TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Pleiotroopia roll geeniekspressiooni regulatsioonis

Bakalaureusetöö (9 EAP)

Pilleriin Jukk

Juhendaja: PhD Kaur Alasoo

Tartu 2021

Pleiotroopia roll geeniekspressiooni regulatsioonis

Lühikokkuvõte:

Geeniekspressioon on protsess, mille käigus avaldatakse geenides sisalduv pärilik materjal RNA või valguna. Uurides geeniekspressiooni on võimalik tuvastada haiguse põhjustajaid ja välja töötada neile sobilikke ravimeetodeid. Küll aga ei ole geeniekspressiooni regulatsioon spetsiifiline. See tähendab, et leidub pleiotroopsust ehk üks geneetiline variant mõjutab mitut tunnust korraga. Lisaks, kuna läheduses asuvad geneetilised variandid on omavahel korreleeritud, siis saadakse selliste uuringute tulemuseks mitmeid tunnusega tugevalt seotud variante, millest on aga keeruline järeldusi teha. Töö eesmärk oli välja selgitada, kuidas mõjutab horisontaalne pleiotroopsus geeniekspressiooni regulatsiooni. Selleks analüüsiti geeniekspressiooni ja transkriptsiooni täppiskaardistamise tulemusi (Kerimov *et al.*, 2020) ning võrreldi, kas mõne protsessi puhul olid variandid seotud vähemate geenidega kui geeniekspressiooni puhul. Leiti, et kõigi nelja uuritud protsessi – transkriptide valiku, RNA splaissimise, promootori ja terminaatori kasutuse – uurimistulemused on oluliselt spetsiifilisemad kui geeniekspressiooni omad. Uurimistulemustest saab järeldada, et põhjuslike geenide väljaselgitamiseks tuleks esmalt hakata uurima transkriptide valikut, RNA splaissimist, promootori ja terminaatori kasutust mõjutavaid variante, sest nende protsesside puhul leidub vähem horisontaalset pleiotroopiat ja seega on nende uurimistulemused täpsemad. Nende protsesside puhul on seetõttu ka lihtsam välja selgitada, läbi millise geeni põhjuslik variant haigust mõjutab.

Võtmesõnad:

eQTL kataloog, pleiotroopia, geeniekspressioon, transkriptsioon, RNA splaissimine

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

The Role of Pleiotropy in the Regulation of Gene Expression

Abstract:

Gene expression is a process by which genetic information is used to produce RNA or protein. By studying gene expression, it is possible to identify the causes and develop suitable treatment methods for diseases. However, the regulation of gene expression is not specific. It means that there is pleiotropy, i.e. one genetic variant affects several features at once. In addition, because nearby variants are correlated, such studies result in multiple causal variants from which it is difficult to draw conclusions. The goal of this paper was to find out how horizontal pleiotropy affects the regulation of gene expression. For this, gene expression and transcription fine-mapping results (Kerimov *et al.*, 2020) were analyzed. It was compared whether in some processes variants were associated with fewer genes than in gene expression. It was found that the research results of all four processes – transcription selection, RNA splicing, promoter and terminator usage – are significantly more specific than those in gene expression. The results suggest that variants affecting transcription selection, RNA splicing, promoter and terminator usage should first be investigated to identify causal genes as these four processes have less horizontal pleiotropy and are therefore more accurate. These processes also make it easier to determine through which gene the causal variant influences the disease.

Keywords:

eQTL catalogue, pleiotropy, gene expression, transcription, RNA splicing

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

1. Sissejuhatus

Kõik organismid sisaldavad endas geneetilist informatsiooni, mis määrab nende ülesehituse ja talitluse (Cooper, 2000). Seda informatsiooni talletatakse rakkudes DNA sees, mis koosneb desoksüribonukleotiididest (Clark, 2009). Geneetiline informatsioon koosmõjus keskkonnateguritega mõjutab pea kõiki haigusi, määrates organismi vastuvõtlikkust või resistentsust haigustele (Thomson & Esposito, 1999). Komplekssete haiguste, nagu näiteks astma ja Alzheimeri tõve puhul, põhjustab haigust ühe geneetilise faktori asemel mitmete geneetiliste ja keskkonnategurite koosmõju (Mitchell 2012; Craig, 2008), mistõttu on keeruline haigustekitajaid välja selgitada. Geneetiliste informatsiooni ja keskkonnategurite koosmõjude põhjalik uurimine võib aidata paremini haiguse põhjuseid ja sobilikke ravimeetodeid välja selgitada (Craig, 2008).

Genoomiülesed assotsiatsiooniuuringud (lüh. GWAS, ingl. *genome-wide association studies*) on leidnud suurel hulgal levinud haigustega seotud geneetilisi variante (The Wellcome Trust Case Control Consortium, 2007). See on uuringutüüp, kus proovitakse välja selgitada, kas mõni geneetiline variant on seotud mõne kindla tunnusega (Hindorff *et al.,* 2009), näiteks mõne haigusega. Geneetilisteks variantideks nimetatakse ühe liigi DNA erinevuste piirkondi (Rahim *et al.*, 2008). Samas on aga kogu genoomi assotsiatsiooniuuringute tõlgendamine ja neist järelduste tegemine keeruline, kuna ei ole täiesti selge, milliseid geene variandid mõjutavad (Cano-Gamez & Trynka, 2020).

Üks võimalus on uurida geeniekspressiooni kvantitatiivsete tunnuste lookusi (lüh. eQTL, ingl. *expression quantitative trait loci*) ehk lühikesi DNA järjestuse piirkondi, millel on geeniekspressiooni uurimisel ja mõistmisel oluline roll (Pickrell *et al.*, 2010; Li *et al.*, 2016). Kvantitatiivsete tunnuste lookusi (QTLe) iseloomustab lühike DNA järjestus, mille ühe või paarinukleotiidilised muutused põhjustavad nähtavaid muutusi isendi fenotüübis (Nica & Dermitzakis, 2013). Uurides eQTLe saab välja selgitada, kas haigust põhjustavad variandid on seotud mõne kindla geeni ekspressiooniga mõnes kindlas rakutüübis. Analüüsides haigust põhjustavate variantide seost geeniekspressiooniga saab aga tulemuseks mitu võimalikku põhjuslikku geeni, sest geeniekspressiooni regulatsioon pole liiga spetsiifiline ning võib esineda pleiotroopiat. Pleiotroopia on nähtus, kus üks geneetiline variant mõjutab samaaegselt mitme tunnuse kujunemist. Horisontaalne pleiotroopia tähendab seda, et üks variant mõjutab paralleelselt mitut geeni või tunnust korraga.

Teine võimalus on vaadata, kas haigustega seotud geneetilised variandid mõjutavad mõnd konkreetsemat geenieksрessiooni etappi. Nendeks on DNA põhjal RNA sünteesimine ehk transkriptsioon, sellest valku mitte kodeerivate lõikude (intronite) väljalõikamine ehk RNA splaissimine (ingl. *RNA splicing*) ja omakorda selle põhjal valkude sünteesimine ehk translatsioon. Lisaks, kuna transkripti sünteesimisel pole üht kindlat algus- ja lõppkohta, siis saab täpsemalt uurida ka transkripti algus- ja lõppkoha valikuid (Pal *et al.*, 2011). Kuna geneetilised variandid ja mutatsioonid mõjutavad transkripti töötlemist (Park *et al.*, 2018), siis võiks selle uurimisel leitud geenid olla täpsemalt seotud haiguse tekke põhjustamisega ning võimaldada konkreetsemaid järeldusi teha.

Töö eesmärk on välja selgitada, kuidas mõjub horisontaalne pleiotroopsus geenieksрessiooni regulatsioonile. Pleiotroopiat hinnates saab välja selgitada, kas transkripti töötlemisega seotud geneetiliste variantide uurimistulemused on spetsiifilisemad võrreldes geeni-ekspressiooniga seotud variantidega. See tähendab, kas ühe protsessi uurimisel saadud variandid mõjutavad väiksema tõenäosusega paralleelselt mitut geeni kui teise protsessiga. Mida spetsiifilisem on tulemus, seda tõenäolisemalt on leitud geenid ka tõenäoliselt haigust põhjustavad. Saades teada, millised protsessid annavad spetsiifilisemaid tulemusi, saab teha järeldusi, milliseid protsesse tuleks järgnevalt kasutada haigust põhjustavate geenide väljaselgitamiseks.

Andmeanalüüsiks kasutatakse eQTL Catalogue andmebaasis (Kerimov *et al.*, 2020) olevaid transkriptsiooni ja geenieksрessiooniga seotud variantide andmeid. Andmeanalüüs teostati kasutades Pythonit[1], selle mooduleid pandas[2], PyRanges (Stovner & Sætrom, 2020) ja python-igraph[3]. Erinevate geenieksрessiooni protsesside spetsiifilisuse hindamiseks leiti Fisheri täpse testiga (Fisher, 1941), kas võrdluses geenieksрessiooniga on teiste protsesside vahel statistiliselt oluline seos olemas. Töö käigus kirjutatud kood ja osa algandmetest on saadaval GitHubi koodivaramus pilleriinjukk/splaissimine-vs-ge (Jukk, 2021).

Bakalaureusetöö koosneb kahest osast. Esmalt antakse taustaülevaade geenieksрessioonist, transkriptsioonist ja omakorda selle alamosadest. Lisaks antakse ülevaade põhjuslike variantide leidmisest, Fisheri täpsest testist ning seni tehtud uurimustest. Teises osas kirjeldatakse kasutatud andmeid, antakse ülevaade teostatud andmetöötlusest, kirjeldatakse tulemusi ning viimaks pakutakse välja uuringusuundi tulevikuks.

