TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Oliver Aasmets

Mikrobioomi andmete analüüs

Magistritöö matemaatilise statistika erialal (30 EAP)

Juhendajad:
PhD Krista Fischer
PhD Elin Org

Tartu 2018

Mikrobioomi andmete analüüs

Lühikokkuvõte

Inimese soolestikus on suur hulk erinevaid baktereid, mis täidavad organismi jaoks mitmeid olulisi funktsioone. Käesoleva magistritöö eesmärk on uurida, kas teist tüüpi diabeedi eelses seisundis indiviidide soolestiku bakterikoosluses on muudatusi võrreldes tervete indiviidide bakterikooslusega. Võrreldakse bakterikoosluse puhul huvitavaid α- ning β- mitmekesisuse näitajaid. Seejärel uuritakse Mendeli randomiseerimise skeemi abil, missugune võiks olla bakterikoosluse liigirikkuse põhjuslik mõju prediabeedile. Lisaks uuritakse, kas leidub üksikuid baktereid, mis esinevad tervete ja prediabeetikute mikrobioomides erineva sagedusega kasutades selleks kompositsionaalsete andmete analüüsimiseks mõeldud meetodeid. Kirjeldatakse kopositsionaalsete andmete jaoks mõeldud seose tugevuse näitajat uurimaks, kas soolestiku mikrobioomis on liike, mis esinevad soolestiku keskkonnas enamasti koos. Lisaks modelleeritakse prediabeedi esinemist logistilise regressiooni ning regulariseeritud logistilise regressiooniga.

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

Märksõnad: põhjuslik analüüs, statistilised mudelid, mikrofloora

Microbiome data analysis

Abstract

There are large number of different bacteria in the human gut that fulfill several important functions for the body. The aim of this Master's thesis is to investigate whether there are differences between the intestinal bacterial community of healthy and prediabetic individuals. Microbiome α- and β- diversity indicators are compared for the groups. Mendelian ranomization scheme is used to investigate the causal effect of the bacterial species richness on prediabetes. In addition, it is examined whether there are isolated bacteria that occur at different frequencies with microbiome from healthy and prediabetes, using methods developed for analyzing compositional data. A strengh of association measure for compositinal data is introduced to investigate whether there is co-existing bacteria in the intestinal environment. In addition, prediabetic state is modeled by logistic regression and regular logistic regression.

CERCS research specialisation: P160 Statistics, operations research, programming, actuarial mathematics

Keywords: causal analysis, statistical models, microflora

Sisukord

1 Sissejuhatus **5**

2 Inimese mikrobioom ja selle uurimise meetodid **6**

2.1 Mikrobioom 6

2.1.1 Mikrobioomi määramine 16S rRNA geeni sekveneerimise teel 7

2.1.2 Mikroobide operatiivsed taksonoomiaühikud (OTUd) . . 7

2.2 Populatsiooni mitmekesisus 8

2.2.1 α-mitmekesisus 8

2.2.2 β-mitmekesisus 9

3 Teist tüüpi diabeet **10**

4 Statistiline metoodika **12**

4.1 Klassikaline- ja mitteparameetriline dispersioonanalüüs (ehk ADONIS) 12

4.2 Regressioonmudelid uuritava tunnuse modelleerimiseks 15

4.2.1 Lineaarne regressioon 15

4.2.2 Logistiline regressioon 15

4.2.3 Regulariseeritud logistiline regressioon (LASSO) 15

4.3 Põhjuslike mõjude hindamine Mendeli randomiseerimise meetodil 16

4.4 Kompositsionaalsete andmete analüüs 18

4.4.1 Kompositsionaalsed andmed 18

4.4.2 Kompositsionaalsete andmete modelleerimine 20

4.4.3 Proportsioonide hindamine Bayesi meetodil 21

4.4.4 Bayesi metoodika rakendamine tarkvara R abil 22

4.5 Seosenäitajad kompositsionaalsete andmete korral 23

5 Mikrobioomi andmete analüüs **25**

5.1 Andmestiku kirjeldus 25

5.1.1 Kasutatav tarkvara ja andmestiku eeltöötlus 25

5.2 Kirjeldav analüüs 26

5.2.1 Fenotüübiandmete kirjeldav analüüs 26

5.2.2 Mikrobioomi kirjeldav analüüs 28
5.3 α-mitmekesisuse analüüs 29
5.4 β-mitmekesisuse analüüs 31
5.5 Prediabeetikute ja kontrollide võrdlus üksikute OTUde osas . . . 32

6 Logistilise regressiooni mudelid teist tüüpi diabeedi esinemisele 33
6.1 Klassikaline logistiline regressioon 33
6.2 LASSO regressioon 34

7 OTUde koosesinemine 38

8 Diskussioon 40

9 Kokkuvõte 44

Kasutatud kirjandus 46

A Joonised 47

B Tabelid 48

1 Sissejuhatus

Inimese organism on äärmiselt keeruline süsteem. Personaalse meditsiini seisukohal on oluline teada, kuidas meie genoom koos elukeskkonna ja elustiiliga üksteisega seostuvad ja kuidas nad mõjutavad erinevate haiguste arengut ning avaldumist. Tartu Ülikooli Eesti Geenivaramu on uurinud inimese geneetilisi andmeid hindamaks haiguse tekkimise riski sooviga rakendada saadud tulemusi personaalses meditsiinis.

Lisaks inimese geneetikale on organismi toimimises oluline roll mikroorganismidel. Soolestikus on suur hulk erinevaid baktereid, mis täidavad organismi jaoks mitmeid olulisi funktsioone, sealhulgas osaledes immuunsüsteemi arengus ning reguleerides ainevahetust. Samuti on leitud seoseid indiviidi bakterikoosluse ja mitmete haiguste vahel. Seetõttu pakub soolestiku bakterikooslus teel personaalse meditsiinini võimaluse haiguseriske täpsemalt hinnata.

Käesoleva magistritöö eesmärk on uurida, kas teist tüüpi diabeedi eelses seisundis indiviidide soolestiku bakterikoosluses on muutusi võrreldes tervete indiviidide bakterikooslusega. Töö esimeses pooles antakse ülevaade mikrobioomist, mikrobioomi andmete kogumise eripäradest ning statistilisest metodoloogiast, mida kasutatakse mikrobioomi andmete analüüsimiseks. Töö teises osas analüüsitakse Soomes kogutud meeste kohordi soolestiku mikrobioomi andmeid.

Autor tänab Krista Fischerit rohkete nõuannete ja suunamise eest statistiliste meetodite rakendamisel ning Elin Orgi huvitava probleemipüstituse ning abi eest mikrobioomi andmete analüüsimisel.

Töö praktilise osa läbiviimiseks on kasutatud statistikatarkvara R, töö on vormistatud kasutades tekstitöötlustarkvara LaTeX.

2 Inimese mikrobioom ja selle uurimise meetodid

2.1 Mikrobioom

Inimese soolestik on väga mitmekesine mikroorganismide- bakterite, seente, ning viiruste kogum. Kogu selle koosluse, mikrobioomi, väljakujunemine algab sünnist: algselt steriilne soolestiku keskkond asustatakse sünnitusteedest kaasa tulnud bakteritega ning juba lapseeas suureneb soolestiku mikrobioomi mitmekesisus märkimisväärselt [16]. Imiku puhul muutub mikrobioomi kooslus märgatavalt rinnapiimatoidult tahkele toidule üle minnes. Lisaks mõjutavad mikrobioomi kooslust eluea vältel paljud erinevad faktorid, sealhulgas vanus, sugu, ravimid, tervislik seisund, elustiil ja -keskkond. Suurim mõju soolestiku mikrobioomile on tarbitaval toidul. Inimese mikrobioom sisaldab hinnanguliselt sama palju rakke, kui on inimese enda kehas ning sadu kordi rohkem geene kui inimese genoomis. Läbi soolestiku mikroorganismide töö toodetakse meie organismis olulisi molekule nagu vitamiinid ja aminohapped, mis reguleerivad ainevahetuseprotsesse ning mõjutavad meie immuunsüsteemi [25].

Mitmed nii inimeste kui ka hiirte peal tehtud uuringud on seostanud muutusi mikrobioomis ainevahetuse haiguste, saelhulgas teist tüüpi diabeedi ning rasvumise kujunemisega [17]. Mikrobioomi efekti ning potentsiaali kirjeldab hiirte peal tehtud uuring, kus steriilses keskkonnas kasvanud hiirtele kanti üle ülekaaluliste hiirte mikrobioom ning vastavalt käitus ka fenotüüp: hiired rasvusid [24]. Hiljem on näidatud, et ka rasvunud inimestelt mikrobioomi üle kandmine steriilsetele hiirtele põhjustab hiirtel ülekaalulisust. Seega, sisuliselt on võimalik mikrobioomi üle kandes fenotüüpi muuta ka erinevate liikide vahel [20].

Inimese mikrobioomi ning terviseseisundite vahel täheldatud seoste tõttu ning uute sekveneerimistehnoloogiate tõttu on mikrobioomialaste uuringute arv jõudsalt kasvanud. Kui ajalooliselt on tulnud baktereid laboris kasvatada, siis nüüd võimaldab sekveneerimistehnoloogia kuluefektiivselt tuvastada erinevate bakterite koeksistentsi ning ka haruldasemata mikroobide olemasolu [16].

2.1.1 Mikrobioomi määramine 16S rRNA geeni sekveneerimise teel

Mikrobioomi andmete analüüsimine on võimalik tänu arengutele tehnoloogias ning bioinformaatika valdkonnas. Siiski on tehnoloogiline võimekus teatud määral piiratud, mistõttu erinevate analüüsimeetodite kasutamine peab arvestama andmete kogumise eripärade ning metoodikaga.

Käesolevas töös kasutatakse mikrobioomi määramiseks 16S rRNA geeni sekveneerimisel põhinevat lähenemist.

Mikrobioomi võib vaadata kui üksikute rakkude kogumit, millel igaühel on oma DNA. Sekveneerimise idee on proovis leiduva DNA abil teha kindlaks, millised mikroobid keskkonnas leiduvad ning kui palju neid on. Et iga raku terve genoomi sekveneerimine on töömahukas, kasutatakse teatavaid markereid, mis unikaalselt võimaldavad erinevaid mikroobe üksteisest eristada [16].

Levinuim selline marker on bakterite spetsiifiline 16S rRNA geen, milles asuvad nii unikaalsed kui ka hüpervarieeruvad piirkonnad. Unikaalsed piirkonnad võimaldavad proovis leiduva DNA järjestusi amplifitseerida polümeraasi ahelreaktsiooni abil. Hüpervarieeruvad piirkonnad aitavad vahet teha eri liikidel. Tulemuseks saadakse hulk erinevaid 16S rRNA geeni järjestusi ning iga järjestuse kohta arv, mitu korda teda detekteeriti [16].

2.1.2 Mikroobide operatiivsed taksonoomiaühikud (OTUd)

Saadud järjestuste puhul tekib küsimus, mis liiki mikroobiga on tegu ning kas teda on võimalik taksonoomiasse ehk liikide süstemaatikasse jaotada. Taksonoomia järgi lähedaste liikide puhul on ka 16S rRNA geeni hüpervarieeruvad osad sarnased ning võttes arvesse ka sekveneerimisel tehtavad vead, jääb iga unikaalse järjestuse erinevaks liigiks lugemine liialt konservatiivseks, mistõttu grupeeritakse erinevad järjestused teatud sarnasuse alusel. Saadakse järjestuste grupid, kus teatud määr erinevust on lubatud, näiteks loetakse ühte gruppi 97% ulatuses sarnased järjestused. Saadud gruppe nimetatakse operatiivseteks taksonoomiaühikuteks ehk OTUdeks [16].

2.2 Populatsiooni mitmekesisus

2.2.1 α-mitmekesisus

Üksikindiviidi mikrobioomi mitmekesisust nimetatakse α- mitmekesisuseks (ingl α- diversity). α- mitmekesisuse all mõistetakse indiviidi soolestikus olevate unikaalsete mikroobide arvu ehk liigirikkust (ingl *richness*) ning nende jaotumist proovis (ingl *evenness*) [16]. α- mitmekesisuse võrdlemine on esimeseks ja lihtsaimaks viisiks mikrobioomi kogukondade erinevuste hindamiseks [26]. Suurus pakub huvi, sest mitmete terviseseisundite puhul on täheldatud, et haiguseseisundiga on seotud vähenenud mikrobioomi mitmekesisus [16].

Unikaalsete mikroobide arvu hindamise puhul tuleb arvesse võtta sekveneerimistehnikate võimekusest tulenevaid piiranguid. Haruldasemate liikide puhul võib juhtuda, et nende amplifitseeritud järjestuste arv jääb sekveneerimisel alla masina võimekuse piiri. Seetõttu kehtib üldiselt järgmine seaduspära: mida rohkem ning täpsemalt proovi sekveneerida, seda liigirikkam on saadud mikrobioom ehk seda rohkem erinevat tüüpi organisme me näeme proovis olevat. Seetõttu on vaadeldud liigirikkuse puhul tegu alahinnanguga tegelikule liigirikkusele [16].

Probleemiga tegelemiseks kasutatakse hõrendamistehnikaid (ingl *rarefaction*) [10]. Selleks valitakse iga indiviidi proovist saadud lugemitest juhuslikult fikseeritud arv lugemeid, mis võimaldavad erinevaid proove üksteisega võrrelda.

Liigirikkuse näitajana kasutatakse nii kokkuloendatud erinevate OTUde arvu proovis kui ka kohandatud mõõdikuid, mis üritavad hinnata populatsiooni tegelikku mitmekesisust. Populaarsemaid kohandatud liigirikkust näitavaid indekseid on *Chao1* indeks [2]. *Chao1* indeks on mõeldud hindama populatsiooni mitmekesisuse alumist piiri. Indeks arvutatakse kujul:

$$S_{Chao1} = S_{obs} + \frac{f_1(f_1 - 1)}{2(f_2 + 1)}$$

Kus S_{obs} on vaadeldud unikaalsete liikide arv, f_1 on liikide arv, mida on nähtud ainult ühes proovis ning f_2 on liikide arv, mida on nähtud kahes proovis. Liidetava liikme idee on järgmine: kui sekveneerimisel leitakse endiselt haruldasi OTUsid, mida esineb proovis ainult korra, siis on tõenäolisem, et meil on veel liike leidmata. Kui kõiki liike on nähtud vähemalt kahe indiviidi proovides, siis on vähemtõenäolisem, et keskkonnas on haruldasi liike, mida proovis ei leitud

[2].

