Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kätlin Kippar

Vigadega sekveneerimisandmete analüüs. Sekveneerimiskatvuse hindamine

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: PhD. Märt Möls

Tartu 2024

Vigadega sekveneerimisandmete analüüs. Sekveneerimiskatvuse hindamine

Bakalaureusetöö

Kätlin Kippar

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on võrrelda aritmeetilise keskmise ja Poissoni segumudeli täpsust keskmise sekveneerimiskatvuse hindamisel erindite olemasolu korral. Simuleeritakse erindeid sisaldavaid andmed, leitakse mõlema meetodi hinnangud ning otsitakse erindite osakaalu, millest alates üks meetod teisest täpsemaks muutub. Töös antakse ülevaade sekveneerimisest ning sellega kaasnevatest bioloogiamõistetest, kirjeldadakse vajalikku metoodikat ning saadud tulemusi. Tulemuseks leiti 9 meetodite lõikumispunkti erinditele vastava parameetri erinevate väärtuste korral. Tulemused näitavad enamikul juhtudest Poissoni segumudeli kasutamise olulisust erindite osakaalu kasvades.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika; B220 Geneetika, tsütogeneetika.

Märksõnad: Genoom, sekveneerimine, lugem, sekveneerimiskatvus, segujaotus, suurima tõepära hinnang, Poissoni segumudel, tõepärasuhte test.

Analysis of sequencing data that contains inconsistencies. Estimating the sequencing coverage

Bachelor's thesis

Kätlin Kippar

Abstract

The aim of this thesis is to compare the accuracy of the arithmetic mean and the Poisson mixture model in estimating the average sequencing coverage, given inconsistencies in the data. Data with inconsistencies is simulated, and both methods are applied to the simulated data to estimate the average sequencing coverage in order to identify the percentage of inconsistencies at which one method surpasses the other in accuracy. The thesis gives a brief overview of sequencing, including relevant biological terminology, then describes the used methodology and results. As a result of this thesis, 9 intersection points are identified given different parameter values that correspond to the inconsistencies. The results show that in most cases, using the Poisson mixture model is important as the percentage of data inconsistencies increases.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics; B220 Genetics, cytogenetics.

Key words: Genome, sequencing, read, sequencing coverage, mixture distribution, maximum likelihood method, Poisson mixture model, likelihood ratio test.

Sisukord

Sissejuhatus

Esineb olukordi, kus kogutud andmed sisaldavad ka soovimatuid vaatluseid. Näiteks vaatluseid, mis pärinevad kohtadest, mis uuritavasse populatsiooni ei kuulu. Sekveneerimisandmete puhul võivad segadust tekitada genoomi ümberkorraldused, näiteks piirkondade duplikatsioonid, mille toimumisest uurija ei ole teadlik. Genoomis unikaalsetes piirkondades tehtud vaatlustele võivad lisanduda mõõtmistulemused mitteunikaalsetest piirkondadest, kuna ei olda teadlik vahepeal toimunud mutatsioonidest. Sellistel juhtudel koosnevad saadud andmed tegelikkusele vastavatest andmetest ja mingi oskaaluga vigastest andmetest, mida ideaaljuhul soovitakse eemaldada enne andmete analüüsi. Vahel on võimalik kirjeldada nii pärisandmete kui ka vigade jaotuseid parameetriliste mudelite abil - teatakse jaotusi, aga mitte nende jaotuse parameetrite täpseid väärtuseid. Sellistel juhtudel saab jaotuse parameetreid hinnata kahel viisil, võttes ja võtmata arvesse vigade olemasolu andmetes.

Käesoleva töö puhul on nii vigade kui ka tegelikkusele vastavate andmete statistilisteks jaotusteks Poissoni jaotus. Jaotuse parameeter on aga mõlemal puhul erinev, ning hinnata soovitakse tegelikkusele vastavate andmete Poissoni jaotuse parameetrit ehk korrektsete vaatluste tegelikku keskväärtust. Kui erindid puuduvad, siis saab huvi pakkuvat parameetrit hinnata aritmeetilise keskmise abil. Arvestades erinditega, võib andmete jaotust modelleerida Poissoni jaotuste segu ehk segumudeli abil. Bakalaureusetöö eesmärk on võrrelda aritmeetilise keskmise ja segumudeli täpsust tegeliku keskväärtuse hinnangu leidmisel olenevalt vigade osakaalust.

Antud töö on jagatud neljaks osaks. Töö esimeses peatükis antakse ülevaade töö bioloogiaalastest taustast ning selgitatakse töö vajalikkust ja eesmärki. Teises peatükis on välja toodud töös kasutatavad statistilised jaotused ja

metoodika. Kolmandas peatükis on kirjeldatud töö eesmärki teoreetilisemast küljest ning seletatud sellest lähtuvalt töö käiku. Lisaks on kolmandas peatükis välja toodud keskmise ruutvea arvutamine ning ruutvigade võrdluseks kasutatava stohhastilisel optimeerimisel põhineva meetodi erinevust tavalistest numbrilise optimeerimise meetoditest. Samuti on välja toodud tõepära suhte testi iseloomustus. Töö viimases, neljandas peatükis on esitatud töö käigus saadud tulemused.

Antud töö on tehtud kasutades rakendustarkvara R, mille peamised programmikoodid on lisatud töö lõppu. Töös olevate jooniste tegemiseks on rakendustarkvarale R lisaks kasutatud MetaPost-keelt. Töö on kirjutatud tekstitöötlusprogrammiga LaTeX.

Käesolevaga tänab autor bakalaureusetöö juhendajat Märt Mölsi väärtuslike nõuannete ja paranduste, selgituste ja näidete, ning abi ja toetuse eest terve töö vältel.

1 Probleemi kirjeldus

DNA on organismi pärilikku informatsiooni sisaldav materjal, juhend organismi talitluseks, mis näeb ülesehituselt välja kui tähejärjestus neljast nukleotiidist: adeniin (A), tsütosiin (C), guaniin (G) ja tümiin (T). DNA ahelas moodustab adeniin paari tümiiniga ja tsütosiin guaniiniga. Moodustunud paare kutsutakse aluspaarideks. (Brown, 2022) Näiteks võib üks DNA fragment välja näha kui AGCCTACG, millele on vastavuses nukleotiidide järjestus TCGGATGC, ning aluspaarideks oleksid sel juhul A-T, G-C, C-G jne. Inimese DNA on jaotatud kahe rakukomponendi vahel. Mingi osa sellest paikneb mitokondris, kuigi peamiselt asub DNA siiski rakutuumas. Kogu pärilikku informatsiooni, nii rakutuumas kui ka mitokondris asuvat DNA-d, sisaldab endas genoom. (Brown, 2022) Hinnanguliselt on inimese genoom moodustatud ligikaudu kolmest miljardist aluspaarist. Selleks, et kindlaks teha DNA-s asuvate nukleotiidide järjestus, kas mingis teatud DNA fragmendis või terves genoomis, kasutatakse sekveneerimist. (NHGRI, 2020)

Sekveneerimine on protsess, mille käigus loetakse juhuslikest alguspunktidest lühikesi piirkondi genoomist. Selliseid sekveneerimistulemusi nimetatakse lugemiteks. (Chauhan, 2022) Pärast sekveneerimist kontrollitakse saadud tulemusi ning seejärel sageli sobitatakse lugemid taas kokku üheks fragmendiks või genoomiks (Venter *et al.*, 2001). Lugemi pikkus sõltub sekveneerimismeetodist. Uuemad sekveneerimismeetodid ehk kolmanda põlvkonna sekveneerimismeetodid ehk ühe molekuli sekveneerimismeetodid (SMRT) lubavad lugeda korraga pikemaid piirkondi. Nende puhul on võimalik saavutada lugemi pikkuseks 500 000 kuni 2 300 000 aluspaari, mis on võrreldes eelmiste põlvkondadega, mille pikimad lugemid jäid 500 aluspaari juurde, märgatavalt parem tulemus. Kuid endiselt pole suudetud kõrvaldada pikemate lugemite

leidmisega kaasnevaid vigu, mis mõjutavad olulisel määral meetodi täpsust. (Burian *et al.*, 2021) Lisaks geneetilise informatsiooni uurimisele, võimaldab sekveneerimine kindlaks teha muutusi inimese genoomis, mis võivad põhjustada haigusi (NHGRI, 2020).

