TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Anna Smertina

Inimgenoomi ühenukleotiidiliste variatsioonide annotatsioon – ülevaade põhimõtetest ning teise põlvkonna sekveneerimise võimalike artefaktsete SNVde annoteerimine

Bakalaureusetöö

Maht: 12 EAP

Juhendaja PhD Ulvi Gerst Talas

TARTU 2016

Inimgenoomi ühenukleotiidiliste variatsioonide annotatsioon – ülevaade põhimõtetest ning teise põlvkonna sekveneerimise võimalike artefaktsete SNVde annoteerimine

Teise põlvkonna sekveneerimine võimaldab tänu oma kiirusele ja suhtelisele odavusele järjestada kiiresti palju genoome, mille baasil on võimalik läbi viia nii ülegenoomseid assotsiatsiooniuuringuid kui ka kasutada andmeid kliinilises praktikas. Mõlemad lähenemised sõltuvad tugevalt SNVde ja teiste variatsioonide õigest tuvastamisest ning täpsest annotatsioonist.

Antud töös tutvustatakse SNVde annoteerimise protsessi ja selle eripärasid, tuuakse välja annotatsiooni tõlgendamise erinevused lähtuvalt erinevatest tööriistadest ning andmebaasidest.

Töö praktilises pooles näidatakse, et valepositiivselt tuvastatud SNVd võivad annoteerimise ja tulemuste tõlgendamise põhjal olla näiliselt füsioloogiliselt olulised. Artefaktsete SNVde tuvastamisega arvestamine võimaldab vältida vigaste andmete põhjal tehtud ekslikke järeldusi.

Märksõnad: teise põlvkonna sekveneerimine, annoteerimine, SNV, bioinformaatika

CERCS: B110 Bioinformaatika, meditsiiniinformaatika

Annotation of single nucleotide variants in human genome: an overview and annotation of artefact SNVs from NGS

Next-generation sequencing allows, due to the soread of high-throughput methods and relatively low cost, to rapidly sequence a number of genomes. Based on sequencing data, it is possible to both conduct genome-wide association studies and use sequenced genomes in clinical practise. Both applications rely heavily on correct SNV and other variants calling as well as detailed variant annotation.

Current thesis gives a systematic overview of variant annotation process and possible shortcomings of different tools, and points out differences between widely used annotation algorithms and databases.

The thesis's practical section shows that false-negatively called SNVs can, based on annotation results, have seemingly important physiological impacts. Consideration of falsely called SNVs can help to avoid misguided conclusions based on error-prone data.

Key words: next-generation sequencing, annotation, SNV, bioinformatics

CERCS: B110 Bioinformatics, medical informatics

SISUKORD

KASUTATUD LÜHENDID

ANNOVAR -	Annotate Variation (annotatsioonitarkvara)
ASCII -	*American Standard Code for Information Interchange* (keelemärkide tabel)
BWA -	Burrows-Wheeler Aligner (joondusalgoritm)
CCDS -	Consensus CDS Project (andmebaas)
CNV -	koopiaarvu variatsioon (ing.k. *copy number variant*)
DSSP -	Dictionary of Secondary Structure in Proteins (andmebaas)
FATHMM -	Functional Analysis through Hidden Markov Models (annotatsioonitarkvara)
GATK (UGT)	Genome Analysis Toolkit (Unified Genotypic caller) (tarkvarapakett)
GRC -	Genome Reference Consortium (referentsgenoome haldav konsortsium)
HAVANA -	Human and Vertebrate Analysis and Annotation grupp (andmebaas ja seda haldav töögrupp)
indel -	insertsioon ja deletsioon
LD -	mittetasakaalustatud aheldatus (ing.k. *linkage desequilibrium*)
NCBI -	National Center for Biotechnology Information (biotehnoloogia infokeskus)
NCBI NR -	NCBI non-redundant protein database (andmebaas)
NGS -	teise põlvkonna sekveneerimine (ing.k. *next generation-sequencing*)
PDB -	Protein Structure Database (andmebaas)
PROVEAN -	Protein Variation Effect Analyzer (annotatsioonitarkvara)
PSI-BLAST -	Position-Specific Iterated BLAST (tarkvara)
PSIC -	Position-Specific Independent Counts (tarkvara)
RefSeq -	NCBI Reference Sequence Database (andmebaas)
SAM -	*Sequence Alignemnt/Map* (kokkuleppeline failiformaadi tüüp)
SIFT -	Sorting Intolerant from Tolerant (tarkvara)
SNP -	ühenukleotiidiline polümorfism (ing.k. *single nucleotide polymorphism*)
SNV -	ühenukleotiidiline variatsioon (ing.k. single nucleotide variant)
TES -	täis-eksoom sekveneerimine
TGS -	täis-genoomi sekveneerimine
TXT -	tekstifail
VCF -	Variant Call Format (kokkuleppeline failiformaadi tüüp)
VEP -	Variant Effect Predictor (annoteerimistarkvara)
VN -	valenegatiivne
VP -	valepositiivne

SISSEJUHATUS

Teise põlvkonna sekveneerimine leiab aina rohkem kasutust nii teaduses kui ka kliinilises praktikas. Sekveneerimisel saadud andmetest huvi pakkuva ning olulise informatsiooni saamine on mitmeastmeline protsess, mis hõlmab endas sekveneerimisel saadud lugemite joondamist, sealt huvi pakkuvate variatsioonide tuvastamist ning nende annoteerimist ehk variatsioonidele tähenduse andmisest.

Kogu protsessi käigus on mitmeid aspekte, millele tuleb tähelepanu osutada, alates õigest joondamistööriista valikust lõpetades sobivate andmebaasidega, mille abil variatsioone annoteerida. Kõik astmed võivad oluliselt mõjutada lõplikke tulemusi, mis kokkuvõttes võib viia vigaste andmete põhjal tehtud ekslike järeldusteni. Seetõttu on oluline teadvustada iga protsessi käigus kasutatava tööriista puudusi ning nendega arvestada.

