempoolsel joonisel on y-telge kujutatud kuni katvuseni 40. Võrreldes simuleeritud andmete ja reaalsete andmete karpdiagramme, on näha, et simuleeritud andmete jaotus on reaalsete andmete sarnane. Reaalsete andmestike korral on samuti teise plasmiidi ja bakteri k-meeride katvuste jaotusel pikk parempoolne saba. Bakteri andmestiku korral on maksimaalne katvus 382, ühe plasmiidi andmestiku korral 28 ning teise korral 348. Mediaankatvus on bakteri andmete jaoks 16 ning mõlema plasmiidi andmete jaoks 11. Keskmised katvused on bakteri ja teise plasmiidi andmete korral mediaankatvustest natuke suuremad, kuid esimese plasmiidi korral natuke väiksem. Nende andmete korral on karpdiagrammide põhjal raske tulemust ette aimata, kuid jääb mulje, et mõlema plasmiidi jaoks võib tulla sama tulemus.



Joonis 6. Reaalsete andmete ülevaade

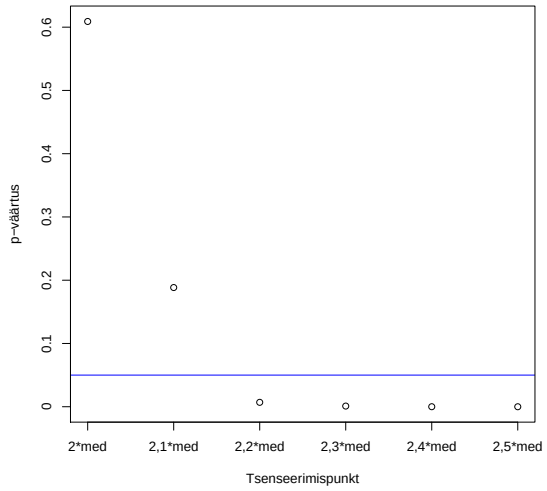
4.3 Simuleeritud andmete analüüsi tulemused

Alustuseks võtsime vaatluse alla bakteri ja plasmiidi andmestikud, mille korral unikaalsete k-meeride keskmine katvus peaks olema võrdne (simuleeritud kui integreerunud plasmiid). Kuigi keskmiste katvuste hinnangud tulid algusest peale ligilähedased, siis bakteri andmestiku tsenseerimata vaatluste põhjustatud pikal sabal oli mõju hinnangule, sest mudelisse olid

kaasatud ainult kuni kaks korda genoomis esinevad *k*-meerid, ja ilma vaatluseid tsenseerimata lükkas test nullhüpoteesi ümber ehk jõudis tulemuseni, et keskmised katvused ei ole võrdsed.

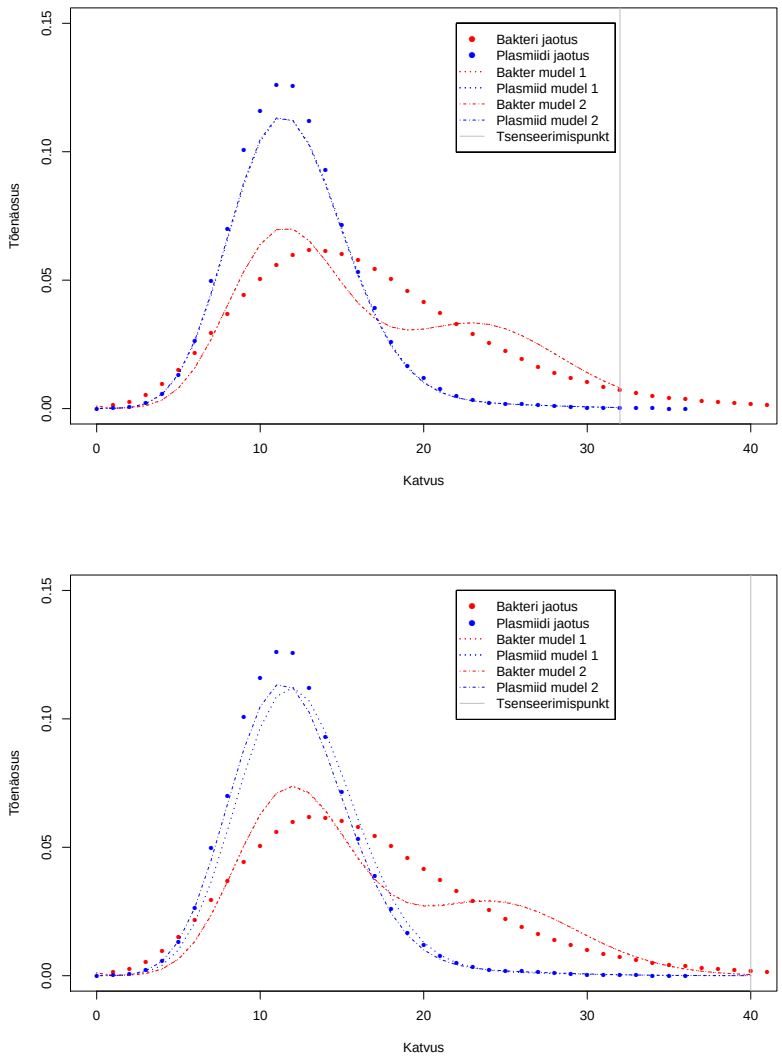
Üldiselt oli nende andmestike põhjal näha, et tsenseerimispunkti valik osutus väga oluliseks, sest kui liiga väike hulk vaatlusi tsenseerida, siis unikaalsete *k*-meeride keskmise katvuse hinnang oli sellest mõjutatud ja hinnatud keskmised katvused olid liiga erinevad, et test jääks nullhüpoteesi juurde. Kasutades tsenseerimiseks kas kahekordsest mediaankatvust või 0,95-kvantiili, tulid tsenseerimispunktid väga lähedased, mistõttu mõlemad meetodid viisid sarnaste tulemusteni.

