TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Peep Kolberg

Ekspressiooni kvantitatiivsete tunnuste lookuste analüüs üksikraku RNA sekveneerimisandmetes

Magistritöö (30 EAP)

Juhendaja: Kaur Alasoo, PhD

Tartu 2023

Ekspressiooni kvantitatiivsete tunnuste lookuste analüüs üksikraku RNA sekveneerimisandmetes

Lühikokkuvõte:
Indiviididevaheline geneetiline variatsioon on rikkalik ning DNA mõju fenotüübile kindlakstegemine keeruline. Selguse toomiseks viiakse läbi uuringuid, kus mõõdetakse inimeste tunnuseid ja kaardistatakse genoom ning nende kahe vahel leitakse põhjuslikke seoseid. Üks mõõdetav tunnus on RNA ekspressioon. Tänapäeval saab RNA ekspressiooni mõõta üksiku raku tasemel, mis loob detailsema pildi rakkude heterogeensusest ja võimaldab luua tugevamaid seoseid genoomiga. Seepärast on üksikraku RNA sekveneerimisandmetega eQTL-analüüs võimas meetod DNA mõju selgitamiseks. Kui aga laborid sooritavad analüüse erisuguste meetoditega, pole publitseeritud tulemused omavahel võrreldavad ega kombineeritavad. Et valimeid maksimaalselt kasutada, peaks andmeid analüüsima üheselt või läbi metaanalüüsi. Lõputöös kasutati andmeid kolmest üksikraku RNA sekveneerimistööst ning näidati, et rakendades kõigile ühte ja sama tegevusahelat, on eQTL-signaalid samamoodi leitavad. Tulemused kinnitavad, et andmestikud sobivad metaanalüüsiks. Kombineerides erinevatest allikatest pärit andmed ja neid koos analüüsides, on suurem statistiline võimsus tuvastada olulisi seoseid.

Võtmesõnad:
üksikraku RNA sekveneerimine, eQTL-analüüs, genotüübi imputeerimine

CERCS:
B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

Expression Quantitative Trait Loci Analysis in Single-Cell RNA-Sequencing Data

Abstract:
Inter-individual genotypic variation is great and the effects of DNA on the phenotype complex. Studies measure individuals' traits, map the genome and find associations between the two in order to gain insights. One quantifiable trait is RNA expression. Presently, RNA expression can be measured at the single cell level which provides a more detailed view of the heterogeneity of cells and allows to find stronger associations with the genome. Thus, eQTL-analysis on single-cell RNA sequencing data is a powerful method for explaining the effects of DNA. However, when independent laboratories use different analysis pipelines, the published results are neither comparable nor combinable. To use samples to their fullest, data should be analyzed uniformly or through metaanalysis. The thesis used data from three single-cell RNA sequencing studies and showed that eQTL signals can be found by using the same pipeline on all datasets. The results confirm that the datasets are suitable for metaanalysis. By combining data from different sources and analyzing it together, there is greater statistical power to find significant associations.

Keywords:
single-cell RNA sequencing, eQTL analysis, genotype imputation

CERCS:
B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

1 Sissejuhatus

Ekspressiooni kvantitatiivse tunnuse lookus (*expression quantitative trait locus*, eQTL) on genoomi piirkond, mis on statistiliselt seotud geeni(de) ekspressiooniga [1]. Selliste lookuste tuvastamine aitab selgitada genotüübi mõju fenotüübile, mida ainult diferentsiaalset geeniekspressiooni uurides ei näe. Kuna eQTL ei pruugi geeniekspressiooni mõjutada igas koes, vaid ainult teatud rakutüüpides [2], on otstarbekas sooritada eQTL-analüüs proovides, mis sisaldavad ainult ühte rakutüüpi.

Selliseid puhtaid rakuproove saab koostada ja nende geeniekspressiooni mõõta üksikraku RNA sekveneerimisega (*single-cell RNA sequencing, scRNA-seq*) [3]. Meetod isoleerib koeproovis olnud rakud ja sekveneerib RNA nii, et igalt rakult pärinenud lugemid on võimalik kindlaks teha [4]. Seejärel saab markergeenide abil identifitseerida iga raku tüübi. Niiviisi kogutakse puhtaid proove mitmest rakutüübist korraga [3].

Kuigi on tehtud palju hulkraku RNA sekveneerimistöid puhastatud koeproovides [2], pole siiski suurele osale haigustega seotud geneetilistele variantidele leitud sihtmärkgeeni [5]. Põhjenduseks tuuakse, et koed sisaldavad heterogeenseid alamrakutüüpe, mida pole hulkrakuproovide puhastamisega ikkagi võimalik eristada [6]. Üksikraku RNA sekveneerimine võimaldab rakutüüpe diferentseerida väga kõrge resolutsiooniga ning tuvastada ka haruldasi rakutüüpe [3].

Üksikraku RNA sekveneerimine loob detailsema pildi geeniekspressioonist üksikute rakkude tasemel. Koos tehnoloogia hinna langusega on üksikrakutööde arv viimastel aastatel oluliselt kasvanud [7]. Aga individuaalsetes uuringutes võib olla statistiline võimsus piiratud ning analüüsimeetodid võivad erineda, mistõttu pole avaldatud statistikud siiski omavahel võrreldavad. Seepärast on loodud ühendus „The single-cell eQTLGen consortium“ [8] metaanalüüsimaks üksikrakuandmeid ning ka lõputöös analüüsiti eri allikatest pärit andmeid üheselt.

Lõputöö eesmärk oli tuvastada eQTL-e kolmes üksikrakuandmestikus: Randolph_2021 [9], OneK1K_2022 [10], Perez_2022 [11]. Kõigile kolmele andmestikule rakendati ühte ja sama analüüsimeetodit, et näha, kui efektiivne on eQTL-analüüs ilma metoodikat ühele konkreetsele andmestikule sobitamata. Lõputöö oli sissejuhatuseks plaanitavale viie üksikrakuandmestiku metaanalüüsile.

Lõputöös joondati üksikraku RNA sekveneerimise andmed, üksikrakumaatriksitest koostati pseudo-hulkrakumaatriksid (*pseudobulk*) ning pseudo-hulkrakumaatriksitega sooritati eQTL-analüüs. Lisaks loodi genotüüpide imputeerimise töövoog, mida kasutati ühest uuringust kõrge tihedusega genotüüpide loomiseks, ja katsetati ka meetodit üksikrakuandmete demultipleksimiseks.

Lõputöö sisu on jaotatud neljaks peatükiks. Tausta peatükis tutvustatakse genotüüpide imputeerimist ning kirjeldatakse, millal on imputeerimine vajalik. Lisaks antakse ülevaade eQTL-analüüsist ning räägitakse üksikraku RNA sekveneerimisandmete kogumisest ja analüüsist. Meetodite peatükk algab lõputöös kasutatud andmestike tutvustusega. Seejärel kirjeldatakse detailselt, kuidas sooritati lõputöös madala katvusega genotüüpide imputee-

rimine, üksikrakuandmete joondamine, demultipleksimine, pseudo-hulkrakumaatriksite loomine ning eQTL-analüüs. Tulemuste peatükis näidatakse iga etapi väljundit ning võrreldakse lõputöö tulemusi teiste autorite töödega. Lõpetuseks lahatakse arutelu peatükis põhjuseid, mis võisid tulemusi mõjutada, ning millised on suunad edasisteks uuringuteks.

2 Taust

Peatükis tutvustatakse lõputööga seotud valdkondi. Esiteks antakse ülevaade genotüüpide imputeerimisest ja eQTL-analüüsist. Seejärel kirjutatakse täpsemalt üksikraku RNA sekveneerimistehnoloogiast ja selle eripäradest. Lisaks kirjeldatakse, kuidas üksikraku-andmeid töödelda, et neid saaks analüüsida levinud hulkraku RNA sekveneerimiseks mõeldud meetoditega.

2.1 Genotüüpide imputeerimine

Imputeerimine tähendab puuduvate või madala katvusega sekveneeritud geneetiliste variantide kindlakstegemist. Imputeerimine on statistiline meetod, kus puuduvad variandid määratakse, kasutades aheldatuse tasakaalutuse (*linkage disequilibrium*, LD) mustreid. Kuna inimpopulatsioonis on teatud haplotüübid rohkem levinud, saab puuduvad variandid määrata lähedalasuvate variantide põhjal. Seega ei pea kõrge tihedusega genotüüpide saamiseks sooritama sügavat sekveneerimist, vaid võib piirduda madala katvusega sekveneerimisega ja pärast genotüübid imputeerida. [12]

Madala katvusega sekveneerimine koos imputeerimisega on odavam sügavast sekveneerimisest. Aga sõltuvalt edasisest analüüsist võib vaja minna just kõrge tihedusega genotüüpe. Geneetilised assotsiatsiooniuuringud (*genome-wide association studies*, GWAS) ja kvantitatiivsete tunnuste lookuste täppiskaardistamine (*fine-mapping*) vajavad suure tihedusega genotüübiandmeid, et tuvastada enam põhjuslikke variante. Imputeerimine täiendab indiviidide genotüüpe variantidega, mida sekveneerimisel või genotüpiseerimisel ei tuvastatud. See tähendab, et imputeeritud genotüüpe kasutades on suurem võimsus statistiliselt oluliste variantide tuvastamiseks. [13]

Valides genotüpiseerimismeetodiks odavama, madala katvusega sekveneerimise, saab uuringusse kaasata rohkem indiviide, mis suurendab testide võimsust veelgi. Inimgeneetikas on genotüüpide imputeerimine usaldusväärne viis suure tihedusega genotüüpide koostamiseks, sest inimese haplotüüpe on põhjalikult kaardistatud ning erinevate populatsioonide kohta on koostatud spetsiifilised viitepaneelid (*reference panel*). [13]

Populaarsed imputeerimisprogrammid Beagle 5.0 [14], Minimac4 [15], GLIMPSE2 [16] töötavad viitepaneeli põhjal. Kuna imputeerimise täpsus sõltub lisaks viitepaneeli sobivusele ka paneeli suurusest, peab imputeerimistarkvara skaleeruma sadu tuhandeid haplotüüpe sisaldavatele paneelidele [13, 16]. Loetletud programmid kasutavad erinevaid algoritme võimalikult kiiresti ja vähese mälukasutusega imputeerimiseks. Samas on kõigi täpsus nii suur, et võrreldes nende imputeeritud genotüüpe kõrge katvusega sekveneeritud genotüüpidega, on genotüübid levinud variantide (MAF > 1%) kohal praktiliselt samad [14, 16, 17].

2.2 Ekspressiooni kvantitatiivsete tunnuste lookuste analüüs

Ekspressiooni kvantitatiivse tunnuse lookus viitab geneetilisele variandile, mis on seotud geeniekspressiooniga. eQTL-analüüs tähendab selliste lookuste ja nende sihtmärkgeenide kindlakstegemist. Tavaliselt eristatakse kahte tüüpi eQTL-e: *cis*-eQTL-id ja *trans*-eQTL-id. Lookus loetakse *cis*-eQTL-iks, kui ta asub oma sihtmärkgeeni lähedal, st samal kromosoomil ja maksimaalselt 1 Mbp kaugusel (1 Mbp on sageli rakendatud lävend, kuid kasutatakse ka väärtusi 100 kbp-st kuni 5 Mbp-ni). [18]

eQTL-analüüs paljastab regulatoorsete variantide mõju fenotüübile. Näiteks saab kõrvutada eQTL-analüüsi ja mõne haiguse GWAS-i tulemusi ehk vaadata eQTL-ide ja GWAS-i variantide kolokaliseerumist. Nii saab tuvastada variandid, mis on samaaegselt seotud mõne geeni ekspressiooniga ja haiguse esinemisega. See aitab tuua päevavalgele haiguse põhjuslikud geenid, mida ainult GWAS-i põhjal on keeruline tuvastada. [19]

eQTL-ide tuvastamine on populaarne analüüsimeetod, kuid erinevad laborid koguvad ja analüüsivad andmeid erinevalt, mistõttu on publitseeritud tulemusi keeruline omavahel võrrelda. Seepärast loodi eQTL Catalogue [20], kuhu koguti indiviiditasemel andmed 112-st andmestikust ning need analüüsiti ühtemoodi. Tulemusena valmis andmebaas, kus on analüüsitud ja kaardistatud eQTL-id 69-st erinevast koest ning tänu ühesele töötlusele saab erinevate kudede eQTL-e otse omavahel võrrelda ja uurida geeniekspressiooni regulatsiooni koespetsiifilisust. [20]

Mida enam sekveneerimisandmeid kogutakse, seda suuremate valimitega saab eQTL-metaanalüüsi läbi viia [21]. Suurem valim tähendab suuremat statistilist võimsust, mis on eriti vajalik *trans*-eQTL analüüsis, kus statistiliste testide arv võib olla 1000 korda suurem kui *cis*-analüüsis [18].

Kui aga metaanalüüsi sooritada, peab jälgima, et erinevate uuringute koeproovid oleksid piisavalt sarnased. Tehniline variatsioon ja proovide erinev rakukooslus muudavad proovid omavahel heterogeensemaks [21]. Tagajärjena on raskem tuvastada koespetsiifilisi eQTL-e. Üksikraku RNA sekveneerimisandmetega on selle võrra lihtsam metaanalüüsi sooritada, kuna rakutüübid saab määrata üheselt üle kõigi uuringute.

2.3 Üksikraku RNA sekveneerimine

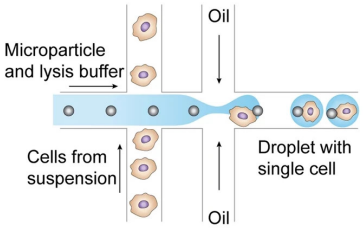
2.3.1 Tehnoloogia tutvustus

Üksikraku RNA sekveneerimine on sekveneerimismeetod, kus RNA sekveneeritakse üksiku raku tasemel [3]. Kui hulkraku RNA sekveneerimine näitab keskmist geeniekspressiooni üle terve koeproovi, siis üksikrakumeetodid toovad päevavalgele iga raku geeniekspressiooni mustri.

Tänapäeval on käsitsi üksikute rakkude eraldamise kõrvale arendatud süsteemid, mis isoleerivad korraga kümneid tuhandeid rakke ning teevad seda kordades kiiremini ja odavamalt kui inimjõududega võimalik [4]. Populaarseimad tehnoloogiad (inDrop,

Drop-seq, 10X) kasutavad rakkude separeerimiseks õliemulsiooni ning RNA püüdmiseks DNA järjestustega kaetud kuulikesi, mis rakkudega segatakse. Rakkude ettevalmistamine on illustreeritud joonisel 1. Rakud liigutatakse ühekaupa läbi mikroskoopilise kanali, kus iga rakk saab endale paariliseks ühe DNA järjestustega kaetud kuulikese. Väljaloetava järjestuse osadeks on tavaliselt

- kuulikesele unikaalne ribakood (*barcode*) raku identifitseerimiseks;
- unikaalne molekulaarne identifikaator (*unique molecular identifier* - UMI) RNA molekulide eristamiseks;
- polü-T järjestus mRNA polü-A sabade püüdmiseks;
- polümeraasi ahelreaktsiooni (PCR) praimer. [3, 4]



Joonis 1. Üksiku raku RNA sekveneerimiseks emulgeeritakse rakud õlis. Rakud liiguvad üksteise järel läbi mikroskoopilise kanali, kus iga rakk saab endale paariliseks ühe DNA järjestustega kaetud kuulikese. Edasi luuakse õliemulsioon nii, et iga tilga sisse jääb üks raku ja kuulikese paar. [3, kohandatud]

Rakk koos oma kuulikesega emulgeeritakse õlis nii, et iga tilga sisse jääb üks rakk koos ühe kuulikesega. Edasi rakud lüüsuvad, mRNA seondub kuulikese küljes olevate järjestustega, transkriptidele pöördtranskribeeritakse komplementaarne DNA ahel, viiakse läbi DNA amplifikatsioon ja sekveneerimine. [3]

Kuigi mRNA sellisel viisil püüdmine võimaldab mõõta geeniekspressiooni ühe raku tasemel, peab arvestama ka tehnoloogia iseärasusi. Esiteks, rakkude segamine kuulikestega ei toimu perfektselt. Tekib ka tilku, kus ei ole täpselt üks kuulike koos täpselt ühe rakuga. Sellised tilgad saab tuvastada alles joondamismisjärgses kvaliteedikontrollis ning need tuleb kindlasti andmestikust eemaldada, kuna need ei esinda ühe raku geeniekspressiooni. [4]

Teiseks, üksikraku RNA sekveneerimisel seondub kuulikestega vaid 10 - 20% mRNA transkriptidest [3]. See tähendab, et loenduste arvud lugemimaatriksis on väga alahinnatud. Lugemimaatriks on väga hõre, sisaldab palju nulle. Lisaks, kuna kuulikestele

seonduvad transkriptid oma 3' polü-A kaudu, saavad sekveneeritud ainult fragmendid geenide 3' otstest.

