TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOTEHNOLOOGIA ÕPPETOOL

Meditsiiniliselt oluliste geenivariantide tõlgendamine RNA ekspressiooniandmete abil

Magistritöö (30 EAP)

Ustina Põnova

Juhendajad: PhD Tarmo Annilo

MSc Viktorija Kukuškina

TARTU 2017

Infoleht

Meditsiiniliselt oluliste geenivariantide tõlgendamine RNA ekspressiooniandmete abil

Kogu transkriptoomi sekveneerimine (RNA-Seq) võimaldab tuvastada patogeensete variantide toimet mRNA avaldumisele ning struktuurile. Käesoleva töö eesmärgiks oli uurida meditsiiniliselt oluliste geenivariantide mõju transkriptoomi tasemel. Selleks viidi läbi RNA-Seq andmete analüüs uuritavate geenide ekspressioonitaseme, splaissingumustri ja alleel-spetsiifilise ekspressiooni tuvastamiseks. Uuringusse sai valitud 21 indiviidi, kellel esinesid mutatsioonid 13 erinevas geenis. Kõikidel uuritavatel mutatsioonidel oli potentsiaalne mõju transkriptoomile, kuna nende hulgas olid raaminihke, enneaegse stoppkoodoni ja splaissingusaitide mutatsioonid. Ekspressioonianalüüsi tulemuste alusel valiti järgnevasse splaissinguanalüüsi 17 indiviidi. Viiel indiviidil tuvastati alternatiivse splaissingu sündmused, mille hulgas olid eksoni vahelejätmine, introni säilitamine ja alternatiivsete splaissingusaitide kasutamine. Alleel-spetsiifilist ekspressiooni (ASE) uuritavatel indiviididel ei tuvastatud.

Märksõnad: RNA-Seq, alternatiivne splaissing

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika

The interpretation of clinically relevant genetic variants through RNA-Seq expression data analysis

RNA sequencing (RNA-Seq) is a powerful tool for detecting the impact of clinically relevant genetic variants on the transcript level. The purpose of the current Master's thesis was to explore the impact of clinically relevant genetic variants on the transcript level. The transcript expression levels were measured for 21 gene donors from Estonian Genome Center, who had mutations in thirteen different genes. All the studied mutations (frameshift, premature stop codon, splice site mutations) could potentially have an impact on the transcript level. Based on the results of expression analysis 17 individuals were selected for further splicing study. Alternative splicing events were found in five genes. Among detected alternative splicing types were exon skipping, intron retention and alternative splice site usage. Allele-specific gene expression (ASE) was not detected.

Keywords: RNA-Seq, alternative splicing

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics

Sisukord

KASUTATUD LÜHENDID

ACMG	Ameerika Meditsiinigeneetika ja -genoomika Kolleegium (*American College of Medical Genetics and Genomics*)
ASE	alleel-spetsiifiline ekspression (*allele-specific expression*)
BBP	harusaidiga seonduv valk (*branch-point binding protein*)
ESE	eksonis paiknev splaissingu võimendaja (*exonic splicing enhancer*)
ESS	eksonis paiknev splaissingu vaigistaja (*exonic splicing silencer*)
FPKM	lugemite arv jagatud geeni pikkusega (tuhandetes nukleotiidides) jagatud lugemite koguarvuga (miljonites) (*fragments per kilobase of transcript per million mapped reads*)
GLM	üldistatud lineaarne mudel (*generalized linear model*)
hnRNP	heterogeenne ribonukleoproteiin (*heterogenous ribonuclear protein*)
ISE	intronis paiknev splaissingu võimendaja (*intronic splicing enhancer*)
ISS	intronis paiknev splaissingu vaigistaja (*intronic splicing silencer*)
KMI	kehamassiindeks
PPT	polüpürimidiinjärjestus (*polypyrimidine tract*)
pre-mRNA	eelas-mRNA (*precursor mRNA*)
PRPF	pre-mRNA protsessimise faktor (*pre-mRNA processing factor*)
RNA-Seq	kogu transkriptoomi järjestamine, RNA sekveneerimine (*RNA sequencing*)
SNP	üksiknukleotiidne polümorfism (*single nucleotide polymorphism*)
snRNA	väike tuuma-RNA (*small nuclear RNA*)
snRNP	väike tuuma ribonukleoproteiin (*small nuclear ribonucleoprotein*)
U2AF	U2 snRNP abifaktor (*U2 auxillary factor*)
UTR	mittekodeeriv regioon (*untranslated region*)
WGS	kogu genoomi sekveneerimine *(whole genome sequencing)*

SISSEJUHATUS

Inimese Genoomi Projekti (*The Human Genome Project*) andmeil on inimese genoomis ligikaudu 20 000 kuni 25 000 valkukodeerivat geeni (Consortium IHGS, 2004). Ühes suuremas proteoomi uuringus *The Human Proteome Map* analüüsiti ligi 17 300 geeni ja leiti, et need võivad kodeerida üle 30 000 erineva valgu (Kim jt., 2014). Selline valkude mitmekesisus on tingitud asjaolust, et ühe geeni põhjal on võimalik sünteesida valgu erinevaid isovorme. Alternatiivne splaissing võimaldab pre-mRNA protsessimise tagajärjel saada algselt transkriptilt kuni mitusada erinevat isovormi (nt geenis *CD44* (Bánky jt., 2012)).

Hoolimata edusammudest haruldaste haiguste diagnoosimisel kogu genoomi või eksoomi sekveneerimisega leitakse genoomipõhise lähenemise korral haigust tekitavaid mutatsioone siiski ainult umbes ühel kolmandikul juhtudest (Cummings jt., 2017). Oluliseks kitsaskohaks on piiratud oskus suure hulga leitud geneetiliste variantide korral nende patogeensuse hindamine. Üheks võimalikuks lähenemiseks oleks kogu transkriptoomi sekveneerimine, mis võimaldab tuvastada patogeensete variantide toimet mRNA avaldumisele ning struktuurile.

RNA splaissing on raku seisukohast oluline protsess transkriptsiooni ja translatsiooni vahel. Korrektne RNA splaissing (sh alternatiivne splaissing) tagab korrektse valgujärjestuse. Ligikaudu 95% inimese geenidest avaldub alternatiivse splaissingu kaudu (Wang jt., 2008; Pan jt., 2008). Mutatsioonid splaissingu mehhanismis või regulatsioonis võivad kaudselt või otseselt põhjustada haiguse teket. On leitud, et kuni 50% mutatsioone haigusseoselistes geenides mõjutavad ka splaissingut (Wang ja Cooper, 2007). RNA splaissingu transkriptoomipõhine analüüs on saanud võimalikuks tänu kiiresti arenevale teise põlvkonna sekveneerimistehnoloogiale.

Käesoleva töö eesmärgiks oli uurida meditsiiniliselt oluliste geenivariantide mõju transkriptoomi tasemel. Selleks viidi läbi RNA-Seq andmete analüüs uuritavate geenide ekspressioonitaseme, splaissingumustri ja alleel-spetsiifilise ekspressiooni tuvastamiseks.

1. KIRJANDUSE ÜLEVAADE

Kirjanduse ülevaade keskendub RNA splaissingu mehhanismi ja regulatsiooni kirjeldamisele. Samuti tutvustatakse alternatiivset splaissingut ning antakse lühiülevaade splaissingu häirumisest põhjustatud haigustest. Teoreetilise osa lõpus kirjeldatakse alternatiivset splaissingut uurivaid programme, mida kasutati käesoleva uurimustöö raames.

1.1. RNA SPLAISSING

RNA splaissing on protsess, mille käigus esmasest RNA transkriptist (pre-mRNA) eemaldatakse introonsed järjestused ja liidetakse kokku allesjäänud eksonid. Splaissing algab vahetult pärast pre-mRNA sünteesi polümeraasi poolt tuumas (Kelemen jt., 2013). Splaissingu viib läbi splaissosoom, mis kujutab endast valkudest ja snRNA-dest (*small nuclear RNA*) koosnevat makromolekulaarset kompleksi. snRNA-d on lühikesed, kuni 200 nukleotiidi pikkused, RNA molekulid. Splaissosoomi kompleksis leidub 5 erinevat snRNA tüüpi: U1, U2, U4, U5 ja U6 (Alberts jt., 2015). Iga snRNA-ga on seondunud vähemalt 7 erinevat valku, mis moodustavad koos snRNP (*small nuclear ribonucleoprotein*) (Alberts jt., 2015). Erinevad snRNP-d moodustavad omakorda splaissosoomi kompleksi. Lisaks sellele osaleb splaissingus üle 150 erineva valgu (Jurica ja Moore, 2003).

Selleks, et splaissing saaks toimuda, peavad splaissosoomi snRNA-d ära tundma pre-mRNA järjestuses kolm spetsiifilist regiooni: 5’ ja 3’ splaissingusaidid ning harusaidi (*branch site*) (Alberts jt., 2015). Introni 5' otsas paiknev GU dinukleotiid on vajalik splaissingu toimumiseks (Cieply ja Carstens, 2015). 3’ splaissingusait asub intron-ekson piiril ning koosneb pürimidiinirikkast järjestusest (*polypyrimidine tract*, PPT) ja splaissingusaidi lõpus asuvast AG dinukleotiidist, mis on korrektse splaissingu jaoks väga oluline (Cieply ja Carstens, 2015). Harusait on spetsiifiline regioon, mis paikneb ligi 40 nukleotiidi kaugusel 3’ splaissingusaidist (Kelemen jt., 2013).

Splaissosoomi üleminek inaktiivsest olekust katalüütiliselt aktiivsesse vormi hõlmab mitmeid snRNP ja RNA ümberkorraldusi (Kornblihtt jt., 2013). Splaissingu jooksul läbib splaissosoom viis erinevat staadiumit (Joonis 1). Splaissing algab splaissosoomi E kompleksi kokkupanemisega (Suñé-Pou jt., 2017). U1 snRNP seondub 5’ splaissingusaidiga (Alberts jt., 2015). BBP (*branch-point binding protein*) valk seondub harusaidiga (Alberts jt., 2015).

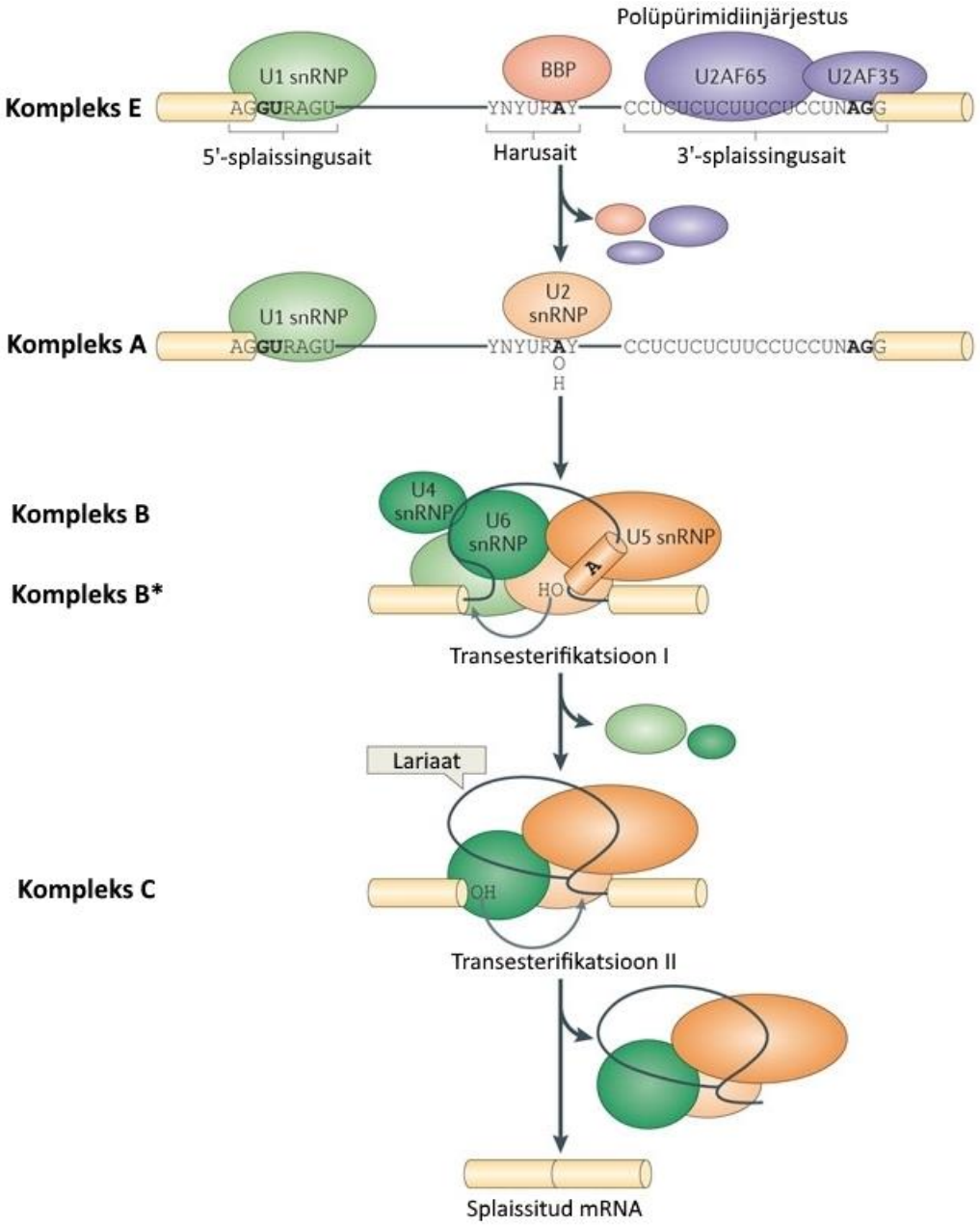
U2AF (*U2 auxillary factor*) valgud tunnevad ära 3’ splaissingusaidi konsensusjärjestuse ja polüpürimidiinjärjestuse (Kornblihtt jt., 2013).

BBP ja U2AF valgud juhivad U2 snRNP harusaidile (Alberts jt., 2015), moodustades splaissosoomi A kompleksi (Suñé-Pou jt., 2017). Pärast U2 snRNP seondumist eemaldatakse BBP ja U2AF valgud pre-mRNA-st (Kornblihtt jt., 2013). pre-mRNA splaissingu kompleksiga liitub U4, U5 ja U6 snRNP-de kolmikkompleks (Alberts jt., 2015). Moodustub splaissosoomi B kompleks (Suñé-Pou jt., 2017).

Introni eemaldamine toimub kahe transesterifitseerimise reaktsiooni jooksul (Kornblihtt jt., 2013). Vahetult enne esimest transesterifitseerimise reaktsiooni tekib splaissosoomi aktiveeritud B* kompleks (Suñé-Pou jt., 2017). Splaissingu esimese transesterifitseerimise reaktsiooni käigus atakeerib harusaidis oleva adenosiinijäägi nukleofiilne 2’OH rühm 5’ splaissingusaiti, mille tulemusena hakkab moodustama lariaat (Kornblihtt jt., 2013). Transesterifitseerimise reaktsiooni käigus toimub ka U4 ja U6 snRNP-de omavaheline lahutamine kolmikkompleksi sees, mille tagajärjel tekib U6 snRNP-sse aktiivsait (Alberts jt., 2015). U1 ja U4 snRNP-d eemaldatakse splaissosoomi kompleksist ning moodustub splaissosoomi C kompleks (Suñé-Pou jt., 2017). Aktiivsaidiga U6 snRNP initsieerib RNA-RNA ümberkorralduse ja teise transesterifitseerimise reaktsiooni. Selle käigus atakeerib 5' suunas paikneva eksoni vaba 3’OH rühm 3’ splaissingusaiti ja toimub introni lõplik väljalõikamine (Alberts jt., 2015; Kornblihtt jt., 2013). Eemaldatud intronjärjestus lariaadi kujul suunatakse degradatsioonile. Lariaadiga seondunud snRNP-d kasutatakse uuesti järgmistes splaissingu reaktsioonides (Alberts jt., 2015). pre-mRNA 3’ ja 5’ otsad ligeeritakse kokku.

1.2. ALTERNATIIVNE SPLAISSING

Alternatiivseks splaissinguks nimetatakse sellist pre-mRNA splaissingut, mille käigus ühest pre-mRNA järjestusest on võimalik saada mitu erinevat mRNA isovormi (Stamm jt., 2013). Alternatiivne splaissing laiendab genoomi kodeerimisvõimalusi ja suurendab mRNA-de mitmekesisust ning võimaldab ühest geenijärjestusest toota mitu erinevat valguvarianti (Kornblihtt jt., 2013; Kelemen jt., 2013). Saadud valgu isovormid võivad erineda omavahel funktsiooni, rakusisese lokalisatsiooni, valk-valk interaktsiooni ja posttranslatsiooniliste modifikatsioonide poolest (Kelemen jt., 2013).



Joonis 1. pre-mRNA splaissingu mehhanism. Splaissosoomi E kompleks: U1 snRNP seondub 5’ splaissingusaidiga. BBP seondub harusaidiga. U2AF valgud tunnevad ära 3’ splaissingusaidi ja polüpürimidiinjärjestuse. Splaissosoomi A kompleks: BBP ja U2AF valgud juhivad U2 snRNP harusaidile ning BBP ja U2AF valgud eemaldatakse pre-mRNA-st. Splaissosoomi kompleks B: splaissosoomi kompleksiga liitub U4, U5 ja U6 snRNP-de kolmikkompleks. Splaissosoomi kompleks B*: toimub esimene transesterifitseerimise reaktsioon, hakkab moodustuma lariaat. U4 ja U6 snRNPd lahutatakse omavahel kolmikkompleksi sees. U6 snRNP-s tekib aktiivsait. Splaissosoomi kompleks C: toimub teine transesterifitseerimise reaktsioon ja introni lõplik väljalõikamine. pre-mRNA 3’ ja 5’ otsad ligeeritakse kokku. (Kornblihtt jt., 2013, kohandatud).

On teada, et inimesel avaldub ligi 95% geenidest alternatiivse splaissingu kaudu (Wang jt., 2008; Pan jt., 2008). Seega sisaldab enamik pre-mRNA-sid alternatiivseid eksoneid, mis võivad ühel mRNA isovormil esineda, teisel aga puududa (Kelemen jt., 2013).

Pre-mRNA-l asuvad eksonid võib jagada kahte rühma: konstitutiivsed ja alternatiivsed eksonid. Konstitutiivsed eksonid on "tugevad" eksonid ja esinevad kõikides vastava mRNA isovormides. Alternatiivsed eksonid on aluseks alternatiivsele splaissingule ning sellised eksonid omavad nõrgemaid äratundmissaite (Kelemen jt., 2013).

