TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

TEHNOLOOGIAINSTITUUT

Margot Saare

Varieeruvus, võimsus ja kvaliteet kvantitatiivses proteoomikas

Bakalaureusetöö

Juhendaja Ülo Maiväli, PhD

Juhendaja David William Schryer, PhD

TARTU 2015

SISUKORD

KASUTATUD LÜHENDID

2D PAGE - *two-dimensional polyacrylamide gel electrophoresis*, kahedimensiooniline polüakrüülamiidgeeli elektroforees

CV – *coefficient of variation*, variatsioonikordaja

EDA - *exploratory data analysis*, andmete esmane analüüs

ESI-MS - *electrospray ionization mass spectrometry,* elektropihustusionisatsioon massispektromeetria

in vitro – katseklaasis

in vivo – elusas, elusal organismil

LC - *liquid chromatography*, vedelikkromatograafia

LC- MS/MS - *liquid chromatography- tandem mass spectrometry*, vedelikkromatograafia tandem mass-spektromeetria

MALDI - *matrix-assisted laser desorption/ionization*, maatriksainega vahendatud laserioniseerimine ja desorptsioon

MS/MS - *tandem mass spectrometry*, tandem mass-spektromeetria

m/z - *mass to charge ratio*, massi ja laengu suhe

SD - *standard deviation*, standardhälve

TOF - *time of flight*, lennuaeg

SILAC - *stable isotope labeling by amino acids in cell culture,* stabiilsete isotoopide aminohapete rakukultuuris märgistamine

SISSEJUHATUS

Valkudel on meie elus suur osatähtsus ja sellepärast tegelevad teadlased aina enam valkude eraldamise ning identifitseerimisega. Proteoomika on suhteliselt uus uurimisvaldkond. Valkude uurimine annab meile infot, kuidas valgud mõjutavad rakulisi protsesse. Valkude uurimisse tõi muutuse tundliku, suure läbilaskevõimega ja pehme mass-spektromeetria meetodite areng.

Mass-spektromeetria andmete analüüsil on tähtis esmane andmete analüüsimine ning katse kvaliteedikontroll. Kõik katsed mõõdavad efekti suurust ning püüavad ennustada, kas efekt on reaalne või tingitud juhusliku hälbega null-efektist. Efekti mõõtmiseks tuleb teada nii tehnilist kui ka bioloogilist varieeruvust.

Andmete esmane analüüs annab parema ettekujutuse uuritavatest andmetest ning võimaldab vähendada katse varieeruvust, et suurema tõenäosusega leida reaalsed efektid. Võimsusanalüüs aitab leida katse võimsust, efekti suurust ning nõutud valimisuurust.

Antud bakalaureusetöö koosneb nii teoreetilisest kui ka eksperimentaalsest osast. Kirjanduslik osa annab lühiülevaate proteoomika ning esmase andmeanalüüsi valdkonnast. Eksperimentaalne osa koosneb tegevuskavandist, mis on loodud mass-spektromeetrial saadud andmete esmaseks analüüsiks, et kirjeldada andmeid ning vähendada katsete varieeruvust.

Käesoleva töö eesmärgiks on koostada tegevuskava mass-spektromeetrial saadud andmete kirjeldamiseks ning esmaseks analüüsiks.

Töö on valminud Tartu Ülikooli Tehnoloogia Instituudis. Tänan oma juhendajat Ülo Maivälja igakülgse abi ja nõu eest.

Märksõnad: proteoomika, andmeanalüüs, statistiline võimsus, mass-spektromeetria

1. KIRJANDUSE ÜLEVAADE

1.1. Valkude identifitseerimise problemaatika

1.1.1. Valgud

Valgud on bioloogilised makromolekulid, mis esinevad kõikides rakkudes. Leidub tuhandeid erinevaid valke, on nii suhteliselt väikeseid kui ka väga suuri polümeere. Valgud määravad bioloogilise mitmekesisuse, nad on molekulaarsed vahendid, mille kaudu geneetiline informatsioon avaldub.

Suhteliselt lihtsad monomeersed subühikud on tuhandete erinevate valkude struktuurivõtmeks. Kõik valgud, nii kõige vanematest bakteriteliinidest kui ka kõige keerulistemast eluvormidest, on kokku pandud sama 20 aminohappe komplektist.

Valkudel on väga palju funktsioone rakus. Nad osalevad informatsiooni dekodeerimises, seondavad teisi molekule transpordiks ja säilitamiseks, reguleerivad hormoonidena biokeemilist aktiivsust märklaudrakus või –koes, esinevad kõrgelt spetsialiseerituna (näiteks antikehadena), pakkuvad rakule tuge ja hoiavad rakkude kuju ning käituvad ensüümina, mis katalüüsivad peaaegu kõiki elusorganismis toimuvaid reaktsioone.

Bioloogiliselt aktiivsete peptiidide ja valkude funktsioonide kohta ei saa teha järeldusi toetudes nende molekulmassile. Looduslikult esinevad peptiidid varieeruvad pikkuses kahest kuni mitme tuhande aminohappejäägini. Isegi kõige väiksematel peptiididel võib olla bioloogiliselt oluline efekt (Nelson & Cox 2005). Tenson ja Ehrenberg (2002) kirjeldasid oma töös peptiide, millel on oma väiksuse tõttu teatud antibiootikumide vastu resistentsus.

1.1.2. Proteoomika

Proteoom on organismi või raku kõikide ekspresseeritud valkude kogum. Proteoomika on proteoomi uurimine, mis keskendub valkude kvantifitseerimisele, lokaliseerimisele, modifikatsioonidele, interaktsioonidele, aktiivsustele ning ka identifitseerimisele**.** Proteoomika oli esmalt suuremjaolt kvalitatiivne protsess, mis koosnes tüüpiliselt võimalikult paljude valkude tuvastamises valkudesegust, koe- või rakulüsaadist. Aastatega on suurenenud proteoomi kvantifitseerimine peamiselt tänu täpsete kvantifitseerimise meetodite arenemisele ning proteoomika algoritmide ja tarkvara loomisele, mis aitavad saadud andmeid analüüsida (Kumar & Mann 2009).

Proteoomi kvantifitseerimist tehakse suhtelisel või absoluutsel tasemel. Suhteline kvalifitseerimine võimaldab võrrelda valgu koguse erinevust erinevate proovide vahel. Absoluutne kvalifitseerimisega mõõdetakse valkude kogust ühikutes ning võrrelda erinevate valkude konsentratsioone (Arike 2012).

Proteoomi analüüs eeldab valkude eraldamist üksteistest (Santoni et al. 2000). Üks rakk koosneb tuhandedest erinevatest valkudest ja sellepärast on ka raske vaid ühe valgu puhastamine. Meetodid, mis eraldavad valke, kasutavad erinevaid valkude omadusi, näiteks nende erinevust üksteisest suuruse, laengu ja sidumisvõime poolest (Nelson & Cox 2005).

1.1.3. Valkude eraldamise meetodid

Valkude eraldamist polüakrüülamiidgeelil läbi isoelektriliste punktide ja molekulaarmassi nimetatakse elektroforeesiks polüakrüülamiidgeeli mittelineaarses pH-gradiendis (2D PAGE, *two-dimensional polyacrylamide gel electrophoresis*). 2D PAGE on üks efektiivsematest meetoditest kompleksete valgusegude eraldamiseks. Meetodi lihtsus peitub selles, et geeli on väga lihtne suurtes kogustes toota ning erinevate värvimismeetoditega saab visualiseerida tuhandeid valke kvantitatiivselt (Blackstock & Weir 1999).

Valkude ja peptiidide eraldamiseks kasutatakse laialdaselt ka vedelikkromatograafiat (LC, *liquid chromatography*), kus liikuva faasina kasutatakse vedelikku ja statsionaaarse faasina kasutatakse poorset tahkist (Capriotti et al. 2011) . Liikuv faas sisaldab valke, mida soovitakse eraldada ning liigub läbi statsionaarse faasi. LC kolonnis saab peptiide eraldada erinevate meetoditega, mille tulemuseks elueeritakse proov erinevatel ajahetkedel (Jemal 2000).

Kõige sagedamini kasutatakse pöördfaasi kromatograafiat, mis põhineb statsionaarses faasis valkude või peptiidide eemaldamises süsisinikahelalatelt nende hüdrofoobsuse järgi. Lühemad süsinikahelad on vähem kinnihoidvamad ja seega kasutusel intaktsete valkude eraldamiseks. Pikemaid ahelaid (C18 kolonn, Thermo Scientific™) kasutatakse peptiidide eraldamiseks (Arike 2012).

Valgu identifitseerimiseks saab kasutada Edmani sekveneerimist või Western blot analüüsi, kuid need on suhteliselt madala läbilaskevõimega meetodid Proteoomikas kasutatakse peaaegu eranditult valkude identifitseerimiseks mass-spektromeetriat. Kaasaegsed mass-spektromeetria meetodid võimaldavad identifitseerida igat valku, mille genoomi järjestus on teada (Guerrera & Kleiner 2005).

1.1.4. Märgistatud ja märgisevaba kvantifitseerimine

Proteoomi kvantifitseerimisel on kaks põhilist lähenemist. Märgistatud kvantifitseerimine seisneb valkude märgistamisel stabiilsete isotoopidega. Märgisevabal kvantifitseerimisel ei märgistata proove. Kõige tavalisemad märgistamise viisid seisnevad peptiidide modifitseerimises isobaariliste märgistega absoluutseks ja suhteliseks kvantifikatsiooniks ja valkude aminohapete metaboolses märkimises stabiilsete isotoopidega (Neilson et al. 2011).

Stabiilsete isotoopide aminohapete rakukultuuris märgistamine (SILAC, *stable isotope labeling by amino acids in cell culture*) on meetod kindlate aminohapete lülitamiseks imetajate kõikidesse valkudesse (Ong et al. 2002). Märgistamist peetakse sageli täpsemaks strateegiaks valkude kvantifitseerimisel, kuid see nõuab kalleid isotoobimärgiseid (Neilson et al. 2011). Märgistamist saab jagada kolmeks sõltuvalt hetkest, millal märgis lisatakse. *In vivo* (elusas, elusal organismil) märgistamist kasutatakse rakkude peal, keda saab kasvatada kultuuris. Siia alla kuulub SILAC. *In vitro* (katseklaasis) enne valkude proteolüüsil ja *in vitro* pärast valkude proteolüüsil lisatud märgistamine sobib igat tüüpi proovidele (Sechi & Oda 2003).

Märgisevaba meetodi eeliseks on asjaolu, et eksperimente saab piiramatult võrrelda, samas kui stabiilsete isotoopidega märgistamise tehnikad on tavaliselt piiratud 2-8 eksperimendiga, mida saab otseselt võrrelda. On tõendeid, et märgisevabad meetodid on kvantifitseerimisel suurema dünaamilise ulatusega. Seega on see kasulik meetod suurte ja globaalsete muutuste jälgimiseks erinevate katsete vahel (Bantscheff et al. 2007).

1.1.5. Alt-üles ja ülalt-alla proteoomika

Proteoomi analüüs algab proovi ettevalmistamisega. Kui valgud lagundatakse ensümaatiliselt peptiidideks, siis on tegu alt-üles analüüsiga. Kui analüüsitakse intaktseid valke, siis on tegu ülalt-alla analüüsiga. Enamasti intaksete ioonidega tehtud analüüsi tulemuseks on valgu ja peptiidi massid ning fragmenteeritud ioonid annavad infot primaarse järjestuse kohta (Yates et al. 2009).

Peaaegu kõik suuremahulised projektid mass-spektromeetria põhises proteoomikas kasutavad trüpsiini, et muuta valgusegud kergemini detekteeritavateks peptiidirühmadeks. Peptiidifragmentide spektrit võrreldakse valgujärjestuste andmebaasidega, et sobitada peptiidijärjestusi valkudesse, mida saab seejärel kvantifitseerida vastavate peptiidide intensiivsuste järgi mass-spektroskoobi detektsioonis (Olsen et al. 2004).

1.2. Mass-spektromeetria

1.2.1. Mass-spektromeetria tööpõhimõte

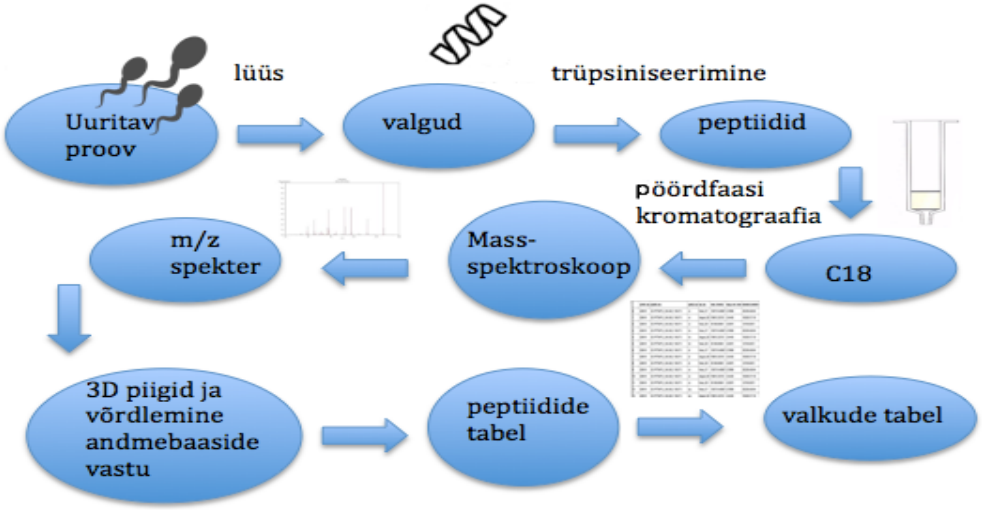
Väga tundliku mass-spektromeetria tehnika arenemine tõi muutuse proteoomikasse (Blackstock & Weir 1999).

Mass-spektromeetria põhine proteoomika on võimalikuks tehtud tänu nii geenide ja genoomijärjestuste kättekättesaadavusele kui ka tehnilistele edusammudele paljudes alades. Kõige märkimisväärsemaks on valgu ionisatsioonimeetodite avastamine ja arendamine, mida tunnustati 2002 aastal Nobeli preemiaga keemias .

Mass-spektromeetria on üha rohkem saanud kompleksete valguproovide analüüsimeetodiks, mis nii identifitseerib kui ka kvantifitseerib tuhandeid valke proovidest (Aebersold & Mann 2003).

Mass-spektromeetriad koosnevad kolmest põhiosast. Esimeseks on ionisatsioonimeetod, mis konverteerib molekulid gaasi-faasi ioonideks. Pärast ioonide tekkimist teine seade, massianalüsaator, eraldab individuaalsed ioonid ning seejärel viidakse ioonid kolmandasse seadmesse- iooni detektorisse. Massi analüsatoor baseerub füüsikalistel omadustel (elektromagnetväli, lennuaeg (TOF, *time of flight*)), et eraldada spetsiifilise massi laengu suhtega ioonid (m/z, *mass to charge ratio*) (Yates 2000).

Kuigi massi analüsaator on oluline mass-spektromeetria osa ja määrab kriitilisi tulemuslikke näitajaid, on oluliseks proteoomika innovatsiooniks kahe uue erineva tehnika väljatöötamine, et suured molekulid ioonifaasi viia (Yates 2000).