Kolmandaks, rakud sekveneeritakse partiide (*batch*) kaupa. Üksikraku RNA sekveneerimise puhul on täheldatud, et partii on tugev tehnilise variatsiooni allikas [3, 4]. See tähendab, et näiv rakkudevaheline diferentsiaalne geeniekspressioon ei ole tingitud erinevast mRNA hulgast, vaid raku partiist. Selline partii efekt (*batch effect*) on vaja välja regresseerida. Tavaliselt indiviidid multipleksitakse, st ühte partiisse pannakse kokku mitme indiviidi rakud, et partii efekti vähendada. Nii toimetades kaob aga informatsioon, milline rakk milliselt indiviidilt pärines. Pärast sekveneerimist on vaja arvutuslikult rakkudele indiviidid määrata ehk indiviidid on vaja demultipleksida.

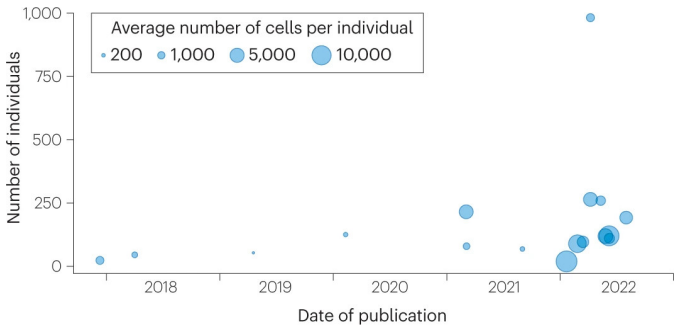
Neljandaks, teadlased peavad leidma tasakaalu uuringusse kaasatud indiviidide arvu, igalt indiviidilt sekveneeritud rakkude arvu ja uuringu eelarve vahel. Kaasates rohkem indiviide, paraneb võimsus geneetiliste variantide (eriti haruldaste) tuvastamiseks. Sekveneerides enam rakke indiviidi kohta, suureneb võimalus tabada rohkem harvaesinevaid rakutüüpe. Need mõlemad tähendavad, et sekveneerida tuleb rohkem rakkude partiisid, mis on aga kulukam. Sekveneerida rohkem rakke korraga ühes partiis on odavam, aga tõstes partii rakkude arvu suureneb ka topeltrakkude (*doublet* - tilk, kus oli mitu rakku) tekkimise oht. [7]

Tehnoloogia miinuseid arvestades, on geeniekspressiooni mõõtmine ühe raku resolutsiooniga siiski kaalukam pluss. Sellest annab tunnistust publitseeritud üksikraku RNA sekveneerimistööde arvu hüppeline kasv viimasel kahel aastal [7], mis on näidatud ka joonisel 2. Randolph jt [9], Yazar jt [10], Perez jt [11] viisid läbi üksikraku RNA sekveneerimise ja kogutud andmete põhjal *cis*-eQTL-analüüsi. Kõik kolm uuringut teostati perifeerse vere mononukleaarsetel rakkudel (*peripheral blood mononuclear cells*, PBMCs). Randolph jt võrdlesid immuunsüsteemi vastust gripiga nakatumisele Euroopa ja Aafrika päritolu indiviidide vahel. Yazar jt kaardistasid eQTL-e immuunrakkudes. Perez jt võrdlesid geeniekspressiooni süsteemse erütematoosse luupuse patsientide ja terve kontrollgrupi vahel. Igas uuringus tuvastati erinevates rakutüüpides tuhandeid eQTL-e. [9, 10, 11]

Et maksimaalselt kasutada erinevates projektides kogutud andmeid, saab sooritada veel metaanalüüsi. Erinevatest allikatest pärit üksikrakuandmed saab kombineerida, üheselt töödelda ning koos analüüsida. Sellisel viisil eQTL-ide tuvastamiseks loodi ühendus „The single-cell eQTLGen consortium" [8]. Ühenduse eesmärk on metaanalüüsiga kaardistada eQTL-e.

2.3.2 Pseudo-hulkrakumaatriksid

Üksikraku RNA sekveneerimisandmed sisaldavad palju tehnilist müra. Loendused on alaesindatud, rakkudevaheline geeniekspressiooni variatsioon on suur, ebakorrektsete tilkade filtreerimine ning rakutüüpide määramine pole veavabad [3]. On arendatud meetodeid [22, 23], mis töötavad otse üksikrakuandmetel, kuid tavaliselt töödeldakse



Joonis 2. Viimasel kahel aastal on üksikraku RNA sekveneerimine kõvasti populaarsust kogunud. Iga ring tähistab ühte publitseeritud tööd. [7]

üksikrakuandmeid nii, et need simuleeriksid hulkrakuproove. Seejärel saab rakendada juba kandakinnitanud hulkrakumeetodeid [7].

Agregeerides ühte tüüpi ja ühe indiviidi rakud kokku, paistavad rakutüübile ja indiviidile omased geeniekspressioonimustrid müra taustal välja. Sekveneerides igalt indiviidilt piisavalt rakke, saab koostada nn pseudo-hulkrakumaatriksid, mis näevad välja justkui hulkraku RNA sekveneerimise loendusmaatriksid, kuid on koostatud ainult ühte kindlat tüüpi rakkudest. [24]

Enne pseudo-hulkrakumaatriksite koostamist on vaja üksikrakuandmestikust eemaldada proovide valmistamisel tekkinud vigased tilgad. Andmestikku tohivad alles jääda ainult tilgad, kus oli täpselt üks rakk täpselt ühe kuulikesega. Eriti valmistavad probleeme tilgad, kus oli üks kuulike mitme eri tüüpi rakuga. Loendusmaatriksis näeb selline tilk välja kui üks rakk, aga raku geeniekspressioon ei ole omane ühelegi rakutüübile. Sellised tilgad segavad piire rakutüüpide vahel ning võivad tekitada soovi defineerida uus vahepealne rakutüüp, mida tegelikult ei eksisteeri. [25]

Vigaste tilkade tuvastamiseks on mitmeid mooduseid. Võib ise defineerida parameetrid, mille järgi tilku filtreerida. Parameetriteks võivad olla näiteks minimaalne transkriptide (UMI) arv, minimaalne ekspresseerunud geenide arv, maksimaalne mitokondriaalsele DNA-le joondunud lugemite osakaal [25]. Aga võib kasutada ka tööriistu, mis filtreerimist sooritavad, nt Cell Ranger [26], dropkick [27].

Kui andmestik on vigastest tilkadest puhastatud, saab koostada pseudo-hulkrakumaatriksid. Erinevaid üksikrakkude agregeerimismeetodeid võrdlesid Cuomo jt [24]. Nad kasutasid samadelt indiviididelt võetud hulk- ja üksikraku RNA sekveneerimise proove. Hulkrakuandmetel sooritati *cis*-eQTL-analüüs. Üksikrakuandmetele rakendati erinevaid agregeerimismeetodeid ning vaadati, millise meetodiga replitseeruvad hulkrakutulemused kõige paremini.

Võrreldi kahte grupeerimisvõimalust: esiteks ainult indiviidi ja rakutüübi kaupa, teiseks indiviidi, rakutüübi ja partii kaupa. Partiid tuli grupeerimisel arvesse võtta, kui

ühe indiviidi rakud jagati sekveneerimisel mitme partii vahel. Grupeeritud rakkude loendustest võeti kas aritmeetiline keskmine, mediaan või summa. Nii sai mitmest rakust üks pseudo-hulkrakuproov. [24]

Parim agregeerimisvõte oli grupeerimine indiviidi, rakutüübi ja partii kaupa ning grupi kaupa loendustest keskmise võtmine. Partii arvestamine grupeerimisel oli vajalik, kuna sekveneerimisel olid osade indiviidide rakud jagatud mitme partii vahel. Autorid põhjendasid aritmeetilise keskmise ülekaalu mediaani ees sellega, et üksikrakumaatriksi hõreduse tõttu on mediaan halb lugemite arvu iseloomustaja. Ja kuna enne agregeerimist normaliseeriti iga raku lugemite arv, oli aritmeetiline keskmine parem geeniekspressiooni kirjeldaja kui summa. [24]

3 Meetodid

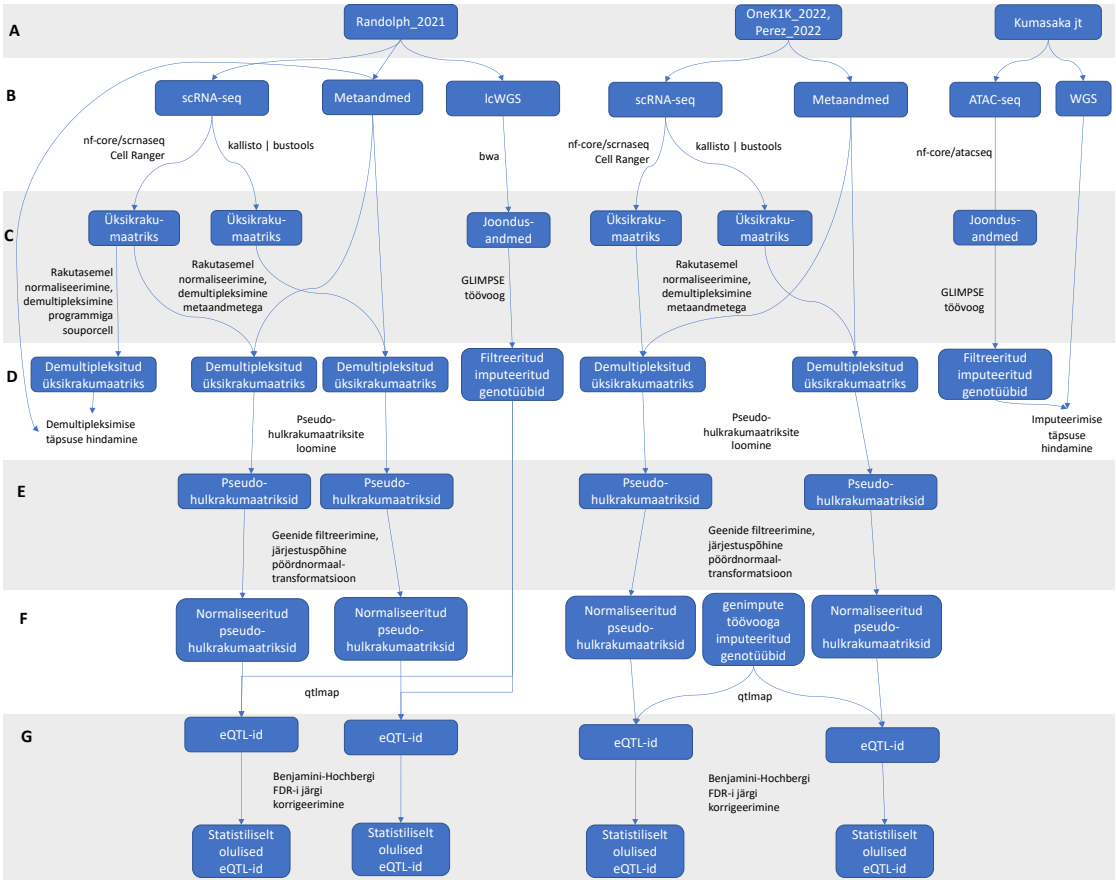
Peatükk algab lõputöös kasutatud andmestike kirjeldusega. Järgnevalt selgitatakse, kuidas rakendati lõputöös loodud genotüüpide imputeerimise töövoogu Randolph_2021 madala katvusega sekveneeritud genotüüpidele. Edasi kirjeldatakse detailselt, kuidas üksikraku-andmed joondati, katsetati demultipleksimist ning koostati pseudo-hulkrakumaatriksid. Viimaseks näidatakse eQTL-analüüsi jooksutamiseks tehtud samme. Kasutatud programmide kohta on kirjutatud ka põgusad tutvustused. Lõputöös ei tehtud üksikrakkudele rakutüüpide määramist, vaid toetuti uuringute metaandmetele. Kogu protsess on kujutatud joonisel 3.

3.1 Andmestikud

Lõputöös tuvastati eQTL-e kolmes kohordis: Randolph_2021 [9], OneK1K_2022 [10], Perez_2022 [11]. Kõigist kolmest uuringust kasutati 10X Chromium Single Cell 3' (chemistry v2) meetodiga kogutud PBMC üksikraku RNA sekveneerimise andmeid. Kokku sisaldasid andmestikud 2.77 miljonit rakku 1332-lt indiviidilt. Täiendavalt, Randolph_2021 uuringust kasutati lõputöös madala katvusega (4X) täisgenoomi sekveneerimisandmeid (*low-coverage whole genome sequencing*, lcWGS), OneK1K_2022 ja Perez_2022 uuringutest genotüübikiibiga genotüpiseeritud ja eQTL-Catalogue/genimpute [31] töövooga imputeeritud genotüübiandmeid. Detailid andmestike kohta on toodud tabelis 1. Lisaks on tabelis 2 toodud portaalid, kust andmed allalaeti.

Tabel 1. Ülevaade andmestikest. Uuringutes määrati indiviidide päritolu vastavalt sellele, millisele 1000 Genoomi Projekti [32] (1000 Genomes Project) metapopulatsioonile oli indiviid kõige sarnasem. Indiviidide, rakutüüpide ja rakkude arvud kajastavad ainult autorite annoteeritud rakke, st autorite kvaliteedikontrollid edukalt läbinud rakke. Sulgudesse on kirjutatud lõputöös kasutatud isendite arv.

Kohort	**Kude**	**Andmed - Tehnoloogia**	**Indiviide**	**Rakutüüpe**	**Rakke (mln)**
Randolph, 2021	PBMC	*scRNA-seq* - 10X 3' v2 lcWGS - DNBseq	89 (89) EUR - 50% AFR - 50%	30 (12)	0.24 (0.23)
OneK1K, 2022	PBMC	*scRNA-seq* - 10X 3' v2 WGS - kiip	982 (981) EUR - 100%	30 (24)	1.27 (1.25)
Perez, 2022	PBMC	*scRNA-seq* - 10X 3' v2 WGS - kiip	261 (193) EUR - 57% EAS - 41% AFR - 1% AMR - 1%	11 (9)	1.26 (0.82)



Joonis 3. Lõputöö analüüsisammud. **A**. eQTL-analüüsiks kasutati kolme andmestikku, imputeerimistöövoo testimiseks ühte. **B.** Üksikraku RNA sekveneerimise andmed joondati kahe programmiga. Randolph_2021 lcWGS andmed joondati programmiga bwa. Kumasaka jt [28] *ATAC-seq* andmed joondati nf-core/atacseq töövooga. **C**. Kõik üksikrakumaatriksid normaliseeriti rakutasemel. Randolph_2021 üksikrakumaatriksid demultipleksiti nii programmiga souporcell kui metaandmetega. Lisaks imputeeriti Randolph_2021 ja Kumasaka jt joondusandmed GLIMPSE töövooga [29]. Ülejäänud üksikrakumaatriksid demultipleksiti vaid metaandmeid kasutades. **D**. Testiti Randolph_2021 souporcelliga demultipleksimise ning Kumasaka jt genotüüpide imputeerimise täpsust. Üksikrakumaatriksitest koostati pseudo-hulkrakumaatriksid. **E**. Pseudo-hulkrakumaatriksites filtreeriti geenid ja rakendati järjestuspõhine pöördnormaal-transformatsioon. **F**. Normaliseeritud pseudo-hulkrakumaatriksite ja imputeeritud genotüüpidega sooritati eQTL-analüüs programmiga qtlmap [30]. **G**. Statistiliselt oluliste eQTL-ide leidmiseks korrigeeriti p-väärtused Benjamini-Hochbergi FDR-i järgi.

Tabel 2. Allalaetud andmete kättesaadavus.