Teine aspekt, mida α-mitmekesisuse all mõistetakse, on liikide jaotuvus valimis. Liikide jaotumise uurimise eesmärk on hinnata, kas keskkonnas domineerivad mõned üksikud liigid või on bakterite arv keskkonnas üksteisele sarnane. Kõige mitmekesisemaks loetakse keskkonda, kus on suur liigirikkus ning kus liigid on arvukuse alusel ühtlaselt jaotunud. Üks populaarsemaid liigirikkust ning liikide jaotuvust arvesse võtvaid statistikuid on *Shannoni* entroopia, mis leitakse kujul:

$$H_{\text{Shannon}} = -\sum_{i}^{S} p_i(log_2(p_i))$$

kus p_i on i. OTU proportsioon valimis ning S on unikaalsete liikide arv valimis. Shannoni entroopia proovib hinnata, kui raske on õigesti hinnata liike indiviidi proovis, kelle jaoks veel mikrobioomi andmeid pole.

Shannoni entroopia käitumist kirjeldav näide. Olgu meil populatsioonis neli unikaalset liiki. Kui proovis leidub kõiki nelja liiki võrdselt ehk vaadeldud liikide proportsioonid on kujul $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, siis Shannoni entroopia $H_{\text{Shannon}} = -4(\frac{1}{4}log_2(\frac{1}{4})) = 2$. Järgnevalt, olgu proovis vaadeldud kõiki liike, aga üks liik on teistest rohkearvulisem, olgu liikide proportsioonid proovis järgmised: $(\frac{5}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Siis $H_{\text{Shannon}} = -\frac{5}{8}log_2(\frac{5}{8}) - 3\frac{1}{8}log_2(\frac{1}{8}) = 1.6$. Leidugu kolmandas proovis ainult kaks liiki ning olgu liikide proportsioonid proovis järgmised: $(\frac{1}{2}, \frac{1}{2}, 0, 0)$. Siis $H_{\text{Shannon}} = -2(\frac{1}{2}log_2(\frac{1}{2})) = 1$. Seega, mida rohkem liike on proovis vaadeldud ning mida ühtlasemalt liikide arvukus proovis jaotub, seda suurem on Shannoni entroopia väärtus.

2.2.2 β-mitmekesisus

Uurimaks haiguseseisundi seoseid mikrobioomiga soovitakse teada, kas tervete ja haigete inimeste mikroobikooslused erinevad. Mitme populatsiooni võrdlemisel räägitakse β- mitmekesisusest [11]. β -mitmekesisust võrreldakse üldiselt sarnasuskordajate abil.

Levinuim β -mitmekesisust võrdlev näitaja on Bray-Curtise eripära, mis hindab kahe indiviidi mikrobioomi koosluste kattuvust skaalal nullist üheni, kus 0 tähendab kahe populatsiooni kokkulangevust. Bray-Curtise eripära indiviidide

i1 ja *i2* vahel leitakse valemiga:

$$S_{BC}(i_1, i_2) = 1 - \frac{\sum_{j=1}^{J} |n_{i_1 j} - n_{i_2 j}|}{\sum_{j=1}^{J} (n_{i_1 j} + n_{i_2 j})}$$

kus n_{ij} on j. OTU arvukus indiviidide i. indiviidi proovis[16].

3 Teist tüüpi diabeet

Teist tüüpi diabeet on haigus, mille korral inimesel areneb insuliiniresistentsus ehk insuliini toime glükoosi lõhustamisel on nõrgenenud ning samuti on insuliini eritumine kõhunäärmest häiritud. See tingib kõrgenenud veresuhkru taseme ning üleliigse suhkru talletamise. Teist tüüpi diabeet on levinuim diabeedivorm, mida põeb Eestis ligikaudu 60-65 tuhat inimest. põhilised riskifaktorid on vanus üle 40 aasta, ülekaalulisus ning samuti on teist tüüpi diabeedi puhul oluline roll pärilikkusel [3].

Teist tüüpi diabeedile eelnev prediabeedi seisund võib kesta aastaid, mistõttu on ohumärkide tuvastamine suure tähtsusega ennetava ravi mõistes [17].

Diabeedieelsete seisundite tuvastamiseks kasutatakse glükoosi tolerantsuse testi OGTT (ingl lühendist *oral glucose tolerance test*). Enne testi sooritamist nõutakse patsiendilt paastumist, misjärel võetakse vereproov glükoositaseme määramiseks. Seejärel manustatakse patsiendile 75g glükoosilahust ning jälgitakse, kuidas muutub glükoositase veres kaks tundi pärast glükoosi manustamist. Pärast glükoosi manustamist hakkab keha tootma insuliini, mis hakkab glükoosi lagundama. Glükoosi manustamise eelse glükoosi konsentratsiooni ning manustamisejärgsete konsentratsioonide alusel määratakse inimesele diabeediseisund vastavalt tabelile 1 nelja kategooria vahel [22].

Tabel 1: Glükoosi taluvuse testi piirmäärad diabeediseisundi hindamiseks

	Paastuglükoos (mmol/l)	OGTT testi 2h glükoos (mmol/l)
Normaalne	$\leq$ 6.0	<7.8
Paastuglükoosi häire (IFG)	6.1 - 6.9	<7.8
Glükoositaluvuse häire (IGT)	<7.0	7.8 - 11.0
Diabeet	>7.0	>11.1

Vastavalt tabelile 1 peaks tervel inimesel glükoositase veres langema kahe

tunni möödudes alla 7.8 mmol/l, glükoositaluvuse häiretega inimestel ning diabeetikutel mitte nii märkimisväärselt. Lisaks käsitletakse prediabeedi vormina paastuglükoosi häiret, mille puhul glükoosi tase veres on normist kõrgem juba enne glükoosilahuse manustamist. Käesolevas tööd käsitletakse prediabeetikutena indiviide, kellel on paastuglükoosi häire, glükoositaluvuse häire või mõlemad.

4 Statistiline metoodika

4.1 Klassikaline- ja mitteparameetriline dispersioonanalüüs (ehk ADONIS)

Uuritava tunnuse keskmiste võrdlemiseks eelnevalt defineeritud gruppide vahel kasutatakse dispersioonanalüüsi. Olgu tegu andmestikuga, kus indiviidid jagunevad k gruppi ning pakub huvi, kas uuritava tunnuse y keskväärtus on grupiti erinev. Testitakse hüpoteeside paari:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \text{leiduvad grupid i ja j nii, et } \mu_i \neq \mu_j$$

kus μ_i on $i.$ grupi keskväärtus.

Dispersioonanalüüsi puhul jagatakse uuritava tunnuse kogu varieeruvus osadeks järgnevalt:

$$SS_T = SS_W + SS_B \tag{1}$$

kus SS_T on uuritava tunnuse varieeruvus üldkeskmisest $\bar{y}$, arvutatud kujul:

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \tag{2}$$

SS_W on uuritava tunnuse varieeruvus grupi keskmisest $\bar{y}_i$, arvutatud kujul:

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \tag{3}$$

ning SS_B on gruppide keskmiste erinevus üldkeskmisest, arvutatud kujul:

$$SS_B = \sum_{i=1}^{k} \sum_{j=1}^{n_i} n_i (\bar{y}_i - \bar{y})^2 \tag{4}$$

kus y_{ij} on uuritava tunnuse väärtus $i.$ grupi $j.$ objektil ning k on gruppide arv.

Põhiidee seisneb uuritava tunnuse grupisisese varieeruvuse SS_W ja gruppide-

vahelise varieeruvuse SS_B võrdlemises. Kui keskmine gruppidevaheline varieeruvus on oluliselt suurem kui keskmine gruppidesisene varieeruvus, on põhjust arvata, et gruppides on erinevad keskmised. Selle testimiseks konstrueeritakse F-statistik kujul:

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

kus k on gruppide arv ning n on vaatluste arv gruppides. Mida suurem on F-statistiku väärtus, seda tõenäolisemalt lükkame ümber H_0, mis väidab, et gruppide keskmised on samad.

Mikrobioomi andmete puhul defineeritakse kahe indiviidi mikrobioomi erinevus teisiti kui eeldab klassikaline dispersioonanalüüs. Eukleidilise kauguse asemel kasutatakse erinevuste kirjeldamiseks sageli teist tüüpi näitajaid nagu Bray-Curtise eripära. Probleem seisneb asjaolus, et näitajate nagu Bray-Curtise eripära puhul on keeruline leida grupi keskmist ning sellest lähtuvalt on probleemne klassikaline varieeruvuse osadeks lahutamine. Seetõttu kasutatakse mikrobioomi andmete puhul dispersioonanalüüsile analoogset mitteparameetrilist meetodit.

Mitteparameetrilise meetodi erisus võrreldes tavalise parameetrilise dispersioonanalüüsi juhuga tuleneb punktidevaheliste kauguste ruutude summa erinevast lahutusest. Järgnevalt avaldame vaatluste y_i hälvete ruutude summa vaatluste omavaheliste kauguste kaudu:

$$\begin{aligned} SS_T &= \sum_{i=1}^{N}(y_i - \bar{y})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(y_i - y_j)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \end{aligned}$$

kus d_{ij} on $i.$ ja $j.$ vaatluste vaheline kaugus ning N on vaatluste koguarv. See tähendab, et ruutude summa kaugus gruppide keskmisest on võrdne punktidevaheliste kauguste ruutude summaga, mis on jagatud punktide arvuga.

Asendades d_{ij} mõne teise kaugusemõõdikuga d_{ij}^*, saame:

$$SS_T^* = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} {d_{ij}^*}^2$$

Analoogiliselt avaldub grupisisene hälvete ruutude summa kujul:

$$SS_W^* = \frac{1}{n} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} {d_{ij}^*}^2 \epsilon_{ij}$$

kus $\epsilon_{ij} = 1$, kui vaatlused i ja j on samast grupist ning 0, kui i ja j on erinevast grupist.

Siis gruppidevaheline hälvete ruutude summa SS_B^* arvutatakse kujul $SS_B^* = SS_T^* - SS_W^*$ ning pseudo F-statistik arvutatakse analoogselt dispersioonanalüüsile kujul:

$$F^* = \frac{SS_B^*/(k-1)}{SS_W^*/(N-k)}$$

Nõnda defineeritud F-statistik annab Eukleidilise kauguse puhul täpselt sama tulemuse kui traditsiooniline dispersioonanalüüsi mudeli esitus. Samuti käitub niinimetatud pseudo F-statistik analoogselt ka teiste kauguse mõõdikutega nagu Bray-Curtis: kui gruppides on erinevused, siis gruppidevaheline kaugus SS_B^* on suhteliselt suur võrreldes gruppidesiseste kaugusega SS_W^* ning F-statistiku väärtus on seetõttu suur [1].

Eelnevalt defineeritud F-statistik ei ole nullhüpoteesi kehtides sama jaotusega, mis konventsionaalne F-statistik, sest Eukleidilise kauguse asemel kasutatakse bioloogiliselt relevantsemaid erinevuse näitajaid nagu Bray-Curtise eripära ning ühtegi jaotuse eeldust pole tehtud [1]. Eelnevalt mainitud põhjustel pole võimalik kasutada F-jaotusel põhinevaid p-väärtuseid.

Olulisustõenäosuste leidmiseks on võimalik luua statistiku jaotus nullhüpoteesi kehtides kasutades vaatluste permutatsioone. Eeldades, et kehtib nullhüpotees ning gruppide vahel pole erinevusi, võiksime vaatlusi erinevate gruppide vahel vahetada, permuteerida. Permuteerimise idee: vahetame suvaliselt gruppi kuulumist näitavaid indikaatoreid ridade vahel ning arvutame iga kord eelnevalt defineeritud pseudo F-statistiku F^{uus}. Kui sellisel kujul gruppe korduvalt redefineerida ning arvutada F-statistik iga võimaliku permutatsiooni jaoks, saame nullhüpoteesi kehtides taasluua F-statistiku jaotuse meie andmete jaoks [1].

Võrreldes esialgse, permuteerimata andmestikul arvutatud F-statistikut F leitud jaotusega, arvutatakse p-väärtus kujul:

$$P = \frac{\#F^{\text{uus}} \geq F}{\#F^{\text{uus}}}$$

4.2 Regressioonmudelid uuritava tunnuse modelleerimiseks

4.2.1 Lineaarne regressioon

Pidevate tunnuste modelleerimiseks kasutatakse käesolevas töös lineaarset regressiooni. Regressioonanalüüs kirjeldab uuritava tunnuse Y keskväärtuse muutumist, kui vaadelda erineva X-tunnuse väärtusega objekte. Lineaarse regressiooni üldkuju on järgmine:

$$EY = \beta_0 + x^T \vec{\beta}$$

kus EY on uuritava tunnuse keskväärtus. Uuritav tunnus Y eeldatakse olevat normaaljaotusega ning vaatlused eeldatakse olevat sõltumatud.

Parameetrid $\beta_0, \vec{\beta}$ hinnatakse vähimruutude meetodil.

4.2.2 Logistiline regressioon

Binaarse tunnuse modelleerimiseks kasutatakse käesolevas töös logistilise regressiooni mudelit:

$$log\frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + x^T \vec{\beta}$$

Kus $Y = 1$ tähistab uuritava sündmuse, käesolevas töös prediabeedi, esinemist. Parameetrite $\beta_0, \vec{\beta}$ hindamiseks maksimeeritakse logaritmilist tõepärafunktsiooni:

$$l(\beta_0, \vec{\beta}^T) = \Big\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \vec{\beta}^T x_i) - log(1 + e^{\beta_0 + \vec{\beta}^T x_i}) \right]$$

4.2.3 Regulariseeritud logistiline regressioon (LASSO)

Lisaks logistilisele regressioonile modelleeritakse käesolevas töös binaarset tunnust regulariseeritud logistilise regressiooniga (LASSO logistiline regressioon).

LASSO on üks regulariseeritud regressiooni meetodeid, mille puhul seatakse piirang hinnatud parameetrite absoluutväärtuste summale:

$$\sum_j |\beta_j| \leq t$$

Sellisel juhul leitakse logistilise regressiooni parameetrite väärtused maksimiseerides niiöelda "karistatud"logaritmilist tõepärafunktsiooni:

$$l(\beta_0, \vec{\beta}^T) = \Big\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \vec{\beta}^T x_i) - log(1 + e^{\beta_0 + \vec{\beta}^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \Big\}$$

Kus λ väärtus saadakse ristvalideerimise teel arvutades teatava hulga λ-väärtuste jaoks hälbimus (ingl *deviance*), mis on defineeritud kui $-2l(\beta_0, \vec{\beta}^T)$, ning valides λ, mille korral vastava mudeli hälbimus partitsioonil on väikseim [7].

