TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOTEHNOLOOGIA ÕPPETOOL

Mario Reiman

RNA INTAKTSUS JA RNA-SEQ EKSPRESSIOONIANDMED

Magistritöö

Juhendaja Siim Sõber, PhD

TARTU 2014

SISUKORD

KASUTATUD LÜHENDID

miRNA – mikroRNA (*microRNA*)

mtRNA – mitokondriaalne RNA (*mitochondrial RNA*)

NPM - Nextera PCR segu (*Nextera PCR mixture*)

PPM – PCR praimerite segu Nextera kitist (Illumina) (*PCR praimer mixture*)

PTC – enneaegne terminatsioonikoodon (*premature termination codon*)

read – sekveneeritud osa raamatukogu fragmendist, tõlgin „lugem"-iks

RIN – RNA intaktsusnumber (*RNA integrity number*)

RNA-seq – Kogu transkriptoomi süvasekveneerimine (*whole transcriptome deep sequencing*)

RQI – RNA kvaliteedi indeks, konkureeriva firma vaste RIN-ile (*RNA quality index*)

RT-qPCR – kvantitatiivne pöördtranskriptaas-PCR (*real-time reverse transcription polymerase chain reaction)*

TD – tagmentatsiooni puhver Nextera kitist (Illumina) (*tagment DNA buffer*)

ddNTP – didesoksünukleotiid (*Dideoxynucleotide*)

PABP – Polüadenüleeritudud järjestusi siduv proteiin (*Poly(A)-binding protein*)

TDE – tagmentatsiooni ensüümi reagent Nextera kitist (Illumina) (*tagment DNA enzyme*)

siRNA – väike segav RNA (*small interfering RNA*)

Q-Q plot – kvantiil-kvantiil plot (*quantile-quantile plot*)

SNV – ühenukleotiidiline muutus (*single nucleotide variant*)

UTR – mittetransleeriv ala transkripti ääres (*untranslated region*)

NMD – enneagsest stoppkoodonist põhjustatud degradatsioon (*nonsense mediated decay*)

l/nt – lugemit transkripti nukleotiidi kohta

lncRNA – pikk mittekodeeriv RNA (*long non-coding RNA*)

SISSEJUHATUS

RNA lagundamine on koes aktiivselt reguleeritav ja varieeruv protsess. Lagundamise kiiruse varieeruvus eri transkriptide ja raku seisundite suhtes tingib selle, et osadel transkriptidel väheneb tervete transkriptide hulk kiiremini kui teistel (Yang jt, 2003). Mõõtes RNA tasemeid osaliselt lagunenud RNA pealt meetodiga, mis eeldab, et proovide ettevalmistamisel alustati tervete RNA-dega, võib geenide ekspressiooni tasemete määramisel tekkida süstemaatiline viga. RNA-seq ehk transkriptoomi süvasekveneerimise andmeanalüüsis transkriptide ja geenide koguse määramisel eeldatakse aga justnimelt seda. Seetõttu on soovitatav kasutada üksnes hea kvaliteediga ja võimalikult vähelagunenud RNA-d. See ei ole aga haruldaste kliiniliste proovide korral alati võimalik, mistõttu võib sääraste RNA-de sekveneerimisel saada aga valesid ekspressiooniväärtusi (Wan ja Yan, 2012).

Bakalaureuseõppes osalesin meie laboris läbi viidud RNA-seq projektis, kus kasutatud RNA oli kuni 8 aastat vana ja mõningal määral lagunenud (keskmine RIN=7,0 (RNA intaktsus number)). Et need proovid olid aga bioloogilises mõttes väga väärtuslikud, valisime RNA intaktsuse suhtes robustse raamatukogu valmistamise meetodi ja sekveneerisime, saades tegelikult suhteliselt hea kvaliteediga lugemid.

Käesolevas töös vaatan ma täpsemalt, kas RNA intaktsuse (mõõdetud RIN väärtusena) ja geenieksрessiooni vahel on kvantifitseeritav seos, missuguste omadustega geenide juures see seos tugevaim on ja kas seda saab ka välja korrigeerida.

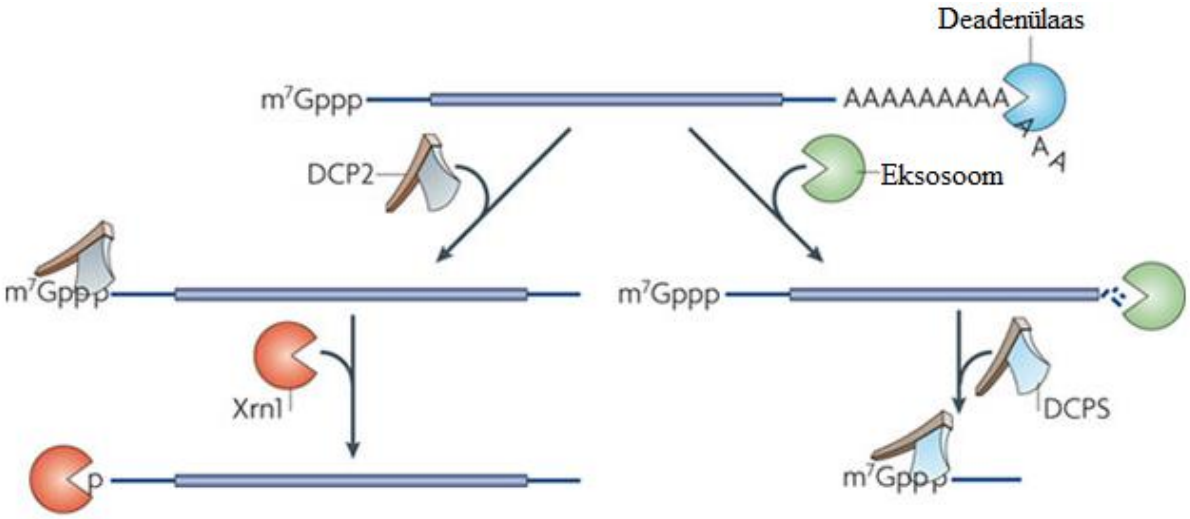
1. KIRJANDUSE ÜLEVAADE

1.1 RNA ja seda lagundavad ensüümid ja protsessid

Rakkudes toimuva transkriptsiooni mõjul tekib sinna suurtes kogustes RNA-d, mis omakorda osaleb translatsioonis ja mõjutab rakku. RNA töötluses osalevad spetsialiseerunud ensüümid, ribonukleaasid ehk RNaasid. RNaasidel on mitmeid bioloogilisi funktsioone, näiteks geenide ekspressiooni regulatsioon, toitainelise RNA seedimine, patogeenide vastane aktiivsus (Sorrentino jt, 2010). Ribonukleaasid on fosforolüütilised ja hüdrolüütilised ensüümid, mis täidavad neid funktsioone lõigates ja lagundades RNA-d. Rakus olevate ribonukleaaside erinev kogus ja spetsiifilisus tingib RNA molekulide äärmiselt erineva eluaea. Näiteks kiiresti jagunevates bakterites võib RNA kestvus varieeruda kümnetest sekunditest kuni tunnini või kõrgemate eukarüootide puhul minutitest nädalateni (Fordyce jt, 2013, Stuart jt, 1999, Reiman, 2013).

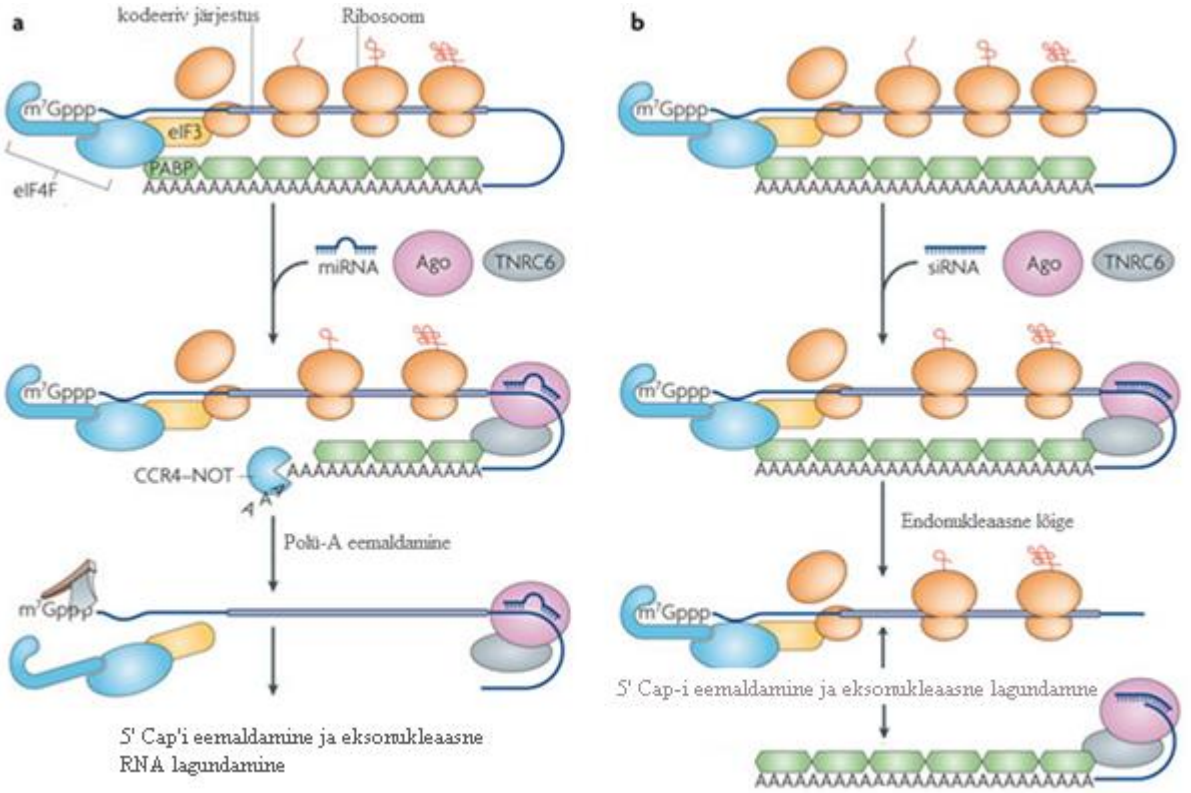
RNA uuringute läbiviimiseks on tihtipeale vajalik koeproove või eraldatud RNA-d pikaaegselt säilitada. Enamasti hoiustatakse proove -80 ^{o}C juures (Riesgo jt, 2012). Koeproovide säilitamise muudab aga keerukaks see, et RNaasid on äärmiselt vastupidavad ja RNA molekulide nukleotiidide vaheliste sidemete hüdrolüüs on termodünaamiliselt soodustatud (fosforolüüs on energeetiliselt neutraalne) (Houseley ja Tollervey, 2009). Kuna täpsed RNA lagundamise mehhanismid *post partum* inimese platsentas ja inimkudedes üldisemaltki ei ole veel selged, kirjeldan ma seni teada olevaid radu ja mehhanisme. Kuigi enamus neist on koostatud elus rakkude põhjal, võivad paljud rajad ka surnud kudedes sarnased olla (Reiman, 2013).

Eukarüootide mRNA lagundamise standardmudel koostati *S. cerevisiae* ja imetajate rakkude põhjal neist leitud ensüümide homoloogide abiga. See kaasab endas nii 3’ kui ka 5’ eksonukleaase, endonukleaase, deadenülaase ja *cap*-struktuuri eemaldus ensüüme. mRNA-de lagunemine algab enamasti 3’ polü-A struktuuri eemaldamisega. See kutsub esile ka 5’ *cap*-struktuuri lagundamise, ning sealt edasi lagundatakse otstest kaitsmata mRNA molekul kiiresti 5’ või 3’ eksonukleaaside poolt (nt: XRN1 või eksosoom (Belasco, 2010), Reiman, 2013) (joonis 1).

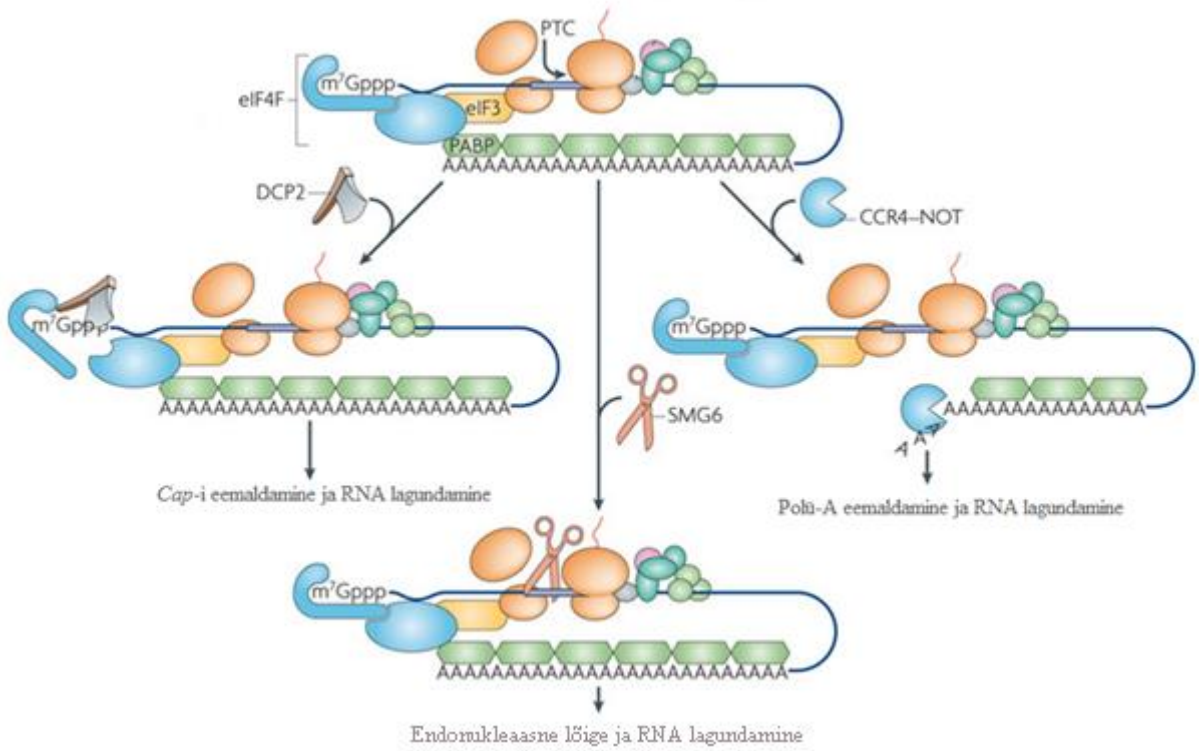


Joonis 1. Traditsiooniline arusaamine mRNA lagundamisest eukarüootide rakkudes. Muudetud Belasco (2010) põhjal.

Ajalooliselt on peetud mRNA lagundamise erinevuseks prokarüootides ja eukarüootides lagundamise alustus protsessi. Nimelt bakterites alustakse seda enamasti mittespetsiifiliste endonukleaasidega ja eukarüootides deadenüleerimisega. Praeguseks on leitud, et mõlemas domeenis võib terve mRNA lagundamine alata nii 3’ kui ka 5’ otsade struktuuride lagundamisega, aga ka spetsiifiliste või mittespetsiifiliste endonukleaase lõikega. Spetsiifilise endonukleaasne lõige tähendab näiteks RNA lagundamist siRNA (väike segav RNA) (joonis 2) või valesse kohta tekkinud terminatsioonikoodoni (*premature terminating codon*, PTC) tõttu (joonis 3). Sellegipoolest on tähendatud, et bakterites on mittespetsiifilise lagundamise osakaal suurem (Belasco, 2010, Reiman, 2013).



Joonis 2. mRNA lagundamine **a.** miRNA (mikroRNA) ja **b.** siRNA tõttu. Muudetud Belasco (2010) põhjal.



Joonis 3. mRNA lagundamine vales kohas oleva stoppkoodoni (PTC) tõttu. Muudetud Belasco (2010) põhjal.

Kuigi RNaase on vägagi põhjalikult uuritud, ei ole veel täielikku arusaamist protsessidest mis RNA-d enne ja pärast raku surma lagundavad (Belasco, 2010, Fordyce jt, 2013). Inimese rakkudes on kirjeldatud hulk RNA lagundamises osalevaid ensüüme (tabel 1), mille koordineeritud tegevus määrab vähemalt osaliselt mRNA eluea. Tingimustes, kus RNA uuringutes kasutatud kudesid tavaliselt hoiustatakse - surnud, sügavkülmutatud ja vahel ka keemiliste inhibiitorite sees - on RNA lagundamise protsessid tõenäoliselt vähem organiseeritud (Houseley ja Tollervey, 2009, Reiman, 2013).

Tabel 1. Tabel inimeses ekspresseeritavatest Rnaasidest (ei ole täielik nimekiri). Koostatud Sorrentino (2010), Belasco (2010) ja Thorn (2012) artiklite põhjal.