Kohort	**Uuring**	**Portaal**	***Accession Number***
Randolph_2021	*scRNA-seq*	GEO	GSE162632
	lcWGS	SRA	PRJNA736483
OneK1K_2022	*scRNA-seq* + genotüübid	GEO	GSE196830
Perez_2022	*scRNA-seq*	GEO	GSE174188
	genotüübid	dbGaP	phs002812.v1.p1

3.2 Genotüüpide imputeerimine

OneK1K_2022 ja Perez_2022 genotüübid olid genotüpiseeritud genotüübikiibiga ning ka juba imputeeritud eQTL-Catalogue/genimpute [31] töövooga. Lõputöös imputeeriti ainult Randolph_2021 madala katvusega genotüübid. Lugemid joondati viitegenoomile (*reference genome*, versioon GRCh38) kasutades tarkvara bwa (v0.7.17) [33]. Täpsemalt, kasutati käsku `bwa mem` vaikeparameetritega. Joondumise kvaliteeti kontrolliti tarkvara samtools [34] (v1.14) käsuga `flagstat`. Kõigi indiviidide puhul oli korrektselt joondunud (tähistatud *properly paired* - lugemipaar joondus samale kromosoomile, oli õigesti orienteeritud ja mõistliku insertsiooni pikkusega) lugemite osakaal üle 97%.

Lõputöö analüüsi sooritamiseks loodi genotüüpide imputeerimise töövoog [29], mis kasutab tarkvara GLIMPSE [35]. GLIMPSE on madala katvusega täisgenoomi sekveneerimisandmete imputeerimis- ja faasimistarkvara. Programm kasutab imputeerimiseks haplotüüpide viitepaneeli. Täpsemalt, programm võtab sisendiks viitepaneeli ja madala katvusega sekveneerimise joondusandmete põhjal arvutatud genotüüpide tõepärad (*genotype likelihoods*). Genotüüpide tõepärad on arvutatud iga viitepaneeli positsiooni kohta ning GLIMPSE parendab iga indiviidi tõepärasid Gibbsi valiku (*Gibbs sampler*) põhimõttel, võttes arvesse viitepaneeli ja teisi indiviide. Iga imputeeritud positsiooni kohta arvutab GLIMPSE ka imputeerimise kvaliteediskoori, mis kirjutatakse väljundi VCF-faili tulpa INFO. [35]

GLIMPSE arendati eesmärgiga luua vähe arvutusressursse nõudev imputeerimisprogramm, mis suudaks imputeerimisel kasutada ka suuri viitepaneele [35]. Tarkvarale loodi ka edasiarendus: GLIMPSE2, mis kasutab veelgi vähem arvutusjõudu ning on haruldaste variantide imputeerimises parem kui esimene versioon [16]. Siiski, lõputöös ei kasutatud tarkvara viimast versiooni, kuna see avaldati vahetult enne töövoo valmissaamist ja esimese versiooniga imputeerimise tulemused olid nii head, et ei nähtud vajadust töövoog ümberkirjutada. Töövoog kasutab tarkvara esimest väljalaset (GLIMPSE v1.1.1), mille lähtekood on olemas ka GitHubis [36].

Kuidas arvutati lõputöös genotüüpide tõepärad ning milliste parameetritega imputeerimisprogramm jooksutati, on leitavad töövoo koodivaramus GitHubis [29]. Viitepaneelina kasutati 1000 Genoomi Projekti sõltumatute indiviidide viitepaneeli [32], millest olid imputeeritavad indiviidid välja võetud, et mitte imputeerijat ülesobitada. Imputeeriti

ainult autosoomid.

Imputeerimise täpsuse kontrolliks kasutati transposaasile ligipääsetava kromatiini sekveneerimise (*assay for transposase-accessible chromatin using sequencing, ATAC-seq*) andmeid Kumasaka jt [28] tööst, kus lisaks *ATAC-seq*-ile sooritati ka täisgenoomi sekveneerimine.

Kromatiini sekveneerimise andmed on sarnased madala katvusega genoomisekveneerimisandmetega. *ATAC-seq* annab umbes sama palju lugemeid üle genoomi kui lcWGS, aga need pole nii ühtlaselt jaotunud. Lisaks, kromatiini sekveneerimisel kasutatakse sama transposaasi, mida madala katvusega genoomisekveneerimiselgi. Seega on *ATAC-seq* andmed sarnased lcWGS andmetega ning imputeerija täpsus *ATAC-seq* andmetel annab ettekujutuse täpsusest lcWGS andmetel.

ATAC-seq andmed joondati programmiga nf-core/atacseq (v1.2.2) [37]. Imputeerimise täpsuse hindamiseks arvutati iga indiviidi igal kromosoomil filtreeritud (MAF > 0.01) imputeeritud genotüüpide ja tegelike genotüüpide vahel alleelidooside (*minor allele dosage* - alternatiivse alleeli esinemise arv positsioonis) lineaarne korrelatsioon.

3.3 Üksikraku RNA sekveneerimine

3.3.1 Joondamine Cell Rangeri ja kallisto | bustoolsiga

Kõik üksikrakuandmestikud joondati kahe erineva programmiga, et võrrelda programmide mõju eQTL-ide tuvastamise efektiivsusele. Esimene programm oli Cell Ranger [26], mis sisaldab erinevaid üksikrakuandmete analüüsi tööriistu. Cell Ranger võtab sisendiks sekveneeritud lugemid ja sooritab lugemite joondamise, tilkade filtreerimise, rakkude klasterdamise (*clustering*) jm normaliseerimise ning geeniekspressiooni uurimise samme [38].

Cell Rangeri jooksutamiseks kasutati nf-core tööriistade kollektsiooni [39]. Kogumikust paigaldati töövoog nf-core/scrnaseq (v2.1.0), mis käivitati vastavalt dokumentatsiooni kasutusjuhistele [40] ning täpsustades parameetri `aligner` väärtusega `cellranger`. Viitegenoom (GRCh38) oli allalaetud Ensembli serverist [41] ning GTF fail EMBL-EBI serverist [42]. Lisaks, Perez_2022 andmeid joondades ei õnnestunud programmil osade rakupartiide puhul automaatselt tuvastada, mis meetodiga olid üksikrakuandmed kogutud. Probleemist ülesaamiseks muudeti Perez_2022 andmete joondamiseks programmi lähtekoodi, täpsustades käsitsi parameetri `chemistry` väärtusega `SC3Pv2` (*Single-Cell 3' v2*). Töövoog nf-core/scrnaseq kasutab ainult Cell Rangeri joondamisfunktsiooni (`cellranger count`), mis koostab rakud $\times$ geenid loendusmaatriksi.

Teiseks joondati üksikrakuandmed ka tarkvaraga kallisto | bustools [43] (edaspidi „kallisto”), mis ei ole nii multifunktsionaalne kui Cell Ranger, aga see-eest on oluliselt kiirem ja väiksema mäluvajadusega. Programmi ajavõit tuleb sellest, et lugemeid ei joondata aluspaari-täpsusega viitegenoomile, vaid sooritatakse pseudojoondus (*pseudoalignment*) viitetranskriptoomile [43]. Meetodi autorid näitasid, et sellisel viisil joondamine on

kiirem ning toodab Cell Rangeri üksikrakumaatriksiga kõrgelt korreleeritud loendused.

kallisto jooksutati kb_python paketi [43] (v0.27.3) kaudu. Kasutati funktsiooni `kb count`, mis loob sama tüüpi loendusmaatriksi kui Cell Ranger. Pseudojoonduseks vajalik indeks loodi vastavalt dokumentatsiooni kasutusjuhistele [44] ning väljundi failitüübiks määrati .h5ad.

3.3.2 Demultipleksimine

Kuna Randolph_2021, OneK1K_2022, Perez_2022 uuringutes sekveneeriti mitme indiviidi proovid ühes partiis, pidid autorid pärast üksikrakkude loendusmaatriksite saamist ka iga raku doonori kindlakstegema. Kuigi autorite lisatud sildid olid iga uuringu metaandmetest kättesaadavad, katsetati lõputöös ka demultipleksimise meetodit. Vajadus demultipleksimise töövoo järele võib tekkida tulevikus, kui uuringutega pole vajalikke metaandmeid kaasapandud. Demultipleksimine sooritati ainult Randolph_2021 andmetel.

Demultipleksimistarkvaraks valiti souporcell [45]. souporcell sooritab igas üksikrakkude partiis doonorite klasterdamise ja iga raku klastrisse määramise. Programm valiti kahel põhjusel. Esiteks, teine suur üksikrakuandmete töövoog - yascp [46] - kasutab demultipleksimiseks just souporcelli. yascp oli lõputöö analüüside ajal alles arendusjärgus, mistõttu ei saanud seda tervenisti rakendada. Teiseks, Randolph jt kasutasid rakkude demultipleksimiseks samuti souporcelli.

Lõputöös kasutati souporcelli singularity konteineri kaudu nagu programmi kasutusjuhises [47] soovitatud. Programm käivitati sisendandmete konfiguratsioonis, mis võimaldas osad ebavajalikud ja ajakulukad sammud vahele jätta. souporcellile anti sisendiks Cell Rangeri loodud üksikrakumaatriks ja joondusandmed (.bam fail), viitegenoom (GRCh38), GLIMPSE töövooga [29] imputeeritud ja filtreeritud (MAF > 1%) Randolph_2021 genotüübid, levinud variantide fail (AF $\geq$ 2%) 1000 Genoomi Projektist, oodatav klastrite arv (6, sest igas partiis olid rakud kuuelt indiviidilt). Lisaks pidi sisendfailis olema loetletud, millise doonori rakud igas partiis olid. See info võeti metaandmetest, kusjuures kahe partii puhul ei olnud doonorite nimekiri täielik, mistõttu neid partiisid ei demultipleksitud.

Teoreetiliselt poleks partii doonorite nimekirja vaja täpsustada. souporcelli peaks saama käivitada ka nii, et määrata iga partii puhul doonoriteks kõik indiviidid. Siis iga partii kohta koostab programm nii palju klastreid kui oli indiviide terves uuringus, aga partiis mitteesinenud indiviididele ei määrata ideaalselt ühtegi rakku. Lõputöös prooviti ka seda varianti, kuid see nõudis oluliselt rohkem aega ja mälu, mistõttu otsustati siiski metaandmetest partii doonorid otsida.

souporcell sooritab küll rakkude jagamist klastrite vahel, kuid programm jätab otsustamata, milline klaster millisele indiviidile vastab. Selle olulise sammu tegemiseks loodi lõputöös skript, mis määrab igale klastrile vastava indiviidi. Idee on analoogne souporcelli autori kommentaarile [48] (kirjutatud 9. augustil 2022), kui temalt küsiti soovitusi klastri ja indiviidi kokkuviimiseks.

Lõputöös loodud skript otsib igale klastrile kõige sarnasema genotüübiga doonori. souporcell väljastab iga klastri genotüübifaili. Skript arvutab alleelidoosid igas klastri genotüübis ja igas partii doonori genotüübis. Seejärel arvutatab programm iga klastri puhul alleelidooside vahe kõigi partii doonoritega. Vähima vahega doonor on genotüübilt kõige sarnasem klastriga, ehk klaster vastab just sellele doonorile. Protsess on illustreeritud joonisel 4.

Klastrite alleelidoosid			
variant	klaster 0	klaster 1	klaster 2
var0	0	2	2
var1	0	2	2
var2	1	2	0
var3	0	0	1
var4	2	2	0
var5	2	2	0

Doonorite alleelidoosid			
variant	doonor 0	doonor 1	doonor 2
var0	2	0	2
var1	1	0	2
var2	0	0	1
var3	1	0	0
var4	0	2	2
var5	0	1	2

Iga klastri puhul arvuta klastri dooside vahe kõigi doonoritega.
Valem: abs(*klastri_doos – doonori_doos*)

Klastri 0 vahed doonoritega			
variant	vahe 0	vahe 1	vahe 2
var0	2	0	2
var1	1	0	2
var2	1	1	0
var3	1	0	0
var4	2	0	0
var5	2	1	0
summa	9	2	4

Klastri 1 vahed doonoritega			
variant	vahe 0	vahe 1	vahe 2
var0	0	2	0
var1	1	2	0
var2	2	2	1
var3	1	0	0
var4	2	0	0
var5	2	1	0
summa	8	7	1

Klastri 2 vahed doonoritega			
variant	vahe 0	vahe 1	vahe 2
var0	0	2	0
var1	1	2	0
var2	0	0	1
var3	0	1	1
var4	0	2	2
var5	0	1	2
summa	1	8	6

Klastrile vastav doonor on vähima dooside vahega doonor (värvitud oranžiks).

Joonis 4. Väikese näitega illustreeritud loogika klastri ja doonori kokkuviimiseks. Olgu partiis kolm doonorit. souporcell väljastas klastrite genotüübid. Doonorite genotüübid on teada. Igale klastrile ja partii doonorile arvutati alleelidoosid. Seejärel arvutati iga klastri dooside vahe iga doonoriga. Vahed summeeriti doonorite kaupa ja vähima vahega doonor määrati klastri doonoriks.

Nii prooviti lõputöös üksikrakuandmeid demultipleksida. Joonistel 8 ja 9 on kujutatud, kui enesekindel demultipleksimise skript oma otsustes oli. Skripti määratud indiviide võrreldi Randolph_2021 metaandmetes määratud indiviididega. Kokkulangevus oli 100%.

3.3.3 Pseudo-hulkrakumaatriksite koostamine

Üksikrakuandmed joondati Cell Rangeri ja kallistoga, aga lõputöös ei sooritatud rakutüüpide ega indiviidide tuvastamist (v.a indiviidide tuvastamine testimise eesmärgil Randolph_2021 andmetes). Üksikrakumaatriksitele lisati rakutüüpide ja indiviidide sildid iga uuringu metaandmetest. Kuna uuringutes olid metaandmed ainult kvaliteetsete rak-

kude (st maatriksist olid ebakorrektsed tilgad eemaldatud) kohta, ei sooritatud lõputöös ka täiendavat rakkude filtreerimist. Kui palju rakke igast uuringust kasutati, on toodud tabelis 1 tulbas "Rakke"sulgudesse kirjutatuna.

Järgmiseks üksikrakumaatriksid normaliseeriti. Erinevaid üksikraku tasemel normaliseerimismeetodeid võrdlesid Ahlmann-Eltze ja Huber [49]. Nad leidsid, et raku loenduste skaleerimine kümne tuhande lugemini raku kohta ning seejärel maatriksi logaritmimine oli üks parimatest normaliseerimismeetoditest. Võrreldes teiste statistiliste meetoditega nõudis see ka kõige vähem arvutusjõudu.

Kuigi Ahlmann-Eltze ja Huber soovitasid loendused skaleerida miljoni asemel kümne tuhandeni, ei oma skaleerimise piir lõputöös tähtsust. Enne eQTL-analüüsi sooritamist normaliseeritakse pseudo-hulkrakumaatriksid meetodiga, mille tulemus sõltub ainult loenduste järgi koostatud indiviidide järjestusest. Kuna skaleerimise piir ei muuda indiviidide järjekorda, ei muuda see ka normaliseerimise tulemust.

Lõputöös normaliseeriti üksikrakumaatriksid rakutasemel skaleerimisega ja logaritmimisega, järgides kallisto koodivaramu õpetust [50]. Esmalt skaleeriti iga raku loendused miljoni lugemini raku kohta (*counts-per-million*, CPM). Seejärel võeti maatriksist naturaallogaritm koos pseudoloendustega (*pseudocounts*): $\ln(X + 1)$, kus X on üksikrakumaatriks. Nii skaleerimiseks kui logaritmimiseks kasutati scanpy paketi [51] (v1.9.3) funktsioone `scanpy.pp.normalize_total` ja `scanpy.pp.log1p`.

Pseudo-hulkrakumaatriksid koostati vaid rakutüüpidest, kus oli vähemalt 1000 rakku. Esimesed testid eQTL-ide tuvastamiseks näitasid, et väiksema rakkude arvuga rakutüüpides ei ole statistilist võimsust tuvastada eQTL-e. Rakkude agregeerimise meetod valiti Cuomo jt [24] parimate tulemuste järgi. Rakud grupeeriti indiviidi ja rakutüübi kaupa ning rakkude loendustest võeti keskmine. Rakkude partiid polnud vaja grupeerimisel arvestada, kuna kasutatud uuringutes olid iga indiviidi rakud ainult ühes partiis. Lõputöös kasutatud agregeerimismeetod on illustreeritud ka joonisel 5.

Viimane pseudo-hulkrakumaatriksi töötlemissamm oli geenide filtreerimine. Kuna rakkudes on ekspresseeritud ainult osa geenidest korraga, tuleb enne eQTL-analüüsi eemaldada mitteekspresseeritud geenid. Sellistele geenidele pole võimalik eQTL-e määrata ning nende alleshoidmine suurendab statistiliste testide arvu, mis omakorda teeb raskemaks ka teiste eQTL-ide nägemise. Lõputöös rakendatud filter oli järgmine: geen jäeti alles, kui selle loenduste arv oli üle nulli vähemalt 60% indiviididest.

