Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Mirian Valk

Eesti geneetiliste emaliinide mitmekesisus, struktuur ja võrdlus valitud naaberpopulatsioonidega

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Märt Möls,
PhD Anne-Mai Ilumäe

Tartu 2024

EESTI GENEETILISTE EMALIINIDE MITMEKESISUS, STRUKTUUR JA VÕRDLUS VALITUD NAABERPOPULATSIOONIDEGA

Bakalaureusetöö

Mirian Valk

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk on Eesti geenivaramu mitokondriaalse DNA andmete põhjal uurida Eesti maakondade geneetilist sarnasust omavahel ja kolme naaberpopulatsiooniga. Esmalt antakse ülevaade mitokondriaalsest DNA-st ning tutvustatakse töös kasutatud statistilisi meetodeid: korrespondentsanalüüs, χ^2 test, Crameri V test. Seejärel uuritakse Eesti maakondade omavahelist geneetilist sarnasust, sarnasust naaberpopulatsioonidega (Soome, Rootsi, Poola) ning Euraasia idaosale omaste geneetiliste emaliinide esinemist Eestis. Viimasena testitakse kasutatud statistilisi meetodeid simulatsioonvalimitel, et teha kindlaks kui hästi need töö alusandmestikus leiduvad seoseid kirjeldavad.

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

Märksõnad: mitokondriaalne DNA, geenivaramu, korrespondentsanalüüs.

DIVERSITY AND STRUCTURE OF ESTONIAN MATRILINEAR GENES AND THEIR COMPARISON TO SELECTED NEIGHBOURING POPULATIONS

Bachelor thesis

Mirian Valk

Abstract

The aim of this bachelor's thesis is to investigate the genetic similarity between Estonian counties based on mitochondrial DNA data from the Estonian Biobank, as well as compare them with three neighbour populations. Firstly, an overview of mitochondrial DNA is provided, and the statistical methods utilized in this study are introduced: correspondence analysis, χ^2 test, Cramér's V test. Then, the genetic similarity between Estonian counties is examined, along with their resemblance to neighbour populations (Finland, Sweden, Poland). The presence of maternal genetic lineages characteristic of Eastern Eurasia in Estonia is also analysed. Finally, the statistical methods are tested on simulation datasets to determine how well they describe the associations present in the underlying data.

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

Key Words: mitochondrial DNA, biobank, correspondence analysis.

Sisukord

Sissejuhatus

Geenid on infoallikas uurimaks kaasaja populatsioonide demograafilist kujunemist. Mitokondriaalne DNA on osa inimese genoomist, mis päritakse ainult emalt. Mitokondriaalse DNA abil saab uurida, kuidas on erinevate populatsioonide naisliinid omavahel seotud nii erinevate populatsioonide vahel kui populatsioonis endas. Kuna mitokondriaalne DNA päritakse vaid ühelt vanemalt, siis see muutub ajas vaid mutatsioonide tekkimisel (ei toimu kahe vanema geenide kombineerumist nagu nukleotiidse DNA puhul), mistõttu saab selle abil uurida täpsemaid ajaloolisi geneetilisi hargnemisi emaliinis.

Kuna Eesti on võrdlemisi väike populatsioon, siis käsitletakse seda populatsioonigeneetilistes uuringutes enamasti ühe populatsioonina ja ei uurita Eesti eri piirkondi eraldi. Käesoleva töö eesmärk on uurida emaliinide struktuuri ja mitmekesisust Eesti maakondades, et hinnata kas Eesti on geneetiliselt homogeenne või on piirkondi, mis teistest erinevad. Kasutatakse Eesti geenivaramu suurandmestikku, kus on pea 50 000 eestlase mitkondriaalse DNA andmed (*Eesti geenivaramu* 2021). Uuritakse ka maakondade populatsioonide geneetilist sarnasust naaberpopulatsioonidega (Soome, Rootsi ja Poola), mille puhul on teada ajaloolised seosed Eesti aladega ning mille mitkondriaalse DNA liinide sageduste andmed on avalikult kättesaadavad. Populatsioonide võrdlemiseks kasutatakse korrespondentsanalüüsi ja Crameri V seosekordajat.

Töö esimeses osas antakse teoreetiline ülevaade mitokondriaalsest DNA-st ning kasutatud matemaatilistest meetoditest. Teises osas viiakse läbi andmeanalüüs ning testitakse kasutatud statistilisi meetodeid simulatsioonandmestikel, et hinnata kui hästi need töös kasutatud andmeid kirjeldavad. Viimasena kirjeldatakse analüüsitulemusi ning tehakse järeldused.

1 Teoreetiline taust

1.1 Mitokondriaalne DNA

Mitokondrid on topeltmembraaniga rakuorganellid, mis esinevad kõigis imetajate rakkudes. Need täidavad mitmeid olulisi funktsioone, millest tuntuim on ATP ehk adenosiintrifosfaadi, rakkude energiaallika tootmine. Mitokondritel on oma genoom, mida kutsutakse mitokondriaalseks DNA-ks (mtDNA). Inimese mtDNA on kinnine ring, pikkusega umbes 16 600 aluspaari. (Bandelt, Macaulay ja Richards, 2006, peatükk 1.1)

Olenevalt rakutüübist on ühes rakus mitu mitokondrit ja igas mitokondris mitu mtDNA molekuli. Seega võib ühes rakus olla umbes 1100 $\pm$ 250 mtDNA molekuli (Bogenhagen ja Clayton, 1974).

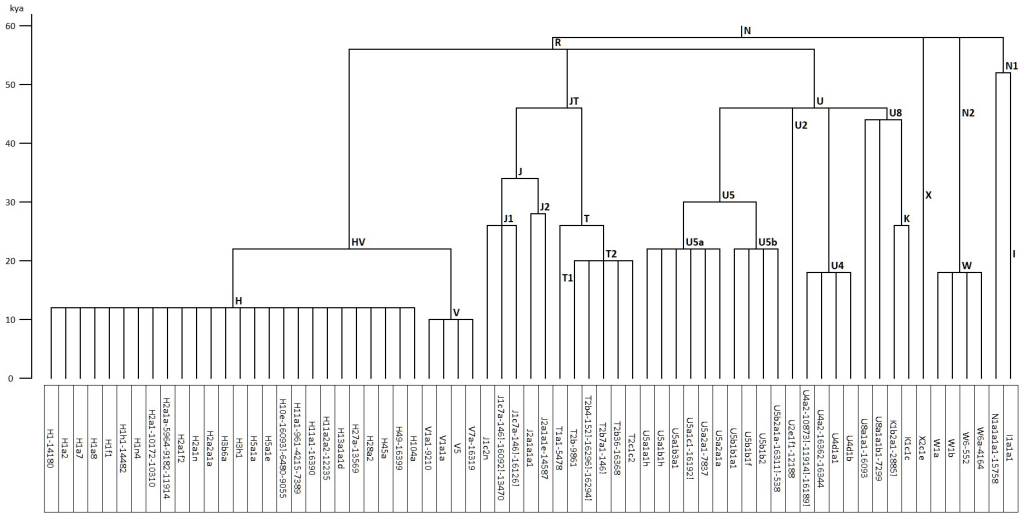
Mitokondriaalne DNA päritakse ainult emalt (Giles *et al.*, 1980). Kuigi ei olda veel täpselt kindlad, miks isalt mtDNA-d ei pärita, siis on teada, et viljastumise hetkel on munarakus rohkem kui 150 000 mtDNA molekuli, samas kui spermis on vaid sadakond (Wai *et al.*, 2010). Loote arengu käigus niigi vähemuses olevad isapoolsed mtDNA molekulid kõrvaldatakse.

Ühes rakus võib esineda erineva genoomiga mtDNA molekule - emalt päritud variandid ja muteerunud variandid. Muteerunud mtDNA hulk rakus võib suureneda raku pooldumise tagajärjel, kui ühte tütarrakku satub rohkem muteerunud kui terveid mtDNA molekule. Kuna mtDNA replitseerub ja hävineb ka sõltumata raku pooldumise tsüklist, siis võib muteerunud mtDNA hulk ka selle käigus rakus muutuda (Stewart ja Chinnery, 2015; Bandelt, Macaulay ja Richards, 2006, peatükk 1.3). Seega toimub mtDNA muteerumine sagedamini kui rakutuuma DNA (nDNA) muteerumine, kusjuures mtDNA mutatsioonide kiirus hinnatakse olevat umbes 10 korda suurem, kui nDNA puhul (Brown, Jr ja C.Wilson, 1979). Seetõttu saab mtDNA abil uurida hiljutisemaid geneetilisi hargnemisi evolutsioonipuul.

1.2 Haplogrupid ja fülogeneesipuu

Mitokondriaalse DNA klassifitseerimiseks määratakse neile haplogrupp. Iga indiviidi mtDNA molekuli iseloomustab teatud mutatsioonide kogum. Iga teatud mtDNA mutatsioonide kogumile on määratud haplogrupp. Haplogruppi tähistatakse suure tähega, selle järgneb alamhaplogruppi tähistav number ja nii edasi tähestikulises järjestuses, kus tähed vahelduvad numbritega. Vajadusel lisatakse ka selgitav mutatsiooninumber.

mtDNA evolutsioneerumist kujutatakse fülogeneesipuul (joonis 1), mis kirjeldab emaliinide hargnemisi haplogruppidena. Fülogeneesipuult saab näha, millised haplogrupid pärinevad samast järjestusest ja mitu hargnemist põhjustanud mutatsiooni on toimunud alates nende viimasest ühisest esivanemast emaliinis. Mutatsioonide arvu järgi saab hinnata ka evolutsiooniliste hargnemiste ajalist kulgu. Kuna molekulaarse evolutsiooni kiirus on võrdlemisi püsiv ehk mutatsioonid toimuvad DNA molekulides üsna konstantse kiirusega, siis on võimalik uurida kui ammu toimusid mtDNA järjestuste hargnemised. Seda töövõtet tuntakse molekulaarse kellana (*Geneetika sõnastik* 2014). Selle abil saab uurida ka erinevate populatsioonide omavahelist geneetilist sarnasust ja millal populatsioonides levinud emaliinid evolutsioonilises mõttes lahku läinud.