Ribonukleaas	Sihtmärgid/funktsioon
Endonukleaasid	
Argonaut	Lõikab miRNA-mRNA ja siRNA-mRNA komplekse
SMG6	Lõikab PTC-d (enneaegne stoppkoodon) sisaldavat mRNA-d
RNase 1 = Rnaas A	Lõikab nii ühe kui kaheahelalist RNA-d, eelistatavalt polü-C järjestuste juurest
RNaasid 2 ja 3	Eelistavad polü-U järjestusi, ei tööta Polü-A järjestustel
RNaas 4	Eelistab polü-U järjestusi
RNaas 5	Väikese aktiivsusega, lagundab tRNA-d ja rRNA-d
RNaasid 6, 7, 8	Vähe katalüütilist informatsiooni, RNaas 8 on platsentaspetsiifiline
RNaas L, 2-5A-dependent RNase	Väikestes kogustes inaktiivses vormis kõikides rakkudes, interferoonide kaudu aktiveeritult hävitab kogu RNA 4-22 nt pikkusteks lõikudeks
RNaas T2	Roll võib olla sarnane RNaas L-ile
3’ eksonukleaasid	
eksosoom	3’ otsad, mida ei kaitse PABP (polüadenüleeritud järjestusi siduv valk)
CCR4-NOT PAN2–PAN3	Valkude kompleksid, deadenülaasid
PARN	Polü-A Spetsiifiline Ribonukleaas
5’eksonukleaasid	
XRN1 ja XRN2	Monofosforüleeritud 5’ otsad
DCP2 ja DCP5	Eemaldab 5’ cap struktuuri

1.2 RNA intaktsuse hindamine

1.2.1 RNA intaktsuse hindamise vajalikus

Geenieksпressiooni uuringute üks etappidest on RNA intaktsuse hindamine. Kuigi RNA kvaliteedi all mõeldakse nii selle puhtust kui ka intaktsust, mõjutas selle töö tulemusi määravalt üksnes RNA intaktsus, seega pean ma RNA kvaliteedi all silmas üksnes selle intaktsust. Kuna RNA ekspressiooni uuringutetehnikates, nagu geeni ekspressioonikiibi analüüs, RT-qPCR (kvantitatiivne pöördtranskriptaas-PCR) ja RNA-seq, tuleb mingis etapis sünteesida cDNA (Metzker, 2010, veebiaadress 1), peab vastavate uuringute planeerimisel olema kindel uuritavate RNA-de kvaliteedis. Põhjus on selles, et enamasti saab proove analüüsiks ette valmistada eri viisidel ja mitte kõik viisid ei ole sama taluvad RNA kvaliteedi suhtes ((Riedmaier jt, 2010), veebiaadress 1). Näiteks kasutavad osad protokollid cDNA sünteesiks polü-T praimereid. Lagunenud RNA-ga proovidest võib saada selliste meetoditega aga äärmiselt lühikesi cDNA juppe, mis ei pruugi esindada transkriptoomi piisava usaldusväärsusega. Samuti on levinud mRNA analüüsides kogu RNA-st mRNA eraldamiseks kasutada nende polü-A saba (polü-A suhtes rikastamine) (veebiaadress 1). Need mõlemad meetodid põhjustavad aga lõplikus raamatukogus transkripti 5’ otsa alaesindatust, ehk 3’ kõrvalekallutatust (ingl. *3’bias*) (Hestand jt, 2010, Wang jt, 2009) Lagunenud RNA puhul on rRNA eemaldamiseks mõistlikum kasutada näiteks ribodepletsiooni põhiseid tooteid (näiteks Ribo-Zero, Epicentre Biotechnologies). Ribodepletsioon eemaldab kogu RNA-st mRNA asemel just soovimatud RNA tüübid (rRNA, mtRNA (mitokondriaalne RNA)) ja seega jätab ka fragmenteerunud või ilma polü-A sabata RNA (NT: miRNA, tsirkulaarne RNA) edasiste etappide jaoks alles (Reiman, 2013).

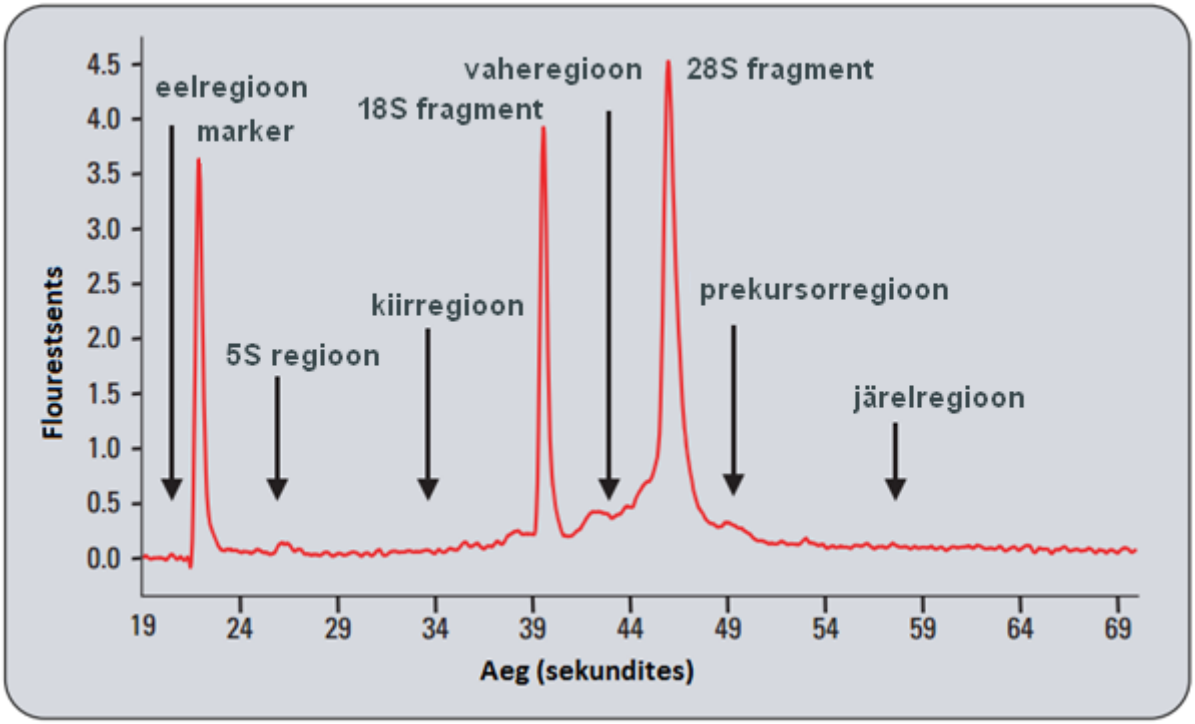
Osaliselt lagunenud RNA korral soovitatakse kasutada cDNA sünteesiks juhuslike praimerjärjestusi (Nardon jt, 2009, Ståhlberg jt, 2004). Need praimerid on kindla pikkusega (6 kuni 21 nukleotiidi (Nardon jt, 2009, Ståhlberg jt, 2004, Stangegaard jt, 2006)) juhusliku järjestusega oligonukleotiidid, mis sisaldavad piisaval hulgal kõikvõimalikke sellise pikkusega järjestusi. Seetõttu suudavad nad ka seonduda ükskõik millisele RNA ahela kohale. Kui viia cDNA süntees läbi selliste praimeritega, siis sünteesitakse juhuslikult fragmente üle kogu algse RNA. Kuna juhuslikult praimeerides saadakse enamasti tulemuseks rohkem ja lühemaid cDNA molekule on võimalus ülehinnata mõndade transkriptide hulka (Stangegaard jt, 2006, Reiman, 2013).

Kuna mõlemal praimeerimismeetodil on omad eelised ja vead, on vaja eri meetodite vahel valides teada võimalikult täpselt algmaterjaliks oleva RNA kvaliteeti.

1.2.2 RNA intaktsuse hindamise metoodika

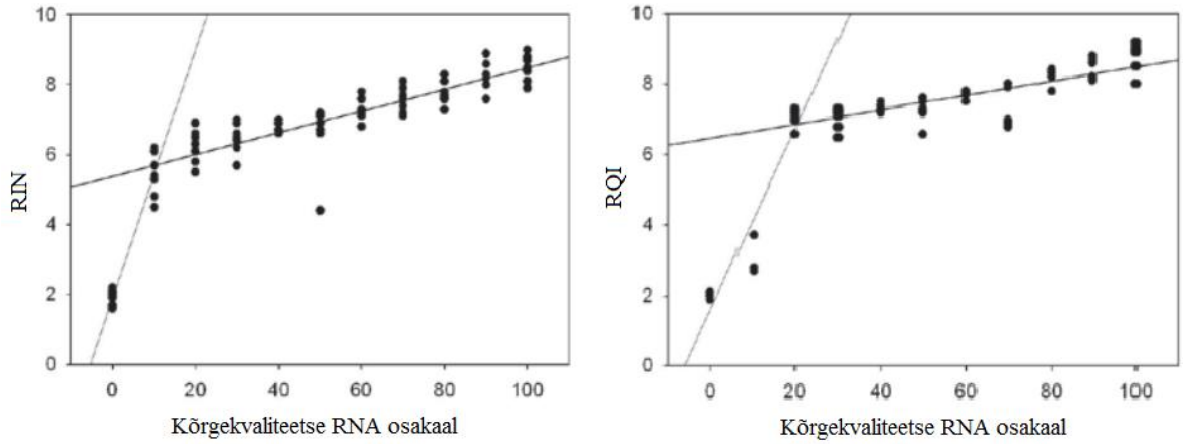
Kõige levinuim meetod kontrollimaks RNA intaktsust on analüüsida RNA-d geelelektroforeesiga denatureerival agaroosgeelil ja võrrelda 18S ja 28S rRNA bändide intensiivsust ja jaotumist (veebiaadressid 2 ja 3). See meetod nõuab enamasti aga suurt hulka RNA-d (EtBr-iga värvides vähemalt 200 ng (veebiaadress 2)) ja intaktsuse hinnang ei ole tegeliku mRNA intaktsusega nii hästi korreleeruv ega ka replitseeritav (veebiaadress 4). Heaks alternatiiviks on sellisel juhul Agilent 2100 Bioanalyzer'i (Agilent Technologies) (veebiaadress 4, Agilent, 2007), poolt välja arvutatav RIN (RNA intaktsusnumber, *RNA Integrity Number*). RIN varieerub ühest kümneni, kus 10 tähistab parimalt säilinud RNA-d ja 1 iseloomustab äärmiselt lagunenud RNA-d. Agilent 2100 Bioanalyzer viib spetsiaalsete kiipide peal läbi kapillaarelektroforeesi ja sealt saadud elektroferogrammi (joonis 4) abil arvutab spetsiaalne tarkvara välja RNA-d iseloomustavad parameetrid nagu kontsentratsioon, 28S/18S rRNA suhe ja RIN (Reiman, 2013).

RIN-i arvutamise algoritm töötati välja adaptiivse õppimise algoritmide abiga, kus algselt eksperdid andsid 1300-le eri määral lagunenud, eri kudedest ja liikidest pärit imetaja RNA-dele, omapoolse kategorisatsiooni (1-st 10-ni). Iseõppimisprogrammid leidsid nende kvaliteediskooride alusel elektroferogrammidest tähtsamad tunnused (signaalpiirkonnad, intensiivsused ja nende suhted), mis siis seoti ühte algoritmi (veebiaadress 4).



Joonis 4. Näide Agilent 2100 Bioanalizer-i elektroferogrammist, ja signaali piirkondadest, mida RNA intaktsusnumbri (RIN) arvutamisel arvesse võetakse. Muudetud veebist saadud pildi põhjal (veebiaadress 4).

Kuna RIN näidu päritolu tuleb inimesepoolsest hinnangust, mitte eelnevatest teoreetilistest mudelitest, ei saa hetkel öelda, et RIN näit on seotud rakulise RNA fragmentide kindla suurusjaotuse funktsiooniga. Seetõttu ei saa ilma varasemate kogemusteta ette arvutada, mis RIN väärtustel eksperiment välja tuleb (veebiaadress 4). Samas on aga RIN ja ka konkureeriva firma toote Experion RQI (RNA kvaliteedi indeks, *RNA Quality Index*) (Bio-Rad Laboratories) algoritmide tulemused omavahel korreleeruvad ja platvormide siseselt hästi replitseeritavad (Riedmaier jt, 2010). Riedmaieri jt (2010) poolt sooritatud eksperimendis näidati aga, et nii RIN kui ka RQI ei hinda hästi rohkem lagunenud RNA kvaliteeti. RNA kvaliteedi muutust simuleeriti, segades eri suhetes tervet RNA-d UV-kiirgusega lagundatud RNA-ga. Sealt saadud tulemustes näidati, et niimoodi segatud RNA proovides oli RIN väärtuste 2,5 kuni 4 vahepealses alas väga vähe proove, kuna nende proovide väärtused hinnati väiksemaks (joonis 5). Muidugi ei pruugi rakuline RNA tavatingimustes sellist moodi laguneda ja RIN-i genereeriv algoritm on loodud loomuliku eukarüoodi RNA analüüsimiseks.

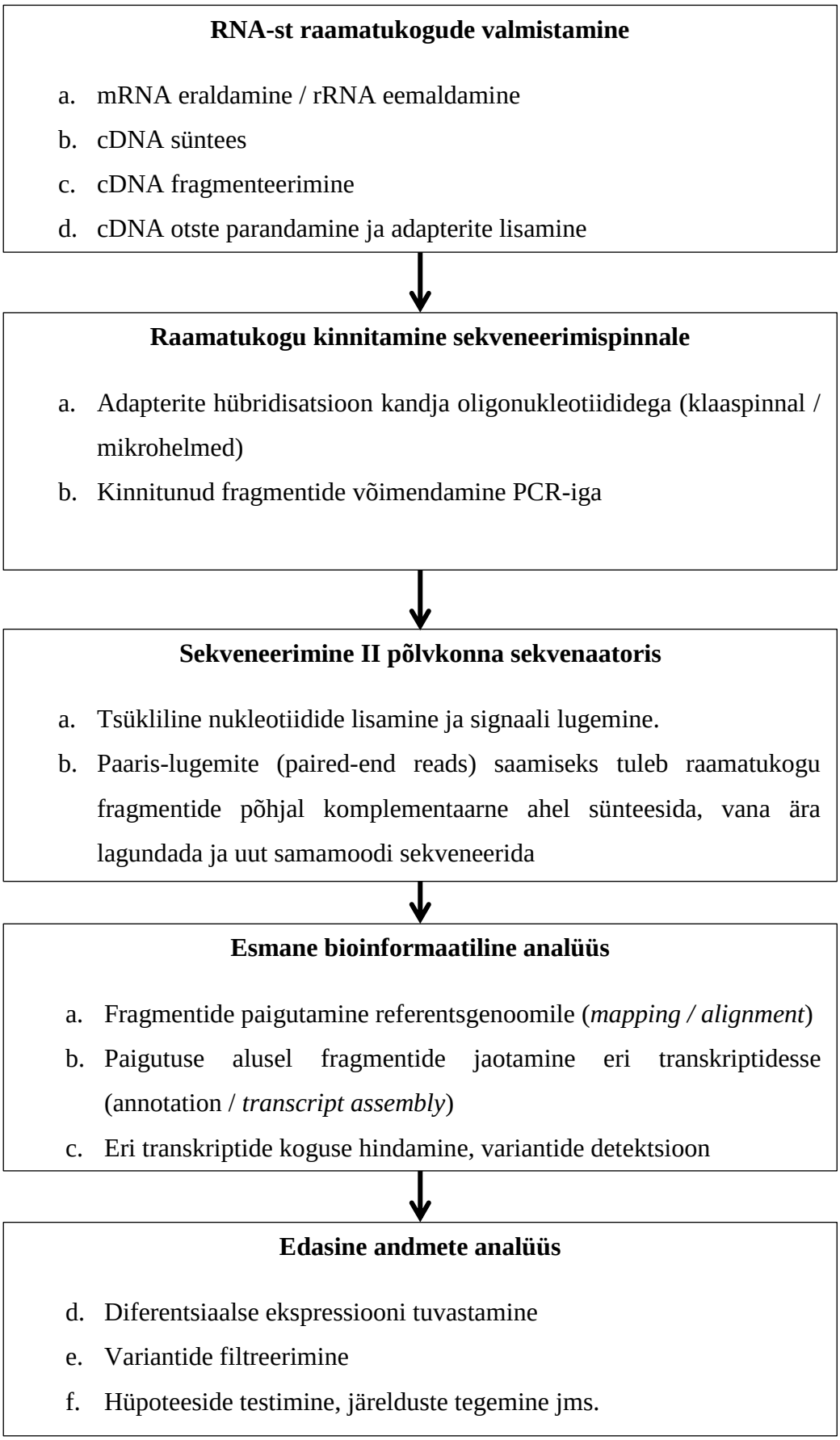


Joonis 5. RIN ja RQI väärtused, kui segada eri suhetes kõrgekvaliteetset ja UV kiirgusega lagundatud veise südame ja soole RNA-d. Alates mingist hetkest muutub kvaliteedi hinnang väga järsult, viidates mõningasele ebausaldusväärsusele madalama kvaliteediga RNA intaktsuse hindamisel. Samas on ka võimalus, et selles uuringus kasutatud erineva intaktsusega RNA simuleerimise viis ei ole kõige parem. Muudetud Riedmaier jt (2010) põhjal.