Joonisel 7 on esitatud teststatistiku abil saadud unikaalsete *k*-meeride keskmiste katvuste võrdluse p-väärtused erinevate tsenseerimispunktide väärtuste korral, *x*-teljel on tsenseerimispunkti väärtus alates kahekordsest mediaanist kuni 2,5-kordse mediaanini ning *y*-teljel on sellisel viisil tsenseeritud andmestiku põhjal leitud p-väärtus. Jooniselt on näha, et valides tsenseerimispunktiks kahekordse mediaani, jääb test kindlalt nullhüpoteesi juurde. Kuna teame, et antud juhul tegelikult kehtib nullhüpotees, siis reaalsete andmete analüüsil kasutame tsenseerimispunktina kahekordset mediaani.



Joonis 7. Tsenseerimispunkti mõju testi p-väärtusele

Vaadates mõlema andmestiku jaotust jooniselt 8 on näha, et kahekordne mediaan paikneb kahe- ja kolmekordse hinnatud keskväärtuse vahel, mis on ka intuitiivselt loogiline punkt tsenseerimiseks. Joonisel 8 on kujutatud bakteri ja plasmiidi jaotused koos lihtsama mudeli (mudel 1) ja keerulisema mudeli (mudel 2) abil hinnatud jaotustega, x-teljel on toodud k-meeri katvus ning y-teljel tõenäosus. Ülemisel joonisel on tsenseerimispunktiks valitud bakteri k-meeride kahekordne mediaankatvus ning lihtsama ja keerulisema mudeli hinnatud jaotused on väga lähedased. Sellel juhul jääb test nullhüpoteesi juurde ja ei saa väita, et plasmiidi ja bakteri keskmised katvused oleksid statistiliselt oluliselt erinevad. Alumisel joonisel on kujutatud bakteri ja plasmiidi jaotusi koos mudelite hinnatud jaotustega juhul kui tsenseerimispunktiks on 2,5-kordne bakteri k-meeride katvuste mediaan. Sellisel juhul plasmiidi jaoks tulevad lihtsama ja keerulisema mudeli poolt hinnatud jaotused erinevad, kuid bakteri jaoks on need endiselt peaaegu võrdsed. Bakteri andmetel on tugev mõju ühisele keskmise katvuse hinnangule, sest bakteri DNA on pikem ning seetõttu on bakteri andmestikus rohkem k-meere, kui plasmiidi andmestikus.



Joonis 8. Bakteri ja esimese plasmiidi *k*-meeride katvuste jaotused koos hinnatud jaotustega (ülemisel joonisel tsenseerimispunktiks bakteri *k*-meeride kahekordne mediaankatvus, alumisel 2,5-kordne)

Võrreldes lihtsama ja keerulisema mudeli unikaalsete *k*-meeride jaoks saadud hinnanguid, mis on toodud tabelis 1, on näha, et keerulisema mudeli korral plasmiidi jaotuse parameetrite hinnangud ei muutu erinevate tsenseerimispunktide korral. Antud juhul plasmiidi *k*-meeride maksimaalne katvus on 36 ning bakteri andmete korral 382, mistõttu bakteri andmestiku abil tsenseerimispunkti määramisel plasmiidi andmestiku vaatlused jäävad suurel määral tsensee-

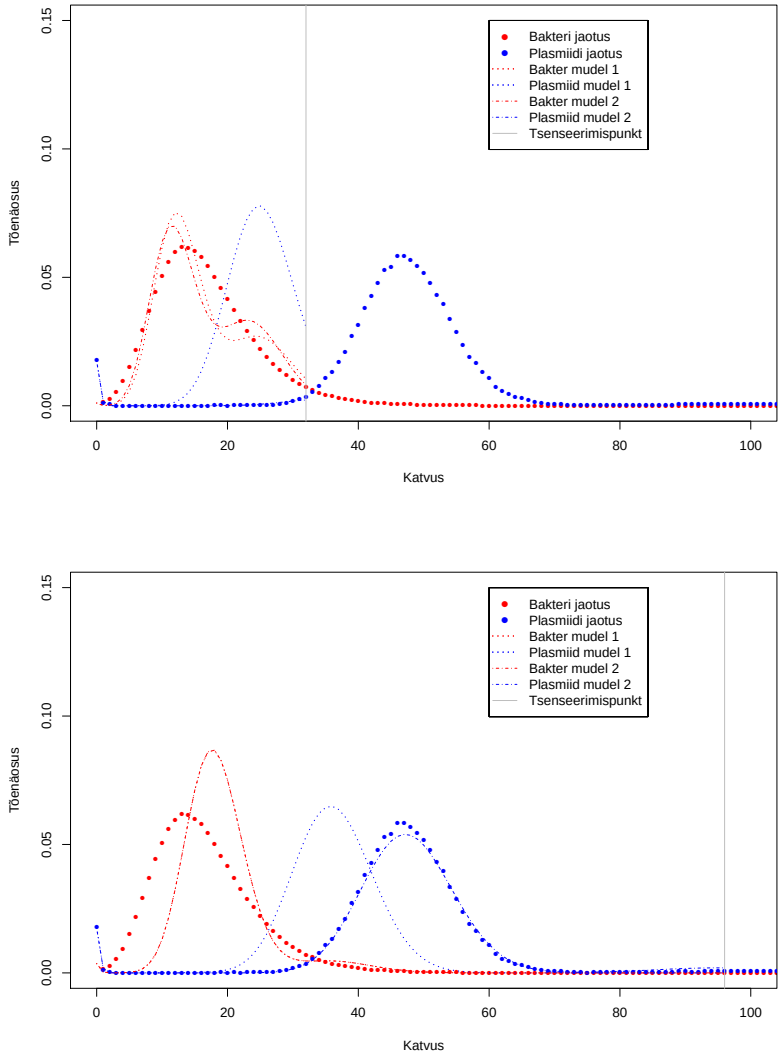
rimata. Tsenseerimispunkti valik mõjutab keerulisema mudeli bakteri jaotuse parameetreid ning samuti lihtsama mudeli parameetreid. Bakteri andmestiku pikk parempoolne saba kallutab lihtsama mudeli keskmise katvuse hinnangut piisavalt, et plasmiidi andmestiku jaoks tulevad keerulisema mudeli saadud parameetrid oluliselt täpsemad. See viib ka nullhüpoteesi kummutamiseni kui bakteri andmestiku vaatlused jätta tsenseerimata või tsenseerida liiga vähesed vaatlused.