1.2.1. Alternatiivse splaissingu tüübid

On olemas neli peamist alternatiivse splaissingu tüüpi: eksoni vahelejätmine, introni säilitamine, alternatiivne 3’ splaissingusait ehk alternatiivne aktseptorsait ja alternatiivne 5’ splaissingusait ehk alternatiivne doonorsait (Joonis 2).

Eksoni vahelejätmine (*exon skipping*)

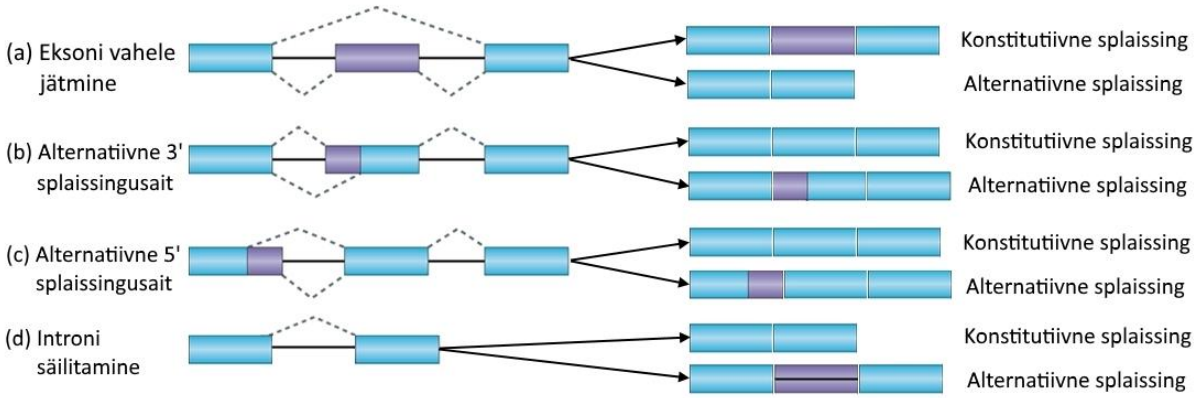
Enamik mutatsioone splaissingu konsensusmotiivides (harusait, 5’ ja 3’ splaissingusaidid) põhjustavad eksoni vahelejätmist, mis on kõige levinum alternatiivse splaissingu tüüp (Krawczak jt., 2007).

Introni säilitamine (*intron retention*)

Introni säilitamise puhul ei lõigata intronit splaissingu käigus splaissosoomi poolt välja ja see jääb valmis mRNA transkripti koosseisu. Enamikel juhtudel leiab introni säilitamine aset mRNA mittekodeerivates regioonides (*untranslated region,* UTR). Juhul, kui intron säilitatakse valkukodeerivas alas, põhjustab see suure tõenäosusega lugemisraami nihke või enneaegse stoppkoodoni tekke. Juhul, kui sellist mRNA-d ei lagundata, võib tagajärjeks olla vigase valgu süntees ja rakuliste protsesside häirumine (Galante jt., 2004).

Alternatiivsed 5’ ja 3’ splaissingusaidid

Mutatsioonid kanoonilistes 3’ või 5’ splaissingusaitides võivad tingida selle, et splaissosoom kasutab alternatiivseid ehk krüptilisi splaissingusaite (Stamm jt., 2012).



Joonis 2. Peamised alternatiivse splaissingu tüübid: eksoni vahelejätmine, alternatiivne 3’ splaissingusait ehk alternatiivne aktseptorsait, alternatiivne 5’ splaissingusait ehk alternatiivne doonorsait ja introni säilitamine. (Keren jt., 2010, kohandatud).

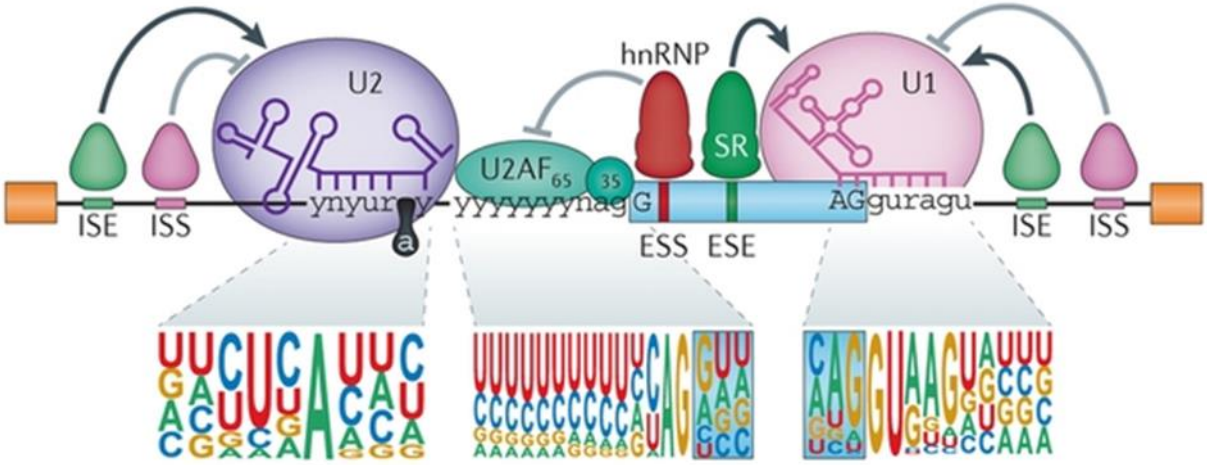
Lisaks ülalnimetatutele on olemas veel teisi, vähem levinud alternatiivse splaissingu tüüpe, nagu näiteks alternatiivne esimene ekson, alternatiivne viimane ekson, teineteist välistavad eksonid või tandeemsed 3’ UTR regioonid (tandeemne polüadenülatsioon).

1.3. SPLAISSINGU REGULATSIOON

Splaissingu korrektse toimumise ning splaissingumustri vastavalt rakutüübile tagavad erinevad splaissingu regulatsioonimehhanismid.

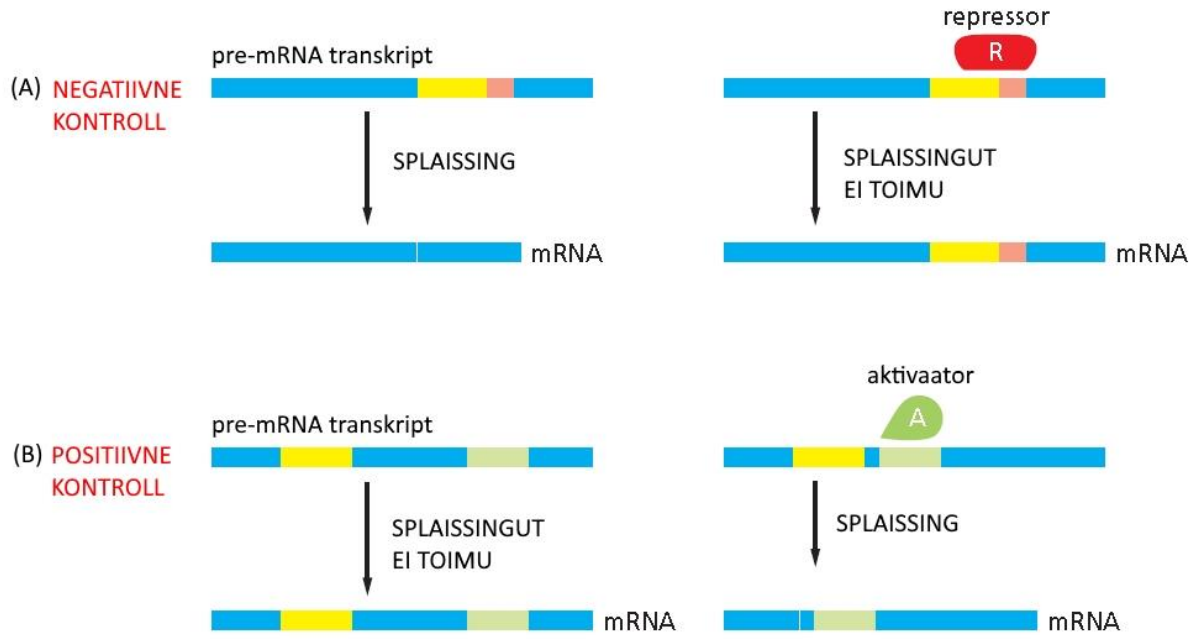
RNA splaissingu peamised regulatoorsed elemendid on eksonis paiknevad võimendajad (*exonic splicing enhancers,* ESE) ja vaigistajad (*exonic splicing silencers,* ESS) ning intronis paiknevad võimendajad (*intronic splicing enhancers*, ISE) ja vaigistajad (*intronic splicing silencers,* ISS) (Joonis 3). Ülalmainitud splaissingu regulaatorid kuuluvad *cis*-regulatoorsete elementide rühma ja on vajalikud splaissingusaidi äratundmiseks ning alternatiivsete transkripti isovormide regulatsiooniks (Cieply ja Carstens, 2015). Splaissingu võimendajad soodustavad eksoni lülitumist mRNA-sse (Kelemen jt., 2013). Võimendajate aktiveerimiseks peavad nendele seonduma spetsiifilised splaissingu abifaktorid, näiteks SR-valgud (Kelemen jt., 2013). Samuti aitavad SR-valgud kaasa U1 ja U2 sRNP-de korrektsele seondumisele (Alberts jt., 2015). Splaissingu vaigistajate peamiseks funktsiooniks on eksoni tuvastamise pärssimine splaissosoomi poolt.

Splaissingu vaigistamine aktiveeritakse samuti spetsiifiliste abifaktorite seondumisega. Tuntumad vaigistamisfaktorid on hnRNP-d (*heterogenous ribonuclear protein*) (Kelemen jt., 2013). Splaissingu vaigistamis- ja abifaktorid kuuluvad *trans*-regulatoorsete elementide hulka.



Joonis 3. Alternatiivse splaissingu regulatsioon. Joonisel on toodud kõik peamised *cis*- ja *trans*-regulatoorsed elemendid: 3’ ja 5’ splaissingusaidid, harusait, eksonis paiknevad võimendajad (ESE) ja vaigistajad (ESS), intronis paiknevad võimendajad (ISE) ja vaigistajad (ISS), SR-valgud, hnRNP-d. Joonise alumises osas on toodud harusait, 3’ ja 5’ splaissingusaidid ning konsensusjärjestused konserveerunud positsioonidega: adenosiin (A) harusaidis, AG dinukleotiid 3’ splaissingusaidis ja GU dinukleotiid 5’ splaissingusaidis. Regulatoorsed splaissingu võimendajad (ESE, ISE) soodustavad splaissosoomi seondumist splaissingusaidile. Splaissingu vaigistajad (ESS, ISS) aga pärsivad splaissosoomi komponentide seondumist splaissingusaitidele. Võimendajad aktiveeritakse SR-valkude seondumisega ning vaigistajad aktiveeritakse hnRNP abivalkude seondumisega. Splaissosoomi U2 snRNP seondub harusaidile. U2AF valk tunneb ära 3’ splaissingusaidi polüpürimidiinjärjestuse. U1 snRNP seondub 5’ splaissingusaidile. (Scotti ja Swanson, 2016, kohandatud).

RNA splaissingu regulatsioon võib olla nii positiivne kui ka negatiivne. Positiivse regulatsiooni korral vajab splaissosoom intronjärjestuse väljalõikamiseks aktivaatorvalku, mis seonduks splaissingu võimendajate (ESE, ISE) piirkonda. Negatiivse regulatsiooni korral seondub repressorvalk splaissingusaidile pre-mRNA-l, takistades splaissosoomi seondumist selle piirkonnaga. Splaissingu negatiivne regulatsioon viib alternatiivsete splaissingu variantide tekkimiseni (Alberts jt., 2015).



Joonis 4. Splaissingu negatiivne ja positiivne regulatsioon. Negatiivne kontroll (A): repressorvalk (punane) seondub pre-mRNA splaissingusaidile (roosa), takistades splaissosoomi seondumist sellele piirkonnale. Positiivne kontroll (B): Introni (kollane) väljalõikamine splaissosoomi poolt on võimalik ainult aktivaatorvalgu (roheline) olemasolul. Splaissingu toimumiseks peab aktivaatorvalk seonduma splaissingu võimendaja (heleroheline) piirkonda. (Alberts jt., 2015, kohandatud).

Splaissingu regulatsioon on tihedalt seotud transkriptsiooniga (Kelemen jt., 2013). Välja on pakutud kaks mudelit, kuidas transkriptsioon võib splaissingut mõjutada: värbamismudel (*recruitment model*) (Das jt., 2006, Kelemen jt., 2013) ja kineetiline mudel (*kinetic model*) (Kornblihtt, 2006). Värbamismudeli järgi kogunevad splaissingu faktorid RNA polümeraas II karboksüterminaalsele domäänile ning seonduvad valmivale pre-mRNA-le kohe pärast selle väljumist transkriptsiooni kompleksist (Das jt., 2006, Kelemen jt., 2013).

Teatud splaissingusaidi kasutamine sõltub selle saidiga seotud splaissingu faktorite hulgast. Kineetilise mudeli järgi vajavad splaissingu regulaatorvalgud rohkem aega, et eksonit ära tunda ja seonduda splaissingusaidiga. Histoonide modifikatsioonid ja nukleosoomide tihe paiknemine teatud DNA regioonis aeglustavad polümeraasi tööd, võimaldades splaissingu faktoritel seonduda splaissingusaidiga ja soodustades selle regiooni kaasamist valmivale mRNA-le (Kornblihtt, 2006).

Splaissingu regulatsioon on tugevalt koespetsiifline, tihtipeale ka rakutüübispetsiifiline protsess. Lisaks standardsetele regulatoorsetele elementidele reguleerivad splaissingut ka koespetsiifilised abifaktorid (Cieply ja Carstens, 2015). Lõplik splaissingu tulemus sõltub paljudest regulatoorsetest elementidest. Nende koosmõju ja kombinatoorne efekt soodustab või inhibeerib splaissingut (Cieply ja Carstens, 2015).

1.4. SPLAISSINGU SEOS HAIGUSTEGA

Transkriptoomi sekveneerimine on tõhus vahend geneetiliste haiguste, sealhulgas harvade haiguste diagnoosimiseks (Cummings jt., 2017). Sarnaselt teistele uue põlvkonna sekveneerimistehnoloogiatele võimaldab RNA-Seq personaalset lähenemist igale haigusjuhtumile. RNA sekveneerimisel põhinevad diagnostilised meetodid on keskendunud muutuste otsimisele transkriptoomi tasemel nii kodeerivates kui ka mittekodeerivates regioonides (Cummings jt., 2017).

RNA-Seq võimaldab vahetult näha transkriptide funktsionaalseid muutusi, mis on põhjustatud geneetilisest varieeruvusest. Lisaks on transkriptoomi sekveneerimise abil võimalik mõõta geenide ekspressiooni taset ning tuvastada alternatiivsete transkriptide olemasolu, mida genoomi sekveneerimine ei võimalda (Byron jt., 2016).

RNA-Seq abiga on lisaks võimalik identifitseerida haigusseoseliste transkriptide isovorme, mis võiksid olla potentsiaalseteks diagnostilisteks markeriteks (Byron jt., 2016).

Splaissingumustri muutuste poolt põhjustatud alternatiivseid mRNA variante on seostatud paljude haigustega, nagu näiteks vähkkasvajad (eesnäärmevähk (Dehm jt., 2008), glioblastoom (Reardon jt., 2015), rinnavähk (Gambino jt., 2015)), neurodegeneratiivsed haigused (taupaatia (Liu ja Gong, 2008), Parkinsoni tõbi (La Cognata jt., 2015) ja arenguhäired (Duchenne'i ja Beckeri lihasdüstroofiad (Magri jt., 2011)).

Haigusseoselisi struktuurseid variatsioone (insertsioonid, inversioonid, deletsioonid, duplikatsioonid, ühenukleodiidsed polümorfismid *(single nucleotide polymorphism*, SNP)), mis esinevad valku kodeerivates geenides, seostatakse tihti splaissingu muutustega, sõltuvalt sellest, mis tüüpi mutatsiooniga on tegemist (Hindorff jt., 2009; Scotti ja Swanson, 2016; Byron jt., 2016). Olenevalt paiknemisest võib regulatoorseid mutatsioone jagada kahte rühma: mutatsioonid *cis*-regulatoorsetes elementides ja mutatsioonid *trans*-regulatoorsetes elementides (Suñé-Pou jt., 2017).

Mutatsioonid *cis*-regulatoorsetes elementides on kõige levinum splaissingu mutatsiooni tüüp (Scotti ja Swanson, 2016). *Cis*-regulatoorsete elementide mutatsioonid mõjutavad otseselt splaissingut ja splaissosoomi tööd. Sellised mutatsioonid asuvad splaissingu võimendajate ja vaigistajate regioonides või tihtipeale ka splaissingu konsensusjärjestustes, tekitades muutusi harusaidis ja 5’ ning 3’ splaissingusaitides (Scotti ja Swanson, 2016; Suñé-Pou jt., 2017). Mutatsioonid 3’ või 5’ splaissingusaidi konserveerunud konsensusjärjestuses blokeerivad täielikult nende saitide kasutamise ning põhjustavad eksoni vahelejätmist (muteerunud 3’ splaissingusaidi puhul), introni säilitamist (muteerunud 5’ splaissingusaidi puhul) või mõne krüptilise splaissingusaidi kasutamist (Cieply ja Carstens, 2015). Teiste *cis*-regulatoorsete elementide mutatsioonid mõjutavad splaissingusaidi tugevust ja võivad osaliselt või täielikult seda inhibeerida (Cieply ja Carstens, 2015). *Cis*-regulatoorsete elementide mutatsioonid võivad initsieerida uute splaissingusaitide teket, mis võib omakorda viia haiguse tekkele (Scotti ja Swanson, 2016; Suñé-Pou jt., 2017). *Cis*-regulatoorsete elementide mutatsioonid võivad põhjustada selliseid haigusi, nagu eelnimetatud Parkinsoni tõbi (La Cognata jt., 2015), Duchenne’i ja Beckeri lihasdüstroofiad (Magri jt., 2011).

Mutatsioonid *trans*-regulatoorsetes elementides mõjutavad splaissingu abifaktoreid, paiknedes näiteks SR-valkude ja hnRNP-de kodeerivates geenides. Sellised mutatsioonid võivad mõjutada splaissingusaitide kasutust, rikkudes ära tavapärase splaissingumustri (Scotti ja Swanson, 2016; Suñé-Pou jt., 2017). *Trans*-regulatoorsete elementide mutatsioonide poolt põhjustatud haiguste näidetena võib nimetada amüotroofse lateraalskleroosi (Sun jt., 2015a) ja leukodüstroofia (Bartoletti-Stella jt., 2015).