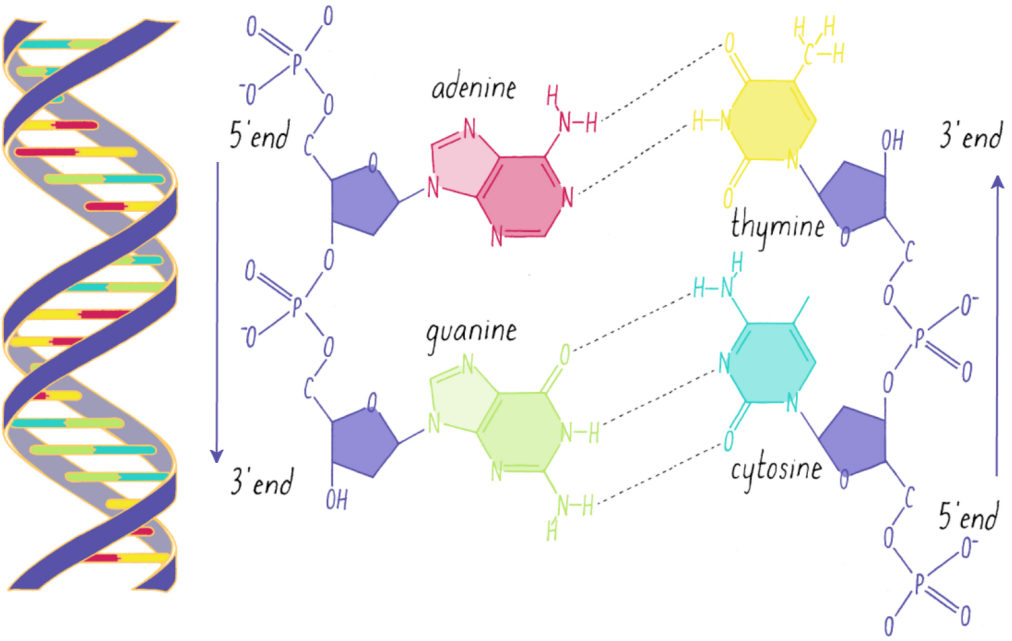
[1] Python versioon 3.8.3
[2] https://pandas.pydata.org/
[3] https://igraph.org/python/

2. Kirjanduse ülevaade

Selles peatükis antakse taustaülevaade geenieksprессioonist, transkriptsioonist ja RNA splaissimisest. Seejärel kirjeldatakse, mis on pleiotroopia ning kuidas on võimalik põhjuslikke variante leida. Täpsemalt kirjeldatakse GWASi ja täppiskaardistamist. Lisaks kirjeldatakse Fisheri täpset testi ja viimases peatükis antakse ülevaade ka seni tehtud uurimustest horisontaalse pleiotroopia kohta.

2.1 Geenieksprессioon

DNA ehk desoksüribonukleiinhape on molekul, mis koosneb kahest pikast komplementaarsest ahelast ning sisaldab nelja põhilist lämmastikalust adeniin (lüh. A, ingl. *adenine*), guaniin (lüh. G, ingl. *guanine*), tümiin (lüh. T, ingl. *thymine*) ja tsütosiin (lüh C, ingl. *cytosine*). Komplementaarsed paarid moodustatakse alati A-T ja G-C vahel. RNA ehk ribonukleiinhape on üheahelaline molekul, kusjuures tümiini asemel on uratsiil (lüh. U, ingl. *uracil*). RNA peamine ülesanne on DNAs paikneva päriliku informatsiooni realiseerimine (Clark, 2009). Nii DNA kui ka RNA molekuli puhul on nukleotiidide ahela kaks otsa erinevad, üht otsa nimetatakse 5’ ja teist otsa 3’. Joonisel 1 on DNA molekul.



Joonis 1. DNA molekuli struktuur koos ahelate suundadega (Vale, 2019, täiendatud).

Geenieksprессioon ehk geeni avaldumine on see, mis muudab organismi genotüübi fenotüübiks (Hill, Vande & Wittkopp, 2020). See on protsess, mille käigus geenides sisalduv pärilik materjal avaldatakse RNA või valguna ning selle kaudu mõjutatakse organismi

ülesehitust ja toimimist (Clark, 2009). Geenieкspressioonil on kolm peamist etappi: DNA põhjal RNA sünteesimine ehk transkriptsioon, sellest intronite väljalõikamine ehk RNA splaissimine ja mRNA põhjal valguahela sünteesimine ehk translatsioon.

Geenieкspressiooni regulatsioon on protsess, mille abil saab kontrollida, milliseid geene ekspresseeritakse, näiteks millistest geenidest toodetakse valke ning mis kogustes seda tehakse (Cooper, 2000). Geenieкspressiooni regulatsiooni uurides on üks võimalus vaadata eQTLe (Pickrell *et al.*, 2010; Li *et al.*, 2016), mille ühe- või paarinukleotiidilised muutused põhjustavad nähtavaid muutusi isendi fenotüübis (Nica & Dermitzakis, 2013). See võimaldaks välja selgitada, kas haigust põhjustavad variandid on seotud mõne kindla geeni ekspressiooniga mõnes kindlas rakutüübis. Selleks uuritakse geneetilisi variante ja nende geeni ekspressiooni tasemeid, mida mõõdetakse tavaliselt kümnetel või sadadel inimestel (Pagán, Holmes & Simon-Loriere, 2012) ehk uuritakse, kui palju toodetakse erinevatel indiviididel geenidest valke.

Lisaks ei piisa sellest, kui uurida geeni ekspressiooni vaid ühes rakutüübis. Näiteks geen CCL16 ei avaldu igas rakutüübis, vaid avaldub peamiselt maksas (Viñuela et al., 2021). Analüüsides geenieкspressiooni ainult ühes rakutüübis, näiteks veres, võivad koespetsiifilised dünaamikad jääda tähelepanuta. Seega tuleb vaid ühe rakutüübi asemel uurida geeni ekspressiooni erinevates rakutüüpides korraga (Viñuela et al., 2021).

2.2 Transkriptsioon

Informatsiooni ülekandmist DNAst RNAks nimetatakse transkriptsiooniks. Transkriptsiooni käigus sünteesitakse DNA 5’ suunast 3’ otsa liikudes sellele vastav RNA molekul ehk transkript (Clark, 2009). Transkriptsiooni käigus saadakse erinevaid RNA molekule, millest mRNA põhjal toodetakse valke (Newman, 1998).

Erinevad geneetilised variandid ja rakus valitsevad tingimused võivad mõjutada seda, kuidas geeni sünteesitakse transkriptiks, näiteks pannakse mõnda geeni liiga palju või liiga vähe transkribeerima (Richards *et al.*, 2017). Sel juhul toodetakse mõnda valku rohkem või vähem. Analüüsides vaid geeni ekspressiooniga seotud geneetilisi variante võivad transkripti töötlemisega seotud geneetilised variandid ja nendevaheline dünaamika märkamata jääda (Yi *et al.*, 2018). Näiteks ei pruugi alternatiivsest splaissimisest (vt ptk 2.2.2) põhjustatud muudatused saada kaardistatud.

2.2.1 Transkriptsiooni algus- ja lõppkoht

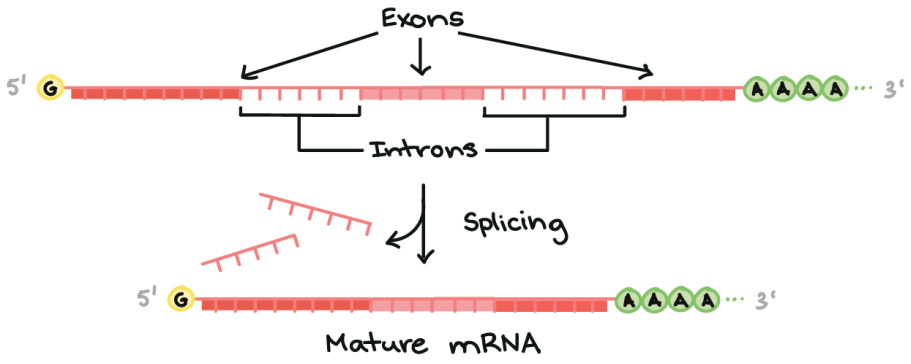
Geneetilised variandid võivad mõjutada ka transkriptsiooni algus- (lüh. TSS, ingl. *transcription start site*) või lõppkoha (lüh. TTS, ingl. *transcription termination site*) valikut. Alustades või lõpetades transkriptsiooni erinevatest kohtadest saab luua erinevaid transkripte, millel võib olla erinev bioloogiline mõju (Pal *et al.*, 2011).

Promootoriks nimetatakse sellist nukleotiidijärjestust DNAs, kuhu seonduvad transkriptsioonifaktorid ja RNA polümeraas ning mis määrab transkriptsiooni alguskoha (Clark, 2009). Paljudel geenidel on rohkem kui üks TSS ja seega ka rohkem kui üks promootor, mis võimaldab luua erinevaid transkripte (Sandelin *et al.*, 2007).

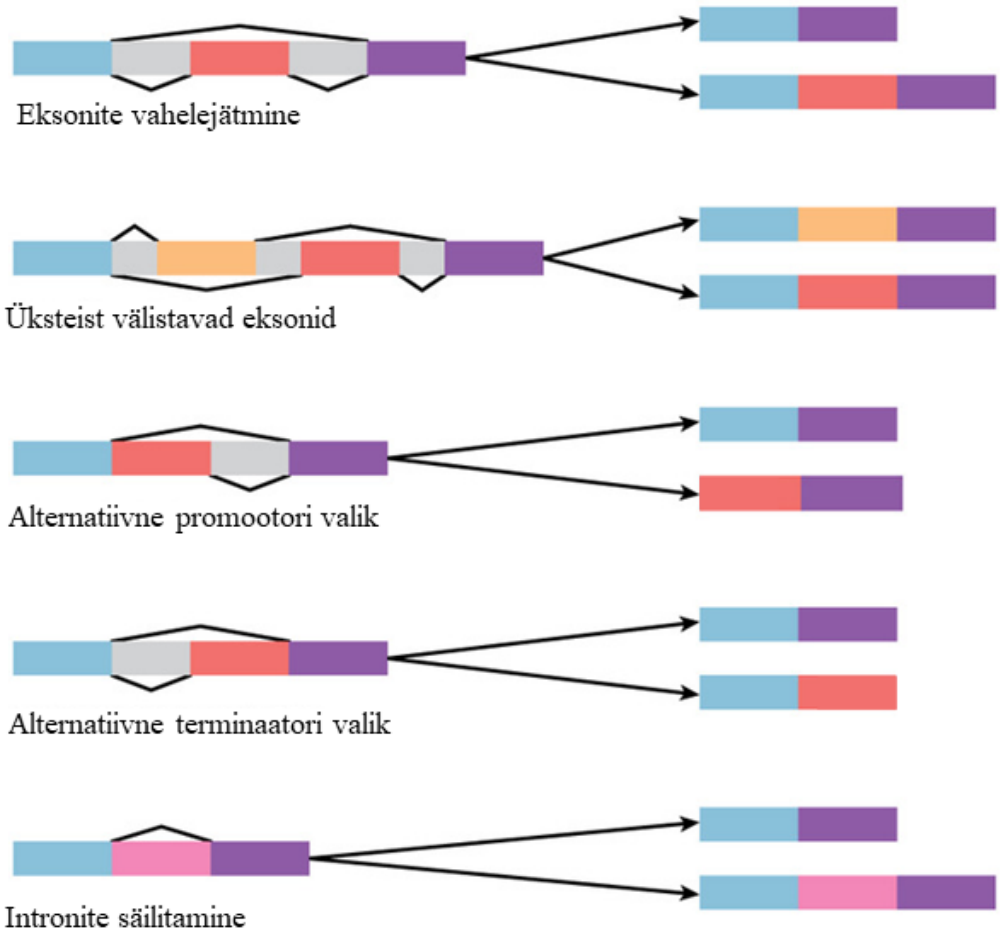
Loodavat transkripti võib mõjutada ka transkriptsiooni lõpp ehk terminaator (Porrua & Libri, 2015). See on DNA järjestus, pärast mida lõpetatakse transkriptsioon. Analoogselt promootoritele on paljudel geenidel mitu võimalikku TTSi, tänu millele on võimalik luua erinevaid transkripte (Lodish *et al.*, 2000). Lisaks võib sellel olla ka oluline mõju loodava transkripti stabiilsusele (Porrua & Libri, 2015).