Tabel 1. Hinnangud parameetritele

	Lihtsam mudel			Keerulisem mudel				
Tsenseerimispunkt	$\hat{\pi}_3$	$\hat{\pi}_4$	$\hat{\mu}_3$	$\hat{\pi}_1$	$\hat{\mu}_1$	$\hat{\pi}_2$	$\hat{\mu}_2$	p-väärtus
2*med	0,600	0,981	11,933	0,601	11,937	0,980	11,881	0,609
2,1*med	0,608	0,982	12,014	0,609	12,024	0,980	11,881	0,188
2,2*med	0,621	0,984	12,153	0,622	12,171	0,980	11,881	$7{,}0 \cdot 10^{-3}$
2,3*med	0,626	0,985	12,211	0,628	12,234	0,980	11,881	$1{,}1 \cdot 10^{-3}$
2,4*med	0,635	0,986	12,308	0,637	12,337	0,980	11,881	$2{,}1 \cdot 10^{-5}$
2,5*med	0,642	0,987	12,388	0,645	12,421	0,980	11,881	$4{,}7 \cdot 10^{-7}$

Teiseks võtsime vaatluse alla bakteri ja plasmiidi andmestikud, mille korral teame, et unikaalsete k-meeride keskmine katvus ei tohiks tulla võrdne. Joonisel 9 on kujutatud bakteri ja plasmiidi jaotused koos lihtsama mudeli (mudel 1) ja keerulisema mudeli (mudel 2) abil hinnatud jaotustega. Ülemisel joonisel on tsenseerimispunktiks valitud bakteri k-meeride kahekordne mediaankatvus. Kuna plasmiidi andmestikul on paljud katvustest sellest tsenseerimispunktist suuremad, siis plasmiidi andmestiku keskmised katvused tulevad oluliselt erinevad tegelikkusest. Alumisel joonisel on kujutatud bakteri ja plasmiidi jaotusi koos mudelite hinnatud jaotustega juhul kui tsenseerimispunktiks on plasmiidi k-meeride kahekordne mediaankatvus. Kuna bakteri andmestikul on sellisel juhul pikem parempoolne saba kui eelmise hinnangu korral, siis sellel juhul nihkub unikaalsete k-meeride keskmise katvuse hinnang natuke paremale. Ühise keskmise katvuse hinnang kattub bakteri andmestiku keskmise katvusega

keerulisema mudeli korral. Plasmiidi andmestiku unikaalsete *k*-meeride katvusi kirjeldab paremini kahekordne lihtsama mudeli abil saadud keskmise katvuse hinnang, kuid ka see satub kaugele keerulisema mudeli hinnangust. Mõlemal juhul lükkab test nullhüpoteesi kindlalt ümber ja kahe hinnatud keskmise katvusega mudel sobib paremini ehk bakteri ja plasmiidi keskmised katvused on statistiliselt oluliselt erinevad.

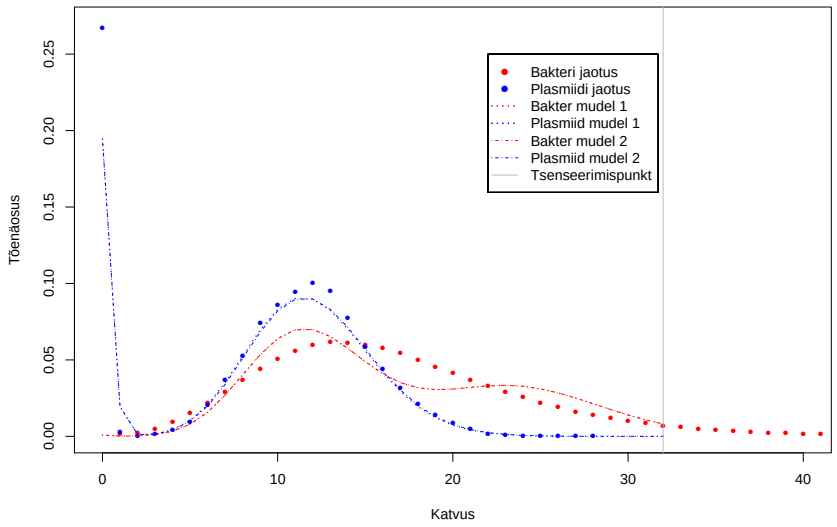


Joonis 9. Bakteri ja teise plasmiidi *k*-meeride katvuste jaotused koos hinnatud jaotustega (ülemisel joonisel tsenseerimispunktiks bakteri *k*-meeride kahekordne mediaankatvus, alumisel plasmiidi *k*-meeride kahekordne mediaankatvus)

4.4 Reaalsete andmete analüüsi tulemused

Simuleeritud andmete korral oli näha, et puuduvate *k*-meeride osakaal on peaaegu olematu, kuid reaalsete andmete korral on puuduvaid *k*-meere märgatavalt rohkem, eriti plasmiidide andmete korral. See on peamine erinevus simuleeritud ja reaalsete andmete jaotuste võrdlusel.