Joonis 1: Eestis, Soomes, Rootsis ja Poolas levinud mtDNA haplogruppide fülogeneesipuu. Vertikaalne skaala näitab kui ammu haplogruppide hargnemine aset leidis (1 kya = 1000 aastat tagasi, (kya = kilo years ago)). Joonise autor Dr. Maere Reidla, Tartu Ülikooli genoomikainstituut.

2 Metoodika

2.1 Korrespodentsanalüüs

Korrespondentsanalüüs on meetod, mille eesmärk on kujutada sagedustabeli ridade ja veergude väärtused madala dimensiooniga, enamasti kahe- või kolmemõõtmelisele graafikule. Selleks leitakse vastava dimensiooniga alamruum, mis paikneks tabeli punktidele võimalikult lähedal. Meetodi rakendamisel läheb osa tabelis olevast informatsioonist kaduma, kuid meetodi eesmärk on, et kaduma läheks võimalikult vähe informatsiooni. Selline andmete esitus aitab näha tabeli tunnuste ja tunnuste väärtuste vahelisi seoseid. See meetod on pigem geomeetriline kui statistiline, seega sobiv andmetes leiduvatele seostele esialgse hinnangu andmiseks ja hüpoteeside genereerimiseks. Meetodi eelduseks on, et kõik tabelis olevad arvud on mittenegatiivsed. (peatükk 3 Greenacre, 1984)

2.1.1 Meetodi ülevaade

Alapetükk on kirjutatud Kalev Pärna uurimuse põhjal (Pärna, 1993).

Olgu antud n vaatlust, mis on klassifitseeritud tunnuste A ja B põhjal. Tunnusel A olgu I kategooriat A_1, A_2,...,A_I ja tunnusel B olgu J kategooriat B_1,B_2,...,B_J. Tähistame $n_{i.}$-ga nende vaatluste arvu, mille tunnuse A väärtuseks on A_i ja $n_{.j}$-ga vaatluste arvu, mille tunnuse B väärtuseks on B_j. Vaatluste arvu, millel on väärtused A_i ja B_j samaaegselt, tähistame n_{ij}. Saame andmed esitada $I \times J$ maatriksina N, kus i-rea j-veeru elemendi väärtuseks on n_{ij}. Tabeli ridade ja veergude marginaalsagedused ning vaatluste kogusumma tähistame vastavalt

$$n_{i.} = \sum_j n_{ij} \qquad n_{.j} = \sum_i n_{ij} \qquad n = \sum_i \sum_j n_{ij}.$$

Maatriksi N põhjal saab leida järgmised suhtelised sagedused:

$f_{ij} = \frac{n_{ij}}{n} \approx P(A_i B_j)$, A_i ja B_j väärtustega vaatluste osakaal kõigist vaatlustest,

$f_i = \frac{n_{i.}}{n} \approx P(A_i)$, A_i väärtusega vaatluste osakaal A väärtuste seas,

$f_j = \frac{n_{.j}}{n} \approx P(B_j)$, B_j väärtusega vaatluste osakaal B väärtuste seas,

$f_j^i = \frac{n_{ij}}{n_{i.}} \approx P(B_j|A_i)$, B_j väärtusega vaatluste vaatluste osakaal i-reas,

$f_i^j = \frac{n_{ij}}{n_{.j}} \approx P(A_i|B_j)$, A_i väärtusega vaatluste osakaal j-veerus.

Veerutunnuse B tinglikku jaotust tingimusel, et $A = A_i$ nimetame (i-nda rea) reaprofiiliks ja tähistame vektoriga $f_B^i = (f_1^i,...,f_J^i)^T$. Sarnaselt tähistame ka reatunnuse A tinglikku jaotust, kui $B = B_j$ vektorina $f_A^j = (f_1^j,...,f_I^j)$ ja nimetame (j-nda veeru) veeruprofiiliks.

Näeme et kokku on I reaprofiili, mis asuvad J-dimensionaalses eukleidilises ruumis. Reaprofiilide hulka nimetame pilveks I-punktist ruumis R^J. Veeruprofiilide hulka, mis koosneb J-punktist I-dimensionaalses ruumis nimetame veeruprofiilide pilveks ruumis R^I. Tähistame rea- ja veeruprofiilide pilved vastavalt

$$N_B(A) = \{f_B^i | i = 1,2,...,I\}$$
$$N_A(B) = \{f_A^j | j = 1,2,...,J\}.$$

Igal punktil reaprofiilide (veeruprofiilide) pilves on mass, mille määrab vastava rea (veeru) marginaaltõenäosus f_i (f_j). Seega koosneb pilv kaalutud punktidest. Defineerime pilvede keskpunktid vastava pilve massikeskmena kujul

$$f_B = \sum_i f_i \cdot f_B^i \text{ , reaprofiilide pilve keskpunkt,}$$
$$f_A = \sum_j f_j \cdot f_A^j \text{ , veeruprofiilide pilve keskpunkt.}$$

Näeme, et f_B on võrdne veergude marginaaljaotustega ja f_A on võrdne ridade

marginaaljaotusega. Seega saame pilvede keskpunktid esitada kujul

$$f_B = (f_1,...,f_J)^T$$
$$f_A = (f_1,...,f_I).$$

Mõlema punktipilve puhul on dimensionaalsus K määratud valemiga

$$K = \text{rank}(N) - 1 \leq \min\{I-1,J-1\}.$$

Korrespondentsanalüüsi graafiku jaoks sobiv madala dimensionaalsusega alamruum peab paiknema kõikidele pilve punktidele võimalikult lähedal. Kuna punktidel on erinevad massid (olenevalt pilvest kas f_i või f_j), siis peaks alamruum paiknema lähemal suurema massiga punktidele.

Vaatame nüüd ainult reaprofiilide pilve $N_B(A)$. Elemendi f_B^i kaugus pilve keskpunktist defineeritakse hii-ruut kaugusena kujul

$$d^2(i) = \sum_j \frac{(f_j^i - f_j)^2}{f_j}.$$

Seda kaugust nimetatakse ka kaalutud eukleidilise kauguse ruuduks.

Pilve $N_B(A)$ inertsiks nimetatakse I reaprofiili kaalutud keskmist kaugust pilve keskpunktist, mis näitab profiilide hajuvust keskpunkti ümber ja leitakse valemiga

$$in(A) = \sum_i f_i \cdot d^2(i).$$

Veeruprofiilide pilve $N_A(B)$ inerts leitakse sarnaselt, valemiga

$$in(B) = \sum_j f_j \cdot d^2(j),$$

kus

$$d^2(j) = \sum_i \frac{(f_i^j - f_i)^2}{f_i}.$$

Mõlema pilve inertsid saab avaldada kujul

$$\text{in}(A) = \text{in}(B) = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{n_{i.}n_{.j}} = \frac{\chi^2}{n} = \lambda.$$

Otsitav madaladimensiooniline alamruum on määratud inertsi peatelgedega. Peatelgedeks on K vektorit, mida rakendatakse pilve keskpunktile ja need näitavad pilve suurimate inertside suunda. Peateljed järjestatakse suurimast inertsist väiksemani ja iga järgnev telg on kõigi eelnevate suhtes ortogonaalne. Korrespondentsanalüüsis valitakse enamasti kaks esimest peatelge, et kujutada rea- ja veeruprofiilid kahemõõtmelisele graafikule. Peatelgede leidmiseks kasutatakse singulaarväärtuslahutust, mida antud töös lähemalt ei tutvustata. Kuna $\text{in}(A) = \text{in}(B)$, siis saame mõlemad pilved kujutada samal graafikul.

2.1.2 Graafikute tõlgendamine

Samasse pilve kuuluvate punktidevahelised kaugused on hinnangulised χ^2 kaugused vastavate profiilide vahel. Kui samale graafikule on kujutatud nii rea- kui veerutunnuste punktid, siis nendevahelistele kaugustele ei saa otseselt omistada sarnasust nende väärtuste vahel, aga saab vaadata kui suur on nurk kahe erinevast pilvest pärit punkti vahel. Kui nurk on väike, siis need tunnuste väärused on omavahel seotud, kui nurk on suur, siis ei ole seotud. Seega paiknevad sarnase profiiliga tunnused joonisel samas piirkonnas ning omavahel erinevad tunnused paiknevad joonisel teineteise suhtes vastaspoolel. (peatükk 9 Greenacre, 1993) Samuti saab joonisel värvide abil näidata, kui hästi valitud kahe telje poolt moodustatud tasand pilve kuuluvate tunnuste profiile kirjeldab.

2.2 Hii-ruut test ja Crameri V

Hii-ruut test on statistiline test, mis hindab kas sagedustabeli rea- ja veerutunnus on sõltuvad või sõltumatud. Testi kasutatakse suurte valimite ja kategooriliste tunnustega tabelitel. Testi puhul eeldatakse, et tunnused on omavahel sõltumatud. Püstitatakse nullhüpotees (H_0) ja alternatiivne hüpotees (H_1):

H_0 : Uuritavad tunnused on omavahel sõltumatud.

H_1 : Uuritavad tunnused on omavahel sõltuvad.

Olgu tabeli reatunnuseks A, millel on i = 1,2,...,I erinevat väärtust ja veerutunnuseks B, millel on j = 1,2,...,J erinevat väärtust. Kui A ja B on sõltumatud, siis iga väärtuspaari esinemissagedus peaks olema võimalikult lähedane oodatavale esinemissagedusele. Sagedustabeli väärtuspaari $(A = a_i, B = b_j)$ oodatav esinemissagedus avaldub kujul

$$E_{ij} = \frac{n_{i.} n_{.j}}{n},$$

kus $n_{i.}$ on tabeli i-rea marginaalsagedus, $n_{.j}$ j-veeru marginaalsagedus ja n on vaatluste koguarv. Hii-ruut statistik arvutatakse valemiga

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

kus O_{ij} on sagedustabeli ij lahtris olev väärtus ehk väärtuspaari $(A = a_i, B = b_j)$ tegelik esinemissagedus valimis.