Sekveneerimise põhjalikkust hinnatakse katvuse järgi. Katvuse all mõistetakse seda, mitu korda ühte ja sama piirkonda on loetud. Mida suurem on katvus, seda kindlamalt saab väita, et selle piirkonna sekveneerimisel pole esinenud vigu. (Illumina, 2014) Näiteks, kui mustrit AGCCTACG loetakse 29 korda, siis võib eeldada, et tõepoolest selles piirkonnas pole esimese adeniini (A) asemel hoopis guaniin (G). Hetke standardite järgi soovitatakse inimgenoomi sekveneerimisel piisavalt hea täpsuse tagamiseks saavutada 30-50 kordne katvus (Illumina, 2024). Keskmist katvust saab arvutada järgnevalt:

$$C = \frac{L \cdot N}{G},$$

kus

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

Teades keskmist katvust, on võimalik välja arvutada, näiteks kui tõenäoline on, et mingit piirkonda katab soovitud arv lugemeid, mille järgi saab hinnata, kas saavutatud katvus vastab katse vajadusele. (Illumina, 2014) Probleem tekib aga siis, kui soovitakse sekveneerida genoomi, mille pikkus pole teada. Kuidas saab sellisel juhul hinnata keskmist katvust? Või kui soovitakse leida vastuseid ka siis, kui lugemite tagasi kokku sobitamine osutub liialt keeruliseks või isegi võimatuks. Näiteks on tsentromeerid ülesehituselt kui samade

või väga väikeste muudatustega DNA fragmentide kordused, mistõttu on nende piirkondade tagasi kokku sobitamine osutunud senimaani võimatuks (Miga *et al.*, 2014). Inimgenoomi teadaolev pikkus (3 miljardit aluspaari) on vaid ligikaudne hinnang, mis võib erineda veel omakorda inimesiti (NHGRI, 2020; Piovesan *et al.*, 2019). Vahel on teada teatud piirkondade sekveneerimiskatvused ehk mitu lugemit antud genoomi piirkonda katab. Üks võimalus keskmise sekveneerimiskatvuse hindamiseks on leida selliste kohtade katvuste aritmeetiline keskmine. Kuid teadaolevalt mõjutavad erindid ehk ekslikult liiga suured või väikesed vaatlused aritmeetilist keskmist väga tugevalt. Seetõttu, kui piirkonna katvuse leidmisel on tekkinud vigu ning teatud piirkondade katvus on tegelikkusest tunduvalt suurem või väiksem, oleks leitud keskmine katvus vale. Katvuse määramiseks on kindlaks tehtud genoomi piirkonnad, milles esinev nukleotiidide järjestus on unikaalne. Kuid teadmata põhjustel võib ette tulla olukordi, kus seda järjestust ehk mustrit esineb genoomis mitmes erinevas kohas. Selliseid olukordi võib esineda mitmel põhjusel, näiteks võib probleem esineda proovi saastusest (proovi sattunud bakteri genoomis esineb sama muster) või võib see olla tingitud genoomis toimunud mutatsioonidest, mille toimumistest ei olda teadlikud (Koboldt *et al.*, 2010). Näiteks pole tuvastatud kõiki ägeda müeloidleukeemia põhjustatud mutatsioone inimgenoomis (Mardis *et al.*, 2009). Sellistel juhtudel, kus unikaalseks peetud muster esineb genoomis kahes või enamas kohas, loetakse selle piirkonna katvuseks vastavalt kahe või enama piirkonna katvuste summa ning seetõttu on saadud katvus ekslikult kõrge.

Teine viis hinnata keskmist katvust on modelleerida seda kasutades segujaotust, ning seeläbi võtta hinnangu leidmisel arvesse ka erinditest põhjustatud vigu ja nende osakaalu.

Antud töö eesmärk on hinnata vigade osakaalu, millest alates muutub üks

keskmise katvuse hindamise meetod teisest täpsemaks, et kindlaks teha, millal tuleks rakendada keskmise katvuse hindamisel viimasena mainitud keerukamat meetodit.

Hetkeseisuga on sekveneerimistehnoloogiate turu liider Illumina (Seeking Alpha, 2023), mistõttu põhinevad antud töö mõned algteadmised Illumina tehnilistes andmetes välja toodud teadmistel.

2 Kasutatud metoodika

Käesolevas peatükis on toodud ülevaade statistilistest jaotustest ja meetoditest, mida kasutatakse antud töö vältel.

2.1 Poissoni jaotus

Poissoni jaotus, teisiti kutsutud ka kui harvaesinevate sünduste jaotus, on diskreetne tõenäosusjaotus. Teades sündmuse toimumise keskmist sagedust, saab Poissoni jaotuse abil kujutada selle sündmuse võimalike väärtuste toimumise sagedust. (Tiit, Parring ja Möls, 1977) „Poissoni jaotus sobib hästi mitmesuguste looduses ja tehnikas esinevate protsesside kirjeldamiseks“ (Tiit, Parring ja Möls, 1977, 71). Kui juhuslik suurus X on Poissoni jaotusega, siis tema tõenäosusfunktsioon avaldub järgmiselt:

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

kus parameeter $\lambda > 0$ ja $k = 0,1,2,\ldots$. Kusjuures juhusliku suuruse X keskväärtus (EX) ja dispersioon (DX) on võrdsed parameetriga λ ehk

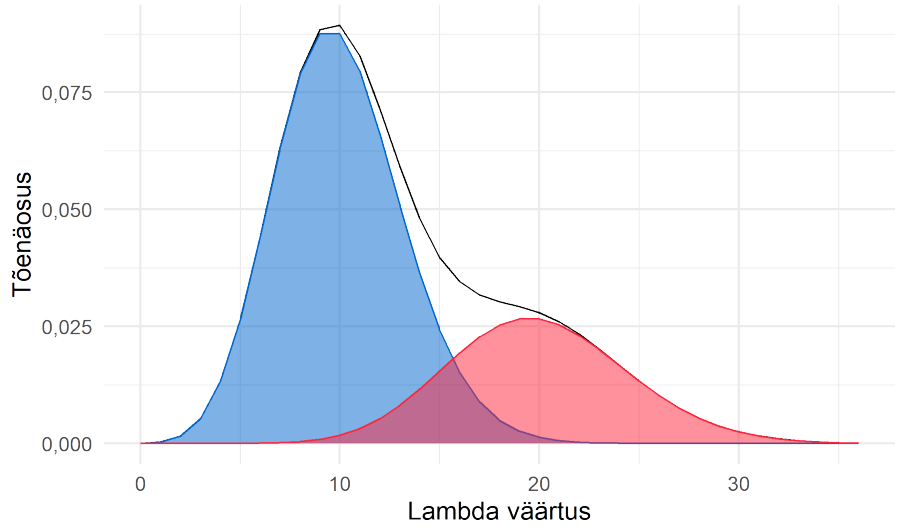
$$EX = DX = \lambda.$$

(Tiit, Parring ja Möls, 1977)

Ühe genoomi piirkonna sekveneerimiskatvus on Poissoni jaotusega juhuslik suurus, kus Poissoni jaotuse parameeter λ on keskmine sekveneerimiskatvus (Illumina, 2014).

2.2 Segujaotus

Segujaotus on jaotus, mis koosneb kahe või enama jaotuse segust. Segujaotused võivad olla kombinatsioonid erinevatest jaotustest või erinevate parameetritega samast jaotusest. Iga segujaotuse moodustamisel osaleva jaotuse kohta on määratud ka selle jaotuse osakaal ehk millise tõenäosusega on segujaotuse vaatlus pärit talle vastavast jaotusest. (Titterington, Smith ja Makov, 1985) Näiteks, olgu $h(x_i; \alpha_1, \ldots, \alpha_m)$ ja $g(y_i; \beta_1, \ldots, \beta_n)$ kaks erinevat tõenäosusjaotust, kus $(\alpha_1, \ldots, \alpha_m)$ ja $(\beta_1, \ldots, \beta_n)$ on vastavate jaotuste parameetrid ning x_i ja y_i tähistavad jaotustel põhinevaid valimeid. Valides tõenäosusega a jaotusega h juhusliku suuruse väärtusi ja tõenäosusega $1 - a$ jaotusega g juhusliku suuruse väärtusi, on tulemuseks kokku nende kahe jaotuste segu. Alloleval joonisel (joonis 1) on toodud lisaks näide kahe erineva parameetriga, $\lambda_1 = 10$ ja $\lambda_2 = 20$ Poissoni jaotuse segust, kus tõenäosus $a = 0{,}3$.



Joonis 1: Segujaotus.