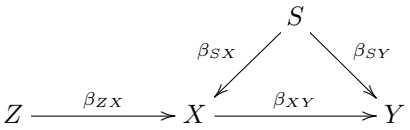
Regulariseeritud logistilise regressiooni puhul kasutatav piirang parameetritele osutub kasulikuks mudeli valiku seisukohalt: piisavalt väike t sunnib mudeli hindama parameetreid selliselt, et osad koefitsiendid β_j on nullid [8]. LASSO regressioon aitab valida mudelisse vaid kõige olulisemad tunnused. Lisaks, LASSO regressioon võimaldab parameetreid efektiivsemalt hinnata juhul, kus hinnatavate parameetrite arv võrreldes valimimahuga on suur ning kui argumenttunnused on omavahel sõltuvad. Mikrobioomi analüüsimisel annab mudeli kasutamine alternatiivse variandi hindamaks, millised OTUd on enim uuritava fenotüübiga seotud.

4.3 Põhjuslike mõjude hindamine Mendeli randomiseerimise meetodil

Mendeli randomiseerimine on uuringukavand, kus geneetilist riskiskoori kasutatakse instrumenttunnusena hindamaks nähtud seose põhjuslikku mõju.

Vaatleme olukorda, kus soovitakse hinnata tunnuse X põhjuslikku mõju tunnusele Y. Olukorras, kus nii X kui Y võivad olla mõjutatud samade niinimetatud segavate tunnuste S poolt, ei saa vaid X ja Y omavahelist seost uurides eristada huvipakkuvat põhjuslikku mõju S kaudu tekkinud korrelatsioonist. Üks võimalus

põhjusliku seose hindamiseks on instrumenttunnuse Z kasutamine. Mendeli randomiseerimise seisukohalt on olulised kaks eeldust: esiteks, instrumenttunnus on sõltumatu segavatest tunnustest S. Teiseks, instrumenttunnus Z peab olema seotud argumenttunnusega X, kuid tal ei tohi olla otsest põhjuslikku uuritavale tunnusele Y. Ainus võimalik seos võib olla läbi tunnuse X [21]. Skemaatiliselt võib antud olukorda esitada järgmiselt [9]:



Lihtsal lineaarsel juhul on toodud skeem esitatav regressioonivõrranditena kujul:

$$Y = \beta_Y + \beta_{XY}X + \beta_{SY}S + \epsilon_Y$$
$$X = \beta_X + \beta_{ZX}Z + \beta_{SX}S + \epsilon_X$$

Siit

$$\begin{aligned} E(Y|X) &= E(\beta_Y + \beta_{XY}X + \beta_{SY}S + \epsilon_Y|X) \\ &= \beta_Y + \beta_{XY}X + \beta_{SY}E(S|X) \end{aligned}$$

Võttes arvesse instrumenttunnuse Z kohta tehtud eeldust, mille järgi instrumenttunnus ei ole seotud segavate tunnustega S, saab kirjutada:

$$\begin{aligned} E(X|Z) &= E(\beta_X + \beta_{ZX}Z + \beta_{SX}S + \epsilon_X|Z) \\ &= \beta_X + \beta_{ZX}Z + \beta_{SX}E(S|Z) \\ &= \beta_X + \beta_{ZX}Z \end{aligned}$$

Eelneva põhjal:

$$\begin{aligned} E(Y|Z) &= E(\beta_Y + \beta_{XY}X + \beta_{SY}S + \epsilon_Y|Z) \\ &= \beta_Y + \beta_{XY}E(X|Z) + \beta_{SY}E(S|Z) \\ &= \beta_Y + \beta_{XY}E(X|Z) \\ &= \beta_Y + \beta_{XY}(\beta_X + \beta_{ZX}Z) \\ &= \beta_Y^* + \beta_{XY}\beta_{ZX}Z \end{aligned}$$

Saame hinnata regressioonmudelid, kus tunnuse X modelleerimiseks kasutatakse argumenttunnusena tunnust Z ning tunnuse Y modelleerimiseks kasutatakse argumenttunnusena tunnust Z. Nimetatud regressioonmudelitega hinnatakse vastavalt suurusi β_{ZX} ja $\beta_{XY}\beta_{ZX}$. Seejärel leitakse kordaja β_{XY} järgnevalt:

$$\beta_{XY} = \frac{\widehat{\beta_{XY}\beta_{ZX}}}{\widehat{\beta_{XY}}}$$

4.4 Kompositsionaalsete andmete analüüs

Üldiselt kasutatakse 16S rRNA analüüside puhul lähenemist, kus OTUde kohta saadud lugemite arvu käsitletakse loendusandmetena ning normaliseeritud lugemite arve modelleeritakse Poissoni või Negatiivse-Binoomjaotusega [5]. Sellisel kujul saadud tulemused ei ole tihtipeale reprodutseeritavad uutel andmetel [5].

Sekveneerimistehnoloogia puudujääkide tõttu lugemite koguarv proovis ei ole huvipakkuv suurus, sest ta sõltub otseselt sekveneerimisplatformi võimekusest [5]. Samal põhjusel ei paku huvi absoluutne erinevus OTUde lugemite arvude vahel. Seetõttu on korrektsem uurida ainult relatiivset informatsiooni, mida OTUde lugemite arvud endas kannavad. Sellisel juhul peetakse andmeid kompositsionaalseks [5].

4.4.1 Kompositsionaalsed andmed

Vektor $\vec{x} = [x_1, x_2, ..., x_D]$ on defineeritud kui D-osaline kompositsioon, kui kõik ta komponendid on rangelt positiivsed reaalarvud ning nad kannavad ainult relatiivset informatsiooni [18]. Sellisel kujul andmed paiknevad simpleksil, mis on defineeritud järgnevalt:

$$S^D = \left\{x = [x_1, x_2, ..., x_D] | x_i > 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x_i = k\right\}$$

Suurim probleem kompositsionaalsete andmetega tegelemisel on asjaolu, et andmestiku punktid ei ole Eukleidilise ruumi osa, vaid punktid asetsevad konstantse summa eelduse tõttu niinimetatud Aitchinsoni simpleksil [5]. John Aitchison oli šoti matemaatik, kes 1986. aastal sõnastas põhiprintsiibid kompositsionaalsete andmete analüüsiks.

Kompositsionaalseid andmeid analüüsides tuleb arvestada teatud probleemidega. Esiteks, kompositsionaalseid andmeid tuleb analüüsida skaalainvariantselt. Seega proportsionaalsete komponentidega vektorid $(\frac{1}{3}, \frac{2}{3})$ ning $(10, 20)$ on ekvivalentsed. Teiseks, elemendid kompositsioonis ei ole omavahel sõltumatud tingituna konstantse summa eeldusest. Kolmandaks, alamkompositsioonil tehtavad järeldused peavad olema kooskõlas kogu kompositsioonilt tehtavate järeldustega [5].

Aitchinson mõistis, et vaadates proportsioonide suhteid, on võimalik eelnevatest takistustest üle saada või nõudeid leevendada. Üks enamkasutatavaid transformatsioone on tsentreeritud log-suhte transformatsioon (ingl *centered log-ratio, CLR*), mille puhul element x_i jagatakse vektori $\vec{x}$ elementide geomeetrilise keskmisega ning seejärel võetakse jagatisest logaritm:

$$CLR(\vec{x}) = (q_1, q_2, ..., q_D)$$

kus element q_i avaldub kujul:

$$q_i = log\frac{x_i}{gm(\vec{x})}$$

kus $gm(\vec{x})$ tähistab vektori $\vec{x}$ elementide geomeetrilist keskmist:

$$gm(\vec{x}) = \sqrt[D]{x_1 x_2 \cdots x_D}$$

Saadakse nulli ümber tsentreeritud väärtused, millest negatiivsed väärtused näitavad valimikeskmisest vähem arvukamaid elemente ning positiivsed väärtused keskmisest arvukamaid elemente [5].

Näide demonstreerimaks CLR-transformatsiooni mõju alamkompositsiooni-

le. Olgu meil OTUde lugemite vektor kolme OTU jaoks kujul $x = [5, 40, 300]$, mida käsitleme proportsioonide vektorina $p_x = [0.0145, 0.1159, 0.8696]$ mille proportsioonide summa on 1. Siin esimese ning teise elemendi vahe on -0.1014. Uurides aga ainult esimest kahte elementi, saame uueks proportsioonide vektoriks $p_x = [0.1250, 0.8750]$, kus elementide vahe on märksa suurem - 0.625, mis jätab tunde justkui oleks nähtud vahe märksa suurem. Vaadates andmeid kui kompositsioone ehk proportsioonide suhteid ning rakendades andmetele CLR teisendust logaritimiga alusel 2, saame esimesel juhul $clr_x = [-2.9690, 0.0310, 2.2938]$ ning eemaldades viimase elemendi: $clr_x = [-1.5, 1.5]$. Tähele tasub panna, et CLR transformatsiooni korral on esimese kahe elemendi vahe mõlemal juhul 3, mistõttu tehtav otsus ei muutu [5].

Mikrobioomi lugemite andmete puhul on probleemiks rohkete nullide olemasolu andmestikus. Seetõttu ei saa lugemite andmetele otseselt rakendada CLR transformatsiooni, mis hõlmab geomeetrilise keskmise arvutamist ning logaritmi võtmist [5]. Järgnevalt kirjeldatud metoodika lahendab nullide probleemi ning teeb võimalikuks CLR transformatsioonil põhinevad analüüsid.

4.4.2 Kompositsionaalsete andmete modelleerimine

Huvitagu meid lugemite suhtelised sagedused p_i, mitte lugemite üldarvud x_i. Iga valimi jaoks lugemite vektor $\vec{x}$ on multinomiaalse jaotusega parameetritega $[p_1, p_2, ..., p_m]$ ja $n = \sum_{i=1} x_i$. Multinomiaalne jaotus kirjeldab olukorda, kus n sõltumatu sündmuse jaoks on m võimalikku tulemit ning p_i vastab i.-tulemi esinemise tõenäosusele [6]. Sellisel juhul avaldub vektori $\vec{x}$ tihedusfunktsioon kujul

$$f(x_1, x_2, ..., x_m | n, \vec{p} = (p_1, p_2, ..., p_m)) = \frac{n!}{x_1! x_2! ... x_m!} \prod_{i=1}^{m} p_i^{x_i} \tag{5}$$

Parameetri p_i hinnanguks suurima tõepära meetodiga tuleb $p_i = x_i / \sum_m x_i$. Probleemiks suurima tõepära hinnangu puhul on suur nullide osakaal andmestikus, mistõttu suurima tõepära hinnangud tulevad ebatäpsed [4]. Nullide puhul andmestikus tuleb ka $x_i = 0$ korral suurima tõepära hinnang parameetrile p_i null, kuid mikrobioomi andmete analüüsimisel on eeldus, et keskkonnas teatud mikroobe ei ole, liiga range. Sekveneerimistehnoloogia tundlikkuse ning võetud

proovi juhuslikkuse tõttu on loomulikum eeldada, et bakter on pigem keskkonnas haruldane, kuid ta ei puudu sealt täielikult.

Seetõttu kasutatakse edasises kirjeldatud meetodi puhul suurima tõepära meetodi asemel Bayes'i statistika tehnikaid tegemaks järeldusi proportsioonide vektori $[p_1, p_2, ..., p_m]$ kohta [4].

4.4.3 Proportsioonide hindamine Bayesi meetodil

Bayesi statistika põhiidee on, et mudeli parameetreid ei saa täpselt hinnata, sest tegu ei ole konstantidega. Seetõttu esitatakse parameetrid jaotusega ning nende kohta tehakse tõenäosuslikke otsuseid [23].

Oluline idee Bayesi statistika vaatenurgast on eelinfo kasutamine hinnatavate parameetrite kohta kasutades Bayesi teoreemi:

$$p(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)p(\theta)}{f(\vec{x})} \tag{6}$$

kus $p(\theta)$ on parameetrite eeljaotus, mis põhineb eelteadmistel parameetrite võimaliku jaotuse kohta, $f(\vec{x}|\theta)$ on vaatlusandmete jaotus, $f(\vec{x})$ on vektori $\vec{x}$ marginaalne tihedusfunktsioon ning $p(\theta|\vec{x})$ märgib parameetervektori järeljaotust, mis kombineerib vaatlusandmete pealt nähtu eelteadmistegaparameetrite kohta [23].

Saadud võrdust esitatakse sageli kujul:

$$\pi(\theta) \equiv p(\theta|\vec{x}) \propto f(\vec{x}|\theta)p(\theta) \tag{7}$$

kus võrdeteguriks on konstant $\frac{1}{f(\vec{x})}$. Alternatiivse esituse põhjuseks on asjaolu, et simuleerimaks andmeid järeljaotusest $\pi(\theta)$, ei ole vaja teada konstanti $\frac{1}{f(\vec{x})}$ [23]. Käesolevaga meetodis kasutatakse parameetrite $[p_1, p_2, ..., p_m]$ kohta eelteadmiste kirjeldamiseks Dirichlet' jaotust [4].

Vektor $\vec{p} = [p_1, p_2, ..., p_m]$, kus $p_i \geqslant 0 \forall i$ ning $\sum_{i=i}^{m} p_i = 1$ on definitsiooni järgi Dirichlet jaotusega parameetriga $\vec{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_m]$ (kus $\alpha_i > 0 \forall i$ ning $\alpha_0 = \sum_{i=i}^{m} \alpha_i$), kui tema tihedusfunktsioon avaldub kujul:

$$f(\vec{p}, \vec{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} p_i^{\alpha_i - 1} \tag{8}$$

kus $\Gamma(x)$ on gamma-funktsioon [6].

Kombineerides multinomiaaljaotust kui vaatlusandmete jaotust Dirichlet' jaotuse kui parameetrite eeljaotusega, räägitakse saadud mudelist kui Dirichlet-multinomiaalmudelist. Kui vaatluste jaotus on multinomiaaljaotusega $(\vec{x}|\vec{p}) \sim Mn(n = \sum_i^m x_i, \vec{p})$ ning parameetervektori $\vec{p}$ eeljaotuseks on Dirichlet jaotus $\vec{p} \sim Dir(\alpha)$, siis ka parameetervektori järeljaotus on Dirichlet jaotusega juhuslik suurus.

