

TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Kaur Tenson
**Inimese genoomis esinevate struktuursete
variantide sageduse modelleerimine**

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Maarja Jõeloo, PhD Märt Möls

TARTU 2026

INIMESE GENOOMIS ESINEVATE STRUKTUURSETE VARIANTIDE SAGEDUSE MODELLEERIMINE

Bakalaureusetöö

Kaur Tenson

Lühikokkuvõte

Käesolevas bakalaureusetöös uuriti inimese genoomis esinevate struktuursete variantide (SV-de) harulduse seost nende paiknemisega genoomi erinevates piirkondades. Eesmärk oli hinnata, kas geenide ekspressiooni mõjutavate eQTL-piirkondadega kattuvad SV-d on tugevama puhastava valiku all kui muudes piirkondades paiknevad variandid. Analüüsis kasutati gnomAD v4.1 struktuursete variantide andmestikku, GENCODE geeni andmestikku ja GTEx eQTL andmeid.

SV-de harulduse modelleerimiseks rakendati logistilist regressiooni koos B-splainidega, mis võimaldas kirjeldada pikkuse ja harulduse vahelist mittelineaarset seost. Tulemused näitasid, et nii SV pikkus kui ka ülekate geenide ja eQTL-piirkondadega mõjutavad oluliselt variantide sagedust populatsioonis. Kõige suurem haruldaste variantide osakaal esines SV-del, mis kattusid samaaegselt nii geeni kui ka eQTL-iga. Ainult eQTL-iga kattuvad SV-d olid keskmiselt haruldasemad kui ainult geeniga kattuvad variandid.

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

Märksõnad: struktuursed variandid, geenid, eQTL, puhastav valik, logistiline regressioon, B-splain.

MODELING THE FREQUENCY OF STRUCTURAL VARIANTS IN THE HUMAN GENOME

Bachelor thesis

Kaur Tenson

Abstract

This bachelor's thesis studied the relationship between the rarity of structural variants (SVs) in the human genome and their positioning in different genomic regions. The aim was to assess whether SVs overlapping eQTL regions affecting gene expression are under stronger purifying selection than variants in other regions. The analysis used the gnomAD v4.1 structural variant dataset, the GENCODE gene dataset, and GTEx eQTL data.

To model the rarity of SVs, logistic regression with B-splines was applied, which made it possible to describe the nonlinear relationship between variant length and rarity. The results showed that both SV length and overlap with genes and eQTL regions significantly affect variant frequency in the population. The highest proportion of rare variants occurred among SVs overlapping both a gene and an eQTL at the same time. SVs overlapping only eQTL regions were on average rarer than variants overlapping only genes.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics and financial mathematics.

Key Words: structuralvariants, genes, eQTL, purifying selection, logistic regression, B-splines.

Sisukord

Sissejuhatus	5
1 Taust	6
1.1 Struktuursed variandid	6
1.2 Mobiilsed elemendid	7
1.3 Referentsgenoom	7
1.4 Geenid ja ekspressiooni kvantitatiivsete tunnuste lookused . .	8
1.5 SV-de sagedus ja kahjulikkus	9
2 Andmed	10
2.1 Genoomsete variantide andmebaas gnomAD	10
2.2 Geenide andmebaas GENCODE	14
2.3 eQTL andmebaas GTEx	15
3 Metoodika	16
3.1 B-splainid	16
3.2 Splaini tuletis	18
4 Analüüs ja tulemused	21
4.1 Töö eesmärk	21
4.2 Andmete esmane töötlemine ja kokkupanemine	21
4.3 SV-de pikkuste jaotused funktsionaalsetes gruppides	23
4.4 Mudel SV harulduse hindamiseks	27

4.4.1	Deletsioonide harulduse modelleerimine	28
4.4.2	Insertsioonide harulduse modelleerimine	30
4.5	Tulemuste interpretatsioon	33
	Kokkuvõte	35
	Kasutatud allikad	36
1	Lisa 1: Mudel deletsioonidele	39
2	Lisa 2: Mudel insertsioonidele	41

Sissejuhatus

Inimeste genoomis esinevad struktuursed variandid (SV-d), mis on vähemalt 50 aluspaari pikkused genoomi ümberkorraldused, mõjutavad organismi tugevamini kui üksikud nukleotiidsed muutused, kuna hõlmavad korraga suuri DNA piirkondi. Samas on SV-de mõju veel suhteliselt vähe uuritud.

Käesoleva töö eesmärk on uurida, kuidas ülekate geenide avaldumist mõjutavate ekspressiooni kvantitatiivsete tunnuste lookustega (eQTL) on seotud SV haruldusega populatsioonis, võttes arvesse SV pikkuse ja sageduse omavahe- list seost. Analüüsis kasutatakse gnomAD v4.1 SV-de andmestikku, geenide asukohti GENCODE v49 andmebaasist ning eQTL kohti GTEx V10 and- mestikust.

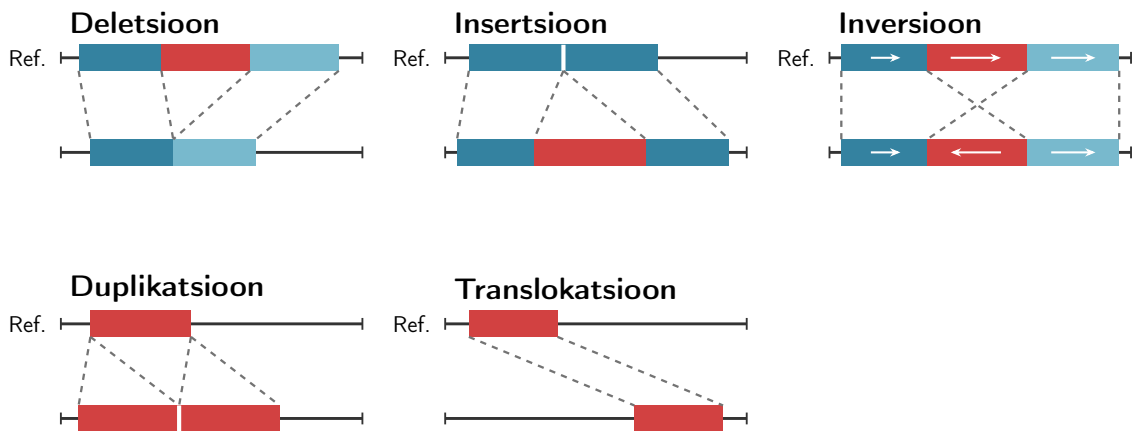
Töö on üles ehitatud järgmiselt: esimene peatükk annab teoreetilise tausta, teine kirjeldab kasutatud andmeid, kolmas metoodikat ning neljas peatükk esitab analüüsi koos tulemustega. Töö raames loodud kood on saadaval Git- Hubis: [https://github.com/kaurxx/T-Inimese-genoomis-esinevate-s- truktuursete-variantide-sageduse-modelleerimine-kood.git](https://github.com/kaurxx/T-Inimese-genoomis-esinevate-struktuursete-variantide-sageduse-modelleerimine-kood.git).

1 Taust

Kuigi kõigi inimeste DNA on keskmiselt umbes 99,6% ulatuses ühesugune ja enamik erinevustest ei oma suurt funktsionaalset mõju, on järelejäänud väike hulk geneetilisi erinevuseid ehk variante üheks inimestevahelise mitmekesisuse põhjuseks. Geneetilisi variante jaotatakse mitmesse kategooriasse. Kõige rohkem on inimesegeneetikas käsitletud üksiknukleotiidseid variante (SNV-d, ingl.k. *single nucleotide variation*), mis tähendab ühe nukleotiidi erinevust inimeste vahel. Palju vähem on uuritud suuremahulisi genoomseid muutuseid, mis hõlmavad kümnete kuni miljonite nukleotiidide erinevusi. Kõiki geneetilisi variante uuritakse eelkõige eesmärgiga leida selliseid muutuseid, mis võivad suurendada osade inimeste vastuvõtlikust mõnele haigusele või ka näiteks ravimikõrvaltoime tekkimisele. [1]

1.1 Struktuursed variandid

Struktuursed variandid (SV-d, ingl.k. *structural variations*) on vähemalt 50 nukleotiidi pikkused suuremahulised genoomi ümberkorraldused. Neid jaotatakse enamasti järgmistesse kategooriatesse: deletsioon (DNA lõigu puudumine), insertioon (uue DNA lõigu lisandumine), duplikatsioon (DNA lõigu kahekordistumine), inversioon (DNA lõigu ümberpööramine) ning translokatsioon (DNA lõikude ümbervahetamine) ja muud. Põhilised SV-de tüübid on toodud joonisel 1.[2] Varasemad genoomi analüüsimise meetodid on olnud väga tõhusad SNV-de tuvastamisel, kuid märksa vähem edukad SV-de leidmisel [3]. Samas on leitud tõendusmaterjali, et SV-d võivad organismi mõjutada oluliselt tugevamalt kui SNV-d, kuna nendega kaasnevad muutused hõlmavad korraga suuremaid DNA piirkondi [4].



Joonis 1: Struktuursete variantide tüübid, koostatud [5] põhjal

1.2 Mobiilsed elemendid

Mobiilsed elemendid (ME-d) on DNA-järjestused, mis suudavad ennast paljundada ja/või muuta oma asukohta genoomis [6]. Seetõttu on nad oma olemuselt sarnased SV-dega ning neid saab tuvastada kui deletsioone või insertioone [7]. Põhilisteks erinevusteks on see, et neil on enamasti fikseeritud teadaoleva pikkusega järjestus ning neid leidub genoomis sadades tuhandetes koopiates [7].

Inimese genoomis moodustavad ME-d ligikaudu poole. Kõige levinumateks ME-deks on LINE-1 (L1) ja ALU elemendid, mis kokku moodustavad kolmandiku kogu genoomist. Tekviklik L1 on umbes 6 000 nukleotiidi pikk. Seevastu keskmine ALU pikkus jääb vaid 300 nukleotiidi ümbrusesse. [7]

1.3 Referentsgenoom

Referentsgenoom on kokkuleppeline organismi DNA järjestus, mida teadlased kasutavad võrdlusalusena teiste sama liigi organismide genoomiuuringute

jaoks. Inimese referentsgenoom on koostatud paljude doonorite geneetilisest materjalist, et esindada nõ keskmise inimese genoomi. Uuritava inimese genoomi kirjeldatakse alati referentsgenoomi suhtes[8]. SV uuringutes on näiteks deletsioon määratud kui DNA piirkond, mis leidub referentsgenoomis, kuid puudub uuritavas indiviidis. Insertioon on määratud kui DNA lõik, mis on olemas uuritavas indiviidis, kuid puudub referentsgenoomis (vt joonis 1). Käesolevas töös on kasutatud inimese referentsgenoomi versiooni GRCh38.

1.4 Geenid ja ekspressiooni kvantitatiivsete tunnuste lookused

Geenid on DNA järjestused, mis sisaldavad valkude valmistamiseks vajalikku informatsiooni. Inimese genoomis on umbes 20 000 valku kodeerivat geeni. Geenijärjestused moodustavad inimese genoomist ligikaudu 1,5%. [9]

Ekspressiooni kvantitatiivsete tunnuste lookused (eQTL-id, ingl.k. *expression quantitative trait locus*) on genoomi piirkonnad, mis mõjutavad geenide ekspressioonitaset ehk mRNA või valgu hulka, mida uuritava geeni järjestyse põhjal toodetakse. Tavaliselt tuvastatakse eQTL-e geneetiliste variantide ja geeniekspressiooni taseme vaheliste seoste analüüsi kaudu. Uuringud on näidanud, et paljud haigustega seotud SNV-d kattuvad sageli eQTL-idega, mis viitab sellele, et geneetilised variandid võivad mõjutada haigusrisi geeniekspressiooni muutmise kaudu. Oluline on ka märkida, et kuigi DNA on inimese kõikide kudede rakkudes identne, siis geenide ekspressioonitasemed ja eQTL-id võivad olla koespetsiifilised, s.t. genoomipiirkond, mis on eQTL näiteks vererakkudes ei pruugi seda olla näiteks maksas. [10]

1.5 SV-de sagedus ja kahjulikkus

Geneetilised variandid, sh SV-d, võivad paikneda nii geenide vahel kui ka geenidega ülekattes. On näidatud, et nii geenisisessed kui ka geeniekspressiooni mõjutavates piirkondades paiknevad struktuursed variandid on populatsioonis haruldasemad kui geenidevahelistes piirkondades paiknevad SV-d [11].