Kahe Poissoni jaotusest moodustatud segujaotuse korral on teada, et esi-

mese jaotuse keskväärtus ja dispersioon on võrdsed Poissoni jaotuse parameetriga λ_1 ja teise Poissoni jaotuse keskväärtus ja dispersioon on võrdsed parameetriga λ_2. Olgu X segujaotusega juhuslik suurus ja näidaku Y kummast Poissoni jaotusest on konkreetne vaatlus pärit, siis $X|Y = 1 \sim \text{Poi}(\lambda_1)$ ja $X|Y = 2 \sim \text{Poi}(\lambda_2)$ ning $E(X|Y = 1) = D(X|Y = 1) = \lambda_1$ ja $E(X|Y = 2) = D(X|Y = 2) = \lambda_2$. Kasutades neid teadmisi, saab kirja panna segujaotuse keskväärtuse ja dispersiooni. Kui parameetriga λ_1 juhuslike suuruseid valitakse tõenäosusega $(1 - a)$, siis segujaotuse keskväärtus on

$$EX = E(E(X|Y)) = (1 - a)\lambda_1 + a\lambda_2. \tag{1}$$

Segujaotuse dispersioon on aga leitav valemi

$$DX = ED(X|Y) + DE(X|Y)$$

järgi, kus

$$ED(X|Y) = (1 - a)\lambda_1 + a\lambda_2$$

ja dispersiooni valemi $DX = E(X^2) - (EX)^2$ järgi on

$$\begin{aligned}
DE(X|Y) &= E\left(E(X|Y)^2\right) - \left[E\left(E(X|Y)\right)\right]^2 = \\
&= (1 - a)\lambda_1^2 + a\lambda_2^2 - \left((1 - a)^2\lambda_1^2 + 2a(1 - a)\lambda_1\lambda_2 + a^2\lambda_2^2\right) = \\
&= (1 - a)\lambda_1^2 - (1 - a)^2\lambda_1^2 - 2a(1 - a)\lambda_1\lambda_2 + a\lambda_2^2 - a^2\lambda_2^2 = \\
&= \lambda_1^2(1 - a)(1 - (1 - a)) - 2a(1 - a)\lambda_1\lambda_2 + a\lambda_2^2(1 - a) = \\
&= a(1 - a)(\lambda_1^2 - 2\lambda_1\lambda_2 + \lambda_2^2) = \\
&= a(1 - a)(\lambda_1 - \lambda_2)^2.
\end{aligned}$$

Järelikult avaldub kahe Poissoni jaotuse segujaotuse dispersioon järgnevalt

$$DX = ((1-a)\lambda_1 + a\lambda_2) + \left(a(1-a)(\lambda_1 - \lambda_2)^2\right). \tag{2}$$

Kui nii korrektselt loetud piirkondade sekveneerimiskatvus kui ka nende piirkondade sekveneerimiskatvus, mille lugemisel on esinenud vigu, on Poissoni jaotusega, siis on vaatlusandmete jaotuseks Poissoni jaotuste segu, kus a määrab vigade osakaalu andmetes.

2.3 Suurima tõepära meetod, Poissoni jaotus

Suurima tõepära meetodi abil saab valimi põhjal hinnata tõenäosusjaotuste parameetrite väärtusi. Leidmaks parameetri(te) hinnangu(id) tuleb maksimeerida tõepärafunktsiooni. Näiteks, olgu teada, et n väärtusest koosnev valim $k_1, k_2, \ldots, k_n$ on Poissoni jaotusega, kuid jaotuse parameeteri λ väärtus olgu teadmata. Hindamaks parameetri λ väärtust, saab kasutada suurima tõepära meetodit. Poissoni jaotuse tõepärafunktsioon avaldub kujul

$$L(\lambda; k_1, \ldots, k_n) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!}.$$

Tõepärafunktsiooni asemel on lihtsam maksimeerida log-tõepära. Log-tõepära saamiseks võtakse tõepärafunktsioonist naturaallogaritm

$$l(\lambda; k_1, \ldots, k_n) = \ln\left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{k_i}}{k_i!}\right) = \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda}\lambda^{k_i}}{k_i!}\right) =$$

$$= \sum_{i=1}^{n} \left(\ln(e^{-\lambda}) + \ln(\lambda^{k_i}) - \ln(k_i!)\right) =$$

$$= \sum_{i=1}^{n} \left(-\lambda \ln(e) + k_i \ln(\lambda) - \ln(k_i!)\right) =$$

$$= -n\lambda + \ln(\lambda) \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} \ln(k_i!). \tag{3}$$

Seejärel maksimiseeritakse saadud avaldist. Selleks leitakse tuletis parameetri λ järgi, pannakse saadud avaldis võrduma nulliga ning avaldatakse parameetri λ järgi

$$\frac{d}{d\lambda} l(\lambda; k_1, \ldots, k_n) = \frac{d}{d\lambda} \left(-n\lambda + \ln(\lambda) \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} \ln(k_i!) \right) = -n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i$$

$$-n + \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} k_i = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i.$$

Lisaks kontrollitakse, et tegemist on maksimumiga, leides selleks teise tuletise parameetri λ järgi

$$\frac{d}{d\lambda} \left(-n + \frac{1}{\lambda} \sum_{i=1}^{n} k_i \right) = -\frac{1}{\lambda^2} \sum_{i=1}^{n} k_i.$$

Teine tuletis kohal $\hat{\lambda}$ tuleb negatiivne, mis näitab, et $\hat{\lambda}$ tõepoolest maksimeerib tõepärafunktsiooni. Leitud tulemus näitab, et hindamaks Poissoni parameetri λ väärtust tuleb leida valimi keskmine.

2.4 Suurima tõepära meetod, Poissoni jaotuste segu

Soovides keskmise sekveneerimiskatvuse hindamisel võtta arvesse võimalikke tekkinud vigu, on andemte jaotus vaadeldav segujaotusena, kus üks segu-

jaotuse komponentidest kirjeldab vigade jaotust. Teades, et andmed pärinevad ühest Poissoni jaotusest parameetriga λ_1 osakaaluga $(1-a)$ ja teisest Poissoni jaotusest parameetriga λ_2 osakaaluga a, saab rakendada suurima tõepära meetodit, et hinnata Poissoni jaotuse parameetreid ning osakaalu a. Tõepärafunktsioon eespool kirjeldatud segujaotusena kirja pandult avaldub järgmiselt

$$L(a, \lambda_1,\lambda_2; k_1, \ldots, k_n) = \prod_{i=1}^{n} \left((1-a) \cdot \frac{e^{-\lambda_1}\lambda_1^{k_i}}{k_i!} + a \cdot \frac{e^{-\lambda_2}\lambda_2^{k_i}}{k_i!} \right).$$

Suurima tõepära meetodil hinnangu leidmiseks viiakse tõepärafunktsioon kõigepealt logaritmilisele kujule ehk leitakse log-tõepära. Seejärel leitakse vastavalt parameetritele osatuletised ning makismeeritakse saadud avaldisi.

$$\begin{aligned} l(a, \lambda_1,\lambda_2; k_1, \ldots, k_n) &= \ln \left(\prod_{i=1}^{n} \left((1-a) \cdot \frac{e^{-\lambda_1}\lambda_1^{k_i}}{k_i!} + a \cdot \frac{e^{-\lambda_2}\lambda_2^{k_i}}{k_i!} \right) \right) = \\ &= \sum_{i=1}^{n} \ln \left((1-a) \cdot \frac{e^{-\lambda_1}\lambda_1^{k_i}}{k_i!} + a \cdot \frac{e^{-\lambda_2}\lambda_2^{k_i}}{k_i!} \right) = \qquad (4) \\ &= \sum_{i=1}^{n} \ln \left(\frac{1}{k_i!} \left((1-a) \cdot e^{-\lambda_1}\lambda_1^{k_i} + a \cdot e^{-\lambda_2}\lambda_2^{k_i} \right) \right) = \\ &= \sum_{i=1}^{n} \left(-\ln(k_i!) + \ln \left((1-a) \cdot e^{-\lambda_1}\lambda_1^{k_i} + a \cdot e^{-\lambda_2}\lambda_2^{k_i} \right) \right) \end{aligned}$$

Leides osatuletised parameetrite a, λ_1 ja λ_2 järgi ning pannes nad võrduma nulliga ehk maksimiseerides, selgub, et parameetrite hinnanguid täpselt leida on keeruline. Seetõttu kasutatakse töös log-tõepära maksimeerimiseks numbrilisi meetodeid.

Saadud tulemus võimaldab leida keskmise sekveneerimiskatvuse ehk λ_1 hinnangut võttes arvesse, et andmetes leidub mingi osakaaluga erindeid, mis vastavad samuti Poissoni jaotusele, kuid mille jaotuse parameeter λ_2 on erinev tegelikust keskmisest sekveneerimiskatvusest.