Mutatsioonid splaissosoomi kompleksi koostisosades võivad samuti olla seotud haigustekkega. Tihtipeale esinevad mutatsioonid geenides, mis kodeerivad splaissosoomi abifaktoreid (nt U2AF rühma abifaktorid, PRPF rühma abifaktorid) (Scotti ja Swanson, 2016). Eelnimetatud abifaktorid seonduvad snRNAdega ja moodustavad koos splaissosoomi põhikomponendi – snRNP-d.

Splaissosoomi abifaktorite mutatsioonidega on seotud järgmised haigused: pigmentretiniit (Tanackovic jt., 2011) ja mitmed vähkkasvajad, näiteks kolorektaalvähk (Adler jt., 2014) ja müeloidne leukeemia (Yoshida jt., 2011).

Splaissingut mõjutavad mutatsioonid põhjustavad sageli mRNA lugemisraami nihke või enneaegse stoppkoodoni tekke. Paljudel juhtudel initsieeritakse sellise transkripti lagundamine *nonsense-mediated decay* (NMD) raja kaudu, mis võib mõjutada antud geeni ekspressioonitaset (Cieply ja Carstens, 2015).

1.5. SPLAISSINGUT UURIVATE PROGRAMMIDE ÜLEVAADE

Alternatiivse splaissingu käigus tekkinud isovormide esinemise hindamiseks RNA-Seq andmetes kasutatakse erinevaid bioinformaatilisi tööriistu.

Alternatiivse splaissingu tuvastamiseks võrdleb enamik programme sekveneeritud lugemeid annotatsioonifailiga, kus on märgitud eksonite ühenduskohad (*splice-junction*). RNA-Seq käigus saadud lugemid jaotatakse kahte rühma: eksonite ühenduskohti katvad (*junction*) ja täielikult ühes eksonis paiknevad (*non-junction*) lugemid. Eksonite ühenduskohti katvad on sellised lugemid, millel on sees lüngad, kui neid referentsgenoomile joondatakse. Selliste lugemite üks osa joondub ühele eksonile ja teine osa järgmisele eksonile. Täielikult ühes eksonis paiknevad lugemid joondatakse referentsgenoomile lünkadeta.

1.5.1. SplicingTypesAnno

SplicingTypesAnno on programm, mille põhieesmärgiks on erinevate alternatiivse splassingu tüüpide tuvastamine ja annoteerimine (Sun jt., 2015b).

Programm on võimeline tuvastama kõiki peamisi alternatiivse splaissingu juhtumeid, nagu näiteks:

- introni säilitamine
- eksoni vahelejätmine
- alternatiivne 5’ splaissingusait (doonorsait)
- alternatiivne 3’ splaissingusait (aktseptorsait)

Igal alternatiivse splaissingu tüübil on kaks alamtüüpi: I ja II alamtüüp. I alamtüüp märgib juhtumit, mis hõlmab ainult ühte eksonit või intronit. II alamtüüp tähistab juhtumit, mis hõlmab mitut eksonit või intronit. Seega jaotab programm alternatiivse splaissingu juhtumid lõppkokkuvõttes kaheksaks tüübiks (Sun jt., 2015b).

Alternatiivse splaissingu tüüpide määramine toimub RNA-Seq andmete võrdlemisel annotatsiooniga. SplicingTypesAnno on võimeline tuvastama ka uudseid alternatiivse splaissingu variante (Sun jt., 2015b).

SplicingTypesAnno on R-i põhine programm. Programmi vajalikud sisendfailid on proovi järjestus BAM formaadis ja anoteerimisfail GTF või GFF formaadis.

Analüüsi võib teostada nii geeni kui ka kogu genoomi tasemel. Analüüsi lõpus genereerib programm veebiaruande, mis sisaldab proovi kirjeldust, lühistatistikat lugemite kohta, leitud splaissingujuhtumite ülevaadet ja BED formaadis faile IGV genoomibrauseris visualiseerimiseks (Sun jt., 2015b).

1.5.2. MISO

MISO (*Mixture-of-Isoforms*) on statistiline mudel, mis kasutab paaris- või üksiklugemite formaadis RNA-Seq andmeid alternatiivse splaissingu analüüsiks nii eksonite kui ka isovormide tasemel (Katz jt., 2010).

MISO modelleerib RNA-Seq-i protsessi ennustades tõenäosust, et teatud lugem pärineb ühest kindlast isovormist. Selleks kasutab MISO Bayesi otsustusmeetodit (veebilehekülg: MISO software documentation, 2017).

MISO tõenäosuslik mudel hindab alternatiivselt splaissitud eksonite ja isovormide ekspressiooni. Iga leitud alternatiivselt splaissitud juhtumile määrab MISO usaldusintervalli (Katz jt., 2010). Hinnangu algoritm põhineb valimivõtul ja Markovi ahelate Monte Carlo algoritmil (“MCMC”) (veebilehekülg: MISO software documentation, 2017). Samuti võimaldab MISO teostada diferentsiaalse kasutuse paarisanalüüsi (kontroll vs uuritav).

Põhilised parameetrid, mida MISO mudel välja arvutab, on Ψ (psi) ehk alternatiivselt splaissitud isovormide osakaal (*percentage spliced in*) ning selle variatsioonid (*unbiased estimators of* Ψ) nagu $\hat{\Psi}_{SJ}$ ja $\hat{\Psi}_{\mathrm{MISO}}$ (Katz jt., 2010).

MISO on võimeline tuvastama kõiki peamisi alternatiivse splaissingu ja pre-mRNA protsessimise tüüpe (veebilehekülg: MISO software documentation, 2017):

- eksoni vahelejätmine
- introni säilitamine
- alternatiivne 3'/5' splaissingusait
- teineteist välistavad eksonid
- tandeemsed 3' UTR regioonid
- alternatiivne esimene ekson
- alternatiivne viimane ekson

MISO on Pythoni-põhine programm. Sisendfailina kasutab programm BAM faile ja annoteerimisfaili GFF formaadis.

Andmete analüüsiks on vaja sisestada lugemite pikkus ning paarislugemite puhul ka keskmine RNA-Seq raamatukogu fragmendi pikkus ja selle standardhälve (veebilehekülg: MISO software documentation, 2017).

Analüüsi lõpus väljastatakse iga proovi jaoks eraldi kokkuvõttev tabel, mis sisaldab olulist infot kõikide leitud isovormide kohta. Diferentsiaalse analüüsi abil on võimalik erinevaid proove omavahel võrrelda (veebilehekülg: MISO software documentation, 2017).

MISO programm võimaldab tulemuste visualiseerimist sashimi graafiku abil. Joonisel visualiseeritakse eksonid ja nende liitekohad uuritavas genoomses regioonis. Eksonite liitekohad on kujutatud kaarega ning kaare peal olev number näitab mitu lugemit antud liitekohaga seotud on (Katz jt., 2015).

1.5.3. JunctionSeq

JunctionSeq on tarkvara, mis võimaldab tuvastada eksonite ja nende ühenduskohtade erinevat kasutamist RNA-Seq andmete põhjal. JunctionSeq-i abil on võimalik tuvastada ka uudsete ekson-ekson liitekohtade diferentsiaalset kasutust nende jaoks spetsiaalse annotatsioonita. JunctionSeq suudab väidetavalt analüüsi läbi viia ka juhul kui isovormide annotatsioon on puudulik (Hartley jt., 2016).

JunctionSeq põhineb QoRT ja suuremal määral ka DEXSeq programmidel, millele on lisatud visualiseerimise tööriistad (Hartley jt., 2016).

DEXSeq on statistiline tarkvara, mis tuvastab eksonite diferentsiaalset kasutamist. DEXSeq programm põhineb lineaarsetel mudelitel (*generalized linear model,* GLM) (Anders jt., 2012). Kuna DEXSeq otsib isovorme ainult eksonite tasemel, siis on see võrreldes JunctionSeq-iga võimeline tuvastama vähem alternatiivsete isovormide tüüpe. Samuti põhineb DEXSeq-i analüüs olemasoleval annotatsioonil ja uudsete isovormide tuvastamine ei ole seega võimalik (Hartley jt., 2016)

QoRT (*Quality of RNA-Seq Toolset*) on funktsionaalne tarkvara RNA-Seq andmete kvaliteedikontrolliks ja protsessimiseks. QoRT programmi kasutatakse lugemite arvu loendamiseks geeni, eksonite ja eksonite liitekohtade tasemel (Hartley jt., 2015; Hartley jt., 2016). Samuti võimaldab QoRT moodul sisendfailide genereerimist genoomibrauserite jaoks (IGV või UCSC).

JunctionSeq on R-i põhine programm. QoRT funktsioonide käivitamiseks on vajalik lisaks Java mooduli olemasolu. Sisendfailidena vajab JunctionSeq proovide faile BAM formaadis ja GTF formaadis annoteerimisfaili.

Analüüsi alguses grupeeritakse geeni tasemel mitte-kattuvad lugemid (või lugemite paarid) eksonite kaupa eraldi rühmadesse. Seejärel loendatakse iga eksoni lugemid (DEXSeq mooduliga) ja eksonite liitekohtadega seotud lugemid (QoRTs mooduliga) nii teadaolevates kui ka uudsetes liitekohtades. Kasutades teise R-paketi DESeq2 moodulit ja üldistatud lineaarseid mudeleid, kontrollitakse seejärel iga eksoni ja eksonite ühenduskoha diferentsiaalset kasutust (Love jt. 2014). Saadud andmetest valitakse välja statistiliselt olulised tulemused, kasutades automatiseeritud sõltumatut filtreerimismeetodit (filtreerimisläve võib programmile käsitsi ette anda), mille peal hakatakse diferentsiaalse kasutuse hüpoteesi kontrollima (Hartley jt., 2016).

Programmi suureks eeliseks on mitmekülgsed visualiseerimisvõimalused, mis aitavad kaasa tulemuste interpreteerimisele. JunctionSeq genereerib joonise iga geeni jaoks, mis sisaldab rohkem kui ühte diferentsiaalses kasutuses olevat eksonit või eksonite liitekohta (Hartley jt., 2016).

2. EKSPERIMENTAALNE OSA

2.1. TÖÖ EESMÄRK

Eksperimentaalse osa eesmärgiks oli hinnata kliiniliselt olulistes geenides leiduvate mutatsioonide mõju transkriptoomi tasemel. Selleks kasutati Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) geenidoonorite kogu genoomi ning transkriptoomi sekveneerimise andmeid.

2.2. MATERJALID JA METOODIKA

2.2.1. Valimi kirjeldus ja kasutatud genoomiandmed

Käesolevas töös kasutati ning analüüsiti TÜ EGV varasemate projektide raames kogutud andmeid.

Uuritav materjal oli kogutud Siirdegenoomika Keskuse (*Centre of Translational Genomics*, CTG, http://www.ctg.ut.ee/) projekti raames aastatel 2011-2015. Uuring on heaks kiidetud Tartu Ülikooli inimuuringute eetika komitee poolt (Protokolli number: 2061/T-4).

Geenidoonorite kogu genoomi sekveneerimine (WGS) viidi läbi keskmiselt 30-kordse katvusega kasutades Illumina HiSeq X platvormi Broad Instituudis. Sekveneerimise metoodika, kvaliteedikontroll ning variantide määramine on täpsemalt kirjeldatud artiklites (Guo jt., 2016; Mitt jt., 2017).

Käesolevas töös analüüsiti selliste indiviidide geenivariante, kellel olid olemas ka RNA sekveneerimise andmed.

2.2.2. RNA sekveneerimine (RNA-Seq) ning andmete esmane analüüs

RNA eraldati Tempus katsutitest (Applied Biosystems, Foster City, CA) TÜ EGV tuumiklaboris kasutades TRIzol reagenti (Invitrogen). Globiinide mRNA hulka vähendati GLOBINclear Kit'i (Invitrogen) abil. Sekveneerimiseks kasutati 200 ng RNA-d ning sekveneerimine viidi läbi EGV tuumiklaboris järgides Illumina TruSeq 50 bp paarislugemite (*paired-end*) protokolli.

Sekveneeritud RNA-de (N = 625) kvaliteedikontroll oli läbi viidud programmiga FastQC (versioon 0.11.3, (Andrews, 2015)) Viktorija Kukuškina poolt. Kontrollitavad parameetrid olid järgmised: „*Per base sequence quality*", „*Per base sequence content*", „*Overrepresented sequences*" ja „*Adapter Content*". Kõik sekveneeritud proovid läbisid FastQC kvaliteedikontrolli.

Esmane toorlugemite joondamine lugemite hulga hindamiseks teostati inimese referentsgenoomil (versioon GRCh37/hg19) põhinevale transkriptoomile kasutades Kallisto tarkvara (versioon 0.42.2.1, (Bray jt., 2016)). Saadud lugemite hulk proovi kohta oli keskmiselt 24,7 miljonit.

Seejärel kontrolliti RNA-Seq proovide vastavust kogu genoomi sekveneerimise andmetele Mixupmapper programmiga (versioon 1.3.7, (Westra jt., 2011)). Selgus, et WGS andmed puudusid 38 proovil. *Mixup*-analüüs identifitseeris 60 probleemset proovi, millest 6 juhtumit lahendati. Lisaks eemaldati edasisest analüüsist 44 proovi, mille puhul oli ilmnenud tehniline tõrge sekveneerimisraamatukogude valmistamisel. Lõplikuks analüüsitavate proovide arvuks käesolevas töös jäi 489.

Geenide ekspressioonitase (FPKM, *fragments per kilobase of transcript per million mapped reads*) leiti programmiga cufflinks (versioon 2.2, (Trapnell jt., 2010)).

BAM failide genereerimiseks joondati paarislugemid referentsgenoomile (versioon GRCh37/hg19). Joondamine viidi läbi Hisat tarkvaraga (versioon 0.1.7-beta, (Kim jt., 2015)). Joondatud andmed konverteeriti SAM formaadist BAM formaati kasutades programmi Samtools (versioon 1.2, (Li jt., 2009)). Sama programmiga teostati ka BAM failide sorteerimine ja indekseerimine.

2.2.3. Kliiniliselt oluliste geenide nimekirjad

Ameerika Meditsiinigeneetika ja -genoomika Kolleegium (*American College of Medical Genetics and Genomics*, ACMG) on välja töötanud kliiniliselt oluliste geenide nimekirja, milles ülegenoomsete uuringute korral tuvastatud leidude puhul on soovituslik anda tagasisidet uuritavatele ja nende pereliikmetele (Green jt., 2013). Esmane ACMG nimekiri avaldati 2013. aastal ning see sisaldab 56 geeni, mis on seotud 24 erineva haigusseisundiga (Green jt., 2013).

Teine kliiniliselt oluliste geenide nimekiri on koostatud Regeneron Geneetikakeskuse (*Regeneron Genetics Center*) ja Geisinger tervishoiusüsteemi (*Geisinger Health System*) poolt 2016. aastal (Dewey jt., 2016). Geisinger nimekiri sisaldab 76 geeni, millest 56 geeni pärinevad ACMG nimekirjast (Dewey jt., 2016).

2.2.4. Alternatiivse splaissingu tuvastamise tarkvara

Alternatiivse splaissingu tuvastamiseks kasutati järgmiseid programme: SplicingTypesAnno, MISO ja JunctionSeq.

2.2.4.1. SplicingTypesAnno (versioon 1.0.2)

Annotatsiooniks kasutati inimese referentsgenoomi (versioon GRCh37/hg19) GTF failiformaadis. Sisendfailina kasutati sorteeritud ja indekseeritud BAM faili. Analüüsi teostati geeni tasemel. Kasutatavad funktsioonid olid järgmised:

- *translateGTF*: annotatsioonifaili ümberkirjutamine ekson-intron struktuuri
- *splicingCount*: lugemite loendamine lähtudes ekson-intron struktuurist
- *splicingGene*: alternatiivse splaissingu tüüpide tuvastamine ja annoteerimine
- *splicingReport*: analüüsi tulemuste kirjutamine väljundfaili

2.2.4.2. MISO (versioon 0.5.3)

Annotatsiooniks kasutati inimese referentsgenoomi (versioon GRCh37/hg19) GFF failiformaadis. Sisendfailina kasutati sorteeritud ja indekseeritud BAM faili. Analüüsi teostati isovormi tasemel. Analüüsi käivitamiseks kasutati standardseid MISO parameetreid:

- minimaalne lugemite arv regiooni kohta: 20
- splaissingu juhtude katvuse filteerimine: *True*
- iteratsioonide koguarv geeni kohta: 5000

Analüüsiks kasutati järgmiseid funktsioone:

- *index_gff*: annoteerimisfaili indekseerimine
- *exon_utils*: RNA-Seq raamatukogu inserdi keskmisest oluliselt pikemate konstitutiivsete eksonite (minimaalne pikkus 1000 aluspaari) filtreerimine annoteerimisfailist. Vajalik etapp järgnevaks inserdi pikkuse arvutamiseks
- *pe_utils*: keskmise inserdi pikkuse ja selle standardhälbe arvutamine
- *run*: analüüsi käivitamine. Lisaparameetrid: lugemite pikkus - 51 nukleotiidi; inserdi pikkus ja selle standardhälve oli iga proovi jaoks individuaalne
- *summarize-samples*: tulemuste summeerimine ja statistikute (psi ja selle usaldusintervallid) arvutamine
- *sashimi_plot*: tulemuste visualiseerimine geeni tasemel

2.2.4.3. QoRTs (versioon 1.1.8)

JunctionSeq töötamiseks vajalik lisatarkvara. Lugemite hulga loendamiseks kasutati järgmiseid funktsioone: *writeKnownSplices*, *writeNovelSplices*, *writeSpliceExon*.

Annotatsiooniks kasutati inimese referentsgenoomi (versioon GRCh37/hg19) GTF failiformaadis. Enne splaissinguanalüüsi teostati annoteerimisfaili indekseerimine *makeFlatGff* funktsiooni abil. Indekseeritud anoteerimisfaili salvestatakse GFF failiformaadis. QoRTs käivitamisel kasutati standardset parameetrit „*stranded*“, kuna RNA-Seq andmed on ahela-spetsiifilised.

2.2.4.4. JunctionSeq (versioon 1.2.4)

Sisendfailina kasutati sorteeritud ja indekseeritud BAM faili. Annotatsiooniks kasutati indekseeritud anoteerimisfaili GFF formaadis, mis oli eelnevalt genereeritud QoRTs programmi poolt. Kasutatavad funktsioonid:

- *runJunctionSeqAnalyses*: põhiline funktsioon, mis teostab splaissinguanalüüsi. Valitud analüüsi tüüp oli "*junctionsAndExons*"
- *writeCompleteResults*: tulemuste kirjutamine

- *buildAllPlots*: analüüsipõhiste kokkuvõtvate graafikute genereerimine. Kohandatud p-väärtuse lävi: 0,01
- *buildAllPlotsForGene*: graafikute genereerimine konkreetse geeni jaoks

2.2.5. Alleel-spetsiifiline ekspressioon

Alleel-spetsiifilise ekspressiooni (*allele-specific expression*, ASE) analüüsi abil uuriti, kas uuritaval geenil võib esineda kõrvelekaldeid mõlema alleeli võrdsest ekspressioonist. Üheks võimalikuks ASE põhjustajaks võib olla alternatiivne splaissing. Enneaegse stoppkoodoni teke võib põhjustada *nonsense mediated decay* raja aktiveerimise, mis viib mutatsiooniga mRNA molekuli lagundamiseni ning antud lookuses alleelide ekspressioonitaseme erinevusele (Rivas jt., 2015).