Nullhüpoteesi kehtimisel järgib χ^2 statistik hii-ruut jaotust, vabadusastmetega $(I-1)\cdot(J-1)$. Vastavalt ülesandele valitakse olulisusnivoo α. Saadud teststatistiku väärtust võrreldakse vastava hii-ruut jaotusega ja leitakse p-väärtus: kui $p < \alpha$, siis jäädakse nullhüpoteesi juurde, kui $p > \alpha$, siis võetakse vastu alternatiivne hüpotees.

Crameri V on statistiline seosekordaja, mis põhineb hii-ruut statistikul ja hindab

kahe nominaalse tunnuse vahelise seose tugevust. Meetodi arendajaks on Harald Cramer (peatükk 21, lk 282 Cramer, 1946). Crameri V väärtused on vahemikus 0 kuni 1, kus väärtus 0 näitab et tunnuste vahel ei ole seost ja 1 näitab et on tugev seos.

Crameri V arvutatakse järgneva valemiga:

$$V = \sqrt{\frac{\chi^2/n}{\min\{I-1, J-1\}}},$$

kus I on ridade arv, J on veergude arv ja n on tabeli elementide kogusumma ning χ^2 on kujul 1.

3 Andmete analüüs

Analüüside läbiviimiseks kasutati rakendustarkvara R (versioon 4.2.2).

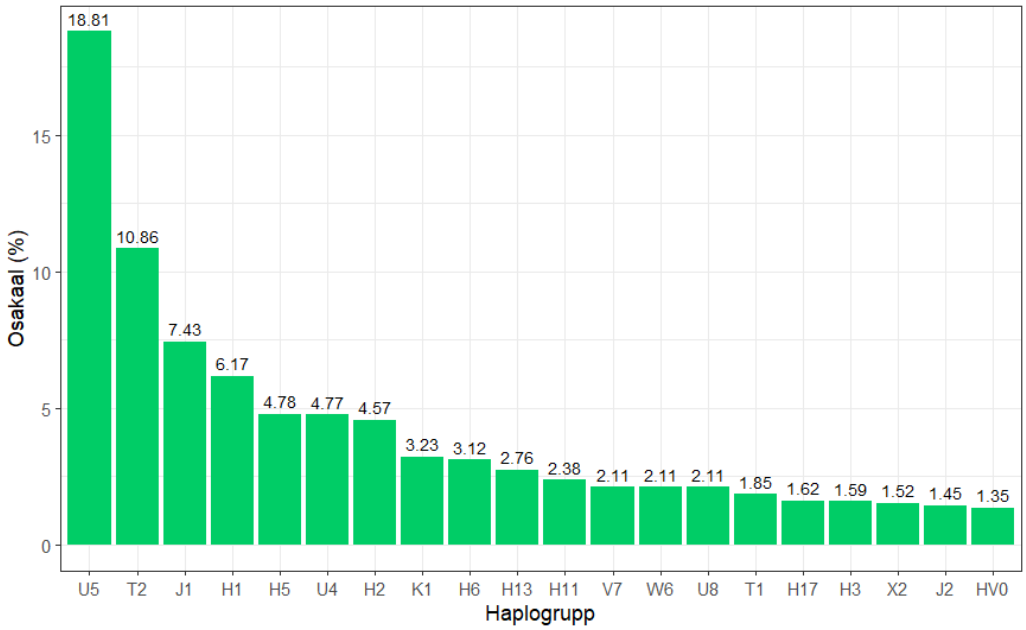
3.1 Andmestiku kirjeldus

Töös kasutatud andmed pärinevad Eesti Geenivaramu suurandmestikust, mis sisaldab üle 200 000 Eesti elaniku genotüübiinfot. (*Eesti geenivaramu* 2021). Kasutatud on 48 628 inimese mtDNA andmed, kes on geenidoonori küsimustikus end rahvuselt eestlaseks märkinud. Kõik mtDNA-d on määratud haplogruppidesse fülogeneentiliselt kõrgeima võimaliku haplogrupi täpsusega. Saadud andmestik koosneb 269st haplogrupist ja nende esinemissagedustest Eesti maakondades (Lisa 1). Eraldi on välja toodud sagedused Tartus ja Tallinnas. Maakonnad on määratud geenidoonori poolt doonori ankeedis märgitud sünnikoha põhjal ja üldistatud vastava maakonna tasemele, välja arvatud Tartu ja Tallinna puhul. Lisaks kaasati ka Eestis esinevate haplogruppide sagedused Soomes, Rootsis ja Poolas (Poola andmed: Kaja *et al.*, 2022; Rootsi andmed: Sturk-Andreaggi *et al.*, 2022; Soome andmed: Clark *et al.*, 2021). Proovide arvud piirkondade kaupa on välja toodud tabelis 1. Naaberpopulatsioonide puhul on sulgudes märgitud saadaval olevate proovide arv, sulgudest väljas on nende proovide arv, mille haplogruppe esines ka Eestis. Naaberpopulatsioonide ülejäänud haplogruppid märgiti kui "Muu".

Tabel 1: mtDNA proovide arvud maakondades, linnades ja riikide populatsioonides (mk=maakond).

Harju mk	Lääne-Viru mk	Ida-Viru mk	Jõgeva mk	Järva mk	Rapla mk	Pärnu mk	Lääne mk	Saare mk	Hiiu mk
5563	3446	1812	1945	1787	1507	4383	1100	1679	534
Viljandi mk	Tartu mk	Põlva mk	Võru mk	Valga mk	Tartu	Tallinn	Soome	Rootsi	Poola
2992	4310	1594	2433	1634	4363	7546	1211(906)	1179(792)	1616(1065)

Kõige rohkem leidus Eestis haplogruppe U5, T2, J1, H1, H5 ja nende alamliine. See on kooskõlas Euroopas esinevate haplogruppide sagedustega. (Richards, Macaulay ja H. J. Bandelt, 2002; Simoni *et al.*, 2000, tabel 2). Eestis enamlevinud 20 haplogruppi ja nende osakaalud on esitatud joonisel 2.



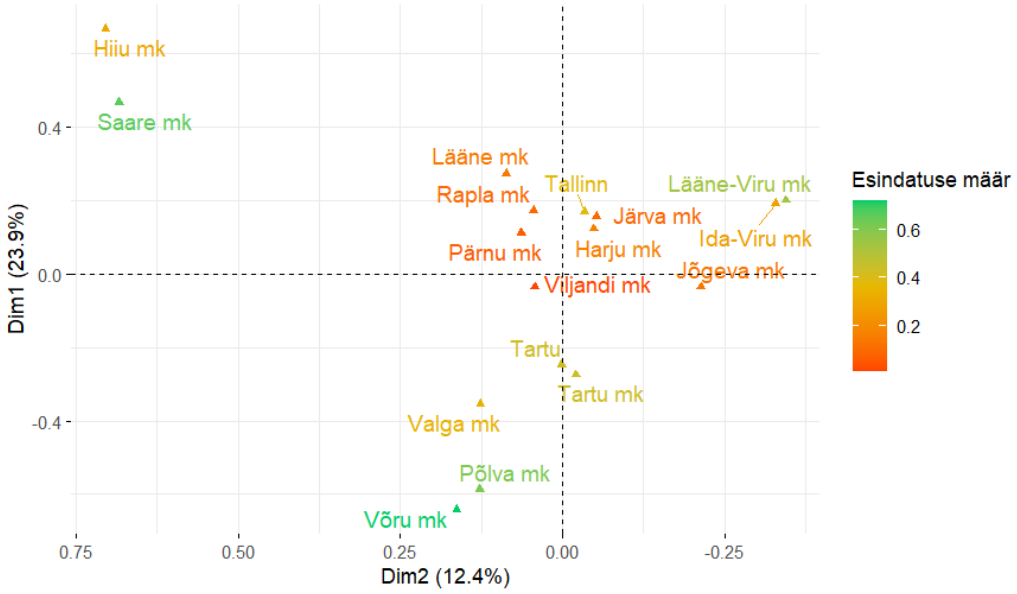
Joonis 2: Kõige suurema osakaaluga haplogrupid Eestis

3.2 Eesti maakondade omavaheline sarnasus

Esmalt uuriti, kas haplogruppide sageduste ja Eesti maakondade vahel leidub statistiliselt oluline seos. Selleks tehti hii-ruut test olulisuse nivool 0,05 ja vabadusastmete arvuga 4288. Tulemuseks saadi $\chi^2 = 15449$ ja p-väärtus $< 2{,}2 \cdot 10^{-16}$. Seega on haplogruppide osakaalud maakondades erinevad, ehk leidub seos haplogruppide ja maakondade vahel.