2.5 Tõepärasuhte test

Tõepärasuhte testi abil on võimalik välja selgitada andmetele kõige paremini vastav mudel kahe erineva mudeli seast, kus üks on teise erijuht. Näiteks hinnatakse sama tunnuse jaotust, kuid üks mudel kasutab selleks rohkem parameetreid kui teine. Tõepära suhte testi teststatistik avaldub kujul

$$LRT = -2 \cdot \ln\left(\frac{L^*(k_1, \ldots, k_n)}{L(k_1, \ldots, k_n)}\right), \tag{5}$$

kus $L^*(k_1, \ldots, k_n)$ on n-ö lihtsama (väiksema parameetrite arvuga) meetodi või mudeli tõepärafunktsiooni maksimaalne väärtus ning $L(k_1, \ldots, k_n)$ on n-ö keerukama (suurema parameetrite arvuga) meetodi või mudeli tõepärafunktsiooni maksimaalne väärtus, kus mõlemad väärtused on leitud kasutades valimit $k_1, \ldots, k_n$. Nullhüpoteesi kehtides on teststatistik asümptootiliselt χ^2-jaotusega ning vabadusastmete arv saadakse leides mõlema meetodi parameetrite arvude vahe. (Casella ja Berger, 2002)

Tõepärasuhte teststatistiku seos suurima tõepära meetodiga avaldub selgelt valemist 5. Seetõttu on ootuspärane, et statistiku käitumine sõltub suurima tõepära meetodi asümptootilisest lähendamisest normaaljaotusega. Selgub, et antud töös, teatud tingimustel, ei saa suurima tõepära meetodit lähendada normaaljaotusega, mistõttu on sellest mõjutatud ka tõepärasuhte teststatistiku asümptootiline lähendamine χ^2-jaotusega. Nimelt on ühe hinnatava parameetri, erindite osakaalu a, puhul rikutud tingimus, et suurima

tõepära meetodiga hinnatava parameetri tegelik väärtus peab olema vastava parameetri-ruumi Ω kuuluva lahtise hulga ω sisepunkt. Erindite osakaalu parameetri-ruumiks Ω_a on vastavalt vahemik [0,1]. Kui erindite tegelikuks osakaaluks on väärtus $a = 0$ või $a = 1$, siis sel juhul on a näol tegemist ω_a ($\omega_a \subseteq \Omega_a$) rajapunktiga. (Casella ja Berger, 2002) Olenemata sellest, et osakaalu $a = 0$ (või $a = 1$) juures on rikutud tingimus, mis mõjutab teststatistiku asümptootilist lähendamist χ^2-jaotusega soovitakse töös siiski vaadata, kuidas töötab test ka selliste andmete puhul, kus selline olukord esineb.

Võrdlemaks omavahel aritmeerilise keskmise ja segumudeli sobivust andmetega, saab tõepärasuhte teststatsitiku leida järgnevalt

$$LRT = 2 \cdot (l(a, \lambda_1,\lambda_2; k_1, \ldots, k_n) - l(\lambda; k_1, \ldots, k_n)),$$

kus $l(a, \lambda_1,\lambda_2; k_1, \ldots, k_n)$ on segumudeli log-tõepära, mis avaldub valemiga 4 ning $l(\lambda; k_1, \ldots, k_n)$ on lihtsama meetodi ehk aritmeetilise keskmise log-tõepära, mis avaldub valemiga 3. Antud juhul võrreldakse teststatistiku LRT väärtust $\chi^2_{df=2}$ väärtusega ning kontrollitav hüpotees on kujul:

H_0 : $\lambda = \lambda_1$ ehk kasutada tuleks aritmeetilist keskmist;

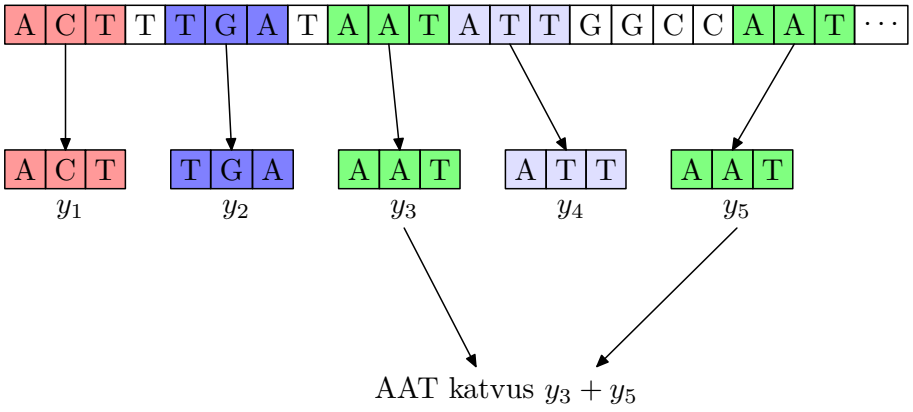
H_1 : $\lambda \neq \lambda_1$ ehk kasutada tuleks segumudelit.

Tõepärasuhte testi abil saab võrrelda erinevate jaotuste sobivust andmetega ning vastavalt valitud mudelile saab otsustada, millist hindamismeetodit tuleks kasutada. Seetõttu on test sobilik otsustamaks, kas keskmise sekveneerimiskatvuse hinnangu leidmiseks tuleks hinnata segumudeli parameetreid või piisab aritmeetilisest keskmisest.

3 Töö käik

3.1 Töö eesmärk

On teada, et valitud piirkondade sekveneerimiskatvus ehk see, mitu korda on ühte ja seda sama piirkonda loetud, on Poissoni jaotusega juhuslik suurus. Sekveneerimisel võib esineda vigu. Vead võivad olla tingitud näiteks sekveneerimisprotsessist endast, ümbritseva keskkonna mõjutustest ja/või muudatustest genoomis, milles polda teadlik (Koboldt *et al.*, 2010). Allolev joonis (joonis 2) kirjeldab viimast olukorda. Nimelt on joonisel toodud neli teadaolevat unikaalset mustrit ACT, TGA, AAT, ATT, kuid toimunud mutatsiooni tõttu pole muster AAT enam unikaalne, ning esineb genoomis kahes erinevas kohas. Kuna toimunud muudatusest ei teata, siis loetakse mustri AAT sekveneerimiskatvuseks joonisel esimesena esineva AAT mustri sekveneerimiskatvuse y_3 ning teisena esineva AAT mustri sekveneerimiskatvuse y_5 summa ehk $y_3 + y_5$. On selge, et sellisel juhul on mustri AAT katvus tunduvalt suurem kui ta tegelikult olema peaks ning tekkinud erind mõjutab keskmise katvuse arvutamisel tulemust.



Joonis 2: Probleemi kirjeldav joonis.

Sarnane olukord võib tekkida ka siis, kui uuritavasse proovi on ümbritsevast keskkonnast sattunud bakterid, mille genoomis esineb samuti sama mustrit. Mõned üksikud erindid küll ei mõjuta keskmise katvuse arvutamisel tulemust, kuid kui erindite osakaal suureneb, hakkavad nad tulemust oluliselt mõjutama.

Juhul kui on tekkinud vigu ning osade piirkondade katvus on loetud tegelikust suuremaks või väiksemaks, on ka nende piirkondade katvus Poissoni jaotusega, küll aga on nende piirkondade puhul jaotuse parameeter tegelikust parameetrist erinev. Olgu tegelik keskmine katvus parameetriga λ_1 ning vigadest mõjutatud piirkondade keskmine katvus parameetriga λ_2. Leides λ_1 ja λ_2 väärtustele vastavalt Poissoni jaotuse valimid, ning osakaalu a järgi mõlemast valimist väärtusi valides, on tulemuseks segujaotus, kus suurema osakaaluga $(1-a)$ valitakse korrektse katvusega vaatlusi, mis vastavad parameetrile λ_1 ning väiksema osakaaluga a tegelikust erineva katvusega vaatlusi, mis vastavad parameetrile λ_2.

Saadud segujaotuse järgi soovitakse hinnata tegelikku keskmist katvust λ_1. Hinnangut saab leida kasutades selleks erinevaid meetodeid. Üheks on arvutada välja segujaotuse väärtuste aritmeetiline keskmine ning teiseks on hinnata keskmist katvust segujaotust modelleerides ehk segumudeli abil, kus on arvesse võetud kõiki parameetreid λ_1, λ_2 ja a. Leidmaks segumudeli hinnangut tegelikule keskmisele katvusele, saab kasutada suurima tõepära meetodit Poissoni jaotuste segu jaoks, pannes selleks kirja avaldise 4 rakendustarkvara R funktsioonina ning seejärel rakendades optim-käsku. Numbrilistel meetoditel põhinev optim-käsk võimaldab leida etteantud segujaotuse ja funktsiooni korral parima ligikaudse lahendi ehk hinnangu soovitud parameetritele. Tulemusena saadakse λ_1, λ_2 ja a hinnangud ühe vaadeldud segujaotuse realisatsiooni jaoks.

Eesmärk on fikseeritud λ_1 ja λ_2 puhul korrata katset erinevate osakaalu a väärtuste korral, et välja selgitada, millal üks meetod teisest täpsemaks muutub, arvutades selleks välja meetodite keskmised ruutvead.

3.2 Keskmine ruutviga ja nihe

Keskmine ruutviga iseloomustab hinnangu ebatäpsust. Mida väiksem on viga, seda täpsem on leitud hinnang. Teades, et aritmeetiline keskmine on tundlik erindite suhtes, on eeldatav, et mida suuremaks erindite osakaal läheb seda ebatäpsemaks muutub meetodiga leitud hinnang ehk seda suurem on keskmine ruutviga. Sarnaselt on eeldatav, et probleemile keerukamalt lähenedes, kasutades selleks segumudelit ning seeläbi võttes lisaks arvesse ka erindeid ja nende osakaalu, on tulemused märgatavalt paremad ning hindamisel tehtav viga tunduvalt väiksem kui aritmeetilise keskmise puhul.