1.3 RNA-seq ehk RNA sügav-sekveneerimine

RNA sügav-sekveneerimine ehk RNA-seq viitab korraga kogu transkriptoomi sekveneerimisele järgmise põlvkonna sekvenaatorites (joonis 6). RNA-seq on hea transkriptoomi analüüsi meetod, sest sellega saab korraga leida pea kõikide transkriptide ekspressioonitasemed, nende variante (näiteks SNV-d (ühe nukleotiidine variatsioonid)), splaiss-saite ja ka organismis seni avastamata RNA liike (Bloom jt, 2009, Burd jt, 2010, Duitama jt, 2012). RNA-seq meetodi kohta on näidatud ka, et võrreldes geenieksprесiooni kiipidega on selle abil on võimalik leida suuremal hulgal diferentsiaalselt ekspresseerunud geene üle suurema ekspressiooni varieeruvusvahemiku ning määrata täpsemalt alternatiivseid transkripte (Gaidatzis jt, 2009). Tugevaks miinuseks võrreldes ekspressioonikiipidega on aga RNA-seq kõrge hind ja selle hinna sõltuvus sekveneerimise sügavusest. RNA-seq metoodika eelised on aga sügavamalt sekveneerides selgemini nähtavad (Perkins jt, 2014).

RNA-seq protokolli põhiliseks objektiks on RNA põhjal sünteesitud suhteliselt lühikeste (1000-200 nt) cDNA fragmentide kogumik ehk raamatukogu. Nende fragmentide otste sekveneerimisel teise põlvkonna sekvenaatorites saadakse veel lühemad (30-200) sekventsid ehk lugemid (*read*). Kui fragmendid sekveneeritakse mõlemast otsast, siis kutsutakse neid lugemeid paaris-lugemiteks (p*aired-end read*). Peale sekveneerimist kontrollitakse lugemeid kvaliteedi suhtes ja kaardistatakse referentsgenoomile. Kaardistatud lugemite



Joonis 6. RNA sügav-sekveneerimise (RNA-seq) põhilised etapid

asukoha, koguse ja olemasolevate transkriptimudelite põhjal hinnatakse, millised transkriptid ja geenid proovis ekspresseerunud on ja geeni piiridesse paigutunud lugemite arvu järgi määratakse ekspressioonitase. Et ühest proovist saab olenevalt kiibile kantud proovide arvust kümneid kuni sadu miljoneid lugemeid, siis satub neid enamusese ekspresseerunud geenidesse piisavalt et nende arvu põhjal usaldusväärselt hinnata, kui suur proovi tegelik geeniekspresioon olla võis (Reiman, 2013).

Kui tahta võrrelda mitme proovi geeniekspressioone, siis tuleb need omavahel normaliseerida. Normalisatsiooni all mõeldakse siin proovide omavahelise hulga „võrdsustamist“. Näiteks juhul kui ühte proovi kantakse eri hulgas (nt: 1:2) kahele erinevale kiibi rajale, siis oleksid nende proovide ekspressiooniväärtused hilisemates analüüsides ikka sama suured, mitte sellised, kus teisel proovil on ekspressioonide hinnangud iga geeni jaoks 2 korda suuremad kui esimesel. Normaliseerimine iga proovi lugemite koguarvu põhjal ei ole osutunud kõige usaldusväärsemaks ja seetõttu kasutatakse tänapäeval teistsuguseid metoodikaid (Dillies jt, 2012). Näiteks DESeq2 normaliseerib proovide vahelist varieeruvust, jagades iga proovi ekspressioonitasemed läbi sellele proovile vastava suurusfaktoriga. Suurusfaktori arvutamiseks leitakse iga geeni ekspressiooni proovide vaheline geomeetriline keskmine ja jagatakse siis iga ekspressioon vastava geeni geomeetrilise keskmisega (kui geomeetriline keskmine ei ole 0). Iga proovi suhete jadast võetakse siis mediaan ja sellest saabki vastava proovi suurusfaktor (Anders, 2010, veebiaadress 5, Reiman, 2013).

2. TÖÖSKEEM JA EESMÄRGID

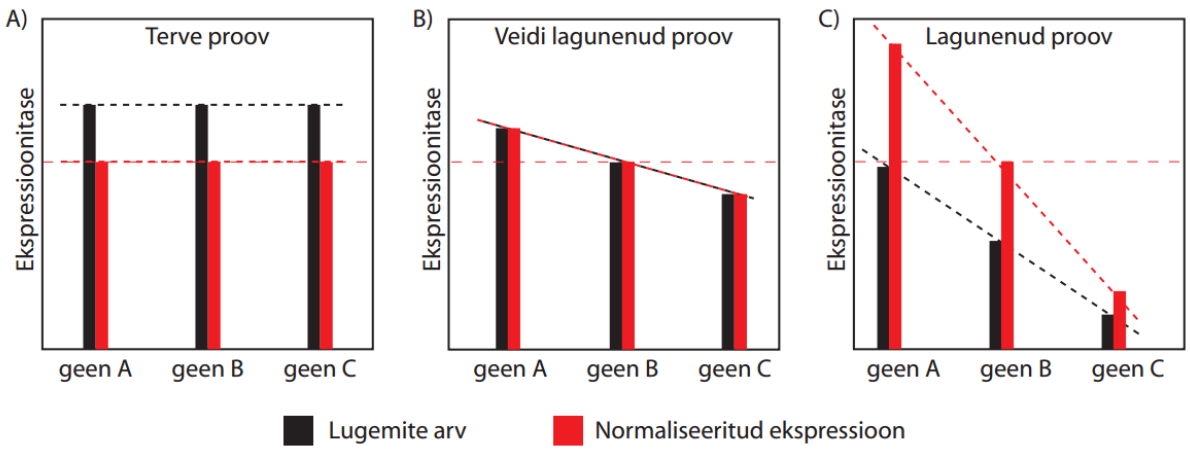
2.1 Töö eesmärgid

Varasemalt olen kirjeldanud oma bakalaureusetöös, kas meie laboris läbi viidud RNA-seq eksperimendis kasutatud proovid olid hoiustamisel kaotanud osa enda intaktsusest ja kas sellest tulenev RIN väärtuste varieeruvus mõjutas geenide vahelisi ekspressiooni tasemeid. Käesolevas töös keskendusin ma spetsiifilisemalt selle mõju kirjeldamisele. Täpsemad uurimusküsimused on järgnevad:

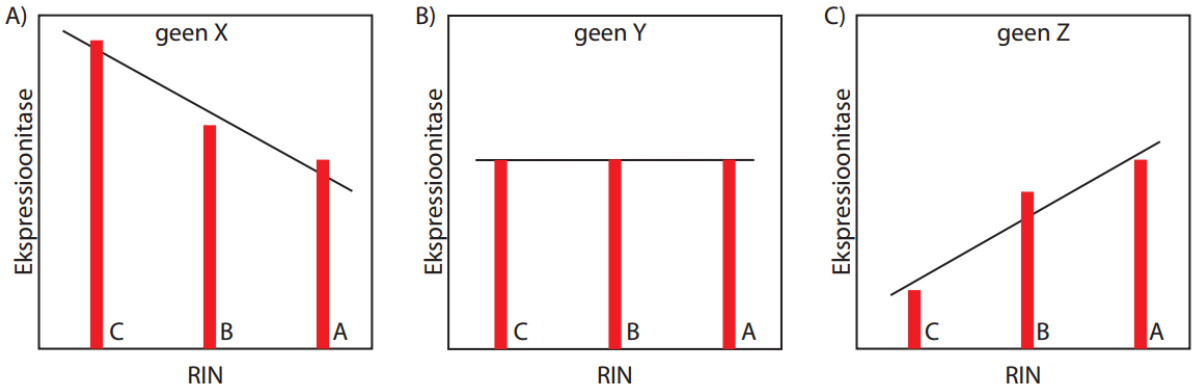
1. Kas on olemas seos proovide RIN väärtuste ja geenide ekspressioonide vahel?
2. Missuguste omadustega geenid on RIN väärtuse poolt enim mõjutatud?
3. Kuidas RIN väärtuse mõju geeni ekspressioonile välja korrigeerida?

2.1.1 Kas on olemas seos proovide RIN väärtuste ja detekteeritud ekspressioonitasemete vahel?

Selles osas vaatasin, kas detekteeritud geeniekspressiooni ja RIN-i vahel on seost. Ma eeldan, et seda seost põhjustav põhiline faktor peaks olema eri RNA molekulide erinev lagunemiskiirus. Joonised 7 ja 8 selgitavad, miks eri kiirusel laguneva RNA tõttu eeldasin näha seost geeniekspressiooni ja RNA intaktsuse vahel.



Joonis 7. Olgu meil kolm kolm hüpoteetilist proovi (A, B ja C), mis on kõik ühe algse proovi võrdse kogusega alikvoodid nii, et proovide B ja C RNA-del on peale alikvootimist lastud laguneda. Proovi RNA lagunemine põhjustab aga detekteeritud ekspressiooni (lugemite arv) langust, sest vähemalt mõned transkriptid, mille pealt raamatukogu sünteesida, on kas osaliselt või täielikult lagunenud. Nende proovide omavaheliseks võrdluseks tuleb proovid normaliseerida (punased tulbad). Lihtsustatult on proovide ekspressiooni normaliseerimine iga geeni ekspressiooni läbijagamist selle proovi ekspressioonitaseme mediaaniga.



Joonis 8. Normaliseeritud ekspressioonidega testin iga geeni ekspressiooni RIN väärtuse suhtes lineaarse korrelatsiooni mudelitega. Testis RIN-i suhtes negatiivses korrelatsioonis olnud geenid on need, mis nii kergesti ei lagunenud ja positiivses korrelatsioonis on proovid, mis lagunesid eriti kergelt. RIN väärtus on seda suurem, mida intaktsem proov.

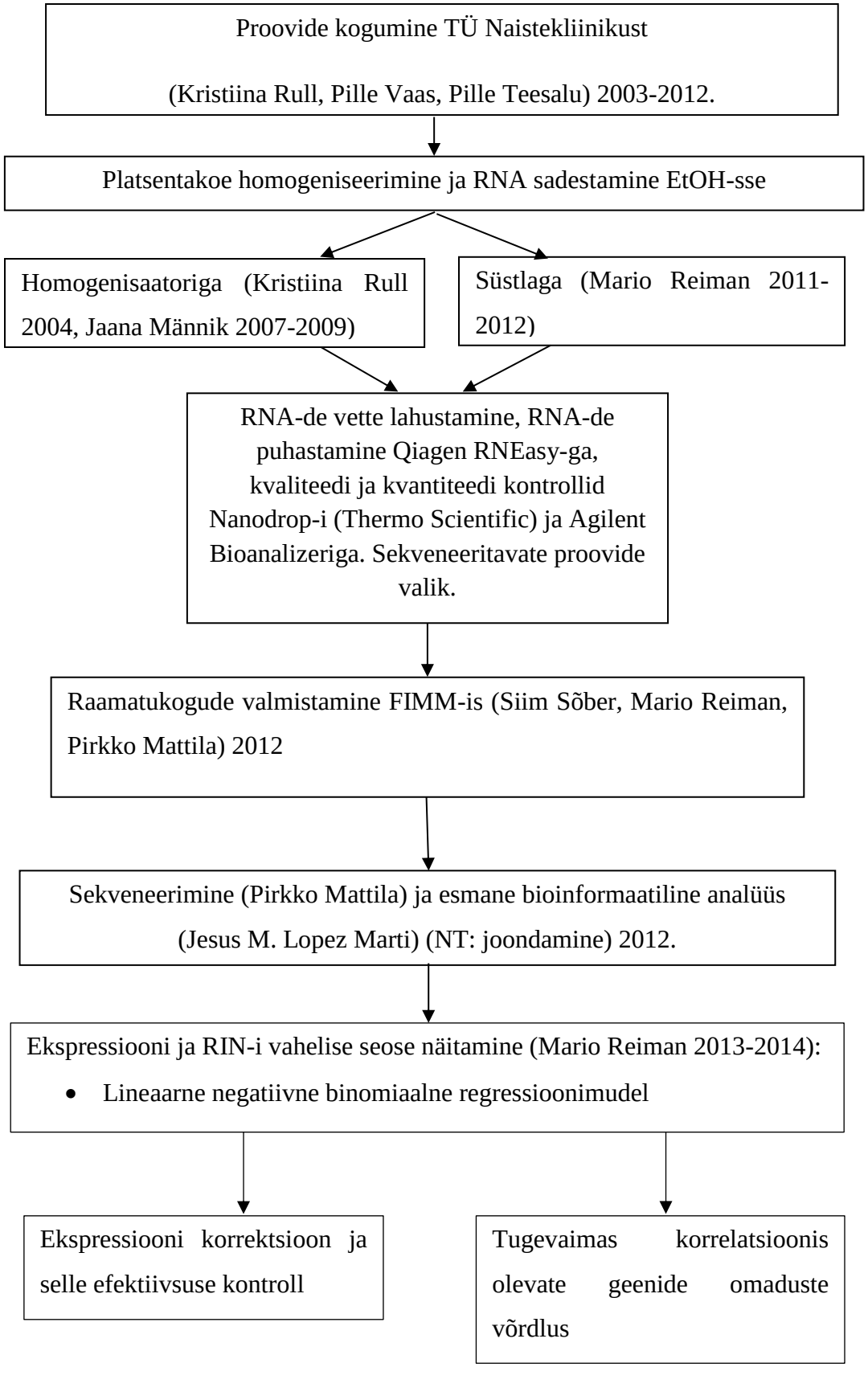
2.1.2 Missuguste omadustega geenid on RIN väärtuse poolt enim mõjutatud?

Et teada saada, kas mingisugused tunnused võivad geenide lagunemiskiirusega oluliselt sõltuvuses olla, võrdlesin tugevaimas positiivses korrelatsioonis olevate geenide omaduste jaotuvuse erinevust tugevaimas negatiivses korrelatsioonis olevate geenide vastavate tunnuste suhtes. Testitud omadused olid valitud enamasti füüsiliselt RNA omaduste põhjal, nagu näiteks transkripti pikkus või selle RNA liik. Täpsed tunnused koos nende võrdlemiseks kasutatud testidega on välja toodud tabelis 2.

2.1.3 Kas RIN väärtuse mõju geeni ekspressioonile saab välja korrigeerida?

Viimaks sooritasin normaliseeritud lugememitele lineaarsete regressioonimudelite põhjal korrektsiooni. See tähendab, ma lahutasin iga geeni regressioonimudelist just sellele geenile leitud RIN-i efekti. Pärast korrektsiooni sooritasin DESeq2-ga uuesti kliiniliste gruppide vahelise diferentsiaalse ekspressiooni leidmise, et leida, kas DESeq2 andis tänu korrektsioonile spetsiifilisemaid tulemusi.

2.2 Tööskeem


Proovide kogumine TÜ Naistekliinikust
(Kristiina Rull, Pille Vaas, Pille Teesalu) 2003-2012.
Platsentakoe homogeniseerimine ja RNA sadestamine EtOH-sse
Homogenisaatoriga (Kristiina Rull 2004, Jaana Männik 2007-2009)
Süstlaga (Mario Reiman 2011-2012)
RNA-de vette lahustamine, RNA-de puhastamine Qiagen RNEasy-ga, kvaliteedi ja kvantiteedi kontrollid Nanodrop-i (Thermo Scientific) ja Agilent Bioanalizeriga. Sekveneeritavate proovide valik.
Raamatukogude valmistamine FIMM-is (Siim Sõber, Mario Reiman, Pirkko Mattila) 2012
Sekveneerimine (Pirkko Mattila) ja esmane bioinformaatiline analüüs (Jesus M. Lopez Marti) (NT: joondamine) 2012.
Ekspressiooni ja RIN-i vahelise seose näitamine (Mario Reiman 2013-2014):
• Lineaarne negatiivne binomiaalne regressioonimudel
Ekspressiooni korrektsioon ja selle efektiivsuse kontroll
Tugevaimas korrelatsioonis olevate geenide omaduste võrdlus
TÜ MRI (TÜ Molekulaar ja Rakubioloogia Instituut)
FIMM (Institute of Molecular Medicine Finland)
TÜ MRI

3. MATERJAL JA METOODIKA

3.1 Proovid

RNA eraldati RNALater (Invitrogen) lahuses hoiustatud platsenta keskregiooni täispaksudest platsenta tükkidest. Proovid koguti SA TÜK Naistekliinikus naistelt, kes olid allkirjastanud vastava informeeritud nõusoleku lehe. Proovid kuulusid REPROMETA valimisse . Seda valimit on varasemalt kasutatud ja detailselt kirjeldatud Männik jt (2010 ja 2012) ja Uusküla jt (2012) töödes. Projekti läbiviimiseks saadi load Tartu ülikooli inimuuringute eetika komiteelt (loa numbrid 117/9, 16.06.2003; 146/18, 27.02.2006; 150/33, 18.06.2006; 158/80, 26.03.2007; 180/M-15, 23.03.2009). Algselt valiti välja 112 patsienti, kes jaotusid seitsmesse alagruppi, kuid käesolevas uuringus kasutati neist vaid 4 gruppi (lisad, tabel 5). Kasutati vaid nelja sellepärast, kuna nende vahelised diferentsiaalse ekspressiooni testid andsid vähem olulisi erinevusi, kui kolm välja jäetud gruppi (esimese trimestri abordid, teise trimestri abordid ja preeklampsiaga rasedused) ja seega segasid teised varieeruvuse allikad RIN efekti määramist väiksemal hulgal. Grupid defineeriti järgnevalt:

- **Hüpotroofia** – vastsündinu, kelle sünnikaal oli väiksem kui Eestis sündinute soole vastav 10-protsentiil. Viimane määrati Karro ja kaasautorite (1997) poolt koostatud sünnikaalu tabelite põhjal.
- **Makrosoomia** – vastsündinu, kelle sünnikaal on suurem, kui Eestis sündinute soole vastav 90-protsentiil. 90-protsentiil määrati Karro ja kaasautorite (1997) poolt koostatud sünnikaalu tabelite põhjal.
- **Gestatsioonidiabeet** – Emal diagnoositi raseduse ajal suhkruhaigus, mis määrati, kui 24.-28. nädalal sooritatud 75 g oraalse glükoositaluvustesti esimeses venoosses vereplasmas oli glükoositase suurem kui 4,8 mmol/l ja/või 1 ja 2 tunni möödudes oli glükoosi tase vastavalt ≥10 mmol/l ja ≥ 8,7 mmol/l. Patsiendid, kellel tekkis diabeet enne 20. nädalat jäeti grupist välja.
- **Normaalsünnid** – normaalse gestatsiooniaja ja sellele vastava kaaluga vastsündinu. Naiste eelnevatel rasedustel ei tohtinud esineda ebaselge etioloogiaga üsasisest kasvupeetust ega preeklampsiat.