Samuti on teada, et geneetiliste variantide kahjulikkus ja sagedus on omavahel tihedalt seotud. Mida suurem on variantide sagedus, seda tõenäolisemalt puudub tal tugev mõju inimese tervisele. Kahjulikud variandid esinevad üldjuhul madalama sagedusega, sest nad „eemaldatakse” evolutsiooni käigus, s.t. nad päranduvad järgmisesse põlvkonda väiksema tõenäosusega. Seda fenomeni nimetatakse evolutsioonigeneetikas puhastavaks valikuks. [12]

2 Andmed

2.1 Genoomsete variantide andmebaas gnomAD

Töös kasutatud SV andmed on pärit gnomAD (Genome Aggregation Database [13]) v4.1 andmestikust (ilmunud aastal 2023), mis on agregeeritud kujul ja avalikult kättesaadavad. Andmed on saadud 63 046 indiviidi põhjal, kelle genoomidest on määratud 2 154 486 SV-d, millest 1 199 117 (55,7%) on läbinud andmestiku koostaja poolse kvaliteedikontrolli [13] (vt tabel 1).

Käesolevas töös on analüüsitud kolme suurimat SV-de kategooriat: deletsioonid (DEL), duplikatsioonid (DUP) ja insertioonid (INS), mida oli andmestikus kokku 1 183 013 kvaliteedikontrolli läbinud varianti. Töös kasutati ainult autosoomseid kromosoome (kromosoomid 1-22), ning jäeti välja soospetsiifilised kromosoomid X ja Y, mille eemaldamisel jäi alles 1 128 660 SV-d.

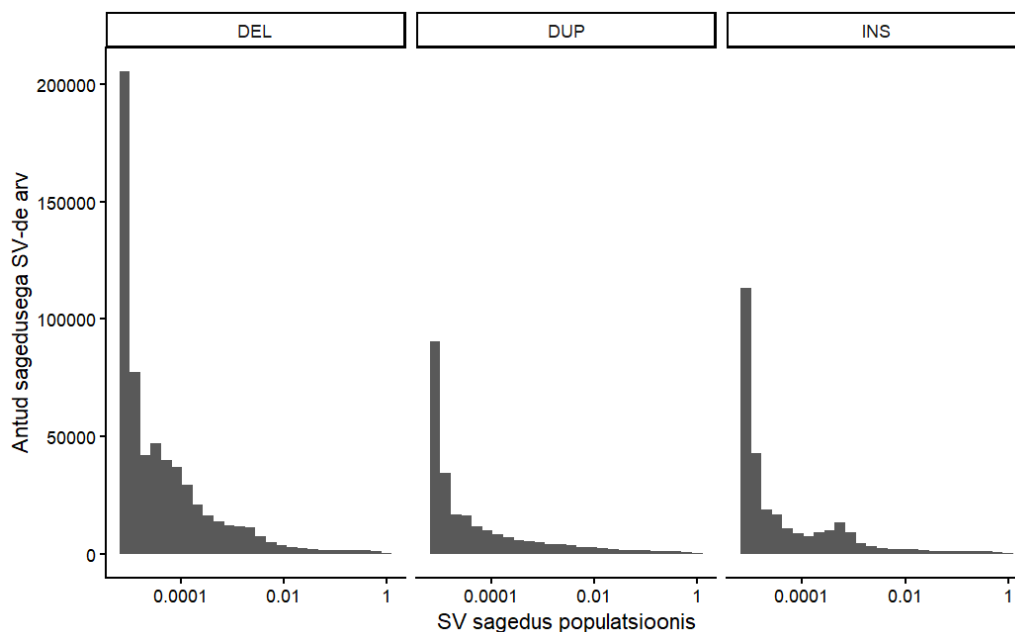
Tabel 1: SV-de arvud gnomAD andmebaasis. Töös kasutatud andmemahud on allajoonitud.

Tüüp Kvaliteet	DEL	DUP	INS	Muud	Summa
Kvaliteetne	627 947	258 882	296 184	16 104	1 199 117
Mittekvaliteetne	578 331	10 444	8 461	358 133	955 369
Summa	1 206 278	269 326	8 461	374 237	2 154 486

Iga SV on andmestikus esitatud kui referentsgenoomi piirkond, mis on defineeritud kolme arvu abil: kromosoomi number, SV alguskoordinaat kromosoomil ja SV lõpp-koordinaat kromosoomil. Lisaks sisaldab gnomAD andmebaas iga SV kohta tunnuseid ka SV pikkuse ja esinemissageduse kohta.

Joonisel 2 on toodud SV-de esinemissageduste jaotus. Kõige enam on väikse sagedusega SV-sid, 89,5% SV-dest on väiksema sagedusega kui 0,001. Jaotus on sarnane kõigi tüüpide puhul, kuid insertioonidel on suurem arv variante

umbes 0,0006 sageduse juures.

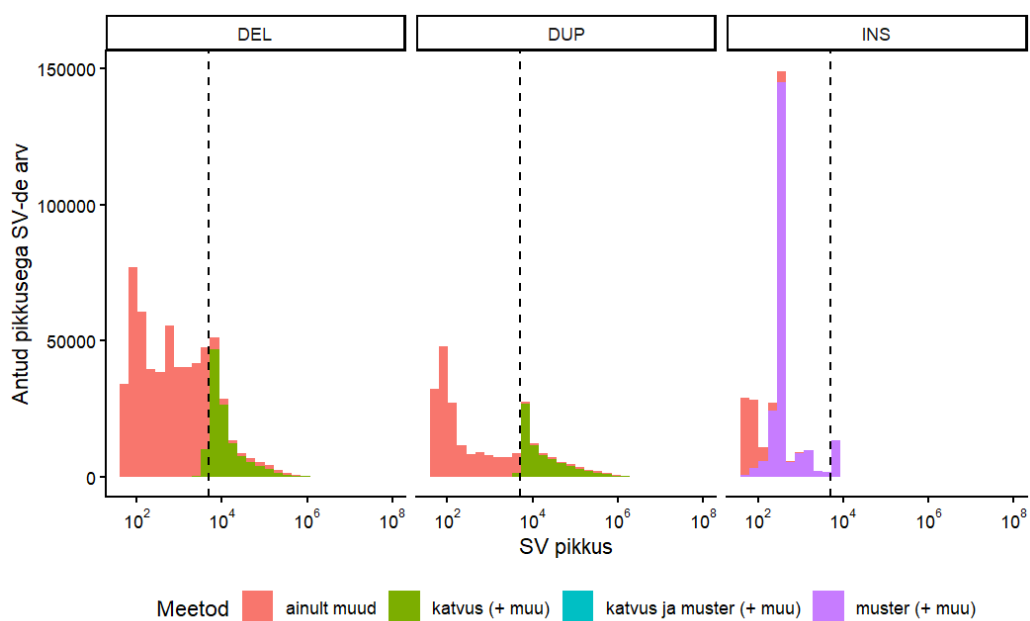


Joonis 2: SV-de sageduste jaotus.

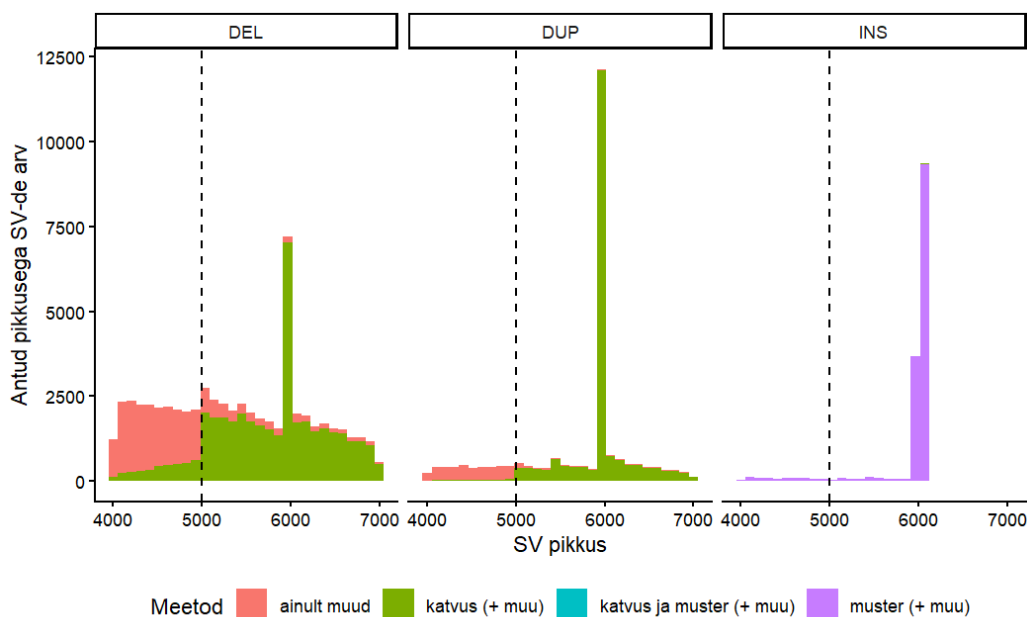
SV-de pikkuste jaotused tüübi järgi on toodud joonisel 3. Enamus deletsioone (54,9%), duplikatsioone (58,2%) ja insertioone (89,6%) on lühikesed (<1000 nukleotiidi). Pikemaid kui 5000 nukleotiidi (joonisel märgitud katkendliku joonega) deletsioone ja insertioone on vaid vastavalt 129 823 (21,7%) ja 13 685 (4,8%). Seevastu duplikatsioonide hulk on 5000 nukleotiidi piirist alates suurem, moodustades 31,6% kõigist andmestikus olevatest duplikatsioonidest, suurem erinevus tekib pikkuse 6000 juures (vt. joonis 4), mis vastab L1 elementide pikkustele. Deletsioonide ja insertioonide jaotuste puhul on näha suuremat variantide sagedust pikkuste 250, mis vastab ligikaudu ALU pikkusele, ja 6000 nukleotiidi juures, mis vastab L1 elementidele. Töös kasutatud gnomAD SV andmebaas on koostatud, kasutades kombi-

natsiooni erinevatest SV tuvastamismeetoditest. Nende meetodite tundlikkus varieerub eri tüüpi ja erineva pikkusega SV-de tuvastamisel. Seega tuleb arvestada ka meetoditega, millega andmed on leitud. Tuvastamismeetodid saab jagada kolmeks: katvusel põhinevad meetodid, mustriotsingul põhinevad meetodid ja muud meetodid. Üks SV võib olla tuvastatud mitme erineva meetodi poolt. Üks SV võib olla tuvastatud mitme erineva meetodi poolt.

Katvusel põhineva meetodiga saab tuvastada vaid deletsioone ja duplikatsioone. Joonisel 3 on näha, et alates pikkusest 5 000 nukleotiidi on valdav enamus SV-sid tuvastatud katvuse meetodil. Lühemate kui 5 000 nukleotiidi pikkade SV-de korral on ainult katvus meetodiga tuvastatavad SV-d loetud ebakvaliteetseks ja seega pole neid kaasatud analüüsi[14]. Mustriotsinguga tuvastatakse ME-sid, mis kuuluvad eelkõige insertioonide kategooriasse. Muude meetoditega saab tuvastada eelkõige lühemaid SV-sid SV tüübis sõltumata.



Joonis 3: SV pikkuse jaotus tüübi ja tuvastamismeetodi kaupa. Katkendliku joonega on märgitud pikkus 5 000 nukleotiidi.