2.2.2 RNA splaissimine

RNA splaissimine (ka splaising, ingl. *RNA splicing)* ehk mRNA töötlemine on päristuumsetes ehk eukarüootsetes, sealhulgas inimeste, organismides üks rakutuuma protsessidest. RNAd splaissides lõigatakse RNA molekulist välja intronjärjestused ning allesjäänud eksonite osad ühendatakse mRNAks (joonis 2), mida translatsioonil kasutatakse proteiini sünteesimiseks (Newman, 1998). Intron on seega mittekodeeriv ja ekson kodeeriv piirkond, mistõttu luuakse just eksonite kombinatsioonidest valke. RNAd splaissitakse splaissosoomis (ingl. *spliceosome)* (Keren, Lev-Maor & Ast, 2010), mis on RNA ja valkude kompleks, kus eel-mRNAst (ingl. *pre-mRNA*) lõigatakse intronid välja (Clark, 2009).



Joonis 2. RNA splaissimine (Eukaryotic pre-mRNA processing).



Joonis 3. Alternatiivne splaissimine (OpenStax College, täiendatud). Eksonid on tähistatud värviliselt ning intronid halli ja roosa värviga.

Ühest geenist saab erinevat moodi eksoneid kokku liites luua RNAst mitmeid erinevate funktsioonidega valke (joonis 3) (Modrek & Lee, 2002). Seda nimetatakse alternatiivseks splaissimiseks ja see võimaldab mõjutada raku arengut ning rakutüübile vastavate spetsiifiliste protsesside toimumist (Pagani & Baralle, 2004; Yang *et al.*, 2016). Alternatiivne splaissimine võimaldab luua vähestest geenidest mitmeid erinevaid spetsialiseeritud valke (Yang *et al.*, 2016), näiteks inimeste ligikaudu 20 000 geenist on alternatiivse splaissimisega võimalik saada 100 000 erinevat valku (Pan *et al,*. 2008).

Splaissimist mõjutavad ka geneetilised variandid ja haigustega seotud mutatsioonid (Park *et al.*, 2018). Need mutatsioonid võivad segada mõne eksoni kombinatsiooni loomist, mille tulemuseks on ebanormaalsed mRNAd ja valgud (Pagani & Baralle, 2004), mis omakorda võivad muuta raku talitlust.

Splaissimist mõjutavaid faktoreid saab jaotada põhi- (ingl. *basal factors*) ja regulatoorseteks (ingl. *regulatory factors*) splaissimisfaktoriteks, kusjuures põhitegurid eelkõige katalüüsivad

splaissimist ning regulatoorsed tegurid soodustavad või pärsivad seda (Dvinge *et al.*, 2019). Põhitegurid kuuluvad splaissosoomi komponentidesse, aga regulatoorsed tegurid ei pruugi (Dvinge *et al.*, 2019). Selle asemel seovad regulatoorsed tegurid tavaliselt võimendajate (ka tugevdajate, ingl. *enhancer*) või vaigistajatena (ingl. *silencer*) eel-mRNA külge, et võimendada või summutada splaissimist (Fu & Ares, 2014).

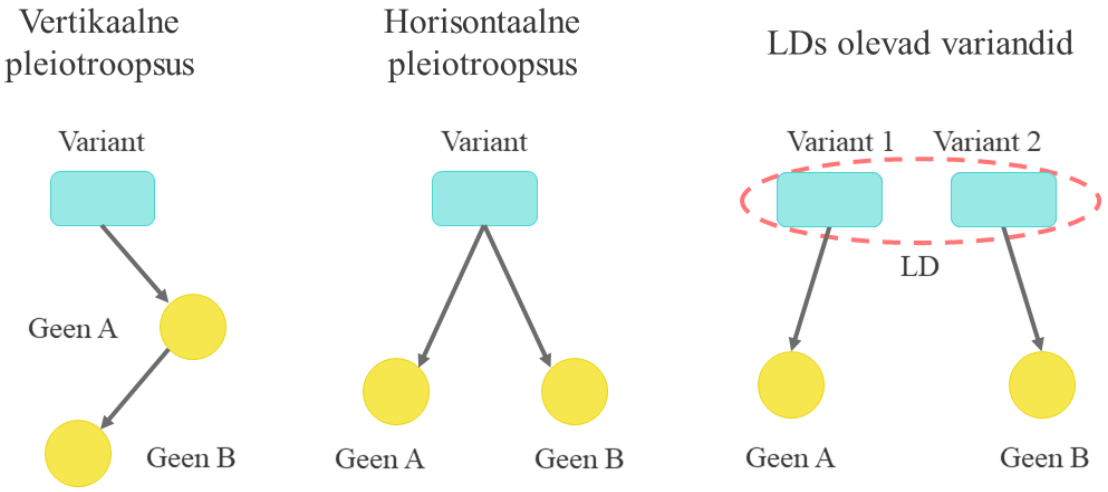
Vääralt reguleeritud geenieksрessioon või transkriptsioonifaktoreid mõjutavad mutatsioonid võivad tavalist splaissimist muuta hoopis haigust põhjustavaks protsessiks (Fu & Ares, 2014). Paljud splaissimist mõjutavad geneetilised variandid avaldavad mõju vaid RNA splaissimisele ja sellele järgnevatele protsessidele, näiteks mõjutavad splaissosoomi või muude splaissimist mõjutavate faktorite seondumist konkreetsest geenist toodetud mRNA molekulile (Alasoo *et al.*, 2019; Garrido-Martín *et al.*, 2021; Li *et al.*, 2016). Geenieksрessiooni mõjutavad geneetilised variandid mõjutavad aga suuresti transkriptsioonifaktorite seondumist DNAle (Laurila & Lähdesmäki, 2009; Li *et al.,* 2016). Seetõttu võivad transkriptsioonifaktorid mõjutada mitut samal DNA molekulil lähestikku asuvat geeni (Spitz & Furlong, 2012). Seega on olemas bioloogiline eeldus, et splaissimise puhul võiks võrreldes geenieksрessiooniga olla väiksem tõenäosus, et sama variant mõjutab mõnda muud geeni.

2.3 Pleiotroopia

Pleiotroopiaks nimetatakse olukorda, kus ühel geneetilisel variandil on mõju mitmele tunnusele või geenile (Lobo, 2008). Näiteks paljud geenid, mis reguleerivad organismi arengut, omavad rolli mitmetes kudedes ja erinevatel arenguetappidel (Paaby & Rockman, 2013). Pleiotroopsed lookused on seotud ka komplekssete tunnustega, nagu kehamassiindeks, pikkus ja skisofreenia (Jordan, Verbanck & Do, 2019). Lookuseks nimetatakse kromosoomi piirkonda, kus paikneb mingi geen või lihtsalt mingi haiguse või tunnusega seotud variante (Clark, 2009).

Vastavalt pleiotroopse geneetilise variandi mõjule saab pleiotroopiat jagada kaheks (joonis 4) (Paaby & Rockman, 2013). Vertikaalseks pleiotroopiaks nimetatakse seda, kui geneetiline variant on mingi geeni põhjuslik variant, mis hiljem omakorda mõjutab mingit muud geeni (Jordan *et al.*, 2019). Horisontaalseks pleiotroopiaks nimetatakse seda, kui geneetiline variant mõjutab mitut geeni paralleelselt (Jordan *et al.*, 2019). Samas aga tuleb eristada neid variante, mille puhul esineb aheldustasakaalutust (lüh. LD, ingl. *linkage disequilibrium*) (Spain & Barrett, 2015). LD tähistab lähestikku paiknevate erinevate geneetiliste variantide

mittejuhuslikku seost populatsioonis (Slatkin, 2008). Lähestikku paiknevate variantide puhul võib uuringutest tulla välja nagu oleks üks variant mingi tunnuse või geeni põhjuslik variant, aga tegelikkuses on see variant põhjusliku variandiga LDs ja ei ole vaadeldava tunnuse või geeni põhjustaja.



Joonis 4. Erinevate pleiotroopia tüüpide skeem koos LDs olevate variantidega.

Näiteks on leitud, et kromosoomi 1p13 lookus, geen SORT1, on inimestel tugevalt seotud madala tihedusega lipoproteiini kolesterooliga (LDL-C) (Musunuru, 2010). Samas aga on teada, et sama variant, mis mõjutab geeni SORT1, mõjutab ka geeni PSRC1 ekspressiooni (Musunuru, 2010). Järelikult on selle variandi puhul tegemist nii horisontaalse kui ka vertikaalse pleiotroopiaga. Horisontaalne pleiotroopia esineb, sest üks variant on tõenäoliselt kahe erineva geeni SORT1 ja PSRC1 põhjuslik variant. Lisaks esineb vertikaalset pleiotroopiat, kuna tõenäoliselt mõjutab vaadeldav variant geeni SORT1, mis omakorda on seotud LDL kolesterooliga.

2.4 Põhjuslike variantide leidmine

Haigusi põhjustavate geneetiliste variantide väljaselgitamine võimaldab kindlaks määrata haiguse iseloomu ning sobilikke ravimeetodeid (Craig, 2008). Järgnevas peatükis on kirjeldatud üht meetodit, kuidas seda on võimalik teostada.

2.4.1 Genoomiülesed assotsiatsiooniuuringud

Genoomiüleste assotsiatsiooniuuringute (lüh. GWAS) käigus uuritakse paljudes isendites üle kogu genoomi seost kindla tunnuse ja geneetiliste variantide vahel, kusjuures sellega on suudetud kaardistada tuhandeid inimeste haigustega seotud lookuseid (The Wellcome Trust

Case Control Consortium, 2007). GWAS on uuringutüüp, kus proovitakse välja selgitada, kas mõni geneetiline variant on seotud mõne kindla tunnusega (Hindorff *et al.,* 2009), näiteks mõne haigusega. Selleks uuritakse tavaliselt üksiku nukleotiidi polümorfisme (lüh. SNP, ingl. *single-nucleotide polymorphism*), mis on DNA järjestuse variatsioonid, mis tekivad ühe nukleotiidi muutumisel (Clark, 2009).