3.4 Ekspressiooni kvantitatiivsete tunnuset lookuste analüüs

Cis-eQTL-analüüsiks kasutati lõputöös töövoogu eQTL-Catalogue/qtlmap [30]. Töövoog võtab sisendiks indiviidide genotüübid ja kvantitatiivse fenotüübi maatriksi ning leiab seosed nende kahe vahel. Täpsemalt, töövoog kasutab geneetiliste variantide nominaalsete p-väärtuste arvutamiseks tööriista fastQTL [52] ning permutatsioonide p-väärtuste arvutamiseks funktsiooni `QTLtools cis` paketist QTLtools [53]. Lisaks arvutab `QTLtools`

Üksikrakumaatriks

Rakk	Rakutüüp	Indiviid	Geen 0	Geen 1	Geen 2	Geen 3
r0	T	i0	1	0	0	15
r1	T	i0	0	0	0	29
r2	T	i1	5	0	0	17
r3	T	i1	17	0	0	3
r4	T	i1	2	0	1	40
r5	T	i2	5	0	2	10
r6	B	i0	0	0	4	8
r7	B	i1	1	0	9	14
r8	B	i2	1	0	3	0
r9	B	i2	0	0	8	6

Grupeeri rakutüübi ja indiviidi kaupa ning võta keskmine.
Salvesta rakutüübid eraldi tabelitena.

T-rakkude pseudo-hulkrakumaatriks

Indiviid	Geen 0	Geen 1	Geen 2	Geen 3
i0	0.5	0	0	22
i1	8	0	0.3	20
i2	5	0	2	10

B-rakkude pseudo-hulkrakumaatriks

Indiviid	Geen 0	Geen 1	Geen 2	Geen 3
i0	0	0	4	8
i1	1	0	9	14
i2	0.5	0	5.5	3

Joonis 5. Demultipleksitud üksikrakkude loendusmaatriksitest koostati pseudo-hulkrakumaatriksid. Üksikrakumaatriks grupeeriti rakutüübi ja indiviidi tasemel ning iga geeni loendustest võeti keskmine üle grupi rakkude. Iga rakutüübi pseudo-hulkrakumaatriks salvestati eraldi tabelina.

`cis` korrigeeritud beeta p-väärtused, kasutades sobitatud beetajaotust [52]. qtlmap väljastab iga testitud geeni kohta statistiliselt oluliseima juhtvariandi, selle variandi nominaalse, permutatsiooni ja beeta p-väärtused.

qtlmap eeldab, et andmetele sobitatud mudeli jäägid on normaaljaotusega. Andmestikes võivad aga esineda erindid, mis seda eeldust rikuvad. Seepärast rakendati pseudo-hulkrakumaatriksi geenidele veel järjestuspõhine pöördnormaal-transformatsioon (*rank-based inverse normal transformation*), tagamaks mudeli jääkide normaaljaotuse. Funktsioon selleks transformatsiooniks kohandati koodivaramust [54]. Enne transformeerimist segati pseudo-hulkrakumaatriksi read (indiviidid), et võrdsed loenduste väärtused esineksid juhuslikus järjekorras.

qtlmap (v23.02.1) laeti alla GitHubist [30]. Genotüübifailidena kasutati imputeeritud (Randolph_2021 genotüübid GLIMPSE töövooga, ülejäänud genimpute töövooga) ja filtreeritud (INFO > 0.4, MAF > 0.01) genotüüpe. Ülejäänud sisendandmed loodi töövoo näidiste järgi. Käsurea parameetrid töövoo jooksutamiseks olid järgnevad: `vcf_has_R2_field FALSE`, `run_nominal FALSE`, `run_permutation TRUE`, `run_susie FALSE`, `cis_window 1000000`. Viimane parameeter määras, kui suurelt alalt geeni ümbert eQTL-e otsiti. Ala suuruseks valiti 1 Mbp, kuna see on standardne *cis*-eQTL-ide tuvastamise regioon.

4 Tulemused

Peatükis kantakse ette kõik lõputöö tulemused. Näidatakse genotüüpide imputeerimise täpsust ja kui hästi töötas demultipleksimise skript. Seejärel tuuakse välja, mitmest rakust koostati ja mitu geeni sisaldasid pseudo-hulkrakumaatriksid. Viimasena kirjutatakse eQTL-analüüsi tulemustest, uuritakse rakutüüpides rakkude arvu mõju tuvastatud statistiliselt oluliste eQTL-ide arvule ning vaadatakse korrelatsiooni kasutatud joondusprogrammide maatriksite põhjal arvutatud eQTL-ide p-väärtuste vahel.

4.1 Genotüüpide imputeerimine

Lõputöö osana loodi madala katvusega täisgenoomi sekveneerimisandmete imputeerimise töövoog [29], mis kasutab imputeerimiseks tarkvara GLIMPSE [35]. Lõputöö eQTL-analüüsiks imputeeriti Randolph_2021 uuringus kogutud genoomisekveneerimisandmed. Imputeeritud genotüübifailis oli kokku 71 mln geneetilist varianti, neist 62 mln SNP-i ja 9 mln *indel*-i. Töövoog väljastab ka MAF-i ja imputeerimise kvaliteedi järgi filtreeritud (MAF > 0.01 ja INFO > 0.4) genotüübid. Filtreeritud failis oli 18 mln geneetilist varianti, neist 15 mln SNP-i ja 3 mln *indel*-i.

Lisaks rakendati töövoogu Kumasaka jt [28] *ATAC-seq* andmetele, testimaks imputeerimise täpsust. Testi tulemused on kokkuvõetud joonisega 6. Imputeeritud genotüübid korreleerusid tegelike genotüüpidega levinud variantide kohal äärmiselt tugevalt. Vähemlevinud variantide puhul korrelatsioon veidi langes, kuid keskmine korrelatsioon püsis siiski üle 0.9. Väga sarnast pilti (toodud joonisel 7) nägid ka GLIMPSE meetodi autorid kui programmi testisid [35].

4.2 Üksikraku RNA sekveneerimine

4.2.1 Demultipleksimine

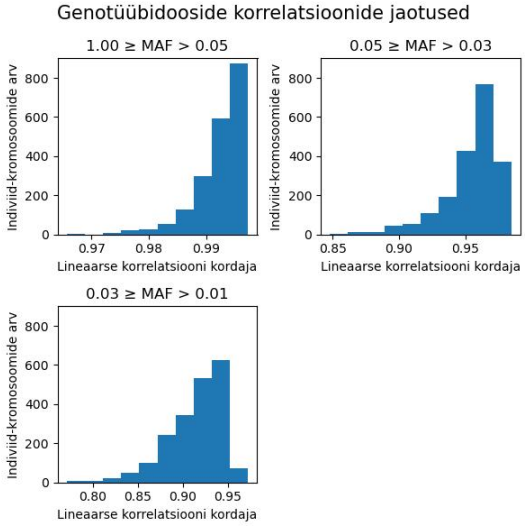
Demultipleksimine sooritati testimise eesmärgil ning ainult Randolph_2021 andmetel. Raku ja indiviidi kokkuviimiseks loodi skript, mis võtab sisendiks souporcelli klastrite genotüübid ja partii indiviidide genotüübid. Skript arvutas iga klastri puhul vahe iga indiviidiga ning määras klastri indiviidiks vähima vahega doonori (vt ka joonis 4). Joonisel 8 on illustreeritud ühes partiis arvutatud vahed ning joonis 9 näitab, kui suure kindlusega klastrid indiviididega kokku viidi. Skripti määratud indiviide võrreldi metaandmetes toodud indiviididega ja leiti, et kõigi rakkude indiviidid kattusid.

4.2.2 Pseudo-hulkrakumaatriksid

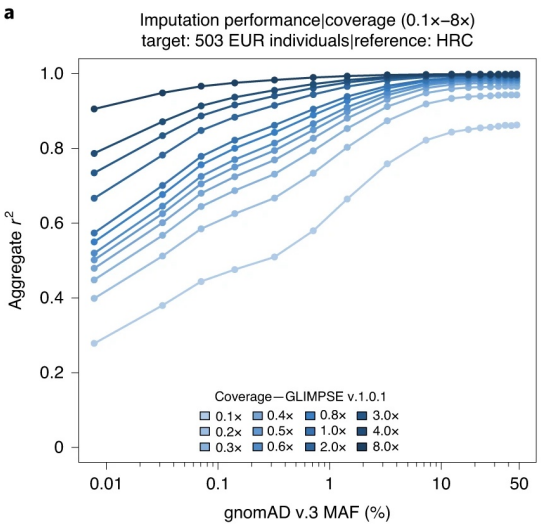
Pseudo-hulkrakumaatriksid koostati vaid rakutüüpides, milles oli vähemalt 1000 rakku. Sellised rakutüübid koos rakkude arvudega on toodud joonisel 10 vasakpoolses tulbas.

T-rakud olid rakkude arvu poolest kõige suurem rakutüüp. Ka B-rakud olid esindatud igas kohordis. Loomulikud tapjarakud (*natural killer cells*, NK rakud) moodustavad suure grupi OneK1K_2022 kohordis, kuid ülejäänutes on nende osakaal väiksem. Sama võib öelda monotsüütide (joonisel märgitud „cM” - *classical monocytes*) kohta Perez_2022 andmetes.

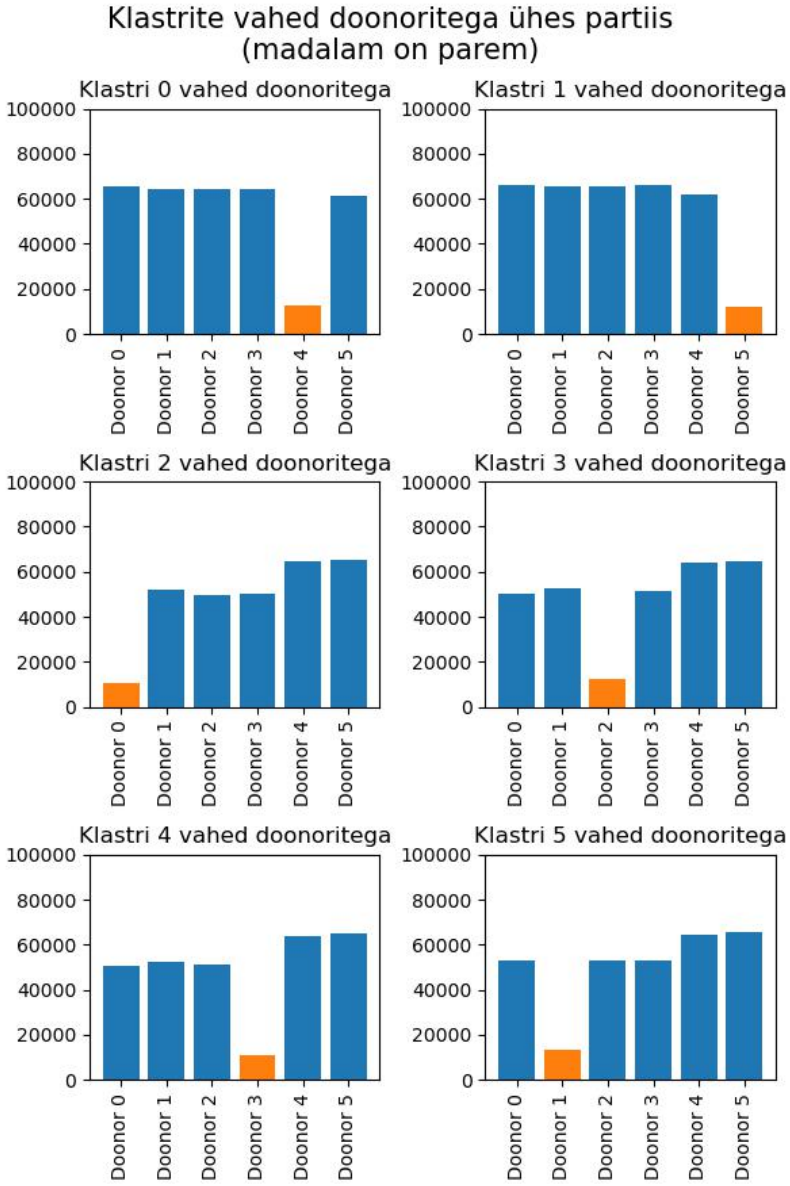
Joondajate väljastatud üksikrakumaatriksites olid loendused toodud umbes 60000 geeni kohta. Kuna rakutüübis on korraga ekspresseeritud vaid osa geenidest, sooritati igas pseudo-hulkrakumaatriksis madala ekspressiooniga geenide väljafiltreerimine. Geen jäeti alles, kui see oli ekspresseeritud (loendus üle nulli) vähemalt 60% indiviididest pseudo-hulkrakumaatriksis. Maatriksitesse allesjäänud geenide arvud on toodud joonisel 10 parempoolses tulbas. Enamlevinud rakutüüpides jäi alles ka rohkem geene.



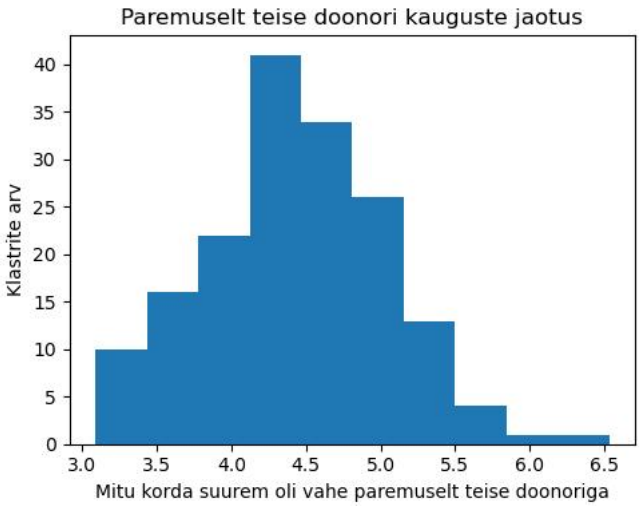
Joonis 6. *ATAC-seq* andmete imputeerimine oli väga täpne. Imputeeritud genotüübid jagati kolmeks MAF-i järgi. Iga indiviidi igas kromosoomis arvutati alleelidoosid. Seejärel leiti nende lineaarne korrelatsioon tegelike alleelidoosidega. Näidatud on korrelatsioonide jaotused. Levinud variantide (MAF > 0.05) puhul oli korrelatsioon äärmiselt tugev: iga kromosoomi puhul üle 0.95, kusjuures enamikul üle 0.99. Alla 0.8 oli korrelatsioon vaid üksikutel juhtudel vähemlevinud (0.03 ≥ MAF > 0.01) variantide puhul.



Joonis 7. GLIMPSE programmi autorid leidsid samuti, et 4X katvusega genotüüpide imputeerimisel on GLIMPSE 1% ja sagedamate variantide puhul väga täpne [35, kohandatud].



Joonis 8. Klastritele indiviidi määramine ühes partiis. Iga klastri puhul on näha, et leidub täpselt üks indiviid (doonor), kellega on alleelidooside vahe palju väiksem võrreldes ülejäänud doonoritega. Joonistel on see doonor värvitud oranžiks. Kõigi teiste doonoritega on alleelidooside vahed kordades suuremad.



Joonis 9. Paremuselt teine doonor oli alati vähemalt 3 korda suurema vahega kui parim doonor. Enamike klastrite puhul oli vahe 4 - 5-kordne.