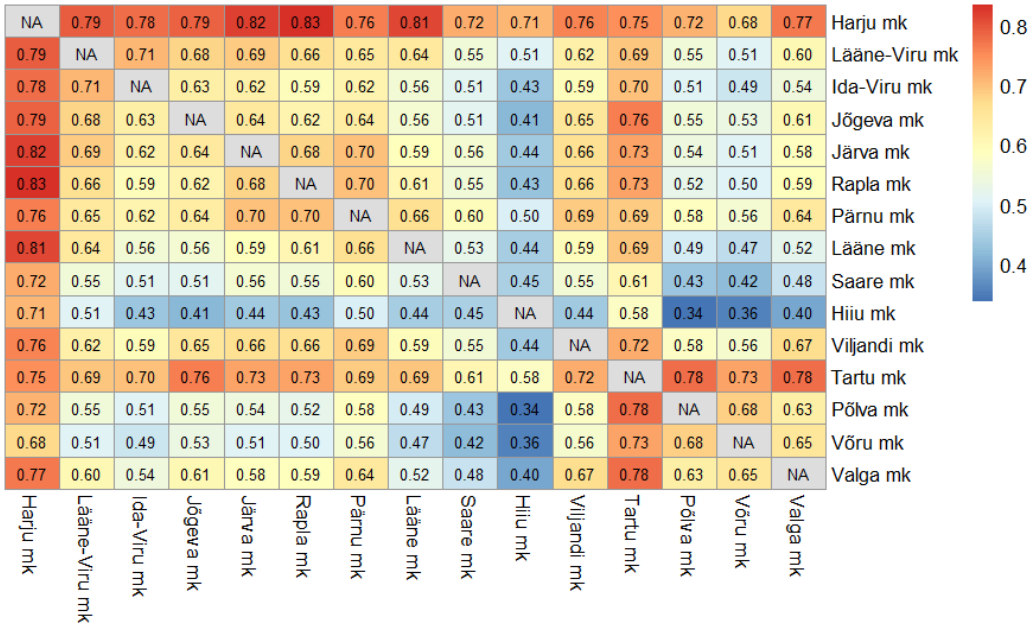
Esialgseks maakonnade omavaheliste sarnasuste hindamiseks tehti korrespondentsanalüüs. Joonisel 3 on näha, et esimese peatelje põhjal (Dim1) on eraldatud Kagu-Eesti maakonnad ja saared. Teine peatelg (Dim2) toob esile erinevuse saarte ja Kirde-Eesti vahel. Ülejäänud maakonand paiknevad joonise keskosas, seega pole neil esimese kahe telje suhtes olulist erinevust. On näha et geograafiliselt lähestikku asuvad maakonnad paiknevad ka korrespondentsanalüüsi graafikul sarnastes suundades telgede suhtes, mis võib viidata nende maakondade omavahelisele geneetilisele sarnasusele. Samas kirjeldavad esimesed kaks telge vaid $23{,}9\% + 12{,}4\% = 36{,}3\%$

andmetes olevast informatsioonist, mistõttu ei saa selle graafiku põhjal kindlaid järeldusi maakondade sarnasuse kohta teha. Maakonnad, mida esimesed kaks telge hästi kirjeldavad, on joonisel kujutatud rohelisega, halvemini kirjeldatud maakonnad on kujutatud punasega.



Joonis 3: Korrespondentsanalüüs Eesti maakondadel. Esindatuse määra skaala on 0..1.

Järgnevalt leiti paarikaupa maakondadevahelised Crameri V väärtused. Algandmestikus liideti Tartu ja Tallinna haplogruppide sagedused vastavalt Tartumaa ja Harjumaa sagedustele. Seejärel valiti saadud tabelist välja kaks maakonda ja leiti Crameri V väärtus tabelile, milles on 269 rida (haplogrupid) ja 2 veergu (valitud maakonnad). Mida suurem on saadud Crameri V väärtus, seda suurem on erinevus nendes maakondades haplogruppide sageduste vahel ehk seda geneetiliselt erinevamad on need maakonnad. Tulemused on esitatud soojustabelina joonisel 4, kus tulemuste paremaks loetavuseks on Crameri V väärtused esitatud kujul $1 - V$ ehk mida väiksem on joonisel maakondade vaheline Crameri V väärtus, seda erinevamad need maakonand omavahel on.



Joonis 4: Crameri V väärtused Eesti maakondade vahel, esitatud kujul $1-V$. Mida suurem on lahtri väärtus, seda sarnasemad on vastavad maakonnad.

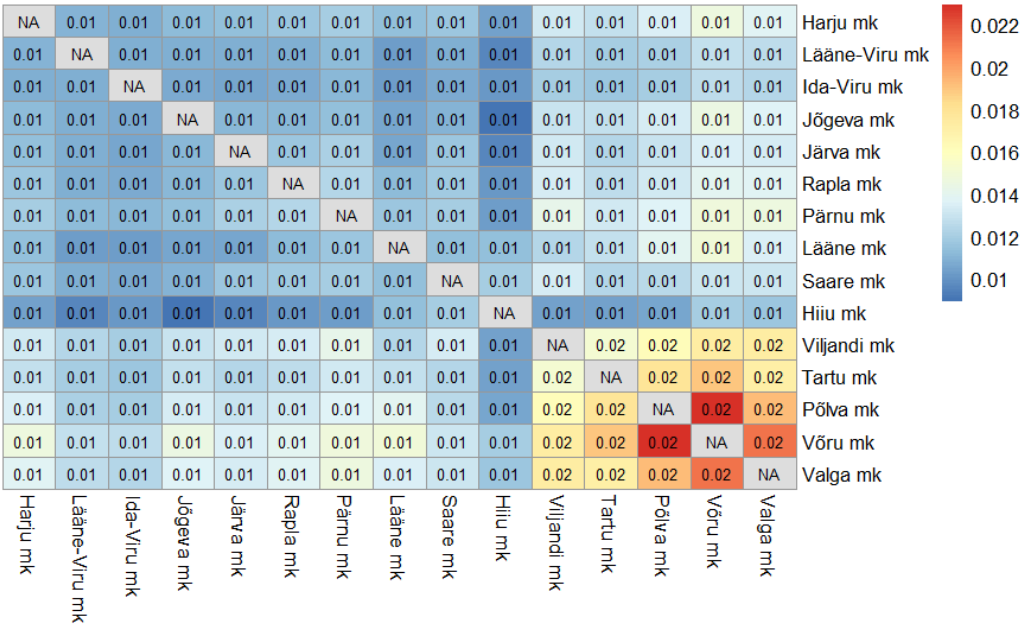
Joonisel on näha, et Tartu- ja Harjumaa on sarnased kõigi maakondadega. Hiiumaa on aga pea kõigist maakondadest erinev, kaasa arvatud Saare- ja Läänemaast. Joonistub välja ka Kagu-Eesti maakondade plokk, kuhu kuuluvad Viljandi, Tartu, Valga, Võru ja Põlva maakond. Sarnaselt korrespondentsanalüüsi graafikule (joonis 3) on ka joonisel 4 näha, et Kagu-Eesti ja saared on omavahel erinevad. Korrespondentsanalüüsi graafikul nähtud saarte ja Kirde-Eesti vaheline erinevus joonisel 4 tugevalt esile ei tule. See võib tuleneda asjaolust, et joonisel 3 olid Kagu-Eesti maakonnad paremini kirjeldatud kui Kirde-Eesti maakonnad, seega võis Kirde-maakondade paiknemine saarte suhtes anda eksitavat informatsiooni. Samuti kirjeldas Kagu-maakondade ja saarte erinevust määrav telg (Dim1) pea kaks korda suuremat osa andmetest leiduvast informatsioonist (23,9%) kui Kirde-maakondade ja saarte vahelist erinevust määrav telg (Dim2) (12,4%). Seega on saarte ja Kagu-Eesti maakondade vaheline erinevus joonisel 3 paremini esindatud kui saarte ja Kirde-Eesti maakondade vaheline seos, mistõttu ei ole tulemused täiesti sarnased

Crameri V tulemustega.

Järgmiseks leiti kõigi maakondade paaride jaoks tõenäosus, et kui võtta kummastki maakonanst üks isik, siis neil on sama haplogrupp. Olgu j-nda haplogrupi osakaal i-ndas maakonnas tähistatud p_{ij}. Kui võtta üks isik maakonnast i_1, olgu tema haplogrupp j_1, ja teine isik maakonnast i_2, olgu tema haplogrupp j_2, siis tõenäosus et neil on sama haplogrupp on

$$P(j_1 = j_2) = \sum_{j=1}^{269} p_{i_1 j} \cdot p_{i_2 j} \quad .$$

Tõenäosus arvutati iga maakonna paari jaoks ja tulemused on esitatud joonisel 5. Sarnaselt joonisega 4 tundub, et Kagu-Eesti maakonnad on omavahel mtDNA suhtes geneetiliselt sarnasemad kui ülejäänud maakonnad. Saarte erinevus Kagu- ja Kirde-Eestist joonisel 5 võrreldes saarte erinevusega muudest Eesti piirkondadest oluliselt esile ei tule.



Joonis 5: Tõenäosus et kahel erinevast maakonnast pärit isikul on sama haplogrupp.

Sellisel maakondade sarnasuse arvutamisel võib aga tekkida ebamäärasusi. Kui kahes maakonnas on üks haplogrupp, mille osakaal on suur, siis see muudab summa suureks, kuigi muude haplogruppide osakaalud nendes maakondades võivad olla väga erinevad. Olgu näiteks 3 maakonda ja igas 3 haplogruppi, mille osakaalud on toodud tabelis 2.

Tabel 2: Näide haplogruppide osakaaludest maakondades

Haplogrupp	hapl 1	hapl 2	hapl 3
maakond A	0,1	0,1	0,8
maakond B	0,4	0,1	0,5
maakond C	0,4	0,2	0,4

Tabeli andmete põhjal on tõenäosused, et maakondadest kahe inimese valimisel on neil sama haplogrupp:

$$\begin{aligned} P(A_j = B_j) &= 0{,}2 \cdot 0{,}5 + 0{,}3 \cdot 0{,}0 + 0{,}5 \cdot 0{,}5 \quad = 0{,}45 \\ P(A_j = C_j) &= 0{,}2 \cdot 0{,}1 + 0{,}3 \cdot 0{,}7 + 0{,}5 \cdot 0{,}2 \quad = 0{,}38 \\ P(B_j = C_j) &= 0{,}5 \cdot 0{,}1 + 0{,}0 \cdot 0{,}7 + 0{,}5 \cdot 0{,}2 \quad = 0{,}38 \end{aligned}$$

Nende tõenäosuste põhjal oleksid kõige sarnasemad maakonnad A ja B, aga vaadates tabelit 2 on näha, et haplogruppide osakaalude proportsioonide poolest on sarnasemad hoopis maakonnad B ja C. Seega peab selliselt arvutatud maakondade sarnasuste interpreteerimisel olema ettevaatlik.

Uuriti ka Euraasia idaosas laialdaselt levinud haplogruppide leidumist ja jaotumist Eesti maakondades. Need haplogrupid on D5a3a1a, Z1a, D4e4b, G2a1 ja M10a2, millest kaks esimest on omased saami rahvastele (Tambets *et al.*, 2004). Nende esinemissagedused ja osakaalud Eestis on esitatud tabelis 3.