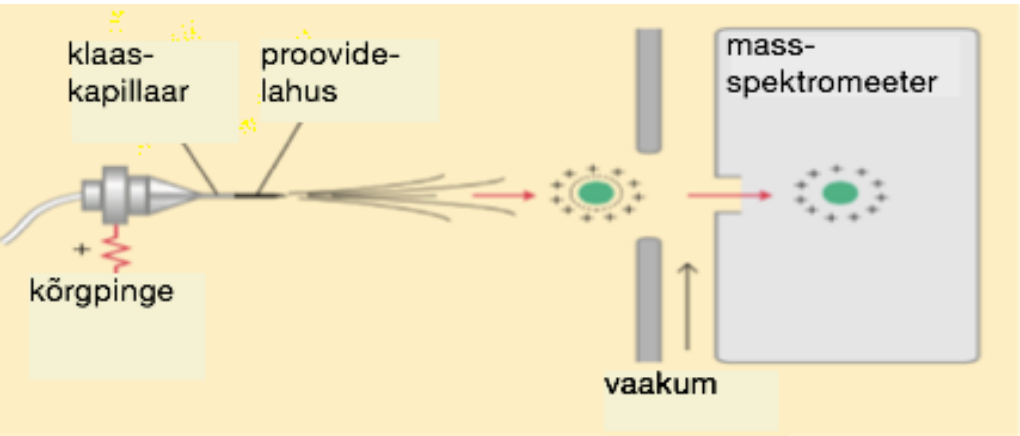
Joonis 1. Antud katses tööprotokoll uuritavast proovist kuni peptiidide tabelini. Uuritav proov lüüsitakse valkudeks ning seejärel trüpsiniseeritakse peptiidideks. Peptiidid lähevad läbi pöördfaasi kromatograafia C18 kolonni ning seejärel mass-spektroskoobi masinasse, kus toimub massi/laengu suhte mõõtmine ning peptiide ja valkude tabeli saamiseks võrreldakse piike andmebaaside vastu.

1.2.2. Elektropihustusionisatsioon

Elektropihustusionisatsioon massispektromeetria (ESI-MS, *electrospray ionization mass spectrometry*) on osutunud kasulikuks nõrgalt seotud, mittekovalentsete komplekside uurimisel, sealhulgas saab uurida valgu interaktsioone inhibiitorite, kofaktorite, metalliioonide, karbohüdraatide, teiste peptiidide ja valkudega. Seega nii ensüüm - substraadi seondumist kui ka nukleiinhappe komplekse (Loo 2000).

Makromolekulid lahuses viiakse otse vedelast faasist gaasifaasi. Proovid viiakse läbi vesilahuses laenguga nõela, mida hoitakse kõrge elektripotentsiaali all ja pritsitakse lahust peene laetud mikrotilkade uduna. Makromolekule ümbritsev lahus aurustub kiiresti ja mitmekordselt laetud makromolekuli ioonid on seega viidud gaasifaasi. Prootonid, mis lisatakse nõela läbimisel, annavad makromolekulile lisalaengu. Massi ja laengu suhet saab analüüsida vaakumkambris.

Joonisel 2 on näidatud elektropihustusionisatsiooni töö põhimõtet. Valgulahus pihustatakse nõela läbimiselkõrgelt laetud tilkadena, mis on kõrgpinge elektrivälja all. Tilgad aurustavad ja ioonid (koos lisatud prootonitega) sisenevad mass-spektromeetriasse massi ja laengu suhte mõõtmiseks (Nelson & Cox 2005).



Joonis 2. Elektropihustusionisatsioon.

1.2.3. Maatriksainega vahendatud laserioniseerimine ja desorptsioon

Maatriksainega vahendatud laserioniseerimise ja desorptsiooni (MALDI, *Matrix-assisted laser desorption/ionization*) eeliseks on kõrge tundlikkus, tolerantsus erinevate puhvrite vastu, kiire andmete kogumine ning lihtne aparatuur (Caprioli et al. 1997).

MALDIs paigutatakse valgud valgust neelavasse maatriksisse. Lühikese laserkiire impulsiga ioniseeritakse valgud ning seejärel desorbeeritakse maatriksist vakuumisse. Maatriksi võtmerolliks on laserikiirguse neelamine ning kaudselt analüüdi aurustumise põhjustamine. Maatriks on ühtlaselt nii prootonidoonor kui ka -aktseptor, ioniseerides analüüti nii positiivselt kui ka negatiivselt.

Tavaliselt kasutatakse MALDI ionisatsiooni puhul massianalüsaatorina lennuaega, mis on kõige lihtsam. TOF analüüs põhineb ioonikogumi kiirendamises detektorisse, kus kõikidele ioonidele antakse sama palju energiat. Kuna ioonidel on sama energia, kuid erinev mass, jõuavad ioonid detektorisse erineval ajal. Väiksemad ioonid jõuavad detektorisse ajaliselt varem ning suuremad hiljem. Detektorisse jõudmise aeg sõltub iooni massist, laengust ning kineetilisest energiast (Lewis et al. 2000).

1.3. Andmeanalüüs

1.3.1. Programmeerimiskeeled

Programmeerimiskeelt saab iseloomustada läbi selle süntaksi ja semantika. Semantika määrab ära, mida on võimalik selles programmeerimiskeeles läbi viia. Keele süntaks määrab selle, kuidas kasutajad saavad ennast väljendada, et uuritav arvutus/toiming läbi viia. Lisaks semantikale ja süntaksile on kasutajale vajalik ka veel töövahendite olemasolu (Ihaka & Gentleman 1996).

Antud töös on kasutatud andmeanalüüsil R-i ning Pythonit. Python on eelkõige inseneride poolt kasutatav, temaga on lihtsam luua täiesti uusi aplikatsioone. R on laialdaselt kasutusel statatistikute poolt. R sisaldab palju rohkem valmis funktsioone andmeanalüüsiks ja statistiliseks modelleerimiseks.

1.3.1.1. Python

Python[1] on programmeerimiskeel, mis on andmete emasel analüüsimisel ja visualiseerimisel sarnaste funktsioonidega nagu R[2]. Pythoni üldisele võimekusele programmeerimisel lisandub hea tarkvara, mis teeb selle tugevaks alternatiiviks andmete analüüsimisel. Ipython on Pythoni töövahend, mis võimaldab mugavat kasutaja keskkonda andmete analüüsiks. Ipython on disainitud, et kiirendada, testida ja puhasta Pythoni koodi. See on eriti kasulik andmete analüüsimiseks ning graafikute joonistamiseks (McKinney, 2013)

Pandas[3] pakett aitab Pythoniga läbi viia andmeanalüüsi terviklikke tööprotokolle. *Pandas* spetsialiseerub andmete analüüsile ning modelleerimisele.

NumPy[4] on põhiline pakett teaduslikuks arvutamiseks.

Matplotlib[5] on pakett andmete visualiseerimiseks jooniste, histogrammide, graafikute ja diagrammide tegemiseks.

[1] https://www.python.org
[2] http://www.r-project.org/
[3] http://pandas.pydata.org/
[4] http://www.numpy.org/
[5] http://matplotlib.org/

SciPy[6] on tarkvara matemaatika, teaduse ja tehnika jaoks.

1.3.1.2. R

R on funktsionaalne keel, mis põhineb S süntaksil. Enamik arvutusi viiakse läbi funktsioonide kasutamisega. R areneb pidevalt, uued võimalused ja funktsioonid ilmnevad iga paari kuu tagant. Arvutusi on lihtne läbi viia erinevate objektide (vektorite, listid, andmefailid, maatriksid) peal tervikuna. R-il on väga võimas tarkvara erinevate graafikute ja histogrammide tegemisel. Andmefaile saab kergelt kasutada modelleerimiseks ning graafiliseks kujutamiseks. R tuleb esialgsete funktsioonidega, kasutajal on võimalik sobivaid pakette juurde laadida (Maindonald 2008).

Dplyr[7] on pakett, mis pakub erinevaid tööriistu andmefailide efektiivseks töötlemiseks. Dplyr on plyri uus versioon, mis keskendub ainult *dataframe* tüüpi andmetabelitele. Andmeanalüüsil kulub palju aega otsustamiseks, mida teha andmetega ning dplyr teeb selle kergemaks individuaalsetefunktsioonidega, mis vastavad kõige tavalistemale andmefailidega tehtavale tehetele (grupeerimine, filtreerimine, muteerimine, summeerimine).

Ggplot2[8] on detailsete võimalustega graafikute/jooniste tegemise pakett.

DataCombine [9] ja zoo on vahendid aegridadega töötamiseks ning andmefailide kombineerimiseks ja puhastamiseks.

Tidyr[10] on pakett, mis aitab andmeid puhastada, et nendega oleks kergem töötada.

Reshape2[11] aitab paindlikult tabelite struktuuri muuta.

1.3.2.Esmane andmeanalüüs

Keerulised proteoomika andmekogud nõuavad analüütilist töötlemist ning visualiseerimist. Proteoomikas rakendatakse statistilisi tehnikaid ning andmekogude statistilisi ja graafilisi aspekte on hakatud üha enam uurima (Kumar & Mann 2009).

[6] http://www.scipy.org/
[7] http://user2014.stat.ucla.edu/abstracts/talks/45_Wickham.pdf
[8] http://ggplot2.org
[9] http://cran.r-project.org/web/packages/DataCombine/DataCombine.pdf
[10] http://blog.rstudio.org/2014/07/22/introducing-tidyr/
[11] http://cran.r-project.org/web/packages/reshape2/index.html

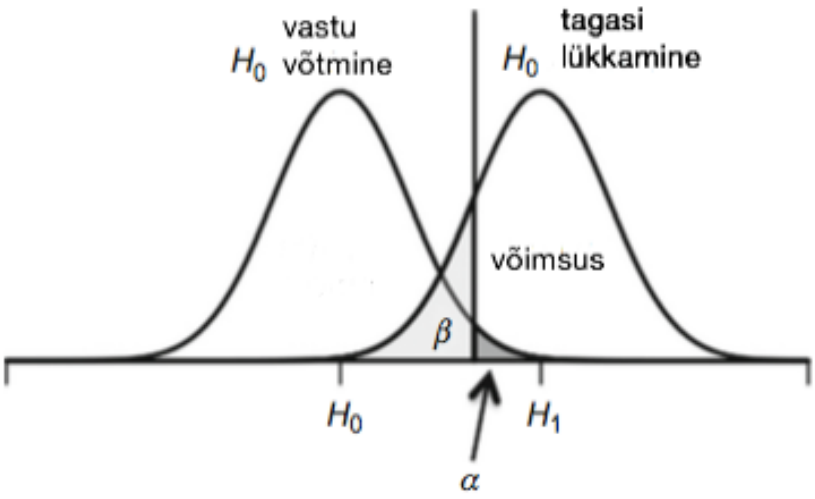
Enamasti on otstarbekas mis tahes statistilist analüüsi alustada esialgsete andmeanalüüsiga, mida nimetatakse andmete ettevalmistavaks analüüsiks (EDA, *exploratory data analysis*). Peamised tegevussuunad on andmete kirjeldamine, kvaliteedikontroll ja esialgsete tööhüpoteeside püstitamine (Chatfield 1986). Esmane andmeanalüüs on peamiselt graafiline meetod.

Antud töös kasutatakse esmasel analüüsil katsete varieeruvuse hindamiseks standardhälvet (SD) ning variatsioonikoefitsienti (CV, *coefficient of variation*), mis mõõdab varieeruvust uuritud tunnuses ning on abiks jaotuste võrdlemisel erinevate proovide vahel. CV arvutamiseks jagatakse standardhälve uuritava tunnuse keskmisega (Abdi 2010). Graafiliselt on töös kirjeldatud jaotusi histogrammide ning karp-vurrud diagrammidega.

1.3.3. Võimsus

Eduka kvantitatiivse uuringu eelduseks on piisav statistiline võimsus, millega uuring on võimeline avastama teaduslikult olulisi efekte (Levin 2011). Statistilise võimsus on tõenäosus, et nullhüpotees lükatakse tagasi kui see nullhüpotees ongi vale.

Võimsus on (1- β) ning β on tõenäosus tüüp II veaks. Tüüp II viga tehakse kui jäädakse nullhüpoteesi juurde, kuigi see tegelikult ei kehti (Faul et al. 2007). Statistiline võimsus sõltub kolmest aspektist: olulisuse nivoost (α) ehk tüüp I vea tõenäosusest, valimisuurusest ning tegelikust efekti suurusest (Coheni d = efekti suurus / SD). Tüüp I viga tehakse kui lükatakse ümber nullhüpotees, kuigi nullhüpotees on tegelikult tõene.



Joonis 3. Võimsusanalüüs (Maiväli, 2015)

Joonisel 3 on näha, et efekti suurusega kasvab ka võimsus. Võimsus näitab tõenäosust, et nullhüpotees lükatakse tagasi kui see nullhüpotees ongi vale, β on tõenäosus, et nullhüpotees võetakse vastu kuigi see tegelikult ei kehti. α näitab olulisuse nivood ehk maksimaalset lubatavat I tüüpi vea tõenäosust.

A priori võimsusanalüüsi peetakse enamike teadlaste poolt ideaalseks võimsusanalüüsi tüübiks. Sellisel analüüsil määrab teadlane sobiva efekti suuruse (d), olulisusenivoo ning soovitud võimsuse. Selliste parameetrite alusel on võimalik arvutada vajalik valimisuurus (N). Tavaliselt kasutatakse olulisuse nivoona 0.05 ja aksepteeritav võimsus algab 80%-st (Erdfelder et al. 1996). Samas on teaduses tegelikult kasutatavad võimsused sageli palju madalamad. Näiteks on tüüpilise neuroteaduste valla uuringu võimsus umbes 25%, mille tagajärjel mitte ainult ei jää enamikud tegelikud efektid avastamata vaid kirjeldatud efektid on ka sageli oluliselt ülehinnatud suurusega või suisa vale suunaga ning tulemustel on madal korratavuse tase (Button et al. 2013, Gelman & Carlin 2014).

Kvantitatiivsetes uuringutes on valimi suurus tähtis aspekt. Valimi suuruse suurenedes suurenevad ka rahalised kulutused katse läbi viimiseks. Lisaks mõningad uuritavad teemad panevad ise piirangu valimisuurusele, mida on reaalselt võimalik saavutada (Fox & Mathers 1997).

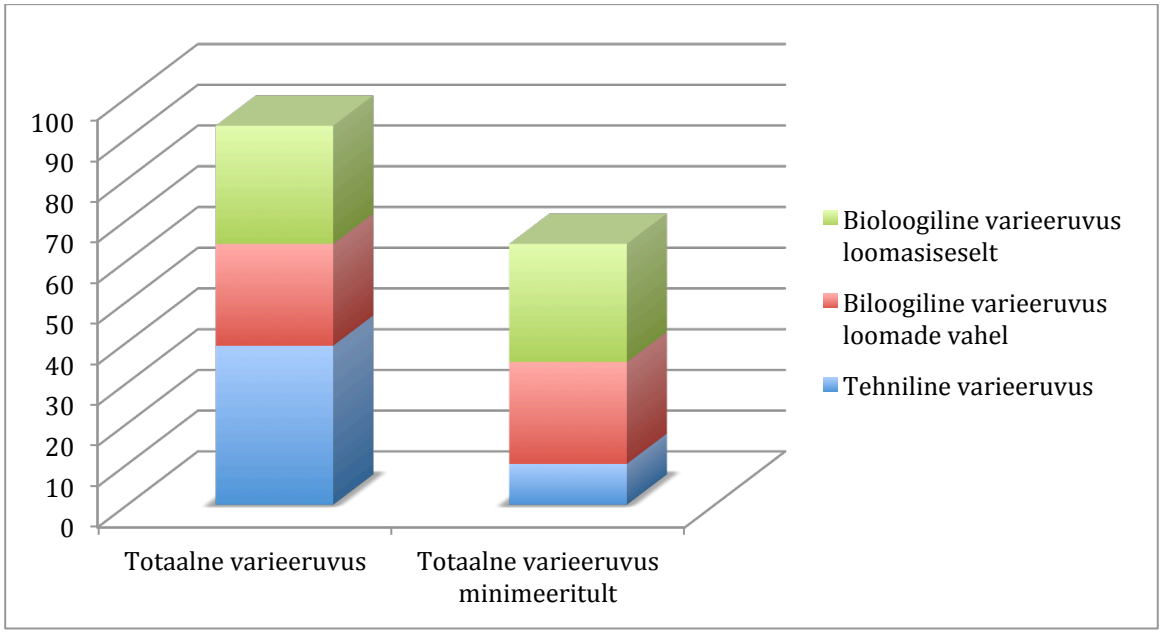
Suure läbilaskevõimega tehnoloogiaid kasutatakse laialdaselt valgueksрessiooni muutuste hindamiseks näiteks mitmesuguste haiguste (eriti vähi) tagajärjel, kuid tihti jäetakse tähelepanuta *batch* efekt, mis esineb siis kui mõõtmistulemused on mõjutatud labori olukorra, reagentide või personali poolt. Proovidel, mida märgistatakse samaaegselt on sama suur lisatud tehniline varieeruvus, kuid proovid, mida märgistatakse erinevatel aegadel on erinevate lisatud tehnilise varieeruvusega.