Olgu $\vec{x} \sim Mn(n = \sum_i^m x_i, [p_1, p_2, ..., p_m])$ ning $[p_1, p_2, ..., p_m] \sim Dir(\alpha_1, \alpha_2, ..., \alpha_m)$. Siis võttes arvesse jaotuste tihedusfunktsioonide kujusid 5 ja 8 ning Bayesi teoreemi esitust kujul 6, saame

$$\begin{aligned}
\pi(\vec{p}|\vec{x}) &= \gamma f(\vec{x}|\vec{p})p(\vec{p}) \\
&= \gamma\Big(\frac{n!}{x_1!x_2!...x_m!}\prod_{i=1}^{m} p_i^{x_i}\Big)\Big(\frac{\Gamma(\alpha_0)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m} p_i^{\alpha_i-1}\Big) \\
&= \tilde{\gamma}\prod_{i=1}^{m} p_i^{\alpha_i+x_i-1} \\
&= Dir(\vec{\alpha}+\vec{x})
\end{aligned}$$

Seega ka parameetrite järeljaotus on Dirichlet jaotus. Saadud jaotusest andmete simuleerimisega saame proportsioonide vektorid, mis on iga indiviidi jaoks kooskõlas nähtud lugemite arvudega.

4.4.4 Bayesi metoodika rakendamine tarkvara R abil

Eelnevalt kirjeldatud sammud on implementeeritud tarkvara *R* paketis *ALDEx2*.

Esmalt muudetakse iga vektori $\vec{p}$ element p_i tõenäosusjaotuseks kasutades selleks 128 Monte Carlo simulatsiooni Dirichlet' jaotusest. Iga elemendi p_i jaoks saadakse 128 replikaati p_{ij}. Parameetrite eeljaotuse parameetrite ehk hüperparameetritena $\alpha = [\alpha_1, \alpha_2, ..., \alpha_m]$ kasutatakse vektorit $\alpha = [0.5, 0.5, ..., 0.5]$. Sellise hüperparameetrite valiku puhul on näidatud, et ta maksimiseerib andmetes leiduva informatsiooni minimiseerides parameetrite eeljaotuse mõju järeljaotusele. Samuti kirjeldab selline hüperparameetrie valik, et kõik elemendid vektorist $\vec{x}$ pakuvad meile ühesuguselt huvi [4].

Seejärel rakendatakse realisatsioonidele p_{ij} CLR-transformatsiooni saades transformatsiooni tulemuseks vektori elementidega q_{ij}..

ALDEx2 pakub kahe grupi võrdlemiseks kahte lähenemist. Esimesel juhul rakendatakse väärtustele q_{ij} t-testi ning Wilcoxoni astaksummatesti. Saadakse olulisustõenäosuste jaotus, mille puhul raporteeritakse jaotuse mediaanväärtus. Samuti raporteeritakse Benjamini-Hochbergi parandusega olulisustõenäosuste mediaanväärtused.

Teise lähenemise puhul arvutab *ALDEx2* lisaks välja dispersioonanalüüsile analoogsed grupisisese ning gruppidevahelise erinevuse näitajad ning vastava efektisuuruse. Olgu elementi kirjeldav indeks i: $i = 1, 2, ..., m$, gruppi indikeeriv indeks j: $j = 1, 2, ..., k$ ning olgu l indeks iniviidi jaoks j. konditsioonis: $l = 1, 2, ..., L_j$.

Kasutatakse järgevalt defineeritud suurusi: Grupisisene summa (ingl *within-condition mixture*):

$$W(i,j) = \sum_{l=1}^{L_j} q_{ijl}$$

Gruppidevaheline erinevus (ingl *absolute fold difference between-condition*):

$$\Delta_A(i, j_1, j_2) = W(i, j_1) - W(i, j_2)$$

Grupisisisene indiviididevaheline erinevus:

$$\Delta_W(i,j) = \max_{l \neq l'} |q_{ijl} - q_{ijl'}|$$

Suhteline efekti suurus:

$$\Delta_R(i, j_1, j_2) = \frac{\Delta_A(i, j_1, j_2)}{\max(\Delta_W(i, j_1), \Delta_W(i, j_2))}$$

Tegu on juhuslike suurustega, mille jaotusi hinnatakse Monte Carlo realisatsioonide kaudu.

4.5 Seosenäitajad kompositsionaalsete andmete korral

Bioloogidele pakub huvi, kas leidub baktereid, mis kipuvad esinema soolestiku keskkonnas samaaegselt. Teadmine bakterite koosesinemisest annaks uurijale

aimu käsitlemaks seotud mikroobide gruppi ühtse kogumina. Lugemite koguarvu sõltuvus amplifitseerimistehnoloogiatest ning valimi kogumisega kaasnevast juhuslikkusest ning saadavate lugemite arvu mitteolulisus mõjutab ka koossesinemise uurimist - proportsioonide sõltuvuse tõttu ei saa antud küsimusele vastamiseks kasutada klassikalisi lähenemisi nagu Pearsoni ja Spearmani korrelatsioon. Samuti pole põhjendatud kasutada lugemite arvu leidmaks konventsionaalsel meetodil korrelatsioone.

Seetõttu tuleb kasutada mikrobioomi andmete puhul koosesinemist uurides teistsuguseid meetodeid: ühe võimaliku variandina Aitchisoni pakutud logsuhte dispersioon $D[log(x/y)]$. Kui x ja y on täpselt proportsionaalsed, siis oleks $D[log(x/y)] = 0$:

Kui $x = ky$, kus $k \in \mathbb{R}^+$, siis $log(x/y) = log(k)$ ning $D(log(k)) = 0$.

Dispersiooni $D[log(x/y)]$ saab lahutada osadeks järgnevalt:

$$\begin{aligned} D[log(x/y)] &= D[log(x) - log(y)] \\ &= D[log(x)] + D[log(y)] - 2cov[log(x), log(y)] \\ &= D[log(x)] * (1 + \frac{D[log(y)]}{D[log(x)]} - 2\sqrt{\frac{D[log(y)]}{D[log(x)]}}\frac{cov[log(x), log(y)]}{\sqrt{D[log(x)] * D[log(y)]}}) \\ &= D[log(x)] * \Phi(log(x), log(y)) \end{aligned}$$

Proportsionaalsuse ulatust kirjeldav statistik Φ avaldub kujul:

$$\Phi(log(x), log(y)) = D[log(x/y)]/D[log(x)]$$

5 Mikrobioomi andmete analüüs

5.1 Andmestiku kirjeldus

Töös on kasutatud METSIM kohordi andmeid 531 indiviidi kohta. METSIM on Ida-Soomes kogutud populatsioonipõhine kohort meestest vanuses 45-70 eluaastat, kelle kohta on kogutud hulk metaboolseid parameetreid ning kellelt on võetud mikrobioomi analüüsiks väljaheiteproov [17].

Mikrobioomi DNA ekstraheeriti iga indiviidi jaoks kasutades PowerSoil DNA isoleerimise komplekti. Seejärel amplifitseeriti 16S rRNA geeni hüpervarieeruvat piirkonda V4 kasutades 515 F/806 praimerit. DNA sekveneerimiseks kasutati Illumina MiSeq platformi, sekveneerimine viidi läbi UCLA ülikooli genotüpiseerimise tuumiklaboris. 16S rRNA geeni sekventside kvaliteedikontroll ning OTUdesse klasterdamine viidi läbi kasutades QIIME vabavaralise platformi versiooni 1.7.0. Sekventsi lugemite koguarv oli 12 785 442 (keskmiselt 21 543 proovi kohta) keskmise pikkusega 153 aluspaari. Sekventsid klasterdati OTUdesse kasutades 97% täpsust ning kõrvutades neid Greengenes referents andmebaasiga [17]. Lisaks on töös kasutada on NMR-spektroskoopia platvormiga mõõdetud aminohapete konsentratsioonid plasmas üheksa aminohappe jaoks, mis on toodud tabelis 3.

5.1.1 Kasutatav tarkvara ja andmestiku eeltöötlus

METSIM kohordi indiviididest 15 olid glükoosi tolerantsuse testi järgi juba diabeetikud. Et töö eesmärk on uurida, kas mikrobioomi puhul on märgatavaid muutusi juba teist tüüpi diabeedile eelnevate seisundite puhul, eemaldati diabeedi diagnoosiga indiviidid edasistest analüüsidest. Lisaks ei kasutatud andmeid kahe indiviidi jaoks, kellel ei olnud fenotüübiandmeid. Ülejäänud indiviididest 164 olid terved, 287 paastuglükoosi häirega, 15 glükoositaluvuse häirega ning 48 indiviidil esines nii paastuglükoosi- kui ka glükoositaluvuse häire. Analüüside jaoks klassifitseeriti paastuglükoosi häire ning glükoositaluvuse häire mõlemad prediabeedina ehk lõplikus andmestikus on andmeid 350 prediabeetiku ning 164 terve indiviidi kohta.

Mikrobioomi mitmekesisuse (α- ja β- mitmekesisus) analüüsis kasutati hõrendamist (ingl *rarefaction*) võttes iga indiviidi lugemite arvudest sama suur arv lu-

gemeid kui minimaalse lugemite arvuga valimis, kokku 10077 lugemit. Muudeks analüüsideks kasutati OTUsid, mis esinesid vähemalt 50 protsendil valimitest, kokku 354 OTUt. Valiku eesmärk on vähendada müra andmetes.

Analüüsiks kasutati statistikatarkvara R (versioon 3.4.3). Põhiliseks andmetöötluseks kasutati paketti *phyloseq*.

5.2 Kirjeldav analüüs

5.2.1 Fenotüübiandmete kirjeldav analüüs

Tabelis 2 on esitatud analüüsitavate fenotüüpide keskväärtused ning standardhälbed tervete indiviidide ning prediabeetikute jaoks. Lisaks on toodud Wilcoxoni astakmärgitesti olulisustõenäosused testimaks, kas vastava fenotüübi keskväärtused gruppides on erinevad. Nähtu on kooskõlas teist tüüpi diabeeti iseloomustavate põhitunnustega, mis on toodud peatükis 3.

Tabel 2: Fenotüübiandmete kirjeldav statistika

Fenotüüp	Terve		Prediabeet		P-väärtus
	Keskmine	Std	Keskmine	Std	
Vanus	61.95	0.44	62.01	0.29	0.9830
Kehamassiindeks	26.79	0.25	28.34	0.20	<.0001
Vöö- ja puusaümbermõõdu suhe	0.98	0.00	1.00	0.00	0.0020
Rasvaprotsent	24.50	0.49	26.34	0.38	0.0041
Glükoosi tase paastuplasmas	5.25	0.02	5.97	0.02	<.0001
Glükoosi tase plasmas 30min	8.56	0.11	9.63	0.07	<.0001
Glükoosi tase plasmas 120min	5.03	0.09	6.17	0.09	<.0001
Insuliini tase paastuplasmas	7.20	0.32	10.29	0.34	<.0001
Insuliini tase plasmas 30min	60.44	3.14	67.42	2.32	0.0450
Insuliini tase plasmas 120min	31.20	2.13	53.45	2.74	<.0001
HbA1c	37.16	0.24	37.92	0.17	0.0107
Süstoolne vererõhk	128.62	1.08	130.97	0.72	0.0825
Diastoolne vererõhk	81.07	0.56	83.03	0.43	0.0176
Geneetiline riskiskoor	1.70	0.04	1.84	0.03	0.0126

P-väärtus on Wilcoxoni astaksumma testi olulisustõenäosus testimaks erinevusi tervete ning prediabeetikute vahel

Esiteks, prediabeetikutel on tervetest inimestest keskmiselt kõrgemad rasvumisele viitavad näitajad: kehamassiindeks, keha rasvaprotsent ja piha-puusa ümbermõõtude suhe. Samuti on prediabeetikutel kõrgemad süstoolse- ehk maksimaalse vererõhu ning diastoolse- ehk minimaalse vererõhu näitajad. Vanuse mõju prediabeedile METSIM andmestikus ei täheldata, kuid arvatavasti on

põhjuseks kohorti kaasatud meeste väike vanusevahe, mistõttu vanus märkimisväärset efekti ei näita. Teist tüüpi diabeedi päritavuse efekti kirjeldab teist tüüpi diabeediga seotud geenimarkeritelt arvutatud riskiskoor, mis on prediabeetikutel tõestatavalt kõrgem kui tervetel inimestel. Geneetiline riskiskoor on arvutatud ligikaugu 7500 ühenukleotiidse polümorfismi kineaarkombinatsioonina nii nagu kirjeldatud artiklis [14].

Erineval ajahetkel glükoosi ning insuliini taset plasmas mõõtvad näitajad ning glükeeritud hemoglobiin HbA1c on toodud kirjeldamaks erinevusi prediabeetikute ning tervete indiviidide glükoosi lõhustamise ja insuliini tootmise protsessis. Glükeeritud hemoglobiin ehk glükohemoglobiin iseloomustab keskmist glükoositaset paari viimase kuu jooksul enne proovi võtmist [12]. Prediabeetikutel on kõrgem keskmine paastuglükoosi tase ning glükoosi tase kaks tundi pärast glükoosi manustamist püsib kauem baastasemest kõrgemal kui tervetel inimestel. Glükoosi manustamise järgselt suureneb kehas insuliini tootmine märgatavalt. Prediabeetikutel on kahe tunni möödudes insuliini tase veres kõrgem kui tervetel inimestel, sest glükoosi lõhustamise protsess on alles pooleli kõrgema glükoositaseme tõttu. Tervetel inimestel on enamus manustatud glükoosist lõhustatud ning insuliinitase veres hakanud langema. Samuti on prediabeetikutel statistiliselt tõestatavalt kõrgem glükohemoglobiini tase veres.

Lisaks mõõdeti kõikidelt uuritavatelt vereplasmast NMR meetodil aminohapete konsentratsioonid. Tabelis 3 on toodud uuritud aminohapete konsentratsioonide keskväärtused ja standardhälved prediabeetikute ning tervete inimeste jaoks. Sarnaselt fenotüübi- andmetele kasutatakse Wilcoxoni mitteparameetrilist testi uurimaks, kas aminohapete konsentratsioonid grupiti erinevad.

