TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika õppekava

Madis Kaasik

Geograafilise päritolu ennustamine geeni-ekspressiooni ja geneetilise varieeruvuse abil

Bakalaureusetöö (9 EAP)

Juhendaja(d): Tauno Metsalu
Tatjana Iljašenko

Tartu 2015

Geograafilise päritolu ennustamine geenieksпressiooni ja geneetilise varieeruvuse abil

Lühikokkuvõte:

Käesoleva bakalaureusetöö eesmärk on uurida, kui palju erineva päritoluga inimesed erinevad üksteisest geenieksпressiooni või geneetilise varieeruvuse mõttes. Selleks kasutatakse avalikke andmeid, kus geenieksпressiooni ja geneetilist varieeruvust on mõõdetud erineva päritoluga ameeriklastel. Andmete analüüsimiseks kasutatakse statistikapakett R'i. Töö käigus tutvutakse erinevate andmeformaatide ja analüüsivõtetega. Antakse ülevaade erinevatest statistilistest meetoditest, masinõppe algoritmidest ning rakendatakse neid eelpool mainitud andmetel. Lõppeesmärgiks on leida, kui täpselt on võimalik ennustada päritolu geenieksпressiooni abil, geneetilise varieeruvuse abil ja kasutades mõlemat korraga ning leida, milline klassifitseerimismeetod sobib kõige paremini päritolu määramiseks.

Võtmesõnad:

Hierarhiline klasterdamine, dispersioonanalüüs, Fisheri test, Random Forest, geenieksпressioon, üksiku nukleotiidi polümorfism

Prediction of geographic origin based on gene expression and genetic variation data analysis

Abstract:

The aim of this thesis is to study, how much do gene expression levels or single nucleotide polymorphisms (SNPs) differ in different ethnical groups. Sample data is publicly accessible gene expression and SNP data, which is collected from americans with different ethincal origin. Statistical analysis software R is used for analysing this data. Thesis aims to give an overview of different statistical methods, machine learning algorithms and apply them on sample data. The end goal is to find out how precisely can origin be predicted using gene expression, genetic variability, gene expression and genetic variability and which classification method is best suited for origin determination.

Keywords:

Hierarchical clustering, analysis of variance, Fisher's exact test, random forest, gene expression, single nucleotide polymorphism.

Sisukord

1. Sissejuhatus

Pärilikkus on läbi aegade pakkunud huvi inimestele ja eriti teadlastele. Tuhande kaheksasaja seitsmekümne esimesel aastal avastati DNA, pärast seda avastust on suudetud mõnedele huvipakkunud teemadele vastuseid leida, aga iga leitud vastus on tekitanud uusi küsimusi. Praeguseks on suudetud kogu inimese genoom kaardistada. Teada on, et igal inimesel on samad geenid, ometi on ka ilmne, et iga inimene on erinev. See on tingitud sellest, et igal inimesel on geenid erinevalt avaldunud ehk ekspresseerunud ning samuti erineb ka DNA järjestus igal indiviidil.

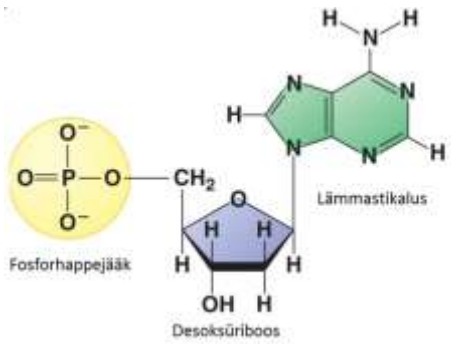
Geenide avaldumist ning järjestuste erinevusi uurides on võimalik mõista, miks me oleme sellised nagu me oleme: pikkus, kaal, juustevärv. Samuti annab see võimaluse mõista paremini haigusi, näiteks millal ning millistes tingimustes need avalduvad, kes kuuluvad riskigruppi, kuidas ennetada ning kuidas ravida.

Käesolev töö üritab leida, milline on seos geenide avaldumise ning nahavärvuse vahel. Millised geenid avalduvad mõnel päritolugrupil rohkem, millised vähem. Samuti otsitakse DNA järjestuses kohti ehk lookuseid, kuhu on sisse kirjutatud inimese nahavärvus. Selleks uurime geenieksпressiooni ja geneetilist varieeruvust ehk DNA järjestust.

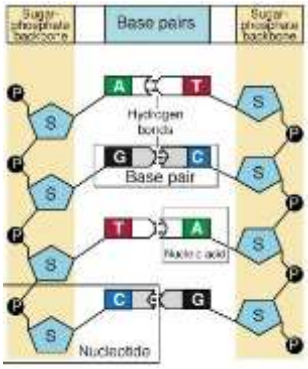
Töö teises peatükis on lühike ülevaade bioloogilisest taustast. Kolmandas peatükis tutvustatakse erinevaid statistilisi analüüsimeetodeid, mis võimaldavad avastada statistiliselt olulisi erinevusi ekspressioonis ning järjestuses populatsioonide vahel. Neljandas peatükis tutvustatakse uurimiseks kasutatud andmeid. Viimases peatükis esitatakse analüüsi käiku, raporteeritakse nii olulisteks osutunud geenid/genoomi lõigud ekspressiooni mõttes kui ka DNA varieeruvuse lookused ning esitatakse päritolu ennustamise tulemused. Samas antakse ka saadud tulemustele hinnang.

2. Bioloogiline Taust

Inimese pärilik info säilitatakse desoksüribonukleiinhappes- DNA's. DNA molekul koosneb nukleotiididest, mis omakorda koosnevad fosforhappejäägist, desoksüriboosist ning lämmastikalusest (vt joonis 1). Erinevaid lämmastikaluseid on neli: adeniin (A), guaniin (G), tsütosiin (C) ja tümiin (T). DNA esineb elusorganismis kahe komplementaarse ahelana ehk kaksikheeliksina, kus iga A nukleotiid paardub T nukleotiidiga ja iga G nukleotiid paadub C nukleotiidiga (vt joonis 2).



Joonis 1. Nukleotiid [1]



Joonis 2. Kaksikheeliks [2]

Geeniekspressioon ehk geeni avaldumine on geenis oleva päriliku info põhjal valgu sünteesimine. Valgud täidavad organismis mitmeid erinevaid funktsioone: kaitse-, struktuuri-, transpordifunktioone. DNA-st valkude sünteesimiseks peab toimuma kaks protsessi: transkriptsioon ja translatsioon (vt joonis 3).

Joonis 3. Seos DNA, RNA ja valgu vahel.

Transkriptsioon on geenide avaldumise regulatsiooni peamiseks tasemeks. Transkriptsiooniks nimetatakse komplementaarse RNA molekuli sünteesi DNA molekuli põhjal. Saadud RNA ahelas on tümiin (T) asendunud uratsiiliga (U). DNA ja RNA vaheline komplementaarsus on: A – U, T – A, C – G, G – C. Näiteks DNA monomeerile ACGCT vastaks RNA, mille lämmastikalused oleksid UGCGA.

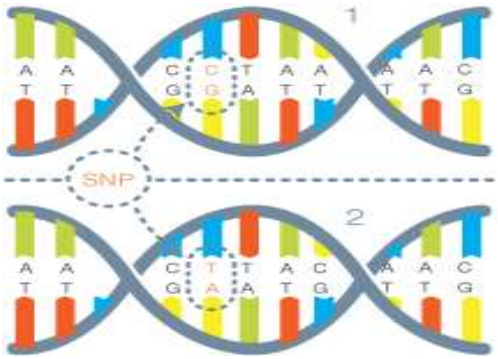
Erinevalt DNA-st, on RNA ebastabiilne ning ei ole rakus püsivalt. Lühikese aja pärast see lagundatakse. RNA-d on 3 erinevat liiki: informatsiooni-RNA (mRNA), transpordi-RNA (tRNA) ja ribosoomi-RNA (rRNA). M-RNA ülesandeks on kanda pärilik info ribosoomi.