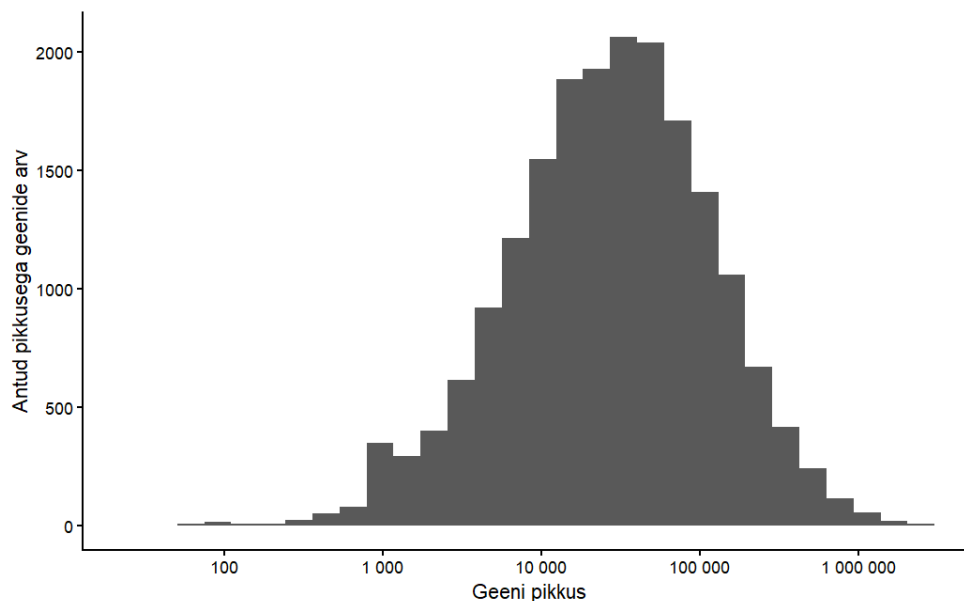


Joonis 4: SV pikkuse jaotus tüübi ja tuvastamismeetodi kaupa. Katkendliku joonega on märgitud pikkus 5 000 nukleotiidi. Detailsem joonis pikkuste vahemiku 4 000—7 000 nukleotiidi jaoks

2.2 Geenide andmebaas GENCODE

Kasutatud geenide asukohad on GENCODE versiooni 49 andmebaasist, mis on pärit aastast 2022[15]. Kokku on andmestikus 78 691 geeni. Töös on kasutatud vaid valku kodeerivaid geene, mida oli kõikide kromosoomide peale 20 097 ja pärast X ja Y kromosoomide eemaldamist 19 156.

Joonisel 5 on toodud valku kodeerivate geenide pikkuste jaotus. Pikkuste jaotus järgib ligikaudu lognormaalset jaotust. Keskmise geeni pikkus on 69 477 nukleotiidi ja 99,5% geenide pikkustest jääb vahemikku 100—1 000 000 nukleotiidi.



Joonis 5: Valku kodeerivate geenide pikkuste jaotus.

2.3 eQTL andmebaas GTEx

Töös kasutatud eQTL-id on võetud avalikult kättesaadavast andmestikust GTEx (Genotype-Tissue Expression [16]) versioon 10. GTEx sisaldab geeniekspressiooni ja eQTL andmeid ligikaudu 1 000 indiviidi ja 54 koe kohta. Iga eQTL on andmestikus esitatud kui üks genoomi positsioon, mis on defineeritud kahe arvu abil: kromosoomi number ja eQTL koordinaat kromosoomil. Keskmiselt on andmestikus 2 031 160 unikaalset eQTL-i positsiooni koe kohta. Töös on vaadeldud vaid valku kodeerivaid geenide eQTL-e, mida oli andmestikus keskmiselt 1 241 714 unikaalset positsiooni koe kohta. Töös kasutati kõiki eQTL-e, mis esinesid vähemalt ühes koes. Sellele tingimusele vastavaid eQTL positsioone oli andmestikus kokku 5 442 410, millest analüüsi jäi 4 494 984 valku kodeeriva geeniga seotud genoomipositsiooni.

3 Metoodika

3.1 B-splainid

Seos struktuurse variandi pikkuse ja tema esinemissageduse vahel on keeruline ja halvasti modelleeritav lineaarse või polünoomse seose abil. Selle tõttu on vaja kasutada paindlikemaid mudeleid. Üks võimalusi sobivama mudeli koostamiseks on kasutada regressioonisplaine (*regression splines*), mis on tükkaaval defineeritud polünoomid. Käesolev peatükk põhineb Thomas W. Yee õpikul [17].

Igal splainil on määratud hulk sõlmpunkte (*knots*), mis jagavad defineerimispiirkonna alamosadeks, kus igas piirkonnas kirjeldab funktsiooni eraldi polünoom. Seejuures säilitatakse funktsiooni ja selle tuletiste pidevus kuni etteantud järguni igas sõlmpunktis.

Praktikas tähendab see, et tundmatu funktsioon esitatakse lineaarse kombinatsioonina splainide baasfunktsioonidest $B_{s,q}(x)$:

$$f(x) = \sum_s \beta_s B_{s,q}(x),$$

kusjuures q on splaini järk ning β_s on baasfunktsioonidele vastavad kordajad, mis hinnatakse andmete põhjal.

B-splainid on üks sagedamini kasutatavaid splainide klasse nende hea numbrilise stabiilsuse tõttu. Splaini siledus sõltub valitud järgust Q , kus kõige levinum valik on $Q = 4$, mis vastab kuupsplainile.

Olgu ξ_s , kus $s = 1, \dots, K$, sisemised sõlmpunktid ning ξ_0 ja ξ_{K+1} kaks äärmist sõlmpunkti.

Seejärel laiendatakse sõlmpunktide hulka veel $2Q$ elemendi võrra, et saada vektor

$$\tau = (\tau_1, \dots, \tau_{K+2Q}),$$

mis rahuldab järgmisi tingimusi:

$$\begin{aligned} \tau_1 &\leq \tau_2 \leq \dots \leq \tau_Q \leq \xi_0 \\ &< \xi_1 \leq \dots \leq \xi_K \\ &< \xi_{K+1} \leq \tau_{K+Q+1} \leq \dots \leq \tau_{K+2Q}. \end{aligned}$$

kus

$$\tau_{Q+s} = \xi_s, \quad s = 1, \dots, K.$$

Tavaliselt valitakse

$$\tau_1 = \dots = \tau_Q = \xi_0 \quad \text{ja} \quad \tau_{K+Q+1} = \dots = \tau_{K+2Q} = \xi_{K+1}.$$

Tähistame $B_{s,q}(x)$ kui s -ndat ja q järku (astmega $q-1$) B-splaini baasfunktsiooni sõlmejada τ korral, kus $q = 1, \dots, Q$.

Need defineeritakse rekursiivselt.

$$s = 1, \dots, K + 2Q - q$$

(1) Kui $q = 1$, siis:

$$B_{s,1}(x) = \begin{cases} 1, & \tau_s \leq x < \tau_{s+1}, \\ 0, & \text{muidu.} \end{cases} \quad (1)$$

(2) Kui $q > 1$, siis

$$\begin{aligned} B_{s,q}(x) &= \omega_{s,q} B_{s,q-1}(x) + (1 - \omega_{s+1,q}) B_{s+1,q-1}(x) \\ &= \frac{x - \tau_s}{\tau_{s+q-1} - \tau_s} B_{s,q-1}(x) + \frac{\tau_{s+q} - x}{\tau_{s+q} - \tau_{s+1}} B_{s+1,q-1}(x). \end{aligned} \quad (2)$$

Seega moodustatakse kõrgemat järku splineid madalamat järku baasfunktsioonide lineaarse kombinatsioonina ning hinnatakse nende baasfunktsioonide kordajad β_s .

3.2 Splaini tuletis

Splaini kasvamis- ja kahanemispiirkondade määramiseks uuritakse selle esimest tuletist. Kui $f'(x) > 0$, siis funktsioon kasvab, ning kui $f'(x) < 0$, siis funktsioon kahaneb. Kuna B-splainid on tükkhaaval polünoomid, siis on ka nende tuletisfunktsioonid tükkhaaval polünoomid, mistõttu kehtib võrdus:

$$f'(x) = \sum_s \beta_s B'_{s,q}(x),$$

Järelikult tuletise leidmiseks tuleb leida baasfunktsiooni tuletise, mis on B-splainide puhul:

$$B'_{s,q}(x) = \frac{q-1}{\tau_{s+q-1} - \tau_s} B_{s,q-1}(x) - \frac{q-1}{\tau_{s+q} - \tau_{s+1}} B_{s+1,q-1}(x)$$

[18]

Kui fikseerida soovitud $Q = 4$, siis saab valemi viia mitterekursiivsele kujule kasutades avaldist (2).

$$\begin{aligned} B'_{s,4}(x) &= \frac{3}{\tau_{s+3} - \tau_s} B_{s,3}(x) - \frac{3}{\tau_{s+4} - \tau_{s+1}} B_{s+1,3}(x) \\ &= \frac{3}{\tau_{s+3} - \tau_s} \left(\frac{x - \tau_s}{\tau_{s+2} - \tau_s} B_{s,2}(x) + \frac{\tau_{s+3} - x}{\tau_{s+3} - \tau_{s+1}} B_{s+1,2}(x) \right) \\ &\quad - \frac{3}{\tau_{s+4} - \tau_{s+1}} \left(\frac{x - \tau_{s+1}}{\tau_{s+3} - \tau_{s+1}} B_{s+1,2}(x) + \frac{\tau_{s+4} - x}{\tau_{s+4} - \tau_{s+2}} B_{s+2,2}(x) \right) \\ &= \frac{3}{\tau_{s+3} - \tau_s} \left[\frac{x - \tau_s}{\tau_{s+2} - \tau_s} \left(\frac{x - \tau_s}{\tau_{s+1} - \tau_s} B_{s,1}(x) + \frac{\tau_{s+2} - x}{\tau_{s+2} - \tau_{s+1}} B_{s+1,1}(x) \right) \right. \\ &\quad \left. + \frac{\tau_{s+3} - x}{\tau_{s+3} - \tau_{s+1}} \left(\frac{x - \tau_{s+1}}{\tau_{s+2} - \tau_{s+1}} B_{s+1,1}(x) + \frac{\tau_{s+3} - x}{\tau_{s+3} - \tau_{s+2}} B_{s+2,1}(x) \right) \right] \\ &\quad - \frac{3}{\tau_{s+4} - \tau_{s+1}} \left[\frac{x - \tau_{s+1}}{\tau_{s+3} - \tau_{s+1}} \left(\frac{x - \tau_{s+1}}{\tau_{s+2} - \tau_{s+1}} B_{s+1,1}(x) + \frac{\tau_{s+3} - x}{\tau_{s+3} - \tau_{s+2}} B_{s+2,1}(x) \right) \right. \\ &\quad \left. + \frac{\tau_{s+4} - x}{\tau_{s+4} - \tau_{s+2}} \left(\frac{x - \tau_{s+2}}{\tau_{s+3} - \tau_{s+2}} B_{s+2,1}(x) + \frac{\tau_{s+4} - x}{\tau_{s+4} - \tau_{s+3}} B_{s+3,1}(x) \right) \right] \end{aligned}$$

Kuna $q = 1$ korral kehtib võrdus (1), kui $\tau_s \leq x < \tau_{s+1}$, siis on mitte nullid ainult baasifunktsioonid $B'_{s-3,4}(x), B'_{s-2,4}(x), B'_{s-1,4}(x), B'_{s,4}(x)$.

$$B'_{s-3,4}(x) = -\frac{3}{\tau_{s+1} - \tau_{s-2}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_{s-1}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_s}$$

$$B'_{s-2,4}(x) = \frac{3}{\tau_{s+1} - \tau_{s-2}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_{s-1}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_s} - \frac{3}{\tau_{s+2} - \tau_{s-1}} \left(\frac{x - \tau_{s-1}}{\tau_{s+1} - \tau_{s-1}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_s} + \frac{\tau_{s+2} - x}{\tau_{s+2} - \tau_s} \frac{x - \tau_s}{\tau_{s+1} - \tau_s} \right)$$

$$B'_{s-1,4}(x) = \frac{3}{\tau_{s+2} - \tau_{s-1}} \left(\frac{x - \tau_{s-1}}{\tau_{s+1} - \tau_{s-1}} \frac{\tau_{s+1} - x}{\tau_{s+1} - \tau_s} + \frac{\tau_{s+2} - x}{\tau_{s+2} - \tau_s} \frac{x - \tau_s}{\tau_{s+1} - \tau_s} \right) - \frac{3}{\tau_{s+3} - \tau_s} \frac{x - \tau_s}{\tau_{s+2} - \tau_s} \frac{x - \tau_s}{\tau_{s+1} - \tau_s}$$

$$B'_{s,4}(x) = \frac{3}{\tau_{s+3} - \tau_s} \frac{x - \tau_s}{\tau_{s+2} - \tau_s} \frac{x - \tau_s}{\tau_{s+1} - \tau_s}$$

.