Sekveneerimiseks valiti välja neist 56 proovi, igast grupist 8. Selekteeriti põhiliselt parema RIN-i ja ekstreemsemate gruppi iseloomustavate tunnuste järgi; igas grupis üritati hoida poiste ja tüdrukute suhet tasakaalus. Samuti püüdsime hoida gestatsiooniajad võimalikult võrdsed nii gruppide sees kui vahel.

3.2 RNA Eraldamine ja kvaliteedi kontroll

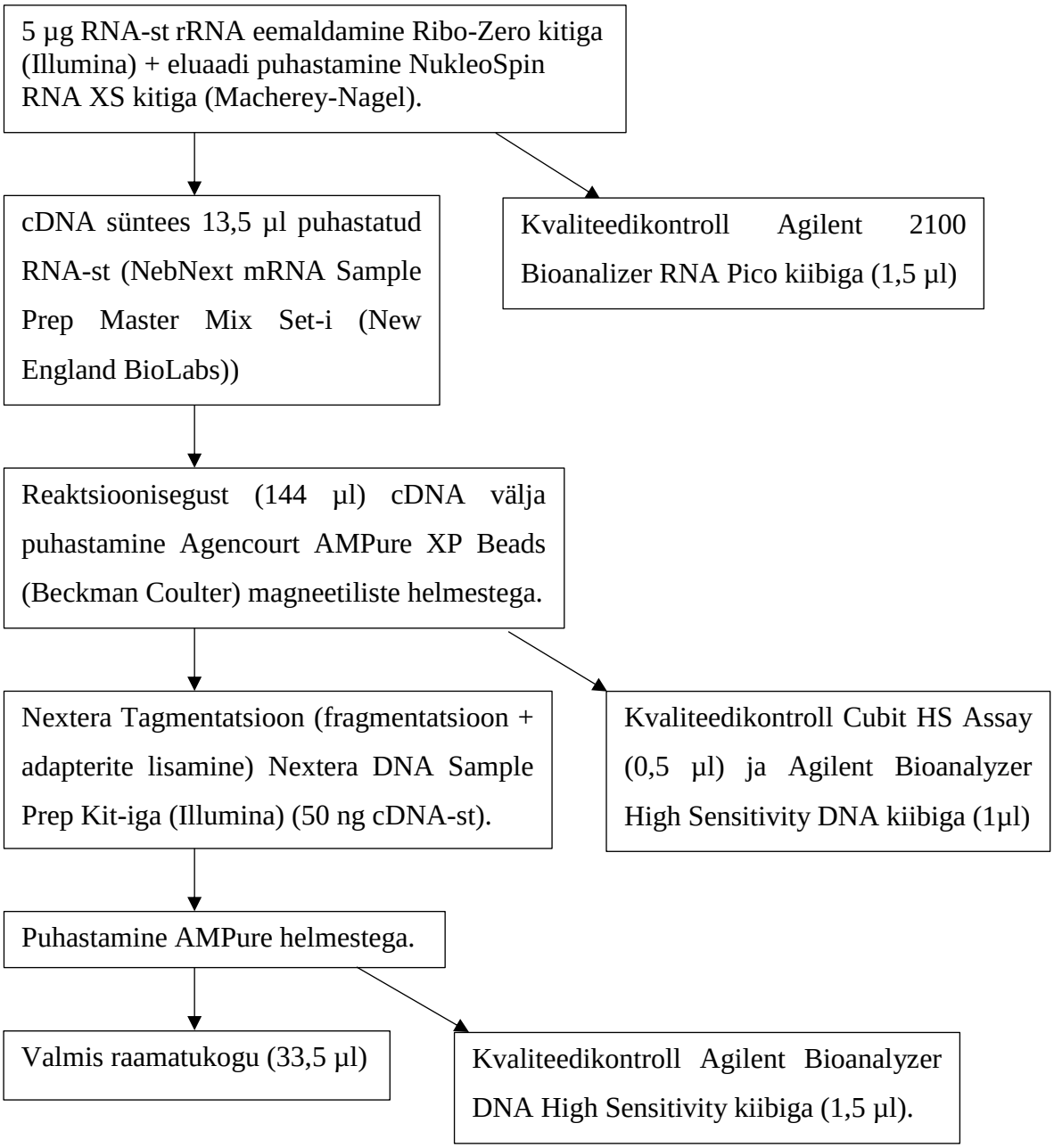
Eraldamise protokolli olen põhjalikult lahti seletanud enda bakalaureusetöös (Reiman, 2013), kuid toon siin välja selle põhipunktid. RNA eraldamiseks kasutatati TRIzol reagenti (Invitrogen). Kuna TRIzol on lenduv ja mürgine, viidi enamus seda kasutavatest etappidest läbi tõmbekapis.

- Esiteks tuli -80^{o} C juures hoiustatud koetükid jää peal ülesse sulatada ja siis TRIzol lahuses homogeniseerida.
- Homogeniseerisime IKA-ULTRA-TURRAX T8 (Ika Works) homogenisaatoriga (enne 2010 eraldatud proovid) või suspendeerides TRIzol-platsenta segu süstlaga (23 gauge nõelaga).
- Homogenaadist eraldasime RNA TRIzol-i juhendit järgides (1-st RNA sadestus etappi pikendasime saagise suurendamiseks 25 minutini).
- Peale RNA puhtasse vette lahustamist mõõtsime Nanodrop-iga (Thermo Scientific) RNA kontsentratsiooni ja puhastasime proove lisaks Qiagen RNeasy MinElute kitiga.
- Viimaks kontrollisime RNA kvaliteeti ja intaktsust Nanodrop-i ja Agilent 2100 Bioanalizeriga (RNA 6000 Nano kiipidega). Saadud kvaliteediandmete ja kliinilise sobivuse põhjal valisime välja lõplikud proovid, mille viisime Soome sekveneerimisele.

3.3 Raamatukogude valmistamine ja sekveneerimine

Raamatukogude valmistamine, sekveneerimine ja esmane bioinformaatiline töötlus viidi läbi FIMM-is (Institute for Molecular Medicine Finland). Meie grupist osalesid raamatukogude valmistamise etapis Siim Sõber ja Mario Reiman, ülejäänud tehti FIMM-i sekveneerimise tuumiklabori poolt (P. Mattila ja Jesus M. Lopez Marti).

Soomes tehti kõigepealt proovidele uued kvaliteedikontrollid Cubit RNA Assay (Life Technologies), Cubit dsDNA BR Assay (Life Technologies) ja Agilent 2100 Bioanalizer 6000 RNA Nano-ga. Selles etapis saadud RIN väärtusi kasutatigi edasistes analüüsides RNA intaktsuse hindamiseks. Raamatukogude valmistmisel käitusime järgneva skeemi põhjal.



Joonis 9. FIMM-is (Institute for Molecular Medicine Finland) kasutatud RNA-seq raamatukogude valmistamise tööskeem.

3.3.2 Sekveneerimine

Kui raamatukogud olid valmis ja puhastatud, kanti nad kiipidele (*flow cell*). Illumina HiSeq 2000 kiipidel on 8 rada, iga rada on võimeline tootma 200 – 400 miljonit paaris-lugemit. Ühele rajale võib kanda mitu proovi, kuid eri proovid peavad olema raja suhtes unikaalsete indeksjärjestustega ja arvestada tuleb ka sellega, et proovi kohta saab seega vähem lugemeid. Meie uuringus kanti radadele 2 – 4 proovi, mis andis keskmiselt 76 miljonit kvaliteedikontrolli läbinud lugemit (täpsemalt saab vaadata lisadest, tabel 5).

Enne sekveneerimist tuleb raamatukogud kinnitada kiibile, neid seal võimendada ja praimeerida sekveneerimiseks, milleks kasutati Illumina cBot-i (TruSeq PE Cluster Kit v3 (Illumina)).

3.4 Esmane andmete analüüs

Esmane andmete analüüs (kvaliteedi kontroll, lugemite ühendamine paaris-lugemiteks, joondamine) ja ekspressioonide hindamine viidi läbi FIMM-is Jesus M. Lopez Marti poolt (FIMM RNA-seq pipeline v2.1).

- Lugemite kvaliteeti kontrolliti FastQC-iga (veebiaadress 11)
- Joondamine Tophat-iga (v2.0.3) (Trapnell jt, 2009) GRCh37 referentsgenoomi vastu
- Geenide ekspressioonitasemed HTSeq-iga (veebiaadress 13) Ensembl-i Homo sapiens geenid versioon 67

HTSeq-ist saadud geenide ekspressioonitasemed kombineerisin üheks tabeliks ja normaliseerisin DESeq2-ega.

3.5 Mudel RIN väärtuse ja geeni ekspressiooni vahel

Et näidata seost geenide ekspressioonide ja RIN väärtuste vahel ning hiljem seda seost ka välja korrigeerida otsustasin koostada igale geenile korrelatsioonimudeli. Igasse mudelisse oli kaasatud 32 proovi ja iga filtratsiooni läbinud geeni kohta oli 2 mudelit. Mõlemis mudelis on seletatavaks tunnuseks antud geeni ekspressioon ja esimeses mudelis on seletavaks tunnuseks RIN väärtus, teises seda ei ole. Sugu oli mõlemisse mudelisse kofaktoriks võetud, kuna paljude geenide korral parandas see tugevalt mudeli sobivust ja teiste geenide korral RIN väärtuse sobivust oluliselt ei häirinud. Neid kahte mudelit võrreldakse siis tõepära suhte

testiga (LRT, ingl. *Likelihood-Ratio Test*)(R-i anova() käsuga). Mudeli jääkide jaotuseks eeldan negatiivset binomiaalset jaotust, kuna seda peetakse RNA-seq ekspressioonimudelite korral sobivaks hinnanguks ja kasutatakse ka DESeq pakettides (Anders ja Huber, 2010, Rapaport jt, 2013, Wan ja Yan, 2012). Mudeli koostamiseks kasutasin R-i MASS paketist pärinevat „nb.glm“ käsku (Venables ja Ripley, 2002), millest on jägnevas reas näiteks toodud täis mudeli avaldis.

```
mudel.nb=try(glm.nb(count~sugu+RIN,data=data.korrel,link=identity))
```

Enne testimist eemaldasin ma kõikide HTSeq-iga annoteeritud geenide hulgast (50807) geenid, mille keskmine ekspressioonitase jäi alla 51 lugemi (jättes alles 14673 geeni), kuna olin juba varasemalt näinud, et selliste geenide korral on müra tase suhteliselt suur (Reiman, 2013) ja madala ekspressiooniga geenide eemaldamist soovitati ka DESeq-i juhendis (Anders ja Huber, 2012). Pärast testimist eemaldasin ka proovid, mille keskmine ekspressioon oli küll üle 50 lugemi, kuid see tulenes põhiasjalikult ühe-kahe proovi äärmiselt kõrgest ekspressioonist (AIC, Akaike informatsiooni kriteeriumi erakordselt kõrge väärtuse põhjal).

3.6 RIN-iga tugevaimat seost näidanud geenide omavaheline võrdlus

Geenide ekspressiooni ja RIN väärtuse vahelise seose tugevuse hindamisest sain iga geeni jaoks korrelatsiooni suuna ja seose olulisust kirjeldava p-väärtuse. Nendest geenidest valisin p-vääruse alusel välja kaks gruppi: 1000 kõige tugevamas positiivses korrelatsioonis olevat geeni ja 1000 kõige tugevamas negatiivses korrelatsioonis olevat geeni. Saadud gruppe võrdlesin siis omavahel EnsEMBL andmebaasist leitud annotatsioonide põhjal (EnsEMBL versioon 67 - mai 2012). Nendest võrdlustest leidsin, kas keskmisest kiiremini lagunevad (positiivses korrelatsioonis olevad) geenid erinevad oma tunnustelt stabiilsematest geenidest (negatiivne korrelatsioon). Võrreldud tunnused ja nende võrdlemiseks kasutatud testid on toodud välja tabelis 2

Tabel 2. EnsEMBL-ist valitud annotatsioonid, mida kasutasin kiiresti ja aeglaselt lagunevate geenigruppide võrdluses. Kõikide tunnuste juures võrreldakse selle tunnuse jaotuvuse või keskmise erinevust kahe geenigrupi vahel: 1000 väikseima p-väärtusega positiivses ja 1000 väikseima p-väärtusega negatiivses korrelatsioonis olevat geeni. Transkripti spetsiifilistes tunnustes valisin vastava geeni transkriptiks HTSeq tabelite põhjal enamus proovide korral oli kõige kõrgema ekspressioonitasemega transkripti.

Tunnus	Test
Transkripti pikkus	Wilcoxon-i järk-summa test (kahesuunaline keskmiste võrdlemine)
Transkripti ekspressioonitase (lugemit transkripti nukleotiidi kohta - l/nt)	
Levinuima transkripti GC%	
RNA liik (NT: mRNA, miRNA jne)	Fisher-i täppistest* (Tunnuste jaotuse võrdlemiseks)
Geeni alternatiivsete transkriptide arv	
Levinuima transkripti eksonite arv	
3'-UTR-i pikkus	
5'-UTR-i pikkus	

* empiiriline p-väärtus, arvutatud 20 miljoni permutatsiooni kohta

3.7 Geenide ekspressiooni korrigeerimine RIN-i suhtes

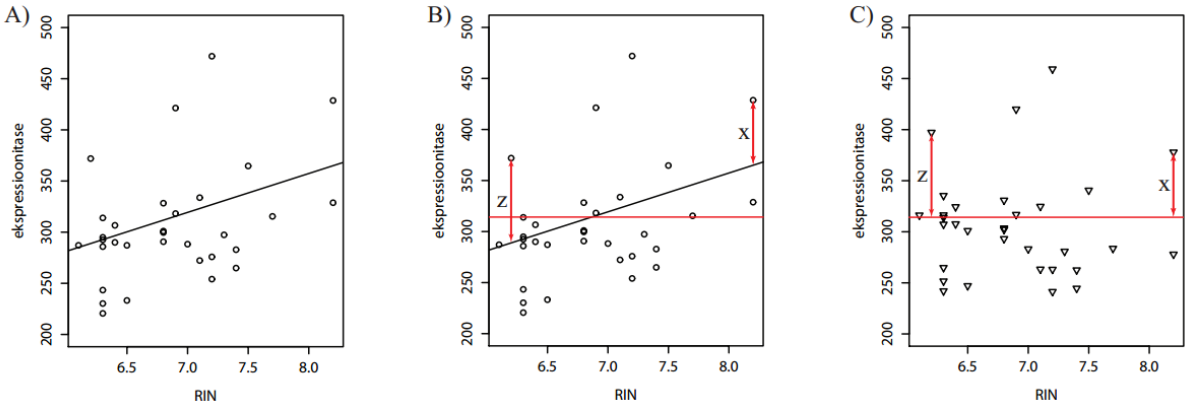
Koostatud geenieksprессiooni ja RIN vahelise mudeli hinnangute põhjal korrigeerisin Popova jt (2008) eeskujul geenide ekspressioonitasemeid Y_{ij} järgneva võrrandi järgi

$$Y_{ij}^{K} = Y_{ij} + \hat{E}(Y|geen = i, RIN = \overline{RIN}) - \hat{E}\left(Y\middle|geen = i, RIN = RIN_j\right) =$$

$$= Y_{ij} + \hat{\beta}_{Ri} * (\overline{RIN} - RIN_j)$$

Y_{ij}^{K} on korrigeeritud ja normaliseeritud geeni i ekspressioonitase proovis j. Y_{ij}on geeni ekspressioon korrigeerimata kujul. $\hat{E}(Y|geen = i, RIN = \overline{RIN})$ on korrelatsioonimudeli hinnang geeni i ekspressiooniväärtusele, kui RIN väärtus oleks võrdne kõikide proovide aritmeetilise keskmisega. $\overline{RIN}$ ja RIN_j tähistavad vastavalt keskmist RIN väärtust ja proovi j RIN väärtust ning $\hat{\beta}_{Ri}$ on mudeli poolt ennustatud tõus RIN-i mõju korrelatsioonisirgele. Sellise korrektsiooni erandiks on juhud, kus ekspressiooniväärtus läheb väiksemaks kui 0,1,

mispuhul see ära piiratakse (sest glm.nb() meetod ei saa võtta algandmeteks efektiivselt negatiivseid lugemeid). Korrektsiooni visuaalne näide on välja toodud joonisel 10.