GWAS uuringute käigus põhjuslike variantide kindlaks tegemist takistab aheldustasakaalutus (lüh. LD) (Spain & Barrett, 2015). GWAS käigus leitud seotud variandid on tihti LDs põhjusliku variandiga, mitte ei oma ise bioloogilist funktsiooni (Spain & Barrett, 2015). Seetõttu peab põhjuslike variantide väljaselgitamiseks tegema järeluuringuid, kuid sadade või tuhandete SNPde vaheliste keeruliste LD mustrite tõttu on see keeruline (Hutchinson, Watson & Wallace, 2020).

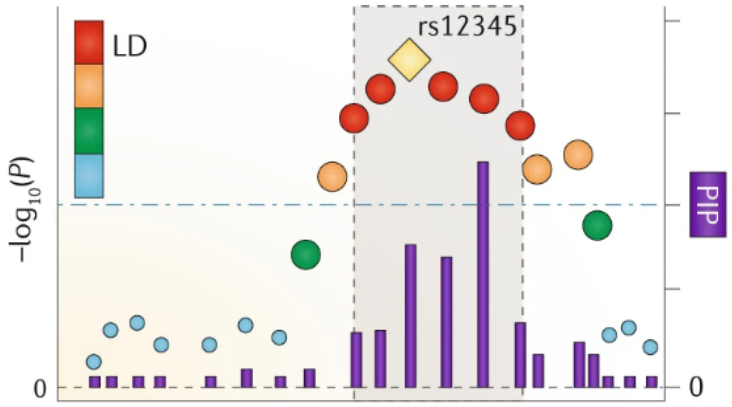
2.4.2 Täppiskaardistamine

Täppiskaardistamise (ingl. *fine-mapping*) uuringute eesmärk on kindlaks teha konkreetsed tunnust põhjustavad variandid, mille käigus analüüsitakse kogu GWAS uuringutega leitud tunnustega seotud piirkondi (Schaid, Chen & Larson, 2018). Seejärel uuritakse iga piirkonna LD struktuuri ja uuritakse kaardistatud geene (Schaid *et al.*, 2018). Kuna lihtsam on korraga kaardistada vaid üht põhjuslikku varianti, siis jaotatakse iga piirkond alampiirkondadeks, kus igal piirkonnal on ligikaudu sõltumatu mõju mingile tunnusele (Schaid *et al.*, 2018).

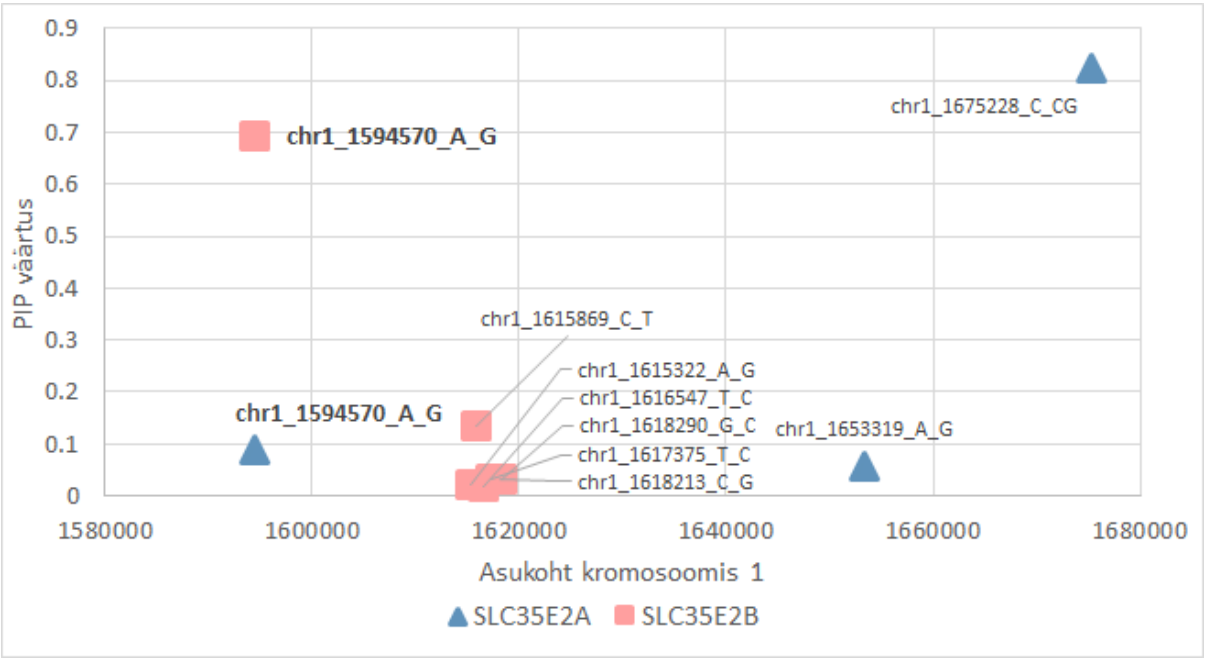
Üks meetod on Bayesi täppiskaardistamine, kus iga tunnuse põhjusliku variandi välja-selgitamiseks kaardistatakse usutavate variantide hulki (lüh. CS, ingl. *credible set*), mis sisaldavad teatud tõenäosusega (näiteks 95% tõenäosusega) endas tunnusega seotud põhjuslikku varianti (Hutchinson *et al.*, 2020).

Joonisel 5 on Bayesi täppiskaardistamise illustratiivne näidis. Väärtus $-\ log10(P)$ tähistab GWAS uuringutega leitud olulisuse tõenäosust (ka p-väärtus, ingl. *p-value*). P-väärtus näitab seda, kui suure tõenäosusega oleks võinud leida sama tugeva või tugevama seose variandi ja tunnuse vahel eeldusel, et seost ei ole (Zhang, 2016). Küll aga võivad põhjusliku variandiga LDs olevate variantide p-väärtused olla ka statistiliselt olulised, olenemata sellest, et tegelikult pole tegemist põhjuslike variantidega (Zhang, 2016). PIP (ingl. *posterior inclusion probability*) tähistab tõenäosust, et vaadeldav variant on põhjuslik variant (Schaid *et al.*, 2018). Igasse CSi kuulub vähemalt nii palju variante, et nende PIP väärtuste summa ületab näiteks 95%. See tähendab, et vähemalt 95% tõenäosusega kuulub põhjuslik variant CSi.

Neist usutavate variantide hulkadest saab laboratoorsete uuringute käigus välja selgitada tegelikud põhjuslikud variandid, mida saab omakorda seostada kindlate geenidega ja sealtkaudu mõista haiguse geneetilist alust, et selgitada välja ravimeetodeid (Hutchinson *et al.*, 2020).



Joonis 5. Bayesi täppiskaardistamine (Schaid *et al.*, 2018). Vasakul on GWAS $-\log_{10}$ p-väärtus iga variandi kohta (kujutatud täppidega) ja paremal PIP väärtus (kujutatud tulp-diagrammina). Rombiga on tähistatud juhtvarianti (statistiliselt olulisima seosega varianti). Teiste variantide puhul on märgitud, kui tugevalt need on juhtvariandiga LDs (värvid skaalal punane kuni sinine). PIP väärtustest moodustatakse usutavate variantide hulki (CSe) kindlaks määratud katvuse tõenäosuse (näiteks 95%) alusel. Hallil taustal punktiirjoonega eraldatult on usutavate variantide hulka valitud variandid.



Joonis 6. Kaks CSi, mis mõlemad sisaldavad varianti chr1_1594570_A_G, kuid ühes on leitud geeni SLC35E2A põhjuslikku varianti ning teises geeni SLC35E2B oma.

Bayesi täppiskaardistamine võib anda tulemuseks sellised CSid, kus sama variant kuulub mitmesse erinevasse CSi (joonis 6). Kui üks variant kuulub mitmesse erinevasse CSi, kus CSides on leitud erinevate geenide põhjuslikke variante, siis on keerulisem sellest variandist järeldusi teha. Oletame, et GWAS käigus on tehtud selgeks, et vaadeldav variant põhjustab haigust. Variant aga kuulub mitmesse CSi ning neis on leitud erinevate geenide põhjuslikke variante. Soovides kindlaks teha, läbi millise geeni põhjustab variant vaadeldavat haigust saadakse seega vastuseks mitu võimalikku geeni. Selline olukord aga pole ideaalne, kuna siis tuleb veel edasi uurida, milline geen tegelikkuses põhjustab vaadeldavat haigust.

2.5 Fisheri täpne test

Fisheri täpse testiga (ingl. *Fisher's exact test*) saab kindlaks määrata, kas uuritavatel protsessidel on statistiliselt oluline seos või mitte (Fisher, 1941). Seda kasutatakse tavaliselt siis, kui võrreldakse kahte gruppi, millel on kaks võimalikku väärtust (tabel 1) (Kim, 2017). Enamjaolt kasutatakse Fisheri täpset testi väikeste valimi suuruste puhul, kuid test toimib ka suurte valimite puhul (Kim, 2017).

Tabel 1. Fisheri täpse testi tegemiseks koostatav tabel.

	Grupp 1	Grupp 2	Kokku
Tunnus 1	a	b	a + b
Tunnus 2	c	d	c + d
Kokku	a + c	b + d	n = a + b + c + d

Tõenäosus, et vaadeldud sündmused toimuvad siis, kui kahe vaadeldava grupi ja kahe tunnuse vahel pole seost, leitakse järgmiselt

$$p = \frac{(a+b)!\,(c+d)! + (a+c)! + (b+d)!}{n!\,a!\,b!\,c!\,d!}.$$

Tõenäosusest tuleb välja, kas vaadeldud sündmuste jaotust on keeruline korrata, kui gruppide vahel pole seost. Kui tõenäosus on väga väike, siis saab väita, et tunnuste ja omaduste vahel on ikkagi seos olemas ja saab nullhüpoteesi kukutada. Nullhüpotees on väide, et tunnuste vahel pole statistilist seost. Seejärel valitakse teatud p-väärtus (tavaliselt 5%). Kui testi tulemus tuleb alla 5%, siis tähendab see seda, et on väga ebatõenäoline, et juhuslikult võiks toimuda sellise jaotusega sündmuseid. Seega saab nullhüpoteesi kummutada ning väita, et tunnuste vahel on statistiliselt oluline seos olemas.