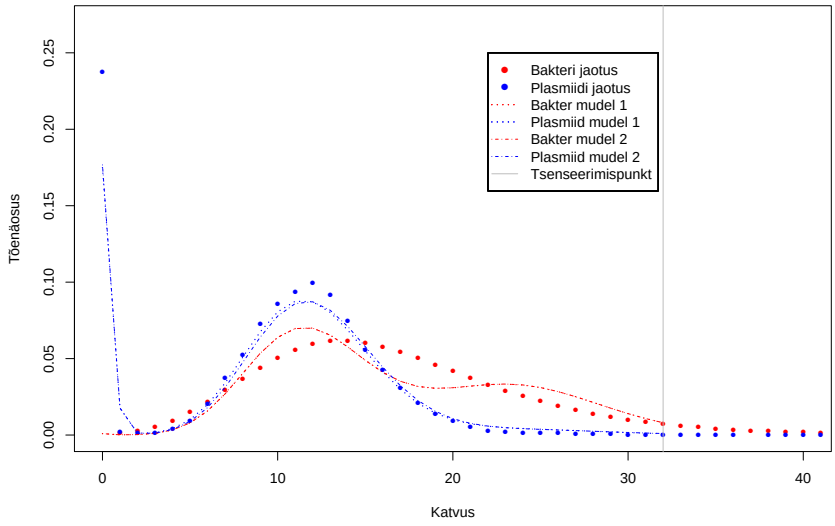
Joonisel 10 on kujutatud bakteri ja esimese plasmiidi jaotused koos lihtsama mudeli (mudel 1) ja keerulisema mudeli (mudel 2) abil hinnatud jaotustega. Nii bakteri kui ka plasmiidi andmete korral tulevad lihtsama ja keerulisema mudeli hinnangud sarnased ning test jääb nullhüpoteesi juurde ehk ei saa öelda, et bakteri ja plasmiidi unikaalsete *k*-meeride keskmised katvused oleksid statistiliselt oluliselt erinevad. Unikaalsete *k*-meeride keskmiste katvuste võrdlemisel saadud p-väärtus oli 0,55.



Joonis 10. Bakteri ja esimese plasmiidi *k*-meeride katvuste jaotused koos hinnatud jaotustega

Joonisel 11 on toodud bakteri ja teise plasmiidi jaotused koos lihtsama mudeli (mudel 1) ja keerulisema mudeli (mudel 2) abil hinnatud jaotustega. Bakteri andmete korral tulevad lihtsama ja keerulisema mudeli unikaalsete *k*-meeride keskmise katvuse hinnangud ligilähedased,

plasmiidi andmete korral tulevad mudelite abil saadud hinnangud natukene erinevad. Keerulisema mudeli hinnang plasmiidi andmete unikaalsete k-meeride keskmisele katvusele on täpsem, kuid võrreldes mõlema andmestiku jaoks saadud hinnanguid, oli testi p-väärtuseks 0,15, mistõttu ei saa öelda, et keerulisema mudeli hinnangud oleksid statistiliselt oluliselt paremad lihtsama mudeli hinnangutest. Seega jääme ka teise plasmiidi korral nullhüpoteesi juurde ja ei saa väita, et bakteri ja plasmiidi unikaalsete k-meeride keskmised katvused oleksid oluliselt erinevad ehk arvatavasti on mõlema plasmiidi korral tegemist integreerunud plasmiidiga.



Joonis 11. Bakteri ja teise plasmiidi k-meeride katvuste jaotused koos hinnatud jaotustega

5 Korduste arvu hindamine

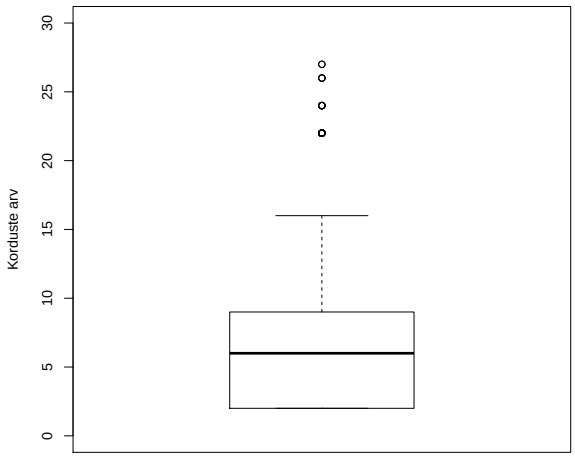
Korduvaks piirkonnaks nimetame lõiku genoomis, kus sama nukleotiidide järjestus kordub väga mitmeid kordi järjest. Korduvad järjestused on olulised piirkonnad inimese geneetika uurimisel. Korduste arv varieerub indiviidide lõikes ning mõjutab erinevate haiguste avaldumist, kuid korduste arvu määramine on keeruline. Näiteks sekveneerimise abil on raske korduvat piirkonda üheks järjestuseks kokku panna, sest sama lugem võib sobituda erinevatesse kohtadesse. Võrreldes indiviidi *k*-meeride katvuse andmeid referentsi *k*-meeride esinemiskorduste andmetega, tahame määrata, kui mitmes korduses vastaval indiviidil esineb korduvat piirkonda.