Joonis 10. Näide geeni ekspressiooni korrigeerimisest. Vasakpoolne joonis (A) on graafik ühe geeni (ENSG00000000460, C1orf112) korrigeerimata ekspressioonist proovide RIN väärtuse vastu ja punktidest läbi tõmmatud joon tähistab mudelist saadud korrelatsioonisirget. Punase joonega on tähistatud horisontaalne sirge, mis lõikab korrelatsioonisirget proovide RIN väärtuste aritmeetilise keskmise kohal ($\overline{RIN}$≈6,9). Korrektsiooni tulemusel (C) asuvad kõik punktid (va need, mis oleksid väiksemad kui 0,1) horisontaalsest joonest sama kaugel, kui nad varem korrelatsioonisirgest olid olnud (NT: Z jz X).

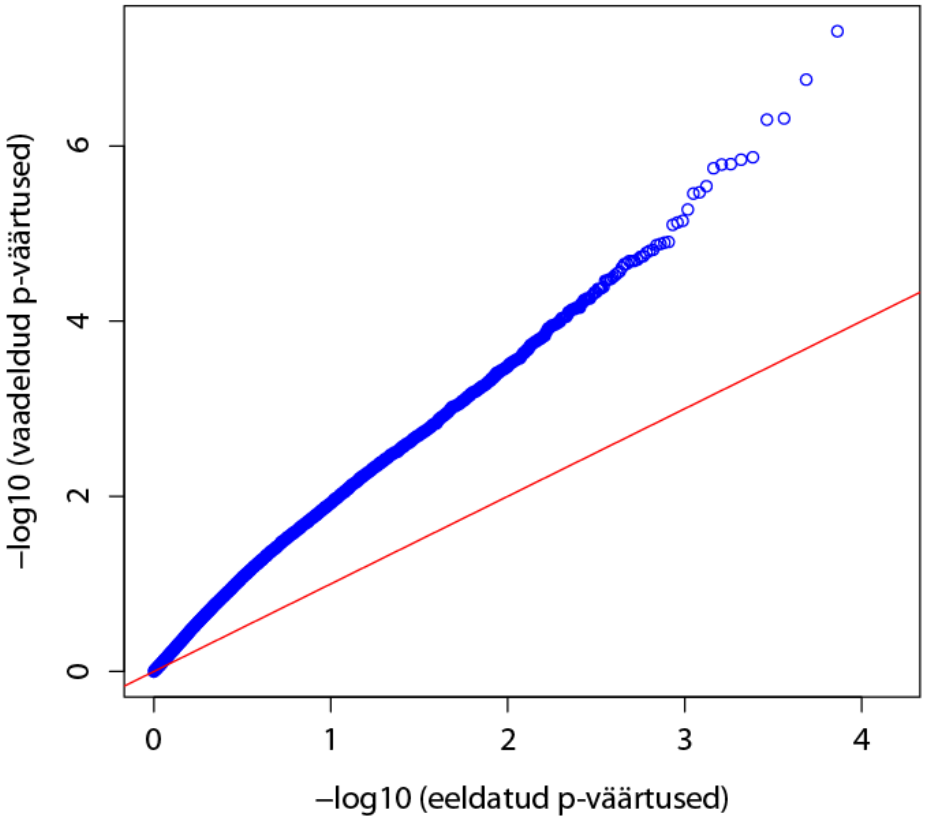
Selles uuringus viisin korrektsiooni läbi selleks, et vaadata, kas korrekteeritud ekspressiooniandmetega saab DESeq2 diferentsiaalse ekspressiooni tuvastamist täpsemaks ja korrektsemaks muuta. DESeq2 nõuab algandmeteks aga normaliseerimata lugemeid, mitte juba normaliseeritud reaalarvulisi ekspressiooniväärtusi. Seetõttu viin korrektsiooni sisse suurusfaktorite abiga. Lõplik geeni ekspressioon saadakse, jagades lugemeid läbi suurusfaktoriga sF_{ji}, ja seega koostasin suurusfaktorite maatriksi, kus iga üksik suurusfaktor sF_{ji} on vastava proovi i ja geeni j normaliseerimata ekspressiooni X_u ning korrigeeritud ekspressiooni X_k jagatis. Erandiks on siinkohal juhud, kus korrigeeritud ekspressioon oleks väiksem kui 0,1, mispuhul see väärtus asendati 0,1-ga.

$$sF_{ij} = \begin{cases} sF_{ij} = \dfrac{X_u}{X_k}, kui\ X_k \geq 0{,}1 \\ sF_{ij} = \dfrac{X_u}{0{,}1};\ kui\ X_k < 0{,}1 \end{cases}$$

4. TULEMUSED

4.1 RNA intaktsus mõjutab detekteeritud geeniekspressiooni

50807-st detekteeritud geenist testisin vaid geene, mille keskmine normaliseerimata ekspressioon oli kõrgem kui 50 lugemit. Madalamate lugemite arvuga geenid tekitavad rohkem mittespetsiifilist müra ja on seetõttu mõttekas analüüsidest välja jätta (Tarazona jt, 2011, Anders ja Huber, 2010). Madala ekspressiooniga geenide eemaldamine jättis testimiseks alles 14673 geeni. Allesjäänud geenide jaoks koostasin meetodites kirjeldatud viisil lineaarsed regressioonimudelid ja nende põhjal määrasin RIN väärtuse ja geeniekspressiooni vahelise seose tugevuse. Seose p-väärtuse leidmiseks võrdlesin täielikku mudelit (s.o. ekspressioon~sugu+RIN) vaesustatud mudeliga (ekspressioon~sugu) tõepära suhte testi abil. Testi tulemused koondasin p-väärtuste Q-Q plotiks (kvantiil-kvantiil plot) (joonis 11).



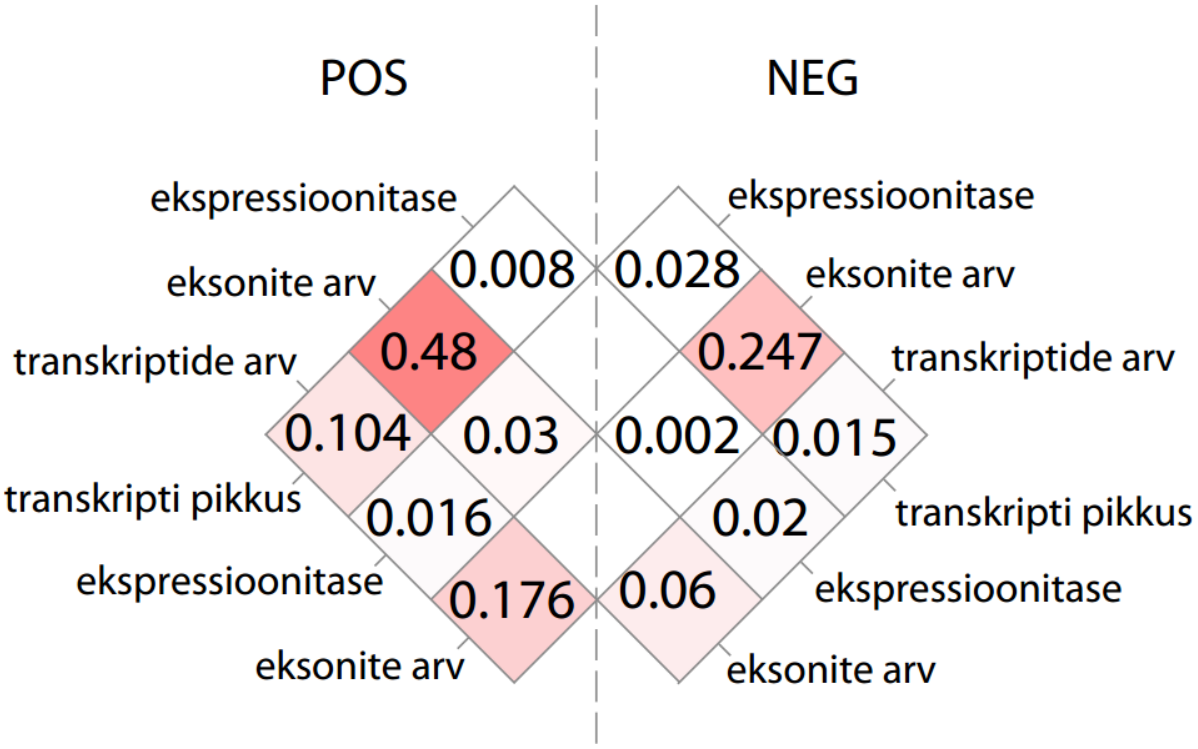
Joonis 11. Regressioonanalüüsi tulemuste põhjal koostatud p-väärtuste Q-Q plot näitab geenide testimisel tulemuseks saadud p-väärtuste vaadeldud (y-telg) ja nullhüpoteesi kehtimise korral eeldatud (x-telg) jaotust. Juhul kui need oleksid täiuslikult ühtinud, siis peaksid kõik punktid paiknema punasel joonel.

Q-Q ploti põhjal saab ütelda, et testide tulemused erinevad nullhüpoteesi kehtimise korral eeldatavast p-väärtuse jaotusest märgatavalt ja ka ei ole need sarnased tulemustele,

mida eeldaks näha näiteks kahe kliinilise grupi võrdlusest juhul kui gruppide vahel on väheste geenide diferentsiaalse ekspressiooniga seletatav põhjus olemas (sinised punktid kattuksid enamuses punase joonega v.a. vähesed geenid ploti paremal ülemises nurgas). Kui korrigeerida p-väärtusi FDR-ideks (valeavastuse määr, *False Discovery Rate*) siis oleks 1122 geenil ekspressioon oluliselt korreleerunud RIN väärtusega, kus FDR<0,1. Selline ühtlaselt kõrvalekallutatud jaotus oli minu hüpoteesi korral aga täiesti oodatav, sest ma eeldasin, et väike erinev lagunemiskiirus oleks kõigil geenidel olemas ja see muutuks üle geenide suhteliselt ühtlaselt. Keskmiselt on väärtuste ühtlasel jaotusel eeldatav p-väärtus 2.15 korda suurem kui vaadeldud väärtus (mediaan=2,34).

4.2 RIN-iga tugevaimat seost näidanud geenide omavaheline võrdlus

Eelmises peatükis leidsin iga geeni jaoks kas ja kui tugevalt RIN väärtus geeni ekspressiooni mõjutada võis. Sealt valisin mõlemast korrelatsiooni suunast välja 1000 olulisima seosega geeni. Seega sain tuhat olulisemat positiivses korrelatsioonis olevat geeni (maksimaalne p-väärtus=0,0205, ehk FDR=0,146, tähistan lühendiga POS) ja tuhat olulisemat negatiivses korrelatsioonis olevat geeni (maksimaalne p-väärtus=0.0198, ehk FDR=0.1451, tähistan lühendiga NEG). Tabelis 2 tähistatud tunnuste testimise tulemused koondasin tabelisse 3 ja igat tunnust kirjeldab täpsemalt ka vastav joonis (Joonised 13 - 18). Kuna eeldasin, et mõned testitud tunnused võivad omavahel korrelatsioonis olla, siis leidsin mõlema grupi jaoks transkripti pikkuse, geeni pikkuse suhtes normaliseeritud ekspressioonitaseme, alternatiivsete transkriptide arvu ja eksonite arvu vahelised Pearsoni korrelatsioonikordajate ruudud (joonis 12).



Joonis 12. Pearsoni korrelatsioonikordajate ruudud transkripti pikkuse, geeni transkriptide arvu, eksonite arvu ja transkripti pikkuse suhtes normaliseeritud ekspressioonitaseme vahel. Vasakpoolsed korrelatsioonikordajad on POS grupi muutujate vahelised seosed ja parempoolsed on NEG grupi sisesed korrelatsioonid. Kõige tugevam seos on mõlemal juhul eksonite arvu ja transkripti pikkuse vahel.

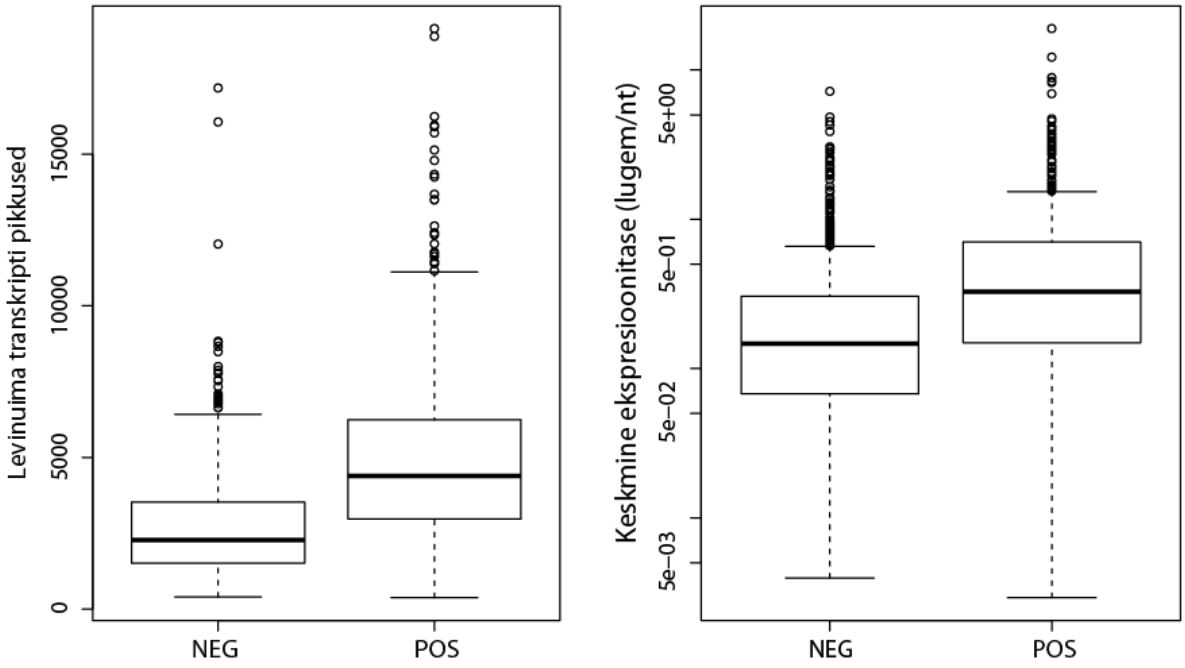
Kõik tabelis 3 nimetatud tunnused andsid testide tulemusteks olulisi p-väärtusi. Tunnused, kus testida tuli otseselt transkripti, mitte geeni tunnust, valisin HTSeq ekspressiooniandmete põhjal välja vastava geeni mažoorse transkripti. Testid jagunesid kaheks, Wilcoxon-i järk-summa test ja Fisheri täppistest. Eelmainituga testisin gruppide keskmiste väärtuste erinevus ja viimatimainituga gruppide väärtuste jaotuse erinevust.

Tabel 3. Tabelis X nimetatud tunnuste testitulemused. POS ja NEG tähistavad siin vastavalt RIN väärtusega positiivse korrelatsiooniga ja negatiivse korrelatsiooniga geenigruppide vastavaid tunnuseid. Märkuste veergu on märgitud kiire ülevaade testi ja joonise tulemusest, täpsema hinnangu jaoks tuleks vaadata viimases veerus märgitud vastavat joonist.