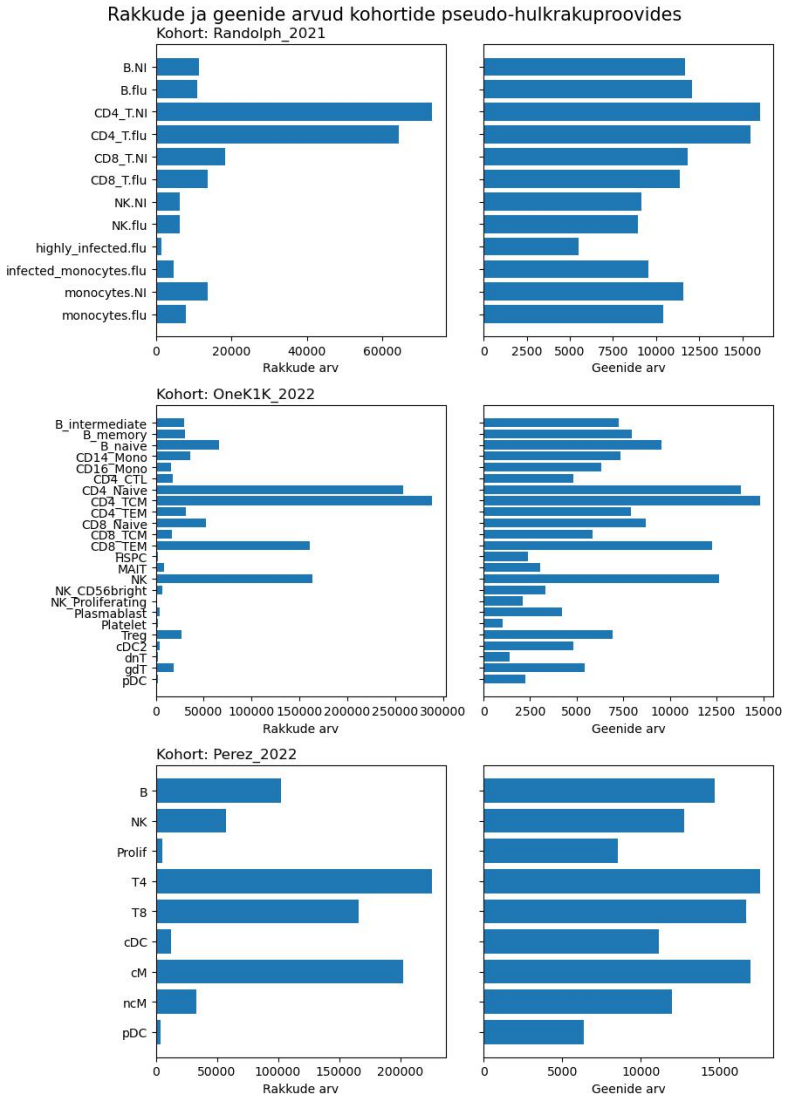
4.3 Ekspressiooni kvantitatiivse tunnuse lookuste analüüs

Üks meetod nägemaks, kas rakutüübis leiti eQTL-e, on vaadata rakutüübis qtlmapi arvutatud beeta p-väärtuste jaotust. Andmestikus, kus eQTL-e ei ole, peaks p-väärtused olema ühtlaselt jaotunud, sest iga juhtvariant on juhusliku p-väärtusega. Seevastu andmestikus, kus esineb eQTL-e, peaks väikseid p-väärtusi olema teistest rohkem, sest tegelikud eQTL-id on madala p-väärtusega ja kõik ülejäänud juhtvariandid juhusliku p-väärtusega.

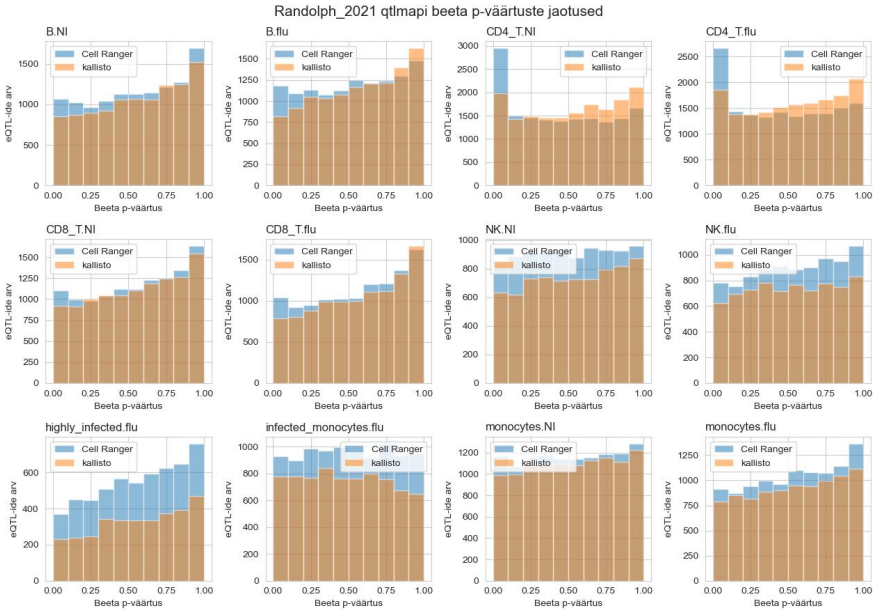
Joonistel 11 - 13 on toodud qtlmapi beeta p-väärtused. Igas kohordis leidus rakutüüp, kus oli eQTL-signaal, kusjuures kõik signaalid olid Cell Rangeriga tugevamad kui kallistoga. T-rakkudes oli signaal kõigis kohortides, B-rakkudes, monotsüütides ja NK rakkudes sõltuvalt kohordist.

Joonistel 14 - 16 on toodud tuvastatud eQTL-ide arvud. Enim eQTL-e leiti T-rakkudes, aga sõltuvalt kohordist olid esindatud ka B-rakud, monotsüüdid ja NK rakud. Suurim statistiliselt oluliste eQTL-ide osakaal oli OneK1K_2022 andmestikus. Lisaks on näha, et igas kohordis leiti Cell Rangeriga joondatud andmetest oluliselt rohkem eQTL-e kui kallistoga.

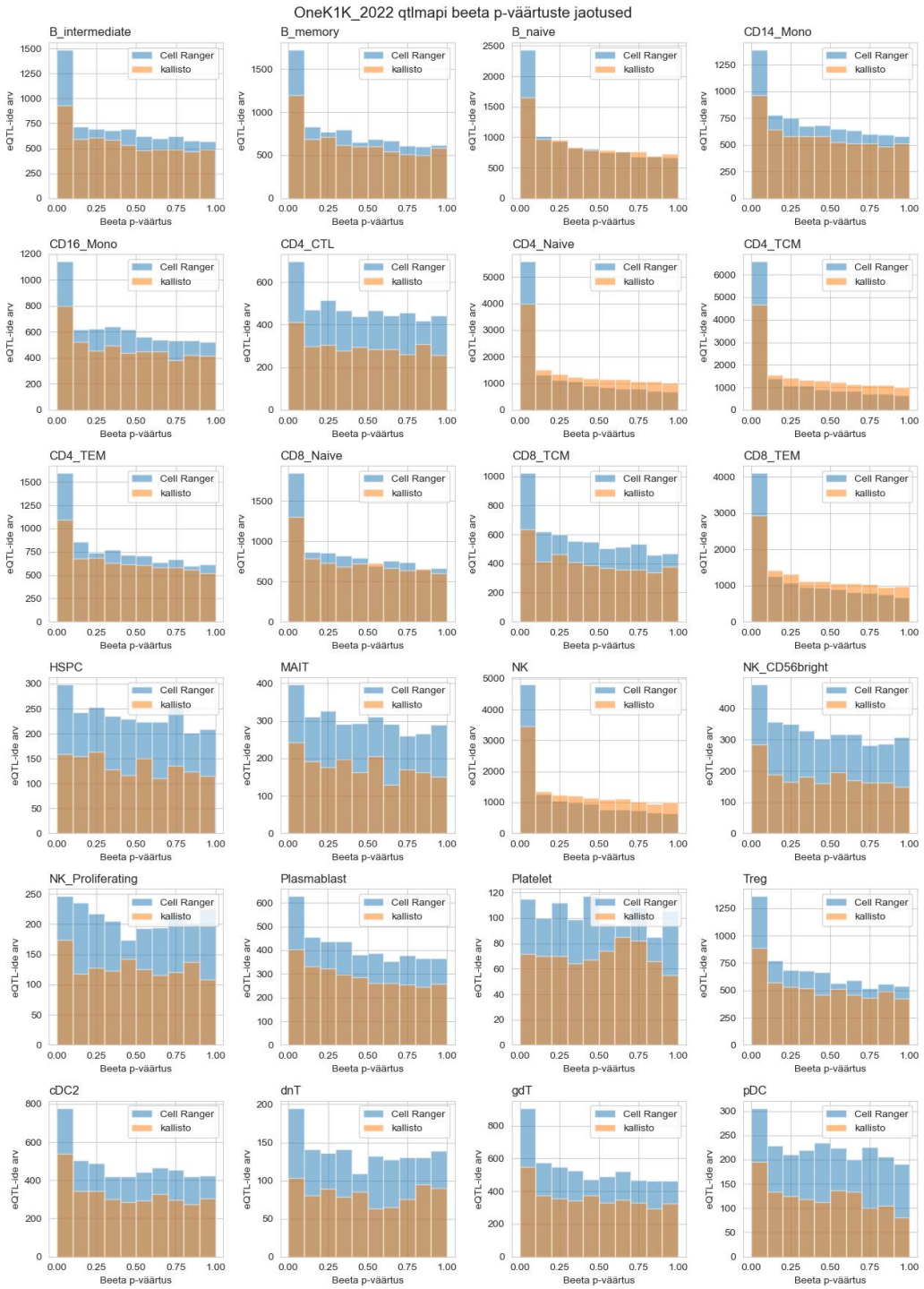
Järgnevalt prooviti võrrelda lõputöö tulemusi andmete autorite tulemustega. Kuna uuringutes kasutati erinevaid eQTL-ide *cis*-akna suurusi, statistilise olulisuse piire ja rakutüüpide kokkusegamisviise, on võrdluste loomine peaaegu võimatu. Ainult Randolphi jt [9] tööst leiti tabel, kus olid toodud osade rakutüüpide kohta eQTL-ide leiud. Tuleb tähele panna, et Randolph jt kasutasid väiksemat *cis*-akent (100 kbp). Seepärast ei ole tabelis 3 toodud arvud üks ühele võrreldavad.



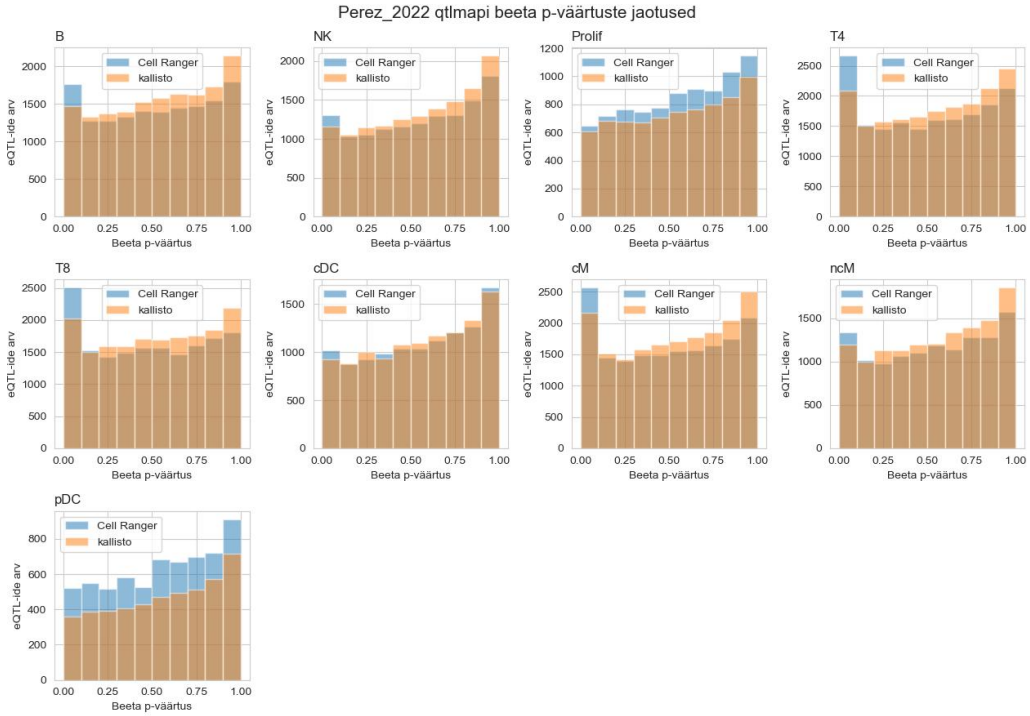
Joonis 10. Üksikraku RNA sekveneerimine on kallis viis koguda T-rakkude andmeid. Joonisel on toodud kõik rakutüübid, millest pesudo-hulkrakuproovid tehti ja millega eQTL-analüüs sooritati. Randolph_2021 rakutüübinimedes viitavad järelliited *.flu* ja *.NI* vastavalt sellele, kas tegu oli gripiviirusega nakatatud koeprooviga või mitte. Geenide filtreerimise eesmärk oli eemaldada väheekspresseerunud geenid, millele eQTL-i tõenäoliselt tuvastada poleks võimalik. Joondaja väljastatud üksikrakumaatriksites oli esialgu umbes 60000 geeni. Filtri lävendi valikul lähtuti sellest, et suurearvulistes rakutüüpides võiks alles jääda umbes 10000 geeni.



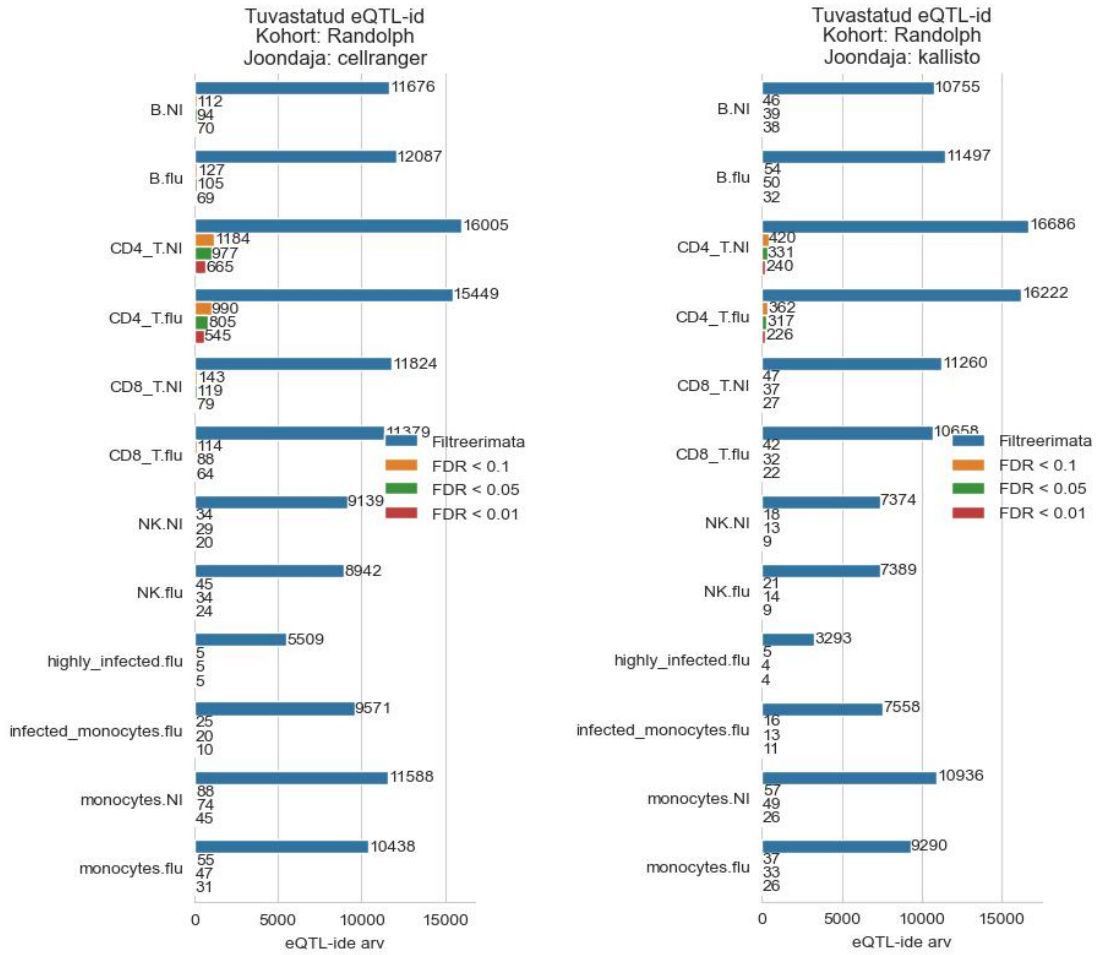
Joonis 11. Randolph_2021 kohordis olid kõige nõrgemad eQTL-signaalid. Esindatud olid ainult T-rakud.