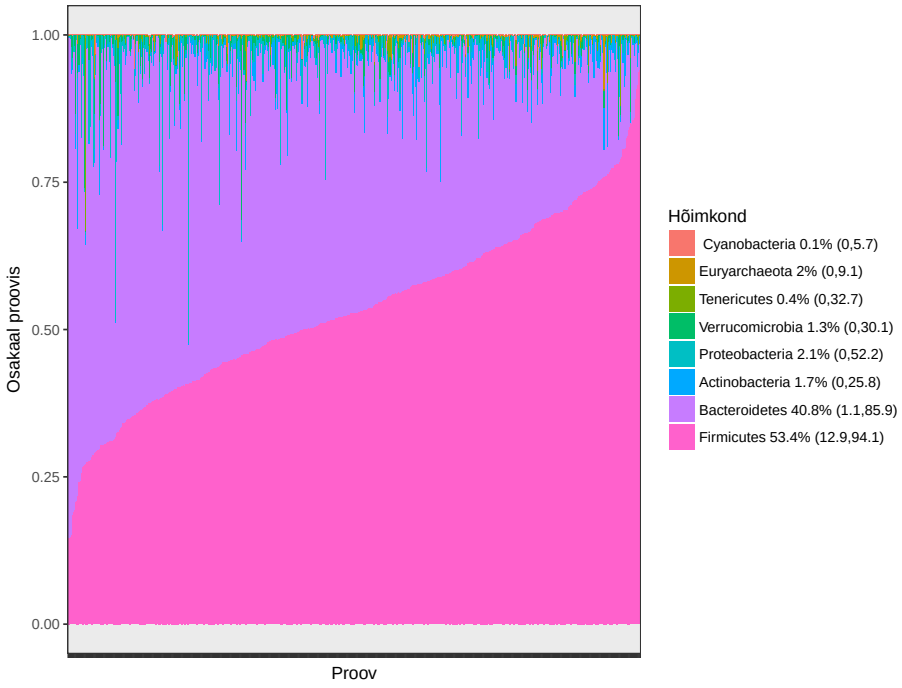
Olulisematest seostest nähtub, et alaniini, leutsiini ning türosiini konsentratsioonid plasmas on tervetel inimestel madalamad kui prediabeetikutel. Veelgi enam, olulisuse nivool 0.05 statistiliselt olulistest erinevustest enamik aminohapetest on tervetel inimestel madalama konsentratsiooniga. Ainult glutamiini konsentratsioon on tervete indiviidide plasmas tõestatavalt kõrgem. On näidatud, et hargnenud ahelaga aminohapete (valiini, leutsiini ja isoleutsiini) tase on kõrgem insuliini resistentsuse ning nende aminohapete kõrgenenud taset on seostatud soolestiku mikrobioomiga [19].

Tabel 3: Aminohapete konsentratsioonid prediabeetikutel ning tervetel inimestel

Aminohape	Terve		Prediabeet		P-väärtus
	Keskmine	Std	Keskmine	Std	
Alaniin	0.356	0.0037	0.379	0.0028	<.0001
Glutamiin	0.577	0.0040	0.564	0.0030	0.0214
Glütsiin	0.181	0.0039	0.189	0.0030	0.2402
Histidiin	0.060	0.0006	0.059	0.0004	0.1678
Isoleutsiin	0.047	0.0008	0.052	0.0007	0.0004
Leutsiin	0.074	0.0010	0.080	0.0008	0.0001
Valiin	0.162	0.0022	0.170	0.0018	0.0133
Fenüülalaniin	0.064	0.0005	0.067	0.0005	0.0014
Türosiin	0.049	0.0007	0.053	0.0005	0.0001

P-väärtus on Wilcoxoni astaksumma testi olulisustõenäosus testimaks erinevusi tervete ning prediabeetikute vahel

5.2.2 Mikrobioomi kirjeldav analüüs



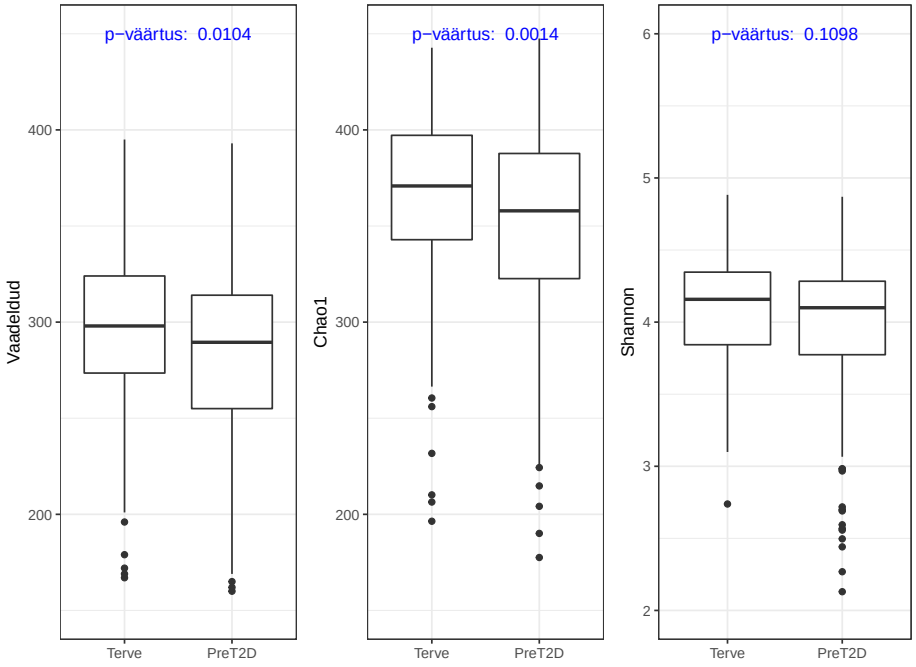
Joonis 1: Erinevast hõimkonnast OTUde osakaalud 531 METSIM proovis. Legendis on toodud iga hõimkonna jaoks keskmine proportsioon proovis ning minimaalne ja maksimaalne vaadeldud proportsioon

Joonis 1 kirjeldab indiviidi proovis leidunud OTUde lugemite jagunemist hõim- konniti. Arvukaimad hõimkonnad on *Firmicutes* ning *Bacteroidetes*. Teistest hõimkondadest OTUde lugemite keskmised proportsioonid valimites on väiksed. Mõningate proovide jaoks on ka teistest hõimkondadest prorportsioonid

siiski märgatavad: näiteks on *Proteobacteria* hõimkonnast mikroobide lugemite maksimaalne osakaal proovist on olnud 52.2%.

Jooniselt 1 nähtub, et indiviidide mikrobioomi kooslus on väga erinev: *Firmicutes* hõimkonda klassifitseeritud lugemite proportsioonid valimites kõiguvad 12.9% kuni 94%-ni. Vähenenud *Firmicutes* hõimkonnast bakterite osakaaluga proovides on suurenenud eelkõige *Bacteroidetes* hõimkonnast bakterite osakaal. Sedavõrd suur kõikumine viitab tugevalt asjaolule, et mikrobioom on indiviididel väga erineva kooslusega ning peronaliseeritud.

5.3 α-mitmekesisuse analüüs



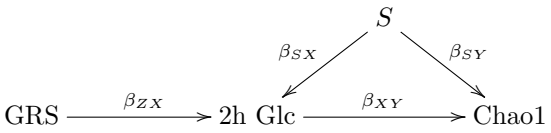
Joonis 2: α- mitmekesisuse näitajad prediabeetikute ning tervete indiviidide jaoks: vaadeldud unikaalsete OTUde arv hõrendatud valimis, *Chao1* indeks populatsiooni liigirikkuse alampiiri hindamiseks ning *Shannoni* entroopia hindamaks nii liigirikkust kui ka liikide jaotuvust valimis

Mikrobioomi mitmekesisuse vähenemist on seostatud mitmete haiguste ning halvenenud tervisenäitajatega, mistõttu pakub huvi, kas ka diabeedieelse seisundi puhul on näha sarnast trendi. Joonisel 2 on toodud prediabeetikute ning tervete inidviidide jaoks karpdiagrammid kolme α- mitmekesisust hindava mõõdiku puhul. Kasutatakse unikaalseid liike hõrendatud valimis, liigirikkuse hindamiseks

Chao1 indeksit ning ühtluse ja liigirikkuse koodnäitajana *Shannoni* entroopiat. Jooniste päises on toodud Wilcoxoni testi olulisustõenäosused võrdlemaks indeksite väärtusi tervete inimeste ning prediabeetikute vahel.

Jooniselt 2 on näha, et kõigi kolme kasutatud α- mitmekesisust hindava näitaja puhul on tervete indiviidide soolestiku mikrobioomi mitmekesisus suurem kui prediabeetikutel. Seos osutub mõlema liigirikkuse näitaja, vaadeldud liikide ning *Chao1* indeksi jaoks olulisuse nivool $\alpha = 0.05$ ka statistiliselt oluliseks. *Shannoni* indeksi puhul ei saa tõestada erinevust gruppide α- mitmekesisuse vahel.

Nähtud statistiliselt oluline seos prediabeedi ning liigirikkuse näitajate vahel tekitab küsimuse, kas vähenenud liigirikkus mõjutab prediabeedi seisundit või vastupidi. Antud küsimuse uurimiseks kasutatud Mendeli randomiseerimise skeem:



Siin *GRS* on METSIM kohordi indiviididele arvutatud geneetiline riskiskoor teist tüüpi diabeedi esinemisele. Tunnus *2h Glc* on prediabeedi indikeerimiseks kasutatud pidev tunnus, mis on glükoosi tase veres OGTT testi puhul 2 tundi pärast glükoosi manustamist ning *Chao1* on liigirikkuse näitaja. Geneetiline riskiskoor arvutatakse vaid teist tüüpi diabeediga seotud geenimarkerite pealt, mistõttu ei ole alust arvata, et leidub otsene seos geneetilise riskiskoori ning α-mitmekesisuse vahel. Tähtsamad teist tüüpi diabeedi esinemist mõjutavad, antud skeemi järgi segavad faktorid on kehamassiindeks ning indiviidi vanus. Ka teist tüüpi diabeedile hinnatud geneetilise riskiskoori ning kehamassiindeks ja vanuse vahel pole põhjust arvata olevat seost, mistõttu on geneetiline riskiskoor kasutatav instrumenttunnusena Mendeli randomiseerimise kontekstis.

Esmalt testiti GRSi mõju tunnusele *2h Glc*. Geneetilise riskiskoori hinnanguks tuli 0.26 ning olulisustõenäosuseks 0.038. Seejärel testiti geneetilise riskiskoori mõju tunnusele *Chao1*. Hinnanguks saadi -0.63 ning olulisustõenäosuseks 0.872. Hinnati kaheastmeline regressioonmudel kasutades paketi *sem* funktsiooni *tsls*, mis väljastab hinnangu ning parameetrile β_{XY}. Parameetri hinnanguks tuli -2.43 ning olulisustõenäosus 0.871, mistõttu andmete põhjal ei saa tõestada,

et prediabeedile viitav kõrgem OGTT testi 2h glükoosi tase omaks põhjuslikku seost soolestiku mikroobilisele liigirikkusele.

Kuigi prediabeedi põhjuslikku mõju liigirikkusele ei õnnestunud näidata, jätab nähtud statistiline seos prediabeedi ning liigirikkuse vahel nii võimaluse, et see mõju on siiski olemas, kuid meie andmestikus puudub piisav võimsus selle tõestamiseks, kui ka võimaluse, et põhjuslik mõju on teistpidine. Kuigi meil puudub tunnus, mis oleks bioloogiliselt põhjendatud valik liigirikkuse instrumenttunnuseks, on võimalik arvesse võtta teadaolevaid segavaid faktoreid S. Et teist tüüpi diabeedi riskifaktorid on suhteliselt hästi teada ning kasutatavas andmestikes mõõdetud, hinnati lineaarne regressioonimudel OGTT testi kahe tunni järgsele glükoositasemele kasutades kovariantidena kehamassiindeksit, indiviidi vanust ning *Chao1* indeksi hinnangut keskkonna liigirikkusele. Saadud tulemused on toodud tabelis 4.

Tabel 4: Lineaarne regressioonmudel prognoosimaks OGTT testi järgset glükoositaset veres teadaolevate segavate tunnuste ning mitmekesisuse indeksi *Chao1* abil

Kordaja	$\hat{\beta}$	$SE(\hat{\beta})$	P(> |t|)
Kehamassiindeks	0.112	0.0195	<.0001
Vanus	0.036	0.0128	0.0054
Chao1	-0.004	0.0014	0.0086

Koostatud mudelis tuleb $Chao1$ näitaja olulisuse nivool 0.05 statistiliselt oluliseks. Seega, modelleerides prediabeeti indikeerivat tunnust *2h Glc* ning võttes arvesse teadaolevaid teist tüüpi diabeeti prognoosivad tunnused, kannab $Chao1$ ikkagi lisainfot fenotüübi prognoosimisel. Kuigi Mendeli randomiseerimist näitamaks mikrobioomi liigirikkuse põhjuslikku mõju prediabeedile ei saa instrumenttunnuse puudumise tõttu formaalselt kasutada, siis andmete põhjal on pigem põhjust arvata, et põhjuslik mõju on just selline.

5.4 β-mitmekesisuse analüüs

Beeta-mitmekesisust kasutati hindamaks, kas keskmised mikrobioomi koosseisud on tervetel inimestel ning prediabeetikutel erinevad. β-mitmekesisuse näitajana kasutati Bray-Curtise eripära. Hinnati mitteparameetrilise dispersioonanalüüsi mudel kasutades funktsiooni *adonis*. Olulisuse nivool 0.05 ei õnnestu tõestada, et

prediabeetikute keskmine mikrobioomi kompositsioon oleks erinev tervete inimeste omast (testi olulisustõenäosus 0.068). Vastav dispersioonanalüüsi tabel on toodud lisas 10.

5.5 Prediabeetikute ja kontrollide võrdlus üksikute OTUde osas

Uurimaks üksikute OTUde erinevat esinemissagedust tervete inimeste ning prediabeetikute vahel kasutati meetodit paketist *ALDEx2*. Meetodi idee järgi genereeritakse Dirichlet jaotusest tõenäosuste järeljaotus, saadud tulemustele rakendatakse CLR-transformatsiooni ning transformeeritud andmetele rakendatakse t-testi ning Wilcoxoni astakmärgitesti. Mõlema testi jaoks saadakse p-väärtuste jaotused, mille puhul raporteeritakse jaotuste mediaanväärtused. Samuti rakendatakse Benjamin-Hochbergi meetodit p-väärtuste kohandamiseks vältimaks kõrget valepositiivsete tulemuste määra. Täpsem meetodi kirjeldus on toodud peatükis 4.4.2.

Tabel 5 kirjeldab meetodi poolt väljastatavaid tähtsaimaid suuruseid. Tabelis on toodud OTUd, mille jaoks Wilcoxoni testi kohandamata olulisustõenäosus $P_{Wilcoxon}$ on alla 0.05. Tabelist nähtub, et Benjamin-Hochbergi meetodi järgi kohandatud p-väärtuste P_{BH} järgi pole ükski seostest olulisuse nivool 0.05 statistiliselt oluline. Samuti on näha iga OTU jaoks, et gruppidevaheline erinevus on palju väiksem kui grupisisene erinevus. Seetõttu on väike nii efekti suurus kui gruppidevahelise erinevuse ning grupisisese varieeruvuse jagatis.