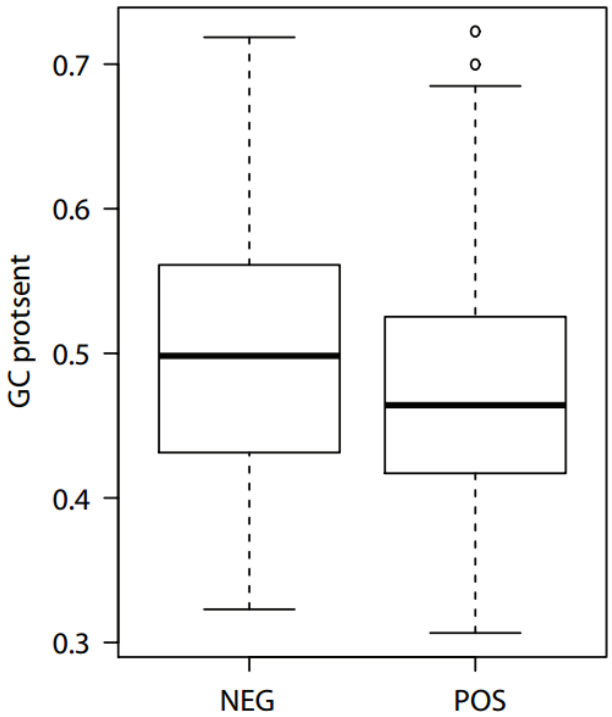
Tunnus	Test	p-väärtus	Efekti suund või märkused	Joonis
Transkripti pikkus	Wilcoxon-i järk-summa test	< 2.2e-16	POS>NEG	14 A
Geeni ekspressioonitase*		< 2.2e-16	POS>NEG	14 B
Levinuima transkripti GC%		8.485e-11	NEG>POS	8
RNA liik (NT: mRNA, miRNA jne)	Fisher-i täppistest**	< 5e-08	POS hulgas rohkem mRNA-sid	3
Geeni alternatiivsete transkriptide arv		3.265e-05	POS hulgas rohkem 1 transkriptiga geene	4
Levinuima transkripti eksonite arv		< 5e-08	POS hulgas rohkem 1 ja väga paljude eksonitega geene	5
3'-UTR-i pikkus		< 5e-08	POS>NEG	6
5'-UTR-i pikkus		0.0009434	NEG hulgas on rohkem lühemaid	7

*ekspressioonitase on läbi jagatud vastava transkripti pikkusega, ühik lugem/nt
** empiiriline p-väärtus, arvutatud 20 miljoni permutatsiooni kohta

Positiivses korrelatsioonis olevate geenide keskmine transkripti pikkus oli 4158 nukleotiidi (mediaan = 3595 nt) ja NEG grupis oli see 2672 nukleotiidi (mediaan = 2266)(joonis 13 A). Samal moel testisin ka geenide keskmist ekspressioonitaset ja levinuima transkripti GC protsenti. POS grupi geenide keskmine ekspressioonitase oli 0.6143 lugemit transkripti nukleotiidi kohta (ilma pikkuse suhtes normaliseerimata 2908) ja NEG grupi keskmine ekspressioonitase oli 0,3156 lugem/nt (ilma normaliseerimata 659,7) (Joonis 13 B). POS ja NEG gruppide keskmised transkriptide GC %-id olid vastavalt 49,98 % ja 47,61 % (Joonis 14).

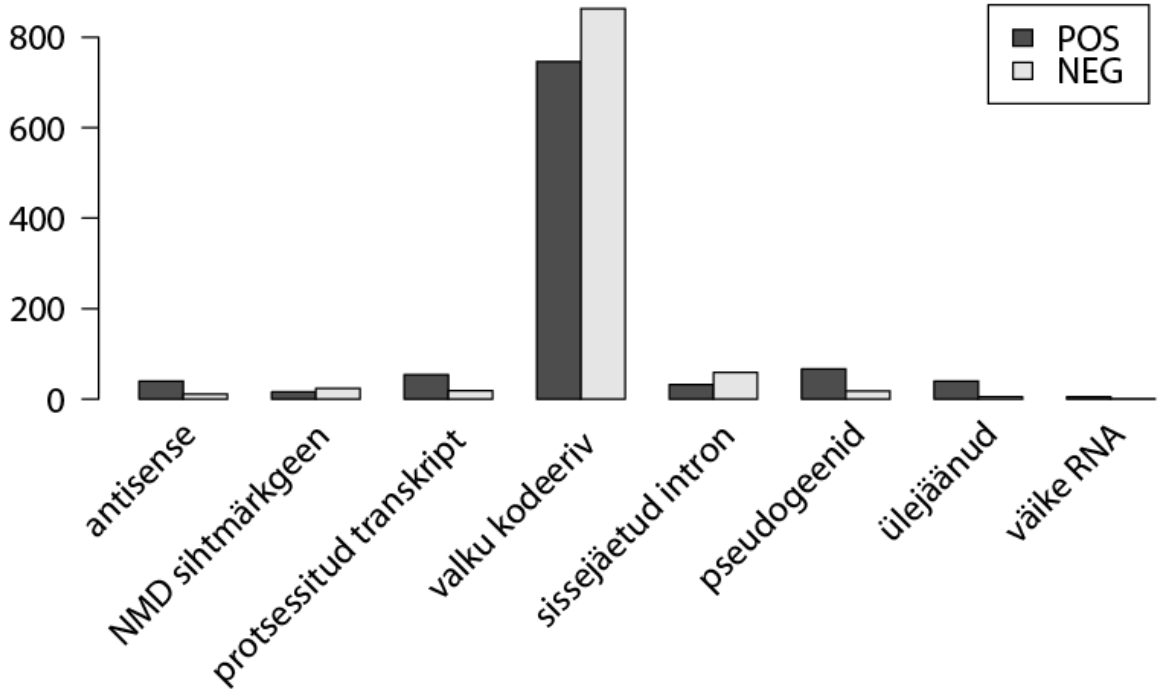


Joonis 13. Karpdiagramm A näitab mis jaotusega olid testitud geenide levinuimate transkriptide pikkused kahes grupis - NEG ja POS, tähistamaks vastavalt tuhandet geeni, mis olid tugevaimas negatiivses korrelatsioonis ja tuhandet geeni, mis olid tugevaimas positiivses korrelatsioonis. B diagramm näitab samade geenide keskmisi ekspressioonitasemeid (normaliseeritud transkripti pikkuse suhtes).



Joonis 14. RIN väärtusega negatiivses ja positiivses korrelatsioonis olevate geenide mažoorsete transkriptide GC %. Keskmised väärtused vastavalt 49,98% (standardhälve=8,28%) ja 47,61% (standardhälve=7,38%)

RNA liikide/biotüüpide võrdlemiseks grupeerisin mõningaid kategooriaid sarnasuse alusel kokku, sest enamus transkripte kuulus valke kodeerivate mRNA-de hulka ja ülejäänud grupid olid valdavalt väga väikesed (Joonis 15). Valdavaks erinevuseks POS ja NEG grupi vahel oli see, et NEG grupis oli suuremal hulgal proteiine kodeerivaid transkripte (863 vs 746) ja POS grupil oli seega rohkem transkripte teistest RNA liikidest (v.a. *retained intron* ja *nonsense mediated decay'*ga seostatud transkriptide gruppides).

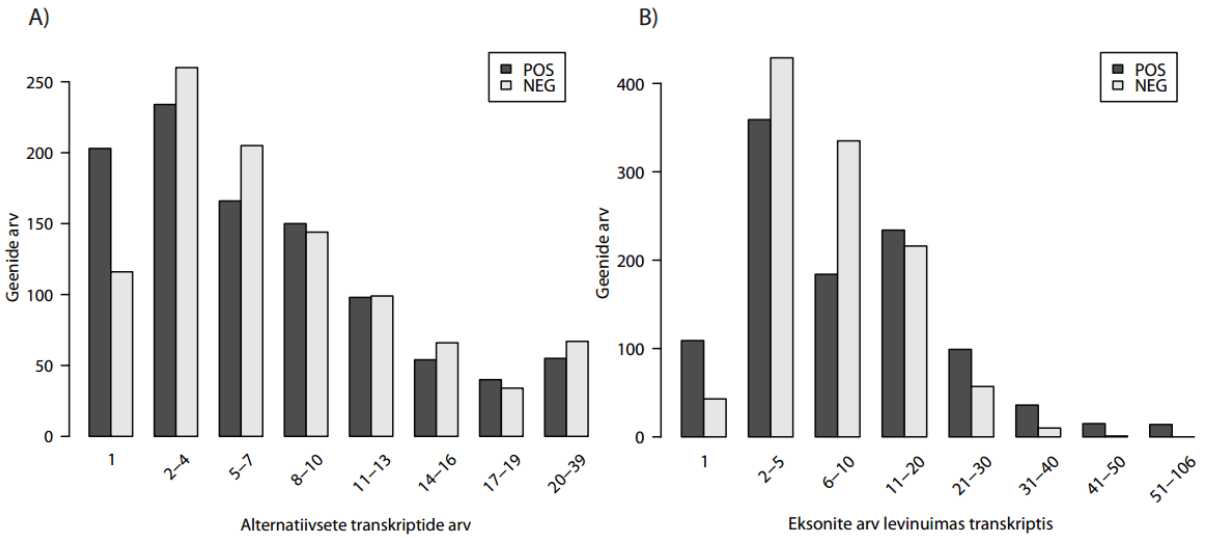


Joonis 15. RNA liikide jaotuse võrdlus POS ja NEG gruppide vahel. Annotatsioonide seletused on võetud Vega/Havana kodulehelt (veebilink 6), toon siin välja lühikese kokkuvõtte.

- Antisense – katab genoomil vähemalt osaliselt vastalahelal asuvat valku kodeerivat lookust.
- NMD (*Nonsense mediated decay*) sihtmärkgeen – mRNA järjestus lõppeb vähemalt 50 nukleotiidi enne järgmist splaiss-saiti
- Protsessitud transkript – lncRNA, mida ei saa grupeerida teiste analoogidega kokku
- Valku kodeeriv – avatud lugemisraamiga mRNA
- Sissejäetud intron– omab alternatiivset transkripti, kus osa kodeerivast alast on välja jäetud (see ala ei ole intronitega piirduv)
- Pseudogeenid – sarnased valku kodeerivatele, kuid sisaldavad ORF-i segavat mutatsiooni
- Väike RNA – miRNA, snRNA snoRNA
- Ülejäänud – lncRNA, eksperimentaalset kinnitust vajav RNA jms

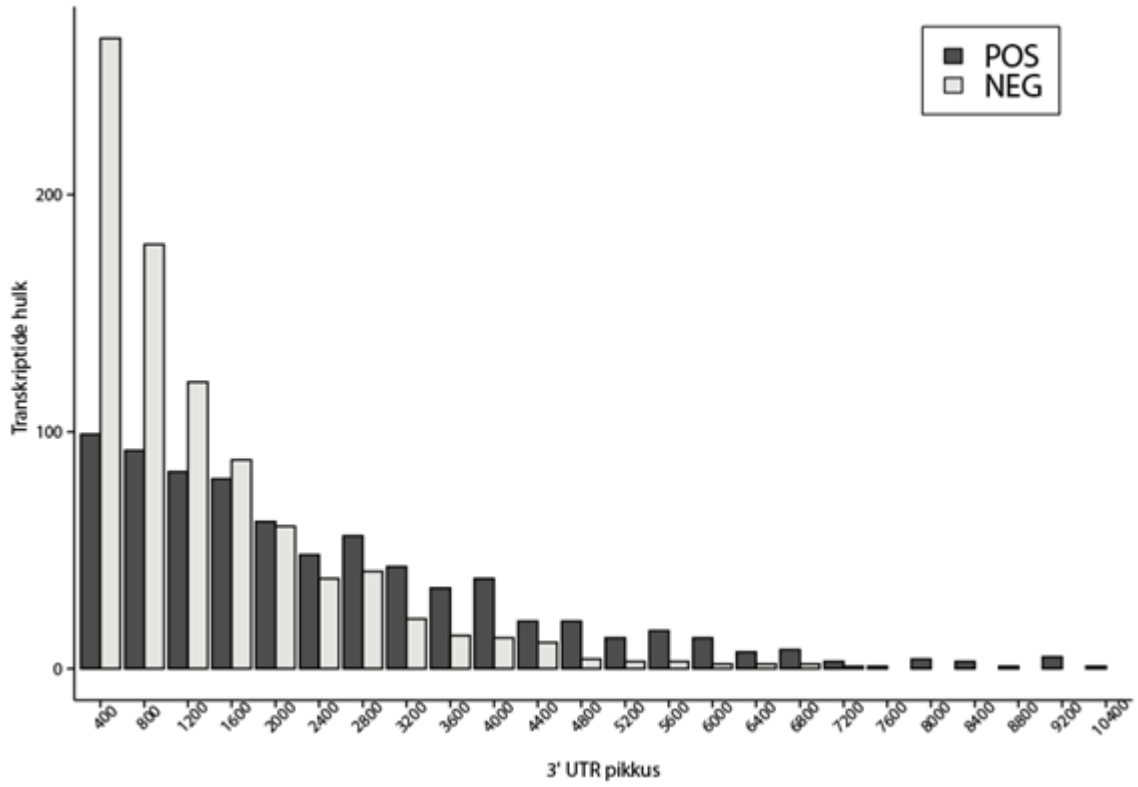
Geeni alternatiivsete transkriptide arvukuse jaotust iseloomustab joonis 16 A. Jaotuselt on näha, et ühe ainsa transkriptiga geenide hulk on suurem POS grupis. NEG grupi keskmine

alternatiivsete transkriptide arv on POS keskmisest suurem (vastavalt 8.048 ja 7.217, Wilcoxon-i testi p-väärtus = 0,0022)

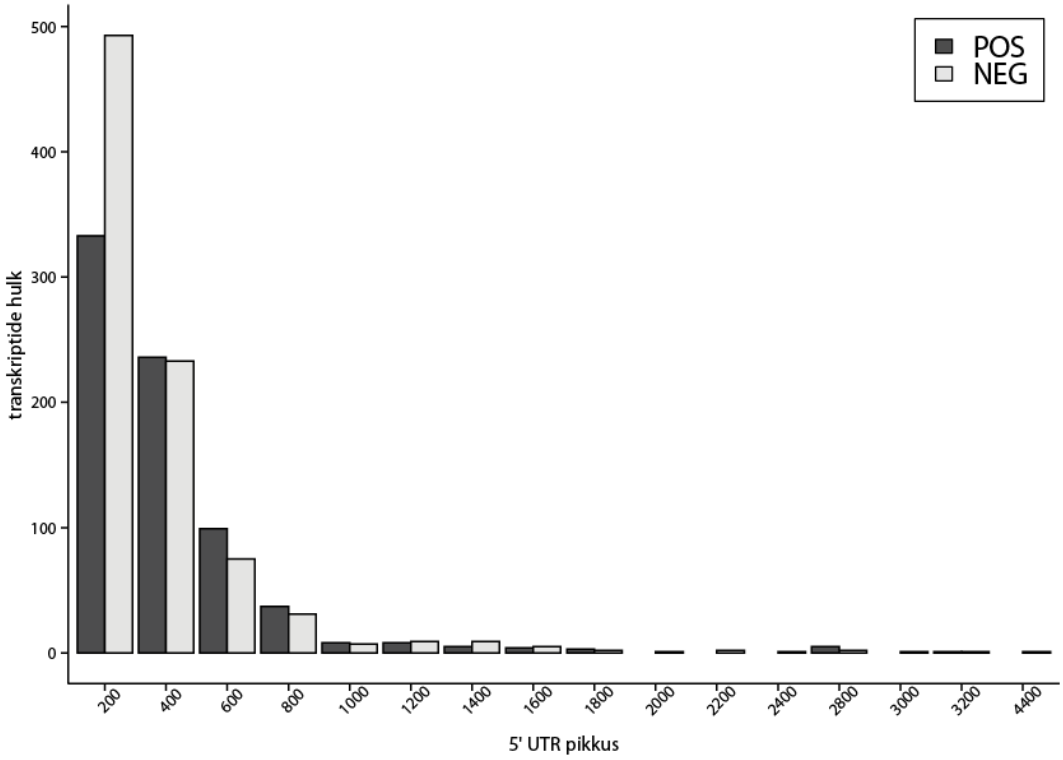


Joonis 16. A POS ja NEG gruppide alternatiivsete transkriptide arv, võetud EnsEMBL andmebaasist. RIN väärtusega positiivses korrelatsioonis olevate geenide (kergemini lagunevad transkriptid) grupis on suuremal hulgal ainult ühe transkriptiga geene, kui negatiivses korrelatsioonis olevate geenide hulgas. B Eksonite arv POS ja NEG grupi geenide mažoorsetes transkriptides. POS grupis on rohkem ühe eksonilisi transkripte ja transkripte, millel on rohkem kui kümme erinevat eksonit, samas kui NEG grupp on arvukamalt esindatud 2-10 eksonilises vahemikus.

3’ ja 5’ UTR-ide pikkuste jaotus on välja toodud joonistel 17 ja 18. Mõlemal korral on NEG grupi transkriptide hulgas lühemaid UTR regioone rohkem. 3’ UTR regioonide juures on see erinevus aga selgemalt eristuv. Keskmised 3’UTR pikkused NEG ja POS jaoks vastavalt 992,5 ja 1665, samas kui 5’ UTR-de keskmised pikkused on 236,0 ja 224,9.