On oluline, et sellist tüüpi tehniline varieeruvus ei segaks bioloogilise varieeruvuse väärtuseid. Probleem tekib kui *batch* efekti andmeid korreleeritakse tulemustega ning selle tagajärjeks on ebakorrektsed järeldused. Efekti vähendamiseks saab proovid blokkidena katse/kontroll süsteemil läbi masina lasta ning planeerides katsed täpselt ette. *Batch* efekte saab tuvastada näiteks märgistades erinevatel päevadel analüüsitud proove erinevate

värvidega, mis aitab visuaalselt avastada efekte (Leek et al. 2010, [12]). Kvantiil normaliseerimine võib *batch* efekti vastu aidata.

1.3.4. Andmeanalüüs sõltuvalt bioloogilistest ja tehnilistest replikaatidest

Olenevalt sellest, millised replikaatide tüüpe on katses kasutatud, saab otsustada, milliseid statistilisi teste on võimalik saadud andmetega läbi viia ja milliseid järeldusi võib teha (Karp & Lilley 2007). Eksperimentaalsed replikaadid jagunevad tehnilisteks ning bioloogilisteks replikaatideks. Replikaat kujutab endast lihtsalt katse mitut kordust.

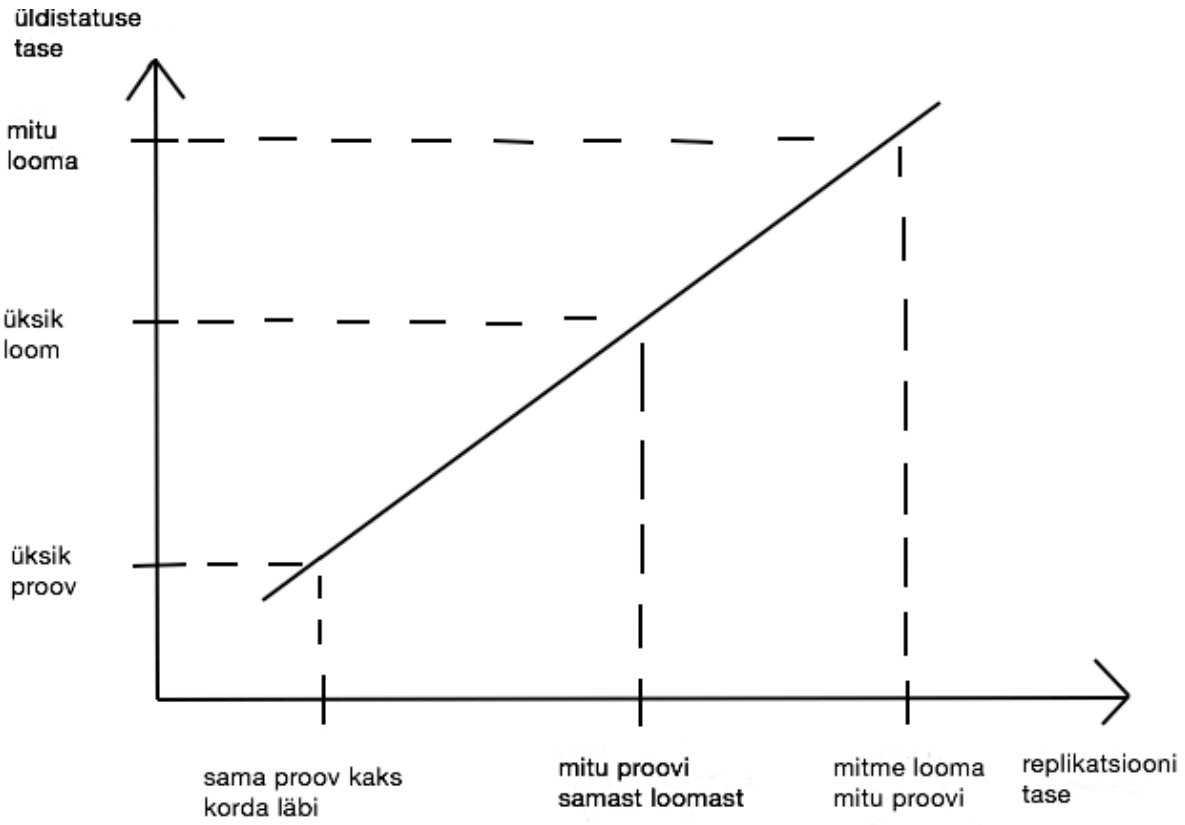


Joonis 4. Totaalne varieeruvus algselt ning totaalne varieeruvus minimeeritult.

Joonis 4 näitab, et totaalne varieeruvus koosneb tehnilisest ja bioloogilisest (loomasisesest ja loomade vahelisest) varieeruvusest. Sealjuures saab tehnilist varieeruvust andmeanalüüsi käigus minimeerida, kuid bioloogilist varieeruvust tuleks optimaalselt kirjeldada.

Tehniline replikaat näitab varieeruvuse suurust katse mõõtmises ja/või proovi ettevalmistamises ning seda kasutatakse enamasti katse võimsuse hindamisel. Bioloogiline replikaat on tähtis bioloogilise varieeruvuse hindamisel ning varieeruvus on kasutusel katse võimsuse hindamisel, et leida reaalseid efekte (Maiväli 2015).

[12] http://www.molmine.com/magma/global_analysis/batch_effect.html



Joonis 5. Katse tulemuse üldistatavuse tase seos replikatsiooni tasemega.

Jooniselt 5 on näha, et tehniline replikaat sisaldab endas sama proovi, mis on kaks korda läbi MS läinud. Bioloogiline replikaat looma tasemel sisaldab endas kõiki ühe looma iseseisvalt kogutud katsepunkte, totaalne varieeruvus on arvutatud kõikide loomade kõikidest katsepunktidest ning sisaldab nii tehnilist, loomasisest kui ka loomade vahelist varieeruvust.

Tehniliste replikaatide põhjal saab järeldusi teha vaid mõõtesüsteemi kvaliteedi kohtamitte aga katse ja kontrolli erinevuse (efekti suuruse) mittejuhuslikkuse kohta. Bioloogilised replikaadid sisaldavad infot süsteemi bioloogilise varieeruvuse kohta, mis võimaldab eristada katsete vahelist eksperimentaalsest töötlusest tingitud efekti katsesisesest loomulikust varieeruvusest.. Ilma bioloogilist varieeruvust hindamata ei ole võimalik põhimõtteliselt otsustada, kas nähtud katse-efekt on tingitud teadlase poolt toime pandud spetsiifilisest eksperimentaalsest töötlusest või hoopis juhulikust valimiefektist.

Katse müra võib defineerida kui andmete mõõtmiste tulemuste kõikumist, mis vähendab signaali puhtust ja sellest tulenevalt ka katse tundlikkust. Katse soovitud võimsuse saamiseks tuleb arvesse võtta seda müra ja katse disaini muuta vastavalt kas rohkem replikaate tehes või muutes bioloogiliste ja tehniliste replikaatide osakaalu (Karp & Lilley 2007).

Replikaadid on justkui sisemised kontrollid, et näha kuidas katse on tehtud, nad võivad esile tuua kõrvalekalduvaid väärtusi, mille esinemisel peaks katset kontrollima või rohkem uurima. (Vaux et al. 2012).

2. EKSPERIMENTAALOSA

2.1. Töö eesmärgid

Käesoleva töö eesmärgiks on koostada tegevuskava, mis võimaldaks kvantitatiivsel märgisevabal mass-spektromeetrial saadud tulemusi ja andmeid hinnata, kirjeldada ning analüüsida.

Selle saavutamiseks püstitatakse järgmised ülesanded:

- Leida optimaalne analüütiline tegevuskava peptiididelt valguintensiivsusteni.
- Kirjeldada katse/proovi tehnilist ja bioloogilist varieeruvust.
- Hinnata katse võimet näidata bioloogiliselt relevantse suurusega efekte.
- Kasutada eelnevat optimaalse analüütilise tegevuskava loomisel kvantitatiivses proteoomikas.

2.2. Materjal ja metoodika

2.2.1. Materjal

Viies katses analüüsiti kokku 13 erineva pulli spermi rakuvälist ja rakusisest komponenti. Katse viidi läbi Triin Tamsalu, Liisa Arikese ja Sergo Kasvandiku poolt enam kui aasta jooksul. P katsed on plasmakatsed, kus vaadeldi spermide rakuvälises keskkonnas leiduvad valke, C katsetes lüüsiti spermid ja uuriti lüsaatides leiduvaid lahustunud valke. Esmane andmeanalüüs toimus Max Quant programmiga, mis annab peptiidide intensiivsuste tabeli ja eraldi valgu intensiivsuste tabelid. Katsed viidi läbi märgisevaba LC/MS/MS meetodiga.

Konkreetsed andmed valiti analüüsimiseks, sest pullide andmefail sisaldab nii tehnilisi kui bioloogilisi replikaate, kusjuures bioloogilised replikaadid on nii loomasisesed kui loomade vahelised. See võimaldab meil üksteisest lahutada erinevat tüüpi varieeruvused, mis omakorda võimaldab anda hinnagu minimaalsetele valimi suurustele, mida on teoreetiliselt parima katsedisaini ja andmeanalüüsiga võimalik saavutada bioloogiliselt relevantsete efektide kirjeldamiseks. Seega pulliandmete valik ei ole seega seotud pullidega, vaid andmete struktuuriga. Tabelis 1 on kokkuvõtlikult esitatud loomade jagunemine viie katse vahel.

Tabel 1. Loomade jagunemine viie erineva katse vahel.

Loom	PA	PB	PC	CA	CB
Ciro	1			1	
Delgado	2			1	
Miracle	3			3	
MisterX	4			3	
Rodeo	2			2	
Welton	4			3	
Boldin		2	2		3
Jerter		1	2		3
Joris		3	1		4
Lukard		3			2
Lukas		2	1		3
Rossen		1	3		4
Sigfrid			3		2

2.2.2. Metoodika

2.2.2.1. Kasutatavad andmeanalüüsi programmid

Programmide ja pakketide lühikirjeldused on esitatud töö kirjanduse ülevaates. R-i ning Pythoni kasutamine andmete analüüsiks on valitud vastavalt aspektidele:

- kättesaadavus;
- kasutamise lihtsus;
- võimalus töötada suurte andmefailidega;
- suur võimsus.

R-i ning Pythoni eelis Exceli ees seisneb asjaolus, et protsesse saab otse käsurealt käivitada, töötamine paljude ridade või veergudega on lihtne, on võimalik valida vaid endale sobivad veerud, millele arvutusi teha. Samuti on lihtne välja valida ja filtreerida read, mida analüüsimisel kasutatakse. Koodi valmis kirjutamisel on seda väga lihtne uuesti kasutada teise andmefaili peal. Vigu tekib vähem, sest koodi kontrollimine on lihtne, vigase koha leiab kergemini, sest koodi vaadates on kõigile täpselt näha, mis operatsioonid on tehtud. R-i ning Pythonisse on võimalik sisse laadida erinevaid failitüüpe. Nende jaoks on tehtud pakette, mida saab kasutada vastavalt sellele, milliseid andmeid analüüsitakse.

2.2.2.2. Andmete analüüs

Antud töös on kasutatud tabelitega töötamisel nii Pythonit kui ka R-i.

Peptiidide ning valkude tasandi analüüsimiseks puhastati eelnevalt andmefailid duplikaatidest ning ebatähtsatest veergudest, jäeti alles vaid veerud, mida antud töös vaja läheb (intensiivsused, peptiidi ID-d, valgu ID-d, loomade nimed, unikaalsed peptiidid, fasta päised).

Analüüsimisel arvutati välja tavalised statistikud nagu mediaan, keskmine ja standardhälve. Katsete varieeruvuse mõõtmiseks leiti CV, mis kujutab endast intensiivsuse standardhälbe ja intensiivsuse keskmise suhet. Chebyshevi teoreemi järgi on sõltumata andmete jaotusest minimaalselt 75% keskväärtustest kahe standardhälbe sees ja 89% kolme standardhälbe sees. Sellega seoses kasutame varieeruvuse mõõtmiseks (CV) lineaarset skaalat. Keskmiste intensiivsuste võrdlemisel kasutame kümnendlogaritmilist skaalat ning võimsusanalüüsides kasutame kahendlogaritmilist skaalat (kuna valguintensiivsuste jaotus on ligikaudu lognormaalne) .

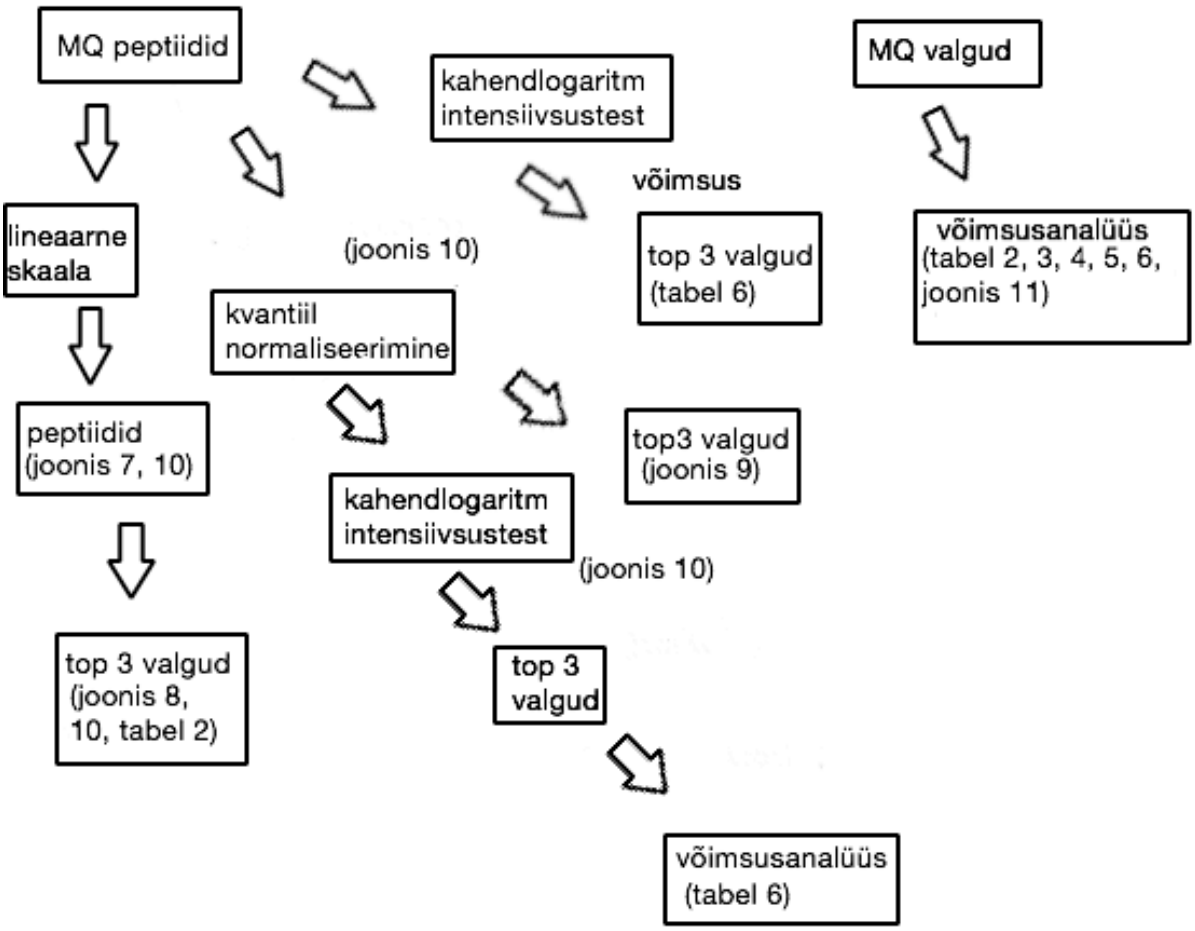
Andmete visualiseerimiseks ning jaotuse analüüsimiseks kasutatakse histogramme. Histogramm näitab tunnuse väärtuste (näiteks valguintensiivsuste) jaotumist nende esinemise sageduse järgi uurija poolt etteantud väärtusvahemikes. Tunnuse (nt normaliseeritud standardhälbe, keskmise intensiivsuse) väärtused on kujutatud x-teljel ja tunnuse sagedused y-teljel.

