TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Kea Mei

Statistiline test *k*-meeride abil DNA tandemkorduse koopiaarvu määramiseks

Bakalaureusetöö

Juhendaja:

Märt Möls

TARTU

2015

Statistiline test *k*-meeride abil DNA tandemkorduse koopiaarvu määramiseks

Käesoleva töö eesmärgiks on välja töötada statistiline test, mille abil on võimalik kindlaks määrata, kas indiviidi DNA tandemkorduse koopiaarv ehk DNA ahelas järjestikku korduva osa korduste arv vastab referentsgenoomis kirjapandud korduste arvule. Kui tandemkorduse koopiaarv varieerub indiviiditi, siis on tegemist varieeruva arvuga tandemkordusega ehk VNTR-iga. Varieeruva arvuga tandemkorduste ülesleidmine võimaldab paremini kirjeldada indiviididevahelisi geneetilisi erinevusi. Samuti kasutatakse neid kriminalistikas kurjategija tuvastamiseks kuriteopaigalt leitud DNA põhjal.

Geeniandmed on tavaliselt väga suured ja mahukad, mistõttu nende töötlemine on aeglane ja kulukas. Käeosolevas töös väljatöötatud testis vaadeldakse tandemkorduse korduvat osa kui *k*-meeri (DNA sekveneeritud jupilt moodustatud väiksemat *k* nukleotiidi pikkust osa) ning teststatistiku leidmiseks loetakse kokku, mitu korda antud *k*-meeri sekveneerimisandmetes esines. Kuna *k*-meeride arvu lugemiseks on olemas kiired algoritmid, siis on ka sellisel meetodil testimiseks kuluv aeg väiksem. Töös arvutatakse teststatistiku jaotus nullhüpoteesi kehtides ning selle põhjal koostatakse statistiline test ning leitakse ka testi võimsus tuvastada tõenäoliseimaid koopiaarvu muutuseid.

Märksõnad: andmeanalüüs, DNA genoom, DNA koopiaarvu variatsioonid, DNA kordusjärjestus, genoomika, *k*-meer, lugem, statistilised meetodid, testimine.

Statistical test with *k*-mers for determining the repeat unit copy number of tandem repeat

The aim of this study is to create a statistical test, which will help determine whether the repeat unit copy number (the number of repetitive parts) of tandem repeat on a specific individual is the same number as in the reference genome. If repeat unit copy number of tandem repeat varies between individuals then tandem repeat is variable number tandem repeat (VNTR). Finding VNTRs helps to describe genetic differences between individuals. They are also used in forensics through DNA fingerprinting.

Genetic data is usually really big which is why analyzing the data is often too slow and expensive. In this test, the repeat unit of tandem repeat is regarded as *k*-mers (subsequences of length *k* from a read obtained through DNA sequencing) and test statistic is the sum of *k*-mers. Finding repeat unit copy number of tandem repeat comes down to counting k-mers with computers. There are fast algorithms to find a lot of *k*-mers and this is why testing with this method is less time-consuming. In this study the distribution of test statistic is calculated and then, on the bases of distribution of test statistic, the statistical test is created. Also, the power of this test is computed to establish the changes of the repeat unit copy number of tandem repeats that are more likely.

Keywords: data analysis, DNA copy number variaitions, DNA genome, DNA tandem repeats, genomics, k-mer, read, statistical methods, testing.

Sisukord

Sissejuhatus

Käesoleva töö eesmärgiks on välja töötada statistiline test, mille abil on võimalik kindlaks määrata, kas indiviidi DNA tandemkorduse koopiaarv ehk DNA ahelas järjestikku korduva osa korduste arv vastab referentsgenoomis kirjapandud korduste arvule. Referentsgenoom on teoreetiline genoom, mille abil on võimalik reaalseid genoome võrrelda. Test aitab leida varieeruva arvuga tandemkordusi ehk VNTR-e. Nende ülesleidmine võimaldab paremini kirjeldada indiviididevahelisi geneetilisi erinevusi, mistõttu saab neid kasutada geneetikas näiteks lapse vanemate identifitseerimiseks või kriminalistikas kurjategija tuvastamiseks kuriteopaigalt leitud DNA põhjal.

Geeniandmed on tavaliselt väga suured ja mahukad, mistõttu nende töötlemine on aeglane ja kulukas. Käeosolevas töös väljatöötatud testis vaadeldakse tandemkorduse korduvat osa kui *k*-meeri (DNA sekveneeritud jupilt moodustatud väiksemat *k* nukleotiidi pikkust osa) ning teststatistiku leidmiseks loetakse kokku, mitu korda antud *k*-meeri sekveneerimisandmetes esines. Kuna *k*-meeride arvu lugemiseks on olemas kiired algoritmid, siis on ka sellisel meetodil testimiseks kuluv aeg väiksem.

Töö koosneb kahest osast – teoreetilisest ning praktilisest. Teoreetilises osas tutvustatakse töös kasutatud geneetikaalaseid mõisteid, et töö lugejale arusaadavamaks muuta. Töö praktilise osa võib jaotada kaheks.

Praktilise osa esimeses osas arvutatakse teststatistiku väärtus nullhüpoteesi kehtides ning koostatakse selle põhjal statistiline test DNA tandemkorduse koopiaarvu määramiseks. Praktilise osa teises osas leitakse testi võimsus tuvastata tõenäolisemaid koopiaarvu muutuseid ning testitakse näiteinimesi Tartu Ülikooli Eesti Geenivaramust saadud andmete põhjal.

Bakalaureuse töö on kirjutatud tekstitöötlusprogrammiga Microsoft Word 2013. Statistiline test ning joonised on koostatud rakendustarkvara paketi R 3.1.3 abil. Töös kasutatud allikatele on viidatud nurksulgude abil.

Autor tänab südamest suure abi ja väärtuslike nõuannete eest juhendajat Märt Mölsi ning Tartu Ülikooli Eesti geenivaramu töötajaid Tarmo Puurandi ja Ulvi Gerst Talasit töös vajalike näiteandmete hankimise eest.

1. Töös kasutatud geneetikamõistete tutvustus

Genoom on organismi geneetilise informatsiooni terviklik koopia [1]. Genoomi iseloomustatakse DNA nukleotiidjärjestuse kaudu, mis koosneb lämmastikalustest (A – adeniin, T – tümiin, G – guaniin, C – tsütosiin) [2].

Sekveneerimine ehk järjendamine on nukleotiidide järjestuse kindlaksmääramine DNA molekulides [3]. DNA sekveneerimine annab tulemuseks neljast erinevast nukleiinhappejäägist (lühendatult A, T, G, C) koosneva tõlgenduse, mida nimetatakse DNA järjestuseks ehk sekventsiks. Sekventsi **katvus** näitab, mitu korda sekveneerimise käigus keskmiselt ühte nukleotiidi loetakse. DNA järjestus hoiab endas organismide ellujäämiseks ja paljunemiseks vajalikku informatsiooni, seega on DNA järjestuse teadmine eriti oluline genoomikas [1].

DNA sekveneerimine on kasutatav meditsiinis näiteks haiguste diagnoosimiseks ja ka personaalse ravi väljatöötamiseks inimese DNA põhjal. Samuti on DNA järjestuse põhjal võimalik kirjeldada inimese teatud väliseid kriteeriume ilma inimest nägemata, näiteks loote DNA järgi öelda, milline on sündiva inimese silma- ja juuksevärv või kui pikaks ta võib kasvada. DNA järjestuse võrdlemine on kasulik ka kriminaalteaduses. Kui näiteks võrrelda kuriteopaika jäetud DNA järjestust kindla kahtlusaluse DNA järjestusega, on võimalik öelda, kas kahtlusalune oli kuriteopaigas. DNA sekveneerimise arendamisel loodetakse tulevikus DNA põhjal paika panna näiteks kurjategija väline mudel koos tema näojoonte, juuste värvi ja kuju ning inimese pikkusega, ka juhul, kui ühtegi kahtlusalust ei ole. [5]

Lugem (ingl. *read*) on DNA üks sekveneeritud jupp. Tänapäeva sekvenaatorid on võimelised ühe masina jooksutamisega tekitama miljardeid lugemeid.

Näiteks juhul, kui 100 nukleotiidi pikkusest DNA molekuli ahelast määratakse sekveneerimisel esimesed 8 nukleotiidi (olgu nende järjestus näiteks AGTTAGCC), siis see 8 aluspaari sellises järjestuses ongi üks lugem. Tähistagu n lugemi pikkust (ehk siin näites n=8).

Lugemist omakorda saab moodustada k nukleotiidi pikkuseid lõike ehk ***k*-meere** (ingl. *k-mer*). Ühelt lugemilt erinevate k-meeride võimalik moodustamise arv K on leitav järgmiselt:

$$K = n - k + 1.$$

Näiteks on võimalik eelmises näites 8 nukleotiidist koosnevalt lugemilt AGTTAGCC moodustada

$$K = 8 - 4 + 1 = 5$$

4-meeri järgmiselt: AGTT, GTTA, TTAG, TAGC ja AGCC.

Tandemkordused on mitme aluspaari pikkused üksteisele järgnevad nukleotiidide järjestuse kordused DNA ahelas [6]. Teisiti öeldes koosneb tandemkordus lühikestest samasuguste *k*-meeride üksteisele järgnevatest kordustest ($k \geq 2$).

Tandemkorduse näitena võib vaadata DNA ahelas lõiku ATTCGATTCGATTCG, kus 5 nukleotiidist koosnev järjestus ATTCG kordub 3 korda. **Tandemkorduse koopiaarv** (ehk korduste arv) on tandemkorduse korduva osa (selles näites lõigu ATTCG) korduste arv. Seega antud näites tandemkorduse koopiaarv on 3.

Kui tandemkorduse koopiaarv (ehk tandemkorduses korduva *k*-meeri korduste arv) on individuaalselt erinev, siis on tegu **varieeruva arvuga tandemkordusega (VNTR - variable number tandem repeat)** [6].

VNTR-i näitena võib vaadata DNA ahelas lõiku ATTCG, mis ühel indiviidil kordub DNA aheldas näiteks 3 korda järjest (ATTCGATTCGATTCG), kuid teisel indiviidil ainult 2 korda järjest (ATTCGATTCG).

Kuna tänapäeval on mitu genoomi juba järjestatud, siis on teada ka varieeruva arvuga tandemkorduseid. Teadaolevate varieeruva pikkusega tandemkorduste abil on aga võimalik prognoosida indiviididevahelisi geneetilisi erinevusi, mistõttu saab neid kasutada näiteks lapse vanemate identifitseerimiseks või kriminalistikas kurjategija tuvastamiseks kuriteopaigalt leitud DNA põhjal. Samuti võivad mõned varieeruva arvuga tandemkordused mõjutada haigestumist.