Tabel 3: Idaliinide haplogruppide esinemissagedused Eestis ja nende osakaalud(%) Eestis esinevatest haplogruppidest.

Haplogrupp	D4e4b	D5a3a1a	G2a1	M10a2	Z1a
Sagedus Eestis	66	154	119	28	85
Protsent haplogrupidest	0.14	0.32	0.24	0.06	0.17

Nende haplogruppide Eesti-sisese jaotumise uurimiseks standardiseeriti iga maakonna valimi moodustavad haplogruppide sagedused valemiga:

$$X_{ij} = \frac{saadud_{ij} - oodatav_{ij}}{\sqrt{oodatav_{ij}}},$$

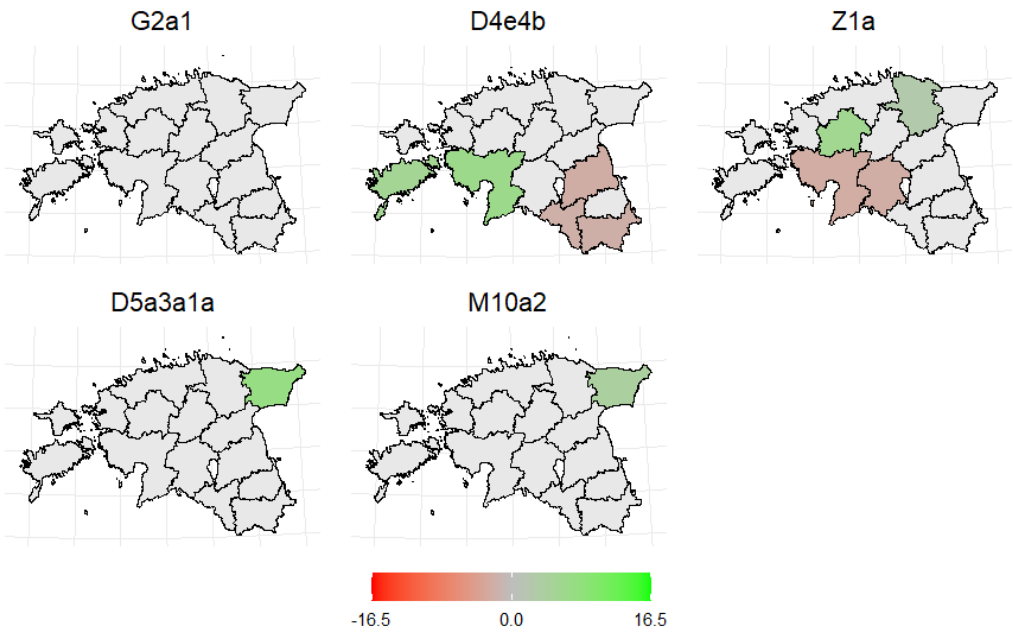
kus $saadud_{ij}$ on valimis vaadeldud j haplogrupi sagedus maakonnas i ja $oodatav_{ij}$ on oodatav j haplogrupi sagedus maakonnas i, kui haplogruppide osakaalud kõigis maakondades oleksid võrdsed. Oodatavad sagedused arvutati valemiga

$$oodatav_{ij} = kokku_i \cdot osakaal_j \ ,$$

kus $kokku_i$ on i-maakonna marginaalsagedus ja $osakaal_j$ on j-haplogruppi marginaalosakaal. Saadud statistikud X_{ij} on standardse normaaljaotusega juhuslikud suurused.

Standardse normaaljaotuse kriitilised väärtused olulisuse nivool $\alpha = 0{,}05$ on vahemikes $-\infty... - 1{,}96$ ja $1{,}96...\infty$. Seega kui statistiku X_{ij} väärtus jäi vahemikku $-1{,}96...1{,}96$, siis ei loetud j haplogrupi sagedust i maakonnas oluliselt erinevaks Eesti keskmisest. Kui väärtus oli väiksem kui -1,96, siis esines haplogruppi maakonnas vähem kui Eestis keskmiselt, kui väärtus oli suurem kui 1,96, siis esines rohkem kui Eestis keskmiselt. Tulemused kanti Eesti kaardile (joonis 6), kus punane värvus tähendab, et maakonnas esines haplogruppi oodatust vähem ja roheline et haplogruppi esines oodatust rohkem.

Joonisel on näha, et haplogruppe D5a3a1a ja M10a2 esines oodatust rohkem Ida-Virumaal. Haplogrupp D4e4b puhul tuli esile ida-lääne suunaline erinevus ja Z1a



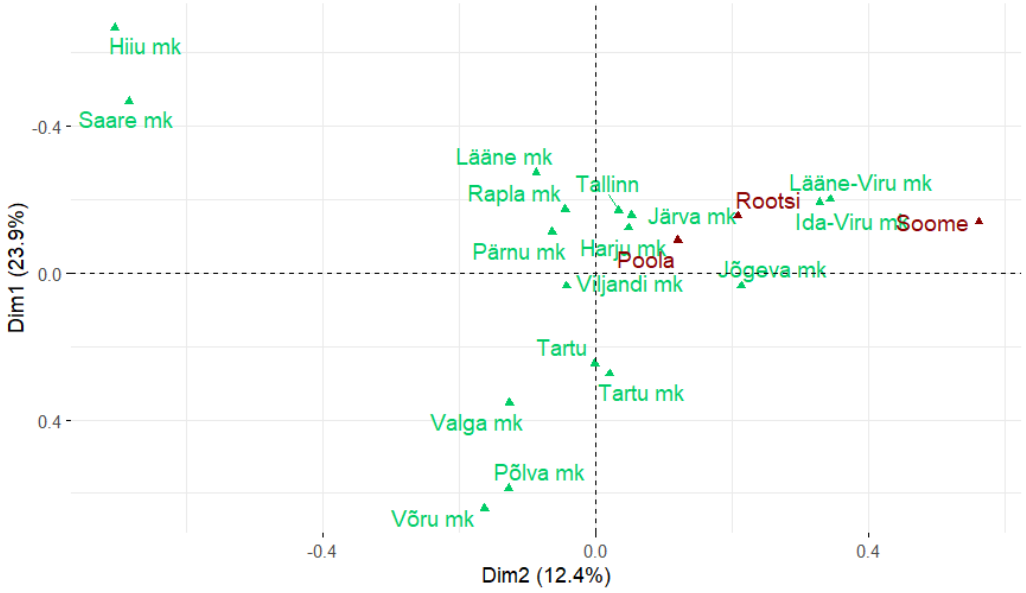
Joonis 6: Ida haplogruppide jaotumine Eesti maakondades. Värviskaala näitab arvutatud statistiku väärtust (kui väärtus on vahemikus -1,96...1,96, siis on maakond kujutatud halliga).

puhul põhja-lõuna suunaline erinevus. G2a1 haplogrupi puhul ei esinenud üheski maakonnas olulist erinevust oodatavast sagedusest. Seega on see haplogrupp kas Eesti maakondades ühtlaselt jaotunud või ei olnud valimis piisavalt andmeid erinevuse tuvastamiseks.

3.3 Eesti maakondade sarnasused naaberpopulatsioonidega

Esmalt koostati korrespondentsanalüüsi graafik (joonis 7), kuhu kaasati lisaks Eesti maakondadele ka Soome, Rootsi ja Poola. Naaberpopulatsioonide info ei ole kaasatud telgede arvutustesse. Esimese peatelje suhtes paiknevad naaberpopulatsioonid joonise keskel, seega sellel suunal neil suuri erinevusi ei ole. Teise peatelje suhtes paiknevad naaberpopulatsioonid graafiku paremal poolel, sarnaselt Põhja- ja

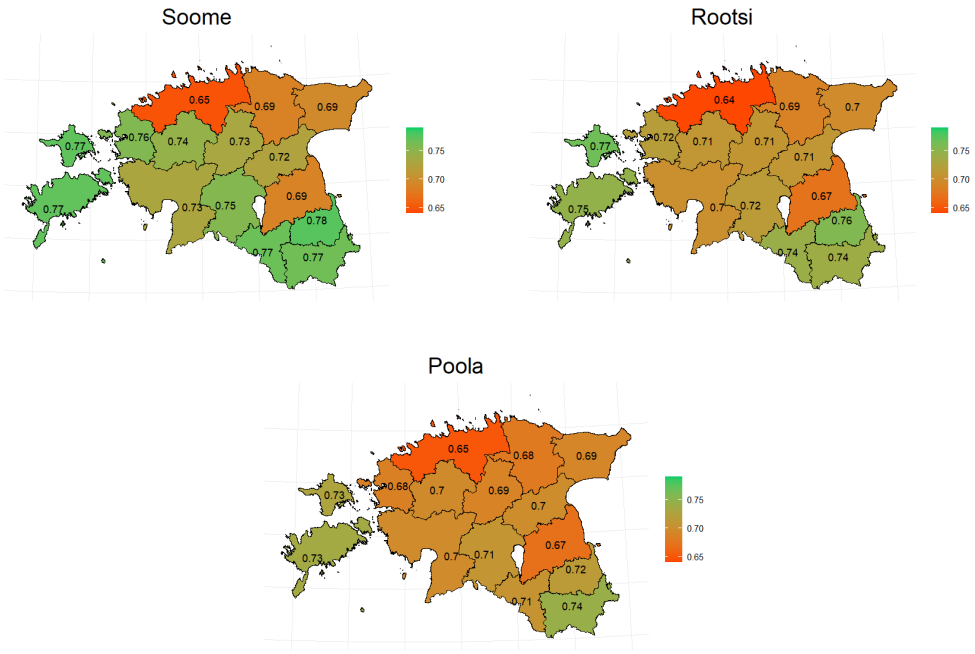
Kirde-Eesti maakondadega.