Tabelis 5 toodud tulemusi kinnitab ka joonis 3. Punasega on toodud Wilcoxoni testi järgi olulisuse nivool 0.05 olulised OTUd. Kohandatud olulisustõenäosuste järgi statistilist erinevust ei õnnestunud ühegi OTU puhul tõestada. Märkimisväärseim on gruppidesisest ning gruppidevahelist erinevust kirjeldav joonis A, mille kohaselt grupisisene erinevus on kordades suurem kui gruppidevaheline. See viitab asjaolule, et OTUde esinemissagedused indiviidide mikrobioomis erinevad eelkõige indiviiditi ning ei ole olulisel määral prediabeedi poolt mõjutatud.

Tabelis 6 on toodud OTUd Wilcoxoni testi järgi kohandamata olulisustõenäosustega alla 0.05 ning nende taksonoomiline liigitus. Leitud markerid kuuluvad hõimkondadesse *Bacteroidetes* ning *Firmicutes*. *Bacteroidetes* hõimkonda kuuluvatest mikroobidest enamus kuuluvad *Bacteroides* perekonda. *Firmicutes*

Tabel 5: Meetodi ALDEx2 väljund

	Δ_A	Δ_W	Δ_R	$P_{wilcoxon}$	P_{BH}
OTU.589071	0.41	2.88	0.12	0.0320	0.4676
OTU.3327894	0.45	3.17	0.12	0.0246	0.4160
OTU.339013	0.51	4.31	0.11	0.0249	0.4334
OTU.181953	0.38	3.02	0.10	0.0439	0.4944
OTU.344525	0.47	4.08	0.10	0.0403	0.4717
OTU.364179	-0.80	5.82	-0.13	0.0436	0.5077
OTU.577294	0.79	5.19	0.14	0.0137	0.3396
OTU.361727	-0.57	3.92	-0.12	0.0253	0.4095
OTU.552380	-0.75	5.29	-0.12	0.0165	0.3930
OTU.1110312	-1.04	6.67	-0.15	0.0117	0.3646
OTU.366237	0.64	3.61	0.15	0.0133	0.3344
OTU.368950	0.47	2.57	0.15	0.0047	0.3168
OTU.3797933	0.74	4.60	0.14	0.0140	0.3102
OTU.368698	0.64	4.49	0.12	0.0280	0.4397
OTU.514996	0.52	3.61	0.12	0.0117	0.3651
OTU.369227	0.30	2.69	0.08	0.0442	0.5107
OTU.584463	-0.93	5.74	-0.14	0.0255	0.4241
OTU.180462	0.63	4.13	0.13	0.0260	0.3919
OTU.1062061	-0.54	4.17	-0.11	0.0239	0.4393
OTU.165935	-0.83	5.13	-0.15	0.0077	0.3306
OTU.564941	-0.89	5.58	-0.15	0.0114	0.3646
OTU.581079	0.86	5.27	0.15	0.0099	0.3415

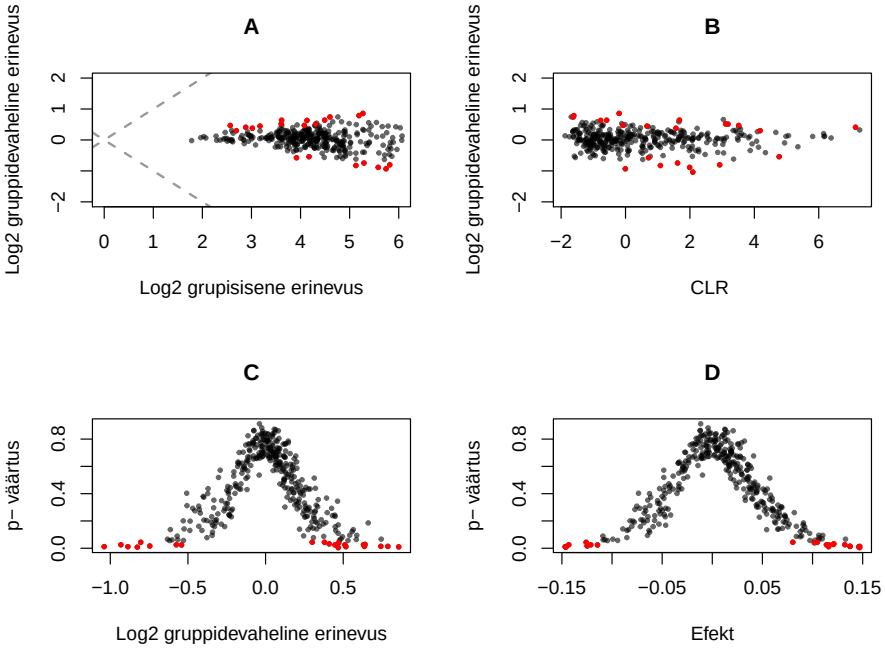
Δ_A kirjeldab gruppidevahelist varieeruvust, Δ_W grupisisest varieeruvust ning Δ_R efekti suurust. $P_{wilcoxon}$ näitab olulisustõenäosuste jaotuse mediaani ning P_{BH} Benjamin- Hochbergi meetodil kohandatud olulisustõenäosuste jaotuse mediaani

hõimkonda kuuluvatest bakteritest on kõik OTUd *Clostridiales* seltsi sugukondade esindajad. Seega on tegu taksonoomiliselt sarnaste OTU-dega.

6 Logistilise regressiooni mudelid teist tüüpi diabeedi esinemisele

6.1 Klassikaline logistiline regressioon

Logistilisse mudelisse kaasati kovariantidena vanus, geneetiline riskiskoor ning kehamassiindeks. Prediabeediga seotud OTUde leidmiseks kasutati ettepoolevalikuga leitud logistilise regressiooni mudelit. Esiteks kaasati indiviidi vanuse, geneetilise riskiskoori ning kehamassiindeksi järgi kaalutud mudelisse väikseima p-väärtusega OTU (OTU_1). Seejärel lisati mudelisse OTU, mis viis väikseima p-väärtuseni mudelis, mis oli kohandatud vanuse, geneetilise riskiskoori, keha-



Joonis 3: ALDEx2 meetodit kirjeldavad joonised. Joonisel punasega on toodud OTUd, mis Wilcoxoni testi järgi olulisuse nivool 0.05 grupiti erinesid. Alamjoonis A kirjeldab gruppidevahelist ja grupisisest erinevust, joonis B näitab CLR transformeeritud OTU esinemissageduste mediaanväärtuse ja gruppidevahelise erinevuse seost, joonise C näitab gruppidevaheliste erinevuse seost Wilcoxoni testi olulisustõenäosustega, joonis D kirjeldab efekti seost Wilcoxoni testi olulisustõenäostega.

massiindeksi ning OTU_1-ga. Protsessi jätkati kuni ükski OTU ei olnud enam olulisuse nivool 0.05 oluline. Kasutati OTUde CLR-transformeeritud väärtusi.

Tabelis 7 on toodud kirjeldatud algoritmi alusel mudelisse valitud OTUd ning nende taksonoomiline liigitus. Leitud markerid kuuluvad sarnaselt meetodiga *ALDEx2* raporteerituile hõimkondadesse *Bacteroidetes* ning *Firmicutes*.

6.2 LASSO regressioon

Lisaks logistilise regressiooni mudelile sobitati andmetele regulariseeritud logistilise regressiooni mudel. LASSO regressiooni rakendamiseks kasutati tarkvara R paketti *glmnet*. Mudelisse kaasati kovariantidena lisaks indiviidi vanusele, geneetilisele riskiskoorile ja kehamassiindeksile ka aminohapete konsentratsioonid ning teised fenotüüpi kirjeldavad tunnused, mis ei ole seotud prediabeedi diagnoosimisega. LASSO regressioon puhul loobutakse nihketuse nõudest limiteerides parameetrite kordajate absoluutväärtuse summat parameetri λ abil. Jooni-

Tabel 6: Meetodiga ALDEx2 leitud erinevalt ekspresseeruvad OTUd taksonoomiaklassiti

	Taksonoomia
OTU.589071	p_Bacteroidetes/g_Bacteroides
OTU.3327894	p_Bacteroidetes/g_Bacteroides
OTU.339013	p_Bacteroidetes/g_Bacteroides
OTU.181953	p_Bacteroidetes/g_Bacteroides
OTU.344525	p_Bacteroidetes/g_Bacteroides
OTU.364179	p_Bacteroidetes/g_Bacteroides
OTU.577294	p_Bacteroidetes/g_Parabact.
OTU.361727	p_Firmicutes/o_Clostridiales
OTU.552380	p_Firmicutes/o_Clostridiales
OTU.1110312	p_Firmicutes/o_Clostridiales
OTU.366237	p_Firmicutes/g_Blautia
OTU.368950	p_Firmicutes/g_Blautia
OTU.3797933	p_Firmicutes/f_Lachnospiraceae
OTU.368698	p_Firmicutes/f_Lachnospiraceae
OTU.514996	p_Firmicutes/f_Lachnospiraceae
OTU.369227	p_Firmicutes/f_Lachnospiraceae
OTU.584463	p_Firmicutes/g_Lachnobacterium
OTU.180462	p_Firmicutes/f_Ruminococcaceae
OTU.1062061	p_Firmicutes/f_Ruminococcaceae
OTU.165935	p_Firmicutes/f_Ruminococcaceae
OTU.564941	p_Firmicutes/f_Ruminococcaceae
OTU.581079	p_Firmicutes/g_Oscillospira

Taksonoomiline liigitus, p - hõimkond (*phylum*), o - selts (ingl *order*), f - sugukond (ingl *family*), g - perekond (ingl *genus*)

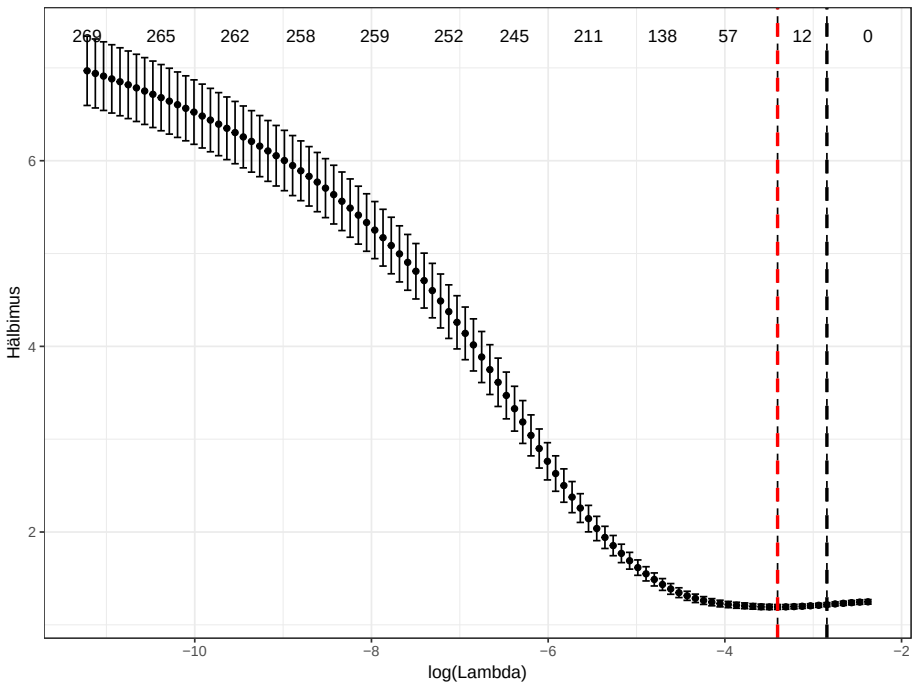
sel 4 on toodud mudeli hälbimuse muutumine varieerides karistusparameetri λ väärtust. Saadud hälbimus on arvutatud ristvalideerimise teel. Joonisel 4 punase vertikaalse joonega on toodud λ väärtus, mille korral mudeli hälbimus on väikseim (edaspidi λ_{min}). Kasutades mudelis parameetrina optimaalset, minimaalset hälbimust tagavat λ-t λ_{min}, saame täpseima mudeli. Joonisel 4 toodud must vertikaalne joon kirjeldab λ väärtust, mille puhul mudeli hälbimus on optimaalse mudeli hälbimusest ühe standardhälve piirides, kuid mis annab kõige lihtsama ehk vähima argumentide arvuga mudeli (edaspidi λ_{1se}). Selliselt defineeritud λ kasutamine lähtub põhimõttest, mille järgi otsime hea täpsusega lihtsaimat mudelit. Optimaalne λ_{min} on seejuures väiksem ning seega on saadud mudeli koefitsiendid sarnasemad tavalise logistilise regressiooni omale. Seda asjaolu kirjeldavad joonised 5 ja 6.