Võrdluseks on valgu tasemel võrreldud nii peptiidi tasemel kvantiil normaliseeritud kui ka ilma normaliseerimata andmeid. Kvantiil normaliseerimine on tehnika mitme jaotuse keskmistamiseks, mille eesmärk on erinevate mass-spektroskoopia katsete vahel juhuslikult tekkivate süstemaatiliste erinevuste normaliseerimine (Bolstad & Bolstad 2001). Kvantiilnormaliseerimine on laialt kasutatud mRNA-de tasemete mõõtmiskatsetes, aga senimaani mitte proteoomikas (Robinson & Oshlack 2010).

Ühe looma tehniliste replikaatidena kasutatakse looma ühe ajapunkti kaht kordusmõõtmist (kaks järjestikkust jooksutamist läbi MS-i). Loomasisese bioloogilise varieeruvuse leidmiseks võetakse analüüsimiseks looma kõik ajapunktid. Täieliku bioloogilise varieeruvuse saame totaalsest varieeruvusest (kõik katsepunktid) maha lahutades tehnilise varieeruvuse. Bioloogiline varieeruvus sisaldab endas ka tehnilise varieeruvuse komponenti.

Võimsusanalüüsi teostamiseks võetakse (1) katse iga peptiidi intensiivsusest kahendlogaritm ning (2) nõnda logaritmitud andmetest valgu- või peptiidiintensiivsuste aritmeetiline kesmine ning standardhälve. Neist statistikutest võetakse seejärel mediaan üle antud proovi kõikide peptiidide või valkude, mida kasutatakse võimsusanalüüsil kõigi valkude esindajana. Selle meetodi eelduseks on, et enamus raku valke omab sarnast suhtelist bioloogilist varieeruvust ning et arvutades paraleelselt paljude väikeste valimite varieeruvust saame me nende varieeruvuste jaotuse keskväärtuse näol adekvaatse hinnangu tegelikule bioloogilisele varieeruvusele, mis ei ole kallutatud väiksest valimist tuleneva valimivea poolt. Kahekordse efekti saamiseks liidetakse esialgsele mediaan-intensiivsuse väärtusele 1, kolmekordse efekti saamiseks aga 1,5. Standardhälbe vastav väärtus leitakse proportsionaalselt keskmise intensiivsusega. Võimsusanalüüsi teostamiseks kasutatakse programmi GPower [13].

[13] http://www.gpower.hhu.de/



Joonis 6. Tegevuskava ülevaade.

Andmeanalüüs viiakse läbi nii MQ peptiidide andmetest kui MQ valkude andmetest (joonis 6). Lineaarsel skaalal grupeeritakse andmed nii peptiidide järgi kui ka pannakse valke kokku kolme kõige suurema intensiivsusega peptiididest. MQ peptiidi andmetest võetakse kahendlogaritm ning seejärel leitakse katse võimsus. MQ peptiididele tehakse ka kvantiil normaliseerimist ning seejärel võetakse kahendlogaritm intensiivsustest ning pannakse valgud kokku tugevamatest peptiidest ning leitakse katse võimsus. MQ valkude andmetest leitakse katse võimsus ning bioloogiliste ja tehniliste replikaatide varieeruvus.

2.2.2.3. Katsete kordused

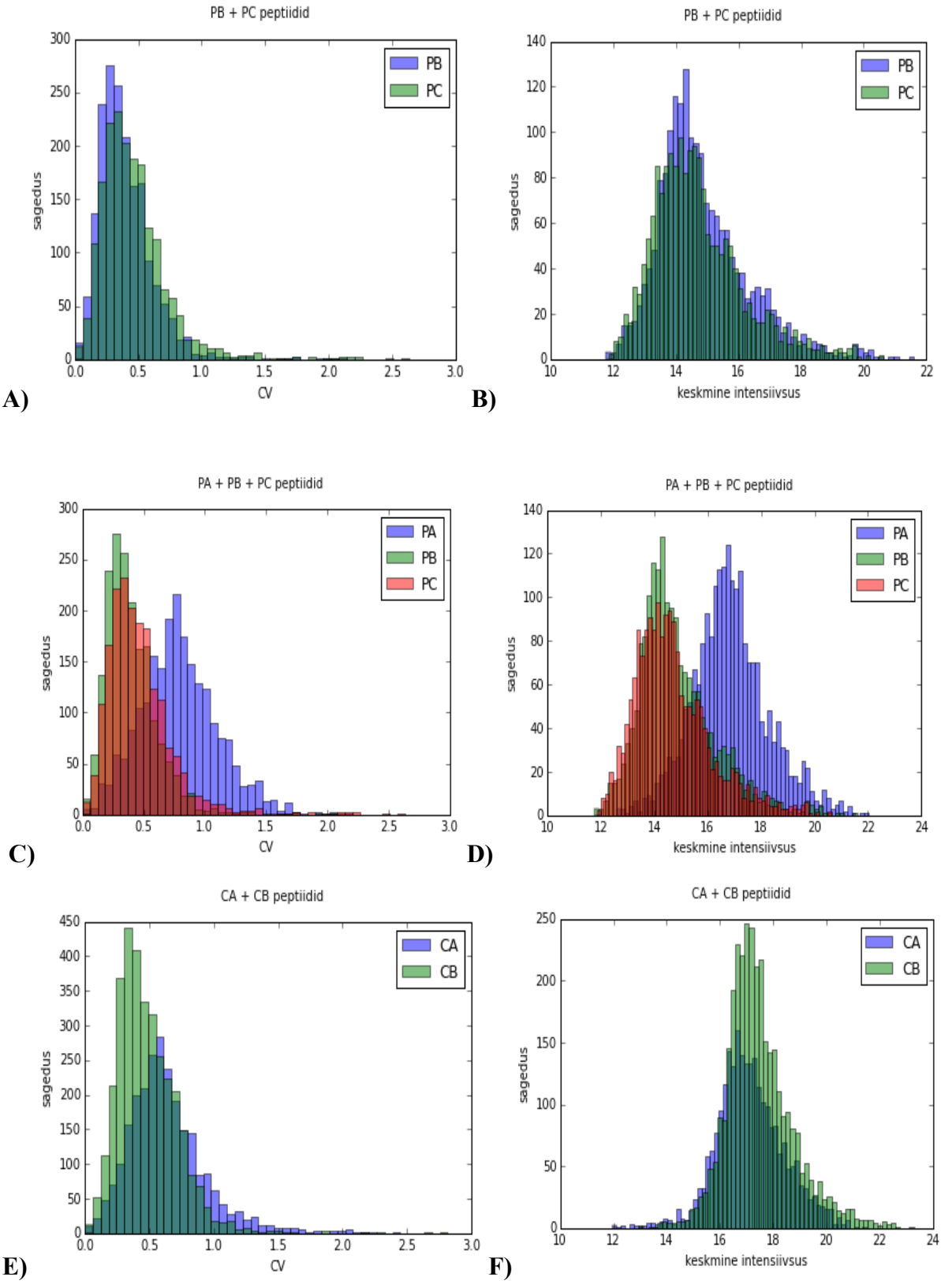
Tabel 1-s on välja toodud loomade jagunemine 5 katseseeria vahel, kus on kokku 13 looma. Iga looma iga katsepunkti kohta on 2 tehnilist kordust, mis võimaldab leida tehnilise varieeruvuse iga proovi piires. Igal loomal on sel viisil analüüsitud keskmiselt 3-4 iseseisvat bioloogilist proovi.

2.3. Tulemused

2.3.1. Peptiidide intensiivsuste jaotus katsetes

Katsete peptiidide jaotuse hindamiseks koostati histogrammid, kus on näha peptiidide CV jaotust sõltuvalt katsest. Analüüs on tehtud ilma andmete esmase normaliseerimiseta ja filtreerimiseta. CV on kasulik statistik selleks, et võrrelda varieeruvusi erinevate katseseeriate vahel isegi siis kui andmeseeriate intensiivsuste keskmised väärtused on erinevad. Histogrammides on erinevate katsete jaotused paigutatud üksteise peale, et katseid omavahel võrrelda. Varieeruvuse hindamine on kasulik selleks, et otsustada milliseid katseid saab paralleelselt analüüsida ning milliseid peaks analüüsima eraldi, et uuritava katse tulemused oleksid usaldusväärsemad.

Joonisel 7 on näidatud, et katse PA on katsest PB ja PC erineva intensiivsuste varieeruvusega. Katse PB ja PC varieeruvused on sarnased ning neid kahte katset saab koos edasi analüüsida. Katse CA ja CB on varieeruvuselt üksteisest erinevad ning neid tuleks analüüsida eraldi.



Joonis 7. A) Katse PB + PB CV väärtuste jaotus peptiidi tasemel **B)** Katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel. **C)** Katse PA + PB + PC CV väärtuste jaotus peptiidi tasemel. **D)** Katse PA + PB + PC keskmiste intensiivsuste logartimitud väärtuste jaotus peptiidi tasemel. **E)** Katse CA + CB CV väärtuste jaotus peptiidi tasemel. **F)** Katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel.

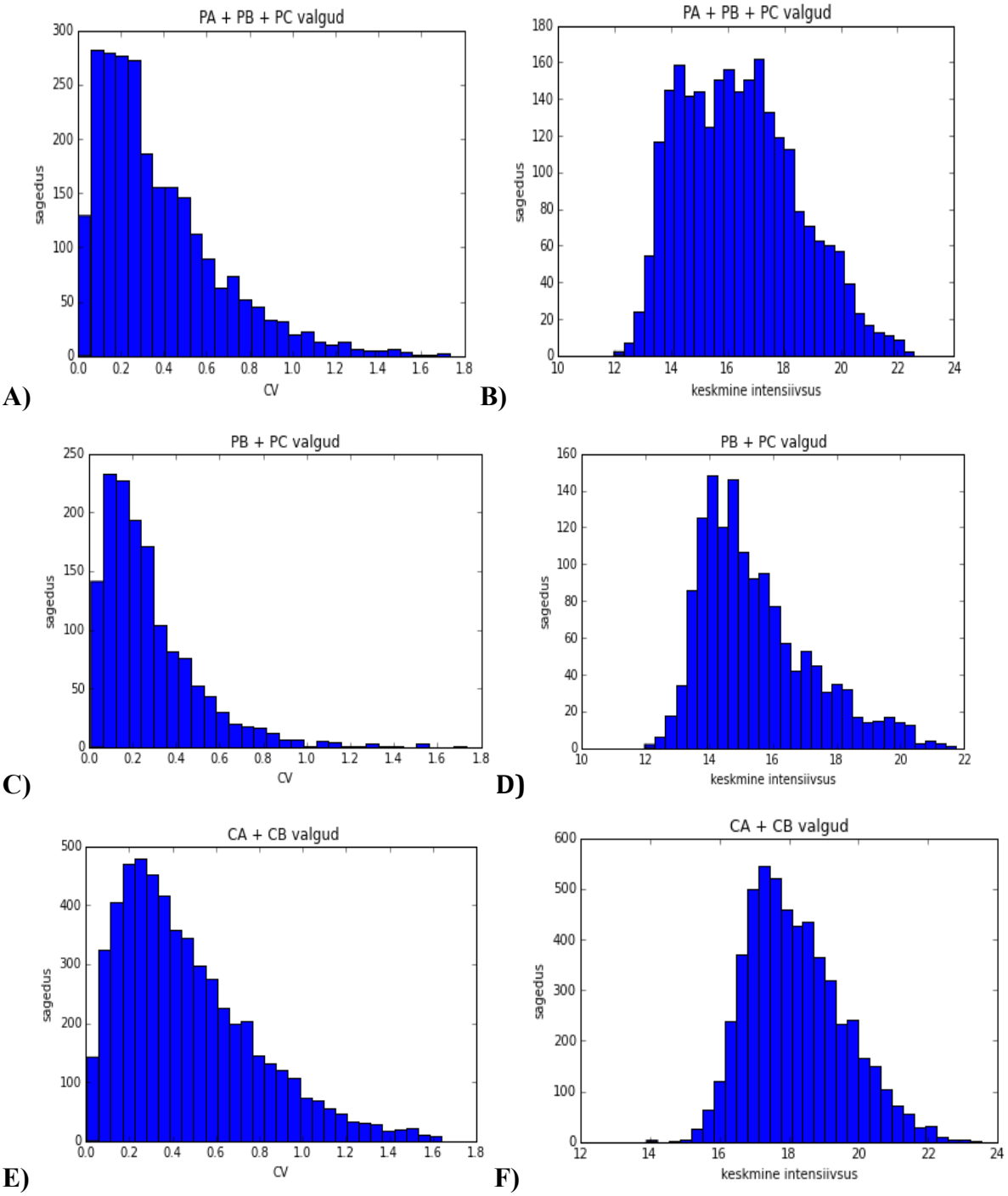
2.3.2. Valkude intensiivsuste jaotus katsetes

Valkude jaotuse hindamiseks katsetes tehti histogrammid, kus on näha valkude normaliseeritud standardhälbe jaotust sõltuvalt katsest. Histogrammide võrdlemisel selgub kui suur on varieeruvus nii katse sees kui ka erinevate katsete vahel.

Valkude kokkupanemiseks pandi valgud kokku kolme kõige kõrgema intensiivsusega peptiidist aritmeetilise keskmisena, sealhulgas filtreeriti välja valgud, millel oli vähem kui kolm unikaalset peptiidi. Selline meetod on laialt kasutatav. Võrdluseks tehti valkude kokkupanek nii kvantiil normaliseeritud kui ka ilma normaliseerimata andmetega. Karp-vurrud diagramm on tehtud võrdlemiseks intensiivsuste jaotust enne ja pärast kvantiil normaliseerimist ning filtreerimist. Karp-vurrud diagramm on joonis, kus on tugevama joonega välja toodud mediaan ning selle ümber on kast, mis näitab andmete jaotuse alumist (25%) ja ülemist (75%) kvartiili (pooled andmepunktid jäävad kasti sisse) ning alla ning üles jääb joon, mis näitab vastavalt valimi – 1,5 x IQR (kvartiilide vaheline kaugus) + 1,5 x IQR. Nendest joontest kaugemal olevad andmepunktid on defineeritud *outlieritena* ja on näidatud ükshaaval punktidena.

2.3.2.1. Normaliseerimata andmetest valkude tase

Joonisel 8 on näidatud, et ka valkude kokkupanekul kolmest suurima intensiivsusega peptiidist katse PA + PB + PC on suurema varieeruvusega kui PB + PC. PB + PC on väiksema varieeruvusega ja neid saab analüüsida koos. Katse CA + CB valkude tasemel varieeruvus on P (plasma) katsetest erineva varieeruvusega.