2. Summaarse *k*-meeride arvu jaotus

Genoomist on võimalik moodustada erinevaid lugemeid, mis koosnevad nukleotiidide järjestusest. Lugemist omakorda saab moodustada erinevaid lõike, pikkusega *k* (ehk *k*-meere).

Kuna katvus näitab, mitu korda sekveneerimise käigus keskmiselt ühte nukleotiidi loetakse, siis kehtib:

$$katvus = \frac{r \times n}{N},$$

kus *r* tähistab lugemite arvu ühel genoomil, *n* tähistab lugemi pikkust ja *N* genoomi pikkust. Seega lugemite arvu ühel genoomil on võimalik arvutada järgmiselt:

$$r = \frac{katvus \times N}{n}$$

Lugemi alguspunktiks loetakse kohti genoomil, kust lugem algab. Leidmaks, mitu korda keskmiselt hakkas lugem ühest alguspunktist (tähis: λ), tuleb lugemite arv ühel genoomil jagada genoomi kogupikkusega ehk:

$$\lambda = \frac{r}{N} = \frac{katvus \times N}{n \times N} = \frac{katvus}{n}.$$

Juhul, kui meile huvipakkuv tandemkorduse koopiaarv on 2 ning lugemi pikkus on suurem kui tandemkorduse kogupikkus, siis lugemite alguspunkte, mille korral sekveneerimise tulemusena saadud lugem sisaldab ainult ühte kordust ehk meid huvitav *k*-meeri (tähis: n_1), on kokku

$$n_1 = 2k \text{ tükki,}$$

kus *k* tähistab korduva *k*-meeri pikkust (ehk korduse pikkust) ning alguspunkte, mille korral sekveneerimise tulemusena saadud lugem sisaldab 2 kordust ehk 2 samasugust *k*-meeri järjest (tähis: n_2), on kokku

$$n_2 = n - 2k + 1 \text{ tükki.}$$

Tähistagu X_1 nende lugemite arvu, milles meid huvitav *k*-meer esineb vaid ühel korral. Juhuslik suurus X_1 võiks olla ligilähedaselt Poissoni jaotusega, sest tulemus on saadud loendamise teel, X_1 väärtuste hulk on mittenegatiivsed täisarvud, katsete arv on suur (lugemeid on palju) ning tõenäosus kohata ühte fikseeritud kordust on väike.

Selle kontrollimiseks on uuritud Tartu Ülikooli Eesti geenivaramust saadud näiteandmeid bakteriofaagi (bakteriviiruse) ΦX174 kohta, kes oli esimene organism maailmas, kelle DNA genoom sekveneeriti. Andmetes on toodud selle organismi kohta sekveneerimise käigus iga nukleotiidi korral seda läbinud lugemite arvud kahe erineva katvuse korral. Andmete põhjal on koostatud jaotusfunktsioonide graafikud ning võrreldud neid vastava Poissoni jaotuse graafikuga.

Graafikud ning nende jaoks kasutatud R koodid paiknevad lisades (lisad 1 ja 2). Antud graafikute põhjal võib öelda, et lugemite arvu tegelik jaotus meenutab mõnevõrra Poissoni jaotust.

Et kinnitada veelgi juhusliku suuruse X_1 puhul Poissoni jaotuse kasutamise õigsust, on uuritud sekveneerimisega tegelevate teadlaste seisukohti. Sekveneerimismasinaid tootvas ettevõttes Illumina töötavate teadlaste arvates võiks ühte nukleotiidi tabanud lugemite arv olla Poissoni jaotusega [7]. Seega võime kasutada X_1 jaotusena Poissoni jaotust.

Tähistagu X_2 nende lugemite arvu, milles meid huvitav *k*-meer esineb kahel korral. Analoogselt X_1-ga võib öelda, et ka juhuslik suurus X_2 on ligilähedaselt Poissoni jaotusega.

Kuna Poissoni jaotuse parameeter on võrdne keskväärtusega, siis võib öelda, et $X_1 \sim \text{Po}(\lambda_1 = \text{E}(X_1))$ ja $X_2 \sim \text{Po}(\lambda_2 = \text{E}(X_2))$, kus keskväärtused λ_1 ja λ_2 on arvutatavad järgmiselt:

$$\lambda_1 = \lambda \times n_1 = \lambda(2k) = \frac{2k}{n} katvus \text{ ja}$$

$$\lambda_2 = \lambda \times n_2 = \lambda(n - 2k + 1) = \frac{(n-2k+1)}{n} katvus .$$

Tähistagu nüüd juhuslik suurus X summaarset otsitavate *k*-meeride arvu. Kuna X_2 korral loetakse igal lugemil 2 kordust (seega 2 *k*-meeri), siis $X = X_1 + 2X_2$. Kuna juhuslik suurus X ei ole enam Poissoni jaotusega, siis on huvipakkuvaks küsimuseks X-i jaotus.

Tabelis 2.1 on toodud juhusliku suuruse X väärtused erinevate X_1 ja X_2 väärtuste korral. Näiteks, kui $X_1 = 0$ ja $X_2 = 1$, siis $X = X_1 + 2X_2 = 0 + 2 \times 1 = 2$.

Tabel 2.1. X-i võimalikud väärtused fikseeritud X_1 ja X_2 korral.

		X_2						
	X	**0**	**1**	**2**	**3**	**4**	**5**	**…**
	0	0	2	4	6	8	10	…
	1	1	3	5	7	9	11	…
	2	2	4	6	8	10	12	…
X_1	**3**	3	5	7	9	11	13	…
	4	4	6	8	10	12	14	…
	5	5	7	9	11	13	15	…
	…	…	…	…	…	…	…	…

Eelneva tabeli põhjal saab kirja panna järgmised seosed:

$P(\{X=0\}) = P(\{X_1=0\} \cap \{X_2=0\})$

$P(\{X=1\}) = P(\{X_1=1\} \cap \{X_2=0\})$

$P(\{X=2\}) = P[(\{X_1=0\} \cap \{X_2=1\}) \cup (\{X_1=2\} \cap \{X_2=0\})]$

$P(\{X=3\}) = P[(\{X_1=1\} \cap \{X_2=1\}) \cup (\{X_1=3\} \cap \{X_2=0\})]$

$P(\{X=4\}) = P[(\{X_1=0\} \cap \{X_2=2\}) \cup (\{X_1=2\} \cap \{X_2=1\}) \cup (\{X_1=4\} \cap \{X_2=0\})]$

$P(\{X=5\}) = P[(\{X_1=1\} \cap \{X_2=2\}) \cup (\{X_1=3\} \cap \{X_2=1\}) \cup (\{X_1=5\} \cap \{X_2=0\})]$

Üldine valem juhusliku suuruse X tõenäosuste leidmiseks on seega:

$$P(\{X = \mathrm{x}\}) = P\left[\bigcup_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} (\{X_1 = \mathrm{x} - 2\mathrm{i}\} \cap \{X_2 = \mathrm{i}\})\right], \qquad (2.1)$$

kus $\mathrm{int}\left(\frac{x}{2}\right)$ tähistab murru $\frac{x}{2}$ täisosa. Seega, näiteks, kui $\frac{x}{2} = 2{,}5$, siis $\mathrm{int}\left(\frac{x}{2}\right) = 2$.

Kasutades tõenäosuse aditiivsuse omadust, saame valemi (2.1) teisendada järgmisele kujule:

$$P(\{X = \mathrm{x}\}) = \sum_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} P(\{X_1 = \mathrm{x} - 2\mathrm{i}\} \cap \{X_2 = \mathrm{i}\}). \qquad (2.2)$$

Kuna lugemeid genereeritakse sõltumatult, on ka X_1 ja X_2 omavahel sõltumatud ning kehtib:

$$P(\{X = \mathrm{x}\}) = \sum_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} [P(\{X_1 = \mathrm{x} - 2\mathrm{i}\}) \times P(\{X_2 = \mathrm{i}\})]. \qquad (2.3)$$

Kuna $X_1 \sim \mathrm{Po}(\lambda_1)$ ja $X_2 \sim \mathrm{Po}(\lambda_2)$ ning Poissoni jaotusega juhusliku suuruse $Y \sim \mathrm{Po}(\lambda)$ puhul kehtib:

$$\mathrm{P}(\{Y = \mathrm{y}\}) = \frac{\lambda^y}{y!} e^{-\lambda},$$

siis seda ära kasutades on võimalik valem (2.3) teisendada järgmisele kujule:

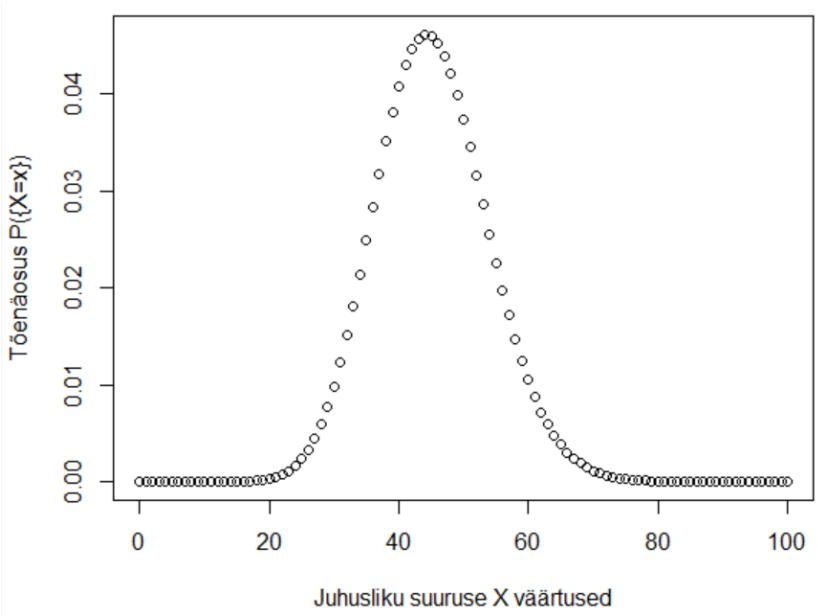
$$\mathrm{P}(\{X = \mathrm{x}\}) = \sum_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} \left(\frac{{\lambda_1}^{x-2i}}{(x-2i)!} e^{-\lambda_1} \times \frac{{\lambda_2}^{i}}{i!} e^{-\lambda_2} \right) = e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} \left(\frac{{\lambda_1}^{(x-2i)} {\lambda_2}^{i}}{(x-2i)!\, i!} \right)$$

Seega oleme valemist (2.1) saanud juhusliku suuruse X tõenäosusfunktsiooni kujul:

$$\mathrm{P}(\{X = \mathrm{x}\}) = e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{\mathrm{int}\left(\frac{x}{2}\right)} \left(\frac{{\lambda_1}^{(x-2i)} {\lambda_2}^{i}}{(x-2i)!\, i!} \right), \qquad (2.4)$$

kus $\lambda_1 = \frac{2k}{n} katvus$ ja $\lambda_2 = \frac{(n-2k+1)}{n} katvus$.