Joonisel 5 on kirjeldatud iga parameetri koefitsiendi muutumist lähtuvalt λ muutumisest. Mida suurem on λ, seda rohkem piiratakse parameetrite absoluutväärtuste summat ning koefitsientide absoluutväärtused kahanevad λ kasvades. LASSO regressiooni puhul karistusparameetri kasvades muutuvad teatud

Tabel 7: Logistilise regressiooniga leitud prediabeetiga seotud OTUd

	$\hat{\beta}$	P(> |z|)	taksonoomia
OTU.193591	0.17	0.0065	p_Bacteroidetes/g_Bacteroides
OTU.198449	-0.15	0.0181	p_Bacteroidetes/g_Bacteroides
OTU.187623	-0.16	0.0017	p_Bacteroidetes/g_Bacteroides
OTU.577294	0.23	0.0010	p_Bacteroidetes/g_Parabact.
OTU.578016	0.09	0.0313	p_Bacteroidetes/g_Parabact.
OTU.361727	-0.13	0.0375	p_Firmicutes/o_Clostridiales
OTU.368950	0.29	0.0017	p_Firmicutes/g_Blautia
OTU.3797933	0.33	0.0005	p_Firmicutes/f_Lachnospiraceae
OTU.360329	0.14	0.0236	p_Firmicutes/f_Lachnospiraceae
OTU.526468	-0.24	0.0098	p_Firmicutes/f_Lachnospiraceae
OTU.584463	-0.10	0.0460	p_Firmicutes/g_Lachnobacterium
OTU.305016	0.15	0.0088	p_Firmicutes/g_Lachnospira
OTU.540862	-0.23	0.0095	p_Firmicutes/f_Ruminococcaceae
OTU.581079	0.27	0.0002	p_Firmicutes/g_Oscillospira
GRS	0.60	0.0026	
KMI	0.12	0.0002	
Vanus	-0.00	0.8873	

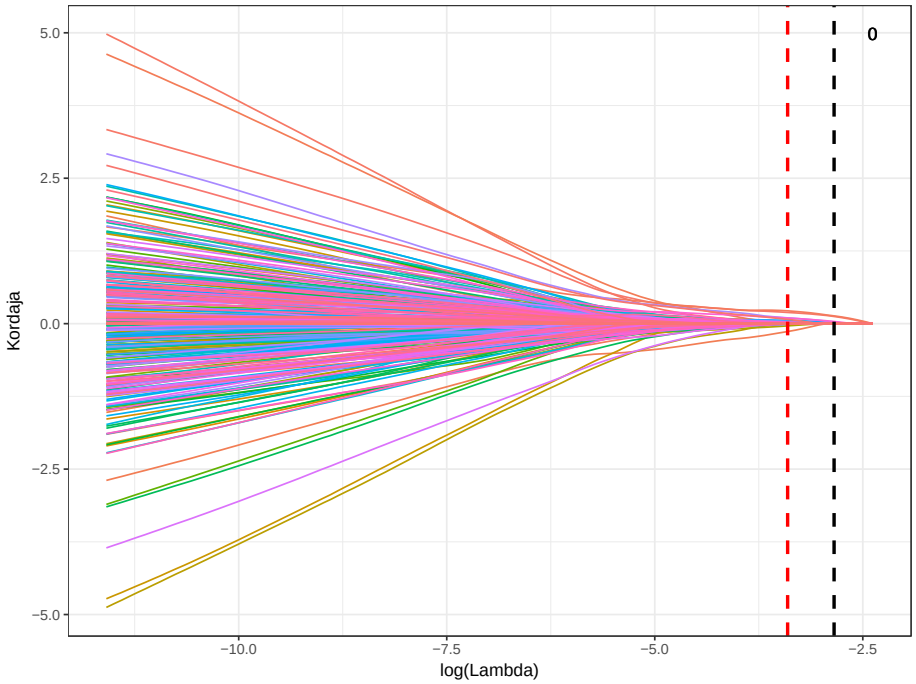
Taksonoomiline liigitus, p - hõimkond (*phylum*), o - selts (ingl *order*), f - sugukond (ingl *family*), g - perekond (ingl *genus*)



Joonis 4: Regulariseeritud mudeli hälbimuse muutumine karistusparameetri λ kasvades. Saadud hälbimus on arvutatud ristvalideerimise teel.

parameetrite koefitsiendid täpselt nullideks.

Joonisel 6 on toodud koefitsientide väärtused sõltuvalt λ-st vahemikus $(e^{-4}, e^{-2}) =$

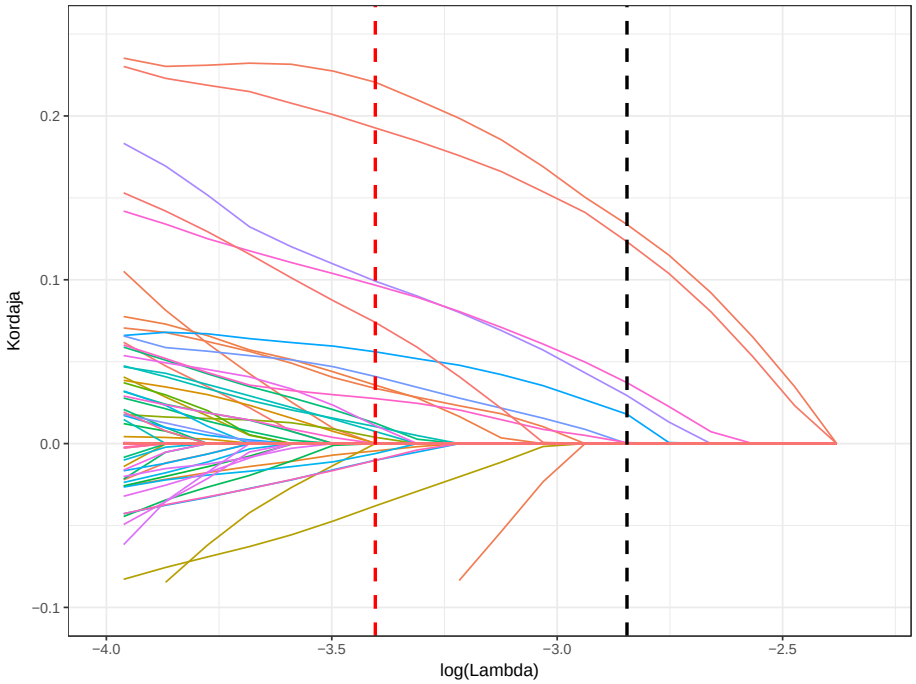


Joonis 5: Regulariseeritud mudeli parameetrite väärtuste muutumine karistusparameetri λ kasvades

$(0.018, 0.135)$. Jooniselt 6 nähtub, et kõik koefitsiendid hinnatakse lõpuks nullideks lähtuvalt karistusparameetri suurusest. Punane vertikaalne joon näitab λ_{min} väärtust ning must joon λ_{1se} väärtust. Mõlema λ väärtuse puhul leidub parameetreid, mille kordajad hinnatakse nullist erinevalt. Samuti, karistusparameetrit λ_{1se} kasutades on nullist erinevaid kordajaid vähem kui karistusparameetri λ_{min} puhul. Seega viib λ_{1se} kasutamine lihtsama mudelini. Samas, joonise 5 alusel võib öelda, et lihtsamat mudelit eelistades ei ole kaotus mudeli täpsuses kuigi suur.

Lõpliku mudeli nullist erinevate kordajatega tulemused on toodud tabelis 8. Tabeli veerg *1se* märgib parameetreid, mis jäid mudelisse mittenulliliste kordajatega alles ka parameetri λ_{1se} korral.

Alaniini, fenüülalaniini ja türosiini kõrgemaid konsentratsioonide on täheldatud metaboolsete haiguste nagu teist tüüpi diabeet puhul [17]. Sarnast efekti on näha ka prediabeetikute puhul, kellel on võrreldes tervete inimestega suurenenud Alaniini ning türosiini konsentratsioon veres. Olulisim seos on sealhulgas suurenenud alaniini tase plasmas.



Joonis 6: Regulariseeritud mudeli parameetrite väärtuste muutumine karistusparameetri λ kasvades

7 OTUde koosesinemine

Mikrobioomi puhul huvitab meid ka, millised OTUd esinevad soolestikus üheaegselt. See võimaldab meil täpsemalt edasistes uuringuetappides kindlaks teha mikrobioomi potentsiaalse kombineeritud efekti olemasolu. Kompositsionaalsete andmete puhul peab kahe OTU proportsioonide suhte dispersioon olema nulli ligidane, et neid lugeda korreleerituks. Selleks kasutatakse peatükis 4.5 defineeritud Φ - statistikut.

Joonisel 7 on välja toodud seotud OTUd ($\Phi < 0.3$), mis on värvitud hõimkondliku kuuluvuse järgi ning nimetatud sugukonna nime järgi. Kollastes ning punastes toonides on hõimkonda *Firmicutes* kuuluvad OTUd ning rohelistes toonides hõimkonda *Bacteroidetes* kuuluvad OTUd.

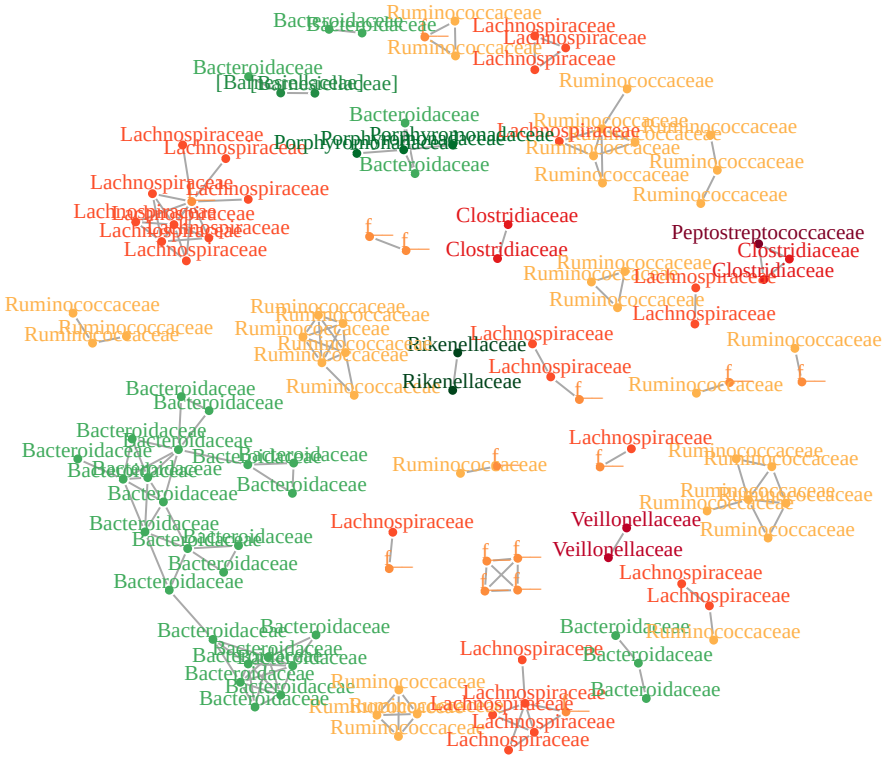
Teiste hõimkondade puhul polnud OTUsid, mis oleks üksteisega seotud $\Phi < 0.3$ korral. Jooniselt nähtub, et seotud OTUd klasterduvad eranditult hõimkonnasiseselt, ei leidu ühtegi paari OTUsid, mis oleks seotud ning oleks taksonoomiliselt erinevatest hõimkondadest. Sama kehtib üksikute eranditega ka hõimkonnasiseselt,

Tabel 8: LASSO regressiooniga leitud erinevalt ekspresseeruvad OTUd

	efekt	taksonoomia	1se
OTU.339013	0.01	p_Bacteroidetes/g_Bacteroides	
OTU.344525	0.01	p_Bacteroidetes/g_Bacteroides	
OTU.193591	0.01	p_Bacteroidetes/g_Bacteroides	
OTU.577294	0.10	p_Bacteroidetes/g_Parabacteroides	X
OTU.177792	0.00	p_Firmicutes/o_Clostridiales	
OTU.366237	0.06	p_Firmicutes/g_Blautia	X
OTU.368950	0.04	p_Firmicutes/g_Blautia	X
OTU.3797933	0.10	p_Firmicutes/f_Lachnospiraceae	X
OTU.305016	0.01	p_Firmicutes/g_Lachnospira	
OTU.551902	0.01	p_Firmicutes/f_Ruminococcaceae	
OTU.581079	0.03	p_Firmicutes/g_Oscillospira	X
Alanine	4.31		X
Tyrosine	3.88		
Leucine	2.34		
GRS	0.13		
KMI	0.05		X

Taksonoomiline liigitus, p - hõimkond (*phylum*), o - selts (ingl *order*), f - sugukond (ingl *family*), g - perekond (ingl *genus*). Veerg 1se näitab parameetreid, mis jäid mudelisse ka karistusparameetri λ_{1se} puhul

kuid erinevaid sugukondi vaadates. Eranditena on seotud klastrites mõned *Ruminoccaceae* sugukonnast OTUd *Lachnospiraceae* sugukonnast OTUdega, *Clostridiaceae* sugukonnast OTUsid *Peptostreptococcaceae* sugukonnast OTUdega ning *Barnesiellaceae* sugukonnast OTU *Bacteriodaceae* sugukonnast OTU-ga. Vaadates gruppidesiseselt taksonoomilist liigitust, nähtub, et seotud OTUd on sugukondadest, mis kuuluvad taksonoomiliselt samasse ülemklassi - seltsi. Samasse seltsi kuuluvad OTUd on fülogeneetiliselt sarnased. Seega METSIM kohordi andmetelt leitud mikrobioomis koosesinevad OTUd on ka fülogeneetiliselt lähedalt seotud, ühtegi fülogeneetiliselt kaugelt seotus mikroobipaari ei leitud.



Joonis 7: Keskkonnas koosesinevad OTUd ($\Phi < 0.3$) sugukonniti. Kollastes ning punastes toonides on hõimkonda *Firmicutes* kuuluvad OTUd ning rohelistes toonides hõimkonda *Bacteroidetes* kuuluvad OTUd. Joonise tegemiseks on kasutatud paketti *igraph*

8 Diskussioon

Mikrobioomi mitmekesisus on oluline indikaator, millega on leitud seoseid haiguseseisunditega. Käesolevas töös nähtus sama efekt: teist tüüpi diabeedi eelses seisundis indiviididel tuvastati madalam liigirikkus, kui tervetel inimestel. Tähelepanuväärne on aga tulemus põhjuslikkuse suuna kohta. Kuigi formaalselt

tulemust tõestada ei saa, võib andmetelt pigem öelda, et vähenenud soolestiku mikrobioomi liigirikkusel on põhjuslik mõju prediabeedile. See tulemus väärib igal juhul edasist uurimist. Kui selline põhjuslikkuse suund leiaks kinnitust, oleks tegu otsese indikaatoriga, mis annaks personaalse meditsiini perspektiivis juurde teguri hindamaks riski teist tüüpi diabeedile ning võimaldaks varasemat ennetustööd.

Mitmekesisuse analüüsimisel on pikka aega kasutusel olnud hõrendamistehnikad (ingl *rarefaction*), kuid aina enam argumenteeritakse teemal, miks hõrendamistehnikate kasutamine ei ole põhjendatud. Lihtsaim argument meetodite mitte kasutamise vastu on asjaolu, et teatud info läheb kaduma, sest iga proovi järjestustest valitakse välja väiksem arv järjestusi, kui proovis on [15, 26]. Eelkõige tekitab hõrendamine võimalikke nihkeid keskkonna liigirikkuse ning β-mitmekesisuse uurimisel. Hõrendamisele alternatiivide otsimine käib, kuid ühtset lähenemist veel ei ole. Edasises töös tasuks uurida veel hõrendamise mõju, näiteks sensitiivsusanalüüsidega. Samuti tasub uurida uuemaid pakutud alternatiive hõrendamisele.