Valemeid kasutades saab teades sõlmi (τ_i) ja splaini parameetreid (β_i) arvutada splaini tuletis kohal x . Lisaks kuna fikseeritud x korral on tegu parameetrite lineaarse kombinatsiooniga on lihtne välja arvutada ka tuletise usaldusintervalli.

4 Analüüs ja tulemused

4.1 Töö eesmärk

On teada, et geenid paiknevad SV-d on võrreldes väljaspool geene paiknevate SV-dega tugevama puhastava valiku all [11]. Seetõttu on geeniga ülekattes olevad SV-d ka keskmiselt harvemad. Vähem on aga uuritud eQTL piirkondadega ülekattes olevaid SV-sid ning neile mõjuvat evolutsioonilist survet ja sagedusi.

Käesoleva bakalaureusetöö eesmärk on uurida, kuidas SV paiknemine eQTL piirkonnaga ülekattes mõjutab tema haruldust. Töös mudeldatakse eri funktsionaalsetesse gruppidesse kuuluvate SV-de haruldust ja vaadeldakse, kas eQTL-iga kattuvate SV-de hulgas on harvade variatsioonide osakaal sarnasem geeniga ülekattes paiknevate või väljaspool geeni/eQTLi paiknevate harvade SV-de osakaaludega. Lisaks tuleb arvesse võtta SV pikkuse kui segava tunnuse. Pikemad SV-d on haruldasemad ning suurema tõenäosusega geenide või eQTLidega ülekattes. Töö tulemused annavad võimaluse hinnata puhastava valiku rolli eQTL piirkondades.

4.2 Andmete esmane töötlemine ja kokkupanemine

Esimene ülesanne oli töös kasutatavad andmestikud kombineerida ja viia mudeldamiseks sobivale kujule. GnomAD andmestikus SV-de algandmestik oli VCF (Variant Call Format [19]) formaadis, mis on sagedasti kasutatav inimloetav failiformaat geneetiliste variantide esitamiseks. Esmane andmete töötlemine tehti Tartu Ülikooli HPC arvutiklastri `rocket.hpc.ut.ee` ning selleks kasutati tööriista `bcftools` [20]. Klastri filtreeriti andmetest välja analüüsi jäävad andmed (vt peatükk 2.1) ja konverteeriti SV andmestik tabeli

formaati, kuhu jäeti ainult mudeldamiseks vajaminevad tunnused. Kõik edasised sammud tehti programmeerimiskeeles R [21] väljaspool HPC keskkonda. Igale SV-le lisati geenide andmebaasist GENCODE juurde tunnus „geenis”, mis kirjeldab, kas vaadeldav SV kattub geeniga või mitte. Sarnaselt lisati, kas on kattumist ühegi koe eQTL-iga. Selleks lisati GTEx andmestikus toodud ühenukleotiidssele eQTL positsioonile ± 100 piirkond, sest tegelikkuses võib eQTL-i täpne positsioon varieeruda.

Mudeldamiseks kasutati üldsageduse asemel tunnust „haruldane”, mis näitab, kas variandi sagedus on suurem või väiksem kui 0,1%. Lisaks pandi kokku duplikatsioonid ja insertioonid, sest tegu on sarnast tüüpi SV-dega ja määramismeetodid ei suuda neid alati hästi eristada. Käesolevas töös nimetatakse neid edaspidi lihtsalt insertioonideks. Peale kirjeldatud tunnuste lisamist alles jäänud tunnused on toodud tabelis 2.

Tabel 2: Tunnused andmetes.

Tunnus	Tunnuse seletus
kromosoom	Kromosoom, kus SV asub
algus	SV alguskoordinaat referentsgenoomi suhtes
lõpp	SV lõppkoordinaat referentsgenoomi suhtes
pikkus	SV pikkus
tüüp	Kas tegu on insertiooni, duplikatsiooni või deletsiooniga
üldsagedus	SV esinemissagedus gnomAD andmestikus
geenis	Kas SV on ülekattes geeniga
eQTL	Kas SV on ülekattes eQTL-iga
tüüp2	Täpsem SV tüüp, kas tegu on mobiilse elemendiga (nt ALUga)
meetod	Meetod, millega SV leiti
haruldane	Kas SV sagedus on suurem kui 0,1%

4.3 SV-de pikkuste jaotused funktsionaalsetes gruppides

Kuna taheti uurida, kuidas SV sagedus on seotud selle paiknemisega, ehk geeni või eQTL-iga ülekattega, siis on SV-d jaotatud nelja nn funktsionaalsesse gruppi:

- SV-d, mis on ülekattes nii geeni kui ka eQTL-iga,
- SV-d, mis on ülekattes ainult geeniga,
- SV-d, mis on ülekattes ainult eQTL-iga,
- SV-d, millel pole ülekatet ei geeni ega eQTL-iga.

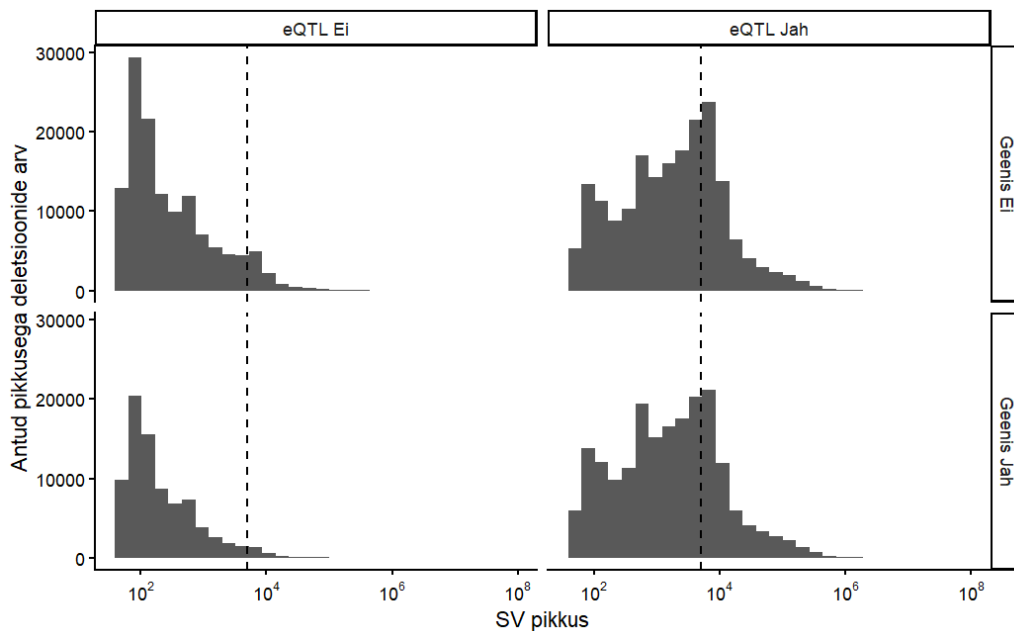
Tabelis 3 on, mitu insertiooni ja deletsiooni kuulub funktsionaalsesse gruppi.

Tabel 3: Deletsioonide ja insertioonide hulk igas funktsionaalses grupis.

	DEL		INS	
	eQTL Ei	eQTL Jah	eQTL Ei	eQTL Jah
Geenis Ei	128 477	192 775	168 118	106 460
Geenis Jah	81 217	196 325	130 261	125 027

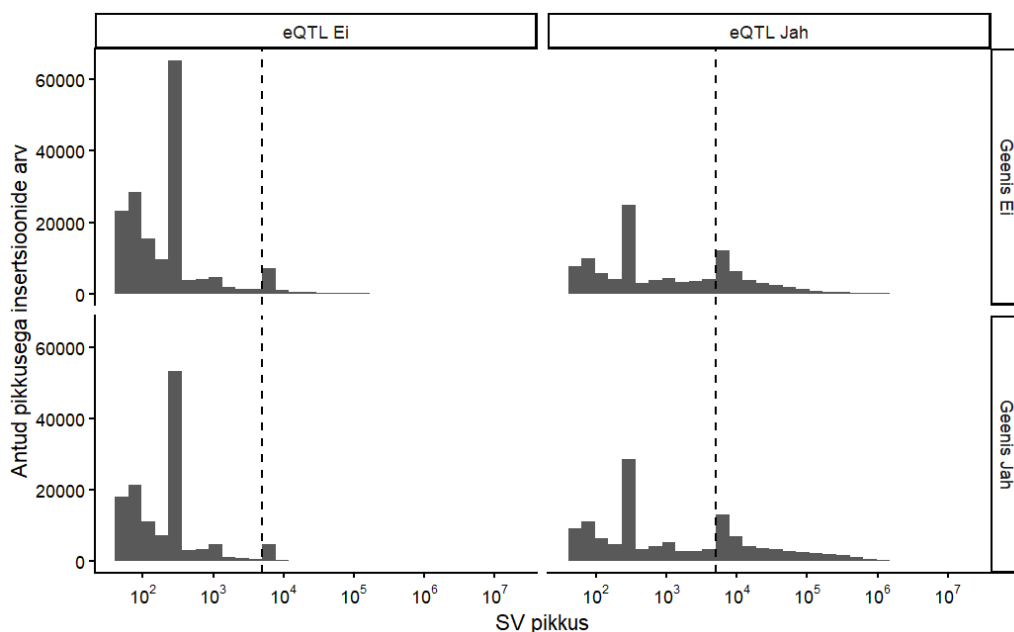
Kuna SV pikkus on segavaks tunnuseks analüüsis, siis esmalt uuriti selle seost funktsionaalsete gruppidega. Joonistel 6 ja 7 on toodud deletsioonide ja insertioonide pikkuste jaotused nii eQTL ja geeniga ülekattes olevate kui ka ülekattes mitteolevate SVde jaoks. Histogrammide põhjal ei ole näha tugevat seost SV pikkuse ja geeniga ülekattumise vahel. Siiski deletsioonidel näitab Mann–Whitney U test, et kui testida mõlemas eQTL-ülekatte grupis eraldi, siis erinevad pikkuse jaotused geeniga ülekattuvate ja mitteülekattuvate SV-de vahel statistiliselt oluliselt ($p < 2,2 \cdot 10^{-16}$). Insertioonide puhul on sama tulemus eQTL-iga ülekattes olevas grupis, kuid grupis, kus eQTL-iga ülekatet

ei ole, ei näita test statistiliselt olulist erinevust pikkuse jaotustes geeniga ülekattuvate ja mitteülekattuvate SV-de vahel ($p = 0,446$). eQTL-iga ülekatte gruppides on näha suuremat erinevust. Nimelt SV-d, mis ei ole eQTL-iga ülekattes, on keskmiselt lühemad võrreldes nendega, mis on eQTL-iga ülekattes. Suurem SV pikkuste erinevus eQTL-ide puhul võib olla tingitud sellest, et pikkuse suurenedes on suurem võimalus SV-l eQTL-iga ülekattes sattuda, vahet pole näha tugevalt geenide puhul, sest neid on kordades vähem.