Statistikapaketiga R valemi (2.4) abil moodustatud X-i tõenäosusfunktsiooni graafikut etteantud lugemi pikkuse n, korduva k-meeri pikkuse k ja *katvuse* korral (mille koostamiseks vajalik kood on lisas 3) kirjeldab joonis 2.1.



Joonis 2.1 X-i tõenäosusfunktsiooni graafik, kui n = 101, k = 26, *katvus* = 30.

3. DNA tandemkorduse koopiaarvu testimine

3.1. Test kahe korduse jaoks

Test kontrollib hüpoteeside paari:

- H_0: Indiviidi DNA tandemkorduse koopiaarv on 2.
- H_1: Indiviidi DNA tandemkorduse koopiaarv ei ole 2.

Kuna test uurib tandemkorduse koopiaarvu ning tandemkordus on esitatav *k*-meeride järjestikuste kordustena, on tandemkorduse koopiaarv võrdne meile huvipakkuva *k*-meeri korduste arvuga. Seepärast kasutame teststatistikuna eelnevalt defineeritud juhuslikku suurust X, mis tähistab summaarstet *k*-meeride arvu indiviidil ning mille saab korduste arvu 2 korral lahti kirjutada Poissoni jaotusega juhuslike suuruste X_1 ja X_2 abil: $X = X_1 + 2X_2$.

Teststatistik X mõõdab erinevust nullhüpoteesis väidetu ja andmetest ilmneva vahel – kui erinevus on piisavalt suur, kummutatakse nullhüpotees. Lisaks on eelmisest peatükist juba teada tõenäosusfunktsioon nullhüpoteesi korral (valem 2.4).

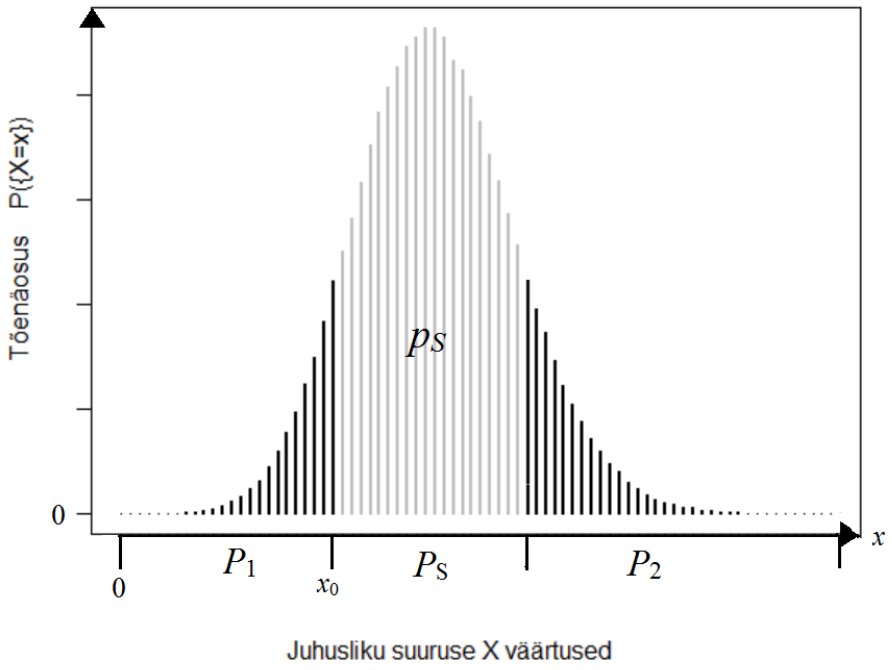
Testi olulisuse tõenäosus p on tõenäosus näha teststatistiku väärtusega $X = x_0$ sama ekstreemset või veel ekstreemsemat (harvemini esinevat) X väärtust. Olulisuse tõenäosuse põhjal saab otsustada, kas konkreetse indiviidi DNA tandemkorduse koopiaarv on 2, nagu seda näeb ette referentsgenoom, või mitte. Selleks on vaja valida sobiv olulisuse nivoo α. Olulisuse nivoo näitab maksimaalset lubatud esimest liiki vea tõenäosust. Esimest liiki viga tekib siis, kui võetakse vastu sisukas hüpotees H_1, aga tegelikult on õige nullhüpotees H_0. Seega valitakse olulisuse nivooks võimalikult väike tõenäosus. Hindamiskriteerium antud testi puhul on järgmine:

- Kui $p < \alpha$, siis võetakse vastu sisukas hüpotees H_1. Seega võib öelda, et indiviidi DNA tandemkorduse koopiaarv ei ole 2.
- Kui $p \geq \alpha$, siis ei ole võimalik tõestada, et indiviidi DNA tandemkorduse koopiaarv ei oleks 2, mistõttu tuleb jääda nullhüpoteesi H_0 juurde, et indiviidi DNA tandemkorduse koopiaarv on 2.

Joonisel 3.1 on X-i tõenäosusfunktsiooni graafikule märgitud võimalik teststatistiku väärtuse x_0 sattumise koht. Teststatistiku väärtusest x_0 sama harva või veel harvemini kohatavad X väärtused (ehk väärtused x, mille korral $\mathrm{P}(\{X = x\}) \leq \mathrm{P}(\{X = x_0\})$) jäävad graafikul

piirkondadesse P_1 ja P_2 ning nende esinemistõenäosused P($\{X = x\}$) on joonisel 3.1 toodud mustade tulpadena.

Ülejäänud piirkond P_S tähistab selliseid väärtusi x, mille korral $\mathrm{P}(\{X = x\}) > \mathrm{P}(\{X = x_0\})$ (ehk x_0-st sagedamini esinevaid X väärtusi, mille esinemistõenäosused joonisel 3.1 on esitatud hallide tulpadena. Olgu p_S teststatistiku X väärtuste sattumise tõenäosus piirkonda P_S. Seega on p_S kõigi hallide tulpade kõrguste summa.



Joonis 3.1 X-i tõenäosusfunktsiooni graafik testi koostamisel.

Kasutades nüüd omadust, et tõenäosusfunktsiooni kõigi tõenäosuste summa (seega joonisel kõigi tulpade kõrguste summa) on võrdne 1-ga, siis on piirkonda $P = P_1 + P_2$ sattumise tõenäosus p avaldatav järgmiselt:

$$p = 1 - p_S \qquad (3.1)$$

Testi olulisuse tõenäosuse leidmiseks tuleb järelikult leida teststatistiku väärtusest x_0 sagedamini esinevate X väärtuste esinemistõenäosused (ehk hallide tulpade kõrgused) ning need kokku liita. Nii saadakse p_S ehk tõenäosus näha teststatistiku väärtusega x_0 sama tihti või veel sagedamini esinevat X väärtust. Kuna aga p on tõenäosus näha teststatistiku väärtusega x_0 sama ekstreemset või veel ekstreemsemat (harvemini esinevat) X väärtust, siis saab p leidmiseks kasutada vastandtõenäosust nii nagu on kirjas valmis (3.1). Täpselt seda ongi koostatud DNA tandemkorduse koopiaarvu testis olulisuse tõenäosuse p arvutamiseks tehtud.

Testile tuleb ette anda konkreetsel indiviidil nähtud summaarne k-meeride arv, sekveneerimisel kasutatud lugemi pikkus, *k*-meeri pikkus ning sekventsi katvus. Test väljastab kõik etteantud parameetrid ning lisaks hüpoteeside paari, testi olulisuse tõenäosuse *p* ning otsustuskriteeriumi, millal võtta vastu sisukat hüpoteesi H_1.

Test on koostatud statistikaprogrammiga R. Lisades on toodud testi R kood (lisa 4) ja testi väljund (lisa 5).

3.2. Üldistatud test DNA tandemkorduse koopiaarvu määramiseks

Eelnevalt vaatasime juhtu, kui korduse arv oli 2 ning kontrolliti hüpoteesipaari, kas indiviidi DNA tandemkorduse koopiaarv on 2 või mitte. Antud testi on võimalik ka üldistada iga korduse arvu jaoks. Seega tuleb testi muuta nii, et kasutaja saaks testile ise parameetrina ette anda korduse arvu, mida soovitakse testida. Kui tähistada indiiviidi DNA tandemkorduse koopiaarv parameetriga Θ, siis uus üldistatud test kontrollib hüpoteesipaari:

- H_0: Indiviidi DNA tandemkorduse koopiaarv on θ ($\Theta = \theta$).
- H_1: Indiviidi DNA tandemkorduse koopiaarv ei ole θ ($\Theta \neq \theta$).

Selleks tuleb leida teststatistiku jaotus nullhüpoteesi kehtides. Kui 2 korduse korral oli teststatistik defineeritud kui $X = X_1 + 2X_2$, siis 3 korduse jaoks tuleks teststatistik järgmine:

$$X = X_1 + 2X_2 + 3X_3,$$

kus X_1 tähistab nende lugemite arvu, milles meid huvitav *k*-meer esineb vaid ühel korral, X_2 tähistab nende lugemite arvu, milles meid huvitav *k*-meer esineb täpselt kahel korral ning X_3 tähistab nende lugemite arvu, milles meid huvitav *k*-meer esineb kolmel korral. Eeldame siinjuures, et uuritav kordus on piisavalt lühike, et esineb vähemalt võimalus paigutada lugemit nii, et kõik 3 kordust mahuvad ühte lugemisse. Analoogselt eelnevale arutluskäigule võib öelda, et juhuslikud suurused X_1, X_2 ja X_3 on ligilähedaselt Poissoni jaotustega, parameetritega vastavalt λ_1, λ_2 ja λ_3.

Tähistagu λ endiselt seda, mitu korda keskmiselt hakkas lugem ühest alguspunktist ning olgu n_i lugemite alguspunktide arv, mille korral sekveneerimise tulemusena saadud lugem sisaldab *i* kordust.

Juhul, kui meile huvipakkuv tandemkorduse koopiaarv on 3 (järjestikuseid *k*-meere tandemkorduses on 3) ning lugemi pikkus on suurem kui tandemkorduse kogupikkus, siis $n_1 = 2k, n_2 = 2k$ ja $n_3 = n - 3k + 1$ ning $\lambda_1 = \lambda_2 = \lambda \times n_1 = \lambda(2k) = \frac{2k}{n} katvus$ ja $\lambda_2 = \lambda \times n_3 = \lambda(n - 3k + 1) = \frac{(n-3k+1)}{n} katvus$.

Seega üldistavalt θ korduse puhul on teststatistik *X* kujul:

$$X = X_1 + 2X_2 + 3X_3 + \cdots + \theta X_\theta = \sum_{i=1}^{\theta} i \times X_i$$

ning kehtivad seosed:

- $n_i = 2k$, kui $i = 1, \ldots, \theta - 1$
- $n_\theta = n - \theta k + 1$.