Kahe populatsiooni β- mitmekesisuse võrlemiseks kasutati Bray- Curtise eripäral põhinevat lähenemist. Olulisuse nivool 0.05 ei olnud saadud erinevus statistiliselt oluline, kuid olulisustõenäosus 0.069 võib anda märku, et mingi seos siiski on. Prediabeedi puhul ei ole organismis toimunud muudatused niivõrd suured kui teist tüüpi diabeeti põdevatel indiviididel, mistõttu bioloogide jaoks väärib nähtud tulemus edasist uurimist. Üks võimalus edasisteks uuringuteks on fülogeense informatsiooni kasutamine. Kasutatud Bray-Curtise eripära võtab arvesse ainult lugemite hõrendatud lugemite arve. Peatükist 7 selgus, et tihtipeale on taksonoomia järgi lähedased liigid ka keskkonnas koos eksisteerivad. Seda asjaolu võetakse β- mitmekesisuse võrdlemisel arvesse UniFrac mitmekesisuse näitajate puhul [13]. UniFrac meetodid mõõdavad OTUde fülogeneetilist kaugust üksteisest ehk võetakse arvesse taksonoomilist sarnasust: taksonoomiliselt sarnaste liikidega mikrobioomid loetakse sarnasemaks, erinevas taksonoomia harus olevad bakterid saavad kauguse arvutamisel erineva kaalu. Sellisel juhul arvutades saadaks suurused, mis kirjeldavad mikrobioomi mitte ainult kvantitatiivselt, vaid ka kvalitatiivselt ning indiviidide mikrobioomi koosluste võrdlemine on bioloogiliselt põhjendatum.

Suur osa mikrobioomi analüüsivatest uurimistöödest käsitlevad OTUde lu-

gemite arve loendusandmetena, kuid praeguste sekveneerimistehnoloogiate kasutamisel pole see põhjendatud. Mikrobioomi andmete analüüsimine nõuab metoodikat, mis võtaks andmete kompositsionaalset olemust arvesse. Peatükis 4.4.2 kirjeldati ühte võimalikku lähenemist kompositsionaalsete andmete analüüsimiseks. Paketis *ALDEx2* implementeeritud meetod on ka mikrobioomi uuringute puhul pigem uus, kuid teadlikkus andmete kompositsionaalsusest on levimas ning aina enam võetakse seda ka arvesse. Siiski on kompositsionaalsete andmete analüüsimise metoodika veel arengujärgus.

Kuigi ühegi kasutatud meetodi puhul ei saa tõestada, et OTUde esinemissagedused proovides oleks grupiti erinevad, leidus neli OTUt, mille leidsid üles nii paketis *ALDEx2* rakendatud meetod, logistilise regressiooni mudel kui ka regulariseeritud logistilise regressiooni mudel. Logistilise regressiooni mudelis kasutati kovariantidena vanust, kehamassiindeksit ning geneetilist riskiskoori, regulariseeritud logistilise regressiooni mudelis kõiki relevantseid fenotüübiandmeid ning aminohapete konsentratsioone. Tabelis 9 on toodud nimetatud neli OTUt ja nende taksonoomiline liigitus. Kõigi nelja OTU proportsioonid valimis olid prediabeetikutel kõrgemad.

Tabel 9: ALDEx2 meetodil, logistilise regressiooni mudelis kui ka regulariseeritud logistilise regressiooni mudelis ühised OTUd

	taksonoomia
OTU.577294	p_Bacteroidetes/g_Parabacteroides
OTU.368950	p_Firmicutes/g_Blautia
OTU.3797933	p_Firmicutes/f_Lachnospiraceae
OTU.581079	p_Firmicutes/g_Oscillospira

Taksonoomiline liigitus, p - hõimkond (*phylum*), f - sugukond (ingl *family*), g - perekond (ingl *genus*)

Kuigi erinevust ei õnnestunud OTUde puhul tõestada, on võimalik, et efekt on siiski olemas. Võimalik, et suure OTUde arvu testimise jaoks pole valim piisavalt suur prediabeetikute ning tervete inimeste võrdlemiseks. Võimalik, et samad OTUd annavad tugevama seose võrreldes terveid indiviide indiviididega, kellel on teist tüüpi diabeet juba välja arenenud.

Soolestiku mikrobioom on keeruline süsteem, kus bakterid koeksisteerivad ning mõjutavad üksteist. Seetõttu on loomulikum uurida OTUsid kui kompositsiooni korraga. Kasutatud regressioonimeetodite, nii logistilise regressiooni kui ka regulariseeritud logistilise regressiooni, eeliseks *ALDEx2* metoodika suh-

tes on asjaolu, et *ALDEx2* testib OTUde erinevusi gruppides ühe OTU kaupa, regresioonimeetodi annavad võimaluse arvestamaks OTUsid kompositsioonina.

Andmestikus, kus on suhteliselt väike arv vaatlusi võrrelduna mudelisse lisatud parameetrite arvuga, on regulariseeritud regressiooni kasutamine huvitav võimalus hindamaks, millised argumendid on uuritava tunnuse modelleerimiseks olulised. Käesolevas töös kasutatud regulariseeritud logistiline regressioon andis OTUde osas sarnaseid tulemusi, kui *ALDEx2* metoodika. Lisaks võimaldas regulariseeritud logistilise regressiooni mudel arvestada kovariantidena aminohapetega. Kuna inimese organismis on protsessid seotud, siis võimaldab OTUde ning aminohapete üheaegne uurimine saada lisainfot metaboolsete protsesside kohta.

Huvitavamaid arenguid regulariseeritud regressioonide vallas on niinimetatud grupeeritud LASSO regressioon [8], mis võimaldab eelnevalt defineerida regressioonis kasutatavad grupid ning seejärel läheneda sarnaselt tavalisele LASSO regressioonile - valitakse välja uuritava tunnuse suhtes kõige rohkem infot kandvamad tunnuste grupid. Selline lähenemine võimaldaks arvesse võtta peatükis 7 nähtud koeksisteerivate bakterite gruppe ning potentsiaalselt parandada meie arusaama mikrobioomist, mistõttu on tegu huvipakkuva lähenemisega edasisteks uuringuteks.

9 Kokkuvõte

Käesoleva töö esmaseks eesmärgiks oli välja selgitada, milliseid statistilisi meetodeid tuleks kasutada kaasaegsete mikrobioomiandmete analüüsiks. Mikrobioomi andmete puhul tuleb arvestada andmete kogumise tehnoloogiaga, mis eeldab spetsiifilisi meetodeid. Mikrobioomi andmeid analüüsides tuleks arvestada andmete kompositsionaalset olemust nii andmete modelleerimisel kui ka seosenäitajate arvutamisel. Töös kirjeldati mõlema juhu jaoks meetodeid, mis arvestavad andmete kompositsionaalset olemust.

Neid meetodeid kasutati METSIM kohordi andmete analüüsil uurimaks, kas prediabeedi seisundis indiviidide soolestiku mikrobioomis on märgatavaid erinevusi võrrelduna tervete inimeste mikrobioomiga. Saadud tulemused kinnitasid mitme haigusega näidatud tendentsi: teist tüüpi diabeedi eelses seisundis indiviididel oli vähenenud soolestiku mikrobioomi liigirikkus. Põhjuslikkuse analüüs viitas aga asjaolule, et vähenenud soolestiku mikrobioomi liigirikkusel on põhjuslik mõju OGTT testi 2h glükoositasemele plasmas, misläbi ka prediabeedile. Saadud tulemusi tuleks kindlasti edasi uurida ning võimalusel ka formaalselt näidata.

Uurides üksikute OTUde avaldumist prediabeetikute ning tervete indiviidide gruppides, ei leitud ühtegi OTUt, mille puhul nähtud erinevus oleks statistiliselt tõestatav. Võimalik, et teist tüüpi diabeedile eelnevate seisundite puhul muutused mikrobioomis nii drastilised ei ole.

Kasutades kompositsionaalsete andmetele mõeldud seose tugevuse näitajat, leiti hulk OTUsid, mis esinesid soolestiku mikrobioomis samaaegselt. Tulemus näitab, et mikrobioom on keeruline süsteem ning üksikute OTUde sageduste analüüsimine on alles vahesamm mikrobioomi mõjudest aru saamisel. Edasistes uuringutes tuleks arvesse võtta mikrobioomi kui kompositsiooni. Ühe huvitava võimaluse selleks annab regulariseeritud regressiooni lähenemine.

Käesolevas töös anti ülevaade mikrobioomi iseloomustatavatest suurustest, mikrobioomi andmete analüüsimise meetoditest ning kompositsionaalsete andmete analüüsimise probleemidest ja võimalustest. Tulemused on näitavad, et indiviidide mikrobioomis esinevad teatud muudatused teist tüüpi diabeedi eelsete seisundite puhul ning mikrobioomi analüüsimine võib anda olulist lisainfot rakendamaks personaalset meditsiini.

Kasutatud kirjandus

[1] M. J. Anderson. A new method for non-parametric multivariate analysis of variance. 2006.

[2] A. Chao, R. K. Colwell, and N. J. Gotelli. Sufficient sampling for asymptotic minimum species richness estimators. 2009.

[3] Eesti diabeediliit. Mis on diabeet? 2015.

[4] A. D. Fernandes, J. M. Macklaim, G. Reid, G. B. Gloor, and T. G. Linn. Anova-like differential expression (aldex) analysis for mixed population rna-seq. 2013.

[5] A. D. Fernandes, J. N.Š. Reid, J. M. Macklaim, D. R. Edgell, and G. B. Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rrna gene sequencing and selective growth experiments by compositional data analysis. 2014.

[6] B. A. Frigyik, A. Kapila, and M. R. Gupta. Introduction to the dirichlet distribution and related processes. 2010.

[7] T. Hastie, R. Tibshirani, and J. Friedman. Regularization paths for generalized linear models via coordinate descent. 2010.

[8] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. 2017.

[9] M. A. Hernan and J. M. Robins. Causal inference. 2018.

[10] J. B. Hughes, J. J. Hellmann, T. G. Ricketts, and B. J. M. Bohannan. Counting the uncountable: Statistical approaches to estimating microbial diversity. 2001.

[11] L. Jost, A. Chao, and R. L. Chazdon. Biological diversity: Frontiers in measurement and assessment. 2014.

[12] S. Leedo. Glükohemoglobiin (b-hba1c). 2017.

[13] C. Lozupone and R. Knight. Unifrac: a new phylogenetic method for comparing microbial communities. 2005.

[14] K. Läll, R. Mägi, and K. Fischer. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. 2016.

[15] P. J. McMurdie and S. Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. 2014.

[16] A. C. Morgan and C. Huttenhower. Human microbiome analysis. *PLoS Comput Biol.*, 2012.

[17] E.Õrg, Y. Blum, S. Kasela, and A. J. Lusis. Relationships between gut microbiota, plasma metabolites, and metabolic syndrome traits in the metsim cohort. *Genome Biology*, pages 1–14, 2017.

[18] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. Lecture notes on compositional data analysis. 2011.

[19] H. K. Pedersen and O. Pedersen. Human gut microbes impact host serum metabolome and insulin sensitivity. 2016.

[20] V. K. Ridaura and J. J. Faith. Gut microbiota from twins discordant for obesity modulate metabolism in mice. 2018.

[21] N. A. Sheehan, V. Didelez, P. R. Burton, and M. D. Tobin. Mendelian randomisation and causal inference in observational epidemiology. 2008.

[22] Synlab. Glükoosi taluvuse proov. URL `https://synlab.ee/arstile/laboriteatmik/tulemuste-interpretatsioonid/kliinilise-keemia-uuringud/glukoosi-taluvuse-proov-75-g-gtt-75-g-gluc-po/`.

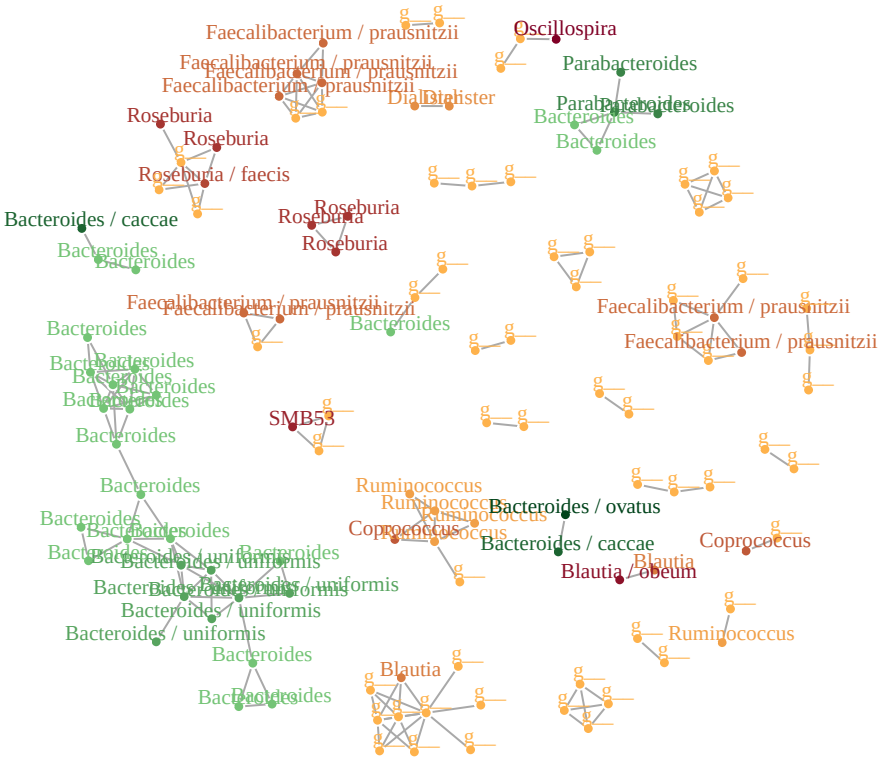
[23] I. Traat. Loengukonspekt bayesi statistika markovi ahelatega. 2017.

[24] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight. Defining the human microbiome. 2013.

[25] Y. Vázquez-Baeza and R. Knight. Impacts of the human gut microbiome on therapeutics. 2018.

[26] A. Willis. Rarefaction, alpha diversity, and statistics. 2017.

A Joonised



Joonis 8: Soolestikus koosesinevad OTUd ($\Phi < 0.3$) perekonniti. Kollastes ning punastes toonides on hõimkonda *Firmicutes* kuuluvad OTUd ning rohelistes toonides hõimkonda *Bacteroidetes* kuuluvad OTUd. Joonise tegemiseks on kasutatud paketti *igraph*

B Tabelid

Tabel 10: Mitteparameetrilise dispersioonanalüüsi *Adonis* väljund

	DF	SS	MS	F	Pr(>F)
OGTT	1	0.343	0.343	1.35	0.068
Jäägid	512	129.790	0.254		
Total	513	130.133			

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Oliver Aasmets,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Mikrobioomi andmete analüüs“, mille juhendajad on Krista Fischer ja Elin Org,

    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, 15. mail 2018