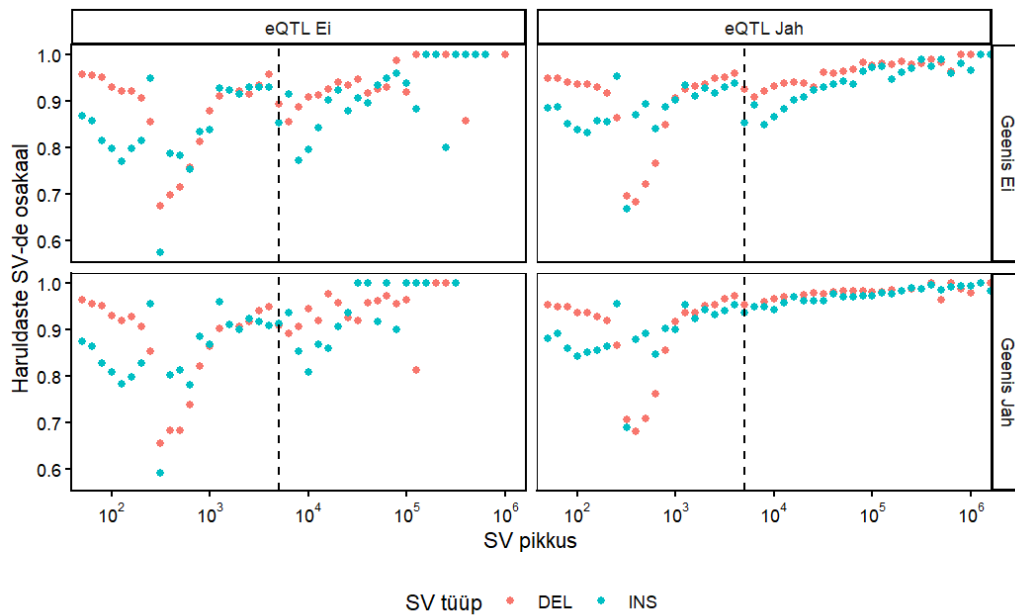
Joonis 6: Deletsioonide pikkuste jaotus funktsionaalsetes gruppides.

Insertsioonidel on kaks suurema sagedusega pikkust. Esimene on umbes 300 nukleotiidi juures, mis vastab ALU pikkusele, ja millest enamus pole ülekattes eQTL-idega. Palju insertsioone on ka pikkusega natuke üle 5 000 nukleotiidi, mis vastab L1 tüüpi mobiilsetele elementidele.



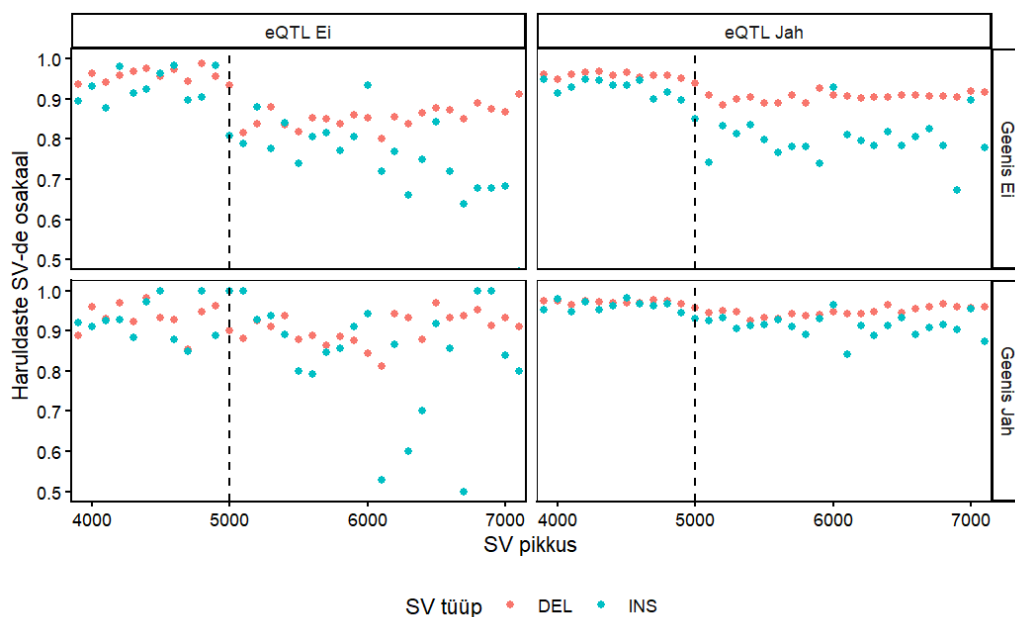
Joonis 7: Insertioonide pikkuste jaotus funktsionaalsetes gruppides.

Joonisel 8 on toodud haruldaste (sagedus $< 0,1\%$) SV-de osakaal iga funktsionaalse grupi kohta. Grupist sõltumata on insertioonidel järsk muutus ALU pikkuse juures. Sarnases piirkonnas on deletsioonidel ka väiksem haruldaste variatsioonide osakaal, aga seda laiemas vahemikus (250–1 000 nukleotiidi). Enne pikkust 5 000 nukleotiidi on vähe erinevust funktsionaalsetes gruppides, kuid 6 000 nukleotiidi juures on näha tugevamat vahet deletsioonidel, mis vastab L1 mobiilsetele elementidele ning natuke pikemate puhul ka insertioonidel. Ainsaks erandiks on grupp, kus SV-l on ülekaal nii geeni kui ka eQTL-iga, kus nimetatud SV pikkuse juures haruldaste SV-de osakaal oluliselt ei muutu. Erinevust haruldaste SV-de osakaaludes pole näha või ei saa hinnata ka pikemate kui 30 000 nukleotiidi SV-de puhul kõigis gruppides.



Joonis 8: Haruldaste SV-de osakaal sõltuvalt pikkusest igas funktsionaalses grupis, pikkus kuni 10^6 .

Selleks, et näha paremini, mis toimub haruldaste variantide osakaaluga pikkuse 5 000 ümber on toodud joonisel 9 haruldaste variantide osakaal pikkuste lõigus 4 000—7 000 nukleotiidi. Geeniga mitte ülekattes olevatel struktuursetel variantidel on näha selgelt muutust pikkuse 5 000 nukleotiidi juures. Igas grupis on näha haruldaste variantide osakaalu tõusu pikkuse 6 000 nukleotiidi juures.



Joonis 9: Haruldaste SV-de osakaal sõltuvalt pikkusest igas funktsionaalses grupis. Detailsem joonis pikkuste vahemiku 4 000—7 000 nukleotiidi jaoks.

4.4 Mudel SV harulduse hindamiseks

Nagu peatükis 1.5 kirjeldatud, on SV haruldus sobiv tunnus hindamaks puhastava valiku tugevust erinevates genoomipiirkondades. Töös tehti selle tunnuse kirjeldamiseks kaks mudelit, üks deletsioonide ja teine insertioonide jaoks. Mudeldamisel kasutati logistilist regressiooni ja lisati sõltuvate tunnustena logaritmitud SV pikkus ja ülekate geeni ning eQTL-iga.

Eelmise alapeatüki tulemustest on näha, et SV harulduse sõltuvust SV pikkusest ei ole võimalik lihtsa polünoomi abil kirjeldada. Seetõttu kasutati pikkuse modelleerimiseks b-splaine, kusjuures iga asukoha (geeni ülekate, eQTLiga ülekate, ülekate mõlemaga või ülekate kummagagi pole) jaoks hinnati erinev

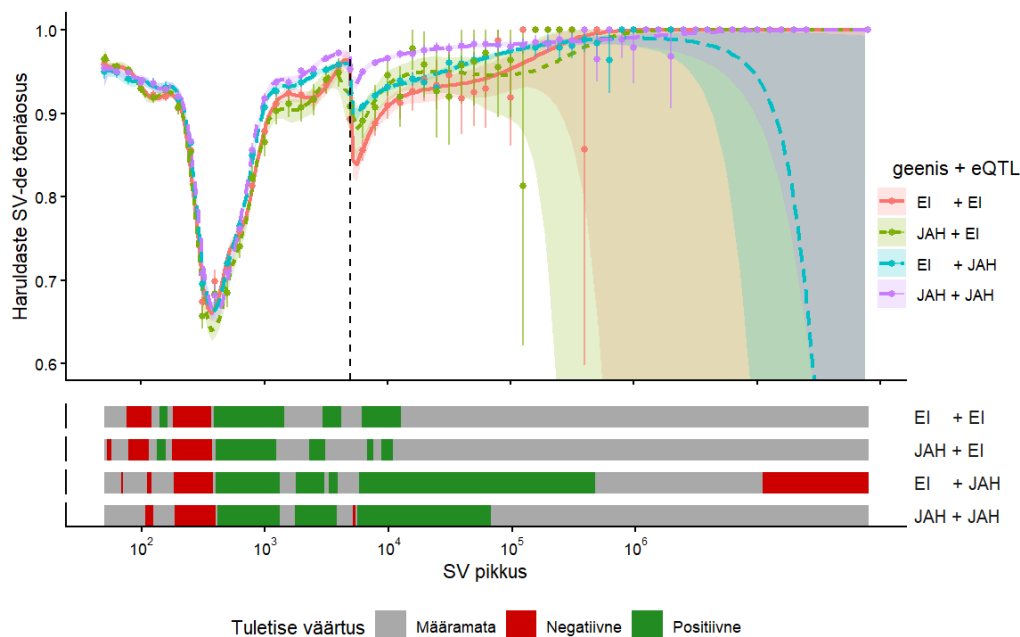
b-splain lisades mudelisse koosmõjud asukoha ja SV pikkuse põhjal arvutatud abitunnuste (splainide baasifunktsioonide) vahele. Kuna pikkuse 5 000 juures muutus SV-de otsimiseks kasutatav metodoloogia, siis lisati mudelisse eraldi ka indikaatorfunktsiooni $I(\text{pikkus} > 5\,000)$ ja SV asukoha koosmõju. Selle SV pikkuse juures toimub oluline muutus SV tuvastamise meetodite tundlikkuses (vt peatükk 2.1) ning töös taheti lubada, et vajadusel saaks spline olla selles punktis katkev. Sobilike splinei sõlmpunktide arvu leidmiseks kasutati AIC-i informatsioonikriteeriumit. Alustati kahe sisemise sõlmpunktiga ja lisati kuni vaadeldavast väärtusest järgmised kaks mudelit olid halvema AIC-i väärtusega.

4.4.1 Deletsioonide harulduse modelleerimine

Deletsioonide mudelis kasutati spline 14 sisemise sõlmega. Mudel on toodud lisas 1. Joonisel 10 on toodud mudeli prognoosid koos haruldaste SV-de osakaaluga andmetes ja splinei tuletisega. Lühemate (pikkus $< 5\,000$ nukleotiidi) deletsioonide hulgas on haruldaste variantide osakaal kõigis funktsionaalsetes gruppides sarnane. Eraldi joonistub välja väiksem harulduse osakaal pikkusega 200-1 000 nukleotiidi deletsioonide hulgas. See on ootuspärane ning vastab ALU elementide suurele sagedusele inimese genoomis.

Pikemate ($\geq 5\,000$ nukleotiidi) deletsioonide korral eristub grupp, mille SV-d on ülekattes nii geeni kui ka eQTL-iga ning, kus haruldaste deletsioonide osakaal on teistest gruppidest oluliselt suurem. Samuti SV-de grupil, millel pole ülekattet ei geeni ega eQTL-iga, on väiksem haruldaste osakaal. Näiteks on mõlema ülekatttega grupis pikkuse 7 000 nukleotiidi juures haruldaste variantide osakaalu 0,95 usaldusvahemik (0,953; 0,961) ja grupis, kus pole ülekatted ei geeni ega eQTL-iga, on usaldusvahemik (0,860; 0,887). Ülejäänud kaks gruppi üksteisest oluliselt ei erine.