2.6 Varasemad uuringud

Jordan *et al.* (2019) on leidnud, et horisontaalne pleiotroopia on laialt levinud kogu genoomis. Lisaks avastasid nad, et horisontaalset pleiotroopiat leidub rohkem aktiivselt transkribeeritud piirkondades ja aktiivsetes regulatsioonipiirkondades. See viitab sellele, et pleiotroopia võib geenieksprессiooni regulatsiooni suurel hulgal mõjutada. Ka Viñuela *et al.* (2021) on avastanud, et pleiotroopia mõjutab eQTLe, mis takistab põhjuslike geenide kindlaks tegemist. Küll aga pole uuritud täppiskaardistamise uuringutest saadud andmete peal, kas horisontaalse pleiotroopia levimus on transkriptide valiku ja RNA splaissimise puhul väiksem kui geenieksprессiooni puhul. Arvestades, et transkriptsiooni ja splaissimise regulatsioon toimub suures osas RNA tasemel, mitte DNA tasemel, nagu geenieksprессiooni puhul, siis võiks transkriptide valiku ja RNA splaissimise puhul leiduda vähem horisontaalset pleiotroopiat.

3. Geneetiliste variantide leidmine

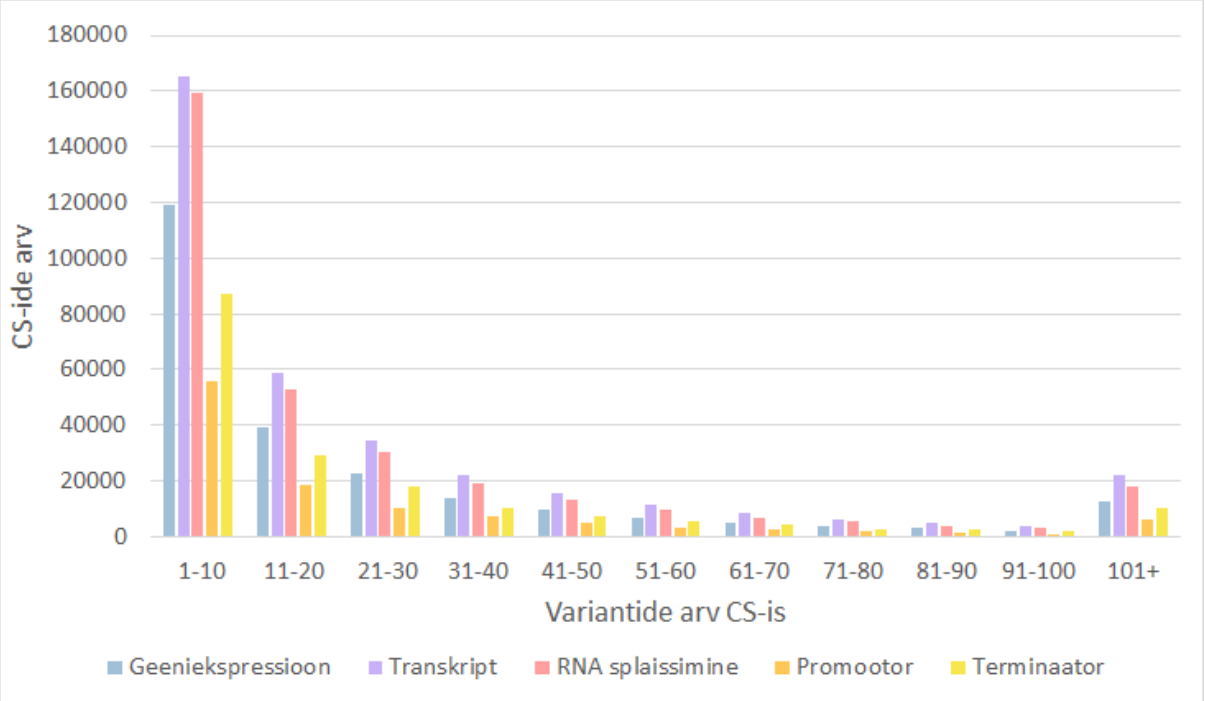
Selles peatükis antakse ülevaade kasutatud andmetest ja andmetöötlusest. Seejärel kirjeldatakse tulemusi. Lisaks pakutakse lõpus välja võimalusi edasisteks uurimusteks.

3.1 Andmed

Andmeanalüüsis kasutatakse eQTL Catalogue andmebaasi (Kerimov *et al.*, 2020) transkriptide valiku, RNA splaissimise, promootori ning terminaatori kasutuse ja geeniekspressiooniga seotud variantide andmeid. Need andmed on kogutud 93st andmestikust üle 16 uuringu. Kokku on geenidoonoreid ligikaudu 4700. Uuritud protsesside algandmete teave on toodud tabelis 2. Uuritud raku- ja koetüüpe oli 66.

Tabel 2. Teave algandmete kohta.

	Geeniekspressioon	Transkript	RNA splaissimine	Promootor	Terminaator
Tabelite pikkus	7 006 139	13 146 973	9 622 475	4 269 109	5 860 610
Geene	21 064	10 857	7555	4572	5253
Variante	1 378 033	1 189 794	827 036	373 464	532 011
CSide arv	237 941	353 168	321 724	112 511	178 518



Joonis 7. Usutavate variantide hulkade suuruste jaotused vastavalt uuritavale protsessile.

Andmebaas keskendub geenikespressiooni QTLidele, kus variandid on seotud lähedalasuvate geenide ekspressioonitasemetega, ja splaissimise QTLidele (sQTL), kus variandid on seotud kindlate splaissimiskohtade, transkriptide, promootorite ja terminaatoritega (Kerimov *et al.*, 2020). Usutavate variantide hulkade suuruste jaotused on näha joonisel 7. Keskmiselt kuulub ühte CSi 33 varianti, kuid igal protsessi puhul leidus ka üle 3000 variandist koosnevaid CSe, suurim oli transkripti valiku uurimistulemustes 3951 variandist koosnev CS.

3.2 Andmeanalüüs

Töö eesmärk on välja selgitada, kuidas mõjub horisontaalne pleiotroopsus geenikespressiooni regulatsioonile. Pleiotroopiat hinnates saab välja selgitada, kas transkripti töötlemisega seotud geneetiliste variantide uurimistulemused on spetsiifilisemad võrreldes geenikespressiooniga seotud variantidega. Mida spetsiifilisem on tulemus, seda tõenäolisemalt on leitud geenid ka tõenäoliselt haigust põhjustavad. Kuna uuriti 66 erineva koe- ja rakutüübi kohta, mitte vaid paari rakutüübi kohta, on suurem tõenäosus, et koespetsiifilised dünaamikad ei jäänud tähelepanuta ning see võimaldab analüüsi tulemusi üldistada.

Andmeanalüüs teostati programmeerimiskeeles Python. Pythoni moodulitest kasutati andmetöötluseks pandast ja spetsiifilisemalt genoomiandmete töötlemiseks PyRanges'it (Stovner & Sætrom, 2020). Lisaks kasutati graafide töötlemiseks paketti python-igraph, mis on võimeline suurte andmehulkadega töötama. Fisheri täpse testi tulemuse leidmiseks kasutati Pythoni moodulit SciPy[4]. Joonised on tehtud Microsoft Exceliga[5].

Andmeid töödeldes käsitletakse geenikespressiooni, RNA splaissimise, promootori kasutuse, terminaatori kasutuse ning transkriptide valiku kohta käivat informatsiooni eraldi. Esmalt loetakse kõik uuringutulemused sisse ning salvestatakse kromosoomipõhiliselt eraldi failidesse. See võimaldab töödelda kogu informatsiooni ühe kromosoomi kohta korraga ning vältida mäluprobleeme.

Seejärel fikseeritakse usutavate variantide hulkade maksimaalne suurus. Väärtuslikum on omandada teadmisi võimalikult väikese CSi kohta (Hutchinson *et al.*, 2020) ehk kuhu kuulub võimalikult vähe variante. Seda seetõttu, et sel juhul tuleb kaaluda väiksemat hulka variante, mille seas on tõenäoliselt põhjuslik variant. Lisaks aitab see vältida põhjuslike variantide ja selle läheduses paiknevate variantide omavahelist LDd. Kui võrrelda kaht suurt CSi

[4] https://www.scipy.org/
[5] Microsoft Office Professional 2016

omavahel, siis on tõenäolisem, et need sisaldavad omavahel sõltumatuid põhjuslikke variante. Järelikult, mida väiksemad on CSid, seda suurem on tõenäosus, et kui kaks CSi on omavahel ülekattes, siis on neil ka sama põhjuslik variant. Seega tehti detailsemat analüüsi andmete peal, kus põhjuslike variantide maksimaalseks suuruseks valiti 50, kuna ühendatud komponentidesse kuuluvate geenide osakaal (vaid üht geeni sisaldavate komponentide arv kogu komponentide arvust) stabiliseerus (vt joonis 10).

Kasutades PyRanges moodulit klasterdatakse andmed asukohapõhiselt (tabel 3), mis annab iga variandi kohta klastri numbri. Samasse klastrisse kuuluvad need variandid, mis paiknevad samas kromosoomis samal asukohal. Selle kaudu saab teada, millistesse usutavate variantide hulkadesse üks variant kuulub.

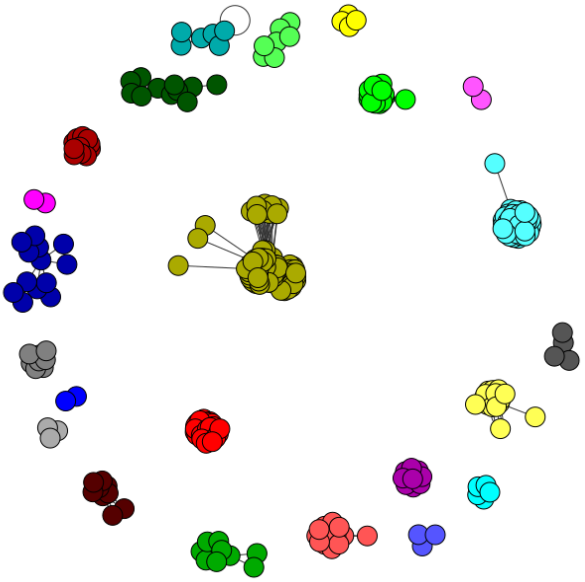
Tabel 3. Klasterdamise tulemus.

geeni_id	variandi_id	chr	asukoht	PIP	cs_id	cs_ suurus	klaster
ENSG00000238009	chr1_108826_G_C	1	108826	0.9043	ENSG00000238 009_L4	3	1
ENSG00000241860	chr1_115746_C_T	1	115746	0.3957	ENSG00000241 860_L1	4	2
ENSG00000238009	chr1_115746_C_T	1	115746	0.459	ENSG00000238 009_L3	4	2
ENSG00000241860	chr1_135203_G_A	1	135203	0.1159	ENSG00000241 860_L1	4	3
ENSG00000238009	chr1_135203_G_A	1	135203	0.136	ENSG00000238 009_L3	4	3

Et välja selgitada, kui palju esineb horisontaalset pleiotroopiat ehk kui tihti mõjutab üks geneetiline variant paralleelselt mitut tunnust korraga, leitakse ühendatud komponente. Ühendatud komponent on selline hulk, kuhu kuuluvad usutavate variantide hulgad, mis sisaldavad vähemalt ühte sama varianti. Kui on kaks CSi, mis mõlemad sisaldavad sama varianti, siis need CSid kuuluvad samasse ühendatud komponenti. Kui ühendatud komponenti kuuluvad CSid, mis on erinevate geenide jaoks leitud, siis on leitud osade variantide puhul, et need on seotud mitme erineva geeniga. Seega saab öelda, et leidub horisontaalset pleiotroopiat.