Ribosoomides toimub valgu süntees ehk translatsioon. Informatsiooni-RNA jagatakse kolmenukleotiidisteks gruppideks ehk koodoniteks. Translatsiooni alustuseks on vaja kindlat koodonit, mida nimetatakse startkoodoniks. Koodoneid on kolme tüüpi:

- Startkoodon, mis määrab translatsiooni alguskoha
- Koodonid, millele vastab kindel aminohape
- Stoppkoodonid, millele ei vasta ühtegi aminohapet

Startkoodonile vastab samuti kindel aminohape, seega algab enamus valke sama aminohappega. Stoppkoodonid tähistavad sünteesi lõppu. Moodustunud aminohappeahelast moodustub hiljem valk.

Inimese genoom koosneb kolmest miljardist DNA aluspaarist [3]. Kahe inimese DNA erineb kuni 0,1%. Kahe inimese genoomi erinevus võib olla põhjustatud ühe nukleotiidi või mitme järjestikuse nukleotiidi asendumisest. Neid järjestuse variatsioone, mis on toimunud ühe nukleotiidi (A,T,G,C) muutumisel, nimetatakse ühenukleotiidseteks polümorfismideks ehk SNP-deks(inglise keeles *single nucleotide polymorphisms,* vahel kasutatakse ka terminit punktmutatsioon). SNP-d võivad esineda ka paariskäivates (homoloogilistes) kromosoomides (vt joonis 4).



Joonis 4. Kaks DNA molekuli, mis erinevad teineteisest ühe nukleotiidi võrra. [4]

SNP-d võivad määrata, kuidas inimestel arenevad haigused ning kuidas reageeritakse ravimitele, kemikaalidele ning vaktsiinidele. Seepärast uuritakse SNP-sid järjest enam, eriti on uurimistööst huvitatud meditsiini valdkond. Eesmärk on muuta meditsiin personaalsemaks. Geenikiibi tehnoloogia on teinud SNP-de analüüsimise väga lihtsaks, korraga on võimalik analüüsida kümneid tuhandeid SNP-sid üle kogu genoomi.

Kui DNA järjestuse väärtus konkreetses lookuses erineb sama liigi erinevatel isenditel või sama isiku erinevates kromosoomides, võime öelda, et tegemist on kahe erineva alleeliga. Alleelid on näiteks GC**C**TA ja GC**T**TA**.** Enamasti on SNP-del kaks alleeli. Alleelid jagunevad omakorda kaheks: dominantseteks ja retsessiivseteks. Dominantse alleeli poolt määratud tunnus avaldub alati, retsessiivse alleeli poolt määratud tunnus avaldub ainult dominantse alleeli puudumisel.

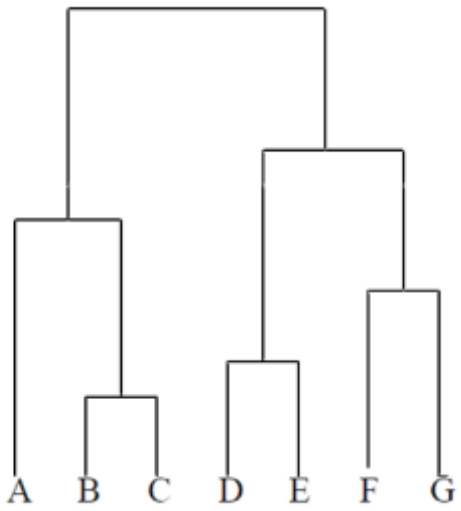
Kui homoloogiliste kromosoomide samades piirkondades on ühel geenil ühesugused (mõlemad dominantsed või mõlemad retsessiivsed) alleelid, nimetatakse seda geenipaari homosügootseks. Kui alleelid on erinevad (üks on dominantne ja teine retsessiivne), nimetatakse geenipaari heterosügootseks.

3. Statistiline taust

3.1 Hierarhiline klasterdamine

Hierarhiline klasterdamine on üks klasteranalüüsi meetoditest. Klasteranalüüs grupeerib objektid määratud tunnuse sarnasuse põhjal. Käesolevas töös kasutatakse R paketi *hclust* meetodit, mis on loomu poolest aglomeratiivne klasteranalüüs. See tähendab, et klasterdamine toimub "alt üles".

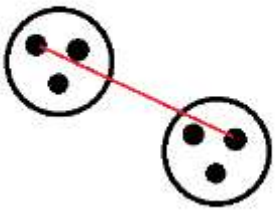
Alguses moodustab iga vaadeldav objekt eraldi klastri ehk grupi, hiljem arvutatakse iga klastri kaugus teistest gruppidest ning lähimal asuvad ehk tunnuse suhtes kõige sarnasemad klastrid ühendatakse [5]. Klastrite ühendamine toimub kuni kõik objektid on ühes klastris [5]. Klastrite vahelisi seoseid näidatakse tavaliselt dendrogrammiga. (vt joonis 5).



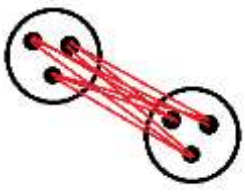
Joonis 5. Hierarhilisel klasterdamisel tekkiv dendogramm.

Meetodeid, leidmaks millised klastrid asetsevad teineteisele kõige lähemal,on mitmeid. Selles töös kasutatakse kahte erinevat R paketi meetodit kauguse arvutamiseks: maksimaalse kauguse meetodit ja keskmise kauguse meetodit.

Maksimaalse (*Complete linkage*) kauguse meetodil leitakse iga klastripaari puhul nende kõige kaugemal asuvate elementide kaugus. Keskmise (*Average linkage*) kauguse meetodi korral arvutatakse klastripaari kõikide objektide kauguse teistest objektidest keskmine. Mõlemal juhul ühendatakse klastrid, mille vaheline saadud kaugus on kõige väiksem.



Joonis 6.
Complete linkage meetod.



Joonis 7.
Average linkage meetod

3.2 Dispersioonanalüüs

Dispersioonanalüüs (ANOVA) inglise keeles *Analysis of variance* uurib gruppidevaheliste erinevuste statistilist olulisust. Meetod arvutab iga grupi keskväärtuse ning võrdleb seda kõikide gruppide keskväärtustega. Eeldused dispersioonanalüüsi kasutamiseks on [6]:

- Uuritavaid gruppe on 3 või rohkem
- Sõltuv tunnus peab olema arvtunnus
- Võrreldavad grupid peavad olema omavahel sõltumatud
- Tulemuste hajuvused võrreldavates gruppides peavad olema sarnased
- Tulemuste jaotus peab olema ligilähedane normaaljaotusele

Loomu poolest on dispersioonanalüüs statistiline hüpoteesi testimise meetod. Püstitatakse kaks hüpoteesi [7]:

$$H_0: \mu_{1=} \mu_{2==\ldots} \mu_n \text{ (nullhüpotees)}$$

$$H_1: \exists \mu_{1\neq} \mu_1 \text{ (sisukas hüpotees)}$$

Oluline on ka paika panna olulisuse nivoo α. Olulisuse nivoo näitab, kui suur on eksimise tõenäosus, mille tõttu saab seda kasutada kui mõõtu, millest alates võib nullhüpoteesi ümber lükata. Enamasti valitakse selleks 0,05. Mõningatel juhtudel ka 0,01 või 0,1.

Eeldades, et nullhüpotees on tõene, peaksid vaadeldavad grupid olema üldkogumist võetud juhuslikud valimid. Võetud valimite hajuvuste põhjal peaks saama hinnata üldkogumi hajuvust. Üldkogumi hajuvust on võimalik hinnata kahel erineval tasemel. [6]

1. Hinnang üldkogumi dispersioonile valimite sisese dispersiooni põhjal.
2. Hinnang üldkogumi dispersioonile valimite vahelise dispersiooni põhjal.