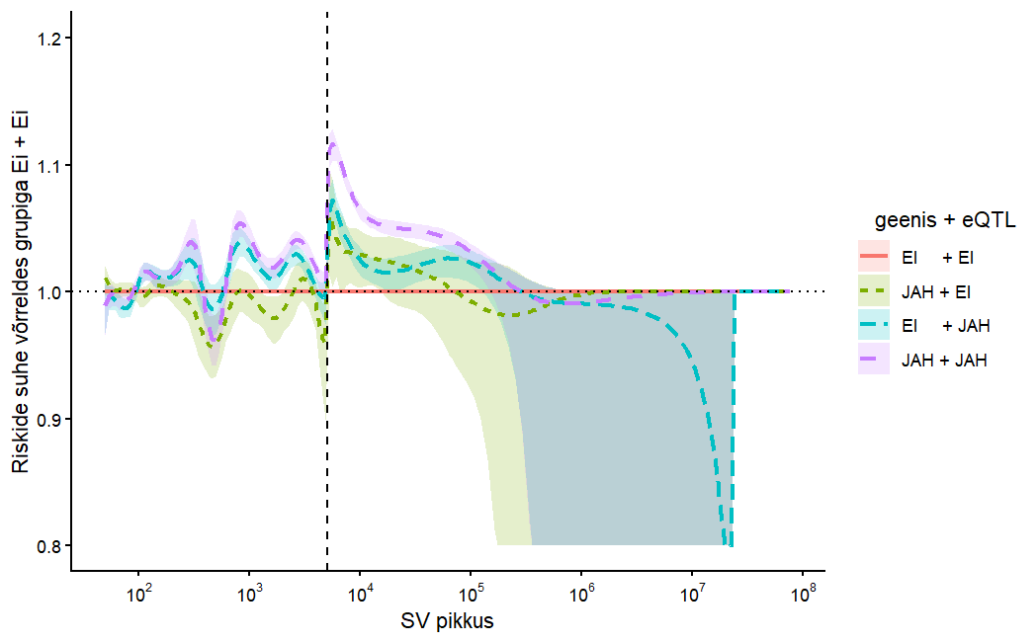
Funktsiooni tuletise abil on selgelt näidatud uuritava mudeli kasvamis- ja kaanemispriirkonnad. Näha on tõusu ja languset pikkuste 300—1 000 nukleotiidi juures, pikkuse 3 000 nukleotiidi juures ja peale pikkust 6 000 nukleotiidi, kus pikemat tõusu on näha gruppides, millel on ülekate eQTL-iga.



Joonis 10: Mudeli prognoosid deletsioonide haruldusele igas grupis koos tuletisega ning haruldaste SV-de osakaaluga andmetes.

Gruppide selgemaks eristamiseks on joonisel 11 toodud riskide suhe baasgrupiga, kus deletsioonid pole ülekattes ühegi geeniga ega eQTL-iga. Lühemate deletsioonide puhul on näha, et ainult geeniga ülekatte puhul on tõenäosuste suhtel erinevus väike ja piirkondades üldiselt statistiliselt mitte oluline, kuid eQTL-iga ülekattet sisaldavad kaks gruppi on mõlemad enamustel pikkustel baasgrupist oluliselt erinevad. Näiteks pikkuse 3 000 juures on mõlema ülekattega grupi riskidesuhte 0,95 usaldusvahemik (1,033; 1,045)

Pikemate deletsioonide puhul on pikkuse vahemikus 5 000—200 000 nukleotiidi näha, et eQTL-iga ülekatte efekt on statistiliselt oluline geeniga ülekattest sõltumatult. Lisaks on näha, et deletsioonid, mis on ülekattes geeniga, aga samaaegselt ka mõne eQTL-iga, on suurema osakaalu haruldaste variantidega kui ainult geeniga ülekattes olevad deletsioonid. Pikemate kui 200 000 nukleotiidi deletsioonide korral pole erinevused tuvastatavad eelkõige andmete vähesuse tõttu.



Joonis 11: Deletsioonide harulduse riskide suhe grupi $E_i + E_i$ suhtes.

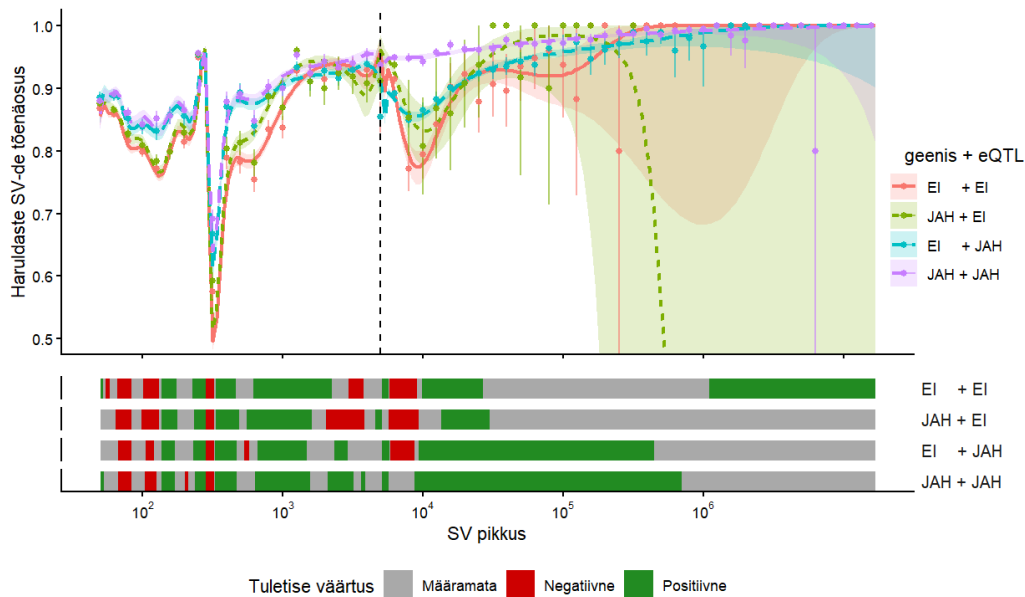
4.4.2 Insertioonide harulduse modelleerimine

Analoogiliselt deletsioonidele tehti eraldi mudel ka insertioonide harulduse modelleerimiseks. Insertioonidel on väga järsk harvade variatsioonide osakaalu muutus ALU pikkuse (ligikaudu 300 nukleotiidi) juures, mis jääb splainidega modelleerimisel halvasti kirjeldatuks. Äkilisema muutuse lubamiseks

mudelis lisati igale SV-le vastavalt tema pikkusele sellise normaaljaotuse tihedusfunktsiooni väärtus, mille parameetrid oli ALU-tüüpi SV-de logaritmitud pikkuse keskmine (2,420) ja standardhälve (0,1011). Uus tunnus pandi mudelisse ilma koosmõjudeta. Mudeli splaini sisemiste sõlmede arvuks võeti AIC-i järgi 18 ning saadud mudel on lisas [2](#).

Saadud mudeli prognoosid on näha joonisel [12](#) koos haruldaste insertsiioonide osakaaluga piirkonnas. Kõigis gruppides on haruldaste SV-de osakaal suurem pikkuse 270 juures ning ootuspäraselt on osakaalude lokaalsed miinimumid ALU ja L1 elementide ümbruses.

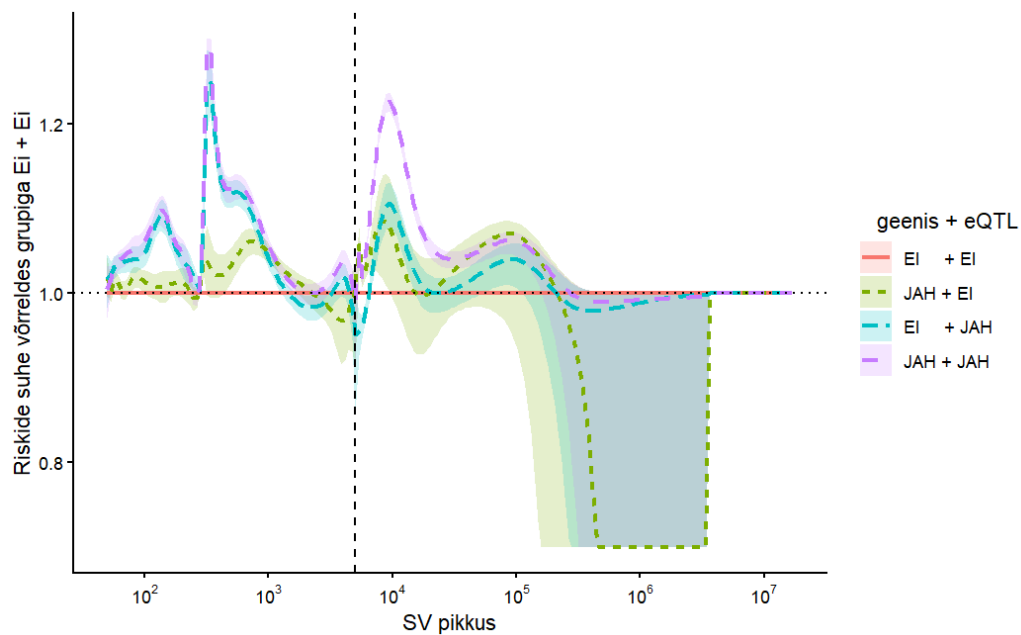
Tuleti kirjeldab kõige paremini erinevust pikkuse 6 000 ümber: mõlema ülekattega gruppis ei toimu haruldaste variantide osakaalu vähenemist kuid kõigis teistes gruppides see toimub. Näha on ka, et pikkuse 3 000 juures eQTL-iga ülekattes olevatel insertsiioonidel harvade variantide osakaal langeb ja teisel kahel grupil osakaal tõuseb.



Joonis 12: Mudeli prognoosid insertioonide haruldusele igas grupis koos tuletisega ning haruldaste SV-de osakaaluga andmetes.

Riskide suhe võrdluses SV-de grupiga, millel pole ülekatet geeni ega eQTL-iga, on toodud joonisel 13. Lühemate kui 2000 nukleotiidi insertioonide puhul on eQTL-iga ülekattes olevad variandid oluliselt suurema haruldaste osakaaluga kui baasgrupis. Vahe on näha ka grupis, kus on ülekate ainult geeniga, kuid see pole suuremas osas piirkonnast statistiliselt oluline.

Sarnane trend on näha ka pikkusega 5 000—200 000 nukleotiidi insertioonide puhul. Siiski riskide suhe grupiga, kus on ülekate ainult eQTL-iga on piirkonnas oluliselt väiksem, kui grupiga, kus on ülekate eQTL-i ja geeniga. Näiteks pikkse 7 000 juures on vastavate riskidesuhete 0,95 usaldusintervallid (1,016; 1,049) ja (1,121; 1,137)



Joonis 13: Insertioonide harulduse riskide suhe grupi $Ei + Ei$ suhtes.

4.5 Tulemuste interpretatsioon

Eelmises peatükis kirjeldatud mudelite põhjal saab esitada järgmiseid väiteid eQTL-iga ülekattes olevate SV-de (nii deletsioonide kui ka insertioonide) kohta:

1. Nad on keskmiselt haruldasemad geeni ja eQTL piirkondadest väljajäävatest SV-dest.
2. Nad on keskmiselt suurema haruldaste osakaaluga ka võrreldes ainult geeniga ülekattes olevate SV-dega.
3. Haruldaste SV-de osakaalud eristuvad kõige selgemini pikkusega 6 000 deletsioonide ja pikkusega 500 insertioonide korral.

Kokkuvõttes saame öelda, et eQTL-iga ülekattes olevad variatsioonid on tugeva puhastava valiku all.

Kokkuvõte

Käesolevas töös uuriti, kas eQTL-idega ülekattes olevad SV-d on puhastava valiku all ning kuidas kuidas SV pikkus on seotud selle haruldusega populatsioonis. Analüüsi kaasati deletsioonid ning insertioonide ja duplikatsioonide ühend autosomaalsetelt kromosoomidelt.

SV harulduse modelleerimiseks kasutati logistilist regressiooni koos B-splainidega, mis võimaldas kirjeldada pikkuse ja harulduse vahelist mittelineaarset seost. Mõlema SV tüübi jaoks koostati eraldi mudel 18 vabadusastmega splineidega. Insertioonide mudelisse lisati täiendavalt ALU-tüüpi mobiilsete elementide pikkuspiirkonna kirjeldamiseks normaaljaotuse tihedusfunktsiooni väärtus.

Tulemused näitasid, et nii SV pikkus kui ka geeni ja eQTL-iga ülekate on seotud SV haruldusega. Kõige suurem haruldaste variantide osakaal on sellistel SV-del, mis kattuvad nii geeni kui eQTL-iga. Huvitaval kombel on SV-de grupil, millel on ülekate ainult eQTL-iga, suurem haruldaste variantide osakaal kui neil SV-del, millel on ülekate ainult geeniga. Insertioonidel esineb pikkuse 300 juures järsk muutus madalama harulduses suunas, mis on selektatav ALU-tüüpi mobiilsete elementidega. Mõlemal tüübil on pikkuse 5000 juures selge murrangupunkt, millest alates tugevneb geeni ja eQTL-iga kattuvuse mõju haruldusele. Tulemused on kooskõlas puhastava valiku teooriaga — funktsionaalselt olulisteseQTL piirkondades paiknevad SV-d on populatsioonis haruldasemad kui väljaspool gene/eQTL-e paiknevad SV-d.