Leidmaks keskmist ruutviga, leitakse kõigepealt tegeliku keskmise katvuse λ_1 hinnang $\hat{\lambda}_1$ mõlema meetodi korral teatud parameetri a väärtuse juures. Seejärel leides tegeliku ja hinnatud parameetri vahe $(\lambda_1 - \hat{\lambda}_1)$, tõstes selle ruutu ning leides siis kõikide osakaalule a vastavate vahe ruutude keskmise, on tulemuseks keskmine ruutviga. Korrates hinnangu leidmise protsessi ühe konkreetse a väärtuse jaoks, saab seeläbi keskmist ruutviga täpsemalt hinnata.

Kui on teada parameetrite λ_1, λ_2 ja a väärtused, saab aritmeetilise keskmisega leitud hinnangul põhineva ligikaudselt arvutatud keskmise ruutvea asemel leida keskmise ruutvea väärtused teoreetiliselt, kasutades selleks segujaotuse dispersiooni ja nihet. Selleks rakendatakse metoodika osas leitud valemeid 1

ja 2. Teoreetiline keskmine ruutviga avaldub järgnevalt

$$MSE = \frac{DX}{j} + (EX - \lambda_1)^2 = \tag{6}$$

$$= \frac{((1-a)\lambda_1 + a\lambda_2) + (a(1-a)(\lambda_1 - \lambda_2)^2)}{j} + (((1-a)\lambda_1 + a\lambda_2) - \lambda_1)^2,$$

kus j tähistab segujaotuse valimi suurust.

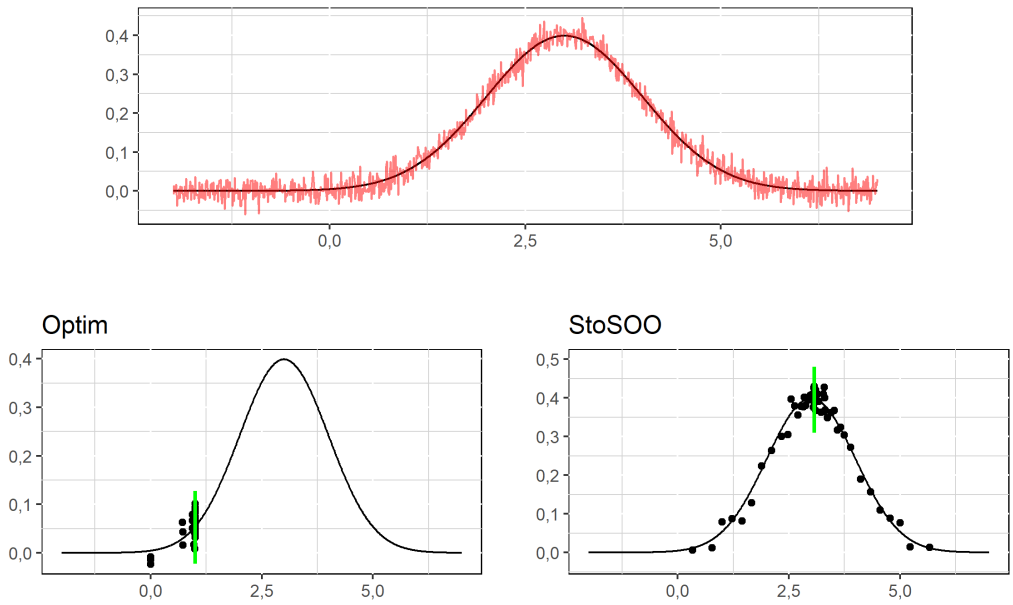
Sekveneerimisel saadavate piirkondade katvuste simuleerimiseks kasutati rakendustarkvara R rpois-käsku, mis tagastab kindla parameetriga Poissoni jaotusele vastavate arvude jada. Et saada segujaotust simuleeriti 10 000 piirkonna katvused vastavalt parameetrile λ_1 ning 10 000 vaatlust vastavalt parameetrile λ_2. Seejärel valiti vastavalt oskaalule a ühtlasest jaotusest simuleeritud arvude järgi kokku 10 000 vaatlust mõlemast Poissoni jaotusest. Saadud segujaotust kasutati tegeliku keskmise katvuse λ_1 hinnangu leidmiseks kahel viisil - leides segujaotuse aritmeetilise keskmise ning leides segumudeli hinnangu. Segumudeli hinnangu leidmiseks kasutati suurima tõepära meetodit Poissoni jaotuste segu jaoks, pannes funktsioonina kirja avaldise 4 ning seejärel rakendades sellel funktsioonil optim-käsku. Seejärel leiti meetodite tehtud keskmised ruutvead. Aritmeetilise keskmisega saadud hinnangu keskmised ruutvead arvutati vastavalt eelpool toodud valemile 6. Segumudeli puhul arvutati keskmine ruutviga ligikaudselt, kasutades selleks eelneval leheküljem toodud arutluskäiku.

3.3 Keskmiste ruutvigade võrdlus

Leidmaks osakaalu a väärtust, millest alates üks meetod teisest täpsemaks muutub kasutati stohhastilisel optimeerimisel põhinevat StoSOO-käsku paketist OOR. StoSOO võimaldab minimiseerida talle etteantavat funktsiooni,

mida pole võimalik täpselt välja arvutada. Seetõttu oskab StoSOO eirata suurima tõepära hinnangu leidmisel tekkivat kõikumist ehk selliseid osakaalu a väärtusi, kus meetodid lõikuvad, küll aga on see lõikumine tingitud vaid hinnangu kõikumisest ning tegemist pole tegeliku lõikumispunktiga.

Alloleval joonisel (joonis 3) on näiteks toodud, kuidas leiab StoSOO võrreldes optim-käsuga funktsiooni maksimumpunkti juhul, kui etteantava funktsiooni väärtused kõiguvad (punane joon joonisel).



Joonis 3: StoSOO ja optim-käsu erinevust kirjeldav joonis.

Joonisel (joonis 3) on näiteks võetud normaaljaotus keskväärtusega 3, kuhu on väärtuste kõikumise loomiseks lisatud juhuslik väikese hajuvusega komponent. Rakendades mõlemat optimeerimiskäsku sellel kõikuvate väärtustega funktsioonil, tuleb selgelt välja nende käskude erinevus. Mustad punktid alumisel kahel joonisel tähistavad väärtusi, mida vastav optimeerija potentsiaalsete maksimum-väärtustena katsetab ning roheline joon tähistab optimeerija leitud ning tagastatud maksimumpunkti. Selgelt on näha, et optim-käsk ei

suuda eirata funktsiooni kõikumist ning seetõttu katsetab väärtusi ainult ühes etteantava piirkonna ([0,6]) osas, tagastades funktsiooni maksimumpunktina väärtuse, mis on tegelikust keskmisest väga kaugel. Seevastu katsetab StoSOO erinevaid väärtusi üle selle sama etteantava piirkonna, eirates funktsiooni kõikumist ning leides lõpuks tegeliku või tegelikkusele väga lähedase funktsiooni maksimumpunkti.

Kuna töö eesmärk on leida meetodite lõikumispunkt, siis on otsitav a väärtus selline, mille korral hinnangute ruutvigade vahe on minimaalne. Pannes eelneva arutluskäigu - segujaotuse simuleerimise, hinnangute ja ruutvigade leidmise - kirja funktsioonina, mis tagastab ruutvigade vahe ning rakendades sellel StoSOO-käsku miinimumpunkti leidmiseks, on võimalik leida soovitud a väärtus. Kõigi tulemuste täpsuste tagamiseks korrati iga a väärtuse jaoks katset 10 000 korda ning StoSOO iteratsioonide arvuks valiti 55 funktsiooni pika tööaja tõttu.

Rakendustarkvara R programmikoodina on lõikumispunkti leidmine esitatud lisades.

3.4 Tõepärasuhte testi võimsus ja kasutatavus

Tõepärasuhte test võimaldab etteantavate andmete põhjal otsustada kumba meetodit andmetele rakendada ehk kas andmetes on erindeid või mitte. Testi kasutamiseks on aga oluline teada, kui hästi suudab test õige otsuse langetada. Üks viis testi iseloomustada on leida testi võimsus. Testi võimsus näitab kui tõenäoliselt suudab test vältida II tüüpi vea tegemist ehk alternatiivse hüpoteesi kehtimisel nullhüpoteesi juurde jäämist.

Testi võimsuse leidmiseks simuleeriti iga osakaalu korral 10 000 andmestikku ja iga simuleeritud andmetiku põhjal tehti tõepärasuhte test. Varasem

teadmine meetodite lõikumispunktist võimaldab hinnata testi otsuse korrektsust ning vaadata, kas ka tõpärasuhte test otsustab segumudeli kasuks nende vigade osakaalude korral, mille puhul varasemate teadmiste järgi tagab segumudeli rakendamine täpsema hinnangu.