Selleks, et nullhüpotees H_0 kehtiks, peavad need hinnangu viisid olema ligilähedased. Juhul, kui need hinnangud on erinevad, võib väita, et nullhüpotees ei kehti. Erinevuste suurust väljendab statistik F, mida arvutatakse järgnevalt [6]:

$$F = \frac{\text{Hinnang üldkogumi dispersioonile valimite vahelise dispersiooni põhjal}}{\text{Hinnang üldkogumi dispersioonile valimite sisese dispersiooni põhjal.}}$$

Ilmne on, et nullhüpoteesi kehtimise korral peab F väärtus olema ligikaudselt 1. Näha on ka, et mida suurem on hinnang üldkogumi dispersioonile valimite vahelise dispersiooni põhjal, seda suurem on ka F [6]. Hüpoteesi testimise eesmärgiga leitakse kriitiline väärtus vastavast tabelist (olgu see f), mida võrreldakse F-statistiku väärtusega. Juhul, kui $F \geq f$, lükkame ümber nullhüpoteesi ning kinnitame sisuka hüpoteesi, mille eksimise tõenäosus on võrdne valitud olulisuse nivooga [7]. Juhul, kui $F < f$, jääme nullhüpoteesi juurde [7]. Praktikas arvutatakse tihti p-väärtus ning võrreldakse seda olulisusnivooga. Kui p-väärtus on suurem kui paika pandud olulisusnivoo, jäädakse nullhüpoteesi juurde. Vastasel juhul kummutatakse nullhüpotees ja sisukas hüpotees loetakse tõestatuks.

Antud töös on kasutatud nn *One-way Anova*, mis on ühefaktoriline dispersioonanalüüs, kus ainsaks seletavaks tunnuseks on kolme tasemeline faktortunnus, mis kirjeldab indiviidide päritolu. Sõltumatuks tunnuseks on numbriline tunnus, mis kirjeldab ekspressiooni taste ning on viidud normaaljaotusele, kasutades log-transformatsiooni.

3.3 Fisheri test

Fisheri test uurib statistilist olulisust, kasutades selleks sagedustabeleid. Fisheri testi saab kasutada diskreetse jaotuse korral. Enamasti on sagedustabeli suuruseks 2 x 2, aga sagedustabel võib olla ka suuremate mõõtmetega. Tabeli veergudes on uuritavad grupid ning ridades tulemused. Käesolevas töös kasutatakse 3 x 3 sagedustabeleid. Fisheri test arvutab, kui suur on tõenäosus saada selline või veelgi ekstremaalsem sagedustabel, eeldusel, et rea tulemus ei sõltu veerust ehk rea tulemused jaotuvad veergude vahel võrdselt [8].

3.4 Random Forest

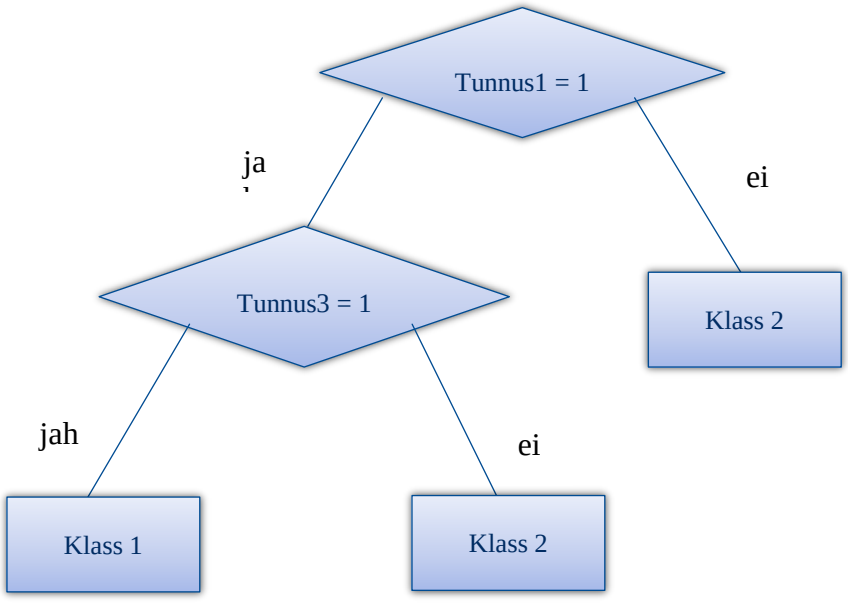
Random Forest on klassifitseerimismeetod, mis kasutab otsustuspuid. Otsustuspuid kasutatakse masinõppes. Otsustuspuu on mudel, mis ennustaks objekti klassi etteantud tunnuste põhjal. Otsustuspuud kujutatakse puu kujul, kus igale tipule vastab test, igale kaarele testi tulemus ja igale lehele klass. Otsusepuude loomiseks on järgnev algoritm:

1. Alusta juurtipust, kõigi treeningandmetega.
2. Kontrolli tippu kuuluvaid elemente, kui kõik elemendid on samast klassist, märgi tipp leheks ning lisa klassitunnus.
3. Leida parim tunnus, mille järgi tipu objektid jagada.
4. Luua kaks alamtippu ja jagada objektid nende vahel valitud parima tunnuse järgi ära.

Vaata allpool olevat näidet.

	Tunnus1	Tunnus 2	Tunnus3
Klass 1	1	0	1
Klass 2	1	0	0
Klass 2	2	0	1

Tabel 1**.** Andmete maatriks



Joonis 8. Tabel1 põhjal moodustunud otsusepuu.

Näites on andmete maatriks, milles on 3 objekti ning igal objektil on 3 tunnust. Selle maatriksi põhjal on koostatud otsustuspuu, kus esimeses tipus jagatakse andmed tunnuse 1 järgi, teises tipus tunnuse 3 järgi ning igas lehttipus on määratud, millisesse klassi selliste tunnustega objekt kuulub. Uue objekti klassifitseerimiseks, alustatakse juurtipust, kus vastavalt tunnusele liigutakse allapoole ning jõudes lehttippu määratakse objekti klassiks lehele vastav klassitunnus.

Random Forest meetodi puhul kasutatakse mitut otsustuspuud, see aitab vähendada üksiku otsustuspuu kasutamisel tekkivat üleõppimise probleemi. *Random Forest*'i loomise algoritm *k* puu loomiseks on järgnev [9]:

- Andmete maatriksist, mis sisaldab *n* objekti (rida) ja *m* tunnust (veergu), valitakse juhuslikult tagasipanekuga *n* objekti (rida), mille tagajärjel tekib uus maatriks (*bootstrap* andmestik), kus mõned objektid korduvad, mõned aga puuduvad (keskmiselt satub loodud maatriksisse 2/3 objektidest [10]). Selliseid *bootstrap* andmestikke moodustatakse *k* tükki. Seda protseduuri nimetatakse *bagging*.
- Igast saadud maatriksist valitakse juhuslikult $\sqrt{m}$ tunnust, mille tagajärjel saadakse *k* alammaatriksit, mida kasutatakse puude moodustamiseks. Puud moodustavad metsa.
- Klassifitseerimise otsus võetakse kõikide puude klassifitseerimiste tulemuste pealt, kus objekti klassiks on klass, mis sai kõige rohkem hääli [10]. Näiteks, kui metsas on neli otsusepuud ning kaks neist klassifitseerivad objekti esimesse klassi, üks teise klassi ja üks kolmandasse klassi, siis meetodi järgi kuulub see objekt esimesse klassi.

Treeningandmete ennustamise täpsuse hindamiseks kasutatakse nn *Out of bag* vea hinnangumeetodit, mis on väga sarnane *leave-one-out* krossvalideerimismeetodiga. Iga objekti puhul proovitakse seda klassifitseerida, kasutades puid, mille *bootstrap* andmestikku objekti ei valitud.