Seega $X_1, X_1, \ldots, X_{\theta\text{-}1} \sim Po\left(\lambda_1 = \lambda(2k) = \frac{2k}{n} katvus\right)$ ja

$X_\theta \sim Po\left(\lambda_2 = \lambda(n - \theta k + 1) = \frac{(n-\theta k+1)}{n} katvus\right)$.

Teades X_1 ja X_2 diskreetseid jaotusfunktsioone, on konvolutsiooni meetodi abil võimalik välja arvutada nende summa jaotus $X = X_1 + X_2$. Olgu X_1 jaotusfunktsioon *f(x)* ja X_2 jaotusfunktsioon *g(x)*, siis:

$$\mathrm{P}(\{X = \mathrm{x}\}) = \sum_{i=-\infty}^{\infty} f(i)g(x - i).$$

Seega teades X_1 ja $2X_2$ jaotusfunktsioone, on konvolutsiooni abil võimalik välja arvutada $X = X_1 + 2X_2$ diskreetne jaotusfunktsioon ning seda teades on võimalik edasi arvutada $X = (X_1 + 2X_2) + 3X_3$ jaotusfunktsioon. Sellist tsüklit läbides on võimalik leida teststatistiku *X* jaotusfunktsioon mistahes korduse korral ning selle abil analoogselt eelmise testiga koostada üldise testi, millele saab lisaks ette anda soovitud korduse arvu.

Lisades on toodud R kood (lisa 6) ja väljund (lisa 7) üldise DNA tandemkorduse koopiaarvu testile, mille korral saab ette anda korduste arvu 1, 2 või 3.

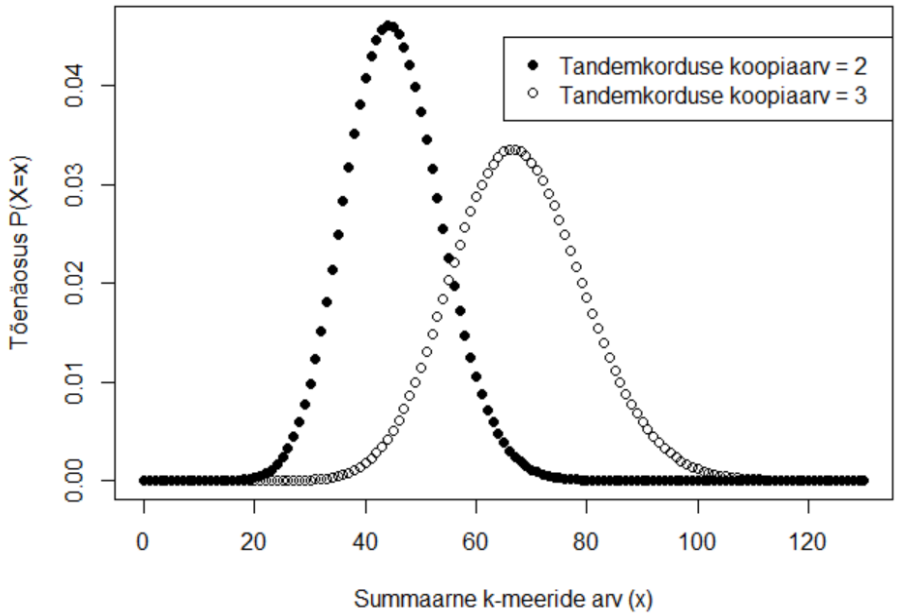
4. Testi rakendamisest praktikas

4.1. Testi võimsusest

Teist liiki viga β tehakse siis, kui testi tulemusel jäädakse nullhüpoteesi H_0 juurde, kuid tegelikult kehtib alternatiivne hüpotees H_1. Testi võimsus $1 - \beta$ on tõenäosus lugeda õigeks ka tegelikult kehtiv alternatiivne hüpotees H_1. Seega, mida suurem on testi võimsus, seda harvemini tehakse teist liiki viga. [8]

Joonis 4.1 kirjeldab teststatistiku jaotuse erinevusi korduste arvu 2 ja 3 korral. Kuna meil on ettekujutus, kuidas kahe erineva tandemkorduse koopiaarvu korral indiviidide summaarsete *k*-meeride arvu jaotused üksteisest erinevad, siis on kõige lihtsam testi võimsust hinnata arvutisimulatsiooni abil.

Selleks tuleb genereerida mingi arv juhuslikke suurusi esimesest jaotusest ning mingi arv juhuslikke suurusi teisest jaotusest, ning vaatata, kui paljude genereeritud juhuslike suuruste korral suutis meie test vastu võtta alternatiivse hüpoteesi.



Joonis 4.1. Summaarne *k*-meeride arvu (ehk teststatistiku) jaotus korduste arvu 2 ja 3 korral (lugemi pikkus *n*=101, *k*-meeri pikkus *k*=26 ning *katvus*=30).

Leidmaks, kui suure tõenäosusega meie test suudab tuvastada seda, kui indiviidil on 2 korduse (st indiviid on pärinud kummaltki vanemalt 2 kordust) asemel 3 kordust (st indiviid on pärinud kummaltki vanemalt 3 kordust), on genereeritud 10 000 teststatistikut $X = X_1 +$

$2X_2 + 3X_3$ jaotusest (jaotusfunktsioon 3 korduse korral). Seejärel on kokku loetud, mitme genereeritud teststatistiku puhul võttis test vastu alternatiivse hüpoteesi usaldusnivoo 0,05 korral. Saadud tulemus on jagatud teststatistikute arvuga (10 000-ga). Tabelis 4.1 on selline olukord ära toodud 3. reas, kus korduste arv on 3 x 3 (ehk mõlemalt vanemalt on saadud 3 kordust). Samuti on tabelis ära toodud testi võimsused olukordades, kus mõlemalt vanemalt on päritud kas 1 kordus või 2 kordust ning olukorrad, kus kummaltki vanemalt on päritud erinev arv kordusi.

Erineva arvu korduste korral on võimsuse arvutamiseks genereeritud 10 000 juhuslikku suurust jaotusest, kus korduste arv on võrdne ühelt vanemalt päritud korduste arvuga ning 10 000 juhuslikku suurust jaotusest, kus korduste arv on võrdne teiselt vanemalt päritud korduste arvuga, kusjuures katvus kummagi jaotuse korral on poole väiksem. Ehk olukorras, kus indiviid on ühelt vanemalt pärinud 2 kordust ja teiselt vanemalt 3 kordust ning soovitakse testi võimsust 30-kordse katvuse korral, on genereeritud 10 000 juhuslikku suurust $X = X_1 + 2X_2$ jaotusest (korduste arv on 2, katvus on 15) ja 10 000 juhuslikku suurust $X = X_1 + 2X_2 + 3X_3$ jaotusest (korduste arv on 3, katvus on 15). Vastavad juhuslikud suurused on kokku liidetud (üks suurus esimesest jaotusest on liidetud teisele suurusele teisest jaotusest ning nii kõigil 10 000 korral) ja saadud 10 000 teststatistikut. Seejärel on kokku loetud, mitme teststatistiku puhul neist võttis test vastu alternatiivse hüpoteesi usaldusnivoo 0,05 korral. Saadud tulemus on jagatud teststatistikute arvuga (10 000-ga).

Tabelis toodud võimsuste arvutamiseks kasutatud R koodid on lisas 8.

Tabel 4.1. Testi võimsus katvuste 30 ja 60 korral. Korduse arv *a* x *b* tähitsab olukorda, kus indiviid päris ühelt vanemalt *a* kordust ja teiselt vanemalt *b* kordust. (Lugemi pikkus *n*=101, *k*-meeri pikkus *k*=26).

Korduste arv	**Võimsus (katvus=30)**	**Võimsus (katvus=60)**
1 x 1	0,8896	0.9987
2 x 2	0,0475	0,0492
3 x 3	0,6611	0,8949
1 x 2	0,2293	0,4584
1 x 3	0,0578	0,0611
2 x 3	0,2758	0,4505

Tabelis toodud võimsused näitavad tõenäosust, et test võtab vastu ka tegelikult kehtiva alternatiivse hüpoteesi H_1 (indiviidil ei ole 2 kordust). Tabeli reas, kus korduste arv on 2 x 2 (ehk olukorras kus indiviid on pärinud kummaltki vanemalt 2 kordust), kehtib tegelikult nullhüpotees H_0 ning seega näitab selles olukorras võimsus tegelikult tõenäosust võtta vastu H_1, kui tegelikult kehtib H_0 (ehk esimest liiki vea tegemise tõenäosust).

Tabelist on näha, et kui indiviidi DNA sekveneerimisel kasutada suuremat katvust, siis testi võimsus suureneb.

4.2. Näiteinimeste testimine geenivaramu andmete põhjal

Testi reaalseks kasutamiseks on Tartu Ülikooli Eesti geenivaramust saadud reaalsed andmed 35 eestlase kohta. Uuritud on DNA tandemkordust, milles 26 nukleotiidist koosnev 26-meer (mille algus on CAATTATAGGAAAGCCAGTCAAAAG...) kordub referentsgenoomi järgi järjest 2.076923 korda ning inimese genoomis teist täpselt sellist nukleotiidide järjestust ei esine. Iga indiviidi kohta on teada nende DNA sekveneerimise käigus kohatud summaarne 26-meeride arv. Seega on teststatistiku X väärtus andmetest saadud summaarne 26-meeride arv. Sekveneerides kasutati lugemeid pikkusega n=101 ja katvust 30.

Et testida kõiki 35 indiviidi korraga, on koostatud R-is funktsioon, mis võtab ette vektori summaarsete k-meeride arvude väärtustega, olulisuse nivoo α, lugemite pikkuse n, k-meeri pikkuse k ja katvuse ning väljastab nende andmete põhjal inimeste arvu, kelle korral suudeti olulisuse nivool α tõestada alternatiivne hüpotees H_1 (indiviidi DNA tandemkorduse koopiaarv ei ole 2). Testimiseks kasutatud R kood on toodud lisas 9.

Antud testis selgus, et olulisuse nivoo α = 0,05 korral suudeti 21 inimese puhul 35-st tõestada, et nende DNA tandemkorduse koopiaarv ei ole 2. Ülejäänud 14 inimese korral oli testi olulisuse tõenäosus $p \geq 0{,}5$ mistõttu jääme nullhüpoteesi juurde, et neil inimestel võib DNA tandemkorduse koopiaarv olla 2.

Rangema olulisuse nivoo α = 0,01 korral suudeti 14 inimese puhul 35-st tõestada, et nende DNA tandemkorduse koopiaarv ei ole 2. Ülejäänud 21 inimese korral oli testi olulisuse tõenäosus $p \geq 0{,}01$ mistõttu jääme nullhüpoteesi juurde, et neil inimestel võib DNA tandemkorduse koopiaarv olla 2.