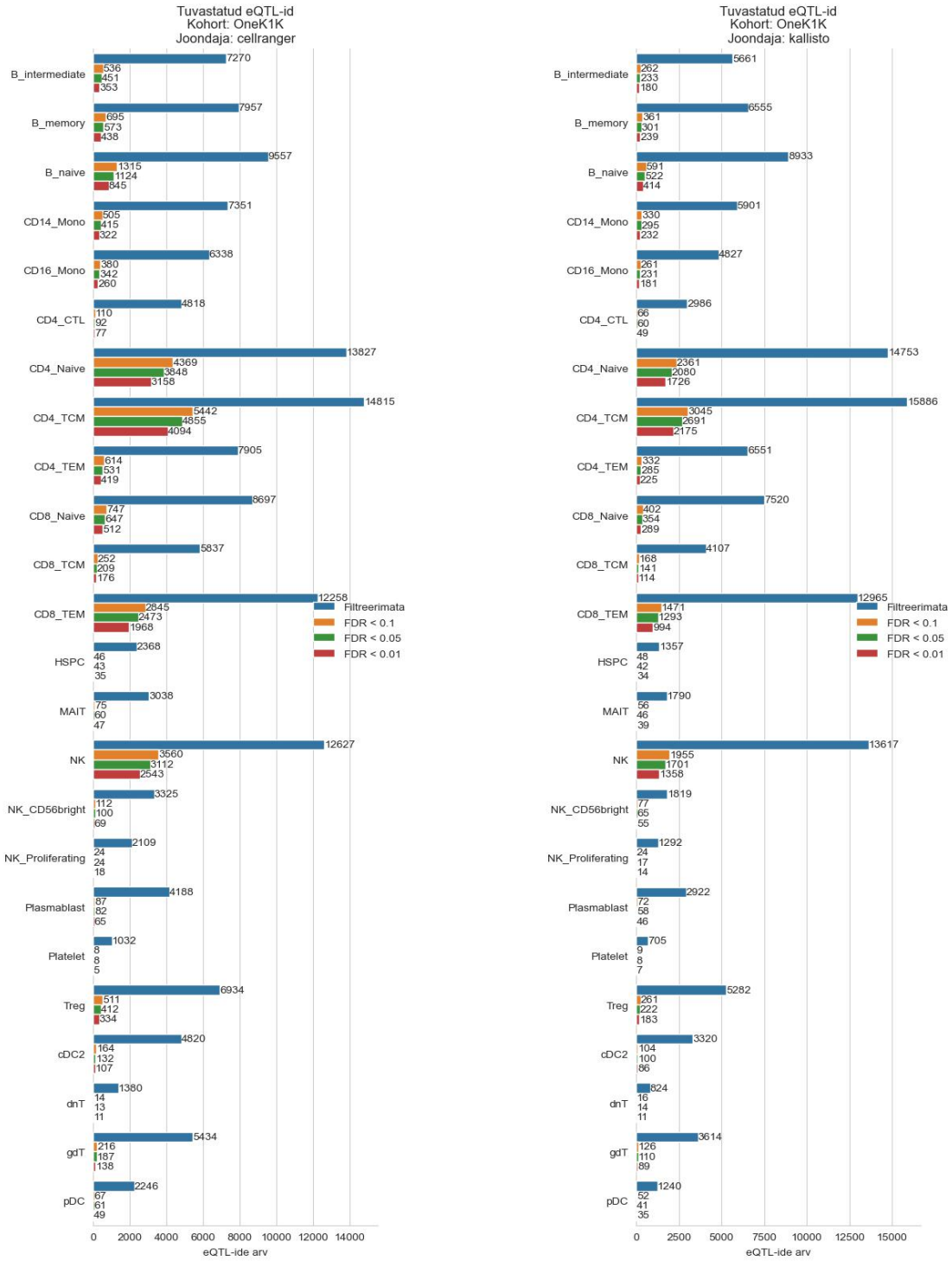
Joonis 12. OneK1K_2022 kohordis olid kõige tugevamad eQTL-signaalid. Näiline signaal oli nii T-, B-, kui ka NK rakkudes ja monotsüütides.



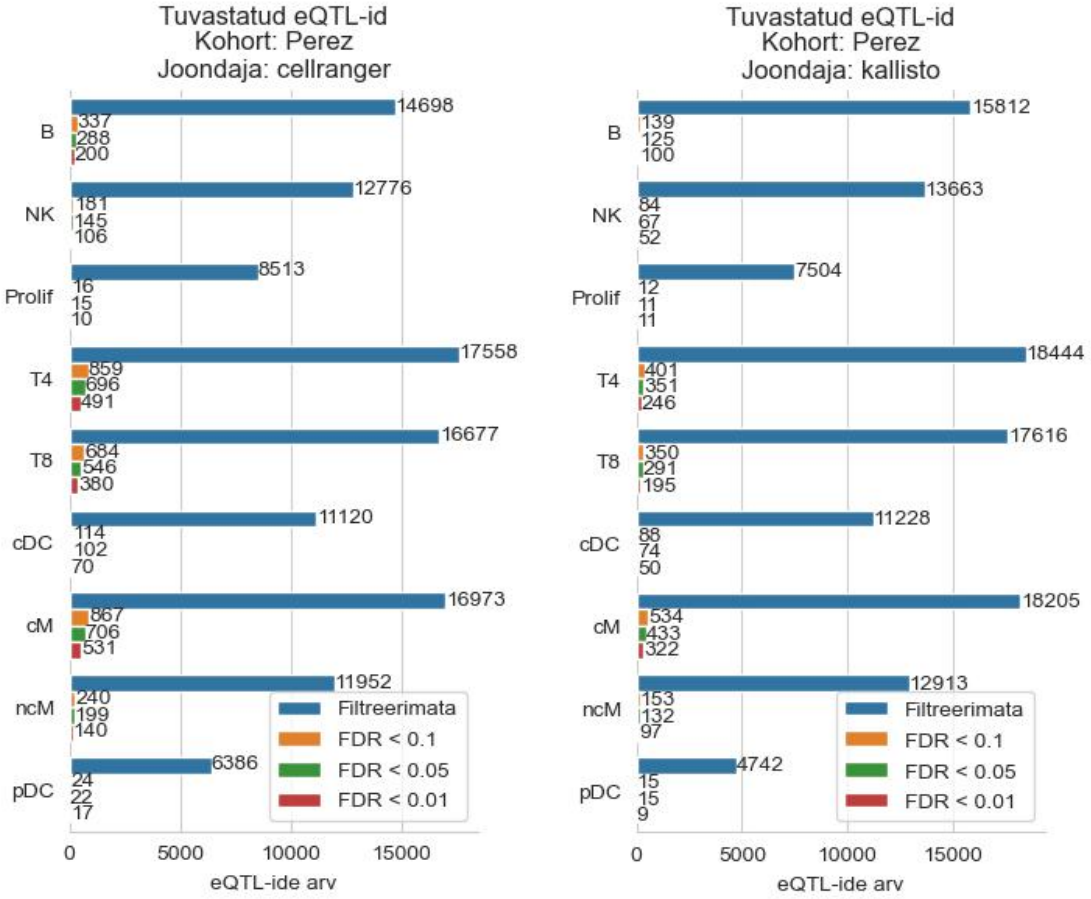
Joonis 13. Perez_2022 kohort oli pigem signaalidevaene. Näiline signaal oli T-rakkudes ja monotsüütides.



Joonis 14. Randolph_2021 kohordis oli signaal vaid T-rakkudes: Cell Rangeriga tuli tugevamalt välja, kallistoga nõrgemalt. Järelliited *.NI* ja *.flu* rakutüüpide nimedes viitavad vastavalt sellele, kas tegu oli naiivsete või gripiviirusega nakatatud koeproovidega.



Joonis 15. OneK1K_2022 kohordis oli kõige suurem statistiliselt oluliste eQTL-ide osakaal. Cell Rangeriga olid kõik signaalid tugevamad kui kallistoga. Signaale on näha nii T- kui ka B-rakkudes, monotsüütides, NK rakkudes.



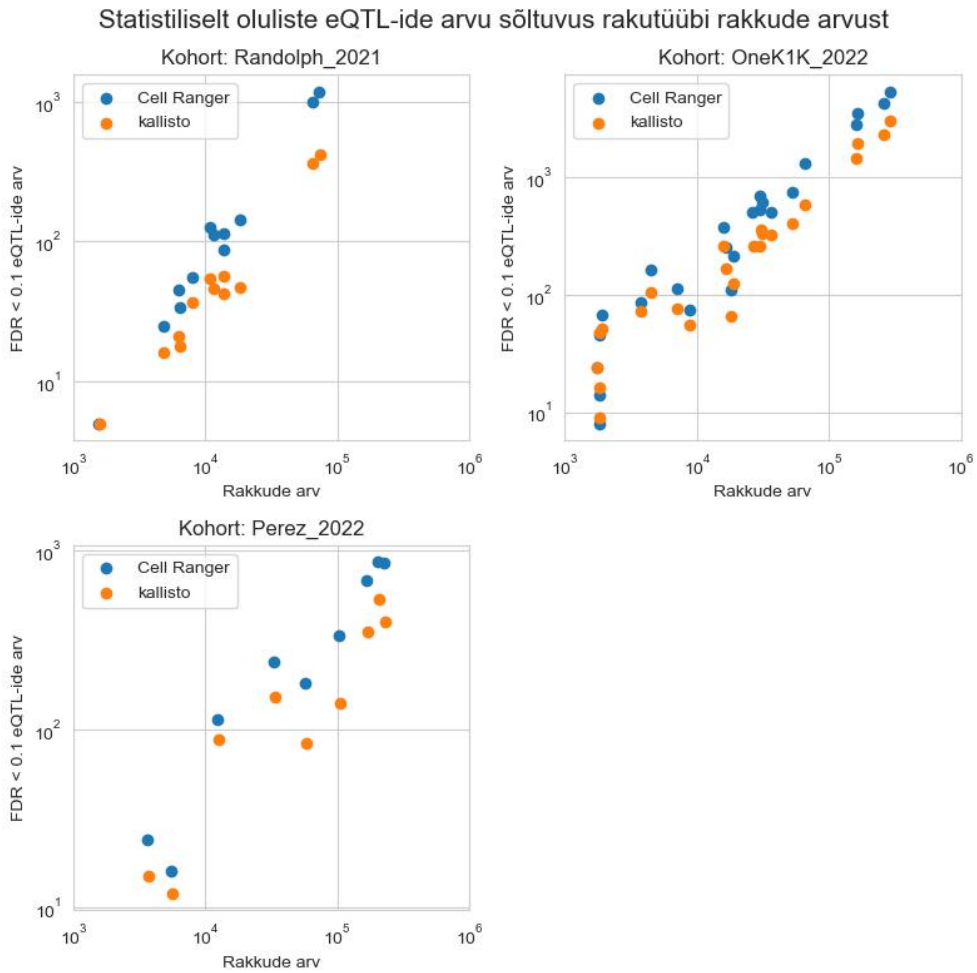
Joonis 16. Perez_2022 kohordis leiti eQTL-e monotsüütides, T- ja B-rakkudes.

Tabel 3. Ka Randolph jt leidsid T-rakkudes oluliselt rohkem eQTL-e kui mujal [9, tabel S11]. Tabelis toodud arvud pole siiski üks ühele võrreldavad, kuna töödes kasutati erinevaid *cis*-aknaid: lõputöös 1 Mbp, esialgses uuringus 100 kbp.

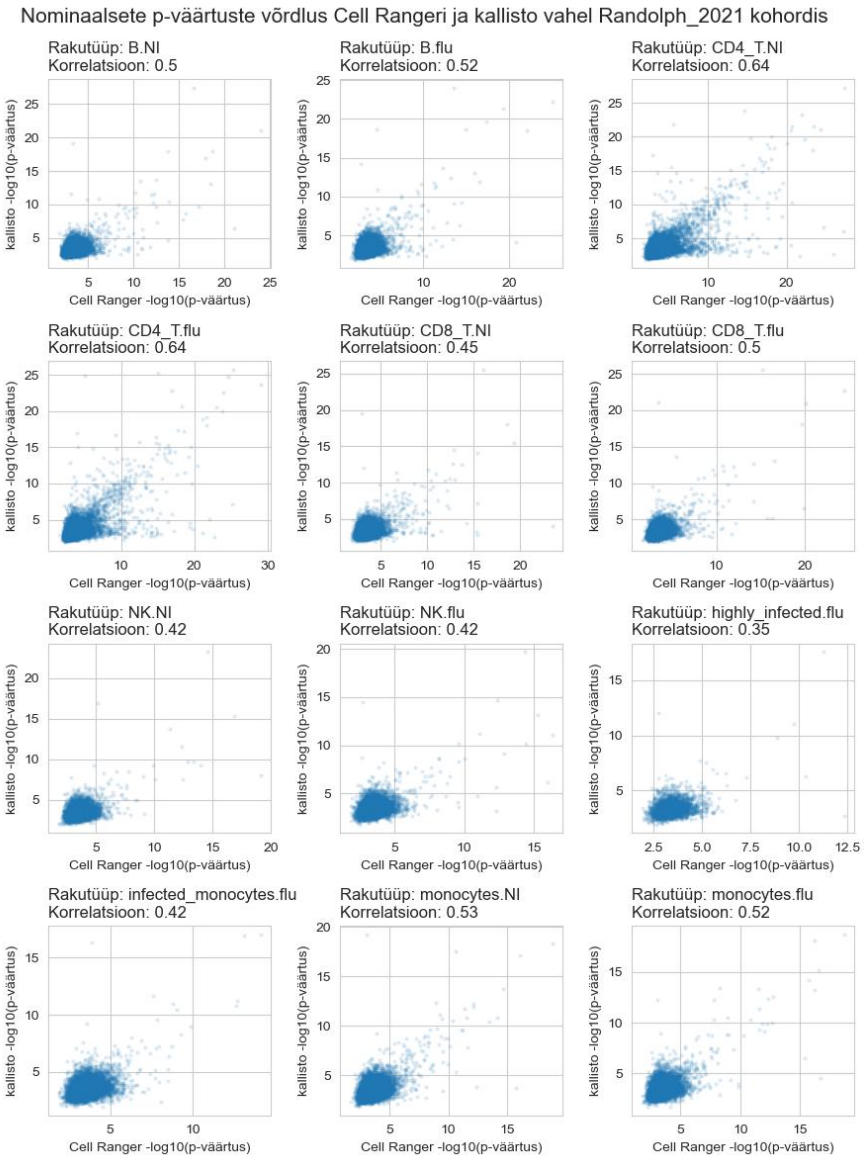
Rakutüüp	**Olulisi eQTL-e originaaltöös**	**Olulisi eQTL-e lõputöös**
CD4 T naiivsed	1377	1184
CD4 T nakatatud	1176	990
B naiivsed	152	112
B nakatatud	196	127
NK naiivsed	68	34
NK nakatatud	76	45
Monotsüüdid naiivsed	265	88
Monotsüüdid nakatatud	251	55
CD8 T naiivsed	204	143
CD8 T nakatatud	178	114

Edasi vaadati, kuidas statistiliselt oluliste eQTL-ide arv sõltus rakutüübis olnud rakkude arvust. Korrelatsioon on toodud joonisel 17. Rakutüübis tuvastatud statistiliselt oluliste eQTL-ide arv sõltus tugevalt rakutüübi rakkude arvust. Joondamisprogrammist ei sõltunud korrelatsiooni tugevus.