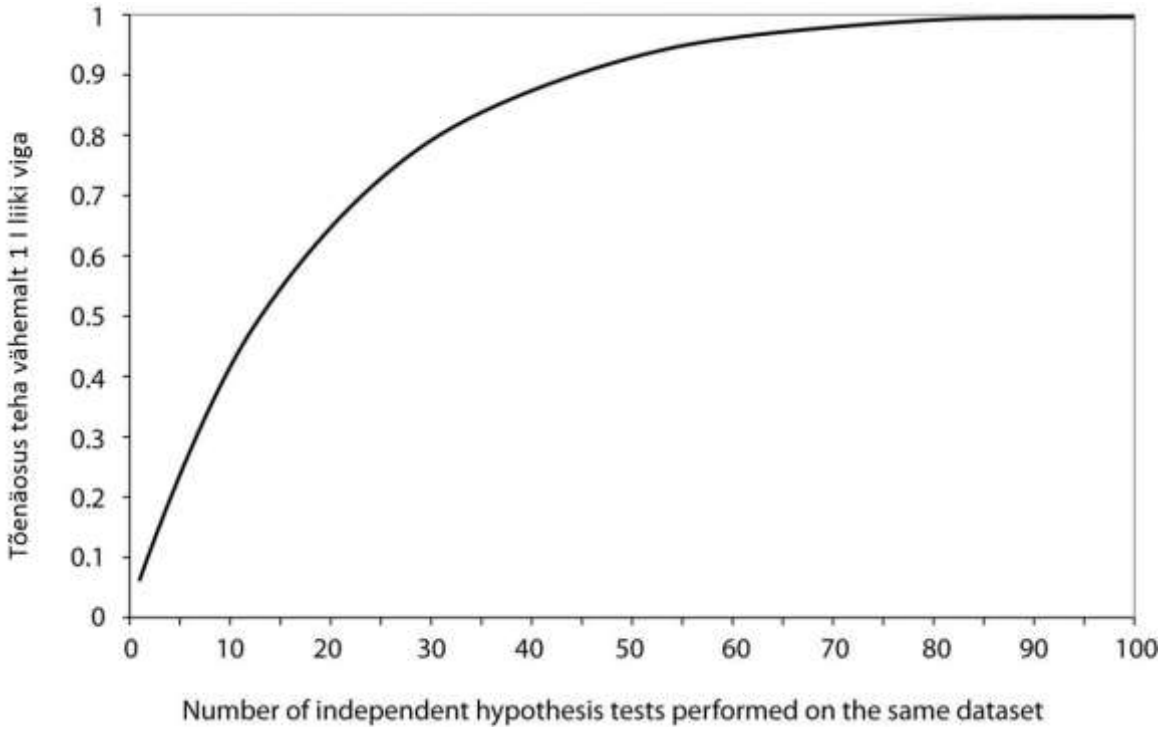
3.5 Mitmese testimise problem

Statistilise hüpoteesi kontrollimise korral üritatakse kummutada nullhüpotees. Tulemus loetakse oluliseks, kui tõenäosus saada selline tulemus juhuslikult on väga ebatõenäoline. Kuna valimi põhjal väidetakse midagi üldkogumi kohta, võib tekkida kahte liiki vigu (I ja II liiki) [11]. Hüpoteeside tegemisel tekkivaid vigu kirjeldab tabel 2. I liiki viga nimetatakse ka valepositiivseks tulemuseks ning II liiki viga valenegatiivseks tulemuseks. I liiki viga tekib siis, kui tõestatuks loetakse sisukas hüpotees, kuigi tegelikult on õige jääda nullhüpoteesi juurde. II liiki viga tekib, kui jäädakse nullhüpoteesi juurde, aga peaks vastu võtma sisuka hüpoteesi. I liiki viga peetakse raskemaks veaks, kui II tüüpi viga, sest sellisel juhul on tõestatud seos, mis on tekkinud juhuse tõttu ning mida päriselt seal ei ole.

	Nullhüpotees kehtib	Sisukas hüpotees kehtib
Nullhüpotees kummutatud	I liiki viga	Õige otsus
Nullhüpotees jääb kehtima	Õige otsus	II liiki viga

Tabel 2. Hüpoteesi kontrollimisel erinevad vead. [11]

Mitmese testimise probleem ilmneb, kui üheaegselt testitakse mitut hüpoteesi. Hüpoteeside arvu suurenedes, suureneb ka tõenäosus juhuslikult tõestada mõni sisukas hüpotees, mis tegelikult ei kehti ehk teha I liiki viga. I liiki vea tegemise tõenäosuse muutumist testitavate hüpoteeside arvu kasvamise korral illustreerib joonis 9.



Joonis 9. Tõenäosuse, teha vähemalt üks I liiki viga, suurenemine, testitavade hüpoteeside arvu kasvamise korral. [12]

Joonisel 9 on kujutatud, kuidas tõenäosus teha I liiki viga kasvab, testitavate tunnuste arvu suurenemise korral. M hüpoteesi kontrollimisel saadakse tõenäosus, et tehakse I liiki viga valemiga: $P(\text{vähemalt 1 I liiki viga}) = 1 - (1-\alpha)^m$. Valides olulisusnivooks 0,05, on 5 hüpoteesi kontrollimisel tehtav viga $1 - (1-0{,}05)^5 \approx 0{,}23$. Viiekümne hüpoteesi korral on I liiki vea tegemise tõenäosuseks $1 - (1-0{,}05)^{50} \approx 0{,}92$. Saja hüpoteesi kontrollimisel on tõenäosuseks $1 - (1-0{,}05)^{100} \approx 0{,}99$.

I liiki vea parandamiseks mitmese testimise korral on mitmeid erinevaid meetodeid. Üks levinumaid on *False Discovery Rate*, lühidalt FDR. FDR meetodit kasutatakse ka käesolevas töös. FDR meetodi algoritmi n sõltumatu hüpoteesi ja vastavalt saadud n p-väärtuste korral võib kirjeldada järgmiselt [13]:

- Järjestada saadud p-väärtused kasvamise järjekorras. $p_1 \leq p_2 \leq \ldots \leq p_n$.
- Leida suurim j, mille puhul $p_j \leq \alpha \frac{j}{n}$.
- Valida hüpoteesid 1,2,…j ning kummutada nullhüpotees ainult nende hüpoteeside puhul.

Teine laialdaselt kasutatud meetod p-väärtuste korrigeerimiseks (eelistatav bioinformaatika valdkonnas) on Bonferroni meetod. Bonferroni meetod ei nõua testide sõltumatust. *N* hüpoteesi korral ei võrrelda p-väärtusi olulisusnivoo α-ga, vaid hoopis $\frac{\alpha}{N}$ –ga. Juhul, kui p-väärtus on väiksem/võrdne, kui $\frac{\alpha}{N}$, võetakse vastu sisukas hüpotees. Vastasel juhul jäädakse nullhüpoteesi jurrde. Bonferroni meetod on väga konservatiivne ning kuigi I liiki vigade arv hoitakse normi piires, suureneb Bonferroni meetodit kasutades II liiki vigade arv [14].

4. Andmete eeltöötlus

4.1 Ekspressiooniandmed

Geeni ekspressiooni andmed on pärit 2010. aastal avaldatud uuringust, kus uuriti patsientide reageerimist keemiaravile. Selleks võrreldi uuringus osalevaid patsiente tervete inimestega. Võrreldavateks andmeteks kasutati 287 inimese mõõdetud geenieksпressiooni andmeid (GEOID: GSE23120). Neist 287-st inimesest 95 olid Aafrika-ameeriklased, 96 Euroopa-ameeriklased ja 96 aasia päritolu ameeriklased. Igal inimesel oli mõõdetud 54613 geeni/genoomi lõigu ekspressiooni, kasutades mikrokiibi tehnoloogiat. Mikrokiibi tehnoloogia võimaldab mõõta ühel patsiendil korraga paljude geenide ekspressiooni või SNP-sid. Mõõdetud geenieksпressiooni andmetemaatriksis võtame igast elemendist naturaallogaritmi, et tagada väärtustele normaaljaotus, mis on ANOVA analüüsi eelduseks.