Tegeliku keskmise sekveneerimiskatvuse $\lambda_1 = 10$ ja erindite keskmise katvuse $\lambda_2 = 7$ puhul, erindite osakaalu $a = 0{,}07$ juures jäi test nullhüpoteesi juurde 7996 korral ning vaid 2004 korral otsustas alternatiivse hüpoteesi kasuks. Tulemustest (vt ptk tulemused) on teada, et selliste parameetrite väärtuste juures peaks test otsustama alternatiivse hüpoteesi kasuks. Seega antud juhul testi võimsus ehk tõenäosus vältida II tüüpi viga on

$$\text{võimsus} = 1 - \mathbb{P}(\text{II tüüpi viga}) = 1 - \frac{7996}{10000} = 0{,}2004$$

ehk kõigest 20,04%, mis viitab sellele, et osakaalu $a = 0{,}07$ juures testi otsustamisvõime pole kuigi hea.

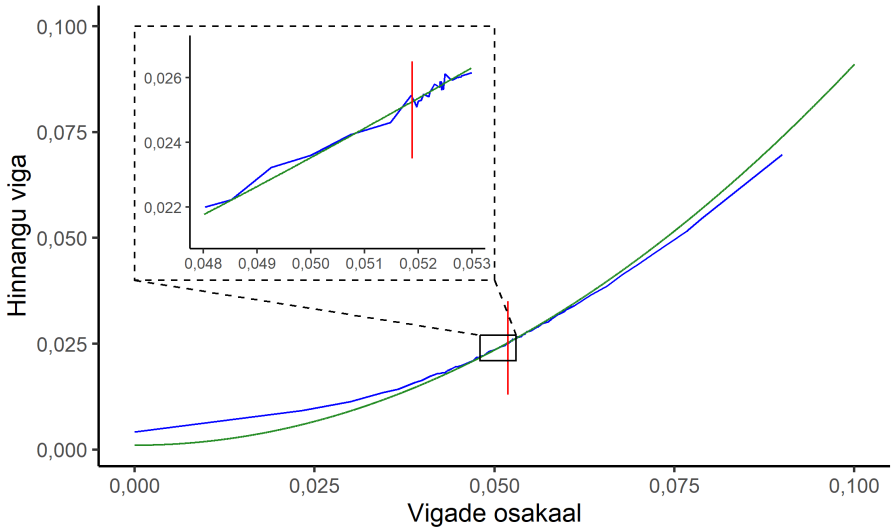
Kuigi eelnev näide viitab sellele, et testi otsustsvõime pole parim, jääb siiski küsimus, kas testi rakendades jõutaks ruutvigade mõttes parema tulemuseni. Ehk kui valida hindamismeetod vastavalt testi otsusele, siis kas saadud hinnang on keskmise ruutvea mõttes täpsem võrreldes sellega, kui leida hinnang alati segumudelit kasutades.

4 Tulemused

Sekveneerimise käigus saadakse teada, mitu korda on mingit konkreetset genoomi piirkonda loetud, ehk mis on selle piirkonna katvus. Kõikide piirkondade katvused moodustavad Poissoni jaotusega valimi, mille keskväärtuseks ehk keskmiseks sekveneerimiskatvuseks on jaotuse parameeter λ_1. Genoomis toimunud mitte teada olevad mutatsioonid ja/või sekveneerimisvead võivad põhjustada olukordi, kus mingite piirkondade puhul on katvus loetud tegelikkusest väiksemaks või suuremaks. Selliste piirkondade katvused on samuti Poissoni jaotusega, kuid nendele andmetele vastav jaotuse parameeter λ_2 on vastavalt kas väiksem või suurem kui tegelik keskmine sekveneerimiskatvus λ_1. Nendest kahest erineva parameetri väärtusega Poissoni jaotusest kokku pandud segujaotus on valim tegelikkusele vastavatest andmetest ja mingi osakaaluga vigastest andmetest. Kui andmed sisaldavad vigu, siis hinnates keskmist sekveneerimiskatvust aritmeetilise keskmise abil, on selge, et hinnatav keskmine katvus pole täpne. Mida suurem on erindite ehk selliste andmete osakaal, kus katvus on vale, seda ebatäpsemaks hinnang muutub. Modelleerides segumudeli järgi, arvestades tekkinud erindeid ning nende osakaalu andmetes, on eeldatavalt vigade osakaalu suurenedes tegelikku keskmise sekveneerimiskatvuse λ_1 hinnang täpsem. Selleks, et välja selgitada erindite osakaal, alates millest tagab segumudel täpsema hinnangu, otsiti erindite osakaalu lõikes meetodite lõikumispunkti.

Alljärgneval joonisel (joonis 4) on hinnatud tegelikku keskmist sekveneerimiskatvust $\lambda_1 = 10$, mis on mõjutatud tegelikkusest väiksema keskmise katvusega andmetest parameetri väärtusega $\lambda_2 = 7$. Roheline joon tähistab aritmeetilist keskmist, sinine suurima tõepära meetodit ning punasega on tähistatud stohhastilise funktsiooni hinnang tõenäosusele, kust üks meetod

teisest täpsemaks muutub. Mida väiksem on meetodi tehtav viga, seda täpsem on ta keskmise hindamisel. Joonise järgi on näha, et vigade osakaalu $0{,}04 - 0{,}06$ puhul on meetodite tulemused üpris sarnased, kuigi osakaalu 0,05 juures hakkab muutuma segumudeli hinnang täpsemaks. Vahemikus $0 - 0{,}04$ osutub aritmeetilise keskmise abil leitud hinnang veidi täpsemaks ning stohhastilise funktsiooni poolt hinnatud meetodite lõikumispunkt asub osakaalu 0,0519 juures.

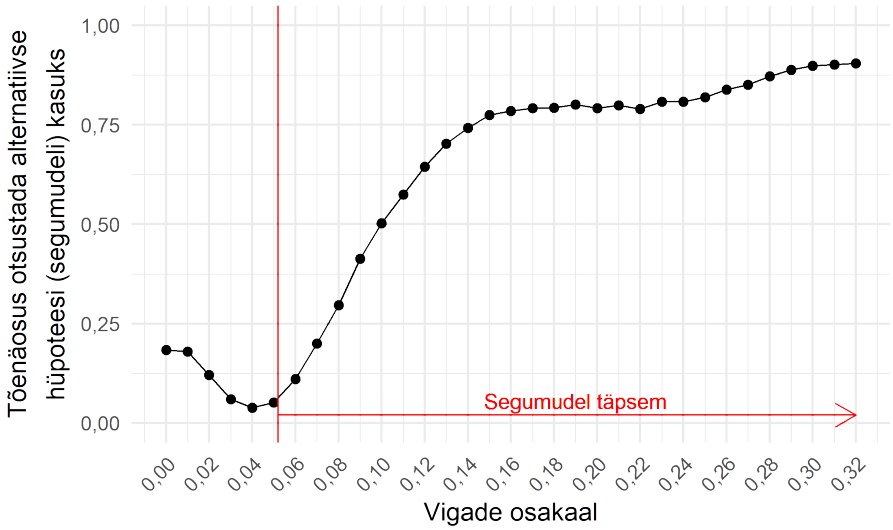


Joonis 4: Meetodite keskmine ruutviga vigade osakaalu a lõikes, kus tegelik keskmine sekveneerimiskatvus on $\lambda_1 = 10$ ning erindid on keskmise sekveneerimiskatvusega $\lambda_2 = 7$.

Ehk selgub, et kui sekveneerimisel tekkinud vigade tõttu on osade piirkondade katvus keskmise katvusega $\lambda_2 = 7$, siis senikaua kuni selliseid väärtusi on vähem kui ligikaudu 5,19% kõikidest andmetest, on aritmeetilise keskmise abil leitud keskmine katvus täpsem. Sealt edasi osutub täpsemaks segumudeli abil leitud keskmine sekveneerimiskatvus.