ASE analüüs oli teostatud CTS-ASE programmiga (veebilehekülg: Github, 2017) Viktorija Kukuškina poolt. ASE mõõdeti SNP-de tasemel, kasutades RNA-Seq andmeid.

2.2.6. Standardskoori arvutamine

Uuritavate geenide ekspressioonitaset mutatsiooniga indiviididel võrreldes kogu valimi keskmisega väljendati z-skoorina (standardskoorina). Standardskoor arvutatakse kasutades valemit:

$$z = (X - \mu)/\sigma$$

kus z on standardskoor, X on uuritav element, μ on valimi keskmine ja σ on standardhälve. Saadud väärtus näitab mitme standardhälbe võrra erineb uuritav element valimi keskmisest.

2.3. TULEMUSED JA ARUTELU

2.3.1. Esmane indiviidide valik uuringusse

Uuringusse valiti indiviidid, kellel olid WGS andmete põhjal tuvastatud funktsioonikaoga (*loss-of-function*) mutatsioonid kliiniliselt olulistes geenides (ACMG nimekirja järgi) ning kellel olid olemas ka RNA-Seq andmed. Indiviidide algne valik WGS andmete põhjal ning geenivariantide kontroll kasutates IGV (Integrative Genomics Viewer) rakendust lugemite visualiseerimiseks konkreetses lookuses oli eelnevalt teostatud Neeme Tõnissoni töörühma ja Tiit Nikopensiuse poolt. Mutatsioonid *BRCA1* ja *BRCA2* geenides olid lisaks kinnitatud Sangeri sekveneerimise abil Karmen Vaiküllі bakalaureusetöö raames (Vaiküll, 2016).

Toodud kriteeriumite põhjal valiti 17 indiviidi, kellel esinesid mutatsioonid 11 erinevas geenis (Tabel 1). Kolmel proovil (GD10, GD12, GD15) esinesid splaissingusaidi mutatsioonid.

Tabel 1. Uuringusse valitud indiviidid ACMG nimekirja järgi.

Indiviid	**Geen**	**Kromosoom: positsioon**	**cDNA mutatsioon**	**Mutatsiooni tüüp**	**Alleelide arv ***
GD1	*BRCA1*	Chr17: 41234520	c.4258C>T	Enneaegne stoppkoodon	1
GD2	*BRCA1*	Chr17: 41245708	c.1840A>T	Enneaegne stoppkoodon	1
GD3	*BRCA1*	Chr17: 41209082	c.5329dupC (5382insC)	Raaminihke mutatsioon	4
GD4	*BRCA1*	Chr17: 41209082	c.5329dupC (5382insC)	Raaminihke mutatsioon	4
GD5	*BRCA1*	Chr17: 41243513	c.4035delA (4154delA)	Raaminihke mutatsioon	6
GD6	*BRCA1*	Chr17: 41243513	c.4035delA (4154delA)	Raaminihke mutatsioon	6
GD7	*BRCA2*	Chr13: 32900279	c.467_468 insT	Raaminihke mutatsioon	1
GD7	*RYR2*	Chr1: 237954744	c.13493_ 13494delAC	Raamminihke mutatsioon	1

Tabel 1 järg

GD8	*APC*	Chr5: 112179247	c.7957delA	Raaminihke mutatsioon	1
GD9	*MSH2*	Chr2: 47703631	c.2131C>T	Enneaegne stoppkoodon	1
GD10	*PMS2*	Chr7: 6029430	c.1144+ 1G>A	5’ splaissingusaidi mutatsioon	1
GD11	*PMS2*	Chr7: 6035203	c.861_864 delACAG	Raaminihke mutatsioon	1
GD12	*PKP2*	Chr12: 32949234	c.2300-2A>T	3’ splaissingusaidi mutatsioon	3
GD13	*KCNH2*	Chr7: 150645949	c.2587C>T	Enneaegne stoppkoodon	1
GD14	*SCN5A*	Chr3: 38655514	c.655C>T	Enneaegne stoppkoodon	1
GD15	*SCN5A*	Chr3: 38663981	c.393-1C>T	3’ splaissingusaidi mutatsioon	2
GD16	*NF2*	Chr22: 30077554	c.1702_1703del AG	Raaminihke mutatsioon	1
GD17	*VHL*	Chr3: 10183539	c.8_9ins18	Raaminihke mutatsioon	1

* Mutatsiooniga alleelide arv kogu WGS andmestikus

2.3.2. Kandidaatgeenide valiku laiendamine

Soovides laiendada uuringut teistele kliiniliselt olulistele geenivariantidele, leiti ClinVar andmebaasist (veebilehekülg: ClinVar, NCBI) kõik kliiniliselt olulised splaissinguga seotud variandid. Otsingu parameetrid olid järgmised: *Organism – Human, Molecular consequence – Splice site, Review status* – vähemalt üks teadusartikkel, mis vastab ClinVar kriteeriumitele. Kokku leiti 3 640 varianti 888 geenis.

Järgmiseks kitsendati uuringut kõrgema ekspressioonitasemega (ja seega kõige informatiivsematele) geenidele ning valiti välja 250 geeni, mille FPKM väärtus EGV valimis oli suurem kui 10. Selline kriteerium seati eelkatsete põhjal, mis näitasid, et madalama ekspressiooni korral ei pruugi lugemite hulk olla analüüsi läbiviimiseks piisav.

Kliiniliselt oluliste geenide valiku aluseks võeti Geisinger'i nimekiri, mille alusel jäi edasisse analüüsi 16 geeni.

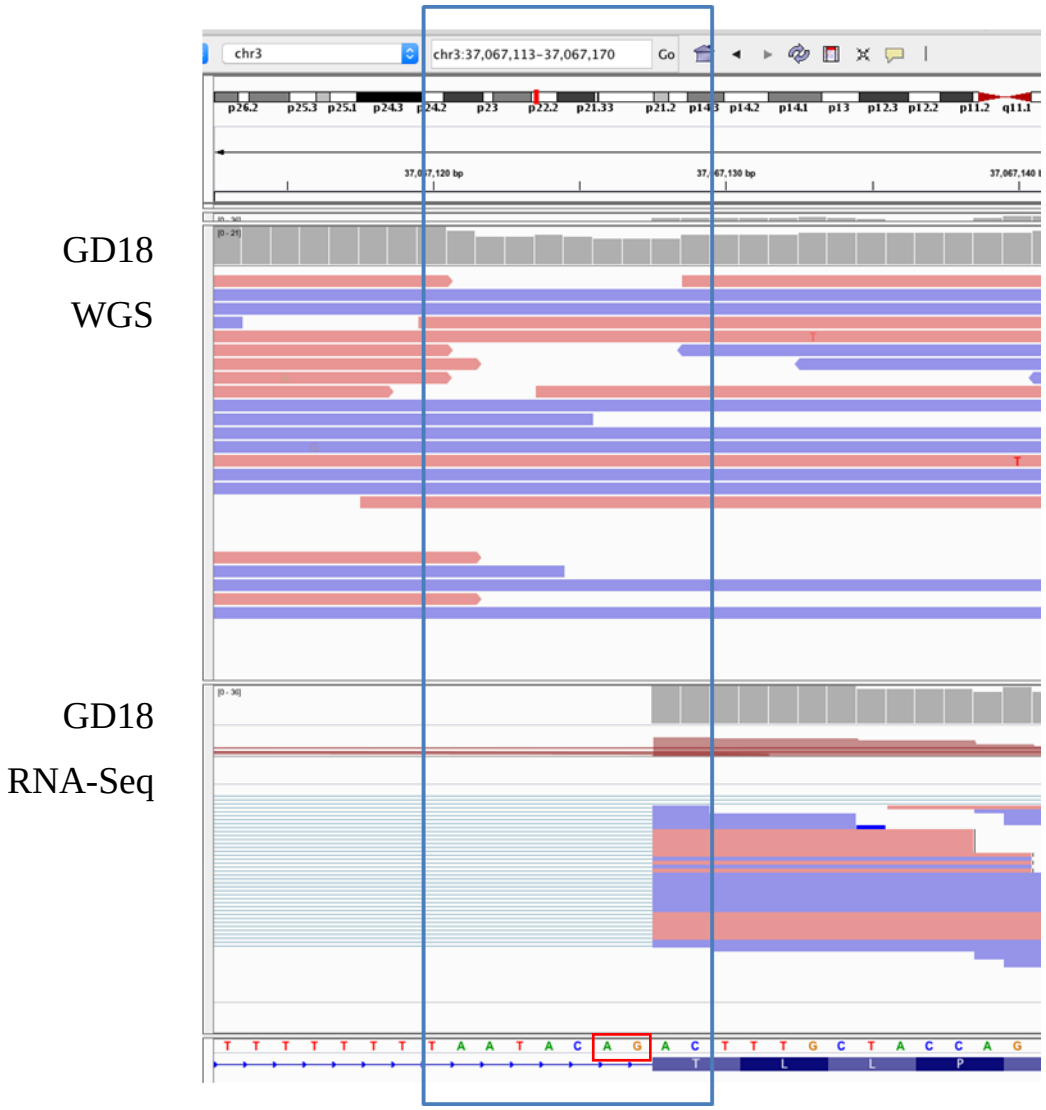
Neid geene kontrolliti funktsioonikaoga mutatsioonide suhtes kogu RNA-Seq valimi hulgas. Kokku leiti neli funktsioonikaoga mutatsiooni kahes geenis: *MLH1* (keskmine FPKM = 11,3) ja *TSC2* (keskmine FPKM = 23,8) (Tabel 2).

Tabel 2. Uuringusse valitud indiviidid Geisingeri nimekirja järgi.

Indiviid	**Geen**	**Kromosoom: positsioon**	**Ref-SNP ID**	**cDNA mutatsioon**	**Mutatsiooni tüüp**	**Alleelide arv ***
GD18	*MLH1*	Chr3: 37067120	-	9-nukleotiidi pikkune deletsioon	3’ splaissingu-saidi mutatsioon	6
GD19	*TSC2*	Chr16: 2137924	rs137854209	34-nukleotiidi pikkune deletsioon	5’ splaissingu-saidi mutatsioon	11
GD20	*TSC2*	Chr16: 2137924	rs137854209	34-nukleotiidi pikkune deletsioon	5’ splaissingu-saidi mutatsioon	11
GD21	*TSC2*	Chr16: 2137924	rs137854209	34-nukleotiidi pikkune deletsioon	5’ splaissingu-saidi mutatsioon	11

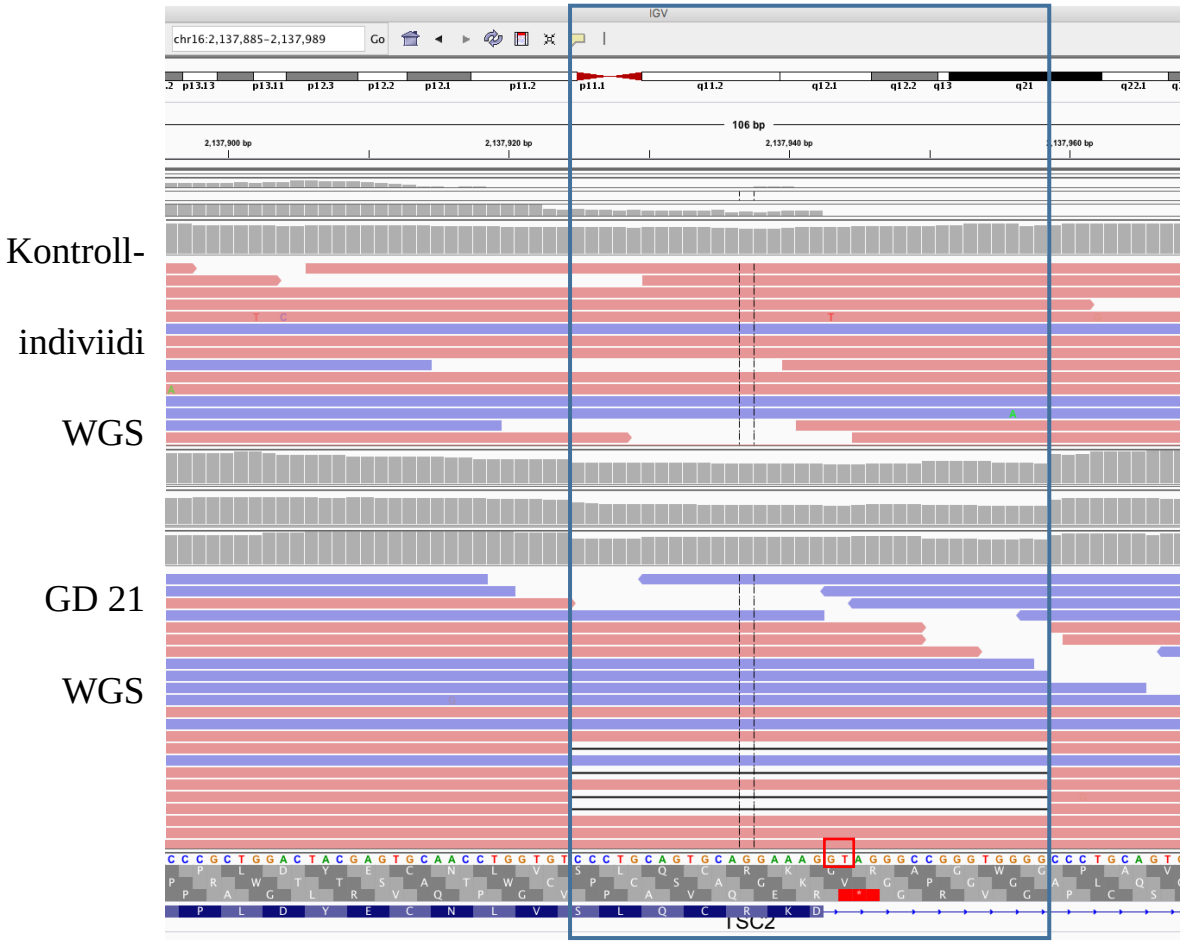
* Mutatsiooniga alleelide arv kogu WGS andmestikus

GD18 indiviidil leitud *MLH1* geeni 9-nukleotiidi pikkune deletsioon asub 3’ splaissingusaidis 11. introni ja 12. eksoni piiril. Mutatsiooni asukoha visualiseerimine indiviidi GD18 WGS ja RNA-Seq andmetes IGV rakenduse abil siiski ei toeta deletsiooni esinemist. WGS andmete järgi kattub 24 lugemit osaliselt või täielikult oletatava deletsiooni asukohaga. Samuti ei ole RNA-Seq andmete kohaselt 3’ splaissingusait rikutud (Joonis 5). Kuna antud alleeli leidub WGS andmestikus 6 korda, siis deletsiooni kinnitamiseks või ümber lükkamiseks oleks vaja läbi viia Sangeri sekveneerimine.



Joonis 5. Geenis *MLH1* indiviidil GD18 leitud splaissingu aktseptorsaidi oletatav mutatsioon TAATACAGAC > T (märgitud sinise kastiga) genoomi (WGS) ja transkriptoomi sekveneerimise (RNA-Seq) andmete põhjal. Leitud deletsioon kaasab splaissingu aktseptorsaidi konsensusjärjestuse (märgitud punase kastiga). Joonis on tehtud IGV rakenduses.

Mutatsioon *TSC2* geenis tuvastati kolmel indiviidil – GD19, GD20 ja GD21. Deletsioon asub 39. eksoni ja 39. introni piiril (Joonis 6).



Joonis 6. Geeni *TSC2* splaissingu doonorsaidi mutatsiooni asukoha (märgitud sinise kastiga) võrdlus indiviidil GD21 sama positsiooniga kontrollindiviidil WGS andmete põhjal. Leitud deletsioon kaasab splaissingu doonorsaidi konsensusjärjestuse (märgitud punase kastiga). Joonis on tehtud IGV rakenduses.

Uuritava *TSC2* mutatsiooni (rs137854209) alleelisagedus EGV valimis on 0,002451. See väärtus on võrreldav Euroopa keskmise aleelisagedusega, mis on 0,002348 ExAC andmebaasi järgi (veebilehekülg: Exome Aggregation Consortium) (Tabel 3).

Tabel 3: *TSC2* geeni mutatsiooni (rs137854209) sagedus erinevates populatsioonides ExAC andmebaasi järgi (veebilehekülg: Exome Aggregation Consortium).

Populatsioon	**Alleeli sagedus**
Ida-Aasia	0,003744
Lõuna-Aasia	0,002668
Euroopa (v.a. Soome)	0,002348
Aafrika	0,001874
Ladina-Ameerika	0,001219
Muu	0,001144
Euroopa (Soome)	0,001079
Keskmine	**0,002265**

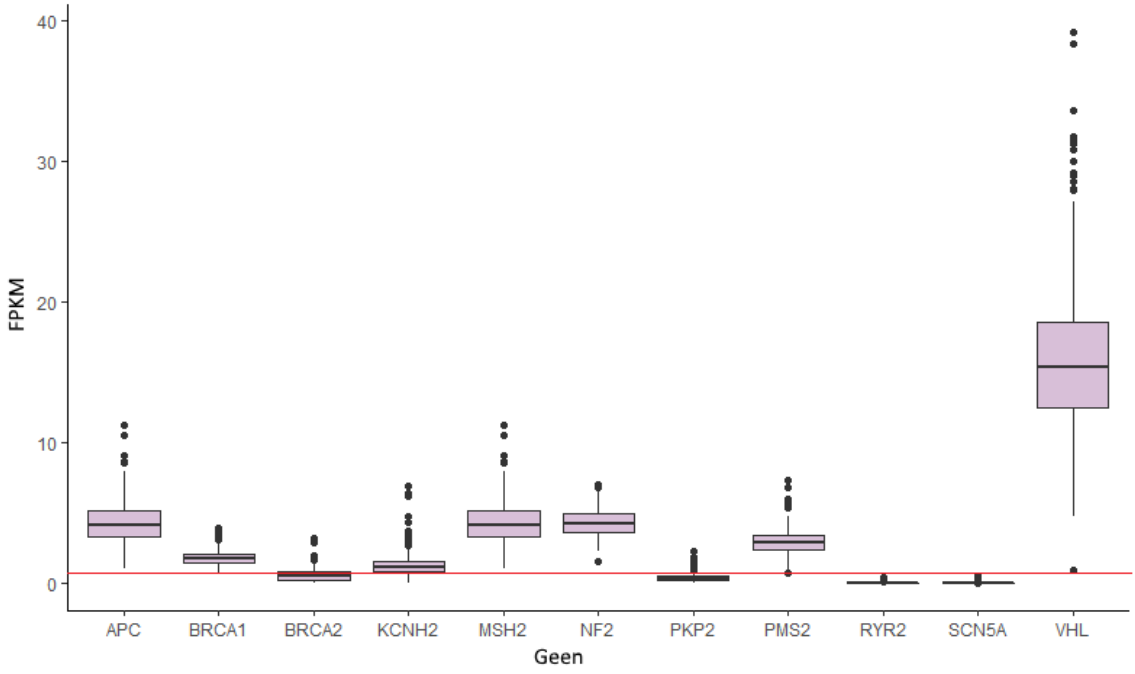
2.3.3. Ekspressioonianalüüs

Selleks, et oleks võimalik teha järeldusi geenivariantide mõju kohta mRNAle, peaks antud geeni avaldumine uuritavas koes olema tuvastatav. Kasutatud RNA-Seq andmestiku põhjal mõõdeti algselt 63 654 transkripti ekspressiooni.