Kasutatud allikad

- [1] National Human Genome Research Institute. *Human Genomic Variation*. URL: <https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genomic-variation> (vaadatud 29.04.2025).
- [2] National Human Genome Research Institute. *Structural Variation*. URL: <https://www.genome.gov/genetics-glossary/Structural-Variation> (vaadatud 29.04.2025).
- [3] Geòrgia Escaramís, Elisa Docampo ja Raquel Rabionet. “A decade of structural variants: description, history and methods to detect structural variation”. *Briefings in Functional Genomics* 14.5 (2015), lk. 305–314. DOI: [10.1093/bfpg/elv014](https://doi.org/10.1093/bfpg/elv014).
- [4] Colby Chiang *et al.* “The impact of structural variation on human gene expression”. *Nature Genetics* 49.5 (2017), lk. 692–699. DOI: [10.1038/ng.3834](https://doi.org/10.1038/ng.3834).
- [5] David Heller. *SVIM-asm - Structural variant identification method (Assembly edition)*. URL: <https://github.com/eldariont/svim-asm?tab=readme-ov-file> (vaadatud 02.05.2025).
- [6] Guillaume Bourque *et al.* “Ten things you should know about transposable elements”. *Genome Biology* 19 (2018), lk. 199. DOI: [10.1186/s13059-018-1577-z](https://doi.org/10.1186/s13059-018-1577-z).
- [7] Richard Cordaux ja Mark A. Batzer. “The impact of retrotransposons on human genome evolution”. *Nature Reviews Genetics* 10.10 (2009), lk. 691–703. DOI: [10.1038/nrg2640](https://doi.org/10.1038/nrg2640).

- [8] National Human Genome Research Institute. *Human Genome Reference Sequence*. URL: <https://www.genome.gov/genetics-glossary/Human-Genome-Reference-Sequence> (vaadatud 29.04.2025).
- [9] National Human Genome Research Institute. *Gene*. URL: <https://www.genome.gov/genetics-glossary/Gene> (vaadatud 02.05.2025).
- [10] Konrad J. Karczewski, Joel T. Dudley ja *et al.* “Genetic data: The new challenge of personalized medicine, insights for rheumatoid arthritis patients”. *Gene* 583.2 (2016), lk. 90–101. DOI: [10.1016/j.gene.2016.02.004](https://doi.org/10.1016/j.gene.2016.02.004).
- [11] Peter H. Sudmant, Tobias Rausch ja Eugene J. *et al.* Gardner. “An integrated map of structural variation in 2,504 human genomes”. *Nature* 526 (2015), lk. 75–81. DOI: [10.1038/nature15394](https://doi.org/10.1038/nature15394).
- [12] M. Cyrus Maher *et al.* “Population genetics of rare variants and complex diseases”. *Human Heredity* 74.3-4 (2012), lk. 118–128. DOI: [10.1159/000346826](https://doi.org/10.1159/000346826).
- [13] Xuefang *et al.* Zhao. *Structural Variants in gnomAD v4*. 2023. URL: <https://gnomad.broadinstitute.org/news/2023-11-v4-structural-variants/> (vaadatud 02.05.2025).
- [14] *gnomAD v4 Downloads*. URL: <https://gnomad.broadinstitute.org/downloads#v4-structural-variants> (vaadatud 02.05.2025).
- [15] GENCODE. *Human Genome Annotation*. URL: <https://www.encodegenes.org/human/> (vaadatud 29.04.2025).
- [16] *GTEx portaal*. URL: <https://www.gtexportal.org/home/> (vaadatud 29.04.2025).
- [17] Thomas W. Yee. *Vector Generalized Linear and Additive Models: With an Implementation in R*. 2. väljaanne. Springer, 2015.

- [18] Carl de Boor. *A Practical Guide to Splines*. Revised Edition. Springer, 2001.
- [19] GATK Team. *VCF - Variant Call Format*. URL: <https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format> (vaadatud 02.05.2025).
- [20] *bcftools Manual Page*. URL: <https://samtools.github.io/bcftools/bcftools.html> (vaadatud 02.05.2025).
- [21] *The R Project for Statistical Computing*. URL: <https://www.r-project.org/> (vaadatud 02.05.2025).

1 Lisa 1: Mudel deletsioonidele

Call:
glm(formula = haruldane ~ (geenis * eQTL * after_bp) * bs(log10_pikkus,
df = 18), family = binomial, data = del_v2)

Coefficients: (60 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.19369	0.11819	27.022	< 2e-16 ***
geenisTRUE	0.31301	0.19336	1.619	0.105488
eQTLTRUE	-0.20896	0.20620	-1.013	0.310891
after_bp	18.40099	28.35792	0.649	0.516414
bs(log10_pikkus, df = 18)1	-0.20169	0.22179	-0.909	0.363153
bs(log10_pikkus, df = 18)2	-0.03266	0.14628	-0.223	0.823322
bs(log10_pikkus, df = 18)3	-0.21433	0.15624	-1.372	0.170134
bs(log10_pikkus, df = 18)4	-0.73191	0.13133	-5.573	2.50e-08 ***
bs(log10_pikkus, df = 18)5	-0.84596	0.14066	-6.014	1.81e-09 ***
bs(log10_pikkus, df = 18)6	-0.41623	0.13763	-3.024	0.002493 **
bs(log10_pikkus, df = 18)7	-3.10598	0.13407	-23.167	< 2e-16 ***
bs(log10_pikkus, df = 18)8	-2.06556	0.13035	-15.846	< 2e-16 ***
bs(log10_pikkus, df = 18)9	-2.07777	0.13289	-15.636	< 2e-16 ***
bs(log10_pikkus, df = 18)10	-0.78310	0.15332	-5.107	3.26e-07 ***
bs(log10_pikkus, df = 18)11	-0.64041	0.17416	-3.677	0.000236 ***
bs(log10_pikkus, df = 18)12	-0.90653	0.20055	-4.520	6.18e-06 ***
bs(log10_pikkus, df = 18)13	-0.21149	0.29815	-0.709	0.478113
bs(log10_pikkus, df = 18)14	0.94161	0.89915	1.047	0.295001
bs(log10_pikkus, df = 18)15	-19.26364	28.35781	-0.679	0.496944
bs(log10_pikkus, df = 18)16	-18.02519	28.39157	-0.635	0.525508
bs(log10_pikkus, df = 18)17	-22.05508	29.61674	-0.745	0.456463
bs(log10_pikkus, df = 18)18	36.24663	75.78692	0.478	0.632458
geenisTRUE:eQTLTRUE	-0.35162	0.30264	-1.162	0.245303
geenisTRUE:after_bp	-34.95581	48.17346	-0.726	0.468069
eQTLTRUE:after_bp	-9.92173	31.14711	-0.319	0.750072
geenisTRUE:bs(log10_pikkus, df = 18)1	-0.30592	0.35634	-0.859	0.390606
geenisTRUE:bs(log10_pikkus, df = 18)2	-0.29626	0.23302	-1.271	0.203596
geenisTRUE:bs(log10_pikkus, df = 18)3	-0.22625	0.25242	-0.896	0.370092
geenisTRUE:bs(log10_pikkus, df = 18)4	-0.44066	0.21185	-2.080	0.037520 *
geenisTRUE:bs(log10_pikkus, df = 18)5	-0.18011	0.22732	-0.792	0.428172
geenisTRUE:bs(log10_pikkus, df = 18)6	-0.35978	0.22192	-1.621	0.104971
geenisTRUE:bs(log10_pikkus, df = 18)7	-0.32113	0.21711	-1.479	0.139103
geenisTRUE:bs(log10_pikkus, df = 18)8	-0.53632	0.21199	-2.530	0.011407 *
geenisTRUE:bs(log10_pikkus, df = 18)9	-0.27696	0.21687	-1.277	0.201571
geenisTRUE:bs(log10_pikkus, df = 18)10	-0.31358	0.25283	-1.240	0.214878
geenisTRUE:bs(log10_pikkus, df = 18)11	-0.71801	0.29065	-2.470	0.013497 *
geenisTRUE:bs(log10_pikkus, df = 18)12	-0.19315	0.35226	-0.548	0.583488
geenisTRUE:bs(log10_pikkus, df = 18)13	0.02007	0.54168	0.037	0.970448
geenisTRUE:bs(log10_pikkus, df = 18)14	-3.19057	1.56604	-2.037	0.041616 *
geenisTRUE:bs(log10_pikkus, df = 18)15	35.03558	48.17263	0.727	0.467047
geenisTRUE:bs(log10_pikkus, df = 18)16	35.14104	48.27642	0.728	0.466667
geenisTRUE:bs(log10_pikkus, df = 18)17	31.32326	51.75644	0.605	0.545043
geenisTRUE:bs(log10_pikkus, df = 18)18	27.10010	161.88460	0.167	0.867052
eQTLTRUE:bs(log10_pikkus, df = 18)1	0.21907	0.38811	0.564	0.572451
eQTLTRUE:bs(log10_pikkus, df = 18)2	-0.01144	0.25496	-0.045	0.964212
eQTLTRUE:bs(log10_pikkus, df = 18)3	-0.10802	0.27078	-0.399	0.689963
eQTLTRUE:bs(log10_pikkus, df = 18)4	0.54388	0.23052	2.359	0.018308 *
eQTLTRUE:bs(log10_pikkus, df = 18)5	0.31160	0.24415	1.276	0.201865
eQTLTRUE:bs(log10_pikkus, df = 18)6	0.39764	0.23395	1.700	0.089188
eQTLTRUE:bs(log10_pikkus, df = 18)7	0.29355	0.22557	1.301	0.193139
eQTLTRUE:bs(log10_pikkus, df = 18)8	0.07374	0.21896	0.337	0.736296
eQTLTRUE:bs(log10_pikkus, df = 18)9	0.40298	0.22073	1.826	0.067902
eQTLTRUE:bs(log10_pikkus, df = 18)10	0.61567	0.23901	2.576	0.009998 **
eQTLTRUE:bs(log10_pikkus, df = 18)11	0.15382	0.25574	0.601	0.547525
eQTLTRUE:bs(log10_pikkus, df = 18)12	0.88310	0.27964	3.158	0.001589 **
eQTLTRUE:bs(log10_pikkus, df = 18)13	0.30459	0.37232	0.818	0.413303
eQTLTRUE:bs(log10_pikkus, df = 18)14	-0.43080	1.01116	-0.426	0.670072
eQTLTRUE:bs(log10_pikkus, df = 18)15	10.42615	31.14743	0.335	0.737824
eQTLTRUE:bs(log10_pikkus, df = 18)16	9.71449	31.18040	0.312	0.755377
eQTLTRUE:bs(log10_pikkus, df = 18)17	19.38953	32.32072	0.600	0.548566
eQTLTRUE:bs(log10_pikkus, df = 18)18	-50.85559	76.98539	-0.661	0.508877
after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA

after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)12	13.30425	40.23112	0.331	0.740875
after_bp:bs(log10_pikkus, df = 18)13	-20.26319	28.20797	-0.718	0.472542
after_bp:bs(log10_pikkus, df = 18)14	-20.82289	29.11522	-0.715	0.474492
after_bp:bs(log10_pikkus, df = 18)15	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)16	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)17	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)18	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp	35.24613	52.18938	0.675	0.499453
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)1	0.40202	0.56797	0.708	0.479054
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)2	0.40280	0.37406	1.077	0.281543
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)3	0.40742	0.39907	1.021	0.307291
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)4	0.47409	0.33699	1.407	0.159473
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)5	0.21192	0.35722	0.593	0.553003
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)6	0.39041	0.34238	1.140	0.254168
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)7	0.49230	0.33119	1.486	0.137160
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)8	0.38778	0.32199	1.204	0.228463
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)9	0.43773	0.32570	1.344	0.178956
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)10	0.47857	0.35850	1.335	0.181901
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)11	0.88515	0.39048	2.267	0.023402 *
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)12	0.41206	0.44653	0.923	0.356109
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)13	0.59445	0.63102	0.942	0.346169
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)14	3.25617	1.71910	1.894	0.058210 .
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)15	-34.56699	52.18905	-0.662	0.507752
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)16	-33.55339	52.28930	-0.642	0.521076
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)17	-37.82117	55.56132	-0.681	0.496055
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)18	-11.71088	163.35032	-0.072	0.942847
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)12	39.68653	85.24509	0.466	0.641532
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)13	35.46767	47.90368	0.740	0.459060
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)14	38.03784	49.50546	0.768	0.442275
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)15	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)16	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)17	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)18	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)12	-6.17747	44.91694	-0.138	0.890611
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)13	10.46011	30.97576	0.338	0.735599
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)14	11.10461	31.99044	0.347	0.728499
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)15	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)16	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)17	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)18	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA

```

geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)12 5.81846 91.44890 0.064 0.949269
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)13 -36.75872 51.89133 -0.708 0.478710
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)14 -37.61026 53.64154 -0.701 0.483215
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)15 NA NA NA NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)16 NA NA NA NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)17 NA NA NA NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)18 NA NA NA NA
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 385534 on 598793 degrees of freedom
Residual deviance: 346802 on 598702 degrees of freedom
AIC: 346986

Number of Fisher Scoring iterations: 7