Jättes parameetrite väärtused samaks ($\lambda_1 = 10,\ \lambda_2 = 7$), on järgneval joonisel (joonis 5) toodud tõepärasuhte testi tõenäosus otsustada alternatiivse

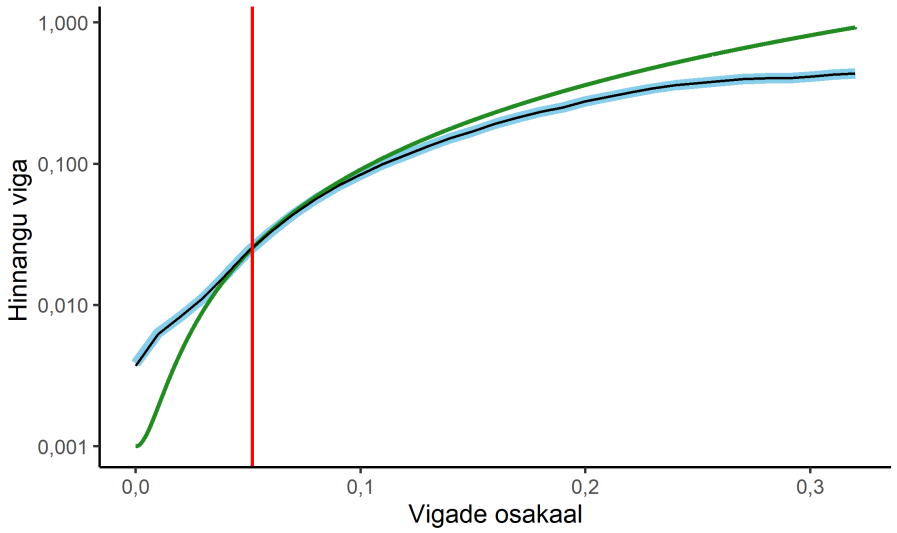
hüpoteesi ehk segumudeli kasuks, lähtudes erindite osakaalust andmetes. Punasest vertikaalsest joonest, mis tähistab eelnevalt leitud meetodite lõikumispunkti, paremale jäävast osast võib välja lugeda ka testi võimsuse. Jooniselt selgub, et vigade osakaalu $0 - 0{,}02$ juures, kus testi otsus peaks kindlalt olema nullhüpoteesi ehk aritmeetilise keskmise kasuks, kaldub test ligikaudu $10 - 20\%$ juhtudest siiski otsustama segumudeli kasuks. Selline olukord on tingitud sellest, et tõepärasuhte testi eeldused selles vahemikus on rikutud (vt ptk tõepärasuhte test), mistõttu teststatistiku χ^2-jaotusega lähendamine osutub valeks. Alates meetodite lõikumispunkti tähistavast vertikaalsest joonest hakkab vigade osakaalu suurenedes kasvama ka testi tõenäosus langetada n-ö õige otsus. Kuigi vigade osakaalu $0{,}06 - 0{,}07$ juures, kus võib pidada testi otsust ehk kõige olulisemaks, jääb see tõenäosus siiski üpris väikeseks, umbes $10 - 20\%$.



Joonis 5: Tõepära suhte testi tõenäosus võtta vastu alternatiivne hüpotees vigade osakaalu a lõikes, kus tegelik keskmine sekveneerimiskatvus on $\lambda_1 = 10$ ning erindid on keskmise sekveneerimiskatvusega $\lambda_2 = 7$.

Saadud tulemus aitab vastata ka töö käigu osas püstitatud küsimusele, kas

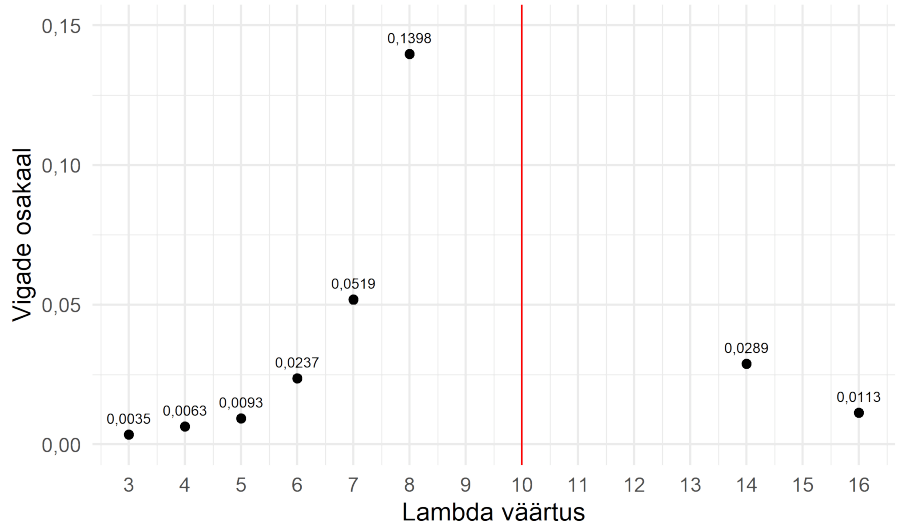
hinnangu viga oleks väiksem, kui valida λ_1 hinnanguks just selle meetodi hinnang, mille kasuks on tõepärasuhte test otsustanud. Vaadates selleks joonist 4 ja joonist 5, siis vahemikus $0 - 0{,}02$ valitakse umbes $10 - 20\%$ juhtudes segumudeli hinnang, mistõttu on ka testi kaasates hinnangu viga kallutatud segumudeli hinnangu vea poole (sinine joon joonisel 4). Meetodite lõikepunkti ümbruses on mõlema meetodi abil saadud hinnangute vead ligilähedased, mistõttu testi kaasamine selles piirkonnas mõjutab tulemust minimaalselt. Sealt edasi otsustatakse enamjaolt segumudeli hinnangu kasuks, mistõttu on ka testi kaasamisel saadud hinnang sellele väga lähedane. Saadud tulemust kirjeldav joonis on toodud jägnevalt (joonis 6), kus on näha, et tõepärasuhte testi otsuse kaasamisel saadav hinnangu viga (must joon) ja segumudeli tehtav hinnangu viga (sinine joon) kattuvad.



Joonis 6: Meetodite keskmine ruutviga logaritmilisel skaalal osakaalu a lõikes, kus tegelik keskmine sekveneerimiskatvus on $\lambda_1 = 10$ ning erindid on keskmise sekveneerimiskatvusega $\lambda_2 = 7$. Roheline joon tähistab aritmeerilise keskmise hinnangu viga, sinine joon segumudeli hinnangu viga, punane vertikaalne joon meetodite lõikumispunkti ning must joon tähistab tõepärasuhte testi otsusel põhinevat hinnangu viga.

Teisisõnu meetodi tehtava keskmise ruutvea mõttes annab tõepärasuhte testi kaasamine otsustusprotsessi ligikaudu sama tulemuse, kui kasutada hinnangu leidmiseks ainult segumudelit.

Juhtudel, kus parameetri λ_2 väärtus on tegelikust keskmisest katvusest λ_1 väiksem, on sekveneerimisel osade piirkondade katvusi loetud tegelikkusest väiksemaks ning juhtudel, kus λ_2 väärtus on suurem, on mõningate piirkondade katvusi loetud tegelikkusest suuremaks. Järgneval joonisel (joonis 7) on toodud erinevate erinditele vastava parameetri λ_2 väärtuste puhul kõik leitud lõikumispunktid, millest alates on segumudeliga saadud kekmise sekveneerimiskatvuse hinnang täpsem. Näiteks, kui vigadest mõjutatud katvused on keskmise katvusega $\lambda_2 = 3$, siis alates osakaalust 0,0035 ehk 0,35% annab segumudel tegeliku keskmise katvuse hindamisel täpsema tulemuse kui aritmeetilise keskmise leidmine.



Joonis 7: Vigade osakaalu a väärtused, millest alates annab aritmeerilise keskmisega võrreldes segumudeli rakendamine täpsema hinnangu keskmisele sekveneerimiskatvusele. Tegelikuks keskmiseks sekveneerimiskatvuseks on $\lambda_1 = 10$ ning parameetri λ_2 väärtustele vastavad erindite keskmised sekveneerimiskatvused.

Jooniselt 7 selgub, et mida lähemale tegelikkusele keskmisele katvusele parameetri λ_2 väärtus jääb, seda suurema osakaalu juurde jääb meetodite lõikumispunkt ehk seda suurema osakaaluni on aritmeetilise keskmise hinnang keskmisele sekveneerimiskatvusele täpsem. Parameetri $\lambda_2 = 9$ korral osutub meetodite lõikumispunkti osakaaluks lausa 0,4966 ehk 49,6%. Samuti selgub, et mida suurem on parameetrite λ_1 ja λ_2 erinevus, seda väiksema osakaalu juures muutub segumudeliga saadud hinnang täpsemaks. Saadud tulemus näitab, et kui andmed sisaldavad vaatlusi, mis on tavapärasest tunduvalt väiksemad või suuremad, on segumudeli kasutamine hinnangu täpsuse saavutamiseks oluline juba vigade väikese osakaalu juures.

Näiteks, pärib laps oma DNA mõlemalt vanemalt. Olukorras, kus ühelt vanemalt pärit DNA ahelas esineb unikaalse mustri piirkonnas mutatsioon (teiselt vanemalt päritud DNA-s sama mutatsiooni pole), siis vastava piirkonna sekveneerimiskatvust DNA-mustri abil tuvastades näeme kõigest pooli - ühe vanema DNA-ahelalt pärit lugemeid. Seetõttu on unikaalsete piirkondade katvusi lugedes mutatsiooniga piirkondade puhul keskmine katvus poole võrra väiksem kui tegelik keskmine sekveneerimiskatvus. Antud näites, kui tegelikuks keskmiseks sekveneerimiskatvuseks on $\lambda_1 = 10$, oleks muteerunud piirkondade keskmiseks katvuseks $\lambda_2 = 5$ ning joonisel 7 toodud tulemuse järgi tuleks leida keskmise katvuse hinnang kasutades segumudelit, kui vigade osakaal ületab 0,9%.