Käesolevates uuringus valiti ekspressioonitaseme mõõtmise ühikuks FPKM, kuna see on sobiv paarislugemite korral. Seades detekteeritava ekspressiooni piiriks FPKM > 1, jäi edasisse analüüsi 13 588 transkripti, mis vastasid 13 553 unikaalsele geenile. See on lähedane geenide arvule, mida on võimalik usaldusväärselt tuvastada ekspressioonikiipidega (nt. Peters jt. uuringus tuvastati veres 15 639 geeni avaldumine (Peters jt., 2015)). Uuritavast 11 geenist nelja (*RYR2*, *BRCA2*, *PKP2* ja *SCN5A*) ekspressioon oli sellest piirist madalam (vastavalt 0,043, 0,573, 0,393 ja 0,02) (Tabel 4, Joonis 7).

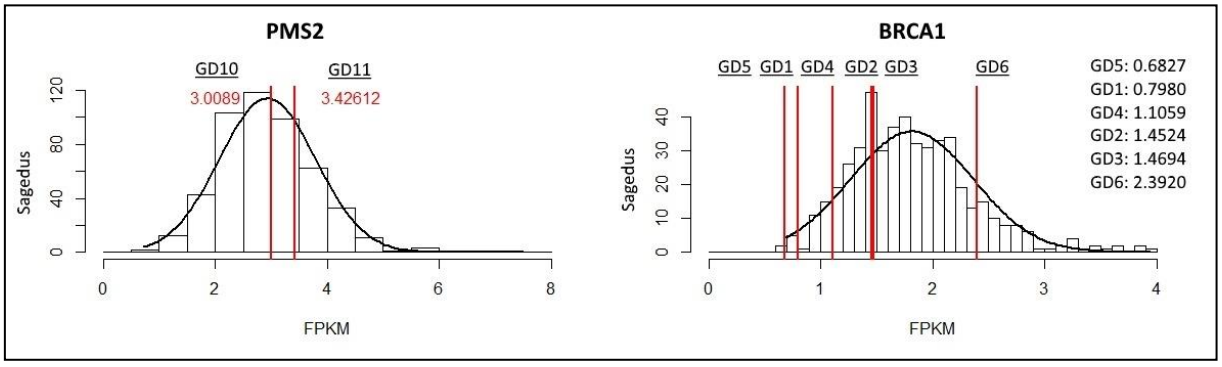
Tabel 4. Uuritavate geenide ekspressioon mutatsiooniga indiviididel. Andmed on järjestatud keskmise ekspressioonitaseme kahanemise järgi. Neljal geenil (*RYR2*, *BRCA2*, *PKP2* ja *SCN5A*) oli keskmine FPKM < 1 ja nad jäid edasist analüüsist välja. FPKM väärtused on ümmardatud täpsusega 4 kohta pärast koma. Keskmine ekspressioonitase on arvutatud kasutades kogu RNA-Seq andmestikku.

Nr	Proov	Geen	Indiviidi FPKM	Keskmine FPKM
1	GD17	*VHL*	11.9991	15,8832
2	GD16	*NF2*	3,4314	4,3240
3	GD11	*PMS2*	3,4261	2,9328
4	GD10	*PMS2*	3,0089	2,9328
5	GD9	*MSH2*	2,9983	4,3462
6	GD6	*BRCA1*	2,3920	1,8128
7	GD8	*APC*	2,1428	2,7310
8	GD3	*BRCA1*	1,4694	1,8128
9	GD2	*BRCA1*	1,4524	1,8128
10	GD4	*BRCA1*	1,1059	1,8128
11	GD1	*BRCA1*	0,7980	1,8128
12	GD5	*BRCA1*	0,6827	1,8128
13	GD13	*KCNH2*	0,5034	1,3007
14	GD12	*PKP2*	0,4613	0,3931
15	GD7	*BRCA2*	0,1459	0,5738
16	GD7	*RYR2*	0,0337	0,0436
17	GD15	*SCN5A*	0,0083	0,0244
18	GD14	*SCN5A*	0	0,0244



Joonis 7. Uuritavas andmestikus funktsioonikaoga mutatsioone sisaldavate geenide ekspressioonitase (FPKM). Joonisel on näha ka neli geeni (*RYR2*, *BRCA2*, *PKP2* ja *SCN5A*), mille keskmine FPKM oli alla 1 (punane joon) ning mis jäid edasist analüüsist välja. Karpdiagrammil on näidatud mediaan, alumine ja ülemine kvartiil. Vertikaaljoontega on tähistatud 95%-line usaldusintervall.

Uuritavate geenide ekspressioonitaseme varieeruvuse võrdlus mutatsiooni kandjatel ja ülejäänud RNA-Seq valimis visualiseeriti histogrammi kujul (Joonis 8).

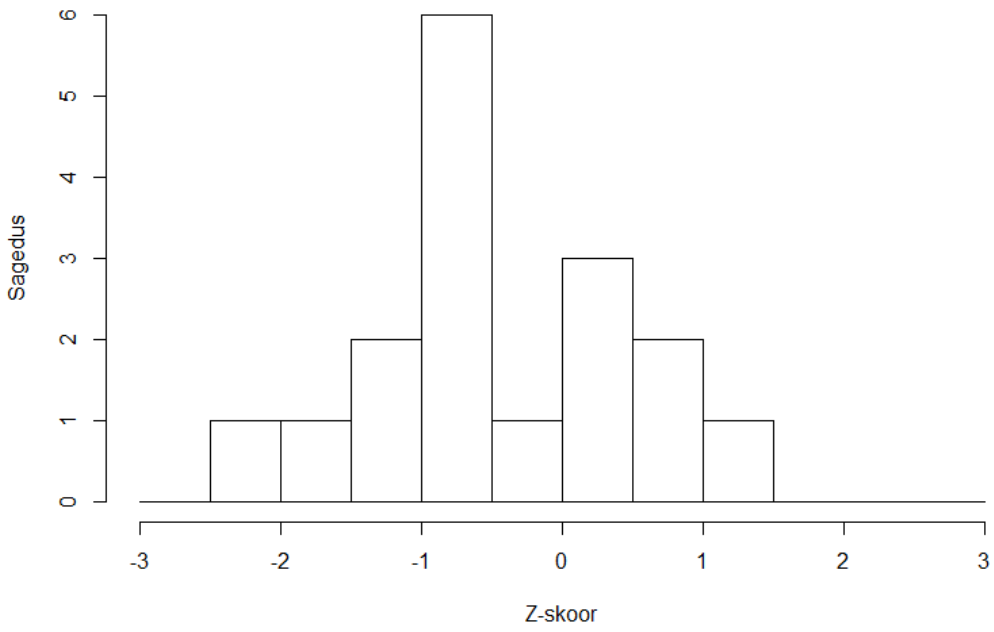


Joonis 8. Näited *PMS2* ja *BRCA1* geenide ekspressioonist. Histogrammil on toodud vastava geeni ekspressioontaseme jaotus kogu RNA-Seq valimi indiviididel. Punaste joontega on näidatud ekspressioonitase uuritavatel indiviididel. Toodud on ka vastavad FPKM numbrilised väärtused.

2.3.4. Mutatsiooniga geenivariantide standardskoorid

Uuritavate geenide ekspressioonitaset mutatsiooniga indiviididel võrreldes kogu valimi keskmisega väljendati z-skoorina (standardskoorina) (Tabel 5, Joonis 9).

Kuna iga konkreetse mutatsiooniga indiviidide arv jäi uuritavas valimis enamasti alla kolme, siis iga mutatsiooni efekti individuaalne hindamine on keerukas. Summaarselt analüüsides oli uuritud mutatsioonide puhul geenide ekspressioonitase keskmisest madalam ($p = 0{,}038$, kahepoolne Wilcoxoni astakmärgitest). Samas tuleb silmas pidada, et koos analüüsiti erinevaid mutatsioone erinevates geenides, mille toime geeni avaldumisele võib samuti olla väga erinev.



Joonis 9. Mutatsiooniga geenide ekspressiooni Z-skooride jaotus uuritavatel indiviididel.

Tabel 5: Uuringusse valitud geenide ekspressioonitaseme kõrvalekalle valimi keskmisest mutatsiooniga indiviididel. Tulemused on järjestatud z-skoori kahanemise järgi. z-skoorid on ümardatud 4 kohta pärast koma.

Nr	**Indiviid**	**Geen**	**cDNA mutatsioon**	**z-skoor**
1	GD6	*BRCA1*	c.4035delA	1,0541
2	GD11	*PMS2*	c.861_864delACAG	0,5763
3	GD10	*PMS2*	c.1144+1G>A	0,0889
4	GD21	*TSC2*	c.5068+27_5068+60del34	0.8760
5	GD20	*TSC2*	c.5068+27_5068+60del34	0.4392
6	GD18	*MLH1*	deletsioon	0.3702
7	GD19	*TSC2*	c.5068+27_5068+60del34	-0.1381
8	GD3	*BRCA1*	c.5329dupC	-0,6250
9	GD2	*BRCA1*	c.1840A>T	-0,6560
10	GD8	*APC*	c.7957delA	-0,6794
11	GD17	*VHL*	c.8_9ins18	-0,7469
12	GD9	*MSH2*	c.2131C>T	-0,9260
13	GD16	*NF2*	c.1702_1703delAG	-0,9806
14	GD13	*KCNH2*	c.2587C>T	-1,0871
15	GD4	*BRCA1*	c.5329dupC	-1,2867
16	GD1	*BRCA1*	c.4258C>T	-1,8472
17	GD5	*BRCA1*	c.4035delA	-2,0570

2.3.5. Splaissinguanalüüs

Esmast splaissinguanalüüsi teostati seitsme geeni (*VHL*, *NF2*, *PMS2*, *MSH2*, *BRCA1*, *APC* ja *KCNH2*) suhtes programmiga SplicingTypesAnno. Analüüs viidi läbi indiviidide kaupa (kokku 13 indiviidil) vastava uuritava geeni suhtes.

GD13 indiviidi *KCNH2* geeni splaissinguanalüüs ebaõnnestus madala ekspressioonitaseme tõttu. Vaatamata sellele, et *KCNH2* keskmine ekspressioonitase (FPKM=1,3007) oli määratud piirist kõrgem, oli uuritava indiviidi GD13 ekspressioonitase sellest palju madalam (FPKM=0,5034).

Allesjäänute 6 geeni splaissinguanalüüsi käigus tuvastati alternatiivset splaissingut ainult GD10 indiviidil *PMS2* geenis.

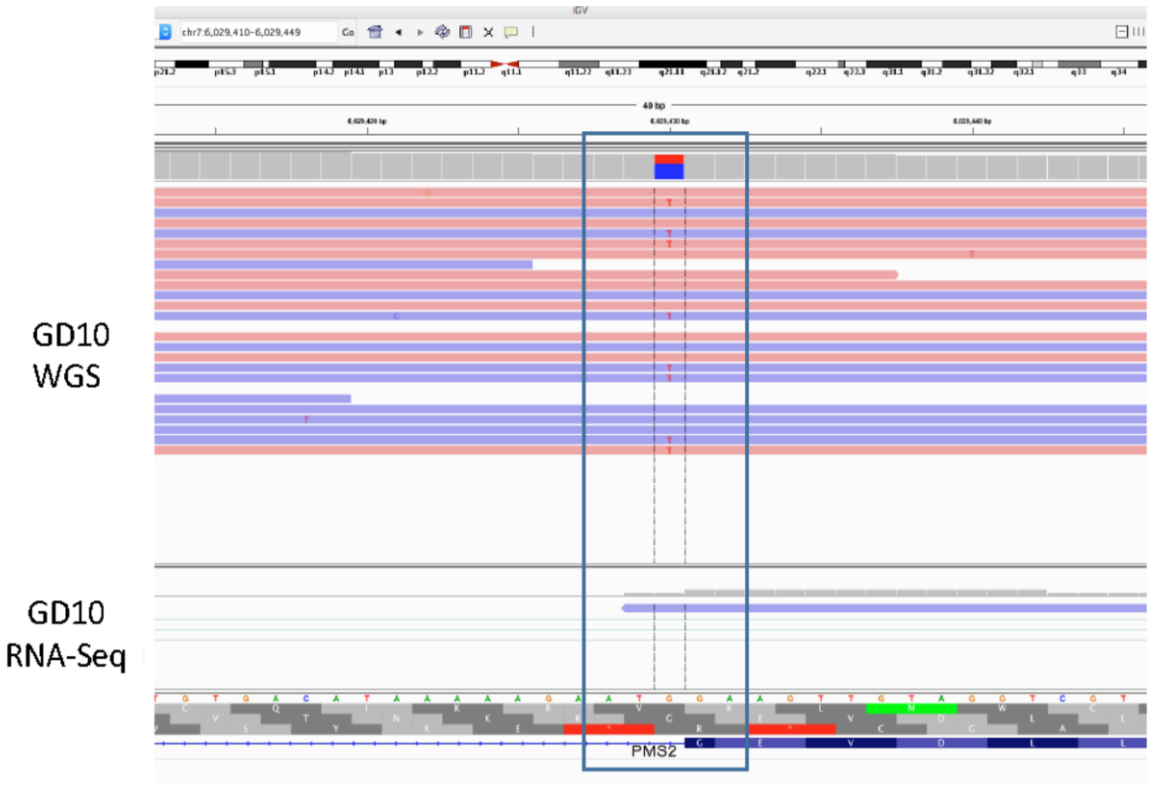
Samuti leiti alternatiivse splaissingu sündmused *MLH1* ja *TSC2* geenides uuritavatel indiviididel GD18, GD19, GD20 ja GD21.

2.3.5.1. *PMS2* geeni splaissinguanalüüs

SplicingTypesAnno tuvastas alternatiivse splaissingu indiviidil GD10, kellel oli WGS andmete põhjal leitud splaissingu doonorsaidi mutatsioon *PMS2* geenis positsioonis Chr7: 6029430 (c.1144+1G>A, Joonis 10). Doonorsaidi konsensusjärjestuseses (GT) 10. eksoni ja 10.introni piiril asuv guaniin (G) asendatakse adeniiniga (A).

On teada, et antud mutatsioon (c.1144+1G>A) põhjustab 10. eksoni vahelejätmist (van der Klift jt., 2015). Hendriks jt. uuringus on seostatud 10. eksoni vahelejätmist päriliku kolorektaalvähiga ehk Lynchi sündroomiga (Hendriks jt., 2006).

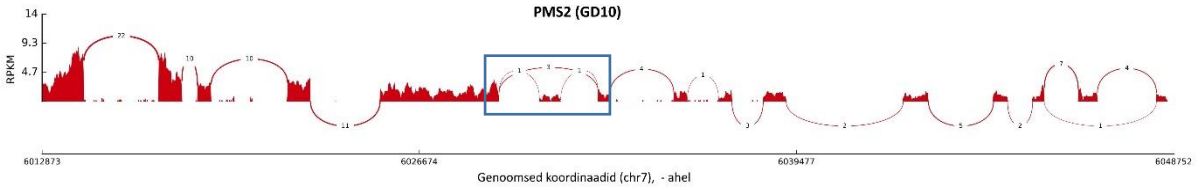
SplicingTypesAnno tuvastas võimaliku viimase introni säilitamise (*intron retention*) *PMS2* geenil (Tabel 5), kuid ei suutnud leida mutatsioonist tingitud eksoni vahelejätmist.



Joonis 10. Geenis *PMS2* indiviidil GD10 leitud splaissingu doonorsaidi mutatsioon c.1144+1G>A (märgitud kastiga) genoomi (WGS) ja transkriptoomi sekveneerimise (RNA-Seq) andmete põhjal. Joonis on tehtud IGV rakenduses.

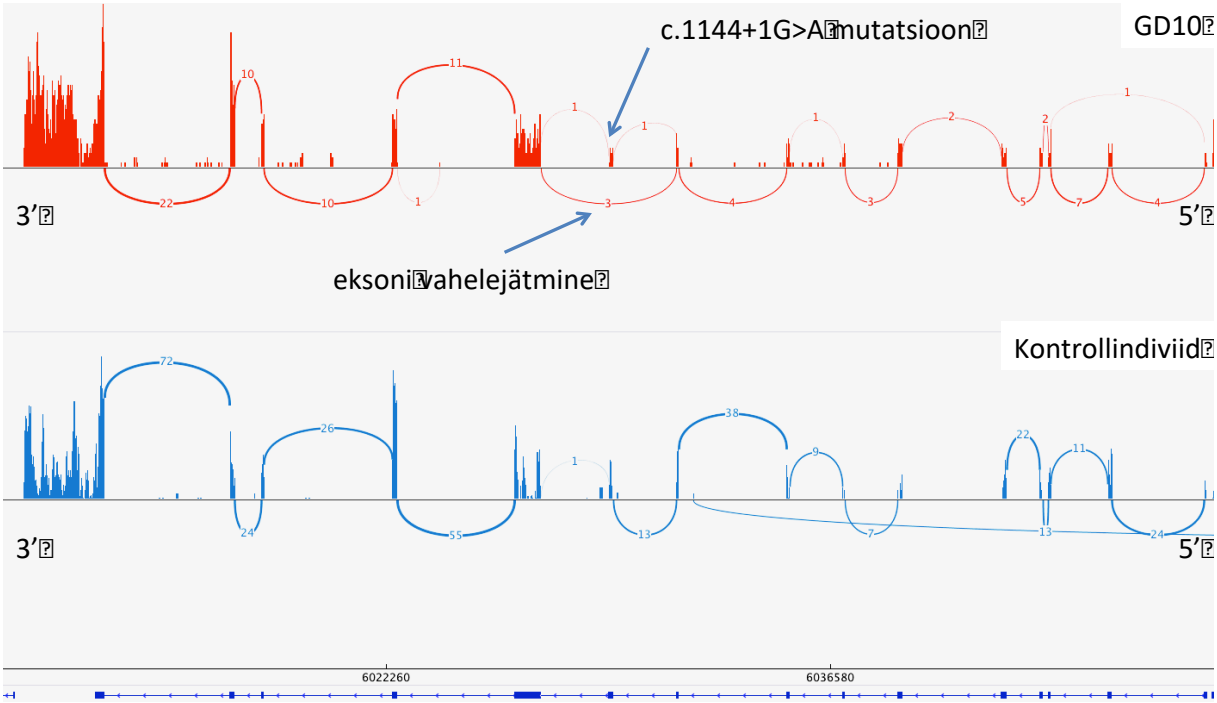
PMS2 introni säilitamise puhul on tegemist viimase introniga. Joonistel 11 ja 12 on näha üksikud lugemid, mis on joondatud viimase introni (asub joonisel vasakul pool) regiooni. Viimase introni säilimist mõnede transkriptide puhul võib selgitada sellega, et raaminihke tõttu tekkinud stoppkoodon jääb viimasesse eksonisse ning seetõttu ei lagundata sellist transkripti *nonsense-mediated decay* vahendusel. Samuti on näidatud, et splaissingu efektiivsus väheneb mRNA 3’ suunas (Tilgner jt., 2012).