4.2 Üksiku nukleotiidi polümorfismi andmed

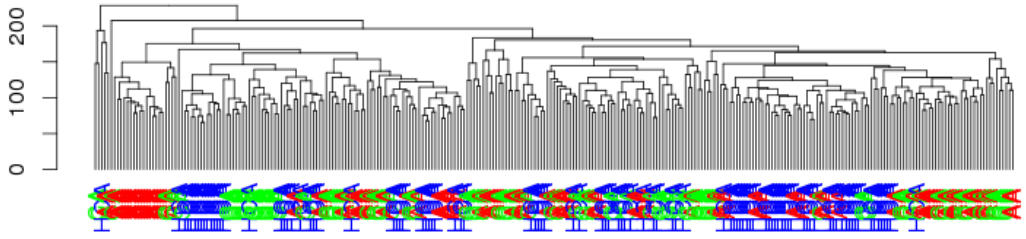
Üksiku nukleotiidi polümorfismi andmed on pärit samast 2010. aastal tehtud uuringust, kus uuriti patsientide reageerimist keemiaravile. Keemiaravi saanud patsientidel mõõdeti SNP-d ning võrreldi neid 288 terve inimese (GEOID: GSE24245) omadega. Käesolevas töös kasutame võrreldavate 288 inimese andmeid. 288-st inimesest 96 olid Aafrika-ameeriklased, 96 Euroopa-ameeriklased ning 96 Aasia-ameeriklased. Igal inimesel oli mõõdetud 511354 polümorfismi: inimesed olid antud lookustes genotüpiseeritd ning nende genotüübid olid kodeeritud vastavalt AA, AB ja BB, olenevalt kas nad olid heterosügoodid, homosügoodid ühe alleeli või teise alleeli suhtes. Kodeeringu A ja B all mõistetakse kahte võimalikku alleeli iga lookuse korral, vaadeldud populatsiooni tasemel. Kuna sõnaline kodeering ei sobi analüüsi tegemiseks, seatakse igale kodeeringule vastavusse numbriline väärtus. Väärtusteks on alleeli B arv iga induvidaalse genotüübi jaoks. Seega genotüübile AA seatakse vastavusse 0, genotüübile AB vastavusse 1 ja genotüübile BB seatakse vastavusse 2. Kui andmed on arvulisel kujul, on võimalik neid statistiliselt analüüsida.

5. Tulemused(Analüüs)

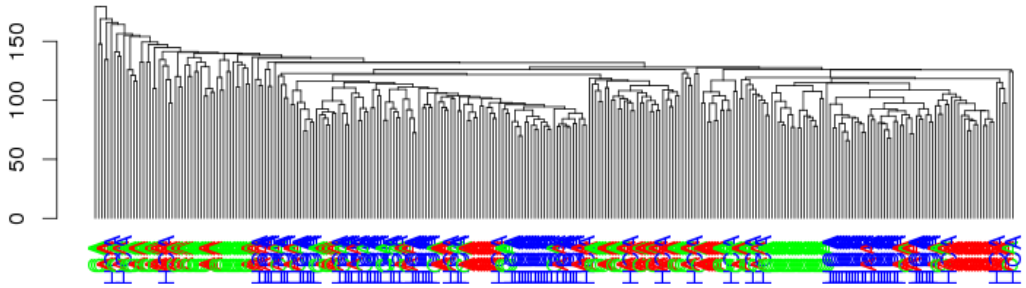
5.1 Hierarhilise klasterdamise tulemused

Enne andmete statistilist analüüsimist proovib töö autor geenieksрessiooni ja SNP andmeid klasterdada kasutades selleks R programmi meetodit *hclust*. Saadud tulemuste põhjal on võimalik saada esmane ülevaade, kas ja millised inimesed on üksteisele sarnasemad. Kui sama päritoluga inimesed koonduvad ühistesse klastritesse, võib eeldada, et nende kohta kogutud info on sarnane ja et suurem osa mõõdetud andmetest on päritoluga seotud. Kui sama päritoluga inimesed ühistesse klastritesse ei koondu, võib eeldada, et suuremas osas on kõik uuritavad inimesed kogutud tunnuste poolest sarnased. Andmete suure hulga tõttu ei saa siiski välistada, et uuritavad mõne tunnuse poolest erinevad.

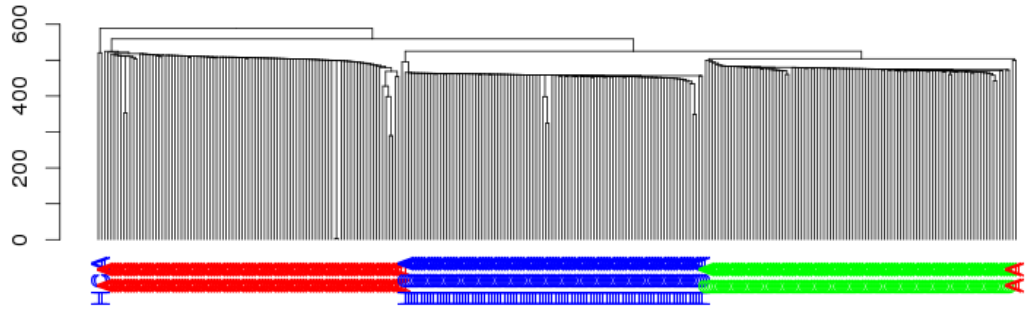
Nii geenieksрessiooni kui SNP andmeid klasterdatakse kaks korda. Esimesel korral võetakse objektide vahelise kauguse arvutamise meetodiks maksimaalse kauguse meetod, teisel korral kasutatakse keskmise kauguse meetodit (vt joonised 10-13).



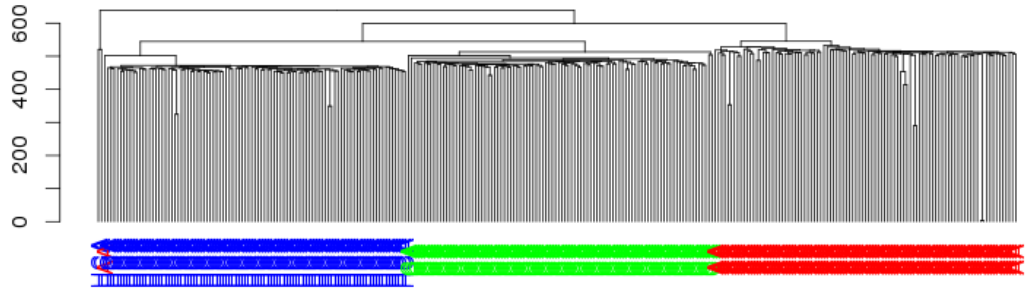
Joonis 10. Geenieksрessiooni klasterdamise tulemused, kasutades kaugusemõõduks maksimaalset kaugust.



Joonis 11. Geenieksрessiooni klasterdamise tulemused, kasutades kaugusemõõduks keskmist kaugust



Joonis 12. SNP andmete maatriksi klasterdamise tulemused, kasutades kaugusemõõduks keskmist kaugust.



Joonis 13. SNP andmete maatriksi klasterdamise tulemused, kasutades kaugusemõõduks maksimaalset kaugust.

Kuna uuritavaid inimesi on palju ja dendogrammilt on raske inimeste päritolu välja lugeda, on joonistel 10-13 märgitud iga päritolugrupp erineva värviga. Roheline värv tähistab inimest, kelle geograafiline päritolu on liigitatud, kui Europpa-ameeriklane (*Caucasian-american*). Punane värv tähistab Aafrika-ameeriklast (*African-american*) ja sinine värv Aasia-ameeriklast(*Han Chinese-American*).

Joonistelt 10-11 selgub, et geenieksрessiooni andmete järgi klasterdades on tekkinud mõned väiksemad grupid, kuhu kuuluvad sama päritoluga inimesed, ent üldpildis selget kolmeks grupiks jagunemist ei toimu. Joonistelt 12-13 ilmneb, et SNP andmete järgi klasterdades, olenemata kaugusemõõdu valikust, on tekkinud kolm gruppi. Igasse gruppi kuuluvad sama päritoluga inimesed, mõne üksiku erandiga. Selline tulemus on mõnevõrra oodatav, sest SNP-de kui DNA polümorfismide liigisisene varieeruvus võib olla suur, samas kui ulatuslik geenieksрessiooni varieeruvus ei ole võimalik. Järgnevates peatükkides kasutatakse erinevaid statistilise analüüsi meetodeid, et leida millised geenid ning SNP-d on vaadeldaval kolmel päritolugrupil erinevad.