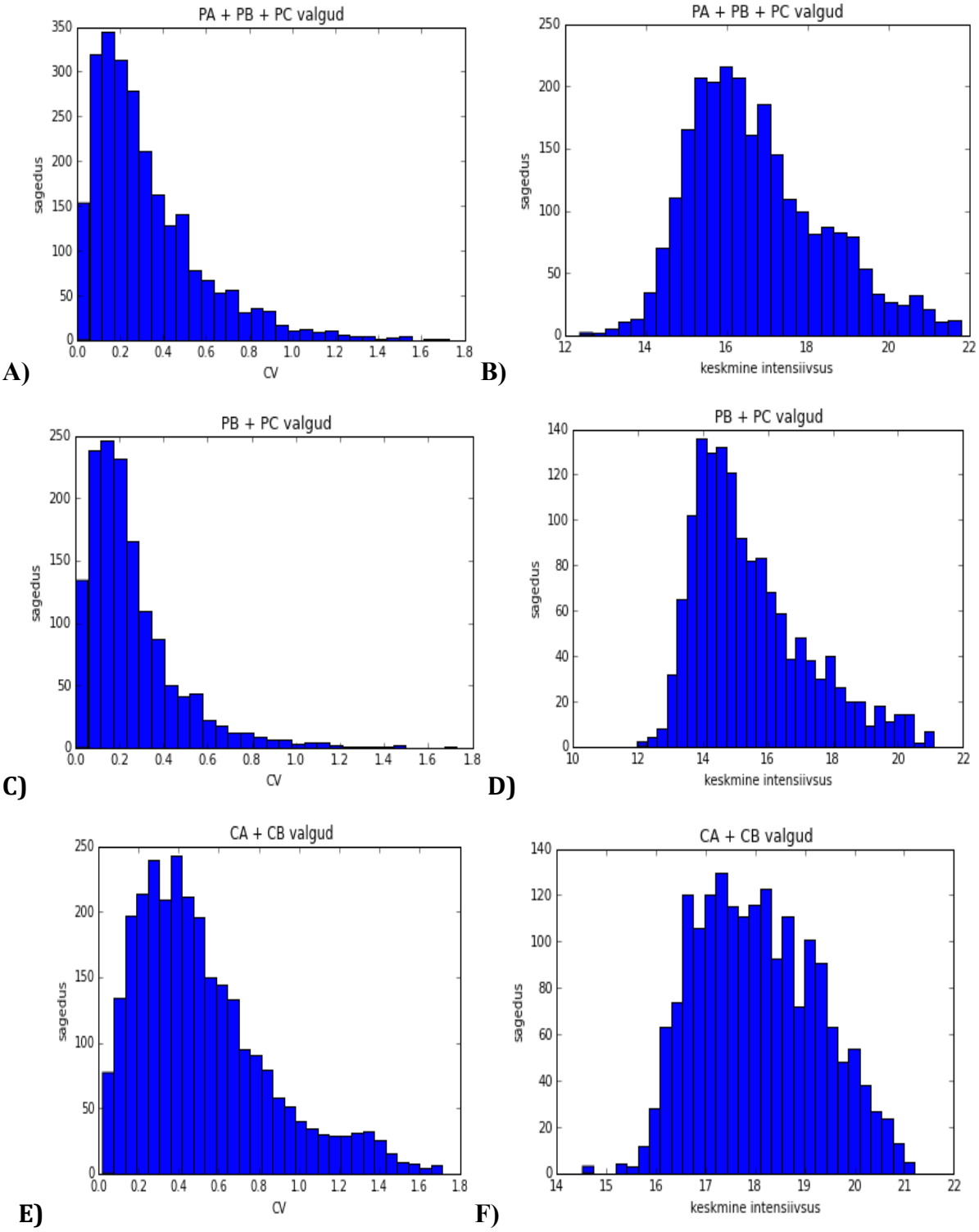


Joonis 8. **A)** Normaliseerimata katse PA + PB + PC CV väärtuste jaotus valgu tasemel. **B)** Normaliseerimata katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel. **C)** Normaliseerimata katse PB + PC CV väärtuste jaotus valgu tasemel. **D)** Normaliseerimata katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel. **E)** Normaliseerimata katse CA + CB CV väärtuste jaotus valgu tasemel. **F)** Normaliseerimata katse CA + CB keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel.

2.3.2.2. Kvantiil normaliseeritud andmetest valkude tase

Jooniselt 9 järeldan, et katse PA + PB + PC on ka kvantiil normaliseeritud andmetest valkude kokkupanemisel kolme suurima intensiivsusega peptiidist kõige suurema varieeruvusega.

Katse PB + PC on sarnase varieeruvusega valkude tasemel. CA + CB varieeruvus on P (plasma) katsete varieeruvusest erinev. Kvantiil normaliseerimine ning kolme tugevaima intensiivsuga peptiidist valkude kokkupanek vähendab varieeruvust eelkõige suurema varieeruvusega katsetel (PA + PB + PC).



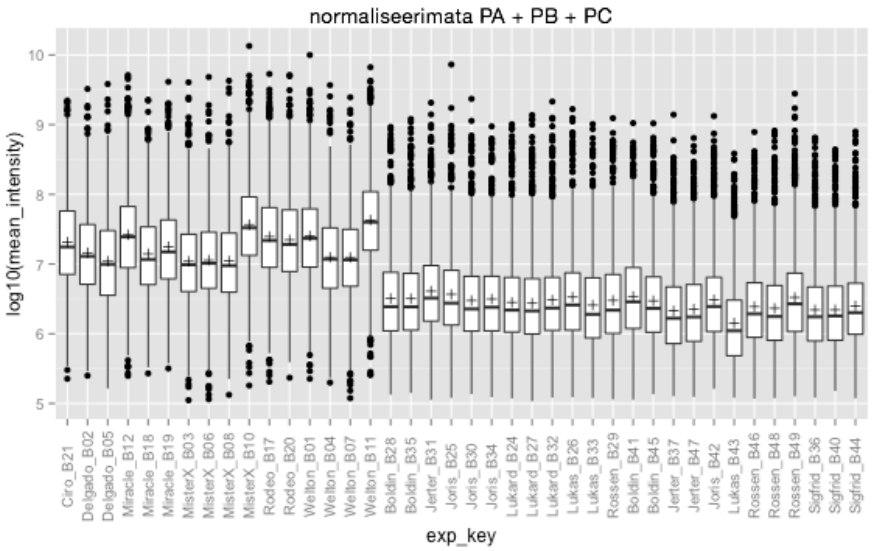
Joonis 9. A) Normaliseeritud katse PA + PB + PC CV väärtuste jaotus valgu tasemel. **B)** Normaliseeritud katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel. **C)** Normaliseeritud katse PB + PC CV väärtuste jaotus valgu tasemel. **D)** Normaliseeritud katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel. **E)** Normaliseeritud katse CA + CB CV väärtuste jaotus valgu tasemel. **F)**

Normaliseeritud katse CA + CB keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel.

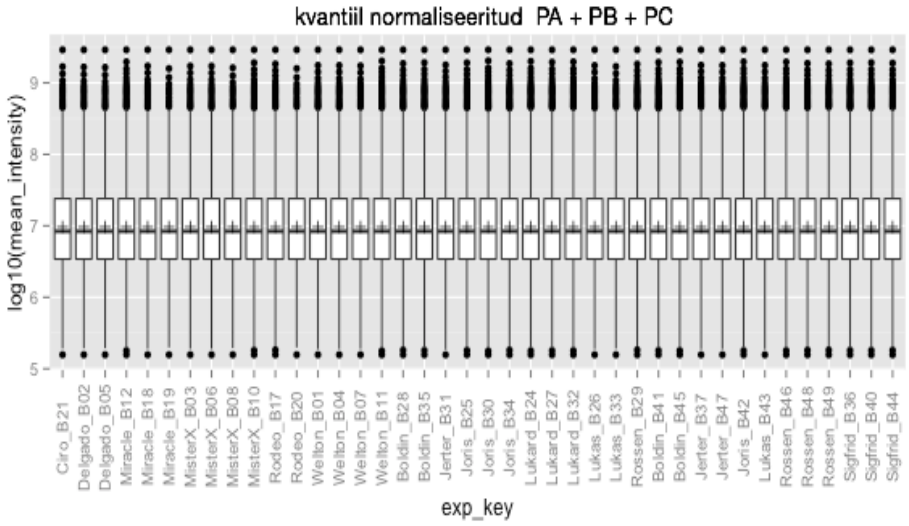
2.3.2.3.Karp-vurrud diagrammid enne ja pärast normaliseerimist ning filtreerimist

Karp-vurrud diagrammid illustreerivad, mida kvantiil normaliseerimine teeb ning kuidas jaotus näeb välja pärast valgu kokku panemist kolmest suurima intensiivsusega peptiidist.

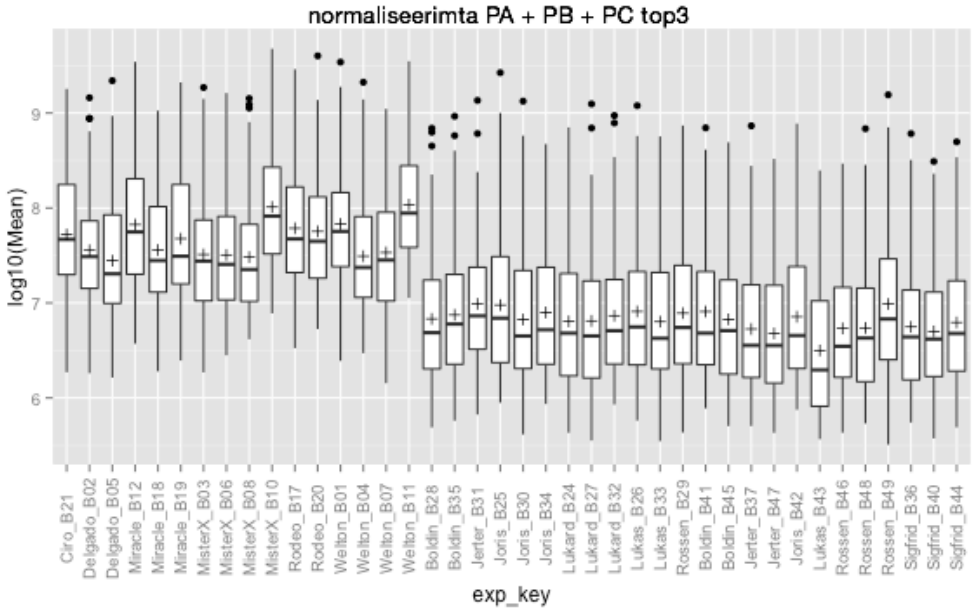
Jooniselt 10 näeb, et peptiidi tasemel katsel PA + PB + PC on intensiivsuste varieeruvus suur ning on näidatud, et normaliseerimine keskmistab peptiidide jaotused, et saaks edasi teha valkude kokkupanemist kolmest kõige suurema intensiivsusega peptiidist. Valgutasemel kvantiil normaliseerimine varieeruvust ei vähendanud, sest filtreerimine kolme suurema intensiivsuse järgi juba selekteerib välja väiksema varieeruvusega peptiidid. Katse PA + PB + PC varieeruvus on suurem kui katsel PB + PC nii peptiidi tasemel kui valgu tasemel.



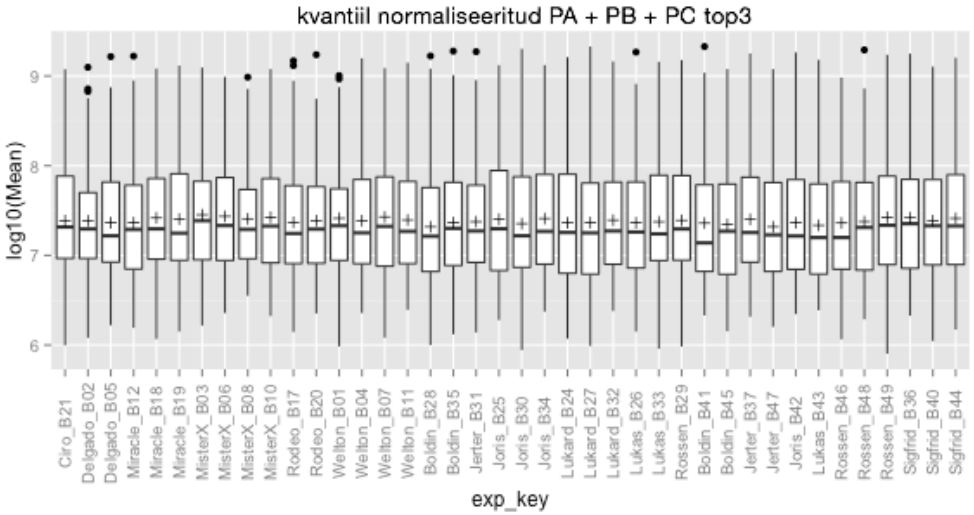
A)



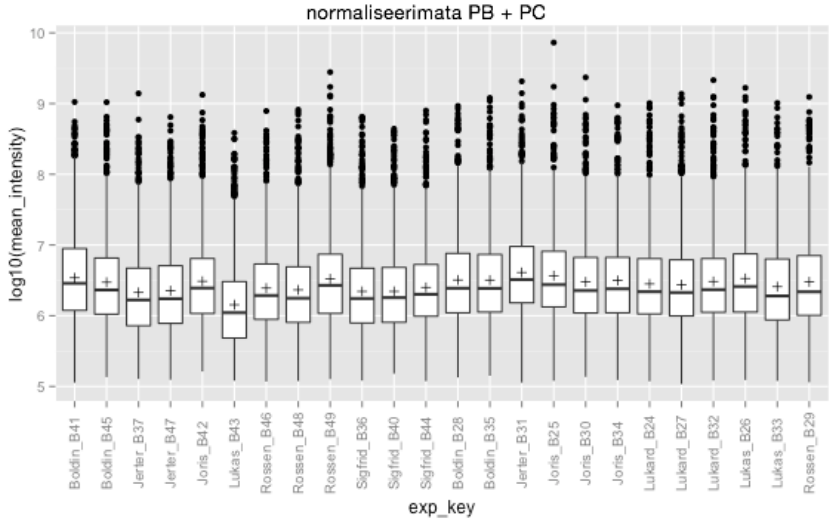
B)


normaliseerimta PA + PB + PC top3
log10(Mean)
exp_key
Ciro_B21
Delgado_B02
Delgado_B05
Miracle_B12
Miracle_B18
Miracle_B19
MisterX_B03
MisterX_B06
MisterX_B08
MisterX_B10
Rodeo_B17
Rodeo_B20
Welton_B01
Welton_B04
Welton_B07
Welton_B11
Boldin_B28
Boldin_B35
Jerter_B31
Joris_B25
Joris_B30
Joris_B34
Lukard_B24
Lukard_B27
Lukard_B32
Lukas_B26
Lukas_B33
Rossen_B29
Boldin_B41
Boldin_B45
Jerter_B37
Jerter_B47
Joris_B42
Lukas_B43
Rossen_B46
Rossen_B48
Rossen_B49
Sigfrid_B36
Sigfrid_B40
Sigfrid_B44

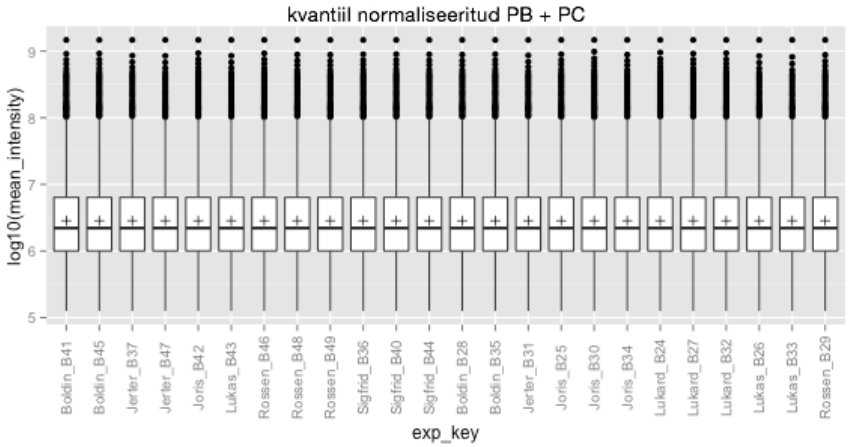
C)


kvantiil normaliseeritud PA + PB + PC top3
log10(Mean)
exp_key
Ciro_B21
Delgado_B02
Delgado_B05
Miracle_B12
Miracle_B18
Miracle_B19
MisterX_B03
MisterX_B06
MisterX_B08
MisterX_B10
Rodeo_B17
Rodeo_B20
Welton_B01
Welton_B04
Welton_B07
Welton_B11
Boldin_B28
Boldin_B35
Jerter_B31
Joris_B25
Joris_B30
Joris_B34
Lukard_B24
Lukard_B27
Lukard_B32
Lukas_B26
Lukas_B33
Rossen_B29
Boldin_B41
Boldin_B45
Jerter_B37
Jerter_B47
Joris_B42
Lukas_B43
Rossen_B46
Rossen_B48
Rossen_B49
Sigfrid_B36
Sigfrid_B40
Sigfrid_B44