5.1 Andmete kirjeldus

Võtame vaatluse alla referentsandmestiku, mis sisaldab kahte tunnust: *k*-meer ja sellele vastav korduste arv, mitu korda antud *k*-meer referentsgenoomis esines. Lisaks vaatleme 43 erineva inimese andmeid, mis sisaldavad samuti kahte tunnust: *k*-meer ja mitu korda antud *k*-meeri esines lugemites ehk *k*-meeri katvus. Referents- ja indiviidide andmed sisaldavad samu *k*-meere, mis olid moodustatud referentsgenoomi põhjal, liikudes ühe nukleotiidi kaupa järjest mööda genoomi edasi. See tähendab, et kui esimene *k*-meer on ACCA...TG, siis järgmised on CCA...TGG, CA...TGGA jne. Referentsgenoomis ühes korduses esinevaid *k*-meere nimetame unikaalseteks, ülejäänud *k*-meere nimetame korduvateks. Eelnevalt kirjeldatud *k*-meeride moodustamise viisi tõttu esineb korduvaid *k*-meere andmestikes mitu korda. Nii referents- kui ka indiviidide andmete korral on teada, et lisaks huvipakkuvale korduste piirkonnale on sekveneeritud 1000 *k*-meeri enne ja pärast huvipakkuvat piirkonda, mistõttu korduvaid *k*-meere andmestikest eraldades, jätame vaatluse alt välja 1000 esimest ning 1000 viimast vaatlust.

Algselt on kõikides vaadeldavates andmestikes 3916 vaatlust, millest unikaalseid *k*-meere on 2461. Jättes kõrvale 1000 vaatlust algusest ning lõpust ja võttes igat *k*-meeri arvesse ühe korra, saame korduvaid *k*-meere 172. Joonisel 12 on toodud referentsgenoomi korral huvipakkuva piirkonna *k*-meeride esinemiskorduste karpdiagramm. Korduvate *k*-meeride minimaalne

korduste arv on 2 ning maksimaalne 27, kuid suurem osa kordusi jääb kuni 9 korduseni. Erinevate indiviidide korral on korduvate k-meeride katvuste jaotus erinev ning samuti on indiviidide lõikes erinev unikaalsete k-meeride katvuste jaotus.

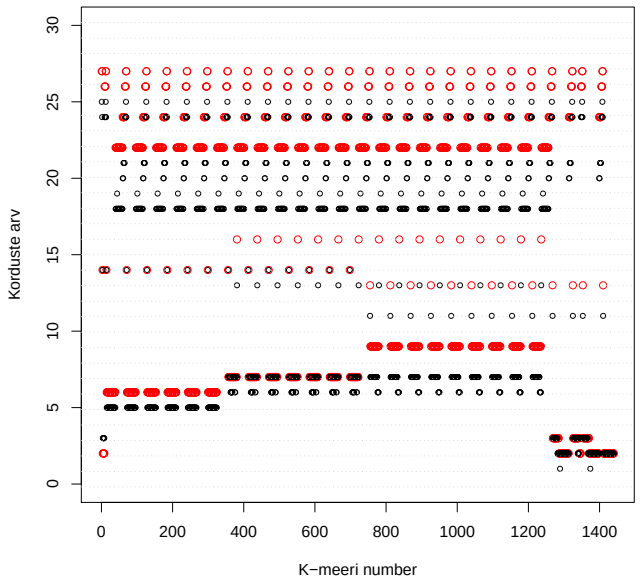


Joonis 12. Huvipakkuva piirkonna k-meeride esinemiskorduste arv referentsgenoomis

5.2 Analüüsi käik

Referentsandmestiku põhjal saame kindlaks määrata unikaalsed k-meerid ehk milliseid k-meere peaks genoomis esinema ühes korduses. Eraldades need k-meerid indiviidi andmestikus saame suurima tõepära meetodi abil hinnata unikaalsete k-meeride keskmise katvuse $\hat{\mu}$. Esialgsed hinnangud, mitmes korduses antud k-meer esines indiviidi genoomi huvipakkuvas piirkonnas, saame, jagades k-meeri katvuse unikaalsete k-meeride keskmise katvuse hinnanguga. Joonisel 13 on x-teljel k-meeri järjekorranumber ning y-teljel korduste arv, mustade punktidena on kujutatud ühe indiviidi korral saadud esialgsed korduste hinnangud ning punaste punktidena referentsgenoomi korduvate k-meeride esinemiskordused. Võrreldes referentsgenoomi k-meeride korduste arve leitud indiviidi k-meeride korduste arvude hinnangutega, on näha, et k-meeride kordused on võimalik jagada nelja põhigruppi: k-meerid, mis esinevad referentsgenoomis kordustega 3, 6, 7 ja 9 ning ülejäänud k-meeride kordused on

avaldatavad põhigruppide korduste kaudu, näiteks *k*-meer, mis esineb nii teises, kolmandas kui ka neljandas grupis, esineb referentsgenoomis 22 korda. Koos põhigruppidega on selliseid gruppe kokku 11. Mõned üksikud *k*-meerid esinevad referentsgenoomis väikeste korduste arvudega, mistõttu selleks, et minimiseerida hinnatavate parameetrite hulka, võtame nende *k*-meeride esinemiskorduse referentsgenoomist ning liidame vastavalt vajadusele vahepeal otsitavatele korduste arvudele juurde ühe või kaks kordust (vt valem 6).



Joonis 13. Korduvate *k*-meeride grupid

Järgnevalt hindame, mitu korda esimesse, teisse, kolmandasse ja neljandasse gruppi kuuluvaid *k*-meere esineb antud indiviidil ning ülejäänud gruppide korduste hinnangud saame summeerides kokku vastavad põhigruppide hinnangud. Parameetrite hinnangud leiame suurima tõepära meetodil, kusjuures tõepära maksimiseerimise juures võtame arvesse, et otsitavatel korduste hinnangutel saavad olla ainult täisarvulised või 0,5-ga lõppevad väärtused, sest ühte *k*-meeri saab esineda ainult kas mõlema vanema või ühe vanema poolt saadud kromosoomides. Tõepära arvutamisel võtame arvesse samaaegselt kõiki gruppe, kus iga grupi keskmiseks katvuseks on otsitava korduste arvu ja hinnatud unikaalsete *k*-meeride keskmise katvuse kor-