Antud töö eesmärgiks on kirjanduse ülevaate osas anda ülevaade annoteerimise põhimõtetest, täpsemalt annoteerimise puhul kasutatavate tööriistade ja vajalike andmebaaside iseärasustest. Lühidalt tuuakse välja annoteeritava andmestiku saamise protsessi ning selle olulisemaid kitsaskohti.

Eksperimentaalosa ülesandeks oli annoteerida erinevate joondusalgoritmide poolt joondatud lugemitest määratud SNVsid ning anda täpsema ülevaate selle kohta, kas artefaktselt määratud variatsioonide seas võib esineda näiliselt füsioloogiliselt olulisi variatsioone või mitte.

1. KIRJANDUSE ÜLEVAADE

1.1. Annoteerimise eesmärgid

Variatsioonide annoteerimine on protsess, mille käigus ennustatakse variatsioonide võimalikku mõju variatsioonidega seotud geenide funktsioonidele. Selleks kasutatakse spetsiifilisi annoteerimistööriistu, mis hindavad variatsioonide võimalikke mõjusid lähtuvalt olemas olevast informatsioonist DNA ja valkude järjestuste ning nende funktsioonide seoste kohta (Aubourg ja Rouzé, 2001).

Üheks sagedamini määratavaks geneetilise variatsiooni tüübiks on ühenukleotiidilised variatsioonid ehk SNVd (inglise keeles *single nucleotide variant*). Nende korrektne tuvastamine ja annoteerimine mängib olulist rolli inimese genoomi analüüsil ja on eelduseks, et selekteerida haiguste või kindlate fenotüübiliste tunnustega seotud SNVsid. Kuigi kindla SNV ja fenotüübi vahelise seose kindlaks määramisel on oluline eksperimentaalne valideerimine, ei ole eksperimentaalset võimalik valideerida kõiki umbes kolme miljonit või suuremat arvu SNVsid ühe indiviidi kohta. Annoteerimine võimaldab välja sorteerida võimalikke uuritavate haiguste või fenotüübiliste tunnustega seostatavaid kandidaat-SNVsid. Lisaks genoomi analüüsil tuvastatud võimalikele genotüüp-fenotüüp seostele aitab annoteerimine täpsustada, millised SNVde poolt põhjustatud võimalikud molekulaarbioloogilised muutused tingivad uuritavaid fenotüüpe (Ritchie ja Flicek, 2014).

Antud töö praktilises osas annoteeritakse SNVsid, keskendudes mudelile, mis jätab lihtsustatult välja insertsioonid ja deletsioonid (indelid), koopiaarvu variatsioonid (CNV, inglise keeles *copy number variant*) ja teised võimalikud raskemini tuvastatavad variatsioonid.

Kirjanduses on mõistete SNV ja ühenukleotiidiline polümorfism ehk SNP (inglise keeles *single nucleotide polymorphism*) kasutuse vaheline piir sageli hägune, eelistatud on SNV kui laiema mõiste kasutamine. See on korrektne tähistus ka juhtudel, kui variatsiooni esinemissagedus ei ole teada ning seda on raske liigitada SNPks või mutatsiooniks.

1.2. Teise põlvkonna sekveneerimine – andmete saamine ja töötlus

1.2.1. NGS toorandmete saamine

Teise põlvkonna sekveneerimine (NGS, inglise keeles *next-generation sequencing*) on kõrge läbilaskevõimega sekveneerimismeetodite üldnimetus. NGS võimaldab sekveneerida suurt hulka DNA ja RNA järjestusi palju kiiremini ja odavamalt kui Sangeri sekveneerimismeetod. Enimkasutatud on firmade Illumina (Solexa) Inc., Roche/454, Ion Torenti ja SOLiD

sekveneerimistehnoloogiad ja –platvormid [1]. Kuna teise põlvkonna sekveneerimise kasutamisel teadusuuringutes ja kliinilises praktikas on ülekaalus Illumina Inc. tehnoloogiad, lähtub antud töö eelkõige Illumina Inc. tehnoloogiate spetsiifikast [2].

Ühe proovi järjestamisel teise põlvkonna sekveneerimistehnoloogiatega toodetakse paralleelselt miljoneid lühikesi DNA järjestusi. Lühikeste järjestuste saamiseks fragmenteeritakse järjestatav proov, valmistatakse ette matriitsjärjestuste raamatukogu, immobiliseeritakse need tahkele kandjale, amplifitseeritakse sellel olevad järjestused koopiate kimpudeks ning üheahelalise matriitsjärjestuse kõrvale komplimentaarse ahela sünteesi ajal registreeritakse iga nukleotiidide liitumise tsükli ajal sünteesitavale ahelale liitunud nukleotiidiga seostud fluorestsentsmärgise-signaal. Registreeritud signaalide põhjal koostatakse lugemite järjestused. Teise põlvkonna sekveneerimistehnoloogiate puhul kasutatakse erinevaid sünteesi ja signaali registreerimise lähenemisi, mis kõik võimaldavad esimese põlvkonna sekveneerimisega võrreldes viia paralleelselt läbi rohkem reaktsioone, järjestades seeläbi proove palju kiiremini ja odavamalt (Metzker, 2010).

Teise põlvkonna sekveneerimisel saadud lugemid on reeglina lühikesed, tavaliselt kuni 300bp pikad ning nende pikkus sõltub sekveneerimisel kasutatud tehnoloogiaplatvormist. Näiteks populaarseima sekveneerimisplatvormide tootja, Illumina sekveneerimisplatvormid toodavad ühe sekveneerimisprotsessi käigus maksimaalselt kas 2 x 150 bp (MiniSeq, NextSeq, HiSeq ja HiSeq X seeria platvormid) või 2 x 300 bp (MiSeq seeria platvorm) pikkusi paarislugemeid [2].