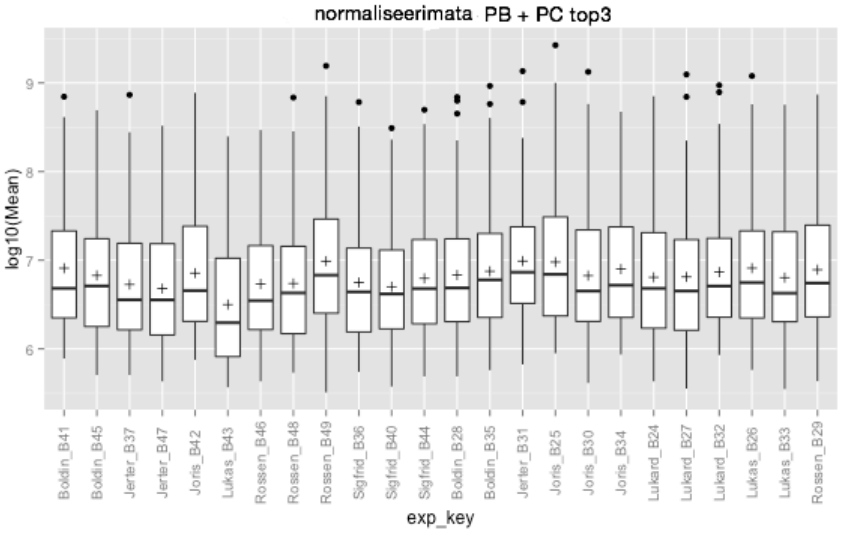
D)


normaliseerimata PB + PC
log10(mean_intensity)
exp_key
Boldin_B41
Boldin_B45
Jerter_B37
Jerter_B47
Joris_B42
Lukas_B43
Rossen_B46
Rossen_B48
Rossen_B49
Sigfrid_B36
Sigfrid_B40
Sigfrid_B44
Boldin_B28
Boldin_B35
Jerter_B31
Joris_B25
Joris_B30
Joris_B34
Lukard_B24
Lukard_B27
Lukard_B32
Lukas_B26
Lukas_B33
Rossen_B29

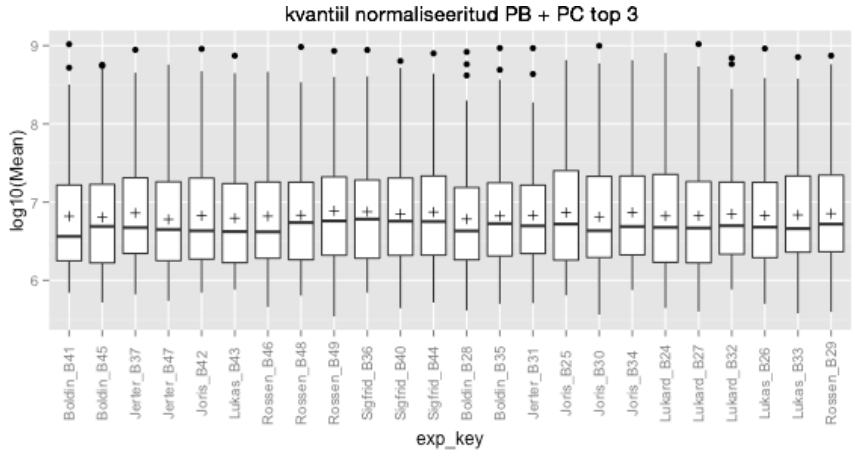
E)



F)



G)



H)

Joonis 10. A) Normaliseerimata katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel. **B)** Normaliseeritud katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel. **C)** Normaliseerimata katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel, mis on kokku pandud kolmest suurima intensiivsusega peptiidist. **D)** Normaliseeritud katse PA + PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel, mis on kokku pandud kolmest suurima intensiivsusega peptiidist. **E)** Normaliseerimata katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel. **F)** Normaliseerimata katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus peptiidi tasemel. **G)**

Normaliseerimata katse PB + PC keskmiste intensiivsustelogaritmitud väärtuste jaotus valgu tasemel, mis on kokku pandud kolmest suurima intensiivsusega peptiidist.**H)** Normaliseeritud katse PB + PC keskmiste intensiivsuste logaritmitud väärtuste jaotus valgu tasemel, mis on kokku pandud kolmest suurima intensiivsusega peptiidist.

Kvantiil normaliseerimisel on suurem mõju suurema varieeruvusega katsete varieeruvuse normaliseerimisel (joonis 10). Normaliseerimine keskmistab katsete peptiide jaotust, et neist saaks edasi valke kokku panna. Kolme tugevama intensiivsusega peptiidist valkude kokkupanek omab jällegi suuremat mõju kui on rakendatud suurema varieeruvusega katsete peal, sest eelnevalt PA + PB + PC katsest PA välja viskamine vähendab valgu tasemel varieeruvust oluliselt.

Tabelis 2 on näidatud, et kui võrrelda omavahel normaliseeritud ning normaliseerimata andmeid, varieeruvuse ei vähene. Põhjus on selles, et filtreerimine kolme suurima intensiivsusega peptiidist selekteerib välja peptiidid, mis on väiksema varieeruvusega. Esialgse masina andmete töötlemise programmiga (Max Quant) on katse PA + PB + PC mediaan CV-st 1.18 ning juba PA katse eraldamine vähendab katse mediaan CV-d 0.42-ni.

Max Quanti andmete kvantiil normaliseerimine ning filtreerimine kolme kõige suurema intensiivsusega peptiidi järgi osutub kasulikuks, sest saab valgu tasemel varieeruvust väiksemaks ning tõenäoliselt pseudoefektidest lahti.

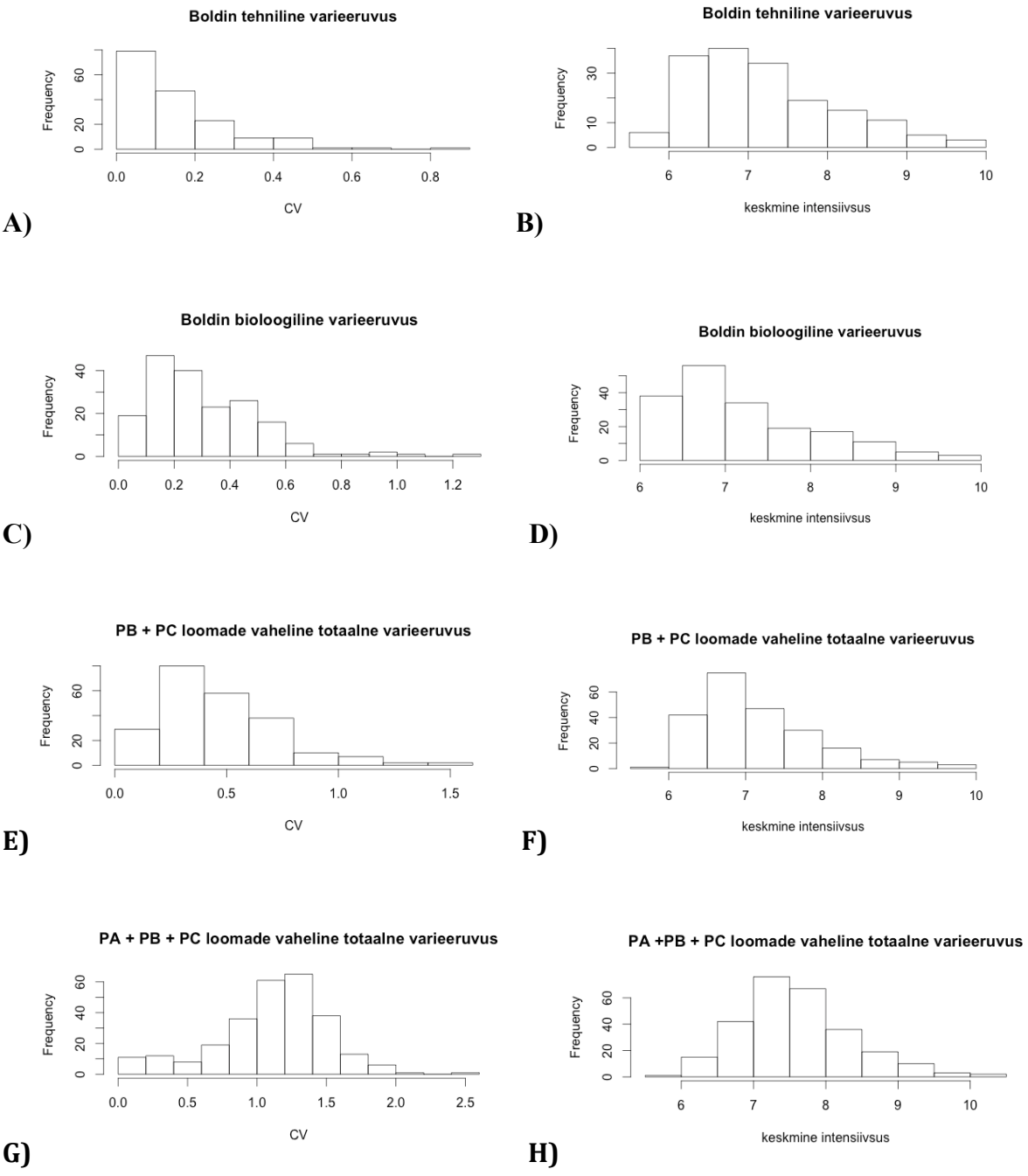
Tabel 2. CV mediaanist erinevate andmetöötluste puhul.

Üle loomade	CV mediaan
Kvantiil normaliseeritud PA + PB+ PC	0.48
Normaliseerimata PA+ PB+ PC	0.47
Kvantiil normaliseeritud PB + PC	0.44
Normaliseerimata PB + PC	0.44
MQ PA + PB + PC	1.18
MQ PB + PC	0.42

2.3.3. Bioloogilised ja tehnilised replikaadid

Ühe looma jaoks leitakse ühe ajapunkti kahe tehnilise replikaadi ühendamisel intensiivsuste CV jaotus ning logaritmitud intensiivsuste jaotus. Seejärel ühendatakse bioloogilised replikaadid ning leitakse CV jaotused ning logaritmitud intensiivsuste jaotus. Bioloogilised ning tehnilised replikaadid leitakse MQ andmetest, mida pole ei normaliseeritud ega filtreeritud.

Joonisel 11 (A-D) on näitena toodud ühe looma (Boldin) tehnilise ja bioloogilise varieeruvuse histogrammid. Tehnilised ja bioloogiliste replikaatide histogrammid on tehtud katsete PB + PC kõikide loomade jaoks ning neid võib leida lisa 1-st. Katse PA, PB ja PC loomade totaalne varieeruvus on laiema ulatusega kui katse PB ja PC totaalne varieeruvus. Joonisel 11 (E-H) on välja toodud totaalne varieeruvus nii PA + PB + PC katses kui ka PB + PC katses. Nagu juba eelnevalt eeldada võib, on PA + PB + PC suurema varieerumusega.



Joonis 11. A) Loom Boldin ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Boldin ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus. **C)** Loom Boldin bioloogilise varieeruvuse neljast ajapunktist CV jaotus. **D)** Loom Boldin bioloogilise replikaadi neljast ajapunktist logaritmitud intensiivsuste jaotus. **E)** PB +PC bioloogilise replikaadi ajapunktidest

CV jaotus. **F)** PB + PC bioloogilise replikaadi ajapunktidest logaritmitud intensiivsuste jaotus. **G)** PA + PB +PC bioloogilise replikaadi ajapunktidest CV jaotus. **H)** PA + PB + PC bioloogilise replikaadi ajapunktidest logartimitud intensiivsuste jaotus.

Tabel 3. Normaliseerimata tehniliste replikaatide CV keskmine ning mediaan. Replikaadid tehtud ainult PB ja PC loomadest.

Tehnilised replikaadid	CV keskmine	CV mediaan
Boldin	0.15	0.11
Joris	0.16	0.12
Lukas	0.17	0.13
Rossen	0.22	0.15
Jerter	0.15	0.10
Sigfrid	0.18	0.10
Lukard	0.18	0.14

Tabel 4. Normaliseerimata bioloogiliste replikaatide CV keskmine ning mediaan. Replikaadid tehtud PB ja PC loomadest

Bioloogilised replikaadid loomasiseselt	CV keskmine	CV mediaan
Boldin	0.30	0.26
Joris	0.30	0.26
Lukas	0.46	0.48
Rossen	0.43	0.37
Jerter	0.44	0.43
Sigfrid	0.24	0.22
Lukard	0.24	0.23

Tabel 5. Normaliseerimata totaalsete varieeruvuse CV keskmine ning mediaan.

Totaalne varieeruvus	CV keskmine	CV mediaan
MQ PB + PC	0.46	0.42
MQ PA + PB + PC	1.12	1.18
Top 3 PA + PB + PC	0.54	0.47
Top 3 PB + PC	0.51	0.44

Tabelis 3 ja 4 on toodud iga looma kohta tehniliste ning bioloogiliste replikaatide keskmine ja mediaan CV-st. Tabelis on näidatud, kuidas loomad üksteisest nii bioloogilise kui ka tehnilise varieeruvuse poolest erinevad (katse PB ja PC loomade bioloogiline ning tehniline varieeruvus on toodud lisas 1). Rossen on tehnilisest replikaatidest kõige suurema varieeruvusega. Bioloogilistest replikaatidest on kõige suurema varieeruvusega Lukas. Tehniliste replikaatide suurusjärgud jäävad iga looma puhul samaks, kuid bioloogilistel replikaatidel on suurusjärgud veidi erinevad. Tabelis 5 on näidatud totaalset varieeruvust loomade vahel nii PB + PC kui ka PA + PB + PC katsele. PA välja jätmine vähendab katse

varieeruvust 1.18 pealt 0.42ni. Tehnilised replikaadid on umbes kaks korda väiksemad kui loomade vaheline totaalne varieeruvus. Osadel loomadel (Jerter, Rossen, Lukas) on bioloogiline varieeruvus sama suur kui loomade vaheline totaalne varieeruvus. Loomade bioloogiline varieeruvus erineb mõningatel juhtudel üksteisest 1/3 võrra.

Tabelil 5 on näidatud, et esialgsete MQ andmete analüüsimisel vähendab varieeruvust oluliselt PA katse välja jätmine. Kvantiil normaliseerimine seda varieeruvust veel omakorda ei vähenda, sest katse PA välja viskamisel on PB + PC valgud sarnase varieeruvusega.

2.3.4. Võimsusanalüüs

Tabel 6. Võimsusanalüüs erinevalt töödeldud andmetega (kahendlogaritm skaalas).

	2x efekt N=2	2x efekt N=3	3x efekt N=2	3x efekt N= 3
Normaliseerimata peptiidi tasemel PA + PB + PC	0.12	0.17	0.18	0.27
Kvantiil normaliseeritud peptiidi tasemel PA + PB + PC	0.22	0.35	0.34	0.58
Normaliseerimata valgu tasemel PA + PB + PC	0.53	0.83	0.78	0.98
Kvantiil normaliseeritud valgu tasemel PA + PB + PC	0.49	0.79	0.74	0.97
Normaliseerimata peptiidi tasemel PB + PC	0.26	0.43	0.42	0.69
Kvantiil normaliseeritud peptiidi tasemel PB + PC	0.30	0.5	0.48	0.78
Normaliseerimata valgu tasemel PB + PC	0.73	0.97	0.93	1
Kvantiil normaliseeritud valgu tasemel PB + PC	0.71	0.96	0.93	1
MQ otse masinast PA+ PB + PC	0.11	0.14	0.14	0.21
MQ otse masinast PB + PC	0.29	0.48	0.46	0.76

Tabelis 6 on näidatud, et kahekordse efekti saamiseks (kui valimisuurus on kaks) suurendab peptiidi tasemel katse PA välja jätmine katsest PA + PB + PC võimsust 0.12 pealt 0.26 peale, kvantiil normaliseerimine katse PA + PB + PC võimsust 0.12 pealt 0.22 peale, kvantiil normaliseeritud PA + PB + PC katsest katse PA välja jätmine võimsust 0.22 pealt 0.3 peale.

Kolmekordse efekti saamiseks (kui valimisuurus on kolm) suurendab kvantiil normaliseeritud PA + PB + PC katsest PA välja jätmine powerit 0.58 pealt 0.78 peale.

Kahekordse efekti puhul (kui valimisuurus on kaks) suurendab valgu tasemel PA + PB + PC katsest PA välja jätmine võimsust 0.53 pealt 0.73 peale. Kolmekordse efekti puhul (kui valimisuurus on kolm) on valgu tasemel normaliseerimata PA + PB + PC võimsus 0.98 ning valgu tasemel normaliseeritud PB + PC võimsus 1.