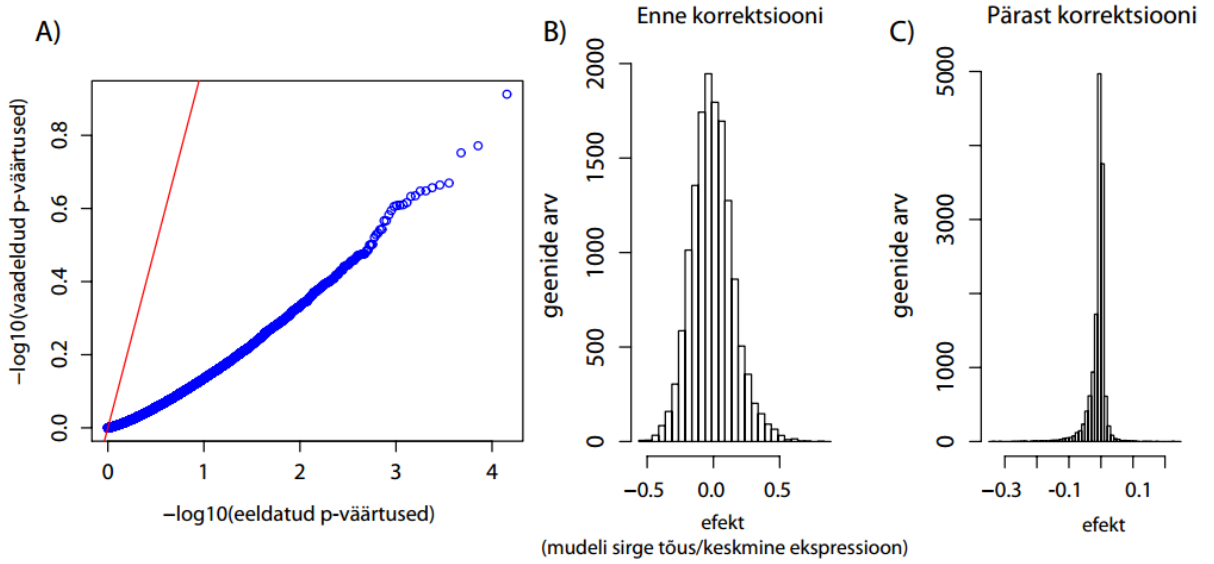
Joonis 17. 3’ UTR-i pikkuste jaotus POS ja NEG gruppide vahel. NEG grupi 3’UTR-id on tavaliselt lühemad kui POS grupi geenide korral.



Joonis 18. 5’ UTR-i pikkuste jaotus POS ja NEG gruppide vahel. Nagu 3’ UTR-i korralgi, on ka siin NEG grupis lühemaid transkripte rohkem (väiksemal määral).

4.3 Ekspresiooni korrekteerimine RNA intaktsuse suhtes

Korrektsiooni viisin läbi peatükis 4.3 kirjeldatud viisil 14673 geenile. Pärast geenide ekspressioonide korrigeerimist tegin kontrolliks uuesti peatükis 4.1 kirjeldatud viisil mudelid ja eemaldasin geenid mille puhul esines ülekorrektsiooni, mudelid ei konvergeerunud või mudeli AIC (Akaike informatiivsuse kriteerium) väärtus tuli keskmisest märgatavalt suurem (AIC>1000). Liigselt ülekorrekteerituks hindasin mudelit juhul, kui hilisema kontrollmudeli mudeli tõusuks hinnati esimesega võrreldes vastasuunalist tõusu, mis oli absoluutväärtuselt suurem kui ¾ esimese tõusust. Pärast filtratsiooni jäi alles 14288 korrigeeritud geeni (välja langes 385). Pärast korrigeerimist toimunud mudeldamise tulemusi kirjeldab joonis 21.



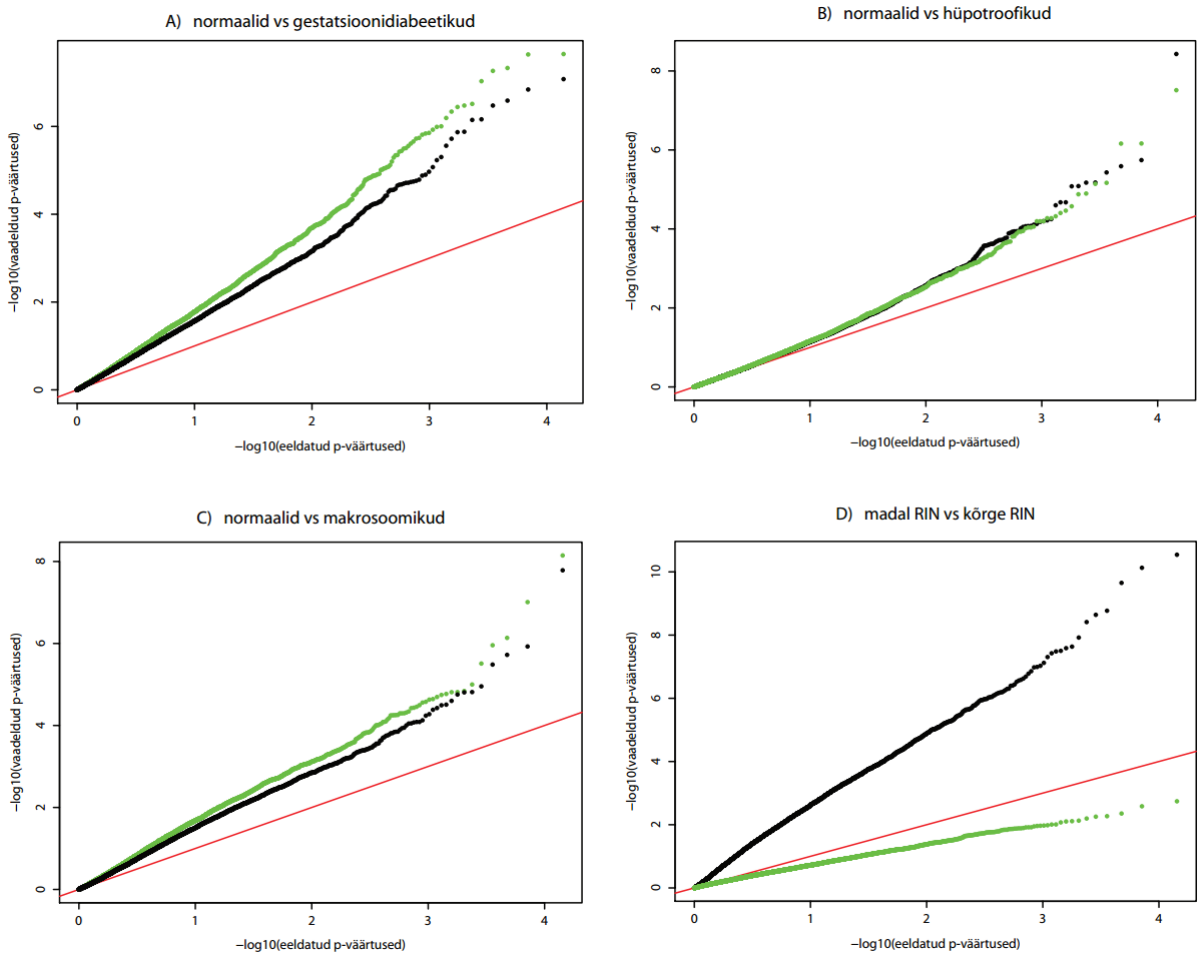
Joonis 19. Korrektsiooni tulemuste kontrollimine põhjal koostatud p-väärtuste Q-Q plot (A) ja histogrammid korrelatsioonimudelite tõusudest enne (B) ja peale korrektsiooni (C). A näitab geenide testimisel tulemuseks saadud p-väärtuste negatiivsete kümnendlogaritmide vaadeldud (y-telg) ja nullhüpoteesi kehtimise korral eeldatud (x-telg) jaotust. Nende logaritmide 1:1 suhet tähistab punane joon. Histogrammide B ja C järgi on näha, et pärast korrektsiooni on enamus korrelatsioonisirgete tõusudest muutunud märgatavalt väiksemaks.

Korrektsiooni tõttu vähenes proovide keskmine standardhälve 3,59% võrra, kuid tõusude standardhälve vähenes 83,0% võrra. Kõikide geenide puhul oli tõusude absoluutväärtus pärast korrektsiooni väiksem kui enne, kuid keskmine efekt nihkus negatiivse tõusu suunas ($\overline{efekt}_{enne\ korrektsiooni} = -0{,}0077$, $\sigma_{enne\ korrektsiooni} = 0{,}1573$ ja $\overline{efekt}_{pärast\ korrektsiooni} = -0{,}0101$, $\sigma_{pärast\ korrektsiooni} = 0.0267$), samas kui mediaan lähenes nullile ($efekti\ mediaan_{enne\ korrektsiooni} = -0{,}0156$, $efekti\ mediaan_{pärast\ korrektsiooni} = -0{,}0027$).

Et hinnata, kas sooritatud korrektsioon muudaks edaspidiseid teste täpsemaks, leidsin diferentsiaalselt ekspresseerunud geene minu uuringus kasutatud proovide kliiniliste gruppide vahel enne ja peale korrektsiooni ja kontrolli mõttes tegin sama analüüsi ka RIN-i põhjal kaheks jaotatud gruppide korral. Seega tegin DESeq2-ga neli kahe grupi võrdlust (normaalsünnid vs gestatsioonidiabeetikud, normaaalsünnid vs hüpotroofikud, normaalsünnid vs makrosoomikud ja kõrge RIN väärtusega vs madala RIN väärtusega proovid). Iga grupp koosnes kaheksast proovist, välja arvatud madala ja kõrge RIN grupid, mis koosnesid vastavalt 17 ja 15 proovist (32 proovi jagatud nende RIN väärtuse mediaani põhjal). Kui minu korrektsioon on tõepoolest mõttekas ja eeldusel, et nende gruppide vahel on suhteliselt väike arv tõelise diferentsiaalse ekspressiooniga geene, siis oleksin eeldanud näha Q-Q plottides (v.a. madal RIN vs kõrge RIN analüüsil) parema nurga juures suhteliselt suuremat tõusu kui ploti keskmises ja vasakus osas (joonis 20). Selle all mõtlen, et keskmine FDR väärtus diferentsiaalsiaalselt ekspresseerunuks tunnistatud (FDR<0.1) geenide vahel oleks peale korrektsiooni väiksem kui enne korrektsiooni. võrdlesin FDR väärtuste keskmisi wilcox'i testiga ja koondasin tulemused tabelisse 4.

Tabel 4. DESeq2-ga võrdlesin omavahel 4 grupi (normaalid, gestatsioonidiabeetikud, makrosoomikud ja hüpotroofikud) geeniekspressioone ja vaatasin, kas diferentsiaalselt ekspresseerunuks tunnistatud geenide keskmine FDR on pärast korrektsiooni madalam. Diferentsiaalselt ekspresseerunud geenid on osaliselt erinevad DESeq2-e analüüsi enne ja peale korrektsiooni variantides.

võrreldavad grupid	keskmine väiksem kui 0,1 FDR enne korrektsiooni	keskmine väiksem kui 0,1 FDR pärast korrektsiooni	wilcox'i testi p-väärtus
normaalid vs gestatsioonidiabeetikud	0.0620	0.0526	$3.033 * 10^{-6}$
normaalid vs makrosoomikud	0.0544	0.0527	0.3923
normaalid vs hüpotroofikud	0.0572	0.0757	1



Joonis 20. Q-Q plotid neljast kahe grupi võrdlusest enne (mustad täpid) ja peale (rohelised täpid) geenieksрessiooni korrigeerimist. Diferentsiaalse ekspressiooni leidmiseks tehti võrdlused järgnevate paaride vahel: A) normaalsed sünnitused vs gestatsioonidiabeetikud, B) normaalsed vs hüpotroofiad, C) normaalsed vs makrosoomikud, D) madala RIN väärtusega proovid vs kõrge RIN väärtusega proovid. A, B ja C gruppide suurused olid 8 proovi, D joonisel oli RIN väärtuse mediaani põhjal jaotatud kaheks 32 proovi (17 madalat vs 15 kõrget).

5. ARUTELU

5.1 RIN-i ja geeni ekspressiooni vahelise seose testimine

Regerssioonanalüüsi tulemusena leidsin, et suhteliselt suurel osal geenidel (1122 geeni 14288-st (7,8%), FDR<0,1) oli RIN väärtus oluliseks geeniekspressiooni mõjutajaks. Ka oli keskmine (mediaan) eeldatud p-väärtus vaadeldud p-väärtusest 2,34 korda suurem. Sarnase profiiliga Q-Q ploti (joonis 20 D) andis ka RIN-i mõju testimine DESeq2-s (RIN väärtuse alusel moodustatud kahe grupi võrdlemine), kuid siis tuli keskmine kõrvalekallutatus veel suurem (mediaan(eeldatud p-väärtused/vaadeldud p-väärtused)=3,50 ning 3863 geenil (27,0%) FDR<0,1).

See annab kindlust väitmaks, et proovide RNA intaktsus tõepoolest mõjutab RNA-seq eksperimendiga detekteeritud geeni ekspressiooni taset. Samale järeldusele olid jõudnud ka Popova jt (2008) ning Wan jt (2012), kuid nad mõlemad kasutasid korrelatsioonimudelis seletavaks tunnuseks transkripti eri osade suhtelist ekspressioonisignaali, mitte RIN väärtust. Popova artikli puhul saadi ekspressiooniandmed ka Affymetrix ekspressioonikiipidelt (HG-UI33_AB ja HG-uI33_Plus_2.0 kiibid), mitte RNA-seq andmetest. Wan jt artiklis sekveneeriti proovid Illumina sekveneerimisplatvormidel (HiSeq 2000 ja GA seadmetega), suurimaks erinevuseks minu materjalidega võrreldes on see, et nende uuringus kasutati rRNA eemaldamiseks polü-T oligonukleotiide (meil aga rRNA spetsiifilisi oligonukleotiide), mis tõenäoliselt ka suurendas degradatsiooni detektsiooniefekti (tänu 3’ *bias*-ile). Need erinevused võivad olla vägagi tähtsad, sest Wan ja Yan artikli tulemustes näidati, et degradatsioonikiiruse hinnang geeni jaoks sõltus sekveneerimisprotokollist rohkem kui kasutatud koest (inimese maks ja neer).

5.2 RIN-iga tugevaimat seost näidanud geenide omavaheline võrdlus

Kõikide tabelis 2 ja 3 välja toodud tunnuste testimise korral sain tulemuseks olulise erinevuse positiivses ja negatiivses korrelatsioonis olevate geenigruppide vahel. Seega järeldan, et kiireimini ja aeglaseimini lagunevate geenide hulka satuvad eelistatult erinevate bioloogiliste tunnustega transkriptid. Nende tunnuste juures on muidugi võimalik, et mõni neist on segavaks muutujaks, sest transkriptide keskmised pikkused on näiteks osalt seletatavad eksonite arvuga, mis omakorda seletab ka alternatiivsete transkriptide koguarvu. Peatükk 4.2 joonisele 12 olen koondanud mõndade vaadeldud tunnuste omavahelised korrelatsioonid, millest tugevaimalt korreleerunuks osutus transkripti eksonite arv ja

transkripti pikkus. See võib seletada ka seda, miks varasemad tööd (Wan jt, 2012, Yang jt, 2003) ei leidnud olulist seost transkripti pikkuse ja selle lagunemiskiiruse vahel, kuid minu tulemustes erines keskmine (ja ka mediaan) transkripti pikkus kiiresti ja aeglaselt lagunevate transkriptide vahel väga oluliselt. Samas vaadati nendes (Wan jt, 2012, Yang jt, 2003) uuringutes tegelikku transkripti lagunemiskiirust (pöördväärtus poolelueast), kuid mina grupeerisin proove p-väärtuse põhjal. Yang jt (2003) ning Wan jt (2012) testisid seost transkripti pikkuse ja selle lagunemiskiiruse vahel korrelatsioonimudelite abil, mitte kahe ekstreemse grupi võrdlusena. Viimases töös kasutati ka erinevaid mudeleid ühe transkriptiga geenide ja mitme transkriptiga geenide lagunemiskiiruste hindamiseks. Minu tulemus on vastuoluline varasemate uuringutulemustega ja kuigi siin on näha selge erinevus RIN väärtusega positiivses ja negatiivses korrelatsioonis olevate geenide transkriptide pikkuses, võib see vastuolu olla tingitud analüüsimeetodite erinevusest.

Minu tulemused 3’ UTR-ide pikkuste on sarnased Yang jt. (2003) töö tulemustega. Nimelt oli RIN väärtusega negatiivses korrelatsioonis (stabiilsemad, aeglasemalt lagunevad) olevate geenide hulgas UTR’ide, eriti märgatavalt aga 3’ UTR-i pikkused tavaliselt lühemad, kui positiivses korrelatsioonis olevate geenide grupis. Yang jt. (2003) leidsid, et mRNA-d mille 3’ UTR-id on pikemad kui 1 kb lagunevad märgatavalt suurema kiirusega. 5’ UTR pikkuse seost transkripti stabiilsusega on uuritud vähem (Hestand jt, 2010) ja selle tulemustes on kas kahtlustatud seose olemasolu (Hestand jt, 2010, Hoen jt, 2010, Hirata jt, 2004) või on kuulutatud see mitteoluliseks (Manful jt, 2011). Siinses eksperimendis oli 5’ UTR-i pikkuse efekt sarnane 3’ omalegi, kuid mitte nii tugevalt kui 3’ UTR-i korral. See tähendab, et kuigi lühemaid 5’ UTR-e on negatiivses korrelatsioonis olevate geenide hulgas rohkem, on neid suhteliselt suures hulgas ka positiivses korrelatsioonis olevate geenide hulgas ja selliste tulemuste replitseerumise tõenäosus ei ole nii suur.