Paarislugemite puhul saadakse ühe sekveneeritava fragmendi mõlema otsa järjestused ehk lugem koosneb kahest kindlaks määratud järjestusega otsaalast ning nende vahele jäävast aimatava pikkusega järjestamata alast. See võimaldab võrreldes üksiklugemitega neid referentsjärjestusele täpsemini joondada, seda eriti genoomi kordusjärjestuste piirkondades. Samuti hõlbustavad paarislugemid ulatuslikemate genoomsete ümberkorralduste, kordusjärjestuste ja uute transkriptide leidmist [3].

Maksimaalselt registreeritakse kõige madalama läbilaksevõimega Illumina platvormiseeriate MiniSeq ja MiSeq puhul 25 miljonit ja kõrgeima läbilaskevõimega platvormiseeria HiSeq X Ten puhul 60 miljardit lugemit. Korraga on ühe sekveneerimisprotsessi puhul võimalik saada 7,5 Gb kuni 1800 Gb järjestuste toorandmeid [2].

Toorandmeid väljastab sekveneerimisplatvorm reeglina FASTQ formaadis. FASTQ on Wellcome Trust Sanger Instituudis (Suurbritannia) Jim Mullikini poolt FASTA-formaadi baasil arendatud järjestuste esitamise formaat, mis sisaldab iga järjestuse kohta eraldi ridadel

järjestuse nime ja täpsustava informatsiooniga rida, nukleotiidide järjestust, informatsioonirea kordust ning igale nukleotiidile vastavat Phred-kvaliteediskoori väärtust esitatuna ühe sümbolina ASCII (*American Standard Code for Information Interchange*) kodeeringus koos nihkega 33 sümboli võrra (Phred+33) (*Cock et al*, 2010).

1.2.2. Inimese täis-genoomi ja täis-eksoomi sekveneerimine

Genoomi sekveneerimisel on võimalik valida kahe levinud lähenemise – täis-genoomi või täis-eksoomi sekveneerimise vahel ning valik mõjutab oluliselt saadavaid tulemusi.

Täis-eksoomi sekveneerimine (TES) on laialdaselt kasutatav meetod nii sagedaste kui ka haruldaste inimese geenivariatsioonide tuvastamiseks. Täis-genoomi sekveneerimine (TGS) omab TESiga võrreldes laiemat katvust, kuid on hinna poolest kallim meetod (Belkadi *et al*, 2015).

Kuna variatsioonide mõju, mis asuvad väljaspoolt valkude kodeerimisalasid, on raske tõlgendada, otsitakse eelistatult valkude kodeerimisalasse jäävaid variatsioone. Neid on võimalik tuvastada nii TESi kui ka TGSiga.

Siiski näitas 2015. aastal Aziz Belkandi koos kolleegidega (Belkadi *et al*, 2015) kuue indiviidi TES ja TGS meetoditega saadud andmete võrdlusel, et TGS tagab ühtlasema sekveneerimis-kvaliteedi parameetrite jagunemise kui TES, mille kallutatus on tõenäoliselt tingitud TES-il kasutatavate proovi hübridisiatsiooni ja PCR-amplifikatsioni meetoditest. Samuti leiti, et TGS detekteerib sadu potentsiaalselt kahjulikke SNVsid (umbes 3% kõigist kõrge kvaliteediga SNVdest, mida TGS puhul määrati) rohkem kui TES, kuigi SNVd asusid TES sihtmärkaladel (Belkadi *et al*, 2015).

1.2.3. NGS andmete töötluse põhisammud

Nagu eelpool mainitud, saadakse NGSil suur kogus toorandmeid, mis koosnevad sadadest tuhandetest või miljonitest lühikestest lugemitest ehk kindlaks määratud järjestusega DNA lõikudest. Andmete edasine analüüs koosneb üldjoontes järgmistest sammudest:

- toorandmete kvaliteedi hindamine,
- lugemite joondamine referentsgenoomile,
- joondatud järjestusest variatsioonide tuvastamine,
- saadud variatsioonide annoteerimine ning andmete visualiseerimine,
- huvi pakkuvate variatsioonide filtreerimine, leitud variatsioonide valideerimine (Pabinger et al. 2014).

1.2.4. NGS lugemite paigutamine referentsjärjestusele

Selleks, et sekveneerimisel saadud toorandmeid analüüsida, on lühikestele lugemitele vaja leida nende positsioon genoomis. Teise põlvkonna sekveneerimisel saadud andmeid järjestatakse eelkõige referentsgenoomile ehk lühikestele lugemitele leitakse asukoht, kus nad on kõige sarnasemad referentsiks kasutatud järjestusele (Trapnell ja Salzberg, 2009).

Inimese genoomse järjestuse kasutatakse referentsina Genome Reference Consortiumi (GRC) poolt kokku pandud referentsgenoomi (Niels*en et al*, 2011). GRC poolt kokku pandud referentsgenoom on pidevalt täiendatav, 21. märtsil 2016 ilmus inimese referentsgenoomi versioon GRCh38.p7. Referentsgenoomi järjestuse pidev täiendamine toob kaasa muutused nii järjestustes endis kui ka referentsgenoomi mahus. Näiteks versioon GRCh38.p7 koosneb 3 232 546 710 järjestatud aluspaarist, versioon GRCh37.p13 koosneb 3 234 834 689 aluspaarist. Muutused järjestustes versioonide vahel toovad kaasa versioonidevahelisi nihkeid koordinaatides. See tähendab, et ühes kromosoomis olev kindel positsioon võib erinevates referentsgenoomi versioonides kanda erineva järjekorranumbri [5]. Seetõttu on oluline panna tähele, et erinevate referentsgenoomist sõltuvate tööriistade kasutamisel oleks kasutusel sama referentsgenoomi versioon. Vajadusel saab ühele referentsgenoomile joondatud järjestuse koordinaate konverteerida ümber teise referentsgenoomi koordinaadistikku National Center for Biotechnology Information (NCBI) Genome Remapping Service'i abil [6].