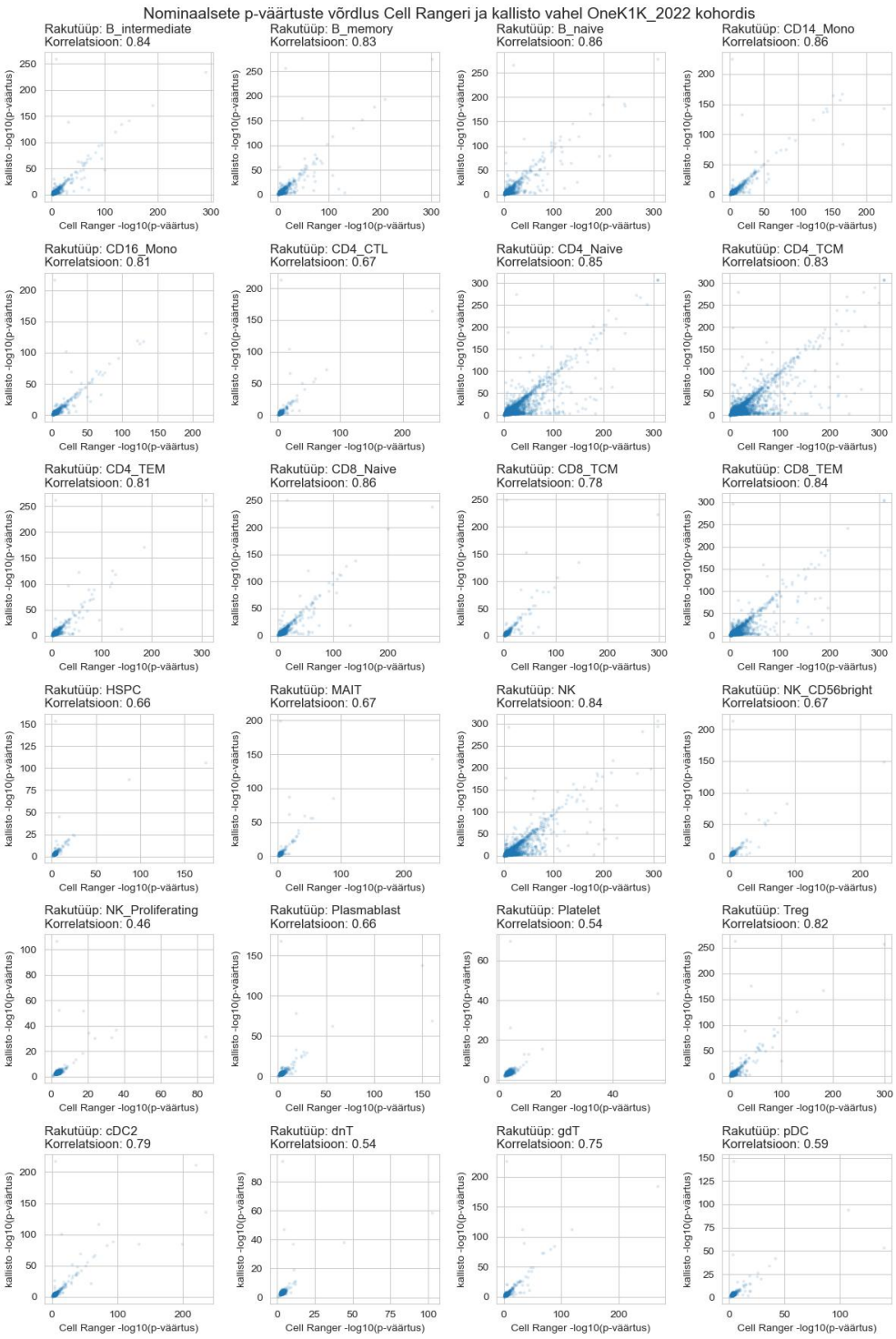
Järgmiseks uuriti Cell Rangeri ja kallisto joonduste põhjal arvutatud eQTL-ide nominaalsete p-väärtuste korrelatsioone, mis on toodud ka joonistel 18 - 20. Perez_2022 ja OneK1K_2022 kohortides oli korrelatsioon tugevam kui Randolph_2021 kohordis. Sellele vaatamata oli korrelatsioon üldiselt oodatust madalam. Tehes järeldusi kallisto autorite sooritatud võrdlusest Cell Rangeriga [43], ei tohiks p-väärtused nii palju erineda.



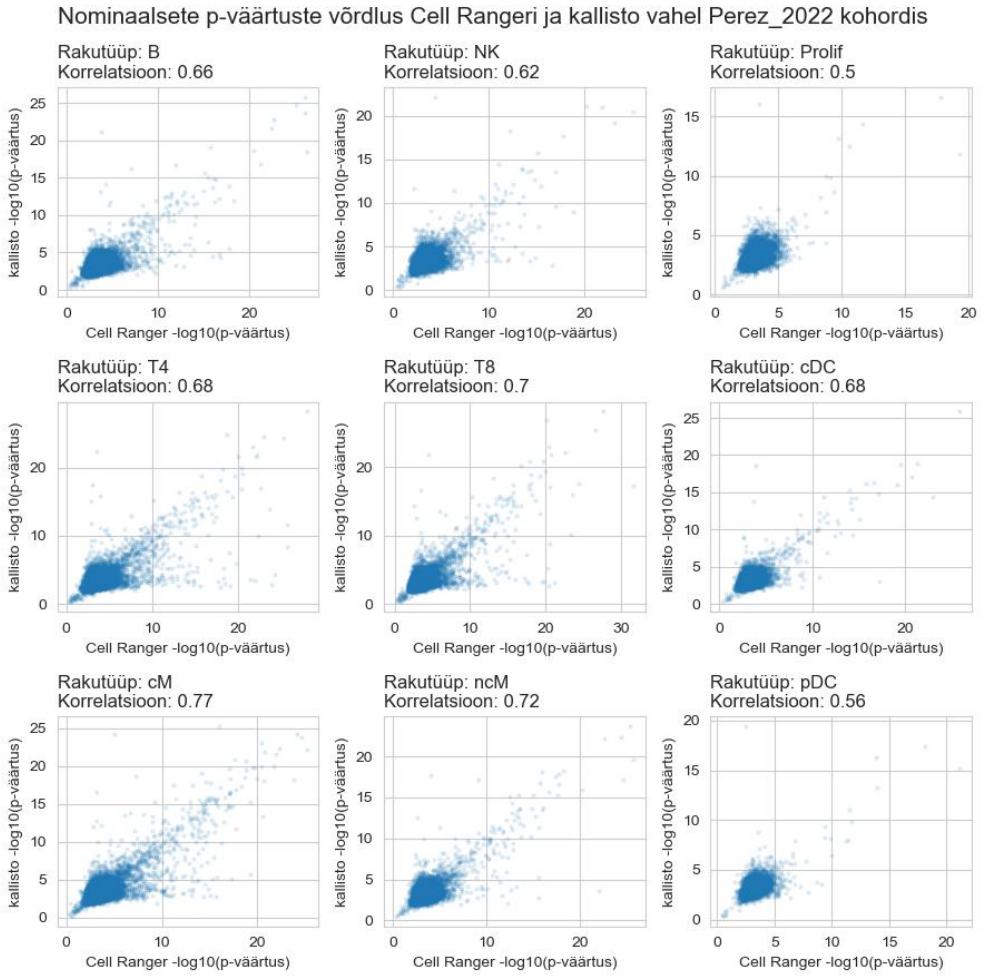
Joonis 17. Rakutüübis tuvastatud statistiliselt oluliste eQTL-ide arv on tugevas sõltuvuses rakutüübi rakkude arvuga. Kuigi kallistoga joondades ei leitud nii palju eQTL-e kui Cell Rangeriga, on korrelatsioon oluliste eQTL-ide arvu ja rakkude arvu vahel sama tugev.



Joonis 18. Randolph_2021 kohordis oli korrelatsioon madalapoolne: maksimaalselt 0.64 T-rakkudes.



Joonis 19. OneK1K_2022 kohordis oli korrelatsioon kõige tugevam: rohkemate rakkudega rakutüüpides 0.7 - 0.8.



Joonis 20. Perez_2022 kohordis oli korrelatsioon keskpärane: 0.5 - 0.7.

5 Arutelu

Peatükis tuuakse välja, mis võis tulemusi mõjutada ning millised on edasised uurimissuunad. Esiteks räägitakse imputeerimistöövoo rakendamisest lõputöös ja mujal. Teiseks arutletakse, mis võis põhjustada erienvusi joondajate tulemuste vahel. Kolmandaks võetakse kokku demultipleksimise testi tulemused ning viimasena tuuakse järeldused eQTL-analüüsi tulemustest.

5.1 Genotüüpide imputeerimine.

Genoomi imputeerimistöövoog [29] oli Kumasaka jt [28] *ATAC-seq* andmetel väga täpne. Lisaks on töövoogu rakendatud veel kahes projektis. Esiteks kasutati töövoogu Kerimovi jt [55] kvantitatiivsete tunnuste lookuste visualiseerimise töös BLUEPRINT andmestike imputeerimiseks. Leiti, et GLIMPSE töövooga imputeeritud paneelid sisaldasid rohkem LD-s olevaid variante, mis viitab kõrgemale imputeerimise kvaliteedile. Teiseks kasutas töövoogu Kuningas [56] oma lõputöös, kus GLIMPSE töövooga imputeeritud genotüüpe kasutati kromatiini avatuse QTL-analüüsis, mille tulemustel valideeriti tehisnärvivõrku Enformer.

Järgmine etapp valminud töövoo edasiarenduseks on lisada ka GLIMPSE2 algoritmiga [16] imputeerimise võimalus. Kuigi imputeerimine oli juba täpne, on GLIMPSE2 kiirem ülisuurte viitepaneelidega imputeerimisel. Suurendades viitepaneeli paraneb ka haruldaste variantide imputeerimise täpsus, nii et GLIMPSE2 on jätkusuutlikum. Lisaks, praegu on töövoog seadistatud imputeerima vaid autosoome, aga tulevikus lisatakse ka sugukromosoomide imputeerimise võimalus.

5.2 Üksikraku RNA sekveneerimine

5.2.1 Joondamine Cell Rangeri ja kallistoga

eQTL-analüüsi tulemused näitavad, et nf-core/scrnaseq töövoog joondajaga Cell Ranger oli parem üksikrakkude joondamismeetod kui kasutatud kallisto töövoog. Erinevused võivad tulla sellest, kuidas joondaja käitub vigaste ribakoodidega. RNA sekveneerimises kasutatavad ribakoodid sisaldavad sünteesil tekkinud vigu. Lisaks teeb sekveneerija vigu ribakoode sekveneerides. Tagajärjena peab joondamisalgoritm otsustama, kuidas käituda sekveneeritud ribakoodidega, mis ei vasta ühelegi oodatud ribakoodile. Ka mitmesse lookusesse joonduvate transkriptide puhul pole konsensust, milline on parim strateegia selliste lugemitega tegelemiseks, ning joondajad võivad neid erinevalt paigutada. Korrelatsioonid eQTL-idele arvutatud p-väärtuste vahel olid oodatust madalamad, mis viitab enam sellele, et kasutatud töövoogude koostatud lugemimaatriksid olid erinevad. Samas, kallisto programmi autorid näitasid, et kallisto ja Cell Ranger toodavad väga sarnaseid tulemusi [43].

Teiseks potentsiaalseks erinevuste põhjuseks on, et kallistoga joondamisel ei järgitud programmi juhiseid õigesti. On võimalik, et tehti viga sisendi loomisel või ei kasutatud andmetega sobivaid parameetreid.

Lõputöös ei uuritud joondajate tulemuste erinevust lähemalt ega vaadatud kattuvust joondajatega tuvastatud statistiliselt oluliste eQTL-ide vahel. Küll aga analüüsitakse kõik andmestikud tulevikus uuesti, kasutades valikut teistest nf-core/scrnaseq töövoo joondamisalgoritmidest: Alevin-Fry, STARSolo, kallisto + bustools. See annab terviklikuma pildi joondaja mõjust eQTL-analüüsi tulemuslikkusele. Lisaks jäi lõputööst välja ise rakutüüpide määramine, mis on vajalik metaanalüüsiks.

5.2.2 Demultipleksimine

Lõputöös loodi programm souporcelli klastritele vastavate indiviidide määramiseks. Visualiseeritud enesekindlustest (joonised 8 ja 9) on näha, et klastrile vastav indiviid erines selgelt ülejäänud partii indiviididest ning programm oli oma otsustes kindel. Kontrollides programmi määratud indiviide Randolph_2021 metaandmete vastu, kattusid rakkudele määratud indiviidid sajaprotsendiliselt.

Nii hästi ühtivad indiviidide annotatsioonid tulevad ilmselt sellest, et autorid kasutasid täpselt sama klasterdamise meetodit. Ainult klastri ja indiviidi kokkuviimise loogika oli lõputöös ja originaaltöös tõenäoliselt erinev, kuna artiklist ei leitud detailset kirjeldust, kuidas autorid selle probleemi lahendasid. Sellegipoolest on tulevikuks teada demultipleksimismeetod, juhtudeks kui vajalikke metaandmeid pole avalikustatud.

5.2.3 Pseudo-hulkrakumaatriksid

Perez_2022 andmetes oli oluliselt suurem monotsüütide osakaal kui teistes andmestikes. Põhjendust sellele otsisid ka andmete autorid [11]. Nad leidsid, et proovide seas oli palju luupuse patsiente ja haigusega kaasnebki monotsüütide arvu tõus. Lisaks mainiti, et erinevad rakkude kogumismeetodid uuringute vahel võivad olla kallutatud erinevate rakutüüpide suhtes.

Pseudo-hulkrakumaatriksite koostamisel ja geenide filtreerimisel kasutati lihtsaid meetodeid, mis on heaks aluseks edasistele uuringutele. Parim üksikrakuandmete eeltöötlusviis sõltub siiski andmestikust ja sooritatavast analüüsist. Lõputöös näidati, et üksikrakuandmete eQTL-analüüsiks ei pea rakendama keerulisi ja ajakulukaid statistilisi mudeleid, vaid tulemusi annavad ka lihtsad meetodid.

5.3 Ekspressiooni kvantitatiivsete tunnuste lookuste analüüs

Tugev korrelatsioon rakutüübi rakkude arvu ja leitud eQTL-ide arvu vahel näitab, et kindel viis rohkem eQTL-e leida on enam rakke sekveneerida. Korrelatsioon võib olla põhjendatud sellest, et üksikraku RNA sekveneerimisandmetes on rakkudevaheline

variatsioon väga kõrge. Seega, mida suurem on rakkude valim, seda paremini paistab rakutüübile omane geeniekspressioon välja. T-rakud on vereproovis arvulises ülekaalus, mistõttu juhuslikult rakke sekveneerides ongi enamik sekveneeritud rakkudest just T-rakud. Et ka teisi rakutüüpe saaks piisav arv sekveneeritud, tuleb sekveneerida rohkem proove, laboris proove puhastada või sooritada metaanalüüs.

Kahjuks ei saanud lõputöös tuua head võrdlust andmete autorite eQTL-analüüsi tulemustega. Leiti ainult Randolphi jt [9] avalikustatud eQTL-ide arvud ning ka need pole üks ühele võrreldavad lõputöö tulemustega, kuna kasutati erinevaid *cis*-aknaid. Siiski on võrdlusest näha, et mõlemas töös tuvastati T-rakkudes kordades rohkem eQTL-e kui muudes rakutüüpides.

6 Kokkuvõte

Lõputöös analüüsiti kolme üksikrakuandmestikku: Randolph_2021 [9], OneK1K_2022 [10], Perez_2022 [11]. Kõik andmestikud joondati kahe erineva programmiga ning koostati pseudo-hulkrakumaatriksid, millega viidi läbi eQTL-analüüs. Lisaks imputeeriti ühe uuringu madala katvusega sekveneeritud genotüübid ning katsetati üksikrakuandmete demultipleksimismeetodit.

Et Randolph_2021 andmeid eQTL-analüüsis kasutada, pidi esmalt imputeerima madala katvusega sekveneeritud genotüübid. Selleks kirjutati tarkvara GLIMPSE [35] kasutav töövoog [29], mida testiti ka *ATAC-seq* andmetel. Testis võrreldi imputeeritud genotüüpe täisgenoomi sekveneerimisel saadud genotüüpidega ja leiti, et genotüübid on väga tugevalt korreleeritud.

Randolph_2021 andmestikul testiti üksikrakkude demultipleksimismeetodit. Indiviidide klasterdamine sooritati samamoodi kui originaaltöös, kuid loodi uus skript klastri ja indiviidi kokkuviimiseks. Skript oli klastrile indiviidi määramisel väga kindel ning tulemus kattus ka originaaltöös määratud indiviididega.

Kõik kolm üksikrakuandmestikku töödeldi ühtemoodi, tänu millele olid eQTL-analüüsi tulemused omavahel võrreldavad. Cell Rangeri ja kallisto töövoogude võrdlus näitas, et nf-core paketi Cell Ranger on parem kui kasutatud kallisto töövoog. eQTL-e tuvastati enim OneK1K_2022 kohordis, kusjuures kõigis kohortides oli tugev seos tuvastatud eQTL-ide arvu ja rakutüübis olnud rakkude arvu vahel.

Lõputöö eesmärk täideti edukalt. Kõigis kolmes andmestikus leiti eQTL-signaalid, mis tähendab, et andmed sobivad kaasamiseks ka tulevikus plaanitud metaanalüüsis.

Üksikraku RNA sekveneerimisandmetest eQTL-ide leidmiseni on mitu sammu ja lõputöös ei testitud näiteks erinevate rakkude ja geenide filtreerimise, pseudo-hulkrakumaatriksite normaliseerimise, genotüüpide imputeerimise ega eQTL-ide tuvastamise meetodite mõju eQTL-analüüsi efektiivsusele. Lisaks ei sooritatud lõputöös üksikrakkudele rakutüüpide määramist, vaid toetuti uuringutes avaldatud metaandmetele. Kõiki neid samme tasub lähemalt vaadata kas andmestikes individuaalselt või metaanalüüsis.