Transkriptides olevate eksonite ja geeni alternatiivsete transkriptide arvu jaotuvust aeglaselt ja kiiresti lagunevate geenide hulgas kontrollisin, kuna eeldasin, et väheste eksonitega geenide hulgas on rohkem regulatoorseid RNA-sid, mille hulka peaks rakus olema võimalik kiiresti muuta. Seda on kerge teha nende lagundamisega. Paljude eksonitega ja alternatiivsete transkriptidega geenide stabiilsust võib aga vähendada NMD kontroll, sest iga splaiss-saidiga suureneks tõenäosust, et transkripti tekib mingi viga, mis viiks kaasa selle lagundamisele (Maquat, 2004). Jooniselt 16 on aga näha, et kiiremini lagunevad transkriptid on suuremas ülekaalus ühe eksoniga transkriptide ja ühe transkriptiga geenide suhtes. Sarnasele tulemusele jõudsid ka Hoen jt (2010), kuid nende uurimuses nähti, et kahe ja kolme eksoniga transkriptid olid enamasti kiiremini lagunevad. Seda minu andmetes näha ei olnud.

On võimalik, et erinevus võis koest põhjustatud olla (nad kasutasid lihaskoe rakke). Minu tulemuste põhjal oli kiiresti lagunevate geenide hulgas ka tõepoolest rohkem paljude eksonitega (rohkem kui 20) transkripte kui aeglaselt lagunevate geenide hulgas.

RNA liikide jaotusest RIN-iga positiivselt ja negatiivselt korreleerunud gruppide vahel nähtus, et kiiresti lagunevate geenide hulgas oli suuremal hulgal pseudogeene, pikkasid mittekodeerivaid transkripte ja antisense RNA-sid, mis täidavad rakus enamasti regulatoorset rolli ja ei ole seetõttu üllatav, et selliseid transkripte oleks kergem lagundada. Üllatav on aga see, et aeglaselt lagunevate geenide hulgas oli rohkem (mitte palju, kuid siiski) NMD-ga seonduvaid transkripte, kuigi eeldaks, et neid lagundatakse rakus väga aktiivselt (Baker ja Parker, 2004).

5.3 Mõõdetud geeniekspressiooni väärtuste korrektsioon RIN väärtuse suhtes

Korrektsiooni rakendasin 14673-st piisavalt kõrgelt ekspresseerunud geenist 14288-le (385, ehk ≈2,6% geenidest ei läbinud mudeli kvaliteedikontrolli). Et selle korrektsiooni kvaliteeti hinnata, testisin DESeq2-s omavahel 4 proovide gruppi (normaalid, gestatsioonidiabeetikud, makrosoomikud ja hüpotroofikud), mille vahel eeldasin, et suhteliselt väike osa geene oleks nende gruppide vahel diferentsiaalselt ekspresseerunud. Kuna ma ei saanud kindel olla, mis geenid päriselt nende gruppide vahel diferentsiaalselt ekspresseerunud on, võrdlesin kui kindlalt DESeq2 enne ja peale korrektsiooni osa geenidest erinevalt ekspresseerituks kuulutab. Selleks võrdlesin wilcox'i testiga, kui suur oli FDR<0,1 nivooga oluliseks kuulutatud geenide keskmine FDR väärtus enne ja pärast korrektsiooni. Sellise meetodiga võrreldes leidsin, et see väärtus oli oluliselt vähenenud vaid kolmest gruppide võrdlusest ühel korral, mistõttu sellisel korrektsiooni hindamise meetodil ei ole näha märgatavat diferentsiaalse ekspressiooni tuvastamise paranemist. Et kindlalt teada saada, kas minu korrektsioon töötab, tuleks korrigeerida proove, mille puhul tegelikult diferentsiaalselt ekspresseerunud geenid ette teada oleksid ja siis hinnata testi spetsiifilisust ja tundlikust. Kindlaks viisiks korrektsioonimeetodi hindamiseks on vaid tulemuste eksperimentaalne kontrollimine

KOKKUVÕTE

Selle magistritöö eesmärgiks oli kinnitada, kas RNA intaktsus mõjutas varieeruva RNA kvaliteediga proovide RNA-seq-iga leitud ekspressiooniprofiile. Lisaks uuriti ka mis tüüpi olid RNA intaktsuse mõjutamisele kõige vastupidavaimad ja kiireimini lagunevamad geenid.

Töö kirjalikus osas tutvustasin RNA lagunemise mehhanisme, RIN väärtuse põhimõtet ja missugune on RNA-seq eksperimendi soovitatav kulg halvema intaktsusega RNA proovide korral. Materjalides ja metoodikas kirjeldatakse, kuidas me RNA-seq analüüsi läbi viisime, mis moodi RIN väärtuse ja geenieksressiooni vahelist seost mudeldasime ja selle tulemusi edasi kasutasime.

Andmete analüüsist järeldus, et kuigi mudel ei sobinud kõikidele geenidele, oli suurel hulgal geenieksressioonidest siiski tugev seos RIN väärtusega (1122 geeni 14288-st olulises seoses (7,8%), FDR<0,1) ja selle seose suhtes läbis geene ühtlane trend tavapärasest väiksemaid p-väärtusi näidata.

RIN väärtusega olulisimas positiivses ja negatiivses korrelatsioonis olevate geenide annotatsioonid erinesid üksteisest märgatavalt. Seejuures on aeglasemalt lagunevate geenide hulgas lühemad transkripti pikkused ja ka nende normaliseeritud ekspressioonitase on madalam. Ka 5’ UTR ja eriti 3’ UTR-i pikkused on selles grupis väiksemad. Ühe eksoniga transkripte ja ühe transkriptiga geene on aga rohkem kiiresti lagunevate geenide hulgas.

Korrigeerisin proovide ekspressioone RIN väärtuse suhtes, kuid selle efektiivsuse testimisel andis DESeq2-ga diferentsiaalse ekspressiooni testimine oluliselt paremaid tulemusi vaid ühel korral kolme testi kohta ja seega ei saa kindlalt väita, kas korrigeerimisest praegusel kujul ka praktilist kasu on.

RNA integrity and RNA-seq expression data

Mario Reiman

SUMMARY

RNA-Seq is a powerful transcriptomics tool that quantitates transcripts by counting library fragments that aligned to transcribed regions of the reference genome. However, depending on the library making protocols and the integrity of the transcript to be sequenced, the number of reads obtained from that fragment can differ greatly.

We conducted Illumina (HiSeq 2000 platform) RNA-seq experiment on the transcriptomes of 56 human placentas. Both the extracted RNA and tissue samples had been stored for varying periods of time (up to 6 years) at -80^{o} C. I used 32 of those samples with less varying gene expression profiles but that displayed relatively wide range of RNA integrity (mean RIN=RNA Integrity Number=6.9, standard deviation=0.6)

In this study I am asking:

- what kind of an effect the sample's RIN value has on the gene expression quantified by Illumina Hiseq 2000 RNA-seq experiment.
- Additionally I am going to see if 1000 genes affected the most by this effect have something different compared to the 1000 that are affected the less.
- Can the effect RIN has on sample gene expression profile be corrected.

Even though the model did not fit for all of the genes, a large majority had stronger than expected correlation to RIN value and 1122 out of the 14288 tested genes showed very significant correlation (FDR<0.1).

1000 genes most affected by RIN and the 1000 least affected had several differences between their annotations – the most affected on average had longer transcript length, and higher average expression per the transcripts nucleotide, more single exon transcripts and more genes with only one alternative transcript. The genes affected the least tended to have shorter 3' and 5'UTR-s.

I corrected gene expressions for RNA integrity, but did not manage to conclusively prove the effectiveness of this correction.

KASUTATUD KIRJANDUSE LOETELU

Agilent, (2007). | Agilent Technologies Announces More Than One Million Bioanalyzer Chips Sold.

Anders, S. (2010). [BioC] Normalization by DEseq.

Anders, S., & Huber, W. (2012). Differential expression of RNA-Seq data at the gene level– the DESeq package. EMBL, Heidelberg, Germany.

Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biology 11, R106. doi:10.1186/gb-2010-11-10-r106

Baker, K.E., Parker, R., 2004. Nonsense-mediated mRNA decay: terminating erroneous gene expression. Curr. Opin. Cell Biol. 16, 293–299. doi:10.1016/j.ceb.2004.03.003

Belasco, J.G. (2010). All things must pass: contrasts and commonalities in eukaryotic and bacterial mRNA decay. Nat Rev Mol Cell Biol 11, 467–478.

Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M., and Caudy, A.A. (2009). Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics 10, 221.

Burd, C.E., Jeck, W.R., Liu, Y., Sanoff, H.K., Wang, Z., and Sharpless, N.E. (2010). Expression of Linear and Novel Circular Forms of an INK4/ARF-Associated Non-Coding RNA Correlates with Atherosclerosis Risk. PLoS Genet 6, e1001233

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Gall, C.L., Schaëffer, B., Crom, S.L., Guedj, M., Jaffrézic, F., 2012. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform bbs046. doi:10.1093/bib/bbs046

Duitama, J., Srivastava, P.K., and Măndoiu, I.I. (2012). Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. BMC Genomics 13, S6.

Fordyce, S.L., Kampmann, M.-L., Doorn, N.L. van, and Gilbert, M.T.P. (2013). Long-term RNA persistence in postmortem contexts. Investigative Genetics 4, 7.

Gaidatzis, D., Jacobeit, K., Oakeley, E.J., Stadler, M.B., 2009. Overestimation of alternative splicing caused by variable probe characteristics in exon arrays. Nucl. Acids Res. 37, e107–e107. doi:10.1093/nar/gkp508

Hestand, M.S., Klingenhoff, A., Scherf, M., Ariyurek, Y., Ramos, Y., Workum, W. van, Suzuki, M., Werner, T., Ommen, G.-J.B. van, Dunnen, J.T. den, Harbers, M., Hoen, P.A.C. 't, 2010. Tissue-specific transcript annotation and expression profiling with complementary next-generation sequencing technologies. Nucl. Acids Res. 38, e165–e165. doi:10.1093/nar/gkq602

Hirata, N., Yonekura, D., Yanagisawa, S., Iba, K., 2004. Possible involvement of the 5'-flanking region and the 5'UTR of plastid accD gene in NEP-dependent transcription. Plant Cell Physiol. 45, 176–186.

Hoen, P.A.C. 't, Hirsch, M., Meijer, E.J. de, Menezes, R.X. de, Ommen, G.B. van, Dunnen, J.T. den, 2010. mRNA degradation controls differentiation state-dependent differences in transcript and splice variant abundance. Nucl. Acids Res. gkq790. doi:10.1093/nar/gkq790

Houseley, J., and Tollervey, D. (2009). The Many Pathways of RNA Degradation. Cell 136, 763–776.

Karro, H., Rahu, M., Gornoi, K. Baburin, A. (1997). Sünnikaalu jaotumine raseduse kestuse järgi Eestis aastail 1992–1994. Eesti Arst, 76(4), 299 - 302.

Manful, T., Fadda, A., Clayton, C., 2011. The role of the 5′–3′ exoribonuclease XRNA in transcriptome-wide mRNA degradation. RNA 17, 2039–2047. doi:10.1261/rna.2837311

Maquat, L.E., 2004. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol 5, 89–99. doi:10.1038/nrm1310

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46.

Nardon, E., Donada, M., Bonin, S., Dotti, I., and Stanta, G. (2009). Higher random oligo concentration improves reverse transcription yield of cDNA from bioptic tissues and quantitative RT-PCR reliability. Experimental and Molecular Pathology 87, 146–151.

Perkins, J.R., Antunes-Martins, A., Calvo, M., Grist, J., Rust, W., Schmid, R., Hildebrandt, T., Kohl, M., Orengo, C., McMahon, S.B., Bennett, D.L., 2014. A comparison of RNA-seq

and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat. Molecular Pain 10, 7. doi:10.1186/1744-8069-10-7

Popova, T., Mennerich, D., Weith, A., and Quast, K. (2008). Effect of RNA quality on transcript intensity levels in microarray analysis of human post-mortem brain tissues. BMC Genomics 9, 91.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology 14, R95. doi:10.1186/gb-2013-14-9-r95

Reiman, M., 2013. Effect of RNA integrity on RNA-seq results, bachelor's thesis

Riedmaier, I., Bergmaier, M., and Pfaffl, M. (2010). Comparison of two Available Platforms for Determination of RNA Quality. Biotechnology & Biotechnological Equipment 24, 2154–2159.

Sorrentino, S. (2010). The eight human "canonical" ribonucleases: Molecular diversity, catalytic properties, and special biological actions of the enzyme proteins. FEBS Letters 584, 2194–2200.

Ståhlberg, A., Håkansson, J., Xian, X., Semb, H., and Kubista, M. (2004). Properties of the reverse transcription reaction in mRNA quantification. Clin. Chem. 50, 509–515.

Stangegaard, M., Dufva, I.H., and Dufva, M. (2006). Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. BioTechniques 40, 649–657.

Stuart, C.A., Wen, G., and Jiang, J. (1999). GLUT3 protein and mRNA in autopsy muscle specimens. Metabolism 48, 876–880.

Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A., 2011. Differential expression in RNA-seq: A matter of depth. Genome Res 21, 2213–2223. doi:10.1101/gr.124321.111

Thorn, A., Steinfeld, R., Ziegenbein, M., Grapp, M., Hsiao, H.-H., Urlaub, H., Sheldrick, G.M., Gärtner, J., and Krätzner, R. (2012). Structure and activity of the only human RNase T2. Nucl. Acids Res. 40, 8733–8742.

Tufarelli, C., Stanley, J.A.S., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., Higgs, D.R., 2003. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat. Genet. 34, 157–165. doi:10.1038/ng1157

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Wan, L., Yan, X., Chen, T., Sun, F., 2012. Modeling RNA degradation for RNA-Seq with applications. Biostat 13, 734–747. doi:10.1093/biostatistics/kxs001

Yang, E., Nimwegen, E. van, Zavolan, M., Rajewsky, N., Schroeder, M., Magnasco, M., Darnell, J.E., 2003. Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes. Genome Res. 13, 1863–1872. doi:10.1101/gr.1272403

KASUTATUD VEEBIAADRESSIDE LOETELU

1) http://www.roadmapepigenomics.org/files/protocols/data/rna-analysis/REMC_RNA-seqStandards_final.pdf

2) http://www.invitrogen.com/site/us/en/home/References/Ambion-Tech-Support/rna-isolation/tech-notes/is-your-rna-intact.html

3) http://www.promega.ee/resources/articles/pubhub/methods-of-rna-quality-assessment/?__utma=1.793970916.1367425961.1367425961.1367425961.1&__utmb=1.1.10.1367425961&__utmc=1&__utmx=-&__utmz=1.1367425961.1.1.utmcsr=google|utmccn=(organic)|utmcmd=organic|utmctr=(not%20provided)&__utmv=-&__utmk=13777463

4) http://www.chem.agilent.com/Library/applications/5989-1165EN.pdf

5) http://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq/inst/doc/DESeq.pdf

6) http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html

LISAD

Tabel 5. Sekveneeritud proovid, nende lugemite pikkused, hulk ja proovile vastavad RIN väärtused

Proov	Jooksu nr	Rada	RIN	Lugemite pikkus	Lugemite arv	Lugemite arv peale filtreerimist
LGA1	1	8	6,8	101	149231230	140015044
SGA1	2	1	6,8	46	86744868	76439428
SGA2	2	1	6,3	46	82683582	76499812
SGA3	2	1	6,8	46	87784390	83789654
LGA2	2	2	7,2	46	73274674	69666348
LGA3	2	2	8,2	46	63767878	60821022
LGA4	2	3	7,4	46	97634818	89628758
LGA5	2	3	7,1	46	72765330	67658656
GD1	2	3	6,3	46	89088710	82540346
LGA6	2	4	6,2	46	103972774	92248112
SGA4	2	5	7,7	46	60164112	57733278
GD2	2	5	6,9	46	96531304	88021416
GD3	2	5	7,2	46	88985914	80257132
GD4	2	6	6,5	46	76479722	72988776
GD5	2	8	6,8	46	87851394	82178022
N1	3	1	7,4	46	60971164	56231264
N2	3	1	6,3	46	72651606	66294416
N3	3	1	8,2	46	56475136	51037306
N4	3	2	6,3	46	69016032	60141782
SGA5	3	2	6,3	46	94033172	82061160
SGA6	3	2	6,4	46	104254734	96924966
LGA7	3	2	6,4	46	88832480	74354670
LGA8	3	3	7,5	46	73282050	67294474
GD6	3	4	6,9	46	64836844	61685892
GD7	3	4	7,3	46	75534046	67901954
GD8	3	4	6,1	46	54516914	50405430
N5	3	7	7,2	46	96996860	85055858
N6	3	7	6,5	46	79871872	70663846
N7	3	7	6,3	46	98147796	88772594
N8	4	4	7,0	46	95516334	85978328
SGA7	4	4	6,3	46	84524868	79133182
SGA8	4	4	7,1	46	85871532	79341688
KESKMINE			6,9		83’509’192	76’367’644

LIHTLITSENTS LÕPUTÖÖ REPRODUTSEERIMISEKS JA LÕPUTÖÖ ÜLDSUSELE KÄTTESAADAVAKS TEGEMISEKS

Mina Mario Reiman (sünnikuupäev: 20. 02. 1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
„RNA intaktsus ja RNA-seq ekspressiooniandmed" mille juhendaja on Siim Sõber, PhD

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 alates 26.05.2015 kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 26. 05. 2013