Lugemite joondamisel referentsjärjestusele on mitmeid väljakutseid. Sekveneerimisel saadud lugemite puhul on enne joondamist teada mitmeid parameetreid – lugemi pikkus, paarislugemi puhul ligikaudne fragmendi pikkus ning vale nukleotiidi määramise tõenäosus ehk kvaliteet. Joondamise eesmärgiks on leida referentsjärjestuse peal koht, mis langeb lugemiga kõige paremini kokku. Kuna eukarüootide genoomides on palju kordusi, referentsjärjestusest erinevaid variatsioone ning lisaks tekib ka sekveneerimisel vigu, on joondamisel vaja lubada nii mittesarnaste tähtede paare (inglise keeles *mismatch*) kui ka joonduse vahesid. Sobiva joondamisvigadega arvestatava algoritmi valik võimaldab joondada lugemeid, mis osaliselt sisaldavad korduselemente või erinevaid genoomivariatsioone (SNVd, insertsioonid ja deletsioonid, koopiaarvu muutused) (Pabinger *et al*, 2014; Reinert *et al*, 2015).

Laialdasemalt kasutatavad veamudelid on *Hamming distance*, mis võtab vea arvutamisel arvesse ainult *mismatch*e lugemi ja valitud genoomse asukoha vahel, ning *edit distance*, mis võtab arvesse *mismatch*e ning indeleid. Lisaks on võimalik kasutada kaalutud *edit distance* mudelit, mis annab kaalutud veahinded ehk teeb vahet erineva pikkusega *mismatch*idel, indelitel ning võib kaaluda positsiooni-spetsiifilisi vigu erineval sõltuvalt nukleotiidi valesti

määramise tõenäosusest. Kuigi joondamistööriistad võivad läheneda joondamise probleemile väikeste erinevustega, näiteks lubades lugemite otstel olla mitte-joondatud, kasutavad joondusalgoritmid üldjoontes siiski sarnaseid lähenemisi.

Lisaks erinevustele lugemite ja referentsjärjestuse vahel tuleb joondusalgoritmidel optimeerida tööriistade poolt kasutatavat arvutusmälu ning joondamisele kuluvat aega, mis kasvavad proportsionaalselt lugemi pikkuse ja genoomijärjestuse suurenemisega. Suurte sisendmahtude, nii lugemite arvu kui referentsjärjestuse suuruse, kiiremaks ja optimaalsemaks joondamiseks on kasutusel kaks põhilist lähenemist: filtreerimine ja indekseerimine.

Filtreerimispõhine lähenemine väldib suuri referentsjärjestuse piirkondi, millele ei leita sarnaseid järjestusi. Referentsjärjestus jaotatakse lühikesteks piirkondadeks ning võrreldakse neid lugemite lühikeste lõikudega. Referentsjärjestuse piirkonnad, mis ei oma lugemiosadega täielikult kattuvaid alasid, jäetakse edasisest joondamisest välja (Reinert *et al*, 2015).

Indekseerimisel põhinev lähenemine eeldab lugemite ning referentsjärjestuse eeltöötlemist stringideks (tähestikusümbolitest koosnevateks sõnedeks). Eeltöötluse järel ei ole vaja joondamisprotsessi käigus skaneerida tervet referentsjärjestust, mis võimaldab algoritmil töötada kiiremini, kasutades seejuures suuremat arvutusmälu mahtu. Stringi indekseerimise lähenemist kasutavad näiteks *suffix array* (Manber ja Myers, 1990), FM-indeksi (Ferragina ja Manzini, 2000) ning Burrow-Wheeleri transformatsiooni (Burrows ja Wheeler, 1995) põhjal koostatud algoritmid.

Enimkasutatud joondamistööriistad, mida kasutatakse teise põlvkonna sekveneerimisel saadud lugemite joondamiseks, on Burrows-Wheeler Aligner (BWA) (Li ja Durbin, 2009) ja Bowtie2 (Langmead *et al.*, 2009).

BWA on mõeldud suurte genoomide joondamiseks ning töötab FM-indekseerimise meetodil, mis baseerub Burrows-Wheeleri transformatsioonil. BWA pakett koosneb kolmest alternatiivsest algoritmist – *align*, SW ja MEM. Esimene on kasutusel Illumina tehnoloogiatega saadud lugemite, mis on kuni 100 nukleotiidi pikad ning teised on pikemate, 70…1000 nukleotiidi pikkade lugemite joondamiseks. Neist kõige uuem, BWA-MEM, on täpsem ja kiirem ning soovituslik tööriist inimese genoomi joondamisel asendamaks BWA *align*i [7].

BWA-MEM võimaldab leida ka kimäärseid lugemeid. Samuti tolereerib tööriist paremini sekveneerimisel tehtud vigu. BWA-*align* on disainitud töötama sekveneerimisvigadega alla 2% ning vajadusel lõikab Illumina Inc. tehnoloogiaga loodud lugemitel ära 3’-otstest madala kvaliteediga nukleotiide. BWA-MEM on aga disainitud tolereerima pikemate lugemite korral

rohkem sekveneerimisel tekkivaid vigu, näiteks 5% veamäära 500 bp pikkuste lugemite ning 10% veamäära 1000 bp pikkuste joonduste korral [7].

BWA-MEM on soovituslik tööriist Broadi Instituudi Genome Analysis Toolkiti (GATK), poolt. GATK on muutunud standardiks variatsioonide tuvastamiseks. GATK koondab enda alla mitmeid kvaliteedkontrolli-, diagnostika-, andmetöötluse ja variatsioonidega töötamise tööriistu [8].

Bowtie2 (Langmead ja Salzberg, 2012) on samuti FM-indekseerimise meetodil põhinev joondamistööriist. See on mõeldud 50…1000 nukleotiidi pikkuste lugemite joondamiseks suurte genoomidele (näiteks imetaja omale). Bowtie2-l on kaks erinevat joondamise viisi – *local* (lokaalne) ja *paired-end* (paarisjoonduse).

Lokaalse joonduse puhul ei pea joondused referentsiga täiesti kattuma, vaid neid võib otsast veidi „kärpida“, kuid samuti on võimalik joondada *end-to-end* viisil ehk lugemid peavad täielikult referentsiga joonduma.