Kvantiil normaliseerimine suurendab katse võimsust peptiidi tasemel, kuid vähendab veidi katse võimsust valkude tasemel, sest valgud on juba kokku pandud kolmest suurima intensiivsusega peptiidist ning seega on juba selekteeritud välja peptiidid, mis on väiksema varieeruvusega.

Otse masinast tulnud proovide puhul (MQ) on katse PA + PB + PC võimsus kahekordse efekti saamiseks (kui valim on kaks) 0.11 ning katse PB + PC puhul 0.29. Kolmekordse efekti saamiseks (kui valim on kolm) on PA + PB + PC võimsus 0,.21 ning katse PB + PC võimsus 0.76.

Katse PA eraldi analüüsimine katsest PB ja PC tõstab katse võimsust igal juhul. Kvantiil-normaliseerimine tõstab katse võimsust peptiidi tasemel.

2.3.5. Arutelu

Enne analüüsi läbi viimist saab disainida bioloogiliste ning tehniliste replikaatide kindla suhte, et saavutada katses vajalik võimsus. Võimalik on ka võimsusanalüüsi teostamine kui katse on juba läbi viidud, et selekteerida välja optimaalne analüütiline meetod, millel puhul varieeruvus on minimeeritud ning võimsus võimalikult suur. Katse tehniline varieeruvus võib olla põhjustatud katse masinas analüüsimise järjekorrast, sest masina enda kalibratsioon aja jooksul muutub. Tehniline varieeruvus võib põhjustatud olla ka süstemaatilisest veast, mida saab minimeerida andmeanalüüsiga. Näiteks kvantiil normaliseerimine muudab erinevate katsete intensiivsuste jaotused peptiidi tasemel ühtlaseks, sest pole põhjust uskuda, et antud töös uuritavad katsed üksteisest intensiivsuste poolest märgatavalt erineksid. Andmeanalüüs aitab vabaneda ka statistiliselt juhuslikust veast. Minimeerides tehnilist varieeruvust tuleb bioloogiline varieeruvus rohkem esile ning reaalselt olulisi efekte leiab suurema tõenäosusega.

Antud töö põhjal saab järeldada, et andmete esmane analüüs vähendab nii katsete varieeruvuste jaotust kui ka tõstab katse võimsust. Väheneb tõenäosus pseudoefektide leidmiseks. Andmete analüüs peptiidide tasemel näitab katsesisest ning -vahelist varieeruvust, mis võimaldab otsustada milliseid katseid saab paralleelselt analüüsida. Võimsusanalüüsi eesmärgiks Peptiidide tasemel normaliseerimine tõstab katse võimsust. Valgu tasemel osutub kasulikuks kolmest suurima intensiivsusega peptiidist valkude kokkupanek, mis tõstab nii võimsust kui vähendab varieeruvust. Bioloogiliste ja tehniliste replikaatide kirjeldamine aitab eristada kui suur osa totaalsel varieeruvusel on põhjustatud tehnilistest replikaatidest ning kui suurelt loomad omavahel bioloogiliselt varieeruvuselt erinevad. Katsetest leiti, et tehniline varieeruvus on umbes pool loomade vahelisest totaalsest varieeruvusest. Bioloogiline varieeruvus varieerub üksikute loomade vahel kuni 1/3 väärtusest.

Andmetega edasi töötamisel on võimalik leida uuritava efekti suurus ning omakorda teha sellest juba konkreetseid järeldusi. Samuti saab andmeid kasutades modelleerida mudeli, et hinnata efektide ülehindamise tõenäosust. Mudel oleks konkreetse eksperimendi kvaliteedikontroll ning annaks võimaluse otsustada, millist võimsust tahame saavutada.

Lisa 2-s on välja toodud koodid, millega analüüse teostati.

KOKKUVÕTE

Proteoomika tehnoloogiate areng on bioteaduste vallas tähtis. Valkude eraldamine ning identifitseerimine on nüüd lihtsalt ning kiirelt tehtav suure läbilaskevõimega mass-spektromeetria masinas. Siiski pole masinast saadud andmetega enamasti läbi viidud edasist andmeanalüüsi, mis kontrolliks andmete varieeruvust ning võimsust. See on mõjutanud omakorda reaalsete efektide leidmist ning katse võimsust. Katsed mõõdavad efekti suurust ning püüavad ennustada, kas efekt on reaalne või seletatav juhusliku hälbega nullefektist.

Käesoleva töö tegevuskavandi eesmärgiks on võrrelda erinevalt töödeldud andmeid, et selekteerida välja optimaalne analüütiline meetod, mis töötaks andmete peal kõige paremini ehk vähendaks katsete varieeruvust ning tõstaks katse võimsust, et teada saada, kas efekt on reaalne. Lisaks on kirjeldatud bioloogilist, tehnilist ja totaalset varieeruvust. Koostati optimaalne tegevuskava andmete esmaseks analüüsiks.

Jõuti järeldusele, et üks plasma katse on teistest plasma katsetest oluliselt erineva varieeruvusega ning selle katse välja jätmine/eraldi analüüsimine parandas võimsust nii peptiidi kui ka valgutasemel. Peptiidi tasemel vähendas varieeruvust kvantiil normaliseerimine. Valgu tasemel suurendas võimsust valgu kokkupanek kolmest tugevaima intensiivsusega peptiidist. Leiti tehniline varieeruvus, mis on umbes pool bioloogilisest varieeruvusest. Bioloogiline varieeruvus (mis sisaldab endas ka tehnilist varieeruvust) on osadel loomadel umbes pool loomade vahelisest totaalsest varieeruvusest ja teistel jällegi umbes sama suur kui loomade vaheline totaalne varieeruvus.

SUMMARY

Variability, power and quality in quantitative proteomics

Margot Saare

Proteomics is a relatively new research area. Proteins have a large impact in our lives and that is why scientists are increasingly identifying and separating proteins. Proteins provide us knowledge about cellular processes. Sensitive, high-throughput and soft ionization mass spectrometry techniques brought a change in proteomics.

Exploratory data analysis and quality control is essential when analysing data from mass spectrometry. All analyses are measuring the size of the effect and try to predict whether the effect is real or a result of a random offset.

Exploratory data analysis gives us a better overview of our data and helps to minimize the variability of test to find real effects. Power analysis is used to test the power of the test, the size of the effect and the required sample size. Power also acts as a quality control for the data set and gives us an objective validation of variability.

The aim of this thesis was to create a workflow to analyse data from mass spectrometry. Several workflows were compared in this study (analysing in peptide or protein level, quantile normalisation, putting proteins together from three peptides that have the highest intensities, Max Quant) in order to find an optimal workflow that minimizes the variability in tests and between the tests so they can be compared with each other, and to find which workflow have the higest power. The variability of biological and technical replicates were also described in this thesis. Power analysis was used to measure the quality of a dataframe.

KIRJANDUSE LOETELU

Abdi, H., 2010. Coefficient of variation. *Encylopedia of Research Design*, lk1–5.

Aebersold, R. & Mann, M., 2003. Mass spectrometry-based proteomics. *Nature*, 422(6928), lk198–207.

Arike, L., 2012. *Quantitative Proteomics of Escherichia coli : From Relative to Absolute Scale*,

Bantscheff, M. et al., 2007. Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, 389(4), lk1017–1031.

Blackstock, W.P. & Weir, M.P., 1999. Proteomics: Quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17(3), lk121–127.

Bolstad, B. & Bolstad, B., 2001. Probe Level Quantile Normalization of High Density Oligonucleotide Array Data. *Cell*, (December), lk1–8.

Button, K.S. et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews. Neuroscience*, 14(5), lk365–76.

Caprioli, R.M., Farmer, T.B. & Gile, J., 1997. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical chemistry*, 69(23), lk4751–4760.

Capriotti, A.L. et al., 2011. Intact protein separation by chromatographic and/or electrophoretic techniques for top-down proteomics. *Journal of Chromatography A*, 1218(49), lk8760–8776.

Chatfield, C., 1986. Exploratory data analysis. *European Journal of Operational Research*, 23(1), lk5–13.

Erdfelder, E., Faul, F. & Buchner, A., 1996. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28(1), lk1–11.

Faul, F. et al., 2007. A Flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, 39(2), lk175–191.

Fox, N. & Mathers, N., 1997. Empowering research: Statistical power in general practice research. *Family Practice*, 14(4), lk324–329.

Gelman, A. & Carlin, J., 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors.

Guerrera, I.C. & Kleiner, O., 2005. Application of mass spectrometry in proteomics. *Bioscience Reports*, 25(1-2), lk71–93.

Ihaka, R. & Gentleman, R., 1996. A language for data analysis and graphics R. *Journal of Computational and Graphical Statistics*, 5(3), lk299–314.

Jemal, M., 2000. High-throughput quantitative bioanalysis by LC/MS/MS. *Biomedical Chromatography*, 14(6), lk422–429.

Karp, N. a. & Lilley, K.S., 2007. Design and analysis issues in quantitative proteomics studies. *Proteomics*, 7 Suppl 1, lk42–50.

Kumar, C. & Mann, M., 2009. Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS letters*, 583(11), lk1703–12.

Leek, J.T. et al., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics*, 11(10), lk733–739.

Levin, Y., 2011. The role of statistical power analysis in quantitative proteomics. *Proteomics*, 11(12), lk2565–2567.

Lewis, J.K., Wei, J. & Siuzdak, G., 2000. Matrix-assisted Laser Desorption / Ionization Mass Spectrometry in Peptide and Protein Analysis. *Encyclopedia of Analytical Chemistry*, lk5880–5894.

Loo, J. a., 2000. Electrospray ionization mass spectrometry: A technology for studying noncovalent macromolecular complexes. *International Journal of Mass Spectrometry*, 200(1-3), lk175–186.

Maindonald, J., 2008. Using R for Data Analysis and Graphics Introduction , Code and Commentary. *Australian Journal of Zoology*, 22(January), lk1–99.

Maiväli, Ü. "Interpreting Biomedical Science". Academic Press, London. 2015, in press.

McKinney, W. “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython”. O’Reilly Media. 2013.

Neilson, K. a. et al., 2011. Less label, more free: Approaches in label-free quantitative mass spectrometry. *Proteomics*, 11(4), lk535–553.

Nelson, D.L. & Cox, M.M., 2005. *Lehninger Principles of Biochemistry*,

Ong, S.-E. et al., 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP*, 1(5), lk376–386.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data.

Santoni, V., Molloy, M. & Rabilloud, T., 2000. Membrane proteins and proteomics: Un amour impossible? *Electrophoresis*, 21(6), lk1054–1070.

Sechi, S. & Oda, Y., 2003. Quantitative proteomics using mass spectrometry. *Current Opinion in Chemical Biology*, 7(1), lk70–77.

Tenson, T. & Ehrenberg, M., 2002. Regulatory Nascent Peptides in the Ribosomal Tunnel. *Cell*, 108(5), lk591–594.

Vaux, D.L., Fidler, F. & Cumming, G., 2012. Replicates and repeats—what is the difference and is it significant? *EMBO reports*, 13(4), lk291–296.

Yates, J.R., 2000. Mass spectrometry from genomics to proteomics. *Outlook*, 16(1), lk5–8.

Yates, J.R., Ruse, C.I. & Nakorchevsky, A., 2009. Proteomics by mass spectrometry: approaches, advances, and applications. *Annual review of biomedical engineering*, 11, lk49–79.

KASUTATUD VEEBIAADRESSID

https://www.python.org

http://www.r-project.org/

http://pandas.pydata.org/

http://www.numpy.org/

http://matplotlib.org/

http://www.scipy.org/

http://user2014.stat.ucla.edu/abstracts/talks/45_Wickham.pdf

http://ggplot2.org

http://cran.r-project.org/web/packages/DataCombine/DataCombine.pdf

http://blog.rstudio.org/2014/07/22/introducing-tidyr/

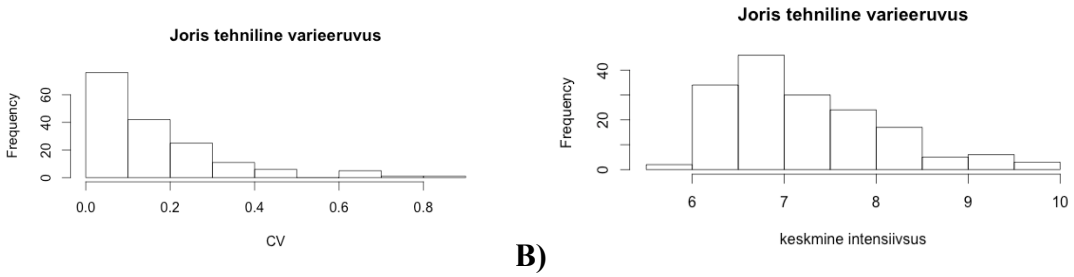
http://cran.r-project.org/web/packages/reshape2/index.html

http://www.gpower.hhu.de/

http://www.molmine.com/magma/global_analysis/batch_effect.html

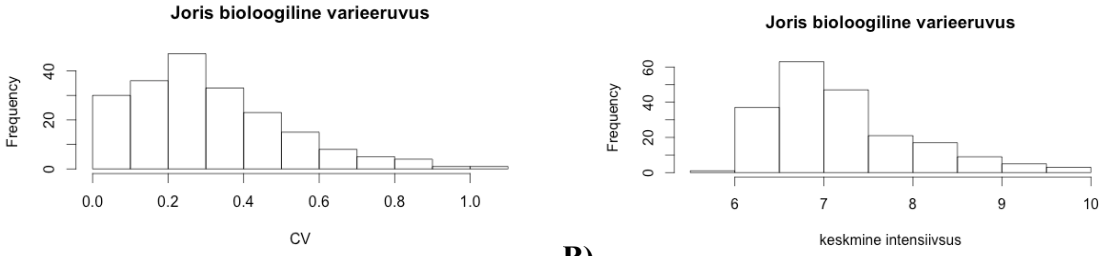
LISA 1.

Joonisel 12-23 on kujutatud katse PB + PC loomade tehnilist varieeruvust (CV ja logaritmitud intensiivsuste jaotus) ning bioloogilist varieeruvust (CV ja logaritmitud intensiivsuste jaotus).



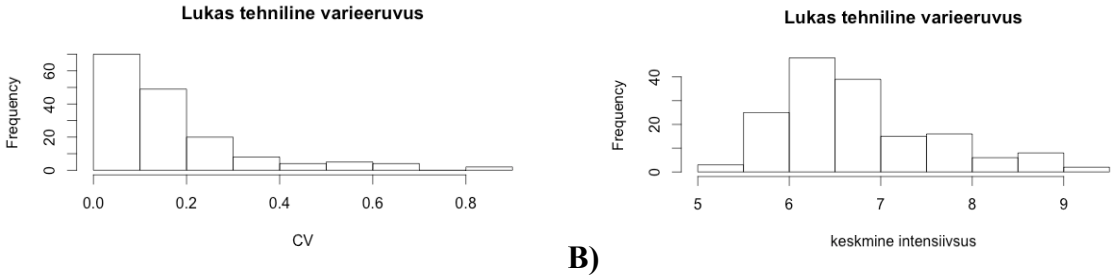
A) **B)**

Joonis 12. A) Loom Joris ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Joris ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



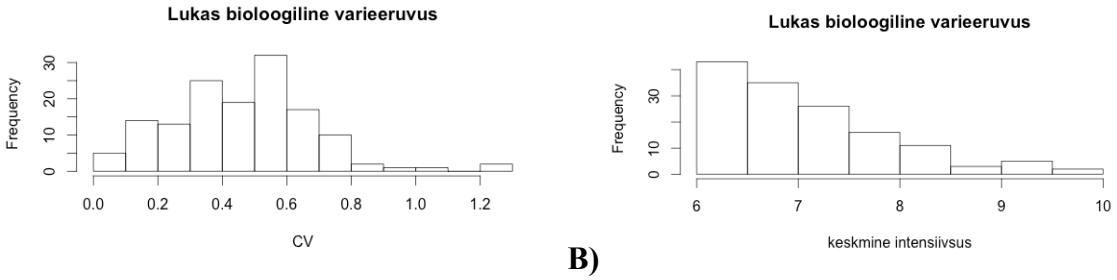
A) **B)**

Joonis 13. A) Loom Joris bioloogilise varieeruvuse neljast ajapunktist CV jaotus. **B)** Loom Joris bioloogilise replikaadi neljast ajapunktist logaritmitud intensiivsuste jaotus.