Veelgi rangema olulisuse nivoo $\alpha = \frac{0{,}05}{35} \approx 0{,}0014$ (Bonferroni meetodi) korral suutis test 8 inimese puhul 35-st tõestada, et nende DNA tandemkorduse koopiaarv ei ole 2. Ülejäänud 27 inimese korral oli testi olulisuse tõenäosus $p \geq \frac{0{,}05}{35}$ mistõttu jääme nullhüpoteesi juurde, et neil inimestel võib DNA tandemkorduse koopiaarv olla 2.

Seega oleme näidanud, et antud tandemkorduse puhul on tegemist varieeruva pikkusega tandemkordusega, sest ka range olulisuse nivoo korral vähemalt osadel eestlastel suudeti tõestada, et nende tandemkorduse koopiaarv ei ole 2, mistõttu on suudetud tõestada, et vähemalt osade inimeste puhul ei ole antud tandemkorduse korduste arv selline nagu väidab selle olevat referentsgenoom.

Kokkuvõte

Käesoleva töö eesmärgiks oli välja töötada statistiline test, mille abil oleks võimalik kindlaks määrata, kas indiviidi DNA tandemkorduse koopiaarv ehk DNA ahelas järjestikku korduva osa korduste arv vastab teoreetilises referentsgenoomis kirjapandud korduste arvule.

Sekveneerimine on nukleotiidide järjestuse kindlaksmääramine DNA molekulides. DNA sekveneerimine on kasutatav meditsiinis näiteks haiguste diagnoosimiseks ja ka personaalse ravi väljatöötamiseks inimese DNA põhjal. Samuti on DNA järjestuse põhjal võimalik kirjeldada inimese teatud väliseid kriteeriume ilma inimest nägemata, näiteks loote DNA järgi öelda, milline on sündiva inimese silma- ja juuksevärv või kui pikaks ta võib kasvada. DNA järjestuse võrdlemine on kasulik ka kriminaalteaduses. Kui näiteks võrrelda kuriteopaika jäetud DNA järjestust kindla kahtlusaluse DNA järjestusega, on võimalik öelda, kas kahtlusalune oli kuriteopaigas.

Kui tandemkorduse koopiaarv varieerub indiviiditi, siis on tegemist varieeruva arvuga tandemkordusega ehk VNTR-iga. Varieeruva arvuga tandemkorduste ülesleidmise abil on võimalik paremini kirjeldada indiviididevahelisi geneetilisi erinevusi. Näiteks kasutatakse neid kriminalistikas kurjategija tuvastamiseks kuriteopaigalt leitud DNA põhjal. Samuti võivad mõned varieeruva arvuga tandemkordused mõjutada haigestumist.

DNA sekveneeritud jupilt moodustatud väiksemat *k* nukleotiidi pikkust osa nimetatakse *k*-meeriks. Geeniandmed on tavaliselt väga suured ja mahukad, mistõttu nende töötlemine on aeglane ja kulukas. Käeosolevas töös väljatöötatud testis vaadeldi tandemkorduse korduvat osa kui *k*-meeri ning teststatistiku leidmiseks loeti kokku, mitu korda antud *k*-meeri sekveneerimisandmetes esines. Kuna *k*-meeride arvu lugemiseks on olemas kiired algoritmid, siis on ka sellisel meetodil testimiseks kuluv aeg väiksem.

Testi koostamine õnnestus. Statistikapaketi R abil koostati statistiline test DNA tandemkorduse koopiaarvu määramiseks koopiaarvu 2 korral ning üldine statistiline test DNA tandemkorduse koopiaarvu määramiseks testija vabal valikul kas 1, 2 või 3 koopiaarvu määramiseks. Esimese testi jaoks arvutati testi võimsus tuvastata tõenäolisemaid koopiaarvu muutuseid. Selgus, et kui indiviidi DNA sekveneerimisel kasutada suuremat katvust, siis testi võimsus suureneb.

Lõpuks testiti koostatud statistilise testi abil näiteinimesi Tartu Ülikooli geenivaramust saadud andmete põhjal. Referentsgenoomi põhjal oleks pidanud nende andmete tulemusena olema

tandemkorduse koopiaarvuks 2. Koostatud test suutis ka rangel olulisuse nivool $0{,}0014$ tõestada alternatiivse hüpoteesi 8 inimese puhul 35-st, kelle korral sai öelda, et nende DNA tandemkorduse koopiaarv ei ole 2. Ülejäänud 27 inimese puhul oli testi olulisuse tõenäosus $p \geq 0{,}0014$ mistõttu ei saanud öelda, et nende inimeste DNA tandemkorduse koopiaarv ei oleks 2 ning jäime nullhüpoteesi juurde, et neil inimestel DNA tandemkorduse koopiaarv on 2.

Kuna ka range olulisuse nivoo korral vähemalt osadel inimestel suudeti tõestada, et nende tandemkorduse koopiaarv ei ole 2 (nagu referentsgenoom seda nõuaks), siis suudeti testi abil näidata, et antud tandemkorduse puhul on tegemist varieeruva pikkusega tandemkordusega.

Kasutatud kirjandus

[1] Heinaru, A. (2012). Geneetika, Tartu, TÜ Kirjastus, lk 319

[2] Viikmaa, M. (1998). Klassikalise geneetika leksikon. http://kodu.ut.ee/~martv/genolex.html (vaadatud 13.04.2015)

[3] Heinaru, A. (2012). Geneetika, Tartu, TÜ Kirjastus, lk 338

[4] Anthony J.F. Griffiths (2013). DNA Sequencing. http://www.britannica.com/EBchecked/topic/422006/DNA-sequencing (vaadatud 13.04.2015)

[5] DNA Sequencing Inc. DNA Sequencing Uses. http://dnasequencing.com/DNA-Sequencing-Uses.html (vaadatud 13.04.2015)

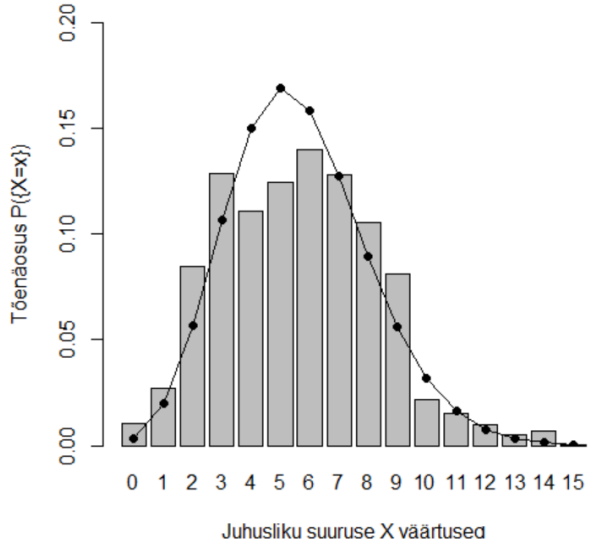
[6] University of Arizona Biology Project (1996). DNA Forensics. http://www.biology.arizona.edu/human_bio/problem_sets/dna_forensics_1/05t.html (vaadatud 13.04.2015)

[7] Illumina, Inc. (2011). Estimating Sequencing Coverage. http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf (vaadatud 27.04.2015)

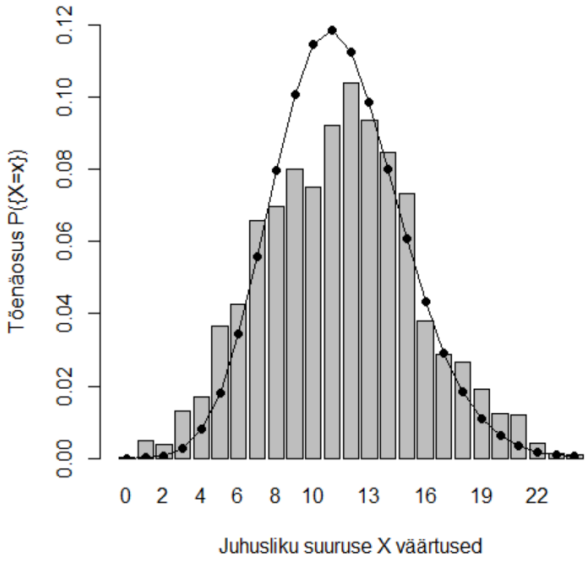
[8] Kaart, T. (2009). Statistiline andmetöötlus. http://ph.emu.ee/~ktanel/VL_0435/magloeng3.pdf (vaadatud 27.04.2015).

Lisad

Lisa 1. Näidisandmestiku põhjal koostatud jaotusfunktsioonid.
Järgnevad graafikud on koostatud reaalse näiteandmestiku põhjal, kus uuringu all oli organism ΦX174. Juhuslik suurus *X* kirjeldab sekveneerimise käigus iga nukleotiidi korral seda läbinud lugemite arvu. Võrdlemiseks on lisatud vastav Poissoni jaotuse tõenäosusfunktsioon.



Lisa 1.1. Lugemite arvu jaotus 6-kordse katvuse korral (tulpadena) ning vastav teoreetiline Poissoni jaotus (punktidena).



Lisa 1.2. Lugemite arvu jaotus 12-kordse katvuse korral(tulpadena) ning vastav teoreetiline Poissoni jaotus (punktidena).