Bowtie2 on loodud suurte genoomide joondamiseks ning optimeeritud pikkadele lugemitele. Samuti on see optimeeritud arvestama vigadega, mida teevad suuremate sekveneerimisfirmade, nagu näiteks Illumina HiSeq või Roche/454, platvormid [9].

Bowtie2-s on kasutusel esmase tööriistana mitmes suuremas *pipeline'is* ning seda kasutatakse nii variatsioonide määramisel, kui ka ChIP-seq, RNA sekveneerimise analüüsidel kui ka bisulfit-sekveneerimisel DNA metülatsioonimustrite uurimisel. Oluline on toonitada selle integreeritust populaarsetesse RNA analüüsitööriistadesse, näiteks TopHat [10], mis on kiire splaissimise ühenduskohtade joondaja [9].

Joondamisel kasutatavad tööriistad saavad teha üldistatult kahte tüüpi vigu – lähtuvalt kas lugemite kvaliteedist või joondusalgoritmi eripäradest. Halvasti määratud (madala kvaliteediga) nukleotiidid joondustes vähendavad nende tõenäosust korrektseks paigutamiseks referentsjärjestuse suhtes. Seejuures tõstab kvaliteediskooriga arvestamine joondamise tundlikkust ning võib põhjustada erinevusi kvaliteediskooriga arvestatavate algoritmide, näiteks BWA, efektiivsuse erinevust kvaliteediskooriga mittearvestavate algoritmidega võrreldes. (Kerpedjiev *et al.*, 2014)

On näidatud, et erinevad tööriistapaketid, mis integreerivad endas joondusalgoritme ning variatsioonide tuvastamise tööriistu, ei tuvasta paljusid SNVsid või indeleid ühtselt. Kokkulangevalt määrasid uuringus kasutatud populaarsed *pipeline*'id ligi 60% SNVdest ning on näidatud, et sõltuvalt metodoloogilistest erinevustest eksivad kõik tööriistad nii ühtemoodi

kui ka omavad neile unikaalseid vigu ehk iga tööriistaga on teatud variatsioonid, mis jäävad tuvastamata. (O'Rawe *et al.*, 2013)

Tööriistad vajavad sisendfailina joondamata lugemeid sisaldavat FASTQ või FASTA failiformaate ning toodavad väljundina joondatud järjestuse, mis on salvestatud reeglina SAM ehk *Sequence Alignemnt/Map* formaadis. Edasiseks kasutamiseks konverteeritakse SAM formaadis failid binaarsesse BAM formaati [7].

SAM formaadis fail on tabuleeritud tekstifail, mis sisaldab päiseosa ning joonduste osa. Joonduste osas on igal real 11 kohustuslikku välja, mis sisaldavad lugemi nime, lugemi positsiooni joondatud järjestuses, informatsiooni nukleotiidide kvaliteedi (Phred-skoor) ning antud lugemi paarilise kohta ning muud joondusalgoritmide-spetsiifilist informatsiooni [11].

1.2.5. SNVde tuvastamine

Genoomianalüüsi üheks eesmärgiks juba sekveneeritud genoomiga organismide puhul on määrata uuritavas genoomis esinevaid variatsioone, sealhulgas SNVsid, mille põhjal on hiljem võimalik otsida seoseid olemas olevate variatsioonide ja fenotüübiliste tunnuste vahel. SNVde tuvastamisel leitakse positsioonid genoomis, kus üks nukleotiididest erineb referentsgenoomis samas positsioonis olevast nukleotiidist.

SNVde tuvastamine ja genotüpiseerimine võib toimuda lihtsal meetodil, võttes arvesse erinevate alleelide esinemist kindlas positsioonis ning rakendades lävendväärtusi määramaks alleeli kas referentsgenotüübiks või SNVks.

Keerukamad SNVde ja genotüüpide tuvastamise algoritmid kasutavad tõenäosuslikku raamistikku ehk arvestavad arvutuslike meetodite abil SNVde ja genotüüpide tuvastamisel võimalikke sekveneerimisel ja joondamisel tekkivaid vigu ning teadaolevaid alleelisagedusi ja mittetasakaalustatud aheldatust (LD, inglise keeles *linkage disequilibrium*) puudutavat informatsiooni (Nielsen *et al*, 2011).

Lisaks võimalikele sekveneerimisel ja joondamisel tekkivatele vigadele mõjutab SNVde tuvastamist sekveneeritud järjestuste katvus. Sekveneerimise katvus näitab keskmist lugemite arvu iga sekveneeritava järjestuse genoomse positsiooni kohta. Mida kõrgem on katvus ehk mida rohkem kordi on sekveneerimisel selle piirkonna nukleotiidi registreeritud, seda suurema kindlusega saab määrata sekveneeritud proovi järjestuse [4]. Kõrge katvusega (üle 20-kordne) sekveneeritud proovide puhul määravad erinevad SNVde tuvastamise tööriistad SNVsid pigem sarnaselt (Adams *et al*, 2012). Madala katvusega (vähem kui 5-kordselt) sekveneeritud proovide puhul määravad tööriistad SNVsid erinevalt. Nelja enimkasutatava tööriista,

SOAPsnp, Atlas-SNP2, SAMtools ja GATK võrdluses oli ühtselt määratud SNVde osakaal umbes 35%...45% dbSNP andmebaasis olevate SNVde puhul ning 19%...28% *de novo* määratud SNVde puhul. Kõige enam mõjutas katvus uudsete SNVde tuvastamist – minimaalse katvuse kriteeriumi tõstmisel kolmekordselt neljakordsele jäi tuvastamata umbes 50% SNVdest ning kümnekordse katvuse puhul määrati ainult 15% SNVdest.

Madala katvusega proovidest SNVde tuvastamisel võib usaldusväärsete tulemuste saavutamiseks kasutada paralleelselt mitut tööriista, kuigi see suurendab samal ajal valenegatiivsete tulemuste saamise ehk tõeliste SNVde mitte-tuvastamise tõenäosust (Yu ja Sun, 2013).

Ülevaade enimkasutatud SNVde tuvastamise algoritmidest on toodud järgnevas tabelis (tabel 1).