Mutatsiooniga seotud 10. eksoni vahelejätmine tuvastati MISO programmi poolt kolme lugemiga (Joonis 11).



Joonis 11. *PMS2* geeni splaissinguanalüüs GD10 indiviidil MISO programmiga. Kastiga on tähistatud 10. eksoni vahelejätmine. mRNA 5’ ots asub joonisel paremal.

10. eksoni vahelejätmine indiviidil GD10 ilmnes ka RNA-Seq andmete kontrollil IGV rakenduses (Joonis 12). Hoolimata lugemite väikesest hulgast antud positsioonis, võib siiski oletada, et tegemist on mutatsiooni poolt põhjustatud sündmusega, kuna kontrollindiviidil sellist eksoni vahelejätmist ei toimu.



Joonis 12. *PMS2* splaissingumuster c.1144+1G>A mutatsiooniga indiviidil GD10 (ülemine paneel) ja ilma mutatsioonita kontrollindiviidil (alumine paneel) RNA-Seq andmete põhjal. Näidatud on mutatsiooni paiknemine ning 10. eksoni vahelejätmine. mRNA 5’ ots asub joonisel paremal. Joonis on tehtud IGV rakenduses.

2.3.5.2. *MLH1* geeni splaissinguanalüüs

Oletatav 9-nukleotiidne deletsioon *MLH1* geenis asub 11. introni ja 12. eksoni piiril (Joonis 5). Uuritav mutatsioon rikub splaissingu aktseptorsaidi ja selle mutatsiooni efekt peaks olema märgatav mRNA tasemel. *MLH1* keskmine ekspressioonitase (FPKM) oli 11,3 ning uuritaval indiviidil 12,6. Splaissinguanalüüs indiviidil GD18 teostati SplicingTypesAnno ja MISO programmidega.

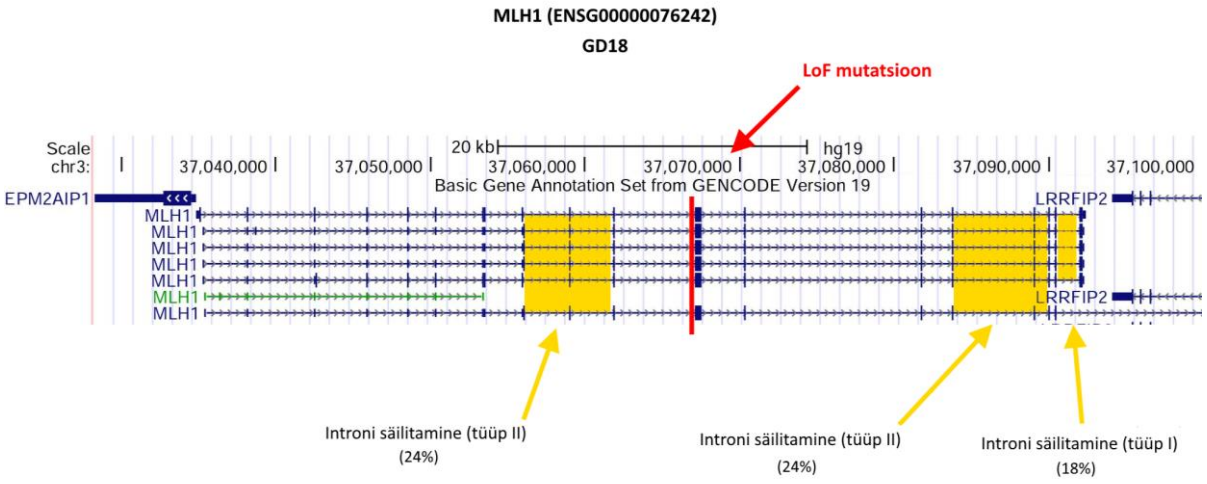
Uuritavale indiviidile valiti soo, vanuse, KMI (kehamassiindeks) ja teiste tunnuste järgi (suitsetamine, alkoholi tarvitamine) neli kontrollindiviidi.

SplicingTypesAnno programm tuvastas uuritaval indiviidil GD18 *MLH1* geenis kolm introni säilitamise sündmust (Tabel 6, Joonis 13). Antud juhul on introni säilitamise tõenäosus alla 25% ja selliseid tulemusi seostatakse tihtipeale mitte-splaissitud või osaliselt splaissitud pre-mRNA fraktsiooniga, mis võib vähesel määral esineda splaissitud mRNA-de hulgas (Galante jt., 2004). Kan jt. uuringus detekteeriti introni säilitamist 36% geenides (N=6 400), kuid alla 5% geenidest sisaldasid säilitatud intronit 95% usaldusväärsusega (Kan jt. 2002).

Tabel 6. SplicingTypesAnno programmiga saadud *MLH1* geeni analüüsi tulemused GD18 indiviidil.

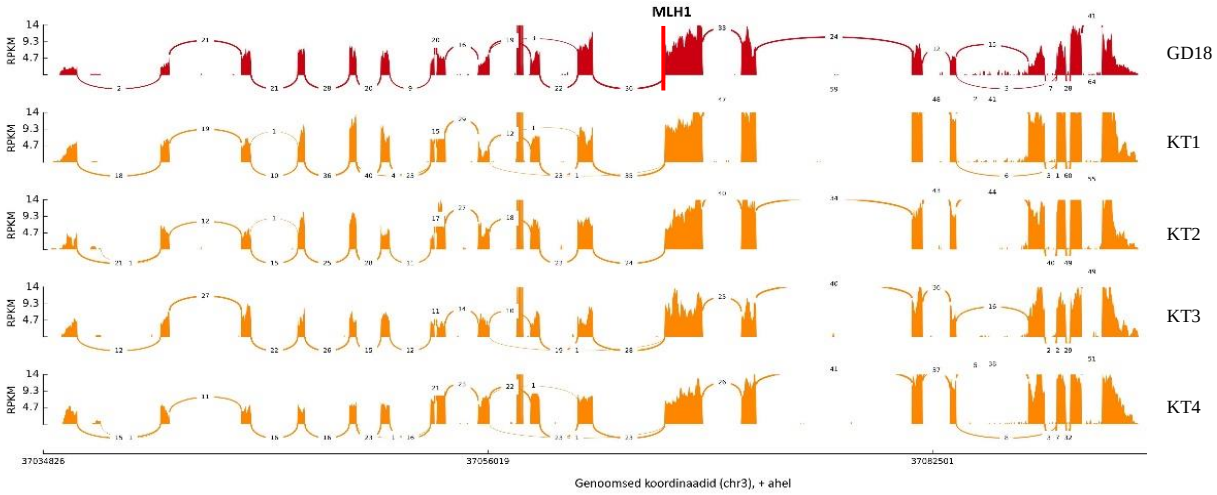
Indiviid	**Alternatiivse splaissingu tüüp**	**Genoomsed koordinaadid**	**Alternatiivse splaissingu tõenäosus**
GD18	Introni säilitamine (tüüp I)	chr3: 37090509-37091976 (intron 22)	18%
	Introni säilitamine (tüüp II)	chr3: 37056036-37061800 (intron 13, 14)	24%
	Introni säilitamine (tüüp II)	chr3: 37083823-37090007 (intron 19, 20)	24%

Samasugune analüüs viidi läbi ka kontrollindiviididel. Kõikidel kontrollindiviididel leiti 22. introni säilitamine (tõenäosusega 21%, 28%, 16% ja 31%). Kolmel kontrollindiviidil leiti 19. introni säilitamine (tõenäosusega 27%, 6% ja 12%). Kontrollindiviidide splaissinguanalüüsi tulemused on esitatud Lisas 1.



Joonis 13. Indiviidi GD18 *MLH1* geeni struktuur. Märgitud on funktsioonikaoga mutatsiooni koht ja SplicingTypesAnno programmiga leitud splaissingusündmuste asukohad ja nende tõenäosus.

Funktsioonikaoga mutatsioon, mis asub splaissingu aktseptorsaidis, peaks otseselt mõjutama splaissingut. Siiski ei tuvastanud *MLH1* geeni analüüs splaissingumustri muutusi, mis oleksid põhjustatud antud mutatsiooni poolt (Joonised 13, 14). Selline tulemus viitab võimalusele, et 9-nukleotiidi pikkune deletsioon positsioonis Chr3:37067120 indiviidil GD18 on genoomsete variantide annoteerimise viga ning tegelikult genoomis ei esine. Kindlasti vajab see hüpotees kontrollimist Sangeri sekveneerimise meetodiga.



Joonis 14. *MLH1* geeni võrdlev splaissinguanalüüs MISO programmiga. Punasega on toodud uuritava indiviidi GD18 geenistruktuur, oranžiga on toodud 4 kontrollindiviidi (KT1, KT2, KT3, KT4) geenistruktuur. Punase joonega on tähistatud funktsioonikaoga mutatsiooni asukoht GD18 indiviidil.

MLH1 geeni splaissingut iseloomustab laialdane mittepatogeensete alternatiivsete isovormide esinemine (Thomson jt., 2015). Üle 60% MLH1 transkripti sisaldavad eksoni vahelejätmist. Kõige levinumad alternatiivsed transkriptid sisaldavad 6., 16., 17. või samaaegselt 9. ja 10. eksoni vahelejätmist (Thomson jt., 2015). 2. või samaaegselt 6. ja 7. eksoni vahelejätmist seostatakse päriliku kolorektaalvähiga ehk Lynchi sündroomiga (Clarke jt., 2000; Tanko jt., 2002).

GD18 ja kontrollindiviidide splaissinguanalüüs MISO programmiga näitas 6. eksoni vahelejätmist kontrollindiviididel KT1 ja KT4 (seda toetasid vastavalt 4 ja 1 lugemit), 10. eksoni vahelejätmist GD18 ja kontrollindiviididel KT1 ja KT4 (sündmust toetasid vastavalt 3, 1 ja 1 lugemit), 10. ja 11. eksoni vahelejätmist kontrollindiviididel KT1, KT3 ja KT4 (sündmust toetasid vastavalt 1, 3 ja 4 lugemit), 16. eksoni vahelejätmist GD18 ja kontrollindiviididel KT1 ja KT4 (sündmust toetasid vastavalt 3, 6 ja 8 lugemit) ja 17. eksoni vahelejätmist kontrollindiviididel KT1 ja KT4 (sündmust toetasid vastavalt 1 ja 7 lugemit) (Joonis 14). Samuti on joonisel 14 näha alternatiivse esimese eksoni olemasolu kõikidel proovidel. Üks lugem kontrollindiviididel KT2 ja KT4 toetab alternatiivse esimese eksoni kasutamist.

2.3.5.3. *TSC2* geeni splaissinguanalüüs

Deletsioon *TSC2* geenil asub 39. eksoni ja 39. introni piiril ja hõlmab splaissingu doonorsaidi konsensusjärjestust (Joonis 6). Deletsioon 39. introni doonorsaidis põhjustab selle introni säilitamist (Roberts jt., 2003).

Umbes 33% *TSC2* patogeenseid splaissingu variante asuvad eksonites 32 - 41, mis kodeerivad tuberiini valgu karboksüterminaalse domääni (Northrup jt., 2015). Patogeenseid mutatsioone *TSC2* geenis seostatakse tuberoosse skleroosiga (*tuberous sclerosis*). Uuritav mutatsioon (rs137854209) ei ole patogenne, kuid toimib modifikaatorgeenina tuberoosse skleroosiga patsientidel (Roberts jt., 2003).

Geeni splaissinguanalüüs indiviididel GD19, GD20, GD21 teostati SplicingTypesAnno ja JunctionSeq programmidega. SplicingTypesAnno tuvastas rohkelt introni säilitamise juhtumeid, mille esinemistõenäosus oli samuti kõrge (Tabel 7, Joonis 15).

Samasugune splaissinguanalüüs viidi läbi ka kontrollindiviididel, kes olid valitud soo, vanuse, KMI ja teiste tunnuste järgi (suitsetamine, alkoholi tarvitamine, naiste puhul võeti arvesse menstruaaltsükli pikkus, raseduste ja sünnituste arv, menopausi saabumise aeg).

Kontrollindiviididel tuvastati samuti alternatiivse splaissingu sündmuseid, mille hulgas olid introni säilitamine ja alternatiivne doonorsait. Kontrollindiviidide splaissinguanalüüsi tulemused on esitatud Lisas 2.

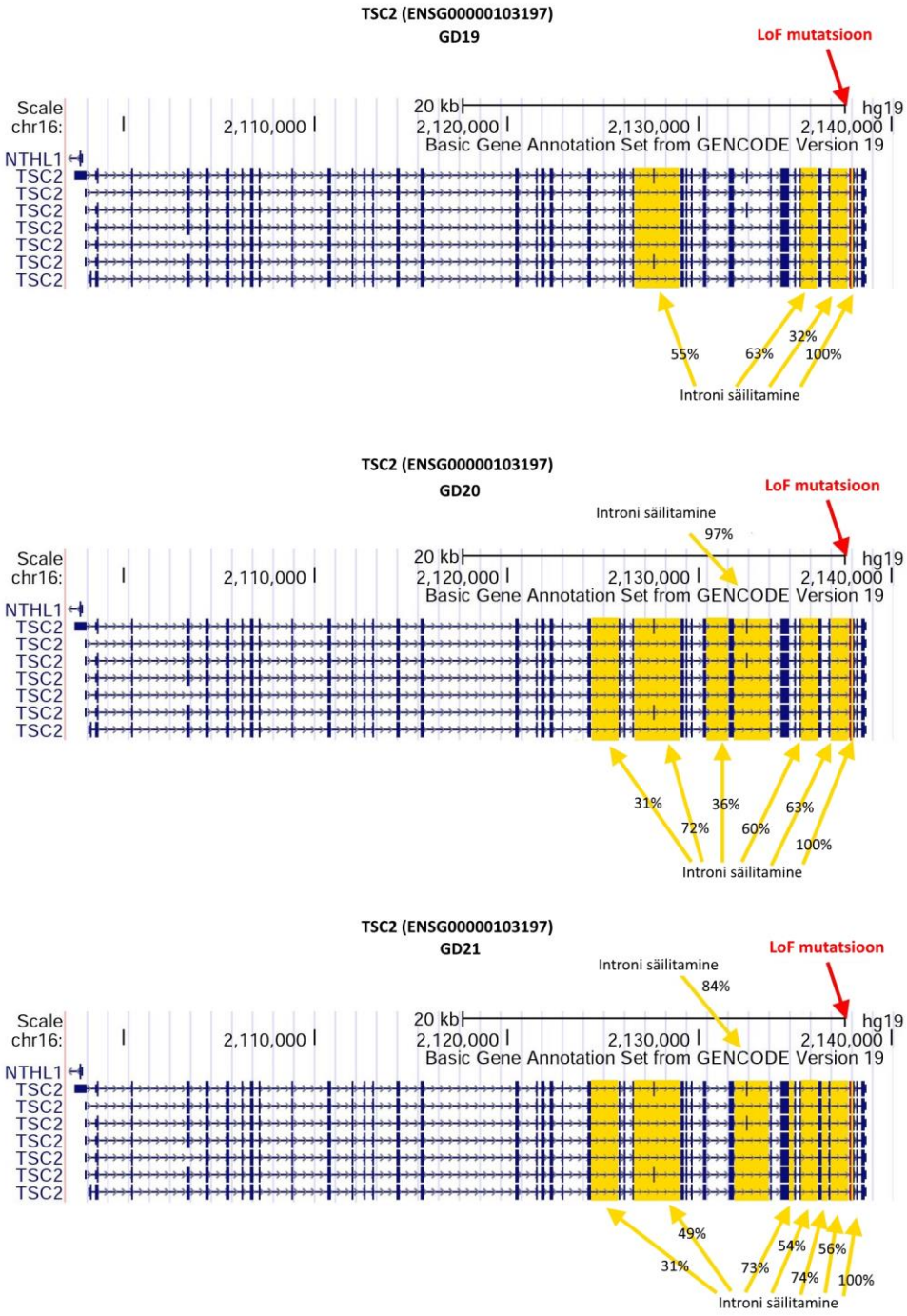
Kõikides uuritavates proovides (GD19, GD20, GD21) tuvastati 41. introni säilitamine, mille esinemise tõenäosus oli 100%. Kontrollindiviididel sellist alternatiivse splaissingu sündmust ei esinenud. Saadud tulemus on kooskõlas Roberts jt. uuringu tulemustega. Erinevus introni järjekorranumbris on tingitud erinevate mRNA isovormide kasutamisest.

Tabel 7. Uuritavate indiviidide (GD19, GD20, GD21) *TSC2* geeni SplicingTypesAnno tulemused.