Selleks, et leida, mitut geeni ühendatud komponendid sisaldavad, koostatakse graaf (joonis 8). See võimaldab jälgida, millised CSid sisaldavad samu variante ja moodustavad seega ühendatud komponente. Loodavas graafis on tippudeks usutavate variantide hulgad, mille

vahele luuakse serv, kui mõlemasse hulka kuulub vähemalt üks jagatav variant. Sellest edasi moodustatakse ühendatud komponente ehk sidusaid alamgraafe, kus ühte ühendatud komponenti kuuluvad sellised tipud, mis on servade kaudu seotud teiste tippudega. Küll ei pea loodavas ühendatud komponendis iga tipu ehk CSi vahel olema serv, vaid iga kahe tipu korral peab leiduma neid tippe ühendav ahel ehk tippude järjend, kus iga kaks järjestikust tippu on ühendatud servadega.



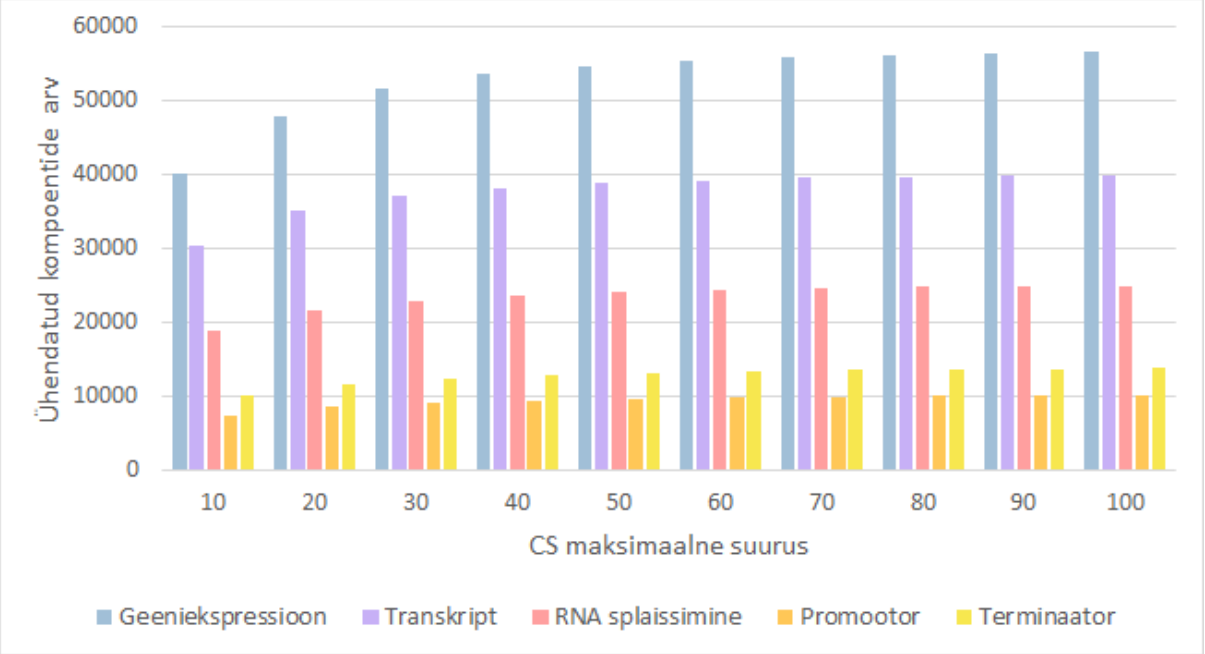
Joonis 8. Näide kromosoomi 1 mõnedest ühendatud komponentidest, kus tipud on CSid ja servad tähistavad sama variandi kuulumist mitmesse CSi.

Seejärel arvutatakse kokku, kui palju erinevaid geene kuulub ühte ühendatud komponenti. Sealt saab teada, kui paljud variandid on seotud mitme geeniga. Tulemustest saab järeldada, kui palju leidub erinevate protsesside puhul pleiotroopiat ehk kui spetsiifilisi tulemusi erinevad põhjuslike variantide uurimismeetodid leiavad.

Statistilise seose kindlaks määramiseks kontrolliti ka Fisheri täpse testiga, kas kahe uuritava protsessi ühendatud komponentide geenide sisalduvused on statistiliselt piisavalt erinevad. Nullhüpoteesiks oli, et ühendatud komponenti kuuluvate geenide arv ei sõltu protsessist ning p-väärtuseks valiti 5%. Kui Fisheri testiga tuleb vastus suurem kui 0,05, siis jääb nullhüpotees kehtima. Kui aga tuleb testi väärtus madalam, kui 0,05, siis saab nullhüpoteesi kummutada ning väita, et on statistiliselt oluline, millist protsessi jälgitakse.

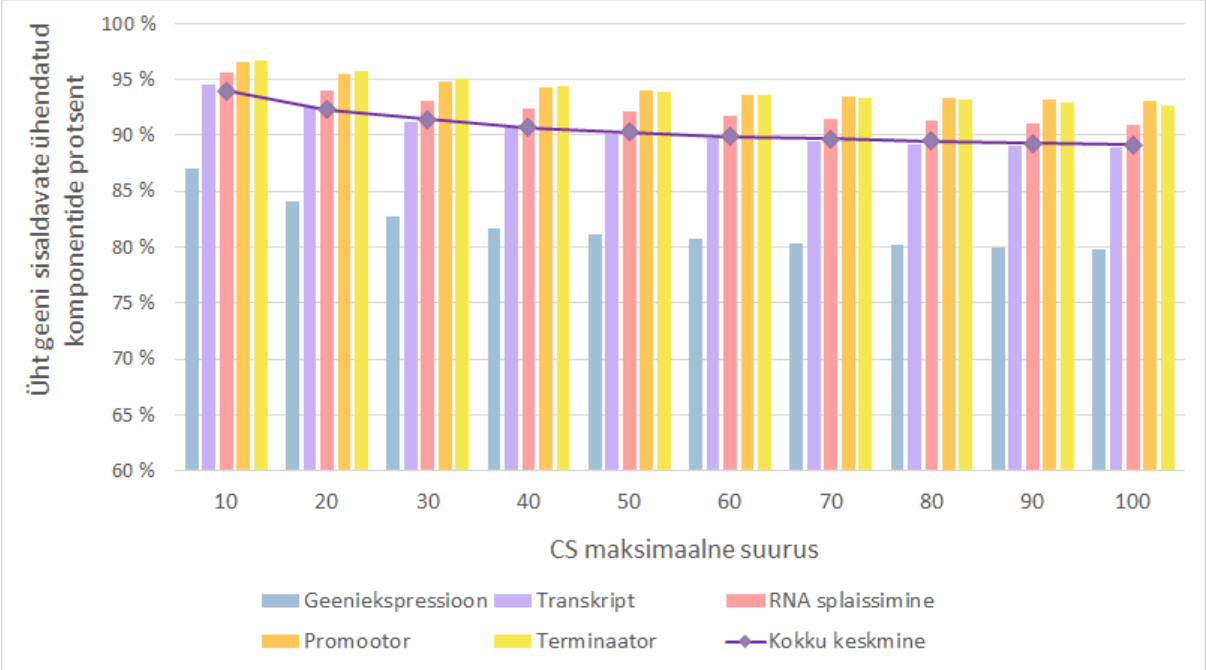
3.3 Tulemused

Maksimaalse CS suuruse muutmise ja leitud ühendatud komponentide arv iga protsessi kohta on toodud joonisel 9. Jooniselt on näha, et nii geenieksрessiooni, transkripti valiku, kui ka RNA splaissimise puhul on leitud rohkem ühendatud komponente, kui promootori ja terminaatori kasutuse kohta. Samas oli aga algandmetes nii promootorite kui ka terminaatori andmete kohta vähem geene kui teiste kohta (vt tabel 2). Kõige rohkem on leitud ühendatud komponente geenieksрessiooni puhul, mille puhul oli ka kõige rohkem geene kaardistatud. Üleüldiselt aga kasvab ühendatud komponentide arv CS maksimaalse suuruse tõstes üsna stabiilselt kuni CS suurus <= 50ni ning jääb siis umbes samasse suurusjärku. See viitab sellele, et üleüldise seisu kajastamiseks piisab vaadata spetsiifilisemalt andmeid, kus CS suurus on fikseeritud kuni 50 peale.



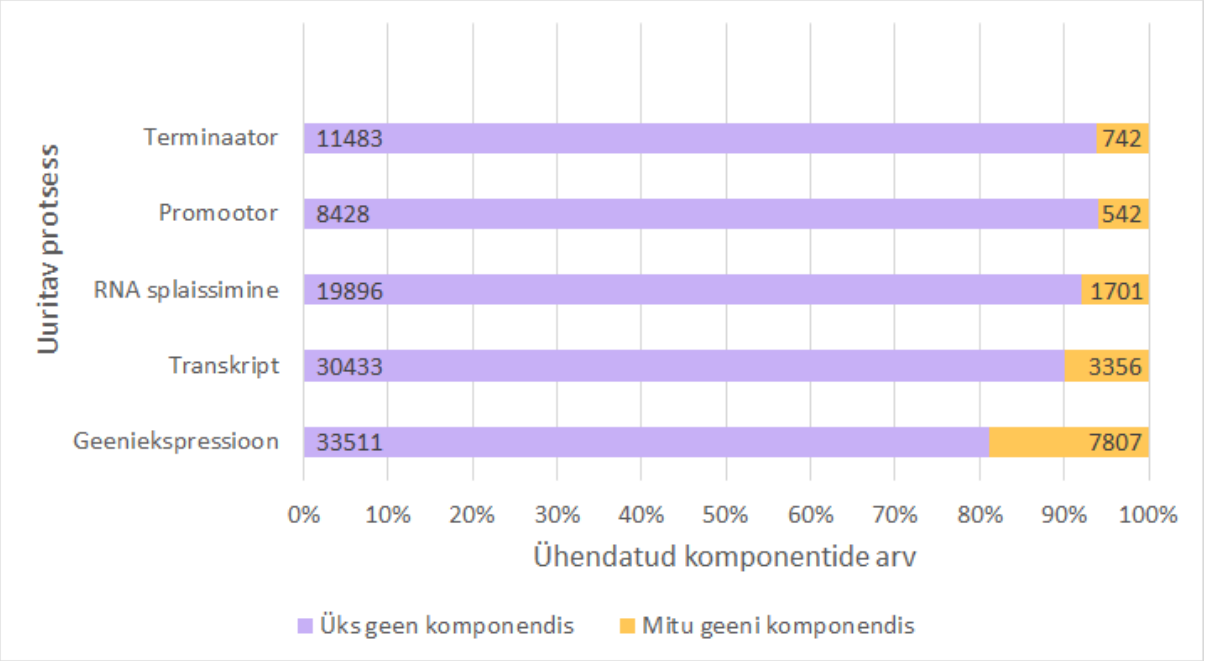
Joonis 9. Ühendatud komponentide arvu muutus erinevates protsessides vastavalt CS maksimaalsele suurusele.