SNVde tuvastamise tööriistad kasutavad sisendina reeglina BAM-formaadis faili ning väljundiks on standardiseeritud Variant Call Format ehk VCF fail. See algab päisest, mis sisaldab informatsiooni faili sisu kohta ning faili sisus kasutatavate lühendite tähendusi. Variatsioonid on fails esitatud tabuleeritud teksti kujul, kus igale real on informatsioon ühe variatsiooni kohta. Variatsioonide kirjeldamiseks faili põhiosas kasutatakse igale variatsioonile vastaval real järgnevaid välju: CHROM – kromosoomi number, POS – variatsiooni alguse positsioon, ID – variatsiooni unikaalne identifitseerimiskood, REF – referentsalleel, ALT – mitte-referentsalleelid, QUAL – SNV tuvastamise kvaliteed Phred-skaalal, FILTER – filtreerimisinformatsioon, INFO – kasutaja või tööriistade poolt lisatav informatsioon näiteks alleelisageduste, katvuse ja genotüübi kvaliteedi kohta. Kui ühe variatsiooni kohta on reas mitme proovi informatsioon, lisatakse juurde FORMAT väli, kirjeldamaks eraldi iga proovi välja sisu (Danecek *et al*, 2011).

1.3. SNVde annoteerimine

Fenotüüp-genotüüp seoste, näiteks kindlate haigustega seotud variatsioonide leidmiseks on pärast SNVde tuvastamist vaja filtreerida suurest hulgast variatsioonidest välja võimalikud huvi pakkuvaid fenotüüpilisi tunnuseid mõjutavad variatsioonid. Selleks loob eeldused annoteerimine ehk variatsioonidele informatsiooni lisamine variatsiooni asukoha kohta genoomis ja/või geenis, informatsiooni (varem tuvastatud) SNVde esinemise kohta kindlates transkriptides ning teadaoleva või arvutuslikult saadud informatsioon nende võimaliku mõju kohta geenide avaldumisele ja sünteesitud valkudele (Cingolani *et al.*, 2012).

Tabel 1 Ülevaade enimkasutatavatest SNVde tuvastamise tööriistadest, mugandatud Nielsen et al, 2011 artikli põhjal. Sisendi formaat tähistab faili tüüpi, milles peavad analüüsitavad joondatud lugemid salvestatud olema. Väljundi formaat tähistab failitüüpi, millesse salvestatakse määratud SNVd. Eelnõuete all on välja toodud olulised sammud, mis peavad olema tehtud enne, kui joondatud lugemitest on võimalik SNVsid määrata, ning tarvilik lisainformatsioon, mida tuleb kasutajal käsitsi täpsustada. Tuvastamise otsuse mõõdik on algoritmi või kriteeriumi otsuse statistiline kirjeldus, mille abil on võimalik hinnata SNV tuvastamise usaldusväärsust. NGS andmete analüüsipakett näitab, kas SNVde tuvastamise tööriist on integreeritud suuremasse analüüsipaketti. Viide on hüperlink veebiaadressile, kust on võimalik vastavat tööriista alla laadida või selle kohta lisainformatsiooni saada.

	Sisendi formaat	Väljundi formaat	Eelnõuded	Tuvastamise otsuse mõõdik	NGS andmete analüüsipakett	Viide
Atlas-SNP2 (Sh*en et al*, 2010)	SAM/ BAM	VCF (*Variant Call Format*)	PCR duplikaatide eemaldaine, nukleotiidide kvaliteedi ümberarvutamine, indelite kohalik ümberkoondus	*Posterior probability*		https://sourceforge.net/p/atlas2/wiki/Atlas-SNP/
Genome Analysis Toolkit Unified Genotypic caller (GATK-UGT) (DePristo *et al*, 2011; McKenna *et al*, 2010)	SAM/ BAM	VCF	Joondatud lugemid	*FisherStrand, Genotype quality, HaplotypeScore, MappingQuality,* QUAL, *RankSumTest, ReadPosRankSumTest*	GATK	https://www.broadinstitute.org/gatk/
SAMtools (Li *et al*, 2009a)	BAM	VCF	Joondatud lugemid	*Genotype quality,* QUAL	*Samtools, bcftools* –	http://www.htslib.org/
SOAPsnp (Li *et al*, 2009b)	SOAPi väljund	VCF	Kõrge kvaliteediga SNVde andmebaas, nt dbSNP	*Consensus score*	SOAP2	http://soap.genomics.org.cn/soapsnp.html

1.3.1. Üldised tööpõhimõtted

Variatsioonide annoteerimisel on võimalik kasutada erinevaid lähenemisi, mida võib jagada järgmistesse gruppidesse: reeglitepõhine, järjestuste konserveeritusel põhinev ja masinõppel põhinev annotatsioon. Selline klassifikatsioon on mugandatud Graham RS Ritchie ja Paul Fliceki 2014. aastal avaldatud klassifikatsioonist (Ritchie ja Flicek, 2014).

1.3.1.1. Reeglitepõhine annotatsioon

Tänapäeva teadmised genoomijärjestuste, kindlate geenielementide funktsioonide ning teadaolevate variatsioonide poolt põhjustatud fenotüübiliste muutuste kohta võimaldavad ennustada variatsioonide poolt põhjustatud muutusi geenide funktsionaalsuses. Variatsioonid põhjustavad muutusi geeni ja seda ümbritsevate alade järjestustes. Kasutades ära teadaolevat informatsiooni geeni struktuuri ja elementide kohta ning teades geneetilise koodi tõlgendamise reegleid, on võimalik ennustada võimalikke muutusi geeni elementides, geeni poolt kodeeritavates valkudes või splaissingus (Cingolani *et al*, 2012; McLaren *et al*, 2010).