Lisa 2. Näidisandmestiku põhjal koostatud jaotusfunktsioonide R kood.

```
#POISSONI JAOTUSE KONTROLLIMINE:

andmestik=read.table("C:/Users/Kea/Desktop/andmestik.depth",head=F,sep="	")

m1_6<-andmestik$V3
m1_12<-andmestik$V4

## väärtuste vahemiku uurimine:
range(m1_6)
range(m1_12)

## jaotusfunktsioonid (tulpdiagrammidena):

sagedused.freq=table(m1_6)
jaotused.relfreq = sagedused.freq / length(m1_6)
bar1_6 <- barplot(jaotused.relfreq, ylim=range(0,0.2), xlab="Juhusliku suuruse X väärtused",
ylab="Tõenäosus P({X=x})")
points(x=bar1_6,y=dpois(0:15,mean(m1_6)), pch=16) ##teoreetilise jaotuse lisamine
lines(x=bar1_6,y=dpois(0:15,mean(m1_6)))

sagedused.freq=table(m1_12)
jaotused.relfreq = sagedused.freq / length(m1_12)
bar1_12 <- barplot(jaotused.relfreq,ylim=range(0,0.12) ,xlab="Juhusliku suuruse X väärtused",
ylab="Tõenäosus P({X=x})")
points(x=bar1_12,y=dpois(0:24,mean(m1_12)), pch=16) ##teoreetilise jaotuse lisamine
lines(x=bar1_12,y=dpois(0:24,mean(m1_12)))
```

Lisa 3. Tõenäosusfunktsiooni graafiku (joonisel 2.1) moodustamiseks vajalik R kood.

```
# Annan väärtused:
n=101    #n - lugemi pikkus
k=26     #k - korduse pikkus
katvus=30 #katvus

lambda=katvus/n          #lambda - mitu korda keskmiselt hakkas lugem ühest alguspunktist?
lambda1<-lambda*2*k      #lambda1 - X1 kesväärtus
lambda2<-lambda*(n-2*k+1)  #lambda2 - X2 keskväärtus

##----1. Tõenäosusfunktsiooni väärtuse P(X=x)arvutamine (ka eelnevalt kirjas):
#Tõenäosusfunktsiooni summa osa arvutav funktsioon:
summa=function(x){
        vec<- numeric()
        i=0
        for (i in 0:trunc(x/2)) vec<- c(vec,((lambda1**(x-2*i))*(lambda2**i))/((factorial(x-
2*i))*(factorial(i))))
        i=i+1
        vastus=sum(vec)
        vastus
}
#Järgnev funktsioon arvutab X=X1+2*X2 tõenäosusfunktsiooni väärtuse kohal x ehk P(X=x):
tn_fun=function(x){
        avaldis<- exp(-(lambda1+lambda2))*summa(x)
        avaldis
}
#X jaotuse graafik:

x<-0:100
#graafik :
plot(data.frame(cbind(0:100,vektor(0,100))),
xlab="Juhusliku suuruse X väärtused", ylab="Tõenäosus P({X=x})")
```

Lisa 4. DNA tandemkorduse koopiaarvu testi R kood.

```
##TEST K-MEERIDE ABIL DNA TANDEMKORDUSE KOOPIAARVU MÄÄRAMISEKS
## (Kea-test hüpoteeside kontrollimiseks)
##----1. Tõenäosusfunktsiooni väärtuse P(X=x) arvutamine:
#Tõenäosusfunktsiooni summa osa arvutav funktsioon:
summa=function(x,lambda1,lambda2){
        vec=0:trunc(x/2)
        vastus<-((lambda1**(x-2*vec))*(lambda2**vec))/((factorial(x-2*vec))*(factorial(vec)))
        sum(vastus)
}

#Järgnev funktsioon arvutab X=X1+2*X2 tõenäosusfunktsiooni väärtuse kohal x ehk P(X=x):
tn_fun=function(x,lambda1,lambda2){
        avaldis<- exp(-(lambda1+lambda2))*summa(x,lambda1,lambda2)
        avaldis
}

##----2. Funktsioon, mis liidab iga a<= x <=b korral kokku kõik tõenäosused P(X=x)
##ning väljastab nende summa
liida=function(a,b,lambda1,lambda2){
        tn_vec<-numeric()
        for (x in a:b) tn_vec<-c(tn_vec,tn_fun(x,lambda1,lambda2))
        x=x+1
        sum(tn_vec)
}

##----3. Funktsioon, mis leiab x-ide vahemiku, mille korral P(X=x)>P(X=x0), kui x0 on etteantud
#ning seejärel arvutab testi p-väärtuse
#x0 on juhusliku suuruse X väärtus konkreetsel inimesel (otsitavate k-meeride summaarne arv sellel
inimesel)

#lõpp-punkti funktsioon lähtudes algusest:
lõpp_punkt=function(x0, algus, lambda1, lambda2){
        i=algus
        while (tn_fun(i, lambda1, lambda2)>tn_fun(x0, lambda1, lambda2)){
                lõpp=i
                i=i+1
        }
        lõpp
}

#alguspunkti funktsioon lähtudes lõpp-punktist:
algus_punkt=function(x0,lõpp,lambda1,lambda2){
        i=lõpp
        algus=lõpp  ##juhul kui x0 ongi kõige suurema P(X=x) väärtusega
        while (tn_fun(i,lambda1,lambda2)>tn_fun(x0,lambda1,lambda2)){
                algus=i
                i=i-1
        }
        algus
}

#----4. Lõplik testi funktsioon:
KeaMeiTest=function(x0,n,k,katvus){ ##x0 -summaarne k-meeride arv, n - lugemi pikkus, k - korduse
pikkus
        lambda=katvus/n         ##lambda - mitu korda keskmiselt hakkas lugem ühest alguspunktist?
```

```
	lam1=lambda*2*k		##lambda1 - X1 kesväärtus
	lam2=lambda*(n-2*k+1)	##lambda2 - X2 keskväärtus
	tõenäosus<-tn_fun(x0,lam1,lam2)
	if(tn_fun(x0+1,lam1,lam2)>tõenäosus) {
		algus=x0+1
		lõpp=lõpp_punkt(x0,algus,lam1,lam2)
		p=1-liida(algus,lõpp, lam1,lam2) ##test p-väärtus
	} else {
		lõpp=x0-1
		algus=algus_punkt(x0,lõpp,lam1,lam2)
		p=1-liida(algus,lõpp,lam1,lam2)  ##testi p-väärtus
	}
	tulemus=list(x0=x0, n=n, k=k, katvus=katvus, p=p)
	class(tulemus)="KeaMeiTulemus"
	return(tulemus)
}

##----5.Testi tulemuste korralik väljastamine:
print.KeaMeiTulemus=function(a){
cat("\nTest k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.\n
Testi tulemus on järgmine:
  Lugemi pikkus: n =",a$n,"
  Korduse (k-meeri) pikkus: k =",a$k,"
  Ühe nukleotiidi keskmine lugemiste arv: katvus =",a$katvus,"
  Summaarne k-meeride arv indiviidil: X =", a$x0,"\n
  Hüpoteesid:
    H0: Indiviidi DNA tandmekorduse koopiaarv on 2.
    H1: Indiviidi DNA tandemkorduse koopiaarv ei ole 2.\n
  Testi p-väärtus: p =",a$p,"\n
  Hindamiskriteerium:
    Kui (p < olulisuse nivoo) ==>H1.\n")
}

KeaMeiTest(35,101,26,30)  #x0=35, n=101, k=26, katvus=30
KeaMeiTest(72,101,26,30)
```

Lisa 5. Näiteid DNA tandemkorduse koopiaarvu testi R väljundist.

```
> KeaMeiTest(35,101,26,30)

Test k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.

Testi tulemus on järgmine:
   Lugemi pikkus: n = 101
   Korduse (k-meeri) pikkus: k = 26
   Ühe nukleotiidi keskmine lugemiste arv: katvus = 30
   Summaarne k-meeride arv indiviidil: X = 35

   Hüpoteesid:
      H0: Indiviidi DNA tandmekorduse koopiaarv on 2.
      H1: Indiviidi DNA tandemkorduse koopiaarv ei ole 2.

   Testi p-väärtus: p = 0.2711096

   Hindamiskriteerium:
      Kui (p < olulisuse nivoo) ==>H1.
```

Lisa 4.1. Näide testi väljundist, kus konkreetsel indiviidil nähtud summaarne *k*-meeride arv x_0= 35, sekveneerimisel kasutatud lugemi pikkus *n* = 101, *k*-meeri pikkus *k* = 26 ja sekventsi *katvus* = 30.

Tulemusest on näha, et p > 0.05, mistõttu jääme nullhüpoteesi juurde, et indiviidi DNA tandemkorduse koopiaarv on 2.

```
> KeaMeiTest(72,101,26,30)

Test k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.

Testi tulemus on järgmine:
   Lugemi pikkus: n = 101
   Korduse (k-meeri) pikkus: k = 26
   Ühe nukleotiidi keskmine lugemiste arv: katvus = 30
   Summaarne k-meeride arv indiviidil: X = 72

   Hüpoteesid:
      H0: Indiviidi DNA tandmekorduse koopiaarv on 2.
      H1: Indiviidi DNA tandemkorduse koopiaarv ei ole 2.

   Testi p-väärtus: p = 0.00362939

   Hindamiskriteerium:
      Kui (p < olulisuse nivoo) ==>H1.
```

Lisa 4.2. Näide testi väljundist, kus konkreetsel indiviidil nähtud summaarne *k*-meeride arv x_0= 75, sekveneerimisel kasutatud lugemi pikkus *n* = 101, *k*-meeri pikkus *k* = 26 ja sekventsi *katvus* = 30.

Testi tulemusest on näha, et p< 0.05, seega võtame vastu alternatiivhüpoteesi ja saame öelda, et indiiviidi DNA tandemkorduse koopiaarv ei ole 2.