Viidatud kirjandus

[1] Frank W. Albert ja Leonid Kruglyak. „The role of regulatory variation in complex traits and disease“. en. *Nature Reviews Genetics* 16.4 (aprill 2015). Number: 4 Publisher: Nature Publishing Group, lk. 197–212. ISSN: 1471-0064. DOI: 10.1038/nrg3891. URL: https://www.nature.com/articles/nrg3891 (vaadatud 08.05.2023).

[2] Nurlan Kerimov *et al.* „A compendium of uniformly processed human gene expression and splicing quantitative trait loci“. en. *Nature Genetics* 53.9 (september 2021). Number: 9 Publisher: Nature Publishing Group, lk. 1290–1299. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00924-w. URL: https://www.nature.com/articles/s41588-021-00924-w (vaadatud 02.05.2023).

[3] Byungjin Hwang, Ji Hyun Lee ja Duhee Bang. „Single-cell RNA sequencing technologies and bioinformatics pipelines“. en. *Experimental & Molecular Medicine* 50.8 (august 2018). Number: 8 Publisher: Nature Publishing Group, lk. 1–14. ISSN: 2092-6413. DOI: 10.1038/s12276-018-0071-8. URL: https://www.nature.com/articles/s12276-018-0071-8 (vaadatud 25.04.2023).

[4] Xiannian Zhang *et al.* „Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems“. en. *Molecular Cell* 73.1 (jaanuar 2019), 130–142.e5. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.10.020. URL: https://www.sciencedirect.com/science/article/pii/S1097276518308803 (vaadatud 25.04.2023).

[5] Benjamin D. Umans, Alexis Battle ja Yoav Gilad. „Where Are the Disease-Associated eQTLs?“ en. *Trends in Genetics* 37.2 (veebruar 2021), lk. 109–124. ISSN: 0168-9525. DOI: 10.1016/j.tig.2020.08.009. URL: https://www.sciencedirect.com/science/article/pii/S0168952520302092 (vaadatud 08.05.2023).

[6] Monique G. P. van der Wijst *et al.* „Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs“. en. *Nature Genetics* 50.4 (aprill 2018). Number: 4 Publisher: Nature Publishing Group, lk. 493–497. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0089-9. URL: https://www.nature.com/articles/s41588-018-0089-9 (vaadatud 08.05.2023).

[7] Anna S. E. Cuomo *et al.* „Single-cell genomics meets human genetics“. en. *Nature Reviews Genetics* (aprill 2023). Publisher: Nature Publishing Group, lk. 1–15. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00599-5. URL: https://www.nature.com/articles/s41576-023-00599-5 (vaadatud 25.04.2023).

[8] MGP van der Wijst *et al.* „The single-cell eQTLGen consortium“. *eLife* 9 (märts 2020). Toim. Helena Pérez Valle *et al.* Publisher: eLife Sciences Publications, Ltd, e52155. ISSN: 2050-084X. DOI: 10.7554/eLife.52155. URL: https://doi.org/10.7554/eLife.52155 (vaadatud 03.05.2023).

[9] Haley E. Randolph *et al.* „Genetic ancestry effects on the response to viral infection are pervasive but cell type specific“. *Science* 374.6571 (november 2021). Publisher: American Association for the Advancement of Science, lk. 1127–1133. DOI: 10.1126/science.abg0928. URL: https://www.science.org/doi/10.1126/science.abg0928 (vaadatud 23.04.2023).

[10] Seyhan Yazar *et al.* „Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease“. *Science* 376.6589 (aprill 2022). Publisher: American Association for the Advancement of Science, eabf3041. DOI: 10.1126/science.abf3041. URL: https://www.science.org/doi/10.1126/science.abf3041 (vaadatud 23.04.2023).

[11] Richard K. Perez *et al.* „Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus“. *Science* 376.6589 (aprill 2022). Publisher: American Association for the Advancement of Science, eabf1970. DOI: 10.1126/science.abf1970. URL: https://www.science.org/doi/10.1126/science.abf1970 (vaadatud 23.04.2023).

[12] Runyang Nicolas Lou *et al.* „A beginner's guide to low-coverage whole genome sequencing for population genomics“. en. *Molecular Ecology* 30.23 (2021). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16077, lk. 5966–5993. ISSN: 1365-294X. DOI: 10.1111/mec.16077. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16077 (vaadatud 27.04.2023).

[13] Shuo Shi *et al.* „Comprehensive Assessment of Genotype Imputation Performance“. Inglise keel. *Human Heredity* 83.3 (2018). Publisher: Karger Publishers, lk. 107–116. ISSN: 0001-5652, 1423-0062. DOI: 10.1159/000489758. URL: https://www.karger.com/Article/FullText/489758 (vaadatud 27.04.2023).

[14] Brian L. Browning, Ying Zhou ja Sharon R. Browning. „A One-Penny Imputed Genome from Next-Generation Reference Panels“. en. *The American Journal of Human Genetics* 103.3 (september 2018), lk. 338–348. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2018.07.015. URL: https://www.sciencedirect.com/science/article/pii/S0002929718302428 (vaadatud 27.04.2023).

[15] Center for Statistical Genetics. *statgen/Minimac4*. en. URL: https://github.com/statgen/Minimac4 (vaadatud 27.04.2023).

[16] Simone Rubinacci *et al. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes*. en. Pages: 2022.11.28.518213 Section: New Results. November 2022. DOI: 10.1101/2022.11.28.518213. URL: https://www.biorxiv.org/content/10.1101/2022.11.28.518213v1 (vaadatud 27.04.2023).

[17] Adriano De Marino *et al.* „A comparative analysis of current phasing and imputation software“. *PLoS ONE* 17.10 (oktoober 2022), e0260177. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0260177. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9581364/ (vaadatud 27.04.2023).

[18] François Aguet *et al.* „Molecular quantitative trait loci“. en. *Nature Reviews Methods Primers* 3.1 (jaanuar 2023). Number: 1 Publisher: Nature Publishing Group, lk. 1–22. ISSN: 2662-8449. DOI: 10.1038/s43586-022-00188-6. URL: https://www.nature.com/articles/s43586-022-00188-6 (vaadatud 01.05.2023).

[19] Peep Kolberg. „Atoopilise dermatiidiga seotud geenide tuvastamine geneetilise kolokalisatsiooni abil“. Bakalaureusetöö. Tartu: Tartu Ülikool, 2021.

[20] Nurlan Kerimov *et al.* „eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs“. en. *bioRxiv* (jaanuar 2021). Publisher: Cold Spring Harbor Laboratory Section: New Results, lk. 2020.01.29.924266. DOI: 10.1101/2020.01.29.924266. URL: https://www.biorxiv.org/content/10.1101/2020.01.29.924266v2 (vaadatud 02.05.2021).

[21] Urmo Võsa *et al.* „Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression“. en. *Nature Genetics* 53.9 (september 2021). Number: 9 Publisher: Nature Publishing Group, lk. 1300–1310. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00913-z. URL: https://www.nature.com/articles/s41588-021-00913-z (vaadatud 01.05.2023).

[22] Anna S E Cuomo *et al.* „CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq“. *Molecular Systems Biology* 18.8 (august 2022). Publisher: John Wiley & Sons, Ltd, e10663. ISSN: 1744-4292. DOI: 10.15252/msb.202110663. URL: https://www.embopress.org/doi/full/10.15252/msb.202110663 (vaadatud 03.05.2023).

[23] Natsuhiko Kumasaka *et al. Mapping interindividual dynamics of innate immune response at single-cell resolution*. en. Pages: 2021.09.01.457774 Section: New Results. September 2021. DOI: 10.1101/2021.09.01.457774. URL: https://www.biorxiv.org/content/10.1101/2021.09.01.457774v1 (vaadatud 03.05.2023).

[24] Anna S. E. Cuomo *et al.* „Optimizing expression quantitative trait locus mapping workflows for single-cell studies“. *Genome Biology* 22.1 (juuni 2021), lk. 188. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02407-x. URL: https://doi.org/10.1186/s13059-021-02407-x (vaadatud 28.04.2023).

[25] Radhika Khetani *et al.* *Single-cell RNA-seq: Quality Control Analysis*. en-US. Veebruar 2020. URL: https://hbctraining.github.io/scRNA-seq/lessons/04_SC_quality_control.html (vaadatud 09.05.2023).

[26] Grace X. Y. Zheng *et al.* „Massively parallel digital transcriptional profiling of single cells“. en. *Nature Communications* 8.1 (jaanuar 2017). Number: 1 Publisher: Nature Publishing Group, lk. 14049. ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: https://www.nature.com/articles/ncomms14049 (vaadatud 03.05.2023).

[27] Cody N. Heiser *et al.* „Automated quality control and cell identification of droplet-based single-cell data using dropkick“. en. *Genome Research* 31.10 (oktoober 2021). Publisher: Cold Spring Harbor Laboratory Press, lk. 1742. DOI: 10.1101/gr.271908.120. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8494217/ (vaadatud 09.05.2023).

[28] Natsuhiko Kumasaka, Andrew J. Knights ja Daniel J. Gaffney. „High-resolution genetic mapping of putative causal interactions between regions of open chromatin“. en. *Nature Genetics* 51.1 (jaanuar 2019). Number: 1 Publisher: Nature Publishing Group, lk. 128–137. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0278-6. URL: https://www.nature.com/articles/s41588-018-0278-6 (vaadatud 06.05.2023).

[29] Peep Kolberg. *peepkolberg/glimpse*. original-date: 2022-12-20T13:29:06Z. Detsember 2022. URL: https://github.com/peepkolberg/glimpse (vaadatud 03.05.2023).

[30] Nurlan Kerimov *et al.* *eQTL-Catalogue/qtlmap: Portable eQTL analysis and statistical fine mapping workflow used by the eQTL Catalogue*. en. URL: https://github.com/eQTL-Catalogue/qtlmap (vaadatud 04.05.2023).

[31] Nurlan Kerimov, Kaur Alasoo ja Ralf Tambets. *eQTL-Catalogue/genimpute: Portable genotype imputation pipeline used by the eQTL Catalogue*. en. URL: https://github.com/eQTL-Catalogue/genimpute (vaadatud 03.05.2023).

[32] Adam Auton *et al.* „A global reference for human genetic variation“. en. *Nature* 526.7571 (oktoober 2015). Number: 7571 Publisher: Nature Publishing Group, lk. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. URL: https://www.nature.com/articles/nature15393 (vaadatud 03.05.2023).

[33] Heng Li ja Richard Durbin. „Fast and accurate short read alignment with Burrows-Wheeler transform“. eng. *Bioinformatics (Oxford, England)* 25.14 (juuli 2009), lk. 1754–1760. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324.

[34] Petr Danecek *et al.* „Twelve years of SAMtools and BCFtools“. en. *GigaScience* 10.2 (jaanuar 2021), giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008. URL: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giab008/6137722 (vaadatud 03.05.2023).

[35] Simone Rubinacci *et al.* „Efficient phasing and imputation of low-coverage sequencing data using large reference panels“. en. *Nature Genetics* 53.1 (jaanuar 2021). Number: 1 Publisher: Nature Publishing Group, lk. 120–126. ISSN: 1546-1718. DOI: 10.1038/s41588-020-00756-0. URL: https://www.nature.com/articles/s41588-020-00756-0 (vaadatud 23.04.2023).

[36] Olivier Delaneau. *odelaneau/GLIMPSE at glimpse1*. URL: https://github.com/odelaneau/GLIMPSE/tree/glimpse1 (vaadatud 23.04.2023).

[37] Harshil Patel *et al. nf-core/atacseq: nf-core/atacseq v1.2.2 - Iron Ossifrage*. Mai 2022. DOI: 10.5281/zenodo.6544493. URL: https://zenodo.org/record/6544493 (vaadatud 09.05.2023).

[38] 10X Genomics. *What is Cell Ranger? -Software -Single Cell Gene Expression -Official 10x Genomics Support*. URL: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger (vaadatud 04.05.2023).

[39] Philip A. Ewels *et al.* „The nf-core framework for community-curated bioinformatics pipelines“. en. *Nature Biotechnology* 38.3 (märts 2020). Number: 3 Publisher: Nature Publishing Group, lk. 276–278. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0439-x. URL: https://www.nature.com/articles/s41587-020-0439-x (vaadatud 04.05.2023).

[40] nf-core. *scrnaseq » nf-core*. URL: https://nf-co.re/scrnaseq/2.3.0/usage (vaadatud 04.05.2023).

[41] Ensembl. *Index of /pub/release-96/fasta/homo_sapiens/dna*. URL: http://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/dna/ (vaadatud 06.05.2023).

[42] EMBL-EBI. *Index of /pub/databases/spot/eQTL/references*. URL: http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/references/ (vaadatud 06.05.2023).

[43] Páll Melsted *et al.* „Modular, efficient and constant-memory single-cell RNA-seq preprocessing“. en. *Nature Biotechnology* 39.7 (juuli 2021). Number: 7 Publisher: Nature Publishing Group, lk. 813–818. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00870-2. URL: https://www.nature.com/articles/s41587-021-00870-2 (vaadatud 04.05.2023).

[44] Pachter Lab. *Building a reference - kallisto | bustools*. URL: https://www.kallistobus.tools/kb_usage/kb_ref/ (vaadatud 04.05.2023).

[45] Haynes Heaton *et al. souporcell: Robust clustering of single cell RNAseq by genotype and ambient RNA inference without reference genotypes*. en. Pages: 699637 Section: New Results. September 2019. DOI: 10.1101/699637. URL: https://www.biorxiv.org/content/10.1101/699637v2 (vaadatud 04.05.2023).

[46] Matiss Ozols. *wtsi-hgi/yascp: scRNA analysis pipeline*. en. URL: https://github.com/wtsi-hgi/yascp (vaadatud 04.05.2023).

[47] Haynes Heaton. *wheaton5/souporcell: Clustering scRNAseq by genotypes*. en. URL: https://github.com/wheaton5/souporcell (vaadatud 04.05.2023).

[48] Haynes Heaton. *Corresponding cluster labels to each individual · Issue #120 · wheaton5/souporcell*. en. URL: https://github.com/wheaton5/souporcell/issues/120 (vaadatud 24.04.2023).

[49] Constantin Ahlmann-Eltze ja Wolfgang Huber. „Comparison of transformations for single-cell RNA-seq data“. en. *Nature Methods* (aprill 2023). Publisher: Nature Publishing Group, lk. 1–8. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01814-1. URL: https://www.nature.com/articles/s41592-023-01814-1 (vaadatud 07.05.2023).

[50] Pachter Lab. *kallistobustools*. original-date: 2019-06-13T03:22:17Z. Aprill 2023. URL: https://github.com/pachterlab/kallistobustools (vaadatud 06.05.2023).

[51] F. Alexander Wolf, Philipp Angerer ja Fabian J. Theis. „SCANPY: large-scale single-cell gene expression data analysis“. *Genome Biology* 19.1 (veebruar 2018), lk. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: https://doi.org/10.1186/s13059-017-1382-0 (vaadatud 11.05.2023).

[52] Halit Ongen *et al.* „Fast and efficient QTL mapper for thousands of molecular phenotypes“. *Bioinformatics* 32.10 (mai 2016), lk. 1479–1485. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv722. URL: https://doi.org/10.1093/bioinformatics/btv722 (vaadatud 06.05.2023).

[53] Olivier Delaneau *et al.* „A complete tool set for molecular QTL discovery and analysis“. en. *Nature Communications* 8.1 (mai 2017). Number: 1 Publisher: Nature Publishing Group, lk. 15452. ISSN: 2041-1723. DOI: 10.1038/ncomms15452. URL: https://www.nature.com/articles/ncomms15452 (vaadatud 04.05.2023).

[54] Ed Mountjoy. *rank-based-INT*. original-date: 2016-03-10T12:59:54Z. Jaanuar 2023. URL: https://github.com/edm1/rank-based-INT (vaadatud 24.04.2023).

[55] Nurlan Kerimov *et al. Systematic visualisation of molecular QTLs reveals variant mechanisms at GWAS loci*. en. Pages: 2023.04.06.535816 Section: New Results. Aprill 2023. DOI: 10.1101/2023.04.06.535816. URL: https://www.biorxiv.org/content/10.1101/2023.04.06.535816v1 (vaadatud 08.05.2023).

[56] Kristiina Kuningas. „Estimating Concordance Between Measured and Predicted Genetic Variant Effect on Chromatin Accessibility“. Magistritöö. Tartu: Tartu Ülikool, 2023.

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Peep Kolberg**,
(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   Ekspressiooni kvantitatiivsete tunnuste lookuste analüüs üksikraku RNA sekveneerimisandmetes,
   (lõputöö pealkiri)

   mille juhendaja on Kaur Alasoo,
   (juhendaja nimi)

   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.

Peep Kolberg
09.05.2023