Iga tööriist kasutab eeldefineeritud variatsiooni tagajärgede (inglise keeles *consequence*) nimekirja ning reegleid nende määramisel. Iga variatsioon kontrollitakse reeglite suhtes ning väljundisse lisatakse variatsiooni juurde kas kõik võimalikud või kõige suurema mõjuga tagajärg. Näiteks tagajärg „*stop-gained“* ehk „stopkoodoni loomine“ on SNV tagajärg, kus aminohapet kodeeriv koodon hakkab selles sisalduva SNV tagajärjel tähistama stopkoodonit ehk on transkriptsiooni lõpetamise signaaliks. See võib tuua kaasa liiga lühikese mRNA transkribeerimise, millelt võidakse transleerida düsfunktsionaalne valk.

Samuti on võimalik lisada täiendavat informatsiooni, nagu näiteks koodoni ja vastava aminohappe muutust, muutuse asukohta cDNAs, valgus või kaugust lähima geenini (Cingolani *et al*, 2012; McLaren *et al*, 2010; Wang *et al*, 2010).

Reeglitepõhist annotatsiooni kasutavad tööriistad rakendavad reeglina kahte sammu: andmebaasi loomist ning variatsiooni mõju arvutamist. Andmebaasi ehitamiseks kasutatakse referentsgenoomi ning annotatsioonitabelit, mis võib olla näiteks RefSeqi (NCBI Reference Sequence Database) või Ensembli andmebaas.

Teise sammuna loeb tööriist sisse nii andmebaasi kui ka variatsioone sisaldava faili (VCF). Seejärel võrdleb tööriist igat variatsiooni loodud andmebaasi vastu ning kui variatsiooni ning andmebaasi vahel on kattuvus, lisatakse variatsioonile selle võimalik mõju. Lisaksinnatakse

eksonis asuvate mitte-sünoüümsete variatsioonide efekte, millele lisatakse koostatud andmebaasis variatsiooni kohta teada olev lisainformatsioon.(Cingolani *et al*, 2012).

Reeglitepõhise annotatsiooni puhul mängib tulemuste saamisel olulist rolli kasutatav andmebaas. Ülevaade andmebaasidest ja nende spetsiifikast on kirjeldatud peatükis 1.5.

Reeglitepõhiste tööriistade annotatsiooniinformatsioon on piiratud praeguste teadmiste ja mudelitega genoomi elementide funktsioonide kohta ning ei suuda ennustada ootamatult käituvate variatsioonide mõju. Samas, selline lähenemine loob hüpoteese variatsioonide mõju kohta, mida on võimalik laboratoorsetes tingimustes kontrollida.

Praegused tööriistad ei arvesta reeglina proovi tüübiga, näiteks koe liigi või arengustaadiumiga, kust järjestus pärineb. Näiteks, kui SNV toob annotatsiooni järgi kaasa enneaegse stopkoodoni loomise ning seeläbi mitte-funktsionaalse valgu tootmise, võib variatsiooni sisaldav transkript uuritavas koes mitte avalduda ning hoolimata ennustuse kaalukusest mõju puudub. Seetõttu on soovitatav lisada annoteerimisinformatsioonile koespetsiifilise ekspressiooni informatsiooi (Ritchie ja Flicek, 2014).

Reeglitepõhise annoteerimise üldreeglid ning võimalikud variatsioonide tagajärjed kolme populaarseima annoteerimistööriista baasil on välja toodud lisas 1.

1.3.1.2. Järjestuste konserveeritusel ja homoloogial põhinev annotatsioon

Kaasaegsete elusorganismide geenid on läbinud loodusliku valiku. Seega on valkude kindlatel positsioonidel olevad aminohapped, mis on ka teistel liikidel konserveerunud, tõenäoliselt olulised ning mutatsioonid neis positsioonides omavad suure tõenäosusega kahjulik mõju (S*im et al*, 2012).

Evolutsioonilise konserveerumise leidmiseks nii DNA kui ka valkude järjestustele on arendatud mitmeid tööriistu, mis põhinevad homoloogsete järjestuste mitmesel joondamisel. Üldjoontes järjestab tööriist mitu homoloogset eri liikidelt pärit joondust. Seejärel analüüsib igas positsioonis esinevaid variatsioone eraldi. Selleks loetakse kokku kõikide asendusnukleotiidide esnemissagedused ning kõrvutatakse eeldatavate tulemusega, mis on fülogeneesipuu haru pikkuste ning neutraalsete järjestuste põhjal välja arvutatud lahknevuse tõenäosus ehk iga nukleotiidi või aminohappe esinemise tõenäosus antud positsioonis. Tõenäosuse põhjal on võimalik hinnata, kas nukleotiidi või aminohappe asendus on tolereeritav või mitte (Cooper *et al*, 2005).

Konserveeritus on annoteerimisel kasutatav oluline tunnus, kuid see ei võta arvesse adaptatsioone, mis on leidnud aset näiteks inimese ja teiste primaatide lahknemisel viimasest teadaolevast ühisest eellasest. Kõige rohkem mõjutab antud puudujääk regulatoorsete piirkondade annoteerimist, mis on evolutsioneerunud palju kiiremini kui valke kodeerivad geenid (Sim *et al*, 2012). Näiteks on transkriptsioonifaktorid liigispetsiifilised, isegi selgroogsete loomade seas (Schmidt *et al.*, 2010). Seetõttu võib olemasolevate tõendite järgi tugevalt konserveerunud järjestus, kus muutused on ennustuste järgi mitte-tolereeritavad, mõningaid muutusi ikkagi tolereerida.

1.3.1.3. Masinõppel põhinev annotatsioon

Alternatiivina bioloogilistest teadmistest sõltuvatele annotatsioonipõhimõtetele, mis vajavad tööriista loojate poolt paika pandud reegleid, on annoteerimisel võimalik kasutada võrdlust kindla funktsiooni või mõjuga variatsioonide ja mõjuta variatsioonide vahel ning töötada võrdluste põhjal välja ennustusalgoritmid.