Joonis 7: Korrespondentsanalüüs Eesti maakondade ja naaberpopulatsioonidega - teljed on arvutatud Eesti maakondade põhjal ja naaberpopulatsioonid on hiljem lisatud

Järgmiseks võrreldi maakondi kolme naaberpopulatsiooniga eraldi. Selleks arvutati iga maakonna ja naaberpopulatsiooni paari jaoks Crameri V väärtus. Tulemused kanti Eesti kaardile (joonis 8), kus väiksem arv näitab maakonna suuremat sarnasust vastava naaberpopulatsiooniga.

Iga naaberpopulatsiooni puhul olid Crameri V väärtused suured, vähemalt 0,64, seega ei ole maakondade sarnasus ühegi naaberpopulatsiooniga suur. Harjumaa oli iga naaberpopulatsiooniga sarnasem kui teised maakonnad. Crameri V väärtuste summad on iga naaberpopulatsiooni puhul sarnased, kuid on kõige madalamad Poolaga ja kõige kõrgemad Soomega, mis võiks viidata sellele, et Eesti on Poolaga geneetiliselt sarnasem kui Soomega. Soome ja Rootsi puhul on näha, et kõige sarnasemad on Põhja-Eesti maakonnad, mis on kooskõlas ka nende piirkondade ajalooliselt tihedama läbikäimisega. Sarnasus Poolaga on üle Eesti ühtlasem.



Joonis 8: Maakondade sarnasused naaberpopulatsioonidega Crameri V põhjal. Mida väiksem on Crameri V väärtus, seda sarnasem on maakond vastava riigi populatsiooniga.

3.4 Simulatsioonid

Uurimaks kui hästi kirjeldavad Crameri V ja korrespondentsanalüüs maakondade sarnasusi naaberpopulatsioonidega, koostati simulatsioon. Esmalt loodi andmestik, mis koosnes haplogruppide esinemissagedustest niinimetatud ürgeestis ja ürgrootsis. Kaasati samad haplogrupid, mis algandmestikus Eestis esinesid. Rootsi osakaaludeks võeti algandmetes olevad haplogruppide osakaalud Rootsis, kaasa arvatud "Muud" alla kuuluvate haplogruppide osakaal. Osakaalud ürgeestis saadi liites kokku algandmete haplogruppide sagedused Eesti maakondades ja leides saadud marginaalsageduste osakaalud.

Seejärel moodustati simuleeritud haplogruppide jaotused kaasaja Eesti maakondades nii, et haplogruppide osakaalud maakondades arvutati seguna ürgrootsi ja ürgeesti osakaaludest, kus ürgrootsi mõju oli igas maakonnas erinev. Olgu maa-

konna indeks $i \in \{1,2,..15\}$ ja haplogruppi indeks $j \in \{1,2,..,269\}$, siis haplogrupi osakaal maakonnas arvutati valemiga

$$p_{ij} = \alpha \cdot p_{\text{Rootsi},i,j} + (1 - \alpha) \cdot p_{\text{ürgeesti},i,j} \ ,$$

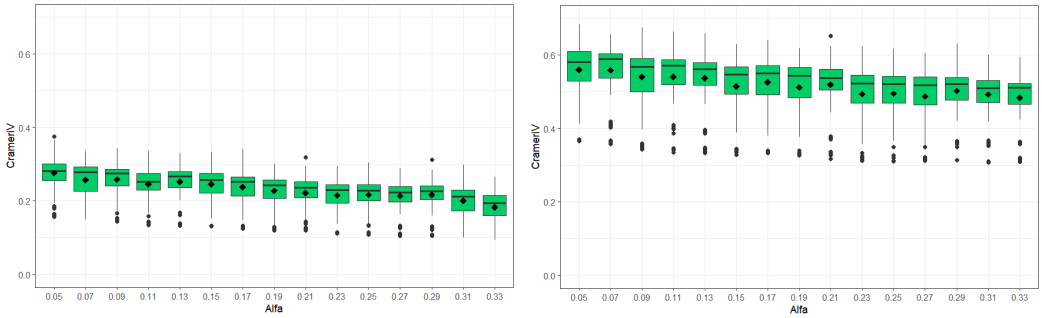
kus $p_{\text{Rootsi},i,j}$ on haplogrupi osakaal Rootsis, $p_{\text{ürgeesti},i,j}$ osakaal Eestis ja $\alpha \in \{0{,}05;$ $0{,}07;$ ...; $0{,}33\}$.

Igast maakonnast võeti multinomiaaljaotusele vastav tagasipanekuga valim, kus valimisuurused olid samad, mis algandmestikus maakondade valimite suurused. Saadi algandmestikuga sarnane andmestik, kus oli sama palju vaatluseid kui algandmestikus. Saadud simulatsioonandmestikule rakendati korrespondentsanalüüsi ja leiti Crameri V seosekordajad, et hinnata kui hästi need leiavad geneetilise sarnasuse määra maakondade ja Rootsi vahel sõltuvalt α suurusest. Kui α on suur, siis peaks maakond olema Rootsiga geneetiliselt sarnasem kui mõni väiksema α väärtusega maakond.

Kirjeldatud viisil võeti 100 valimit, millele leiti Crameri V väärtus nagu alapeatükkides 3.2 ja 3.3. Iga valimi puhul jaotati α väärtused ja valimite suurused maakondade vahel juhuslikult. Iga valimi puhul järjestati Crameri V väärtused vastavalt α väärtustele - väikseimale α väärtusele vastavast Crameri V-st suurima α väärtusele vastava Crameri V-ni. Arvutati igale α-le vastava Crameri V keskmine väärtus üle kõigi valimite. Protsess tehti läbi nii algsete haplogruppidega kui ka kahe esimese sümbolini ümardatud haplogruppidega. Tulemused on kujutatud joonisel 9.

Joonisel on näha pöördvõrdelist seost α ja Crameri V väärtuste vahel: mida suurem on α seda väiksem on Crameri V ehk seda sarnasemad on võrreldud populatsioonid. Järelikult kirjeldab Crameri V seost maakondade ja naaberpopulatsiooni geneetilise sarnasuse vahel. Samas on näha, et nii ümardatud kui ka algsete haplogruppidega arvutades on valdava osa α-de hajuvusvahemikud kattuvad, välja arvatud ümardatud haplogruppide puhul väikseimate ja suurimate α väärtuste hajuvusvahemikud. Seega on Crameri V abil raske maakondi naaberpopulatsiooniga sarnasuse põhjal

järjestada.



(a) Ümardatud haplogrupid (b) Algsed haplogrupid

Joonis 9: Crameri V väärtuste 100 valimi keskmised. Joon karpdiagrammi sees näitab mediaani ja romb keskmist.

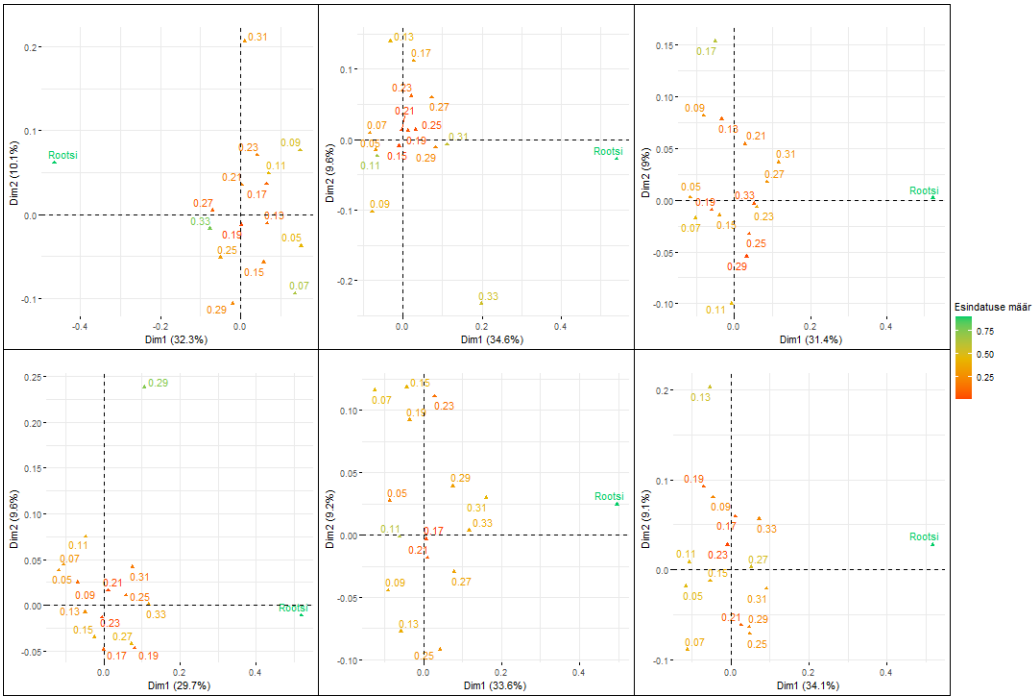
Igale valimile arvutati ka Crameri V ja α vaheline Pearsoni korrelatsioonikordaja, mille varieeruvust simuleeritud 100-s valimis on kirjeldatud tabelis 4. Ka korrelatsioonikordaja näitab, et väärtuste vahel on keskmise tugevusega negatiivne seos, üksikute eranditega. Seda nii ümardatud kui ka algsete haplogruppide puhul. Ümardatud haplogruppidega olid aga seosekordajate kõik kvartiilid, keskmine väärtus ja hajuvus madalamad kui algsete haplogruppidega . Seega oli ümardatud haplogruppidega seos α ja Crameri V väärtuse vahel tugevam.

Tabel 4: 100 valimi korrelatsioonikordajate keskmised kvantiilid, keskmine, vähim ja suurim väärtus.