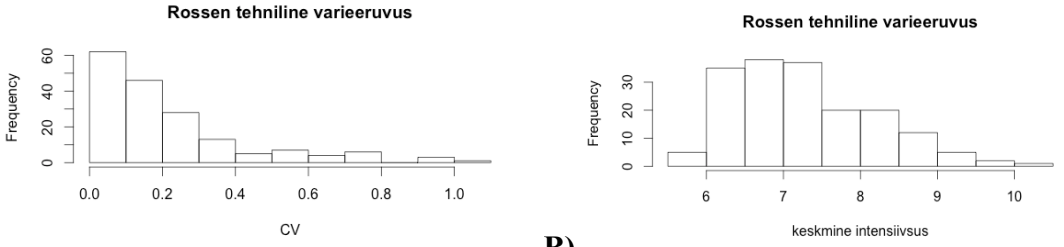
A) **B)**

Joonis 14. A) Loom Lukas ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Lukas ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



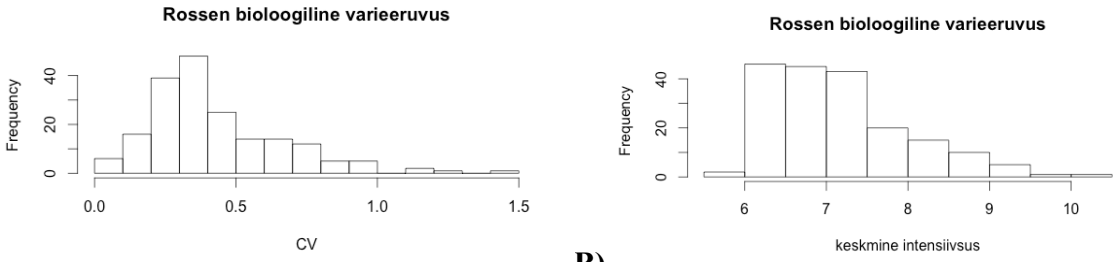
A) **B)**

Joonis 15. A) Loom Lukas bioloogilise varieeruvuse kolmest ajapunktist CV jaotus. **B)** Loom Lukas bioloogilise replikaadi kolmest ajapunktist logaritmitud intensiivsuste jaotus.



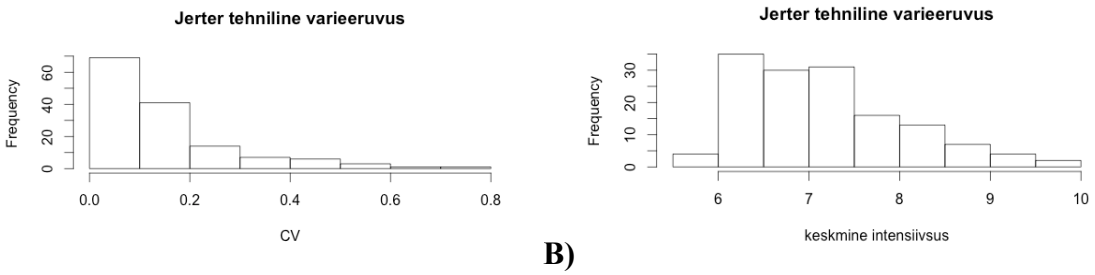
A) **B)**

Joonis 16. A) Loom Rossen ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Rossen ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



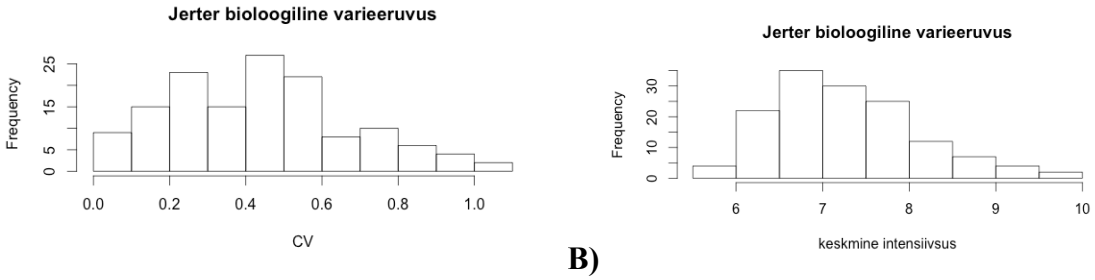
A) **B)**

Joonis 17. A) Loom Rossen bioloogilise varieeruvuse neljast ajapunktist CV jaotus. **B)** Loom Rossen bioloogilise replikaadi neljast ajapunktist logaritmitud intensiivsuste jaotus.



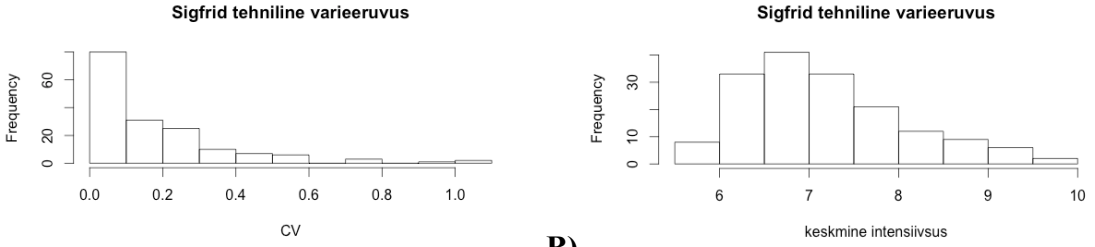
A) **B)**

Joonis 18. A) Loom Jerter ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Jerter ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



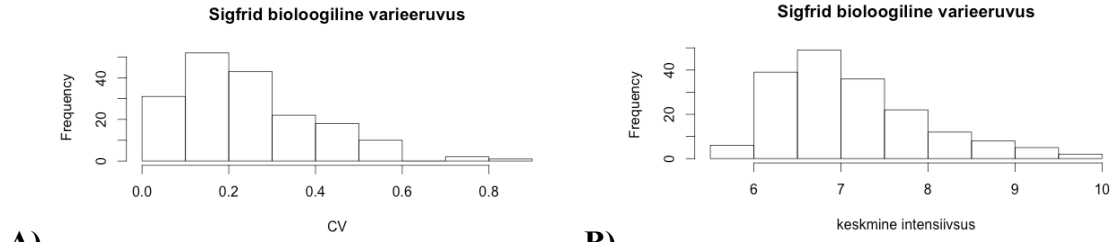
A) **B)**

Joonis 19. A) Loom Jerter bioloogilise varieeruvuse kolmest ajapunktist CV jaotus. **B)** Loom Jerter bioloogilise replikaadi kolmest ajapunktist logaritmitud intensiivsuste jaotus.



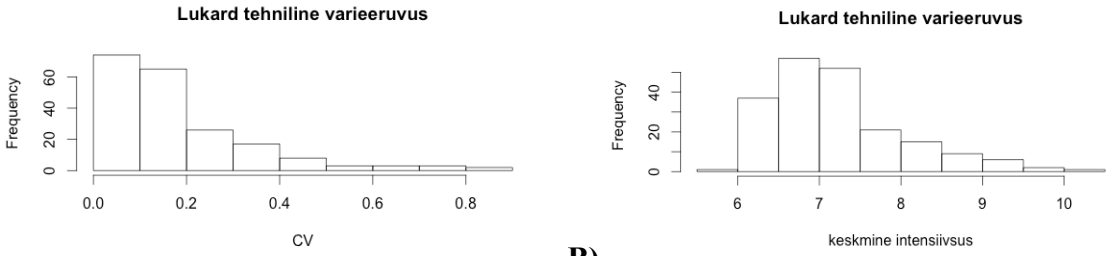
A) **B)**

Joonis 20. A) Loom Sigfrid ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Sigfrid ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



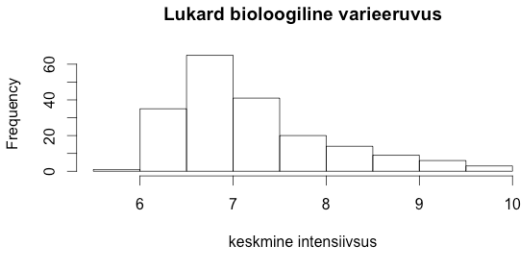
A) **B)**

Joonis 21. A) Loom Sigfrid bioloogilise varieeruvuse kolmest ajapunktist CV jaotus. **B)** Loom Sigfrid bioloogilise replikaadi kolmest ajapunktist logaritmitud intensiivsuste jaotus.

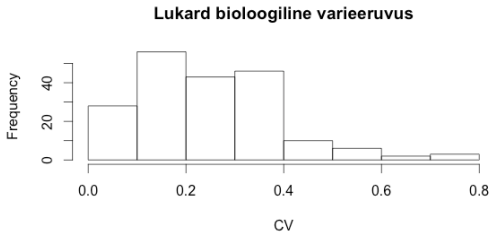


A) **B)**

Joonis 22. A) Loom Lukard ühe ajapunkti kahe replikaadi CV jaotus. **B)** Loom Lukard ühe ajapunkti kahe replikaadi logaritmitud intensiivsuste jaotus.



A)



B)

Joonis 23. A) Loom Lukard bioloogilise varieeruvuse kolmest ajapunktist CV jaotus. **B)** Loom Lukard bioloogilise replikaadi kolmest ajapunktist logaritmitud intensiivsuste jaotus.

LISA 2.

Lisa 2-s on koodid, millega andmete esmane analüüs tehti. Andmete analüüs viidi läbi nii Pythonis kui R-s.

Python

- Peptiidi tase, kus võeti üle kõikide peptiide keskmiste intensiivsuse CV ning keskmise intensiivsuse väärtus. Esialgne *dataframe* puhastati replikaatidest ning eemaldati ebavajalik veerud.

```
import numpy as np
import pandas as pd      #vajalike pakettide installimine
from pandas import read_csv


dfPA = pd.read_csv("PA_raw_data.csv", sep="\t") # andmete sisselugemine

df6PA = dfPA.drop('quality_key', 1)          # veeru 'quality_key' eemaldamine
df7PA = df6PA.drop_duplicates()              #duplikaatidest puhastamine
df8PA = df7PA.drop('protein_key', 1)       # veeru 'protein_key' eemaldamine
df8PA['exp_key'] = df8PA['exp_key'].map(lambda x: str(x)[:-4]) # looma ajapunkti numbri
eemaldamine

groupedPA= df8PA.groupby(['peptide_key'], as_index=False)  # grupeerimine peptiidi järgi

ffPA=groupedPA.agg([np.sum, np.mean, np.std], as_index=False)
ffPA.head() # peptiidi järgi keskmisele intensiivsusele standardhälbe ja keskmise
intensiivsuse arvutamine

df5PA=(ffPA.mean_intensity['std'] / ffPA.mean_intensity['mean'])
df5PA.head() # CV arvutamine

df5PBmean= np.log(ffPB.mean_intensity['mean']) #keskmise intensiivsuste väärtuste
logaritmimine

%matplotlib inline
import matplotlib.pyplot as plt   # histogrammi tegemine sõltuvalt (CV) sagedusest
fig, ax = plt.subplots()
plt.hist(df5PA, bins=30)
ax.set_xlabel('keskmine intensiivsus')
ax.set_ylabel('sagedus')
ax.set_title('CA + CB valgud')
plt.show()
```

- Valgu tase, kus valgud pandi kokku kolme kõige suurema intensiivsusega peptiidist.

```
dfPAnoqnt = pd.read_csv("PA_raw_data.csv", sep="\t")
df6PAnoqnt = dfPAnoqnt.drop('quality_key', 1)
df6PAnoqnt = df6PAnoqnt.drop_duplicates()
df7PAnoqnt = df6PAnoqnt.groupby(['protein_key','exp_key'],as_index=False) #grupeerimine valgu ja looma ajahetke järgi

df7PAnoqnt=df7PAnoqnt['mean_intensity'].apply(lambda x: x.nlargest(3)) #kolme suurima intensiivsuse selekteerimine
df7PAnoqnt=df7PAnoqnt.reset_index()
df7PAnoqnt=df7PAnoqnt.groupby('level_0').filter(lambda x: x['level_1'].nunique() > 2)
# valkude valimine, millel  on vähemalt 3 unikaalset peptiidi

df1PAnoqnt=df7PAnoqnt.groupby(['level_0'],as_index=False).mean() #keskmisest intensiivsusest keskmise arvutamine
df2PAnoqnt=df7PAnoqnt.groupby(['level_0'],as_index=False).std() #keskmisest intensiivsusest standardhälbe arvutamine

df5PAnoqnt=df2PAnoqnt/df1Panoqnt  # CV arvutamine
df5PAnoqnt=df5PAnoqnt.dropna()
dfPAnoqnt=df5PAnoqnt[0]   # CV väärtused
dfPAnoqntmean=df1PAnoqnt[0]  # keskmiste intensiivsuste väärtused
```

R

- R-is tehti *dataframe*-ide kvantiil normaliseerimine ning karp-vurrud diagrammid.

```
qnpb <-pb %>%                  # andmefaili PB normaliseerimine
  select(peptide_key,exp_key,mean_intensity) %>%
  dcast(peptide_key~exp_key) %>%
{peptide_key <- use_series(.,"peptide_key")
 cln <- select(.,-peptide_key) %>% colnames
 select(.,-peptide_key) %>% as.matrix %>%
  normalize.quantiles %>% data.frame %>%
  set_colnames(cln) %>% cbind(peptide_key,.)} %>%
  melt(value.name = "mean_intensity",variable.name = "exp_key") %>%
  filter(complete.cases(.)) %>% {
    protein_key <- pb %>% select(protein_key,peptide_key,exp_key)
    merge(.,protein_key)
  }

qnpb %>%       #normaliseeritud andmefaili PB andmetest karp-vurrud diagrammi tegemine
 ggplot(aes(x=exp_key,y=log10(Mean))) +
 geom_boxplot() +
 stat_summary(fun.y=mean,geom="point",shape=3) +
 theme(axis.text.x=element_text(angle = 90, vjust = 0.5))
```

- Bioloogiliste ja tehniliste replikaatide leidmine

```
Lukard_tehniline_varieeruvus<-full_join(B32A, B32B)%>%na.omit%>%
```

```
  transmute(prot, int_b32=(int_b32a+int_b32b)/2,
          relative_aveDev=((abs(int_b32a - int_b32) + abs(int_b32b - int_b32))/2)/int_b32,
          CV=sqrt((int_b32a - int_b32)**2+(int_b32b - int_b32)**2/2)/int_b32) # tehnilise
varieeruvuse leidmine ühendades ühe ajapunkti kaks replikaati

hist(Lukard_tehniline_varieeruvus$CV,xlab= "CV", main ="Lukard tehniline varieeruvus")
hist(log10(Lukard_tehniline_varieeruvus$int), xlab= "keskmine intensiivsus", main ="Lukard
tehniline varieeruvus")   # histogrammide tegemine

Luk1.2<-full_join(B24P, B27P)    #ajapunktide ühendamine
Lukard_full_Pl <- full_join(Luk1.2, B32P)
Lukard_bioloogiline_varieeruvus<-Lukard_full_Pl%>%group_by(prot)%>%
  mutate(int=mean(c(int_b24, int_b27, int_b32), na.rm=T),
         CV=sd(c(int_b24, int_b27, int_b32), na.rm=T)/int)%>%filter(!is.na(CV))  # 3
bioloogilise replikaadi ühendamine, et leida bioloogiline varieeruvus (mis sisaldab siin endas
ka tehnilist varieeruvust)
```

LIHTLITSENTS LÕPUTÖÖ ELEKTROONILISEKS AVALDAMISEKS

Mina Margot Saare

(sünnikuupäev: 23. 02. 1991)

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

   Varieeruvus, võimsus ja kvaliteet kvantitatiivses proteoomikas,

   mille juhendaja on Ülo Maiväli ja David William Schryer

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, 25.05.2015