rutis. Log-tõepära arvutamiseks on kasutataud järgnevat valemit

$$\sum_{ref_i=3} \ln f(katvus_i, param_1 * \hat{\mu}) + \sum_{ref_i=6} \ln f(katvus_i, param_2 * \hat{\mu}) + \sum_{ref_i=7} \ln f(katvus_i, param_3 * \hat{\mu}) +$$
$$\sum_{ref_i=9} \ln f(katvus_i, param_4 * \hat{\mu}) + \sum_{ref_i=13} \ln f(katvus_i, (1 + param_1 + param_4) * \hat{\mu}) +$$
$$\sum_{ref_i=14} \ln f(katvus_i, (1 + param_2 + param_3) * \hat{\mu}) + \sum_{ref_i=16} \ln f(katvus_i, (param_3 + param_4) * \hat{\mu}) +$$
$$\sum_{ref_i=22} \ln f(katvus_i, (param_2 + param_3 + param_4) * \hat{\mu}) + \sum_{ref_i=24} \ln f(katvus_i, (2 + param_2 +$$
$$param_3 + param_4) * \hat{\mu}) + \sum_{ref_i=26} \ln f(katvus_i, (1 + param_1 + param_2 + param_3 + param_4) * \hat{\mu}) +$$
$$\sum_{ref_i=27} \ln f(katvus_i, (2 + param_1 + param_2 + param_3 + param_4) * \hat{\mu}), \quad (6)$$

kus $f(\cdot)$ on Poissoni jaotuse tõenäosusfunktsioon, $\hat{\mu}$ on unikaalsete k-meeride jaoks hinnatud keskmine katvus, $ref_i = s$ näitab, et summeeritud on üle selliste k-meeride katvuste, kus referentsgenoomi esinemiskordus on s ning *param* on vektor, mis sisaldab suurima tõepära meetodil hinnatavaid parameetreid.

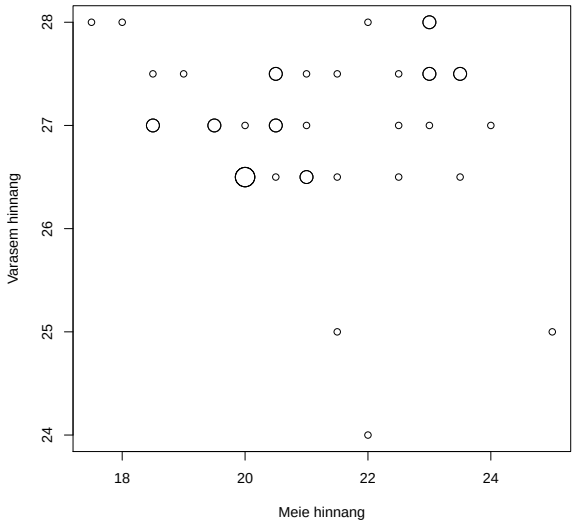
Teades hinnanguid kõikide põhigruppide korduste arvudele, on teada ka korduste hinnangud k-meeridele, mis esinevad mitmes grupis. Huvipakkuva piirkonna korduste arvu hinnanguks on selliste k-meeride grupi korduste arv, mis kuuluvad kõikidesse põhigruppidesse, ehk sellised k-meerid, mida esineb referentsgenoomis 27 korda.

5.3 Tulemused

Hindasime suurima tõepära meetodil kõigi 43 indiviidi korduvate k-meeride esinemiskordused, saadud hinnanguid võrdlesime varasemast teadaolevate hinnangutega. Võrdluseks on tabel, kus on iga indiviidi kohta kirjas, mitmes korduses esines korduvat piirkonda ema poolt saadud kromosoomis ja mitu korda isa poolt saadud kromosoomis. Seega selle tabeli kahe tunnuse keskmine peaks olema võrreldav meie saadud hinnanguga.

Võrreldes unikaalsete k-meeride keskmise katvuse hinnangu abil saadud esialgseid indivii-

di *k*-meeride esinemiskorduste hinnanguid suurima tõepärameetodi abil leitud hinnangutega, mis samuti sõltuvad hinnatud unikaalsete *k*-meeride keskmisest katvusest, tunduvad tulemused ootuspärased. Kuid siiski ei tule meie saadud hinnangud huvipakkuva piirkonna korduste jaoks võrreldavad varasemate hinnangutega. Joonisel 14 on *x*-teljel meie saadud hinnangud huvipakkuva piirkonna korduste arvule ning *y*-teljel varasemalt saadud hinnangud, punkti suurus on vastavuses selliste hinnangute paaride arvuga. Hajuvusdiagrammilt on näha, et meie hinnangud jäävad enamasti teistest hinnangutest väiksemaks, mõnel üksikul juhul tulevad ligilähedaselt sarnased.



Joonis 14. Kahe meetodi hinnangute võrdlus

Meie saadud hinnangute ning varasemate hinnangute vaheline korrelatsioonikordaja on -0,19, seega need hinnangud ei ole omavahel seotud. Kuigi meie saadud hinnangud ei tule varasemate hinnangutega võrreldavad, siis täielikku tõde nende andmete korral ei ole teada, mistõttu ei saa lõpliku kindlusega öelda, millised hinnangud on õiged.

Kokkuvõte

Käesolevas magistritöös hinnati sekveneerimisandmete põhjal, kas plasmiid on integreerunud bakteri kromosomaalsesse DNAsse või mitte. Selleks pakuti välja modifitseeritud tõepärasuhte teststatistik ning hinnati selle sobivust ja testi võimsust statistikapaketi R abil genereeritud simulatsioonide põhjal. Lisaks reaalsetele andmetele analüüsiti ka sekveneerimissimulaatori MetaSim poolt simuleeritud andmeid ning otsustati nende põhjal, milline oleks sobiv tsenseerimispunkt. Töö viimases peatükis hinnati, referentsgenoomi abil saadud unikaalsete k-meeride keskmise katvuse hinnangu kaudu, huvipakkuva piirkonna korduste arvu erinevate indiviidide jaoks.