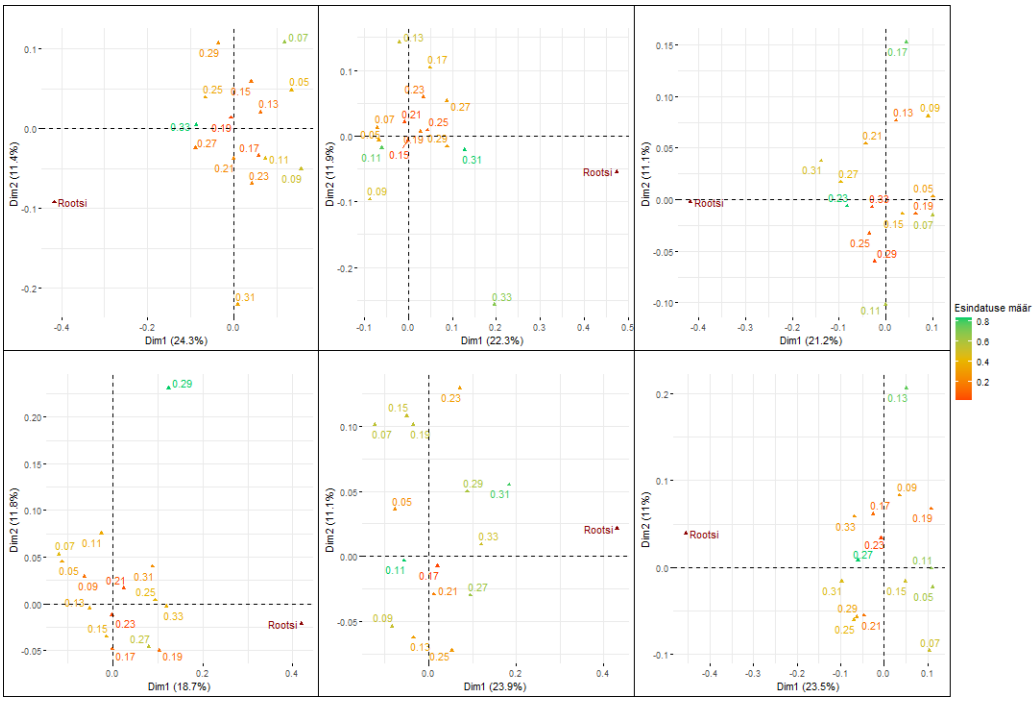
	0,25-kvantiil	mediaan	0,75-kvantiil	keskmine	min	max
Algsed haplogrupid	-0.48	-0.34	-0.17	-0.31	-0.88	0.52
Ümardatud haplogrupid	-0.57	-0.46	-0.31	-0.43	-0.86	0.11

Korrespondentsanalüüsi simuleerimiseks genereeriti 6 valimit sarnaselt Crameri V valimitele, kasutati ümardatud haplogruppe. Iga valimiga tehti kaks korrespondentsanalüüsi graafikut: koos Rootsiga arvutatud dimensioonidega ja ainult Eesti maakondadega arvutatud dimensioonidega, kuhu Rootsi juurde arvutati. Graafikutel 10a ja 10b on näha, et väiksemate ja suuremate α väärtustega maakonnad on esimese kahe peatelje poolt määratud tasandil paremini kirjeldatud ja graafik

on nende maakondade suhtes polariseerunud. Keskmiste α väärtustega maakonnad paiknevad graafiku keskel. Samuti on näha, et Rootsi paikneb graafiku esimese peatelje suhtes sarnaselt suuremate α väärtustega maakondadega. Seega eristab korrespondentsanalüüsi graafik maakondade suurt sarnasust ja erinevust naaberpopulatsiooniga, kuid täpset maakondade sarnasuse järjekorda see edasi ei anna.



(a) Koos Rootsiga arvutatud dimensioonid.



(b) Ilma Rootsita arvutatud dimensioonid.

Joonis 10: Simulatsioonvalimite korrespondentsanalüüsi graafikud. Iga maakond on märgitud sellele määratud α väärtusena.

4 Tulemused

Eesti maakondade omavahelistest võrdlustest tuli välja eelkõige Kagu-Eesti eraldiseisvus ning saarte, eriti Hiiumaa erinevus muudest maakondadest. Kagu-Eesti omapära on täheldatud ka varasemates uuringutes (Pankratov *et al.*, 2020). Hiiumaa valim oli oluliselt väiksem ülejaanud maakondade valimitest, mis võis tulemusi mõjutada. Samas tuli Hiiumaa erinevus kõigi kasutatud meetodite puhul esile (vt joonised 3, 4, 5). Crameri V väärtuste (joonis 4) järgi olid nii Harju- kui Tartumaa kõigi Eesti maakondadega sarnased.

Antud bakalaureusetöö tulemusi võiks edasistes uuringutes kõrvutada Eestisisese rände infoga, et näha kas piirkondade geneetiline segunemine on sellega kooskõlas. Kui inimesed kolivad suurematesse asulatesse, siis kas need paigad muutuvad geneetiliselt mitmekesisemaks ja teiste maakondadega sarnaseks. Teisalt kui kolitakse ära saartelt ja väiksematest asulatest ning nendesse paikadesse kolib asemele vähe inimesi, siis kas need piirkonnad muutuvad geneetiliselt ühetaolisemaks ja muudest maakondadest eraldiseisvaks.

Euraasia idaosa populatsioonide seas levinud haplogruppide puhul esines saamidele omaseid haplogruppe rohkem Ida- ja Lääne-Virumaal. See tulemus on kooskõlas ka geograafiliste eeldustega, sest Virumaad asuvad Soomele ja Venemaale lähemal. Samuti on haplogrupi M10a2 suurem esinemine Ida-Virumaal kooskõlas selle haplogrupi laialdase levikuga Venemaa ja Aasia populatsioonides (Derenko *et al.*, 2012). Ebaootuspärane oli D4e4b jaotumine Eesti maakondades: haplogruppi esines rohkem Lääne-Eestis ja vähem Ida-Eestis, mis läheb vastuollu haplogrupi geograafilise levikuga Venemaal (Derenko *et al.*, 2010). Haplogrupi G2a1 puhul ei tulnud ükski Eesti piirkond esile, seega oli kas informatsiooni liiga vähe või on see haplogrupp Eesti piires ühtlaselt jaotunud. Võimalik, et selle emaliini kandjad jõudsid Eesti alale palju varem, mistõttu on järeltulijad jõudnud ühtlaselt jaotuda kõikide maakondade vahel.

Eesti maakondade võrdluses naaberpopulatsioonidega näis Soome ja Eesti erinevus

olevat suurem kui Eesti erinevus Rootsist või Poolast. Seda võis mõjutada Soome omapärane ajalooline geneetiline kujunemine, mille tõttu on riigi populatsioon ka üldiselt muudest Euroopa populatsioonidest geneetiliselt eraldiseisev (Kere, 2001). Lisaks oli Soome valimisse kaasatud ka saami rahvastele omased geenid, mis omakorda suurendavad Soome erinevust muudest Euroopa populatsioonidest.

Simulatsioonide tulemusena saab väita, et nii Crameri V kui ka korrespondentsanalüüs kirjeldavad geneetilise sarnasuse määra Eesti maakondade ja naaberpopulatsioonide vahel. Crameri V puhul on mõistlik kasutada ümardatud haplogruppe, sest nendega saadud korrelatsioonikordaja väärtused olid keskmiselt oluliselt suuremad kui algsete haplogruppidega saadud väärtused (tabel 4). Seega, kuigi kasutatud meetodid on võimelised tuvastama naaberpopulatsioonide mõju maakondade populatsioonides, siis antud valimisuuruste puhul ei ole need suutelised andma täpset maakondade sarnasuse järjestust naaberpopulatsiooni suhtes.

Simulatsioonide tulemusi arvestades võib Crameri V põhjal (joonis 8) arvata, et Põhja-Eesti maakonnad ja Tartu maakond on Soomega sarnasemad kui ülejäänud maakonnad. Kagu-Eesti ja Hiiumaa on Rootsiga vähem sarnased kui muud maakonnad. Poolaga tundub Eesti olevat ühtlaselt sarnane. Ka korrespondentsanalüüsi jooniselt 7 saab teha samad järeldused: Soome ja Rootsi paiknevad teise peatelje suhtes saartest kaugel, Poola aga asub graafiku keskpunktile lähemal ehk ei ole ühegi Eesti piirkonnaga oluliselt sarnasem kui mõne teise piirkonnaga.

Kokkuvõte

Bakalaureuse töö eesmärgiks oli uurida Eesti maakondade populatsioonide geneetilist sarnasust mitokondriaalse DNA põhjal. Samuti hinnati nende sarnasust naaberpopulatsioonidega.

Eestis esinenud haplogrupid ühtisid üldjoontes Euroopale omaste haplogruppidega. Nii korrespondentanalüüs, Crameri V kui ka sama haplogrupi esinemise tõenäosus näitasid, et Eesti siseselt tuli esile Kagu-Eesti erinevus ülejäänud Eesti maakondade populatsioonidest. Samuti oli eraldiseisev Hiiumaa populatsioon. Eesti suurimate linnadega maakonnad, Harjumaa ja Tartumaa, olid kõigi maakondadega sarnased. Euraasia idaosale omased haplogrupid olid Eestis enamjaolt oodatavalt jaotunud: neid esines rohkem idapool ja vähem läänepool. Ühte haplogruppi esines rohkem läänes ja üks oli Eestis ühtlaselt jaotunud.

Naaberpopulatsioonidest oli Eesti kõige sarnasem Poolaga, vähem Rootsi ja Soome populatsioonidega. Põhjapoolsete naaberpopulatsioonidega tuli esile põhja-lõuna suunaline gradient, kus Põhja-Eesti oli mõlema naaberpopulatsiooniga sarnasem kui Lõuna-Eesti. Poola populatsiooniga oli Eesti maakondade sarnasus ühtlasem. Kasutatud statistiliste meetodite headuse hindamisel simulatsioonvalimitega saadi tulemuseks, et need kirjeldavad töös kasutatud algandmestikku hästi, kuid ei anna täpset järjestust, millise maakonna populatsioon on millise naaberpopulatsiooniga kõige sarnasem.