Lisa 6. Üldise DNA tandemkorduse koopiaarvu testi R kood.

```
##ÜLDINE TEST K-MEERIDE ABIL DNA TANDEMKORDUSE KOOPIAARVU MÄÄRAMISEKS
#kasutaja saab valida ise korduse arvu

##----1.TÕENÄOSUSFUNKTSIOON ERINEVATE KORDUSTE KORRAL (kas 1, 2 või 3)
#Diskreetne konvolutsioon:

#Olgu diskreetse juhusliku suuruse X tõenäosusfunktsioon: fun1
#ning diskreetse juhusliku suuruse Y tõenäosusfunktsioon: fun2
#Järgnev funktsioon arvutab (X+tegur*Y) tõenäosusfunktsiooni kohal x.
#tegur tähistab siin tandemkorduse koopiaarvu

konvolutsioon=function(x,fun1,fun2,tegur){
        i<-0:x
        ii<- i[which((x-i) %% tegur == 0)]
        vastus1=numeric()
        vastus2=numeric()
        for (j in ii) {vastus1<-c(vastus1,fun1(j))
        vastus2<-c(vastus2,fun2((x-j)/tegur))}
        vastus=vastus1*vastus2
        sum(vastus)
}
konvolutsioon(30,pois_jaotus1,pois_jaotus2,2)

#Olgu X-summaarne k-meeride arv indiviidil
#Järgnev funktsioon arvutab tõenäosusfunktsiooni P(X=x)etteantud
#tandemkorduse koopiaarvu (tegur), lugemi pikkuse (n), k-meeri pikkuse (k) ja katvuse korral
tõenäosus=function(x,tegur,n,k,katvus){   #tegur - suurim tegur(väärtused: 1,2 või 3), sama mis
korduse arv

        lambda=katvus/n
        lambda1=lambda*2*k
        lambda2=lambda*(n-tegur*k+1)
        pois_jaotus1=function(x){
                avaldis=dpois(x,lambda1)
                avaldis
        }
        pois_jaotus2=function(x){
                avaldis=dpois(x,lambda2)
                avaldis
        }

        if (tegur==1){      ##juhul Y, korduse arv on 1
                vastus<-dpois(x,lambda2)
                vastus
        }
        else if (tegur==2){   ##juhul X+2Y, korduse arv on 2
                vastus<-konvolutsioon(x,pois_jaotus1,pois_jaotus2,2)
                vastus
        }
        else if (tegur==3) {  ##juhul X+2X+3Y,korduse arv on 3
                jaotusfun=function(x){
                        konvolutsioon(x,pois_jaotus1,pois_jaotus1,2)
                }
                konvolutsioon(x,jaotusfun,pois_jaotus2,3)
        }
```

```
}

tõenäosus(x=30,tegur=2,n=101,k=26,katvus=30)

##----2. Funktsioon, mis liidab iga a<= x <=b korral kokku kõik tõenäosused P(X=x)
##etteantud tandemkorduse koopiaarvu (kordus), n, k, katvuse korral ning väljastab nende summa
liida2=function(a,b,kordus,n,k,katvus){
	tn_vec<-numeric()
	for (x in a:b) tn_vec<-c(tn_vec,tõenäosus(x,kordus,n,k,katvus))
	x=x+1
	sum(tn_vec)
}

##----3. Funktsioon, mis leiab x-ide vahemiku, mille korral P(X=x)>P(X=x0), kui x0 on etteantud
#ning seejärel arvutab testi p-väärtuse
#x0 on juhusliku suuruse X väärtus konkreetsel inimesel (otsitavate k-meeride summaarne arv sellel
inimesel)

#lõpp-punkti funktsioon lähtudes algusest:
lõpp_punkt2=function(x0, algus, kordus,n,k,katvus){
	i=algus
	while (tõenäosus(i,kordus,n,k,katvus)>tõenäosus(x0,kordus,n,k,katvus)){
		lõpp=i
		i=i+1
	}
	lõpp
}

#alguspunkti funktsioon lähtudes lõpp-punktist:
algus_punkt2=function(x0,lõpp,kordus,n,k,katvus){
	i=lõpp
	algus=lõpp  ##juhul kui x0 ongi kõige suurema P(X=x) väärtusega
	while (tõenäosus(i,kordus,n,k,katvus)>tõenäosus(x0,kordus,n,k,katvus)){
		algus=i
		i=i-1
	}
	algus
}

##----4. Lõplik testi funktsioon:
KeaMeiTest2=function(x0,kordus,n,k,katvus){ ##x0 -summaarne k-meeride arv, kordus -
tandemkorduse koopiaarv
##n - lugemi pikkus, k - korduse pikkus, kordus peab olema 1, 2 või 3.

	tn<-tõenäosus(x0,kordus,n,k,katvus)
	if(tõenäosus(x0+1,kordus,n,k,katvus)>tn) {
		algus=x0+1
		lõpp=lõpp_punkt2(x0,algus,kordus,n,k,katvus)
		p=1-liida2(algus,lõpp, kordus,n,k,katvus) ##testi p-väärtus
	} else {
		lõpp=x0-1
		algus=algus_punkt2(x0,lõpp,kordus,n,k,katvus)
		p=1-liida2(algus,lõpp,kordus,n,k,katvus)  ##testi p-väärtus
	}
	tulemus=list(x0=x0, kordus=kordus, n=n, k=k, katvus=katvus, p=p)
	class(tulemus)="KeaMeiTulemus2"
	return(tulemus)
```

```
}

##----5.Testi tulemuste korralik väljastamine:

print.KeaMeiTulemus2=function(a){
cat("\nTest k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.\n
Testi tulemus on järgmine:
  Lugemi pikkus: n =",a$n,"
  Korduse (k-meeri) pikkus: k =",a$k,"
  Ühe nukleotiidi keskmine lugemiste arv: katvus =",a$katvus,"
  Summaarne k-meeride arv indiviidil: X =", a$x0,"\n
  Hüpoteesid:
    H0: Indiviidi DNA tandmekorduse koopiaarv on", a$kordus,"
    H1: Indiviidi DNA tandemkorduse koopiaarv ei ole ", a$kordus,"\n
  Testi p-väärtus: p =",a$p,"\n
  Hindamiskriteerium:
    Kui (p < olulisuse nivoo) ==>H1.\n")
}

KeaMeiTest2(35,2,101,26,30)  #x0=35, kordus=2, n=101, k=26, katvus=30
KeaMeiTest2(72,2,101,26,30)  #x0=72, kordus=2, n=101, k=26, katvus=30
KeaMeiTest2(35,3,101,26,30)  #x0=35, kordus=3, n=101, k=26, katvus=30
KeaMeiTest2(72,3,101,26,30)  #x0=72, kordus=2, n=101, k=26, katvus=30
```

Lisa 7. Üldise DNA tandemkorduse testi väljund.

```
> KeaMeiTest2(35,2,101,26,30)  #x0=35, kordus=2, n=101, k=26, katvus=30

Test k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.

Testi tulemus on järgmine:
   Lugemi pikkus: n = 101
   Korduse (k-meeri) pikkus: k = 26
   Ühe nukleotiidi keskmine lugemiste arv: katvus = 30
   Summaarne k-meeride arv indiviidil: X = 35

   Hüpoteesid:
      H0: Indiviidi DNA tandmekorduse koopiaarv on 2
      H1: Indiviidi DNA tandemkorduse koopiaarv ei ole  2

   Testi p-väärtus: p = 0.2711096

   Hindamiskriteerium:
      Kui (p < olulisuse nivoo) ==>H1.
```

Lisa 7.1. Näide testi väljundist, kus konkreetsel indiviidil nähtud summaarne *k*-meeride arv x_0= 35, korduse arv nullhüpoteesi korral on 2, sekveneerimisel kasutatud lugemi pikkus *n* = 101, *k*-meeri pikkus *k* = 26 ja sekventsi *katvus* = 30.

Tulemusest on näha, et p > 0.05, mistõttu jääme nullhüpoteesi juurde, et indiviidi DNA tandemkorduse koopiaarv on 2.

```
> KeaMeiTest2(35,3,101,26,30)  #x0=35, kordus=3, n=101, k=26, katvus=30

Test k-meeride abil DNA tandemkorduse koopiaarvu määramiseks.

Testi tulemus on järgmine:
   Lugemi pikkus: n = 101
   Korduse (k-meeri) pikkus: k = 26
   Ühe nukleotiidi keskmine lugemiste arv: katvus = 30
   Summaarne k-meeride arv indiviidil: X = 35

   Hüpoteesid:
      H0: Indiviidi DNA tandmekorduse koopiaarv on 3
      H1: Indiviidi DNA tandemkorduse koopiaarv ei ole  3

   Testi p-väärtus: p = 0.003514392

   Hindamiskriteerium:
      Kui (p < olulisuse nivoo) ==>H1.
```

Lisa 7.2. Näide testi väljundist, kus konkreetsel indiviidil nähtud summaarne *k*-meeride arv x_0= 35, korduse arv nullhüpoteesi korral on 3, sekveneerimisel kasutatud lugemi pikkus *n* = 101, *k*-meeri pikkus *k* = 26 ja sekventsi *katvus* = 30.

Tulemusest on näha, et p < 0.05, seega võtame vastu alternatiivhüpoteesi ja saame öelda, et indiiviidi DNA tandemkorduse koopiaarv ei ole 3.