Simulatsioonide põhjal ilmnes, et väljapakutud teststatistiku kohta püstitatud hüpotees on tõene ning testi võimsus on piisav. Sekveneerimissimulaatori abil simuleeritud andmete põhjal otsustati, et tsenseerimispunktiks on sobilik valida bakteri andmete kahekordne mediaankatvus. Selliselt valitud tsenseerimispunkti korral saadi modifitseeritud teststatistiku abil ootuspärased tulemused simuleeritud andmete jaoks. Reaalsete andmete korral ei olnud teada, kas antud plasmiidid olid integreerunud bakteri kromosomaalsesse DNAsse või mitte, kuid rakendades simulatsioonide põhjal toimivat testi, jõuti järeldusele, et mõlemad plasmiidid olid integreerunud. Indiviidide huvipakkuva piirkonna korduste arvu hinnangud ei tulnud eelnevalt teadaolevate hinnangutega lähedased, kuid täielikku tõde ei olnud ka nende andmete korral teada.

Kokkuvõtlikult võib öelda, et unikaalsete k-meeride keskmise katvuse hindamisel võib olla mitmeid erinevaid rakendusi, kuid reaalset situatsiooni teadmata, on vahepeal raske hinnata, kui hästi selle abil on võimalik tegelikkust kirjeldada. Töö käigus genereeritud andmete ning sekveneerimissimulaatori abil simuleeritud andmete põhjal võib eeldada, et võrreldes bakteri ja plasmiidi unikaalsete k-meeride keskmisi katvus, on võimalik välja selgitada, kas plasmiid on integreerunud või mitte. Kuid huvipakkuva genoomi piirkonna korduste arvu hindamiseks võiks proovida mingit teistsugust lähenemist, kui töös väljapakutud meetod.

Kasutatud allikad

1. Wikipedia, *DNA construct*. [URL] https://simple.wikipedia.org/wiki/DNA_construct (vaadatud: 29.03.2018)

2. Wikipedia, *Genomics*. [URL] https://en.wikipedia.org/wiki/Genomics (vaadatud: 28.03.2018)

3. Marchini, J., 2008. *Lecture 5 : The Poisson Distribution*. [PDF] http://www.stats.ox.ac.uk/~marchini/teaching/L5/L5.notes.pdf (vaadatud: 28.03.2018)

4. Wikipedia, *K-mer*. [URL] https://et.wikipedia.org/wiki/K-mer (vaadatud: 28.03.2018)

5. Couturier, M., Bex, F., Bergquist, P. L., Maas, W. K., 1988. *Identification and classification of bacterial plasmids*. [PDF] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC373151/pdf/microrev00046-0077.pdf (vaadatud: 28.03.2018)

6. Zhang, D., 2005. *Likelihood and Censored (or Truncated) Survival Data*. [PDF] https://www4.stat.ncsu.edu/~dzhang2/st745/chap3.pdf (vaadatud: 22.02.2018).

Lisa. Programmikoodid

Andmete genereerimine

```
N = 10000 #simulatsiooni pikkus
L=100 #lugemi pikkus
simulats <- function(Akv,osak) {
   rand.samples = rep(NA,N)
   Ad=rpois(N,Akv)
   Ad2=rpois(N,Akv)
   for (j in 1:(N/10)) {
      d2=j:(j-L+1)
      vektor=d2[d2>0]
      rand.samples[j]=sum(Ad[vektor])
   }
   #duplitseerunud k-meerid
   for (j in (N/10):(N/10+osak)) {
      d2=j:(j-L+1)
      vektor=d2[d2>0]
      rand.samples[j]=sum(Ad[vektor])+sum(Ad2[vektor])
   }
   for (j in (N/10+osak):N) {
      d2=j:(j-L+1)
      vektor=d2[d2>0]
      rand.samples[j]=sum(Ad[vektor])
   }
   #ühe simulatsiooni vektor
   return(rand.samples)
}
```

Poissoni jaotuste segu tõenäosusfunktsioon

```
f_pois=function(x, piis, lambdas) {
   return (sum(piis*dpois(x,lambdas)))
}
```

Log-tõepära keerulisema mudeli jaoks

```
l_pois <- function(param, andmed, C) {
   #C - tsenseerimispunkt
   #parameetrite teisendused, et püsiksid vajalikes piirides
   param[1] <- exp(param[1])/(1+exp(param[1])) #puuduvate k-meeride osakaal
   param[2] <- exp(param[2])/(1+exp(param[2])) #unikaalsete k-meeride osakaal
   param[3] <- exp(param[3]) #unikaalsete k-meeride keskmine katvus
   if (1-param[1]-param[2]<0) {
      return(-Inf)
   }
   piis <- c(param[1], param[2], 1-param[1]-param[2])
   lambdas <- c(0.1,param[3],2*param[3])
   else{
      return(-(sum(andmed$n[andmed$katvus<=C]*log(sapply(andmed$katvus
      [andmed$katvus<=C],f_pois,piis,lambdas))))-(sum(andmed$n[andmed$katvus>C])*
      log(sum(piis*(1-ppois(C,lambdas))))))
   }
}
```

Keerulisema mudeli parameetrite hindamine

```
optimeerimine <- function(alg1, alg2, alg3, andmed, C) {
   param=optim(par=c(alg1,alg2,alg3), fn=l_pois, andmed=andmed, C=C)
   #parameetrite ülehindamine
   param=optim(par=c(param$par[1],param$par[2],param$par[3]), fn=l_pois,
       andmed=andmed, C=C)
   param=optim(par=c(param$par[1],param$par[2],param$par[3]), fn=l_pois,
```

```
      andmed=andmed, C=C)
  return(param)
}
```