5.2 Dispersioonanalüüsi tulemused

Geeniekspressiooni andmete statistiliseks analüüsimiseks kasutatakse ühefaktoriaalset dispersioonanalüüsi. Dispersioonanalüüs on valitud, sest kõik eeldused dispersioonanalüüsi kasutamiseks on täidetud: uuritavaid gruppe on 3, sõltuv pidev tunnus on logaritmimisega viidud normaaljaotusesse ja sõltumatu tunnus on faktortunnus.

Iga maatriksis olevat geeni/genoomi lõiku vaadeldakse eraldi ning püstitatakse kaks hüpoteesi:

- H_0: Vaadeldava geeni/genoomi lõigu ekspressiooni tase ei ole sõltuvuses päritoluga.
- H_1: Vaadeldava geeni/genoomi lõigu ekspressiooni tase on sõltuvuses päritoluga.

Olulisusenivooks määratakse $\alpha = 0.05$. Kasutades R programmi *aov* meetodit leitakse p-väärtus igale vaadeldavale geenile/genoomi lõigule. Pärast 54613 geeni/genoomi lõigu analüüsimist, arvutas R programm välja 17003 geeni/genoomi lõiku, mille puhul p-väärtus tuleb väiksem, kui paika pandud olulisusenivoo. Pärast p-väärtuste korrigeerimist *false discovery rate* meetodiga märgitakse oluliseks 11675 geeni/genoomi lõiku. Moodustatakse uus andmete maatriks, kuhu valitakse geenid/genoomi lõigud, mis on statistiliselt olulised.

5.3 Fisheri testi tulemused

Arvestades, et uuritavad andmed on diskreetse, kasutatakse SNP andmete statistiliseks analüüsimiseks Fisheri testi, et leida SNP-d, mille genotüüpide jaotus on statistiliselt olulise erinevusega populatsioonide vahel. Iga maatriksis olevat SNP-d vaadeldakse eraldi ning saadud p-väärtusi korrigeeritakse *false discovery rate* meetodiga. Iga SNP testimisel saadakse sagedustabel, mis näeb välja järgnev (vt tabel 3).

	CA	HCA	AA
AA	a	b	c
AB	d	e	f
BB	g	h	i

Tabel 3. Iga SNP põhjal loodud sagedustabeli üldkuju.

Tabelis 3 on kujutatud iga SNP kohta loodud sagedustabelit, milles ridades on erinevad SNP genotüübid. Veergudes on geograafiline päritolu (CA – Euroopa-ameeriklane, HCA-Aasia-ameeriklane, AA- Aafrika-ameeriklane).

Pärast mitmese testimise p-väärtuste korrigeerimist on statistiliselt olulisteks osutunud 102757 üksiknukleotiid polümorfismi.

5.4 Random Foresti tulemused

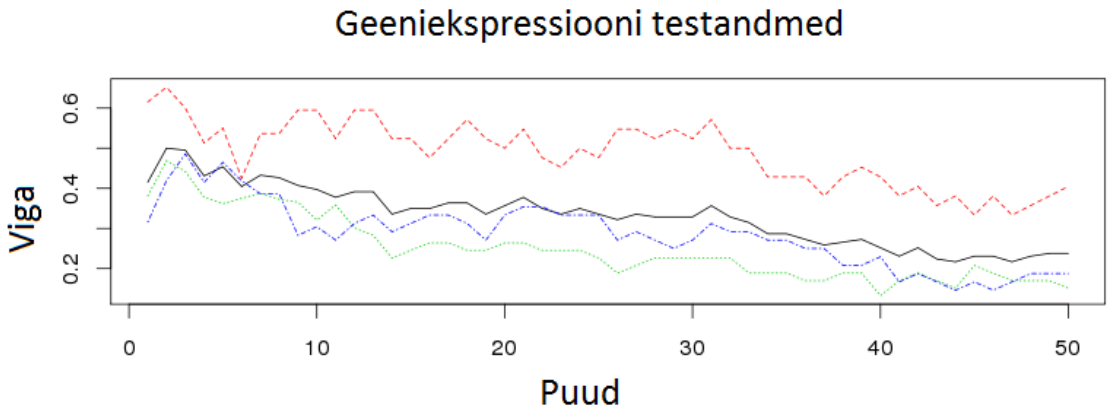
5.4.1 Geenieksпressiooni klassifitseerimise tulemused

Selles peatükis uuritakse, kuidas on võimalik klassifitseerida uuritavaid objekte, kasutades selleks peatükis 5.2 leitud olulisi geene ning peatükis 5.3 leitud olulisi SNP-sid. Klassifitseerimiseks kasutatakse masinõppe meetodit *Random Forest*, mille tööpõhimõtet on selgitatud peatükis 3.4.

Esmalt luuakse uus maatriks, kuhu jäetakse ainult nende geenide/genoomi lõikude ekspressiooni väärtused, mis osutusid olulisteks eelneva statistilise analüüsi põhjal (ANOVA). Saadud maatriks jagatakse kaheks, treeningandmete maatriksiks ja testandmete maatriksiks. Treeningandmete maatriksisse jääb 144 ning testandmete maatriksisse 143 indiviidi. Treeningandmete klassifitseerimisel saadakse *Confusion matrix*, mille kõikide gruppide keskmine OOB viga, kasutades 50 otsusepuud, on 23,78% (vt. tabel 4). Treeningandmete OOB klassifitseerimisvigade muutust loodavate puude arvu muutumisel illustreerib joonis 14.

	AA	CA	HCA	Klassi viga
AA	25	11	6	0,4048
CA	6	45	2	0,1509
HCA	6	3	39	0,1875

Tabel 4. Geenieksпressiooni treeningandmete klassifitseerimisel tekkinud *confusion matrix*



Joonis 14. Klassivea muutmine puud arvu kasvades geenieksпressiooni treeningandmete korral.

Tabelis 4 on geenieksпressiooni treeningandmete klassifitseerimisel tekkinud *confusion matrix*. Ridades on inimeste tegelik päritolu ja tulpades on päritolu, mis saadi klassifitseerimise tulemusel. Viimases tulbas on grupi klassifitseerimisel tehtud viga. AA tähistab Aafrika-ameeriklaste gruppi, CA Euroopa-ameeriklaste gruppi ning HCA Aasia-ameeriklaste gruppi. Joonis 14 kujutab geenieksпressiooni treeningandmete kasutamisel klassivea muutumist, puude arvu kasvades. Punase joonega on märgitud AA klass (aafri-

ka-ameeriklased), rohelisega HCA (aasia-ameeriklased) ning sinisega CA **(**euroopa-ameeriklased**).** Must joon tähistab kolme klassi keskmist viga. Nii tabelist kui jooniselt on kerge märgata, et AA klassi klassifitseerimisel tehtav viga on üle kahe korra suurem, kui teiste klasside klassifitseerimisel tehtav viga.

Treeningandmete põhjal loodud juhusliku metsa testimisel testandmetel saadakse tulemus, mida kujutab tabel 5.

	AA	CA	HCA	Klassi viga
AA	41	7	5	0,2264
CA	1	39	3	0,0930
HCA	1	1	46	0,0417

Tabel 5. Geenieksрessiooni testandmete klassifitseerimisel tekkinud segadustabel.