Lisa 8. Testi võimsuste arvutamiseks kasutatud R kood.

```
##TESTI VÕIMSUSTE ARVUTUSED
#GRAAFIK (Joonise 4.1 jaoks kasutatud kood):
#Järgenv funktsioon leiab vektori jaotuste X=Y1+2*Y2 või X=Y1+2*Y2+3*Y3
# tõenäosusfunktsiooni kohal x
#etteantud vahemikus x=algus kuni x=lõpp etteantud korduse korral
genereeri=function(algus,lõpp,kordus, n,k,katvus){
	vec<-numeric()
	for (x in algus:lõpp) vec<-c(vec,tõenäosus(x,kordus,n,k,katvus))
	x=x+1
	vec
}

#tõenäosusfunktsioonide graafikud (graafikud koos 2 ja 3 korduse korral):
plot(data.frame(cbind(0:130,genereeri(0,130,2,101,26,30))),
xlab="Summaarne k-meeride arv (x)", ylab="Tõenäosus P(X=x)",pch=16)
points(data.frame(cbind(0:130,genereeri(0,130,3,101,26,30))),pch=1)
legend(65,0.045,legend=c("Tandemkorduse koopiaarv = 2","Tandemkorduse koopiaarv =
3"),pch=c(16,1))

##TESTI VÕIMSUSED:
##Testi tulemuste väljastamine korraga:
#Testi tulemuste (p-väärtuste) vektor etteantud X väärtuste vektori korral:
p_vektor=function(Xvec,n,k,katvus){
	vastus=numeric()
	i=1
	for (i in 1:length(Xvec)) vastus<-c(vastus,KeaMeiTest(Xvec[i],n,k,katvus)$p)
	i=i+1
	vastus
}

##Võtame neist välja kõik need, mille korral p-väärtus<0.05 ja arvutame saadud vektori pikkuse ehk
##järgnev funktsioon arvutab,kui palju oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid=function(v){
	vec=numeric()
	for (i in v){
		if (i<0.05){
			vec<-c(vec,i)
		}
	}
	length(vec)
}

##VÕIMSUS 1x1 JUHUL KUI KATVUS=30:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*(n-k+1)
vec=0:100 #võimalikud x väärtused, mille vahel valida
#arvutame vektori teoreetilistest tõenäosustest:
tn_vektor1=function(v,n,k,katvus){
	vec<-numeric()
	for (x in v) vec<-c(vec,tõenäosus(x,1,n,k,katvus))
	x=x+1
	vec
}
#leiame valimi teoreetilisest jaotusest:
v1x1_30 <- sample(vec, size=10000, replace=TRUE, prob=tn_vektor1(vec,n,k,katvus))

##p-väärtused korduste arvu 1 korral on:
```

```
p1x1_30<-p_vektor(v1x1_30,101,26,30) ##katvus=30
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x1_30) #vastus: 8896 tk, seega võimus=8896/10000=0.8896


##VÕIMSUS 1x1 JUHUL KUI KATVUS=60:
n=101;k=26; katvus=60; lambda=katvus/n; lambda1=lambda*(n-k+1)
vec=0:100 #võimalikud x väärtused, mille vahel valida

#leiame valimi teoreetilisest jaotusest:
v1x1_60 <- sample(vec, size=10000, replace=TRUE, prob=tn_vektor1(vec,n,k,katvus))

##p-väärtused korduste arvu 1 korral on:
p1x1_60<-p_vektor(v1x1_60,101,26,60) ##katvus=60
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x1_60) #vastus: 9987 tk, seega võimus=9987/10000=0.9987


##VÕIMSUS 2x2 JUHUL KUI KATVUS=30:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
vec=0:100 #võimalikud x väärtused, mille vahel valida
##Järgnevad funktsioonid arvutavad X=X1+2*X2 tõenäosusfunktsiooni väärtuse kohal x ehk P(X=x):
summa=function(x,lambda1,lambda2){
        vec=0:trunc(x/2)
        vastus<-((lambda1**(x-2*vec))*(lambda2**vec))/((factorial(x-2*vec))*(factorial(vec)))
        sum(vastus)
}
tn_fun=function(x,lambda1,lambda2){
        avaldis<- exp(-(lambda1+lambda2))*summa(x,lambda1,lambda2)
        avaldis
}
##arvutame vektori teoreetilistest tõenäosustest:
tn_vektor=function(v,lambda1,lambda2){
        vec<-numeric()
        for (x in v) vec<-c(vec,tn_fun(x,lambda1,lambda2))
        x=x+1
        vec
}
#leiame valimi teoreetilisest jaotusest:
v2x2_30 <- sample(vec, size=10000, replace=TRUE, prob=tn_vektor(vec,lambda1,lambda2))

##p-väärtused korduste arvu 2 korral on:
p2x2_30<-p_vektor(v2x2_30,101,26,30) ##katvus=30
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p2x2_30) #vastus: 475 tk, seega võimus=475/10000=0.0475


##VÕIMSUS 2x2 JUHUL KUI KATVUS=60:
n=101;k=26; katvus=60; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
#võimalikud x väärtused, mille vahel valida
vec=0:170

##leiame valimi teoreetilisest jaotusest:
v2x2_60 <- sample(vec, size=10000, replace=TRUE, prob=tn_vektor(vec,lambda1,lambda2))

##p-väärtused korduste arvu 2 korral on:
p2x2_60<-p_vektor(v2x2_60,101,26,60) ##katvus=60
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
```

```
alternatiivseid(p2x2_60) #vastus: 492 tk, seega võimus=492/10000=0.0492


## VÕIMSUS 3x3 JUHUL KUI KATVUS=30:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
##Simulatsioonid (genereerime 10000 vaatlust jaotusest, kus korduse arv on 3)
y1<- rpois(10000,lambda1)
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v3x3_30<-y1+2*y2+3*y3

#p-väärtused korduste arvu 3 korral on:
p3_3_30<-p_vektor(v3x3_30,101,26,30) ##katvus=30
#Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p3_3_30) #vastus: 6611 tk, seega võimus=6611/10000=0.6611


## VÕIMSUS 3x3 JUHUL KUI KATVUS=60:
n=101;k=26; katvus=60; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
##Simulatsioonid (genereerime 10000 vaatlust jaotusest, kus korduse arv on 3):
y1<- rpois(10000,lambda1)
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v3x3_60<-y1+2*y2+3*y3
max(v3x3_60)

#p-väärtused korduste arvu 3 korral on:
p3x3_60<-p_vektor(v3x3_60,101,26,60) ##katvus=60
#Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p3x3_60) #vastus: 8949 tk, seega võimus=8949/10000=0.8949


##VÕIMSUS 1x2 JUHUL KUI KATVUS=30:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 1:
vec1<- 0:120
v1x2_30_1 <- sample(vec1, size=10000, replace=TRUE, prob=tn_vektor1(vec1,101,26,15))
##kontrolliks graafik:
plot(data.frame(cbind(v1x2_30_1,tn_vektor1(v1x2_30_1,101,26,15))),
xlab="Juhusliku suuruse X väärtused", ylab="Tõenäosus P({X=x})")

##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 2:
n=101;k=26; katvus=15; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
vec2<-0:120
v1x2_30_2 <- sample(vec2, size=10000, replace=TRUE, prob=tn_vektor(vec2,lambda1,lambda2))
##kontrolliks graafik:
plot(data.frame(cbind(v1x2_30_2,tn_vektor(v1x2_30_2,lambda1,lambda2))),
xlab="Juhusliku suuruse X väärtused", ylab="Tõenäosus P({X=x})")

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib ühelt vanemalt
##ühe korduse ja teiselt 2 kordust
v1x2_30 <- v1x2_30_1 + v1x2_30_2

##p-väärtused korduste arvu 1x2 korral on:
p1x2_30<-p_vektor(v1x2_30,101,26,30) ##katvus=30

##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x2_30) #vastus: 2293 tk, seega võimus=2293/10000=0.2293
```

```
##VÕIMSUS 1x2 JUHUL KUI KATVUS=60:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 1:
vec1<- 0:120
v1x2_60_1 <- sample(vec1, size=10000, replace=TRUE, prob=tn_vektor1(vec1,101,26,30))
##kontrolliks graafik:
plot(data.frame(cbind(v1x2_60_1,tn_vektor1(v1x2_60_1,101,26,30))),
xlab="Juhusliku suuruse X väärtused", ylab="Tõenäosus P({X=x})")

##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 2:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
vec2<-0:170
v1x2_60_2 <- sample(vec2, size=10000, replace=TRUE, prob=tn_vektor(vec2,lambda1,lambda2))
##kontrolliks graafik:
plot(data.frame(cbind(v1x2_60_2,tn_vektor(v1x2_60_2,lambda1,lambda2))),
xlab="Juhusliku suuruse X väärtused", ylab="Tõenäosus P({X=x})")

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib ühelt vanemalt
##ühe korduse ja teiselt vanemalt 2 kordust
v1x2_60 <- v1x2_60_1 + v1x2_60_2

##p-väärtused korduste arvu 1x2 korral on:
p1x2_60<-p_vektor(v1x2_60,101,26,60) ##katvus=60

##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x2_60) #vastus: 4584 tk, seega võimus=4584/10000=0.4584


##VÕIMSUS 1x3 JUHUL KUI KATVUS=30:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 1:
vec1<- 0:120
v1x3_30_1 <- sample(vec1, size=10000, replace=TRUE, prob=tn_vektor1(vec1,101,26,15))

##genereerime 10000 juhuslikku suurust jaotusest, kus korduste arv on 3:
n=101;k=26; katvus=15; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
y1<- rpois(10000,lambda1)
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v1x3_30_3 <- y1+2*y2+3*y3

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib
##ühelt vanemalt ühe korduse ja teiselt 3 kordust
v1x3_30 <- v1x3_30_1 + v1x3_30_3

##p-väärtused korduste arvu 1x3 korral on:
p1x3_30<-p_vektor(v1x3_30,101,26,30) ##katvus=30
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x3_30) #vastus: 578 tk, seega võimus=578/10000=0,0578


##VÕIMSUS 1x3 JUHUL KUI KATVUS=60:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 1:
vec1<- 0:120
v1x3_60_1 <- sample(vec1, size=10000, replace=TRUE, prob=tn_vektor1(vec1,101,26,30))

##genereerime 10000 juhuslikku suurust jaotusest, kus korduste arv on 3:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
y1<- rpois(10000,lambda1)
```

```
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v1x3_60_3 <- y1+2*y2+3*y3

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib
##ühelt vanemalt ühe korduse ja teiselt 3 kordust
v1x3_60 <- v1x3_60_1 + v1x3_60_3

##p-väärtused korduste arvu 1x3 korral on:
p1x3_60<-p_vektor(v1x3_60,101,26,60) ##katvus=60
##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p1x3_60) #vastus: 611 tk, seega võimus=611/10000=0,0611


##VÕIMSUS 2x3 JUHUL KUI KATVUS=30:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 2:
n=101;k=26; katvus=15; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
vec2<-0:120
v2x3_30_2 <- sample(vec2, size=10000, replace=TRUE, prob=tn_vektor(vec2,lambda1,lambda2))

##genereerime 5000 vaatlust jaotusest, kus korduse arv on 3:
n=101;k=26; katvus=15; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
y1<- rpois(10000,lambda1)
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v2x3_30_3<-y1+2*y2+3*y3

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib
##ühelt vanemalt 2 kordust ja teiselt 3 kordust
v2x3_30 <- v2x3_30_2 + v2x3_30_3

##p-väärtused korduste arvu 2x3 korral on:
p2x3_30<-p_vektor(v2x3_30,101,26,30) ##katvus=30

##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p2x3_30) #vastus:  2758 tk, seega võimus=2758/10000=0,2758


##VÕIMSUS 2x3 JUHUL KUI KATVUS=60:
##leiame valimi suurusega 10000 teoreetilisest jaotusest, kus korduste arv on 2:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda2=lambda*(n-2*k+1)
vec2<-0:170
v2x3_60_2 <- sample(vec2, size=10000, replace=TRUE, prob=tn_vektor(vec2,lambda1,lambda2))

##genereerime 10000 vaatlust jaotusest, kus korduse arv on 3:
n=101;k=26; katvus=30; lambda=katvus/n; lambda1=lambda*2*k; lambda3=lambda*(n-3*k+1)
y1<- rpois(10000,lambda1)
y2<- rpois(10000,lambda1)
y3<- rpois(10000,lambda3)
v2x3_60_3<-y1+2*y2+3*y3

##liidame valimite vastavad väärtused, et saada jaotust, mille korral indiviid pärib
##ühelt vanemalt 2 kordust ja teiselt 3 kordust
v2x3_60 <- v2x3_60_2 + v2x3_60_3

##p-väärtused korduste arvu 2x3 korral on:
p2x3_60<-p_vektor(v2x3_60,101,26,60) ##katvus=60
```

##Kui palju neist oli neid juhuslikke suurusi, mille korral võeti H1 vastu:
alternatiivseid(p2x3_60) #vastus: 4505 tk, seega võimus=4505/10000=0,4505

Lisa 9. Näidisinimeste testimine geenivaramu andmete põhjal.

```
#GEENIVARAMUST SAADUD INDIVIIDIDE TESTIMINE

##Testi tulemuste väljastamine korraga:

#Summaarsete k-meeride arvudega vektor:
sum_k=c(61,78,73,66,71,44,48,73,74,46,66,46,83,56,86,77,67,48,67,51,83,69,62,65,57,53,68,55,78,5
8,73,54,84,76,67)
length(sum_k)

#testi tulemuste (p-väärtuste) vektor etteantud X väärtuste vektori korral:
p_vektor=function(Xvec,n,k,katvus){
	vastus=numeric()
	i=1
	for (i in 1:length(Xvec)) vastus<-c(vastus,KeaMeiTest(Xvec[i],n,k,katvus)$p)
	i=i+1
	vastus
}

#Järgnev funktsioon võtab ette p väärtuste vektori ja väljastab, kui paljude korral neist
#võttis test vastu sisuka hüoteesi H1 etteantud olulisuse nivoo alpha korral.
sisukaid=function(vektor,alpha,n,k,katvus){
	p_väärtused <- p_vektor(vektor,n,k,katvus)
	vec=numeric()
	for (i in p_väärtused){
		if (i<alpha){
			vec<-c(vec,i)
		}
	}
	length(vec)
}
#testime geenivaramust saadud inimesi:
sisukaid(sum_k,0.1,101,26,30) ##inimeste arv, kelle korral suudeti vastu võtta sisukas hüpotees
olulisuse nivool 0.1
sisukaid(sum_k,0.05,101,26,30)
sisukaid(sum_k,0.01,101,26,30)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kea Mei,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Statistiline test *k*-meeride abil DNA tandemkorduse koopiaarvu määramiseks“, 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, 29.04.2015.