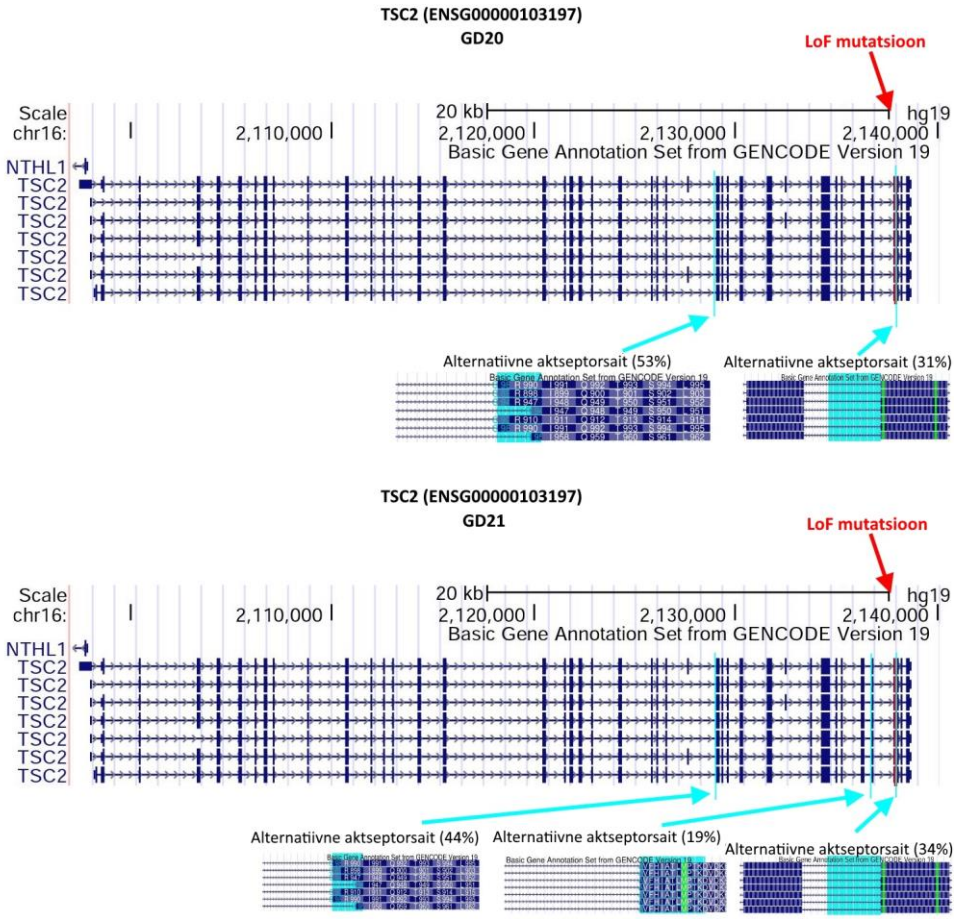
Indiviid	**Alternatiivse splaissingu tüüp**	**Genoomsed koordinaadid**	**Alternatiivse splaissingu tõenäosus**
GD 19	Introni säilitamine (tüüp I)	chr16: 2135324-2136193 (intron 38)	63%
	Introni säilitamine (tüüp I)	chr16: 2136873-2137863 (intron 40)	32%
	Introni säilitamine (tüüp I)	chr16: 2137943-2138048 (intron 41)	100%
	Introni säilitamine (tüüp II)	chr16: 2126587-2129035 (intronid 28, 29)	55%
GD 20	Introni säilitamine (tüüp I)	chr16: 2124391-2125799 (intron 25)	31%
	Introni säilitamine (tüüp I)	chr16: 2130379-2131595 (intron 32)	36%
	Introni säilitamine (tüüp I)	chr16: 2135324-2136193 (intron 38)	60%
	Introni säilitamine (tüüp I)	chr16: 2136873-2137863 (intron 40)	63%
	Introni säilitamine (tüüp I)	chr16: 2137943-2138048 (intron 41)	100%
	Introni säilitamine (tüüp II)	chr16: 2126587-2129035 (intronid 28, 29)	72%
	Introni säilitamine (tüüp II)	chr16: 2131800-2133695 (intron 33, 34)	97%
	Alter.aktseptorsait (3’)	chr16: 2129033-2129036 (ekson 27)	53%
	Alter.aktseptorsait (3’)	chr16: 2138049-2137977	31%

Tabel 7 järg

		(ekson 39)	
GD21	Introni säilitamine (tüüp I)	chr16: 2124391-2125799 (intron 25)	31%
	Introni säilitamine (tüüp I)	chr16: 2134463-2134967 (intron 36)	73%
	Introni säilitamine (tüüp I)	chr16: 2135324-2136193 (intron 38)	54%
	Introni säilitamine (tüüp I)	chr16: 2136381-2136767 (intron 39)	74%
	Introni säilitamine (tüüp I)	chr16: 2136873-2137863 (intron 40)	56%
	Introni säilitamine (tüüp I)	chr16: 2137943-2138048 (intron 41)	100%
	Introni säilitamine (tüüp II)	chr16: 2126587-2129035 (intronid 28, 29)	49%
	Introni säilitamine	chr16: 2131800-2133695 (intron 33, 34)	84%
	Alter.aktseptorsait (3’)	chr16: 2129033-2129036 (ekson 27)	44%
	Alter.aktseptorsait (3’)	chr16: 2138049-2137977 (ekson 39)	34%
	Alter.aktseptorsait (3’)	chr16: 2136733- 2136768 (ekson 37)	19%



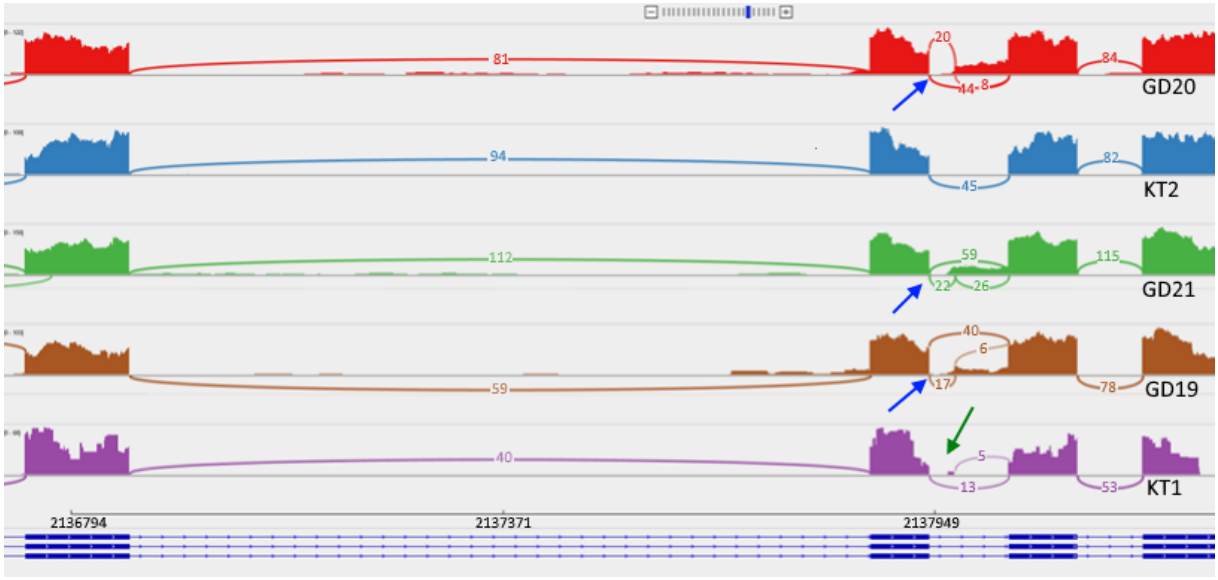
Joonis 15. Skemaatiline *TSC2* geeni struktuur, kus on märgitud indiviidide GD19, GD20 ja GD21 funktsioonikaoga mutatsiooni koht ja SplicingTypesAnno programmiga leitud introni säilitamise sündmused.



Joonis 16. Skemaatiline *TSC2* geeni struktuur, kus on märgitud indiviidide GD19, GD20 ja GD21 funktsioonikaoga mutatsiooni koht ja SplicingTypesAnno programmiga leitud alternatiivsed splaissingusaidid.

Splaissinguanalüüsi käigus tuvastati GD20 ja GD21 indiviididel alternatiivsed aktseptorsaidid 27. ja 39. eksonis. Indiviidil GD21 oli lisaks veel 37. eksonis alternatiivne aktseptorsait. Kõikidel uuritavatel indiviididel kontrolliti IGV rakenduses ka regiooni, kus SplicingTypesAnno tuvastas 41. introni säilitamise 100% tõenäosusega (Joonis 17). On näha, et ainult umbes pooled toorlugemid toetavad introni säilitamist. Vastuolu SplicingTypesAnno tulemuste ja IGV andmete vahel võib olla põhjustatud splaissingu doonorsaiti hõlmavast deletsioonist (Joonis 6), mis ei võimalda programmil korrektset analüüsi läbi viia.

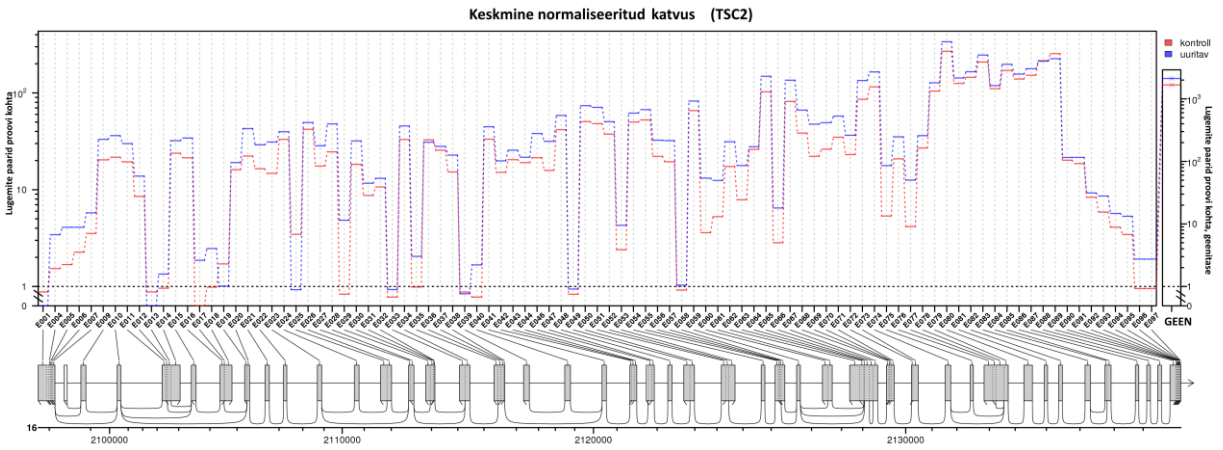
Kontrollindiviidil KT1 tuvastati alternatiivse doonorsaidi esinemist 39. eksonis tõenäosusega 25%. Joonisel 17 on näha antud doonorsaidi esinemist ja selle kasutamist.



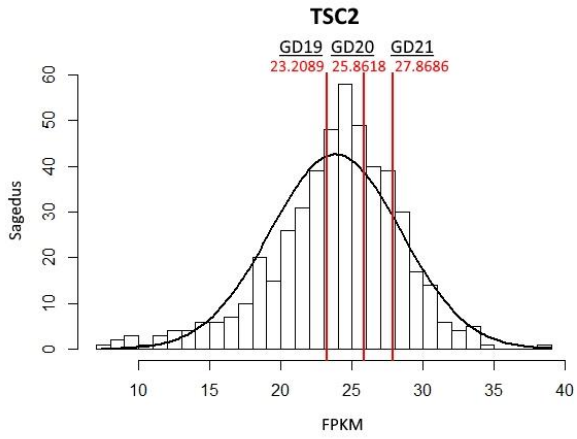
Joonis 17. *TSC2* geeni uuritav regioon indiviididel GD19, GD20, GD21 ja kontrollindiviididel KT1 ja KT2. *TSC2* geeni alternatiivse splaissingu regioon on märgitud sinise noolega. Tuvastatud alternatiivne doonorsait KT1 kontrollindiviidil on märgitud rohelise noolega. Joonis on tehtud IGV rakenduses.

Uuritavatel indiviididel on selgelt näha uue alternatiivse isovormi teket 39. intronis (41. intronis SplicingTypesAnno tulemuste järgi). Introni säilitamine valkukodeerivas alas põhjustab tõenäoliselt lugemisraami nihke või enneaegse stoppkoodoni tekke, mis võib märgatavalt mõjutada mRNA ekspressiooni. *TSC2* ekspressioonianalüüsi andmete põhjal uuritav mutatsioon geeni ekspressioonitaset ei mõjuta. *TSC2* geeni ekspressioonitase mutatsiooniga indiviididel on võrreldav ilma mutatsioonita indiviidide omaga (Joonis 19).

Splaissinguanalüüsi *TSC2* geenil teostati ka JunctionSeq programmiga. Selle jaoks moodustati kaks rühma: uuritavad, kuhu kuulusid indiviidid GD19, GD20, GD21 ja kontrollid, kuhu kuulusid kuus kontrollindiviidi. JunctionSeq kontrollis isovormide diferentsiaalset kasutust kahe rühma vahel. Antud analüüsi puhul ei leidnud programm statistiliselt olulisi erinevusi uuritavate ja kontrollide vahel (Joonis 18). Graafikul on näha väikesed erinevused positsioonides E060-E061, E075, mis vastavad eksonitele 27. – 28. ja 30.



Joonis 18. *TSC2* geeni alternatiivsete ja konstitutiivsete eksonite ekspressioonitasemed uuritavate ja kontrollrühma vahel. Graafiku alumisel paneelil on toodud kõik teadaolevad eksonivariandid *TSC2* geenil vastavalt nende genoomsetele koordinaatidele. Graafiku põhipaneelil on toodud iga testitud eksoni keskmine ekspressioonitase uuritaval (sinine) ja kontrollrühmal (punane). Graafiku parempoolsel paneelil on toodud keskmine *TSC2* ekspressioonitase uuritavatel ja kontrollrühmal.



Joonis 19. *TSC2* geeni ekspressioon. Histogrammil on toodud geeni ekspressioontaseme jaotus kogu RNA-Seq valimi indiviididel. Punaste joontega on näidatud ekspressioonitase uuritavatel indiviididel. Toodud on ka vastavad FPKM numbrilised väärtused.

2.3.6. Alleel-spetsiifiline ekspressioon

Antud uuringu jaoks kasutati ainult eksonites asuvaid SNP-e. Analüüsi käigus mõõdab programm lugemite arvu, mis toetavad ühte või teist alleeli. Minimaalseks lugemite arvuks ASE analüüsiks määrati 5 lugemit. Heterosügootse variandi korral on oodatav, et mõlema alleeli esindatus on võrdne. Homosügootse variandi puhul peaksid kõik lugemid toetama ühte kindlat alleeli. Homosügootse variandi esinemine võib olla põhjustatud alleel-spetsiifilisest ekspressioonist.

ASE analüüs viidi läbi *TSC2, MLH1* ja *PMS2* geenides.

TSC2

ASE analüüs viidi läbi kolmel indiviidil (GD19, GD20, GD21), kellel esines mutatsioon *TSC2* geenis (Tabel 8). Kuus ASE markerit 84-st paiknesid eksonites. Ainult kahel ASE markeril oli lugemite arv määratud piirist kõrgem kõikides uuritavates proovides. Uuritavad indiviidid olid kõikides markerite positsioonides homosügootsed.

Tabel 8. *TSC2* geeni alleelspetsiifilise ekspressiooni markerid uuritavatel indiviididel GD19, GD20, GD21. Iga SNP-i kohta on toodud referentsalleeliga lugemite arv / alternatiivse alleeliga lugemite arv / alternatiivse ega referentsalleeli mittesisaldavate lugemite arv. Nurksulgudes on toodud uuritava indiviidi genotüüp vastavas positsioonis.

ASE marker	Markeri positsioon	Refrents-alleel	Altern. alleel	GD19	GD20	GD21
rs34012042	2114407	T	C	0/18/0[C,C]	0/29/0[C,C]	0/19/0[C,C]
rs13337626	2125834	T	C	19/0/0[T,T]	39/0/0[T,T]	38/0/0[T,T]
rs45474795	2134508	G	T	35/0/0[G,G]	0/0/0[G, G]	0/0/0[G, G]
rs1748	2138269	T	C	85/0/1[T,T]	0/0/0[T, T]	0/0/0[T, T]
rs45517419	2138546	G	A	76/0/0[G,G]	0/0/0[A, G]	0/0/0[G, G]
rs1051771	2138584	G	C	26/0/0[G,G]	0/0/0[G, G]	0/0/0[G, G]

MLH1

Indiviidi GD18 uuriti *MLH1* geeni ASE suhtes (Tabel 9). Eksonites paiknesid kaks ASE markerit, mille ekspressioon oli määratud piirist kõrgem. Uuritav indiviid ja kaks kontrolli olid ühes positsioonis (rs1799977) heterosügootsed. Kuigi referentsalleeli lugemite arvu suhe mõlema alleeli lugemite summa suhtes (REF/(ALT + REF)), kus REF tähistab referentsalleeli lugemite arvu ja ALT alternatiivse alleeli lugemite arvu) oli kõrgem kui oodatav (vastavalt 0,67 ja 0,5), pole tulemus statistiliselt oluline ($p = 0.2379$, täpne test hüpoteeside testimiseks tõenäosuse kohta, *exact binomial test*). Samuti oleks ainult ühe SNP põhjal 2600 nukleotiidi pikkuse mRNA kohta järelduste tegemine ennatlik.

Tabel 9. *MLH1* geeni alleelspetsiifilise ekspressiooni markerid uuritaval GD18 ja kontrollindiviididel. Iga SNP-i kohta on toodud referentsalleeliga lugemite arv / alternatiivse alleeliga lugemite arv / alternatiivse ega referentsalleeli mittesisaldavate lugemite arv. Nurksulgudes on toodud uuritava indiviidi genotüüp vastavas positsioonis.

ASE marker	Markeri positsioon	Referents-alleel	Altern. alleel	GD18	Kontroll 1	Kontroll 2
rs1799977	37053568	G	A	12/6/0[G,A]	10/16/0[G,A]	19/16/0[G,A]
rs1800146	37090070	G	T	25/0/0[G,G]	21/22/0[T,G]	51/0/0[G,G]

PMS2

PMS2 geeni ASE analüüs teostati indiviidil GD10 (Tabel 11). Eksonite piirkonnas paiknesid kolm ASE markerit, mille ekspressioon oli määratud piirist kõrgem. Uuritavad indiviidid olid kõikides markerite positsioonides homosügootsed.

Tabel 10. *PMS2* geeni alleelspetsiifilise ekspressiooni markerid uuritaval GD10 ja kontrollindiviididel. Iga SNP-i kohta on toodud referentsalleeliga lugemite arv / alternatiivse alleeliga lugemite arv / alternatiivse ega referentsalleeli mittesisaldavate lugemite arv. Nurksulgudes on toodud uuritava indiviidi genotüüp vastavas positsioonis.

ASE marker	Markeri positsioon	Refrents-alleel	Altern. alleel	GD10	Kontroll 1	Kontroll 2
rs1805324	6026530	C	T	6/0/0[C,C]	11/0/0[C,C]	8/0/0[C,C]
rs63750668	6026708	C	A	8/0/0[C,C]	17/0/0[C,C]	11/0/0[C,C]
rs2228007	6026865	T	C	5/0/0[T,T]	7/0/0[T,T]	5/0/0[T,T]

Kuna *TSC2* ja *PMS2* geenides olid kõik ASE markerid homosügootsed nii uuritavatel kui ka kontrollindiviididel, siis nendel geenidel alleel-spetsiifilist ekspressiooni analüüsida ei olnud võimalik. *MLH1* geenis oli küll üks ASE marker heterosügootne, kuid alleelide ekspressiooni erinevus ei olnud statistiliselt oluline.