Joonisel 10 on toodud, kuidas erinev CS suurus mõjutab ühte geeni sisaldavate ühendatud komponentide osakaalu. Nii transkriptide valiku, RNA splaissimise, promootorite kui ka terminaatorite kasutuse puhul on vaid üht geeni sisaldavate ühendatud komponentide osakaal palju suurem kui geenieksрessiooni puhul. Neis neljas jääb vaid üht geeni sisaldavate ühendatud komponentide osakaal 89% kuni 96% vahemikku. Geenieksрessiooni puhul on see 80% kuni 87%.



Joonis 10. Ühte geeni sisaldavate ühendatud komponentide osakaalu muutus vastavalt CS maksimaalsele suurusele.

Ühendatud komponentide geenide sisaldus, kus CSi maksimaalseks suuruseks on fikseeritud 50, on näha joonisel 11. Tulemustest joonistub välja, et võrreldes geenieksprессiooniga on transkriptide valiku, RNA splaissimise, promootorite ja terminaatorite kasutuse uuringutulemuste puhul leitud rohkem ühendatud komponente, kuhu kuulub vaid ühe geeniga seotud variante. Geenieksprессiooni puhul on vaid üht geeni sisaldavate ühendatud komponentide osakaal kõigist ühendatud komponentidest 81%. Ülejäänud nelja protsessi puhul on see suurem, neil jääb see vahemikku 90% kuni 94%. Kõige suurem üht geeni sisaldavate ühendatud komponentide osakaal on promootori ja terminaatori kasutuse puhul, mõlemal on see 94%. Järgmisena on RNA splaissimine, mille puhul on osakaal 92%. Lõpuks on transkripti valik, mille puhul on vaid üht geeni sisaldavate ühendatud komponentide osakaal 90%.



Joonis 11. Ühendatud komponentide arv vastavalt sellele, kas komponent sisaldas ühe geeni või mitme geeni erinevaid variante erinevatel protsessidel, CS maksimaalne suurus 50.

Täpne ühendatud komponentide geenide arvu sisaldus, kus CSi maksimaalseks suuruseks on fikseeritud 50, on toodud tabelis 4. Promootorite ja terminaatorite puhul on ühendatud komponentidesse kuuluvaid geene maksimaalselt vastavalt kuni 7 ja 9. Geenieksрessiooni, transkriptide valiku ja RNA splaissimise puhul kuulub ühendatud komponentidesse üldiselt kuni 12 erinevat geen. Kõigi kolme puhul leidub aga rohkem geene sisaldavaid komponente.

Kui CS maksimaalne suurus on 50, siis transkriptide valiku kõige suurema geenide arvuga ühendatud komponendi suurus on 46, mis paikneb kromosoomis 6. Teadaolevalt esineb selles piirkonnas palju geene, palju LDd ja ulatuslikku polümorfismi ehk esineb palju erinevaid geneetilisi variante (Choo, 2007; Mungall *et al.,* 2003). Seal asuvad HLA-antigeenid (ka inimese koesobivuse antigeenid, ingl. *Human-leucocyte-associated antigens*), mis reguleerivad immuunsüsteemi (Choo, 2007). HLA-antigeenide levik ja sagedus varieerub suuresti erinevate etniliste rühmade vahel, mis on tõenäoliselt tingitud nakkustekitajate erinevast levikust piirkonniti (Choo, 2007). Kuna standardmeetodite abil ei suudeta selles piirkonnas kõiki võimalikke seoseid tuvastada, ei ole täppiskaardistamise tulemused selles piirkonnas tingimata usaldusväärsed (Kennedy, Ozbek & Dorak, 2017). Seetõttu on paljudest geenidest koosnev ühendatud komponent transkripti valiku puhul pigem erandlik juhus.

Tabel 4. Ühendatud komponentide geenide sisaldus, CS maksimaalne suurus 50.

Geenide arv	Geenieksprессioon	Transkript	RNA splaissimine	Promootor	Terminaator
1	33511	30433	19896	8428	11483
2	5042	2448	1273	423	614
3	1533	561	274	75	88
4	619	153	81	26	27
5	281	103	40	12	7
6	138	41	14	4	5
7	81	16	7	2	
8	40	16	4		
9	23	8	3		1
10	12	4	2		
11	9	2	1		
12	11	1	1		
13	3				
14	4				
15		1			
16	7	1			
17	1				
19	1				
21	1				
24	1				
31			1		
46		1			

Ühendatud komponentide geenide sisalduse statistilise seose kindlaks määramiseks kontrolliti ka Fisheri täpse testiga, kas geenieksprессiooni tulemused võrdluses teistega on statistiliselt erinevad. Seda uuriti andmete peal, kus CS maksimaalseks suuruseks oli määratud 50 ning võrreldi seda, kas ühendatud komponent sisaldas üht või mitut geeni. Tulemused on toodud tabelis 5. Fisheri täpse testi tulemustest saab järeldada, et leidub statistiline seos iga protsessi vahel võrreldes geenieksprессiooniga (kõikidel on Fisheri testi tulemused $p < 0.00001$, mis on alla 5%). Ehk saab öelda, et ühendatud komponentide geenide sisalduse leidmisel on oluline, millist protsessi uuritakse. Seega transkripti valiku, RNA splaissimise, promootorite kui ka terminaatorite kasutuse puhul saab palju spetsiifilisemad tulemused. Järelikult, kuna nende puhul leidub vähem horisontaalset pleiotroopiat, tuleks põhjuslike geenide

väljaselgitamisel geenieksрessiooni asemel uurida hoopis RNA splaissimist, transkripti valikut, promootorite ja terminaatorite kasutust.

Tabel 5. Fisheri täpse testi tulemused, kus võrreldi, kas leidus statistilist seost ($p < 5\%$) ühendatud komponentide geenide sisalduses geenieksрessiooni ja teiste protsessi tulemuste vahel, CS maksimaalne suurus 50.

	Transkript	RNA splaissimine	Promootor	Terminaator
Fisheri testi tulemus	$p < 0.00001$ (1.2e-266)	$p < 0.00001$ (1.3e-321)	$p < 0.00001$ (2.3e-233)	$p < 0.00001$ (1.6e-300)

Analüüsi tulemusi võib mõjutada see, et uuritud protsesse mõjutavaid geene on juba algselt erinevalt leitud. Geenieksрessiooni mõjutavate geenide suur arv (vt tabel 2) võib tuleneda suuremast statistilisest võimsusest või sellest, et geenieksрessiooni puhul leidub rohkem horisontaalset pleiotroopiat. Statistilist võimsust mõjutab see, et geenieksрessiooni mõõtmine on oma olemuslikult täpsem kui transkripti valiku mõõtmine. CS suurus <= 50 korral on transkripti valiku puhul leitud 38 848 ühendatud komponenti ning geenieksрessiooni puhul 54 636. Küll aga on transkripti valikuga seotud geene leitud poole vähem kui geenieksрessiooni puhul. Seega ei kajastu poole väiksem seotud geenide arv täielikult ühendatud komponentide arvus (st transkripti valikuga seotud ühendatud komponentide arv pole poole väiksem geenieksрessiooni omast). See viitab sellele, et hoolimata poole väiksemast geenide arvust võimaldavad transkripti valikuga seotud variandid kaardistada tihemini ühte põhjuslikku geeni kui geenieksрessiooni puhul.

3.4 Edasi uurimiseks

Veendumaks, et transkripti valiku, RNA splaissimine, promootorite ja terminaatorite kasutuse uurimistulemused on spetsiifilisemad kui geenieksрessiooni puhul tuleb teha järeluuringuid. Esiteks, tuleks sama uuringut korrata teiste andmete peal. Teiseks, tuleks uurida sama tugevate signaalidega eQTLe ja sQTLe ja vaadata, kas need mõjutavad sama suurt hulka geene või mitte. Kui sQTLid mõjutavad väiksemal hulgal variante kui eQTLid, siis on alust arvata, et splaissimisega seotud variandid on spetsiifilisemad kui geenieksрessiooniga. Ühe andmestiku sees saaks eQTLid ja sQTLid jagada p-väärtuste põhjal vahemikesse ja vaadata, kui paljudel juhtudel mõjutavad variandid mitut geeni. Lisaks saaks uurida, kui palju leidub eksonite kasutust mõjutavaid pleiotroopseid variante ning võrrelda, kas selle tulemused on spetsiifilisemad kui geenieksрessiooni puhul.

4. Kokkuvõte

Uurides geenieksпressiooni on võimalik haiguse põhjustajaid ja neile sobilikke ravimeetodeid välja selgitada. GWAS uuringutega on kaardistatud suurel hulgal tunnustega seotud geneetilisi variante. Uurides eQTLe on suudetud kaardistada geenieksпressiooni regulatsiooni mõjutavaid põhjuslikke variante. Küll aga ei ole geenieksпressiooni regulatsioon spetsiifiline. See tähendab, et leidub pleiotroopsust ehk üks geneetiline variant mõjutab mitut geeni korraga. Lisaks, kuna esineb LDd (lähedalasuvaid geene reguleeritakse tihti koos), siis saadakse selliste uuringute tulemuseks mitmeid võimalikke põhjuslikke variante, millest on aga keeruline järeldusi teha.