Edasistes uuringutes võiks kõrvutada Eesti geneetilist struktuuri ajalooliste migreerumistega Eesti aladel. Samuti Eesti maakondade vaheliste liikumiste infot, mis aitaks mõista, miks on haplogruppide jaotumine Kagu-Eestis ja Hiiumaal muudest maakondadest erinev. Samuti, miks esines haplogruppi D4e4b rohkem just Lääne-Eestis, mitte Ida-Eestis.

Kasutatud allikad

Bandelt, H.-J., Macaulay, V. ja Richards, M. (2006). *Human mitochondrial DNA and the evolution of Homo sapiens*. Springer.

Bogenhagen, Daniel ja Clayton, David A. (1974). „The Number of Mitochondrial Deoxyribonucleic Acid Genomes in Mouse L and Human HeLa Cells: QUANTITATIVE ISOLATION OF MITOCHONDRIAL DEOXYRIBONUCLEIC ACID“. *Journal of Biological Chemistry* 249.24, lk. 7991–7995. ISSN: 0021-9258. URL: https://doi.org/10.1016/S0021-9258(19)42063-2.

Brown, Wesley M., Jr, Matthew George ja C.Wilson, Allan (1979). „Rapid evolution of animal mitochondrial DNA“. *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.76.4.1967. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC383514/.

Clark, K., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J. ja Sayers, EW. (2021). *GenBank, Nucleic acids research*. DOI: 10.1093/nar/gkv1276.

Cramer, Harald (1946). *Mathematical methods of statistics*. Princeton paperbacks Princeton landmarks in mathematics and physics. Princeton University Press.

Derenko, M., Malyarchuk, B., Denisova, G., Perkova M.and Rogalla, U., Grzybowski, T., Khusnutdinova, E., Dambueva, I. ja Zakharov, I. (2012). „Complete mitochondrial DNA analysis of eastern Eurasian haplogroups rarely found in populations of northern Asia and eastern Europe.“ *PLoS One* 7.2. URL: https://doi.org/10.1371/journal.pone.0032179.

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., Dambueva, I. ja Zakharov, I. (2010). „Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in northern Asia.“

PLoS One 5.12. URL: https://doi.org/10.1371/journal.pone.0015214.

Eesti geenivaramu (2021). URL: https://genomics.ut.ee/et/eesti-geenivaramu (vaadatud 13.03.2024).

Geneetika sõnastik (2014). URL: https://geneetika.ee/encyclopedia/molekulaarne-kell/ (vaadatud 10.03.2024).

Giles, Richard E., Blanc, Hugues, Cann, Howard M. ja Wallace, Douglas C. (1980). „Maternal inheritance of human mitochondrial DNA“. DOI: 10.1073/pnas.77.11.6715. URL: https://doi.org/10.1073/pnas.77.11.6715.

Greenacre, Michael (1993). *Correspondence analysis in practice*.

Greenacre, Michael J. (1984). *Theory and Applications of Correspondence Analysis*.

Kaja, E., Lejman, A., Sielski, D., Sypniewski, M., Gambin, T., Dawidziuk, M., Suchocki, T., Golik, P., Wojtaszewska M.and Mroczek, M., Stępień, M., Szyda, J., Lisiak-Teodorczyk, K., Wolbach, F., Kołodziejska, D., Ferdyn, K., Dąbrowski, M., Woźna, A., Żytkiewicz, M., Bodora-Troińska, A., Elikowski, Waldemar, Król, Zbigniew J, Zaczyński, Artur, Pawlak, Agnieszka, Gil, Robert, Wierzba, Waldemar, Dobosz, Paula, Zawadzka, Katarzyna, Zawadzki, Paweł ja Sztromwasser, P. (2022). „The Thousand Polish Genomes-A Database of Polish Variant Allele Frequencies.“ *International journal of molecular sciences* 23.9. URL: https://doi.org/10.3390/ijms23094532.

Kere, Juha (2001). „Human Population Genetics: Lessons from Finland“. *ANNUAL REVIEW OF GENOMICS AND HUMAN GENETICS* 2. URL: https://doi.org/10.1146/annurev.genom.2.1.103.

Pankratov, V., Montinaro, F., Kushniarevich, A., Hudjashov, Georgi, Jay, Flora, Saag, Lauri, Flores, Rodrigo, Marnetto, Davide, Seppel, Marten, Kals, Mart, Võsa, Urmo, Taccioli, Cristian, Möls, Märt, Milani, Lili, Aasa, Anto, Lawson, Daniel John, Esko, Tõnu, Mägi, Reedik, Pagani, Luca, Metspalu, Andres ja Metspalu, Mait (2020). „Differences in local population history at the finest level: the case of the Estonian population.“ *European journal of human genetics* 28. URL: https://doi.org/10.1038/s41431-020-0699-4.

Pärna, Kalev (1993). *Correspondence Analysis: an introduction and some examples*. Tehniline raport. Stockholm University.

Richards, M., Macaulay, V. ja H. J. Bandelt, A. Torroni ad (2002). „In search of geographical patterns in European mitochondrial DNA.“ *American journal of human genetics* 71.5, 1168–1174. URL: https://doi.org/10.1086/342930.

Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J. ja Barbujani, G. (2000). „Geographic patterns of mtDNA diversity in Europe.“ *American journal of human genetics* 66.1, 262–278. URL: https://doi.org/10.1086/302706.

Stewart, J. ja Chinnery, P. (2015). „The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease.“ *Nature Reviews Genetics*. URL: https://doi.org/10.1038/nrg3966.

Sturk-Andreaggi, K., Ring, JD., Ameur, A, Gyllensten, U, Bodner, M., Parson, W, Marshall, C. ja Allen, M. (2022). „The Value of Whole-Genome Sequencing for Mitochondrial DNA Population Studies: Strategies and Criteria for Extracting High-Quality Mitogenome Haplotypes.“ *International Journal of Molecular Sciences* 23.4. URL: https://doi.org/10.3390/ijms23042244.

Tambets, Kristiina, Rootsi, Siiri, Kivisild, Toomas, Help, Hela, Serk, Piia, Loogväli, Eva-Liis, Tolk, Helle-Viivi, Reidla, Maere, Metspalu, Ene, Pliss, Liana, Balanovsky, Oleg, Pshenichnov, Andrey, Balanovska, Elena, Gubina, Marina, Zhadanov, Sergey, Osipova, Ludmila, Damba, Larisa, Voevoda, Mikhail, Kutuev, Ildus, Bermisheva, Marina, Khusnutdinova, Elza, Gusar, Vladislava, Grechanina, Elena, Parik, Jüri, Pennarun, Erwan, Richard, Christelle, Chaventre, Andre, Moisan, Jean-Paul, Barac´, Lovorka, Peric˘ic´, Marijana, Rudan, Pavao, Terzic´, Rifat, Mikerezi, Ilia, Krumina, Astrida, Baumanis, Viesturs, Koziel, Slawomir, Rickards, Olga, Stefano, Gian Franco De, Anagnou, Nicholas, Pappa, Kalliopi I., Michalodimitrakis, Emmanuel, Fera´k, Vladimir, Füredi, Sandor, Komel, Radovan, Beckman, Lars ja Villems, Richard (2004). „The Western and Eastern Roots of the Saami—the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes". *The American Journal of Human Genetics* 74.4. DOI: 10.1086/383203. URL: https://www.researchgate.net/publication/8675817_The_Western_and_Eastern_Roots_of_the_Saami_-_The_Story_of_Genetic_Outliers_Told_by_Mitochondrial_DNA_and_Y_Chromosomes/citations.

Wai, Timothy, Ao, Asangla, Zhang, Xiaoyun, Cyr, Daniel, Dufort, Daniel ja Shoubridge, Eric A. (2010). „The Role of Mitochondrial DNA Copy Number in Mammalian Fertility1". *Biology of Reproduction* 83.1. DOI: 10.1095/biolreprod.109.080887. URL: https://doi.org/10.1095/biolreprod.109.080887.

Lisa 1. Andmestik

	Haplogrupp	Harju mk	Lääne-Viru mk	Ida-Viru mk	Jõgeva mk	Järva mk	Rapla mk	Pärnu mk	Lääne mk	Saare mk	Hiiu mk	Viljandi mk
1	D4e4b	13	9	1	1	7	7	32	4	15	0	5
2	D4j, D4j2a D4j5	3	0	2	0	2	3	0	0	1	0	0
3	D5a3a1a	11	6	14	1	1	1	4	0	0	0	2
4	F1b1 (c)	7	1	0	2	0	0	3	0	0	1	0
5	G2a1	23	7	6	9	7	2	7	5	1	1	6
6	G2a2	5	5	0	0	0	1	1	0	1	2	1
7	H10e	4	5	3	0	1	3	10	0	1	0	14
8	H11a (v.a H11a1 H11a2)	72	51	30	18	13	16	48	18	33	46	28
9	H11a1	18	9	1	1	0	7	27	4	2	0	7
10	H11a2 (v.a. H11a2a2)	2	1	1	1	0	0	1	1	1	1	1
11	H11a2a2	16	8	4	7	8	8	34	9	32	1	9
12	H11b1(13528. 14148??)	5	5	4	0	3	2	10	1	1	0	14
13	H13a (v.a H13a1a1a H13a1a1c H13a1a2+16311 H13a1a2a)	36	29	27	7	11	7	24	8	6	36	14
14	H13a1a1a	12	6	2	5	0	4	4	4	5	0	3
15	H13a1a1c	93	33	10	23	9	12	49	11	6	5	23
16	H13a1a2+16311	18	21	16	6	0	7	10	7	11	7	4
17	H13a1a2a	3	3	1	2	3	0	1	0	1	2	17
18	H13b1	8	5	3	1	3	1	10	1	0	0	8

Joonis 11: Andmestiku esimesed 18 rida ja 12 veergu

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Mirian Valk,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Eesti geneetiliste emaliinide mitmekesisus, struktuur ja võrdlus valitud naaberpopulatsioonidega", mille juhendajad on Märt Möls ja Anne-Mai Ilumäe, 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 4.0, 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 intellektuaaloman-di ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Mirian Valk

20.05.2024