KOKKUVÕTE

Käesoleva töö eesmärgiks oli uurida meditsiiniliselt oluliste geenivariantide mõju transkriptoomi tasemel. Kirjanduse ülevaate osas kirjeldati splaissingu mehhanismi, regulatsiooni ja selle mõju mRNAle. Eksperimentaalses osas hinnati mRNA ekspressioonitaseme ja splaissingumustri muutusi Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) geenidoonoritel, kellel esinesid mutatsioonid kliiniliselt olulistes geenides ACMG ja Geisinger geeninimekirjade järgi. Uuringusse valiti 21 indiviidi, kellel esinesid mutatsioonid 13 erinevas geenis. Kõikidel uuritavatel mutatsioonidel oli potentsiaalne mõju transkriptoomile, kuna nende hulgas olid raaminihke, enneaegse stoppkoodoni ja splaissingusaitide mutatsioonid. Uuritavate indiviidide transkriptoomi analüüs põhines RNA sekveneerimise käigus saadud andmetel. Ekspressioonianalüüsi käigus leiti, et paljudel kliiniliselt olulistel geenidel on ekspressioonitase veres üsna madal ja need ei ole piisavalt informatiivsed järgnevaks splaissinguanalüüsiks. Alternatiivse splaissingu sündmused tuvastati viiel indiviidil. Alleel-spetsiifilist ekspressiooni (ASE) ei leitud, kuna peaaegu kõik uuritavad ASE markerid olid homosügootsed ning ainuke heterosügootne marker ei näidanud statistiliselt olulist erinevust.

Splaissingu uurimiseks kasutati SplicingTypesAnno, MISO ja JunctionSeq programme. Igal programmil on omad eelised ja puudused. SplicingTypesAnno on tundlik intronite säilitamise ja alternatiivsete splaissingusaitide tuvastamise suhtes, kuid ei suutnud leida eksonite vahelejätmist. Samuti puuduvad programmil tulemuste visualiseerimisvõimalused. MISO on laialt kasutatav programm alternatiivsete isovormide tuvastamiseks. Programm sobib hästi eksonite kaasamise ja vahelejätmise tuvastamiseks. JunctionSeq on uuem programm erinevate transkripti isovormide diferentsiaalse kasutamise tuvastamiseks kahe uuritava rühma vahel. Antud uuringus kasutati seda programmi vähe, kuna uuritavate indiviidide valim oli väike.

Antud uurimustöö tulemustest võib järeldada, et RNA-Seq on küll sobiv kliiniliselt oluliste geneetiliste variantide mõju tuvastamiseks transkriptoomi tasemel, kuid eelduseks on uuritava geeni piisavalt kõrge ekspressioonitase uuritavas koes.

The interpretation of clinically relevant genetic variants through RNA-Seq expression data analysis

Ustina Põnova

Summary

The human genome encodes approximately 20,000 – 25,000 protein-coding genes. The Human Proteome Map project has identified more than 30,000 proteins coded by approximately 17,300 human genes. It is estimated that the diversity of these protein-coding genes in humans is greatly increased by the presence of numerous protein isoforms. Alternative splicing is a pre-mRNA processing mechanism, which is used to generate multiple mRNA transcripts. Approximately 95% of human genes undergo alternative splicing. Misregulation of splicing by mutations that affect the splicing signals or the splicing machinery itself is the cause of multiple human diseases. Up to 50% of disease-causing mutations are found to affect splicing. Due to development of the RNA sequencing (RNA-Seq) technology, it has become possible to analyze and interpret changes in the splicing pattern and transcript expression.

The purpose of the practical part of the current Master's thesis was to explore the impact of clinically relevant genetic variants on the transcriptome level. The transcript splicing and expression levels were measured for 21 gene donors from Estonian Genome Center, University of Tartu. All the studied individuals had mutations in clinically relevant genes according to ACMG (American College of Medical Genetics and Genomics) and Geisinger Health System's clinically actionable gene lists.

During the expression analysis, FPKM (fragments per kilobase of transcript per million mapped reads) mean level were measured for each studied gene and individual FPKM level for each studied gene per gene donor. From the expression analysis we concluded that many clinically relevant genes have low expression in the blood samples and are not suitable for subsequent splicing analysis. In the further analysis 9 genes were selected, which had mean FPKM level higher than 1.

Splicing analyses were performed by using SplicingTypesAnno, MISO and JunctionSeq programs. For this analysis 17 individuals were selected, who had mutations in nine different genes. All the studied mutations (frameshift, premature stop codon, splice site mutations) could potentially have an impact on the transcript level.

Alternative splicing events were found in five probes. Among detected alternative splicing types were exon skipping, intron retention and alternative splicing sites.

In conclusion, RNA-Seq is a powerful tool for detecting the impact of clinically relevant genetic variants on the transcript level. For accurate analysis results, it is better to use genes with a higher expression level.

TÄNUSÕNAD

Soovin tänada oma juhendajaid Tarmo Annilot ja Viktorija Kukuškinat suurepärase juhendamise, väärtuslike nõuannete, huvitavate arutelude ning abivalmiduse eest.

Suured tänud kõigile Geenivaramu inimestele (eelkõige Kelli Grand, Mart Kals, Neeme Tõnisson, Tiit Nikopensius, Reedik Mägi, Karmen Vaiküll ja Marili Palover), kes aitasid kaasa erinevates töö etappides. Lisaks tänan oma peret ja lähedasi suure toetuse eest töö valmimisel.

KIRJANDUSE LOETELU

Adler, A.S., McCleland, M.L., Yee, S., ... , Firestein, R. (2014). An integrative analysis of colon cancer identifies an essential function for PRPF6 in tumor growth. Genes Dev. 28: 1068–1084.

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, R., Walter P. 2015 Molecular Biology of the Cell (6th ed.). New York: Garland Science Taylor & Francis.

Anders S., Reyes, A., Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. 22(10): 2008–2017.

Andrews S. (2015). FastQC: a quality-control tool for high-throughput sequence data. Saadaval veebilehel http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bánky, B., Rásó-Barnett, L., Barbai, T., Tímár, J., Becságh, P., Rásó, E. (2012). Characteristics of CD44 alternative splice pattern in the course of human colorectal adenocarcinoma progression. Mol Cancer.11:83.

Bartoletti-Stella, A., Gasparini, A., Giacomini, C., ... , Capellari, S. (2015). Messenger RNA processing is altered in autosomal dominant leukodystrophy. Hum. Mol. Genet. 24: 2746–2756.

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology. 34: 525–527.

Byron, S.A., Van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., Craig, D.W. (2016). Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. Nat. Rev. Genet. 17: 257–271.

Cieply, B., Carstens, R.P. (2015). Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA. 6(3): 311–326.

Clarke, L.A., Veiga, I., Isidro, G., Jordan, P., Ramos, J.S., Castedo, S., Boavida, M.G. (2000). Pathological exon skipping in an HNPCC proband with MLH1 splice acceptor site mutation. Genes Chromosomes Cancer. 29: 367–370.

Consortium IHGS. (2004). Finishing the euchromatic sequence of the human genome. Nature.431: 931–945.

Cummings, B.B., Marshall, J.L., Tukiainen, T., ... , MacArthur, D.G. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med.9(386).

Das, R., Dufu, K., Romney, B., Feldt, M., Elenko, M., Reed, R. (2006). Functional coupling ofRNAP II transcription to spliceosome assembly. Genes Dev. 20: 1100–1109.

Dehm, S.M., Schmidt, L.J., Heemers, H.V., Vessella, R.L., Tindall, D. J. (2008). Splicing of a novel androgen receptor exon generates a constitutively active androgen receptor that mediates prostate cancer therapy resistance. Cancer Res. 68: 5469–5477.

Dewey, F.E., Murray, M.F., Overton, J..D., ... , Carey, D.J. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR Study. Science. 354(6319)

Galante, P.A., Sakabe, N.J., Kirschbaum-Slager, N., de Souza, S.J. (2004). Detection and evaluation of intron retention events in the human transcriptome. RNA. 10: 757–765.

Gambino, G., Tancredi, M., Falaschi, E., Aretini, P., Caligo, M. A. (2015). Characterization of three alternative transcripts of the BRCA1 gene in patients with breast cancer and a family history of breast and/or ovarian cancer who tested negative for pathogenic mutations. Int. J. Mol. Med. 35: 950–956.

Green, R.C., Berg, J.S., Grody, W.W., … , American College of Medical Genetics and Genomics. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 15: 565–574.

Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., ... , Sankaran, V.G. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc Natl Acad Sci U S A. 114(3): E327-E336.

Hartley, S.W., Mullikin, J.C. (2015). QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. BMC Bioinformatics.16:224.

Hartley, S.W., Mullikin, J.C. (2016). Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res.44(15): e127.

Hendriks, Y.M., Jagmohan-Changur, S., van der Klift, H.M., ..., Wijnen, J.T. (2006). Heterozygous mutations in PMS2 cause hereditary nonpolyposis colorectal carcinoma (Lynch syndrome). Gastroenterology.130(2): 312-22.

Hindorff, L. A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA.106: 9362–9367.

Jurica, M.S., Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. Mol Cell. 12: 5–14.

Kan, Z., States, D., Gish, W. (2002). Selecting for functional alternative splices in ESTs. Genome Res.12(12): 1837-45.

Katz, Y., Wang, E.T., Airoldi, E.M., Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods.7(12): 1009-15.

Katz, Y., Wang, E.T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., Airoldi, E.M., Burge, C.B. (2015). Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics. 31(14): 2400-2402.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., Stamm, S. (2013) Function of alternative splicing. Gene. 514: 1–30.

Keren, H., Lev-Maor, G., Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet.11(5): 345-55.

Kim, D., Langmead, B., Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements.Nat Methods.,12(4): 357-60.

Kim, M.S., Pinto, S.M., Getnet, D. (2014). A draft map of the human proteome. Nature. 509: 575-581.

van der Klift, H.M., Jansen, A.M., van der Steenstraten, N., Bik, E.C., Tops, C.M., Devilee, P., Wijnen, J.T. (2015). Splicing analysis for exonic and intronic mismatch repair gene variants associated with Lynch syndrome confirms high concordance between minigene assays and patient RNA analyses. Mol Genet Genomic Med.3(4): 327-45.

Kornblihtt, A.R. (2006). Chromatin, transcript elongation and alternative splicing. Nat. Struct. Mol. Biol. 13: 5–7.

Kornblihtt, A.R., Schor, I.E., Allo, M., Dujardin, G., Petrillo, E., Muñoz, M.J. (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nature. 14: 153–165.

Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., Cooper, D.N. (2007) Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum. Mutat. 28: 150–158.

La Cognata, V., D'Agata, V., Cavalcanti, F., Cavallaro, S. (2015). Splicing: is there an alternative contribution to Parkinson's disease? Neurogenetics. 16: 245–263.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. 25: 2078-9.

Liu, F., Gong, C. X. (2008). Tau exon 10 alternative splicing and tauopathies. Mol. Neurodegener. 3: 8.

Love, M.I., Huber, W., Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 15: 550.

Magri, F., Del Bo, R., D'Angelo, M.G., ... , Comi, G.P. (2011). Clinical and molecular characterization of a cohort of patients with novel nucleotide alterations of the Dystrophin gene detected by direct sequencing. BMC Med. Genet. 12: 37.

Mitt, M., Kals, M., Pärn, K., ... , Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet. doi: 10.1038/ejhg.2017.51.

Northrup, H., Koenig, M.K., Pearson, D.A., Au, K-S. (1999 [Last updated 2015 Sep 3]). Tuberous Sclerosis Complex. In: Pagon, R.A., Adam, M.P., Ardinger, H.H., et al., editors. GeneReviews® [Internet].

Saadaval veebilehel: https://www.ncbi.nlm.nih.gov/books/NBK1220/

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 40: 1413–1415.

Peters, M.J., Joehanes, R., Pilling, L.C., … , Johnson, A.D. (2015). The transcriptional landscape of age in human peripheral blood. Nat Commun., 6: 8570.

Reardon, D. A., Schuster, J, Tran, D.D., ... , Hampton, N.J. (2015). 107 ReACT: overall survival from a randomized Phase II study of rindopepimut (CDX-110) plus Bevacizumab in relapsed glioblastoma. Neurosurgery. 62 (Suppl. 1): 198–199.

Rivas, M.A., Pirinen, M., Conrad, D.F., … , MacArthur, D.G. (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 348(6235): 666-9.

Roberts, P.S., Ramesh, V., Dabora, S., Kwiatkowski, D.J. (2003). A 34 bp deletion within TSC2 is a rare polymorphism, not a pathogenic mutation. Ann Hum Genet.67(6): 495-503.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P. (2011). Integrative Genomics Viewer. Nature Biotechnology. 29: 24–26.

Scotti, M.M., Swanson, M.S. (2016). RNA mis-splicing in disease. Nat Rev Genet.17(1): 19-32.

Stamm, S., Smith, C.W.J., Lührmann, R. 2012. Alternative Pre-mRNA Splicing: Theory and Protocols. Weinheim: Wiley-VCH.

Sun, S., Ling, S.C., Qiu, J., ... , Cleveland, D.W. (2015a). ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. Nat. Commun. 6: 6171.

Sun, X., Zuo, F., Ru, Y., Guo, J., Yan, X., Sablok, G. (2015b). SplicingTypesAnno: annotating and quantifying alternative splicing events for RNA-Seq data. Comput Methods Programs Biomed.119(1): 53-62.

Suñé-Pou, M., Prieto-Sánchez, S., Boyero-Corral, S., Moreno-Castro, C., El Yousfi, Y., Suñé-Negre, J.M., Hernández-Munain, C., Suñé, C. (2017). Targeting Splicing in the Treatment of Human Disease. Genes (Basel). 8(3).

Tanackovic, G. Ransijn, A., Thibault, P., Abou Elela, S., Klinck, R., Berson, E.L., Chabot, B., Rivolt,a C. (2011). PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. Hum. Mol. Genet. 20: 2116–2130.

Tanko, Q., Franklin, B., Lynch, H., Knezetic, J. (2002). A hMLH1 genomic mutation and associated novel mRNA defects in a hereditary non-polyposis colorectal cancer family. Mutat. Res. 503: 37–42.

Tazi, J., Bakkour, N., Stamm, S. (2009). Alternative splicing and disease. Biochim Biophys Acta, 1792(1): 14–26.

Thompson, B.A., Martins, A., Spurdle, A.B. (2015). A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. Clin Genet.87(2): 100-8.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res. 22(9): 1616-25.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol.28(5): 511-5.

Wang, G.S., Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet.8(10): 749-61.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature. 456: 470–476.

Westra, H.J., Jansen, R.C., Fehrmann, R.S., te Meerman, G.J., van Heel, D., Wijmenga, C., Franke, L. (2011). MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics.27(15): 2104-11.

Yoshida, K., Sanada, M., Shiraishi, Y., Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 478: 64–69.

BAKALAUREUSETÖÖ:

Vaiküll, K. Meditsiinilist sekkumist vajavate geenivariantide leidmine Eesti Geenivaramu geenidoonoritel. 2016. Tartu Ülikool.

KASUTATUD VEEBILEHED

ClinVar, NCBI. Külastatud 24. märtsil, 2017, aadressil: https://www.ncbi.nlm.nih.gov/clinvar/

Exome Aggregation Consortium. Külastatud 9. mail, 2017, aadressil: http://exac.broadinstitute.org

Github. ASE. Külastatud 18. aprillil, 2017, aadressil: https://github.com/molgenis/systemsgenetics/wiki/ASE

MISO (Mixture of Isoforms) software documentation. Külastatud 18. jaanuaril, 2017, aadressil: http://miso.readthedocs.io/en/fastmiso/

LISA 1

MLH1 geeni kontrollindiviidide splaissinguanalüüsi tulemused, mis olid saadud SplicingTypesAnno programmiga.

Indiviid	**Alternatiivse splaissingu tüüp**	**Genoomsed koordinaadid**	**Alternatiivse splaissingu tõenäosus**
KONTROLL 1	Introni säilitamine (tüüp I)	chr3: 37090509-37091976 (intron 22)	21%
	Introni säilitamine (tüüp II)	chr3: 37083823-37090007 (intron 19, 20)	27%
KONTROLL 2	Introni säilitamine (tüüp I)	chr3: 37056036-37058996 (intron 13)	16%
	Introni säilitamine (tüüp I)	chr3: 37083823-37089009 (intron 19	6%
	Introni säilitamine (tüüp I)	chr3: 37090509-37091976 (intron 22)	28%
KONTROLL 3	Introni säilitamine (tüüp I)	chr3: 37090509-37091976 (intron 22)	16%
	Introni säilitamine (tüüp II)	chr3: 37089175-37090394 (intron 20, 21)	19%
KONTROLL 4	Introni säilitamine (tüüp I)	chr3: 37090509-37091976 (intron 22)	31%
	Introni säilitamine (tüüp II)	chr3:37081786-37090394 (intron 18, 19)	12%

LISA 2

TSC2 geeni kontrollindiviidide splaissinguanalüüsi tulemused, mis olid saadud SplicingTypesAnno programmiga.

Indiviid	**Alternatiivse splaissingu tüüp**	**Genoomsed oordinaadid**	**Alternatiivse splaissingu tõenäosus**
KONTROLL 1	Introni säilitamine (tüüp I)	chr16: 2135324-2136193 (intron 38)	39%
	Introni säilitamine (tüüp I)	chr16: 2136873-2137863 (intron 40)	41%
	Introni säilitamine (tüüp I)	chr16: 2138327-2138446 (intron 43)	100%
	Introni säilitamine (tüüp II)	chr16: 2131800-2133695 (intron 33, 34)	48%
	Alter.doonorsait (5')	chr16: 2137942-2137976 (ekson 38)	25%
KONTROLL 2	Introni säilitamine (tüüp I)	chr16: 2136873-2137863 (intron 40)	29%
	Introni säilitamine (tüüp II)	chr16: 2126587-2129035 (intronid 28, 29)	39%
	Introni säilitamine (tüüp II)	chr16: 2131800-2133695 (intron 33, 34)	72%
KONTROLL 3	Introni säilitamine (tüüp II)	chr16: 2126587-2129035 (intronid 28, 29)	6%

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Ustina Põnova (sünnikuupäev: 02.01.1993)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Meditsiiniliselt oluliste geenivariantide tõlgendamine RNA ekspressiooniandmete abil“, mille juhendajad on Tarmo Annilo, PhD ja Viktorija Kukuškina, MSc,

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, 28. mail 2017