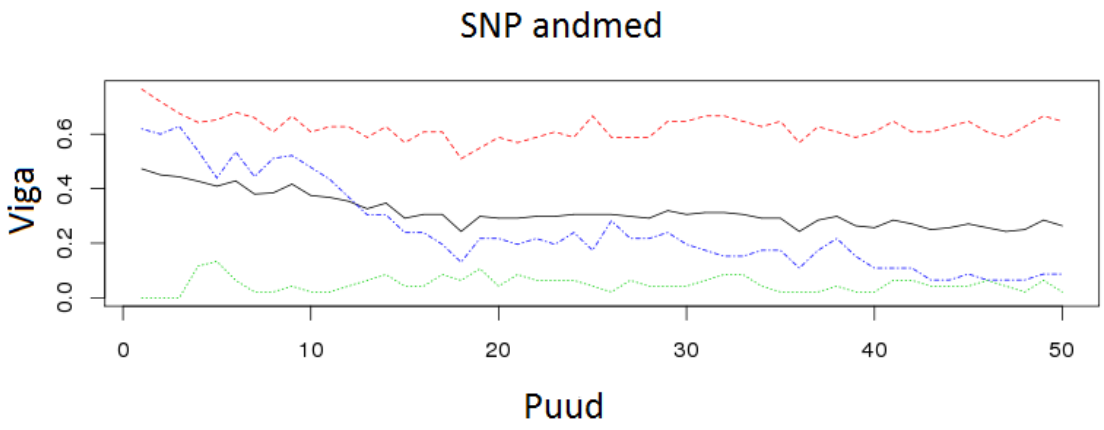
Testandmete klassifitseerimisel saadi Aafrika-ameeriklaste (AA) määramisel veaks 0,2264, Euroopa-ameeriklastel 0,093, Aasia-ameeriklastel 0,0417 ning keskmiseks klassi määramisel tehtavaks veaks 0,125. Jällegi on Aafrika-ameerika päritolu inimeste klassifitseerimisel tehtav viga umbes kaks korda suurem kui Euroopa-ameeriklaste klassifitseerimisel tehtav viga. Kõikide testandmete korral on klassifitseerimisel tehtavad vead väiksemad, kui treeningandmete puhul.

5.4.2 Üksiku nukleotiidi polümorfismi andmete klassifitseerimise tulemused

Inimeste klassifitseerimiseks üksikute nukleotiidi polümorfismi andmete järgi jaotatakse inimesed taas kahte gruppi: treeningandmed ja testandmed. Kuna inimesi on 288, siis on mõlema grupi suuruseks 144 inimest. Kuna peatükis 5.3 saadud oluliste SNP-de arv on liialt suur, et klassifitseerida kõikide oluliste SNP-de järgi, valitakse nende hulgast välja 5000, mille variatsioon on suurim. Treeningandmeid kasutades saadakse segadustabel, mille kõikide gruppide keskmine OOB viga, kasutades 50 otsusepuud, on 26,39% (vt tabel 6). Treeningandmete klassifitseerimisvigade muutust loodavate puude arvu muutumisel illustreerib joonis 15.

	AA	CA	HCA	Klassi viga
AA	18	23	10	0,6471
CA	0	46	1	0,0213
HCA	0	4	42	0,0870

Tabel 6. SNP treeningandmete klassifitseerimisel tekkinud segadustabel.



Joonis 15. SNP treeningandmete klassifitseerimisel tehtava vea muutumine puude arvu kasvades.

Tabelis 6 on polümorfismi treeningandmete klassifitseerimisel tekkinud segadustabel. Aafrika-ameeriklaste grupi vea suurus on 0,64, ehk siis ligikaudu kahe inimese puhul kolmest määratakse inimene Euroopa-ameeriklaste või Aasia-ameeriklaste gruppi. Euroopa-ameeriklaste ja Aasia-ameeriklaste puhul on tehtava vea suurus vähemalt 6 korda väiksem.

Joonisel 15 on kujutatud klassifitseerimisel tehtava vea muutmist puude arvu kasvades. Roheline joon tähistab Euroopa-ameeriklasi, sinine Aasia-ameeriklasi ning punane joon Aafrika-ameeriklasi. Musta joonega on kujutatud kolme grupi keskmist viga. Ka jooniselt 15 on näha, et hea tulemus on saadud Euroopa-ameeriklaste ja Aasia-ameeriklaste klassifitseerimisel. Aafrika-ameeriklaste puhul on valesti klassifitseeritud peaaegu 2 korda enam inimesi kui õigesti (vastavalt 18 ja 33 inimest). Treeningandmete põhjal loodud juhuslikus metsas testandmeid klassifitseerides saadud tulemust väljendab tabel 7.

	AA	CA	HCA	Klassi viga
AA	24	15	6	0,4667
CA	0	49	0	0
HCA	0	0	50	0

Tabel 7. SNP testandmete klassifitseerimisel saadud segadustabel.

Testandmete klassifitseerimisel eksiti umbes poolte Aafrika-ameeriklaste puhul. Vastavalt saadud viga oli 0,4667. See on küll väiksem, kui treeningandmete korral, aga siiski liialt suur, et lugeda klassifitseerimist edukaks. Euroopa-ameeriklaste ja Aasia-ameeriklaste puhul klassifitseeriti kõik testandmestikus olevad inimesed õigesti. Sarnaselt geenieksspressiooni tulemustele on ka SNP-de puhul näha, et Aafrika-ameeriklaste klassifitseerimisel tehtav viga on oluliselt suurem, kui Euroopa- või Aasia-ameeriklaste puhul. Näha on ka, et SNP-de puhul on klassifitseerimisel tehtav viga Euroopa- ja Aasia-ameeriklaste korral väiksem, kui geenieksspressiooni kasutades. Aafrika-ameeriklaste puhul on SNP-de põhjal klassifitseerimine ebatäpsem, kui geenieksspressiooni kasutades. See võib tuleneda

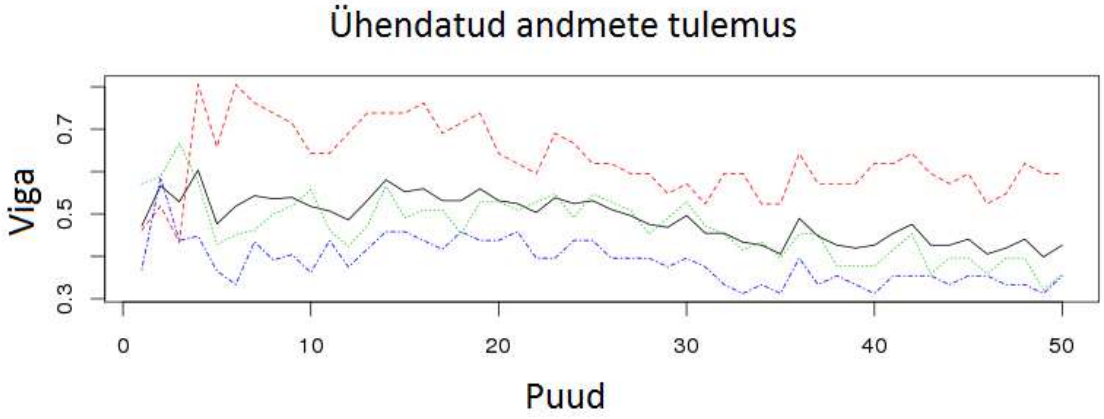
sellest, et geeniekspressiooni andmestikus oli tunnuseid umbes 2 korda rohkem, kui polümorfismi andmestikus.

5.4.3 Geeniekspressiooni ja SNP-de ühise klassifitseerimise tulemused

Eelmistes peatükkides avastati, et geeniekspressiooni andmete kasutamisel on Aafrika-ameeriklaste klassifitseerimisel tehtav viga väiksem, kui polümorfismi andmete kasutamisel. Euroopa- ja Aasia-ameeriklaste klassifitseerimisel on seis vastupidine. Nüüd proovitakse inimesi klassifitseerida, kasutades nii geeniekspressiooni kui ka polümorfismi andmeid. Selleks valime välja oluliste geenide hulgast 2500 geeni, mille standardhälve on suurim. Sama teeme ka saadud oluliste SNP-dega. Valiku geenide ja SNP-de hulgast tehti, sest *Random Forest* kasutamine 102757 tunnusega võtab liiga kaua aega ja tekib oht saada väga keerulisi ning üleõpitud puid. Suurima standardhälbega tunnused valiti, sest nende järgi peaks klassifitseerimine tulema välja kõige paremini. Taaskord jagatakse inimesed kahte gruppi: treeningandmed ja testandmed. Treeningandmete gruppi paigutatakse 143 inimest ning testandmete gruppi 144 inimest. Treeningandmete põhjal moodustades juhusliku metsa 50 puuga, saadakse segadustabel, mille OOB viga on 42,66%. (vt tabel 8). Treeningandmete klassifitseerimisvigade muutust loodavate puude arvu muutumisel illustreerib joonis 16.

	AA	CA	HCA	Klassi viga
AA	17	19	6	0,5952
CA	12	34	7	0,3585
HCA	8	9	31	0,3542

Tabel 8. Treeningandmete klassifitseerimisel tekkinud segadusmaatriks



Joonis 16. Klassivea muutumine puude arvu kasvades, ühendatud treeningandmete korral.