Kokkuvõte

Selleks, et välja selgitada aritmeetilise keskmise ja segumudeli täpsus keskmise sekveneerimiskatvuse λ_1 hindamisel ning hinnata vigade osakaalu, millest alates segumudel muutub täpsemaks, simuleeriti kahe erineva parameetri, λ_1 ja λ_2, väärtusega Poissoni jaotuse segujaotusega andmed, kus tõenäosusega a (vigade osakaal) pärinesid andmed parameetrile λ_2 vastavast Poissoni jaotusest. Simuleeritud andmestiku kasutades leiti mõlema meetodi hinnangud keskmisele sekveneerimiskatvusele λ_1. Vigade osakaalu leidmiseks, millest alates segumudeli hinnang muutub täpsemaks kui aritmeetiline keskmine (n-ö lõikepunkt), kasutati stohhastilisel optimeerimisel põhinevat rakendustarkvara R käsku StoSOO. Samuti hinnati tõepärasuhte testi võimet langetada otsus vastavalt etteantud andmetele sobiva meetodi kasuks.

Töö tulemusena leiti 9 erineva situatsioon jaoks lõikepunkti väärtus. Ootuspäraselt selgus, et mida suurem on parameetrite λ_1 ja λ_2 erinevus, seda väiksem on erindite osakaal, millest alates segumudeli hinnang täpsemaks osutub. Saadud tulemus viitab sellele, et segumudeli rakendamine muutub oluliseks, mida suuremad või väiksemad on ekslikud vaatlused andmetes, isegi kui selliste andmete osakaal on väike. Samuti selgus, et otsustades tõepärasuhte testi järgi kumba meetodit rakendada, tehakse hinnangu keskmise ruutvea mõttes ligikaudu sama suur viga, kui alati segumudelit kasutades.

Kasutatud allikad

Brown T., A. (2022). *Genomes, 2nd edition*. Garland Science. Ptk 1. ISBN: 0471250465.

Burian, A. N., W. Zhao, T.-W. Lo ja D. M. Thurtle-Schmidt (juuli 2021). "Genome sequencing guide: An introductory toolbox towhole-genome analysis methods". *Biochemistry and molecular biology education* 49(5), lk. 815–825. DOI: 10.1002/bmb.21561.

Casella, G. ja L. Berger R. (2002). *Statistical Inference*. 2. väljaanne. United States: Thomson Learning, Australia, lk. 488–490, 516. ISBN: 0534243126.

Chauhan, T. (august 2022). *What Is 'Sequencing Read' In NGS?* URL: https://geneticeducation.co.in/what-is-sequencing-read-in-ngs/ (vaadatud 06.05.2023).

Illumina (jaanuar 2014). *Estimating Sequencing Coverage*. URL: https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf (vaadatud 10.10.2023).

– (2024). *Sequencing Coverage for NGS Experiments*. URL: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html (vaadatud 08.04.2024).

Koboldt D., C., L. Ding, R. Mardis E. ja R. K. Wilson (2010). "Challenges of sequencing human genomes". *Briefings in Bioinformatics* 11(5), lk. 484–498. DOI: 10.1093/bib/bbq016.

Mardis, E. R., L. Ding, D. J. Dooling, D. E. Larson, M. D. McLellan *et al.* (2009). "Recurring mutations found by sequencing an acute myeloid leukemia genome". *The New England journal of medicine* 361(11), 1058–1066. DOI: 10.1056/NEJMoa0903840.

Miga, K. H., Y. Newton, M. Jain, N. Altemose, F. Willard H. *et al.* (2014). "Centromere reference models for human chromosomes X and Y satellite arrays". *Genome Research* 24, lk. 697–707. DOI: 10.1101/gr.159624.113.

NHGRI (august 2020). *DNA Sequencing Fact Sheet*. URL: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet (vaadatud 05.05.2023).

Piovesan, A., M. C. Pelleri, Strippoli P. Antonaros F., M. Caracausi ja L. Vitale (2019). "On the length, weight and GC content of the human genome". *BMC research notes* 12(1), lk. 106. DOI: 10.1186/s13104-019-4137-z.

Seeking Alpha (veebruar 2023). *Illumina, Pacific Biosciences, And Oxford Nanopore Market Position Comparison*. URL: https://seekingalpha.com/article/4575505-gene-sequencing-market-illumina-pacific-bio-oxford-nanopore-comparison (vaadatud 10.10.2023).

Tiit, E., A. Parring ja T. Möls (1977). *Tõenäosusteooria ja matemaatiline statistika*. Tallinn: Kirjastus Valgus, lk. 69–72.

Titterington, D. M., A. Smith ja U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons Ltd., lk. 1–2. ISBN: 0471907634.

Venter J., C., D. Adams M., W. Myers E., W. Li P., J. Mural R. *et al.* (2001). "The Sequence of the Human Genome". *Science* 291, lk. 1304–1351. DOI: 10.1126/science.1058040.

Lisa. Programmikoodid

```
library(OOR)
library(dplyr)
library(rje)
library(Metrics)



fun <- function(parameetrid){
  väärtused <- segujaotus
  osakaal <- expit(parameetrid[1])
  lambda1 <- exp(parameetrid[2])
  lambda2 <- exp(parameetrid[3])
  summa <- sum(log((1-osakaal)*dpois(väärtused, lambda1)+
       osakaal*dpois(väärtused, lambda2)))
  return(-summa)
}



segujaotuse_hinnang <- function(osakaal, lambda1, lambda2, n){
  k <- 1
  andmestik <- matrix(NA, nrow = n, ncol = 6)
  for(j in 1:n){
       poisson1 <- rpois(10000, lambda1)
       poisson2 <- rpois(10000, lambda2)
       vektor <- runif(10000, min=0, max=1)
       segujaotus <- rep(NA, 10000)
       for (l in 1:length(vektor)){
         vaartus <- ifelse(vektor[l] < osakaal, poisson2[l], poisson1[l])
         segujaotus[l] <- vaartus
       }
       segujaotus <<- segujaotus
       tulemused <- optim(c(logit(0.1), log(4), log(5)), fun, method
= "BFGS")
       andmestik[k,1] <- osakaal
```

```
      andmestik[k,2] <- tulemused$convergence
      andmestik[k,3] <- expit(tulemused$par[1]) # hinnatud osakaal
      andmestik[k,4] <- exp(tulemused$par[2]) #lambda1
      andmestik[k,5] <- exp(tulemused$par[3]) #lambda2
      andmestik[k,6] <- mean(segujaotus)
      k <- k+1
  }
  andmestik <- as.data.frame(andmestik)
  colnames(andmestik) <- c("tegelik osakaal","koonduvus", 'osakaal',
         'lambda1','lambda2',"keskmine")


  andmestik <- andmestik %>% mutate(lambda = ifelse(0.5 <= osakaal,
         lambda2, lambda1)) %>% rowwise() %>%
         mutate(segumudel_erinev_tegelik = abs(10-lambda),
         aritm_erinev_tegelik = abs(10-keskmine))
  return(andmestik)
}



DX <- function(osakaal, lambda1, lambda2) {
  tulem <- (1-osakaal)*lambda1 + osakaal*lambda2 +
       (osakaal*(1-osakaal)*(lambda1-lambda2)**2)
  return(tulem)
}



nihe <- function(osakaal, lambda1, lambda2){
  EX = ((1-osakaal)*lambda1 + (osakaal)*lambda2)
  return(EX - lambda1)
}
```

```
tapsustevahe <- function(osakaal, lambda1=..., lambda2=..., n=10000){
  andmestik <<- segujaotuse_hinnang(osakaal, lambda1, lambda2, n)
  koik_tulemused <<- rbind(koik_tulemused, andmestik)

  vead <<- andmestik %>% select(‘tegelik osakaal‘,
       segumudel_erinev_tegelik, aritm_erinev_tegelik) %>%
       group_by(‘tegelik osakaal‘) %>% summarise(segumudel_MSE =
       mean(segumudel_erinev_tegelik**2)) %>%
       arrange((‘tegelik osakaal‘))

  vead2 <- vead %>% mutate(MSE_tapne = DX(osakaal, lambda1, lambda2)/n
       + (nihe(osakaal, lambda1, lambda2))**2,
       vahe_hinnang = segumudel_MSE-MSE_tapne)

  vahe <- vead2$vahe_hinnang
  vahe_ruut <- vahe**2
  return(vahe_ruut)
}



koik_tulemused <- data.frame(matrix(nrow = 0, ncol = 9))
veerud <- c("tegelik osakaal","koonduvus", ösakaal","lambda1",
  "lambda2", "keskmine", "lambda", "segumudel_erinev_tegelik",
  äritm_erinev_tegelik")
colnames(koik_tulemused) = veerud



lopptulemus <- StoSOO(par = NA , fn = tapsustevahe, lower = c(...),
  upper = c(...), nb_iter = 55)
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kätlin Kippar,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Vigadega sekveneerimisandmete analüüs. Sekveneerimiskatvuse hindamine", 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.

Kätlin Kippar
15.05.2024