Töö eesmärk oli välja selgitada, kuidas mõjutab horisontaalne pleiotroopsus geenieksпressiooni regulatsiooni. Selle jaoks analüüsiti geenieksпressiooni ja transkriptsiooni täppiskaardistamise tulemusi (Kerimov *et al.*, 2020) ning võrreldi, kas mõne protsessi puhul olid variandid seotud vähemate geenidega kui geenieksпressiooni puhul. Selleks grupeeriti usutavate variantide hulki (CSe) kokku, kui need sisaldasid samu variante, ning loodi ühendatud komponente. Seejärel uuriti, kui palju erinevaid geene kuulusid ühendatud komponentidesse. Mida vähem geene kuulus ühte komponenti, seda vähem variante oli seotud erinevate geenidega ehk seda tõenäolisemalt olid leitud geenid ka tunnust põhjustavad. Seejärel kontrolliti uurimistulemusi Fisheri täpse testiga, et teha kindlaks, kas leitud seosed on ka statistiliselt olulised.

Leiti, et kõigi nelja uuritud protsessi – transkriptide valiku, RNA splaissimise, promootori ja terminaatori kasutuse – uurimistulemused on oluliselt spetsiifilisemad kui geenieksпressiooni omad. Seega saab öelda, et geenieksпressiooni puhul leidub rohkem horisontaalset pleiotroopsust kui teiste nelja protsessi puhul. Uurimistulemustest saab järeldada, et põhjuslike geenide väljaselgitamiseks tuleks esmalt hakata uurima transkripti valikut, RNA splaissimist, promootori ja terminaatori kasutust mõjutavaid variante, sest nende protsesside puhul leidub vähem horisontaalset pleiotroopiat ja seega on nende uurimistulemused täpsemad. Nende protsesside puhul on seega ka lihtsam välja selgitada, läbi millise geeni põhjuslik variant haigust mõjutab. Edasi saab veel uurida, kui palju leidub eksonite kasutuse puhul horisontaalset pleiotroopsust ning võrrelda sama tugevaid geenieksпressiooni ja splaissimisega seotud variante.

5. Viidatud kirjandus

Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D. F., Paul, D. S., Gaffney, D. J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife, 8*, e41673. https://doi.org/10.7554/eLife.41673

Cano-Gamez, E., Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics, 11*, 424. https://doi.org/10.3389/fgene.2020.00424

Choo, S. Y. (2007). The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei medical journal, 48*(1), 11–23. https://doi.org/10.3349/ymj.2007.48.1.11

Clark, D. P. (2009). *Molecular biology.* Burlington: Elsevier Science & Technology.

Cooper, G. M. (2000). *The Cell: A Molecular Approach.* 2nd edition. Sunderland (MA): Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK9944/ (05.05.2021)

Craig, J. (2008). Complex diseases: Research and applications. *Nature Education, 1*(1), 184. https://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748/ (05.05.2021)

Dvinge, H., Guenthoer, J., Porter, P. L., Bradley, R. K. (2019). RNA components of the spliceosome regulate tissue- and cancer-specific alternative splicing. *Genome Research, 29*(10), 1591-1604. https://doi.org/10.1101/gr.246678.118

Eukaryotic pre-mRNA processing. *Khan Academy.* https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/transcription-and-rna-processing/a/eukaryotic-pre-mrna-processing (11.04.2021)

Fisher, R. A. (1941). *Statistical Methods for Research Workers.* Edinburgh: Oliver & Boyd.

Fu, X. D., Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nature reviews Genetics, 15,* 689–701. https://doi.org/10.1038/nrg3778

Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., Guigó, R. (2021). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature Communications, 12*, 727. https://doi.org/10.1038/s41467-020-20578-2

Hill, M. S., Vande, Z. P., Wittkopp, P. J. (2020). Molecular and evolutionary processes generating variation in gene expression. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-00304-w

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences, 106*(23), 9362-9367. https://doi.org/10.1073/pnas.0903103106

Hutchinson, A., Watson, H., Wallace, C. (2020). Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLoS Computational Biology, 16*(4), e1007829. ttps://doi.org/10.1371/journal.pcbi.1007829

Jordan, D. M., Verbanck, M., Do, R. (2019). HOPS: a quantitative score reveals pervasive horizontal pleiotropy in human genetic variation is driven by extreme polygenicity of human traits and diseases. *Genome Biology, 20*, 222. https://doi.org/10.1186/s13059-019-1844-7

Jukk, P. (2021). Github koodivaramu. https://github.com/pilleriinjukk/splaissimine-vs-ge (05.05.2021)

Kennedy, A. E., Ozbek, U., Dorak, M. T. (2017). What has GWAS done for HLA and disease associations? *International journal of immunogenetics, 44*(5), 195–211. https://doi.org/10.1111/iji.12332

Keren, H., Lev-Maor, G., Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews Genetics, 11*, 345-55. https://doi.org/10.1038/nrg2776

Kerimov, N., Hayhurst, J. D., Manning, J. R., Walter, P., Kolberg, L., Peikova, K., ...Alasoo, K. (2020). eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. *BioRxiv*. https://doi.org/10.1101/2020.01.29.924266

Kim, H. Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics, 42*(2), 152–155. https://doi.org/10.5395/rde.2017.42.2.152

Laurila, K., Lähdesmäki, H. (2009). Systematic analysis of disease-related regulatory mutation classes reveals distinct effects on transcription factor binding. *In silico biology, 9*(4), 209–224. https://doi.org/10.3233/ISB-2009-0398

Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., ...Pritchard, J. K. (2016). RNA splicing is a primary link between genetic variation and disease. *Science (New York), 352*(6285), 600–604. https://doi.org/10.1126/science.aad9417

Lobo, I. (2008). Pleiotropy: One Gene Can Affect Multiple Traits. *Nature Education 1*(1), 10. https://www.nature.com/scitable/topicpage/pleiotropy-one-gene-can-affect-multiple-traits-569/ (26.04.2021)

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., Darnell, J. E. (2000). Transcription Termination. *Molecular Cell Biology. 4th edition* (Section 11.1). New York: W. H. Freeman; https://www.ncbi.nlm.nih.gov/books/NBK21601/

Mitchell, K. J. (2012). What is complex about complex disorders? *Genome biology, 13*(1), 237. https://doi.org/10.1186/gb-2012-13-1-237

Modrek, B., Lee, C. (2002). A genomic view of alternative splicing. *Nature reviews Genetics, 30*, 13–19. https://doi.org/10.1038/ng0102-13

Mungall, A., Palmer, S., Sims, S., Edward C. A., Ashurst, J. L., Wilming , L., ...Beck, S. (2003). The DNA sequence and analysis of human chromosome 6. *Nature 425*, 805–811. https://doi.org/10.1038/nature02055

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., ...Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature, 466*(7307), 714–719. https://doi.org/10.1038/nature09266

Newman, A. (1998). RNA splicing. *Current Biology, 8*(25), 903-905. https://doi.org/10.1016/S0960-9822(98)00005-0

Nica, A. C., Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *368*(1620). https://doi.org/10.1098/rstb.2012.0362

OpenStax College. Eukaryotic Post-transcriptional Gene Regulation. https://courses.lumenlearning.com/boundless-biology/chapter/eukaryotic-gene-regulation/ (15.04.2021)

Paaby, A. B., Rockman, M. V. (2013). The many faces of pleiotropy. *Trends in genetics, 29*(2), 66–73. https://doi.org/10.1016/j.tig.2012.10.010

Pagán, I., Holmes, E. C., Simon-Loriere, E. (2012). Level of gene expression is a major determinant of protein evolution in the viral order Mononegavirales. *Journal of virology, 86*(9), 5253–5263. https://doi.org/10.1128/JVI.06050-11

Pagani, F., Baralle, F. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Reviews Genetics, 5*, 389–396. https://doi.org/10.1038/nrg1327

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L. C., ...Davuluri, R. V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome research, 21*(8), 1260–1272. https://doi.org/10.1101/gr.120535.111

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Reviews Genetics, 40*, 1413–1415. https://doi.org/10.1038/ng.259

Park, E., Pan, Z., Zhang, Z., Lin, L., Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *American journal of human genetics, 102*(1), 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., ...Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature, 464*(7289), 768–772. https://doi.org/10.1038/nature08872

Porrua, O., Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature Reviews Molecular Cell Biology, 16*, 190–202. https://doi.org/10.1038/nrm3943

Rahim, N. G., Harismendy, O., Topol, E. J., Frazer K. A. (2008). Genetic determinants of phenotypic diversity in humans. *Genome Biology, 9*, 215. https://doi.org/10.1186/gb-2008-9-4-215

Richards, A. L., Watza, D., Findley, A., Alazizi, A., Wen, X., Pai, A. A., ...Luca, F. (2017). Environmental perturbations lead to extensive directional shifts in RNA processing. *PLoS genetics, 13*(10), e1006995. https://doi.org/10.1371/journal.pgen.1006995

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews Genetics, 8*(6), 424–436. https://doi.org/10.1038/nrg2026

Schaid, D. J., Chen, W., Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature reviews Genetics, 19*, 491–504. https://doi.org/10.1038/s41576-018-0016-z

Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature reviews Genetics, 9*(6), 477–485. https://doi.org/10.1038/nrg2361

Spain, S. L., Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics, 24*(R1), R111–R119. https://doi.org/10.1093/hmg/ddv260

Spitz, F., Furlong, E. (2012) Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics 13*, 613–626. https://doi.org/10.1038/nrg3207

Stovner, E. B., Sætrom, P. (2020). PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics, 36*(3), 918–919. https://doi.org/10.1093/bioinformatics/btz615

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*, 661–678. https://doi.org/10.1038/nature05911

Thomson, G., Esposito, M. S. (1999). The genetics of complex diseases. *Trends in Cell Biology, 9*(12), 17-20. https://doi.org/10.1016/S0962-8924(99)01689-X

Vale, R. (2019). The Structure of DNA. The Explorer's Guide to Biology. https://explorebiology.org/summary/genetics/the-structure-of-dna (06.04.2021)

Viñuela, A., Brown, A. A., Fernandez, J., Hong, M., Brorsson, C. A., Koivula, R. W., ...Dermitzakis, E. T. (2021). Genetic analysis of blood molecular phenotypes reveals regulatory networks affecting complex traits: a DIRECT study. *medRxiv.* https://doi.org/10.1101/2021.03.26.21254347

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., ...Vidal, M. (2016). Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell, 164*(4), 805-817. https://doi.org/10.1016/j.cell.2016.01.029

Yi, L., Pimentel, H., Bray, N. L., Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology, 19*, 53. https://doi.org/10.1186/s13059-018-1419-z

Zhang, Y. (2016). On The Use of P-Values in Genome Wide Disease Association Mapping. *Journal of Biometrics & Biostatistics, 7*(3), 1000297.
https://doi.org/10.4172/2155-6180.1000297

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Pilleriin Jukk,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Pleiotroopia roll geenieksрessiooni regulatsioonis", mille juhendaja on Kaur Alasoo, 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 Commonsi 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.

Pilleriin Jukk

06.05.2021