```

2 Lisa 2: Mudel insertsioonidele

```

Call:
glm(formula = haruldane ~ (geenis * eQTL * after_bp) * bs(log10_pikkus,
  df = 18) + ME_N, family = binomial, data = ins_v2)

Coefficients: (60 not defined because of singularities)

(Intercept) Estimate Std. Error z value Pr(>|z|)
geenisTRUE 1.782e+00 5.742e-02 31.034 < 2e-16 ***
eQTLTRUE    1.501e-01 9.011e-02 1.666 0.095770 .
after_bp    1.234e-01 1.205e-01 1.024 0.305763
bs(log10_pikkus, df = 18)1 -1.415e+02 7.093e+01 -1.994 0.046109 *
bs(log10_pikkus, df = 18)2 3.026e-01 1.081e-01 2.800 0.005105 **
bs(log10_pikkus, df = 18)3 -9.671e-02 8.138e-02 -1.188 0.234703
bs(log10_pikkus, df = 18)4 1.788e-01 8.244e-02 2.169 0.030117 *
bs(log10_pikkus, df = 18)5 -5.364e-01 7.403e-02 -7.246 4.29e-13 ***
bs(log10_pikkus, df = 18)6 -1.291e-01 8.220e-02 -1.571 0.116147
bs(log10_pikkus, df = 18)7 -1.407e+00 1.036e-01 -13.586 < 2e-16 ***
bs(log10_pikkus, df = 18)8 3.340e+00 2.164e-01 15.436 < 2e-16 ***
bs(log10_pikkus, df = 18)9 6.380e+00 3.244e-01 19.666 < 2e-16 ***
bs(log10_pikkus, df = 18)10 6.371e+00 2.710e-01 23.510 < 2e-16 ***
bs(log10_pikkus, df = 18)11 6.444e+00 2.736e-01 23.555 < 2e-16 ***
bs(log10_pikkus, df = 18)12 2.053e+00 2.440e-01 8.414 < 2e-16 ***
bs(log10_pikkus, df = 18)13 -1.311e+00 1.366e-01 -9.596 < 2e-16 ***
bs(log10_pikkus, df = 18)14 3.461e-01 1.737e-01 1.992 0.046342 *
bs(log10_pikkus, df = 18)15 1.484e+00 2.508e-01 5.916 3.30e-09 ***
bs(log10_pikkus, df = 18)16 -6.017e-01 5.952e-01 -1.011 0.312035
bs(log10_pikkus, df = 18)17 1.489e+02 7.092e+01 2.099 0.035809 *
bs(log10_pikkus, df = 18)18 1.231e+02 7.129e+01 1.726 0.084296 .
ME_N        2.312e+02 8.234e+01 2.807 0.004993 **
geenisTRUE:eQTLTRUE -2.402e-01 1.688e-01 -1.423 0.154719
geenisTRUE:after_bp -7.754e+01 1.182e+02 -0.656 0.511817
eQTLTRUE:after_bp 2.153e+02 7.956e+01 2.706 0.006805 **
geenisTRUE:bs(log10_pikkus, df = 18)1 -2.408e-01 1.679e-01 -1.434 0.151522
geenisTRUE:bs(log10_pikkus, df = 18)2 8.888e-02 1.258e-01 0.706 0.479989
geenisTRUE:bs(log10_pikkus, df = 18)3 -2.017e-01 1.278e-01 -1.579 0.114403
geenisTRUE:bs(log10_pikkus, df = 18)4 -1.214e-02 1.152e-01 -0.105 0.916077
geenisTRUE:bs(log10_pikkus, df = 18)5 -9.379e-02 1.262e-01 -0.743 0.457350
geenisTRUE:bs(log10_pikkus, df = 18)6 -1.602e-01 1.485e-01 -1.078 0.280899
geenisTRUE:bs(log10_pikkus, df = 18)7 3.131e-02 1.700e-01 0.184 0.853832
geenisTRUE:bs(log10_pikkus, df = 18)8 -3.035e-01 1.452e-01 -2.090 0.036615 *
geenisTRUE:bs(log10_pikkus, df = 18)9 -4.400e-02 1.020e-01 -0.431 0.666292
geenisTRUE:bs(log10_pikkus, df = 18)10 9.788e-02 1.035e-01 0.946 0.344112
geenisTRUE:bs(log10_pikkus, df = 18)11 -7.944e-02 1.104e-01 -0.720 0.471648
geenisTRUE:bs(log10_pikkus, df = 18)12 -1.483e-01 1.686e-01 -0.879 0.379262
geenisTRUE:bs(log10_pikkus, df = 18)13 6.804e-01 2.474e-01 2.750 0.005960 **
geenisTRUE:bs(log10_pikkus, df = 18)14 -3.801e-01 3.889e-01 -0.977 0.328385
geenisTRUE:bs(log10_pikkus, df = 18)15 -8.247e-01 9.265e-01 -0.890 0.373375
geenisTRUE:bs(log10_pikkus, df = 18)16 7.244e+01 1.182e+02 0.613 0.539834
geenisTRUE:bs(log10_pikkus, df = 18)17 1.078e+02 1.199e+02 0.899 0.368510
geenisTRUE:bs(log10_pikkus, df = 18)18 -7.467e+01 1.584e+02 -0.471 0.637297

```

eQTLTRUE:bs(log10_pikkus, df = 18)1	5.708e-02	2.284e-01	0.250	0.802667
eQTLTRUE:bs(log10_pikkus, df = 18)2	1.200e-01	1.719e-01	0.698	0.485322
eQTLTRUE:bs(log10_pikkus, df = 18)3	1.916e-01	1.731e-01	1.107	0.268168
eQTLTRUE:bs(log10_pikkus, df = 18)4	6.535e-02	1.540e-01	0.424	0.671274
eQTLTRUE:bs(log10_pikkus, df = 18)5	1.346e-01	1.695e-01	0.794	0.427007
eQTLTRUE:bs(log10_pikkus, df = 18)6	4.961e-01	1.963e-01	2.526	0.011521 *
eQTLTRUE:bs(log10_pikkus, df = 18)7	-4.132e-02	2.215e-01	-0.187	0.851998
eQTLTRUE:bs(log10_pikkus, df = 18)8	8.465e-02	1.937e-01	0.437	0.662021
eQTLTRUE:bs(log10_pikkus, df = 18)9	-5.771e-02	1.356e-01	-0.426	0.670314
eQTLTRUE:bs(log10_pikkus, df = 18)10	-5.041e-02	1.361e-01	-0.370	0.711067
eQTLTRUE:bs(log10_pikkus, df = 18)11	3.456e-01	1.421e-01	2.432	0.015009 *
eQTLTRUE:bs(log10_pikkus, df = 18)12	6.845e-01	1.914e-01	3.576	0.000349 ***
eQTLTRUE:bs(log10_pikkus, df = 18)13	4.276e-01	2.420e-01	1.767	0.077279 .
eQTLTRUE:bs(log10_pikkus, df = 18)14	-1.060e+00	3.159e-01	-3.356	0.000790 ***
eQTLTRUE:bs(log10_pikkus, df = 18)15	1.958e+00	7.070e-01	2.769	0.005622 **
eQTLTRUE:bs(log10_pikkus, df = 18)16	-2.200e+02	7.955e+01	-2.766	0.005677 **
eQTLTRUE:bs(log10_pikkus, df = 18)17	-1.972e+02	7.990e+01	-2.468	0.013571 *
eQTLTRUE:bs(log10_pikkus, df = 18)18	-2.953e+02	8.999e+01	-3.282	0.001032 **
after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)12	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)13	-1.543e+03	1.643e+02	-9.394	< 2e-16 ***
after_bp:bs(log10_pikkus, df = 18)14	1.484e+02	7.079e+01	2.096	0.036074 *
after_bp:bs(log10_pikkus, df = 18)15	1.400e+02	7.139e+01	1.961	0.049860 *
after_bp:bs(log10_pikkus, df = 18)16	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)17	NA	NA	NA	NA
after_bp:bs(log10_pikkus, df = 18)18	NA	NA	NA	NA
geenisTRUE:eQTLTRUE:after_bp	5.153e+01	1.323e+02	0.389	0.696918
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)1	2.575e-01	3.196e-01	0.806	0.420323
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)2	7.529e-03	2.414e-01	0.031	0.975116
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)3	3.470e-01	2.441e-01	1.421	0.155190
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)4	1.680e-01	2.179e-01	0.771	0.440591
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)5	3.038e-01	2.392e-01	1.270	0.204125
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)6	1.955e-01	2.778e-01	0.704	0.481725
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)7	2.466e-01	3.131e-01	0.787	0.431041
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)8	1.020e-01	2.693e-01	0.379	0.704991
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)9	2.254e-01	1.905e-01	1.183	0.236856
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)10	1.085e-01	1.922e-01	0.564	0.572415
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)11	2.987e-01	2.004e-01	1.490	0.136142
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)12	1.917e-01	2.748e-01	0.698	0.485423
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)13	-3.367e-01	3.580e-01	-0.941	0.346910
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)14	6.846e-01	4.913e-01	1.394	0.163435
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)15	1.363e+00	1.114e+00	1.223	0.221407
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)16	-4.709e+01	1.323e+02	-0.356	0.721850
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)17	-7.851e+01	1.339e+02	-0.587	0.557514
geenisTRUE:eQTLTRUE:bs(log10_pikkus, df = 18)18	9.572e+01	1.692e+02	0.566	0.571665
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)5	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)6	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)7	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)8	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)9	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)10	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)11	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)12	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)13	1.079e+03	4.301e+02	2.509	0.012122 *
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)14	7.587e+01	1.180e+02	0.643	0.520153
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)15	7.941e+01	1.189e+02	0.668	0.504328
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)16	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)17	NA	NA	NA	NA
geenisTRUE:after_bp:bs(log10_pikkus, df = 18)18	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)1	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)2	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)3	NA	NA	NA	NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)4	NA	NA	NA	NA

```

eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)5      NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)6      NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)7      NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)8      NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)9      NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)10     NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)11     NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)12     NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)13     5.343e+02  1.876e+02  2.848  0.004394 **
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)14     -2.201e+02  7.940e+01 -2.772  0.005576 **
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)15     -2.159e+02  8.011e+01 -2.694  0.007050 **
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)16     NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)17     NA      NA      NA      NA
eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)18     NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)1  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)2  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)3  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)4  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)5  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)6  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)7  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)8  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)9  NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)10 NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)11 NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)12 NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)13 -3.917e+02  4.619e+02 -0.848  0.396448
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)14 -5.142e+01  1.320e+02 -0.389  0.696976
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)15 -5.230e+01  1.332e+02 -0.393  0.694526
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)16 NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)17 NA      NA      NA      NA
geenisTRUE:eQTLTRUE:after_bp:bs(log10_pikkus, df = 18)18 NA      NA      NA      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370317 on 529865 degrees of freedom
Residual deviance: 339486 on 529773 degrees of freedom
AIC: 339672

Number of Fisher Scoring iterations: 7

```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kaur Tenson,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Inimese genoomis esinevate struktuursete variantide sageduse modelleerimine”, mille juhendajad on Maarja Jõeloo ja 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 digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, 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.

Kaur Tenson

12.10.2026