Joonisel 16 on kujutaud klassifitseerimisel tehtava vea muutumist puude arvu kasvades. Roheline joon tähistab Aasia-ameeriklasi, sinine joon Euroopa-ameeriklasi, punane joon Aafrika-ameeriklasi ning must joon kolme grupi keskmist. Ka ühendatud andmete korral on Aafrika-ameeriklaste klassifitseerimisel eksitud kõige rohkem. Euroopa- ja Aasia-

ameeriklaste puhul on eksitud umbes iga kolmanda inimesega. Testandmete põhjal loodud juhuslikus metsas testandmeid klassifitseerides saadud tulemust väljendab tabel 9**.**

	AA	CA	HCA	Klassi viga
AA	38	10	5	0,2830
CA	6	31	6	0,2791
HCA	2	6	40	0,1667

Tabel 9. testandmete klassifitseerimisel saadud segadustabel

Testandmete klassifitseerimisel on gruppide keskmiseks määramisel tekkinud veaks 24,31% ehk siis umbes iga neljas inimene määratakse valesti. Kõige suurem viga tekib Aafrika-ameeriklaste puhul ning väikseim taaskord Aasia-ameeriklaste puhul. Tähele tasub ka panna, et Aafrika-ameeriklaste puhul on tekkiv viga suurem kui ainult geeniekspressiooni kasutades, aga väiksem, kui ainult SNP-sid kasutades. Euroopa- ja Aasia-ameeriklaste puhul on tekkinud viga suurem, kui kumbagi andmestikku eraldi kasutades.

5.4.4 Järeldused

Random forest'i algoritmi kasutati kolme erineva mudeli peal. Esmalt ekspressioonianmetel, seejärel polümorfismi andmetel ning lõpuks nii ekspressiooni kui polümorfismi andmetel. Ekspressioonimudelisse oli valitud tunnusteks kõik statistiliselt oluliseks osutunud geenid/genoomi lõigud. Kuna statistiliselt olulisi polümorfisme oli palju, valiti polümorfismi mudelisse oluliste hulgast 5000, mille variatsioon oli kõige suurem. Viimasesse mudelisse valiti leitud oluliste tunnuste hulgast 2500 suurima variatsiooniga geeni ningi 2500 suurima variatsiooniga polümorfismi.

Võrreldes kahte esimest mudelit, saab järeldada, et mõlema mudeli klassifitseerimise tulemused Euroopa-ameeriklaste ja Aasia-ameeriklaste puhul on sarnaselt head (klassifitseerimise viga on nullilähedane). Aafrika-ameeriklased klassifitseeriti ekspressiooni mudelit kasutades peaaegu kaks korda täpsemalt.

Kolmanda ehk ühismudeli klassifitseerimisviga jaotus populatsioonide vahel ühtlasemalt, see kajastus Euroopa-ameeriklaste ning Aasia-ameeriklaste klassifitseerimisvea suurenemisel ning Aafrika-ameeriklaste klassifitseerimisvea langemisel, võrreldes kahe eelmise mudeliga.

6. Kokkuvõte

Pärilikkuse ja DNA uurimisega on viimasel sajandil tegeletud väga intensiivselt. Käesolev töö keskendus seoste otsimisele geenide avaldumise, DNA järjestuse ning geograafilise päritolu vahel. Olles leidnud statistiliselt olulised geenid (11665 54613-st) ja polümorfismid (102757 511354-st), prooviti inimese klassifitseerida nende järgi, kasutades selleks *random forest* meetodit.

*Random forest'*i algoritmi implementeeriti kolme erineva mudeli peal: esimene mudel sisaldas ainult ekspressiooniandmeid (kõik statistiliselt olulisteks osutunud geenid/genoomi lõigud), teine mudel sisaldas 5000 (102757 statistiliselt olulise seast) polümorfismi, mille variatsioon oli kõige suurem ning kolmas mudel ühendas nii ekspressiooni kui ka polümorfismi andmeid, kus ekspressiooni andmetest valiti 2500 geeni ning polümorfismi andmetest valiti 2500 lookust. Mõlemal juhul valiti kõige suurema variatsiooniga tunnused.

Võrreldes kahte esimest mudelit, saab järeldada, et mõlema mudeli klassifitseerimise tulemused Euroopa-ameeriklaste ja Aasia-ameeriklaste puhul on sarnaselt head (klassifitseerimise viga on nullilähedane). Aafrika-ameeriklased klassifitseeriti ekspressiooni mudelit kasutades peaaegu kaks korda täpsemalt.

Kolmanda ehk ühismudeli klassifitseerimisviga jaotus populatsioonide vahel ühtlasemalt, see kajastus Euroopa-ameeriklaste ning Aasia-ameeriklaste klassifitseerimisvea suurenemisel ning Aafrika-ameeriklaste klassifitseerimisvea langemisel, võrreldes kahe eelmise mudeliga.

Antud analüüsi põhjal pole võimalik kindlalt eelistada ühte mudelit teistele. Üks võimalik seletus on, et mudelid pole tunnuste arvu poolest võrreldavad ning tunnuste arv on märkimisväärselt suurem kui valimi maht. Järgmise sammuna võiks proovida valida väiksema tunnuste arvu, põhinedes variatsioonile.

7. Tsiteeritud teosed

[1] N. K. Hidayat, „What is Three Parts of Nucleotide ?,“ 13 April 2013. [Võrgumaterjal]. Available: http://dnarnanews.blogspot.com/2013/04/what-is-three-parts-of-nucleotide.html. [Kasutatud 14 May 2015].

[2] „Nucleotides and Bases,“ [Võrgumaterjal]. Available: http://knowgenetics.org/nucleotides-and-bases/. [Kasutatud 14 May 2015].

[3] T. Maimets, Molekulaarne rakubioloogia, Tartu: Ilmamaa, 1999.

[4] [Võrgumaterjal]. Available: http://www.siriusgenomics.com/technology/. [Kasutatud 14 May 2015].

[5] T. Hastie, R. Tibshirani ja J. Friedman, The Elements of Statistical Learning, New York: Springer, 2009.

[6] K. Niglas, „Dispersioonanalüüsi õppematerjal,“ November 2013. [Võrgumaterjal]. Available: http://www.cs.tlu.ee/~katrin/wp/wp-content/uploads/2013/11/dispersioon.pdf. [Kasutatud 25 April 2015].

[7] A.-M. Parring, M. Vähi ja E. Käärik, Statistilise andmetöötluse algõpetus, Tartu: Tartu Ülikooli Kirjastus, 1997.

[8] J. H. McDonald, Handbook of biological statistics, Baltimore, Maryland, U.S.A.: Sparky House Publishing, 2014, pp. 77-86.

[9] A. Loos, „Machine Learning for k-in-a-row Type Games,“ Tartu Ülikool, Tartu, 2012.

[10] G. James, D. Witten ja T. Hastie, An Introduction to Statistical Learning with Applications in R, New York: Springer, 2013.

[11] I. Traat, Matemaatilise statistika põhikursus, Tartu: Tartu Ülikooli kirjastus, 2006.

[12] V. Ilakovac, „Statistical hypothesis testing and some pitfalls,“ [Võrgumaterjal]. Available: http://www.biochemia-medica.com/content/statistical-hypothesis-testing-and-some-pitfalls. [Kasutatud 14 May 2015].

[13] D. Y. Yoav Benjamini, „The Control Of The False Discovery Rate In Multiple Testing Under Dependency,“ Tel Aviv University, Tel Aviv, 2001.

[14] T. V. Perneger, „What's wrong with Bonferroni adjustments,“ *BMJ,* pp. 1236-1238, 1998.

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina **Madis Kaasik** (sünnikuupäev: 18.05.1992)
(*autori nimi*)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **Geograafilise päritolu ennustamine geeniekspressiooni ja geneetilise varieeruvuse abil**,
(*lõputöö pealkiri*)

mille juhendajad on Tauno Metsalu, Tatjana Iljašenko
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **14.05.2015**