Parameetrite tagasiteisendamine

```
teisendus <- function(param) {
  param$par[1] <- exp(param$par[1])/(1+exp(param$par[1]))
  param$par[2] <- exp(param$par[2])/(1+exp(param$par[2]))
  param$par[3] <- exp(param$par[3])
  return(param)
}
```

Log-tõepära lihtsama mudeli jaoks

```
l_pois2 <- function(param, q1, q2, C) {
  #q1 ja q2 on andmestikud
  param[1] <- exp(param[1])/(1+exp(param[1]))
  param[2] <- exp(param[2])/(1+exp(param[2]))
  param[3] <- exp(param[3])/(1+exp(param[3]))
  param[4] <- exp(param[4])/(1+exp(param[4]))
  param[5] <- exp(param[5])
  if (1-param[1]-param[3]<0 || 1-param[2]-param[4]<0) {
    return(-Inf)
  }
  else{
    piis <- c(param[1], param[2], param[3], param[4], 1-param[1]-param[3],
        1-param[2]-param[4])
    lambdas <- c(0.1, param[5], 2*param[5])
    return(-(sum(q1$n[q1$katvus<=C]*log(sapply(q1$katvus[q1$katvus<=C],f_pois,
        piis[c(1,3,5)],lambdas)))+sum(q2$n[q2$katvus<=C]*log(sapply(q2$katvus[
        q2$katvus<=C],f_pois,piis[c(2,4,6)],lambdas))))-(sum(q1$n[q1$katvus>C])*
```

```
        log(sum(piis[c(1,3,5)]*(1-ppois(C,lambdas))))+sum(q2$n[q2$katvus>C])*
        log(sum(piis[c(2,4,6)]*(1-ppois(C,lambdas))))))
  }
}
```

Lihtsama mudeli parameetrite hindamine

```
lihtsam_mudel <- function(andmed1,andmed2,param1,param2,C) {
  #algväärtusteks ühe osakaalud, teise osakaalud ja bakteri keskmine
  param=optim(par=c(param1$par[1],param2$par[1],param1$par[2],
     param2$par[2],param1$par[3]),fn=l_pois2, q1=andmed1, q2=andmed2, C=C)
  #parameetrite ülehindamine
  param=optim(par=c(param$par[1],param$par[2],param$par[3],param$par[4],
     param$par[5]), fn=l_pois2, q1=andmed1, q2=andmed2, C=C)
  param=optim(par=c(param$par[1],param$par[2],param$par[3],param$par[4],
     param$par[5]), fn=l_pois2, q1=andmed1, q2=andmed2, C=C)
  #parameetrite tagasiteisendused
  param$par[1] <- exp(param$par[1])/(1+exp(param$par[1]))
  param$par[2] <- exp(param$par[2])/(1+exp(param$par[2]))
  param$par[3] <- exp(param$par[3])/(1+exp(param$par[3]))
  param$par[4] <- exp(param$par[4])/(1+exp(param$par[4]))
  param$par[5] <- exp(param$par[5])
  return(param)
}
```

Modifitseeritud tõepärasuhtetest

```
toepara_test <- function(param, param1, param2, L=100, k=25) {
  log_lik1=-param$value #lihtsama mudeli tõepära
  log_lik2=-param1$value-param2$value #keerulisema mudeli tõepära
  #tõepärasuhte teststatistik
  D=2*(log_lik2/(L-k+1)-log_lik1/(L-k+1))
```

```
  df=6-5 #vabadusastmete arv
  return(list(D,1-pchisq(D,df)))
}
```

Korduste hindamisel kasutatud log-tõepära

```
l_pois3 <- function(param, andmed=data1_kord2) {
     return(sum(log(sapply(andmed$V2[andmed$ref=="6"], dpois, param[1]*
        param1_t$par)))+(sum(log(sapply(andmed$V2[andmed$ref=="7"],dpois,
        param[2]*param1_t$par))))+(sum(log(sapply(andmed$V2[andmed$ref=="9"],
        dpois,param[3]*param1_t$par))))+(sum(log(sapply(andmed$V2[andmed$ref=="3"],
        dpois,param[4]*param1_t$par))))+(sum(log(sapply(andmed$V2[andmed$ref=="14"],
        dpois,(1+param[1]+param[2])*param1_t$par))))+(sum(log(sapply(andmed$V2[
        andmed$ref=="13"],dpois,(1+param[3]+param[4])*param1_t$par))))+(sum(log(
        sapply(andmed$V2[andmed$ref=="16"],dpois,(param[2]+param[3])*
        param1_t$par))))+(sum(log(sapply(andmed$V2[andmed$ref=="22"],dpois,
        (param[1]+param[2]+param[3])*param1_t$par))))+(sum(log(sapply(andmed$V2[
        andmed$ref=="24"],dpois,(2+param[1]+param[2]+param[3])*param1_t$par))))+
        (sum(log(sapply(andmed$V2[andmed$ref=="26"],dpois,(1+param[1]+param[2]+
        param[3]+param[4])*param1_t$par))))+(sum(log(sapply(andmed$V2[
        andmed$ref=="27"],dpois,(2+param[1]+param[2]+param[3]+param[4])*
        param1_t$par)))))
  }
```

Suurima tõepära meetodil korduste arvu hindamine

```
max_l <- -Inf
for (i1 in seq(1,10,0.5)) {
  for (i2 in seq(0,11,0.5)) {
    for (i3 in seq(4,17,0.5)) {
      for (i4 in seq(1,7,0.5)) {
        log_tp <- l_pois3(param=c(i1,i2,i3,i4))
```

```
            if (log_tp > max_l) {
               max_l <- log_tp
               max_i1 <- i1
               max_i2 <- i2
               max_i3 <- i3
               max_i4 <- i4
}}}}}
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kristel Luik,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Unikaalsete k-meeride keskmise katvuse hindamine ja saadud hinnangu rakendamisnäiteid", mille juhendaja on Märt Möls,
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, 15.05.2018