Masinõppe meetodi puhul kasutatakse üldjoontes treeningvariatsioonide komplekti, kus järjestused on klassifitseeritud näiteks tolereeritavateks, kahjulikeks ja neutraalseteks. Olemas olevate andmete põhjal töötatakse välja algoritmid, mis ennustavad, millise kategooria variatsioonidega on uus variatsioon kõige sarnasem ning liigitavad selle vastavatesse kategooriatesse. Selline lähenemine on kasutusel näiteks üheaminohappeliste asenduste ehk *missense* või mitte-sünonüümsete variatsioonide annoteerimisel (Adzhubei *et al*, 2010; Ramensky *et al*, 2002).

Masinõppel põhinevad tööriistad võivad identifitseerida füsioloogiliselt olulisi uusi variatsioone, mida praeguste teadmiste juures ei ole võimalik teadmispõhiste meetoditega annoteerida. Kuid antud tööriistad väljastavad ennustusi skoori-põhiselt, mis ei ole kergesti tõlgendatav. Samuti võivad sellised tööriistad kasutada ära kallutatust andmestikus, näiteks mõne geeni variatsioonide üleesindatuse treeningmudelis, ning luua kallutatud algoritmi (Ritchie ja Flicek, 2014).

1.4. Ülevaade enimkasutatavatest annoteerimistööriistadest

1.4.1. Ülevaade enimkasutatavatest reeglitepõhistest annoteerimistööriistadest

Ühed enimkasutatud annotatsioonitööriistad on reeglitepõhise annotatsiooni tööriistad Variant Effect Predictor ehk VEP (McLar*en et al*, 2010), Annotate Variation ehk ANNOVAR (Wang

et al., 2010) ja SnpEff (Cingolani *et al.*, 2012). Kokkuvõtlik ülevaade enimlevinud reeglitepõhistest annoteerimistööriistadest on toodud tabelis 2.

Variant Effect Predictor on avaldatud aastal 2010 William McLareni ja tema kolleegide poolt Euroopa Bioinformaatika Instituudist ja Wellcome Trust Sangeri Instituudist. VEP on hallatav Ensembli poolt.

VEP võimaldab kasutada tööriista nii veebi- kui ka käsureapõhiselt. Mõlemal juhul on sisendiks annoteeritavate variatsioonide nimekiri koos kromosoominime ning veebipõhise tööriista väljundiks on Ensembl Genome Broweriga integreeritud annotatsioonidega tabel. Käsureapõhise programmi väljundiks on lisandunud annotatsiooniga laiendatud VCF.

Variatsiooni koordinaatide abil otsib tööriist Ensembl Core andmebaasist, mis on erinevatest allikatest pärit andmete tuumikandmebaas, välja variatsiooniga kattuvad transkriptid. Kui variatsioon langeb eksoni piirkonda, tuletatakse iga variatsiooni alleeli kohta uus koodon ja võrreldakse seda referentsi koodoniga. Variatsiooni asukohta regulatoorsete regioonide suhtes kontrollitakse Ensembl Functional Genomics andmebaasist. Annotatsiooni tulemusele lisatakse transkripti identifitseerimiskood, variatsiooni suhtelise positsioon cDNA ja valgujärjestuses (McLaren *et al*, 2010).

VEP on integreeritav mitmete tööriistadega, näiteks efekti ennustajate SIFTi („Sorting Intolerant from Tolerant“) ja PolyPhen-2-ga. Samuti võimaldab see kasutada erinevaid transkriptikomplekte teistest allikatest lisaks Ensembl'ile [12].

ANNOVAR on 2010. aastal Kai Wangi, Mingyao Li ja Hakon Hakonarsoni (Children's Hospital of Philadelpia ja University of Pennsylvania) poolt avaldatud annoteerimistööriist, mis esialgu töötas vaid käsureapõhiselt ning 2015. aastal lisandus ka veebitööriist.

ANNOVARi sisendiks on mitmed formaadid, sealhulgas standardne VCF, ning väljundiks nii täiendatud VCF kui ta tabuleeritud või komaga eraldatud tekstifail.

ANNOVAR pakub kolme tüüpi annotatsioone: geenipõhist, piirkonnapõhist ja filtripõhist. Geenipõhine annotatsioon annab informatsiooni selle kohta, kuidas variatsioon mõjutab teadaolevat geeni, näiteks olles eksoni-, introni- või splaisinguvariant, sünonüümne või mittesünonüümne või muu sarnane variant. Samuti lisandub informatsioon selle kohta, millised transkriptid on mõjutatud ning millised on muutused aminohappejärjestuses.

Tabel 2. Ülevaade enimlevinud variatsioonide annotatsioonitööriistade omadustest. Tabelis on välja toodud tööriistade kasutamise võimalused, kasutamiseks vajalikud failiformaadid, põhilised kasutatavad andmebaasid ning olulised annoteerimise lisavõimal

		VEP (McLaren *et al*, 2010) [12]	ANNOVAR (Wang *et al*, 2010)	SnpEff (Cingolani *et al*, 2012) [14]
Viide	Koduleht	http://www.ensembl.org/info/docs/tools/vep/index.html	http://www.openbioinformatics.org/annovar/ http://wannovar.usc.edu	http://snpeff.sourceforge.net
Kasutus	Viimane versioon	84 (märts 2016)	Uusim veebruar 2016	4.2 (12. 2015)
	Veebiliides	Jah	Jah	Ei
	Käsurea programm	Jah	Jah	Jah
Toetatud failitüübid	Sisendi fail	VCF, mpileup, HGSV notation	VCF	VCF, BED
	Väljundi fail	VCF	VCF, TXT (tekstifail)	VCF
Kesksed andmebaasid	RefSeq	Ensembl+RefSeq	Jah	Jah
	ENCODE	GENCODE Basic	Jah	ENCODE
	dbSNP	Jah	Jah	Jah
	Ensembl	Jah	Jah	Jah
Annotatsiooni tüübid	Geenipõhine annotatsioon	Jah	Jah	Jah
	Piirkonnapõhine annotatsioon	Ei	Jah	Jah
	Filtripõhine annotatsioon	Ei	Jah	Ei
	Muu	Regulatoorsete elementide. transkriptsioonifaktorite motiive JASPARist	Kasuab USCE referentsgenoomi ja terminoloogiat	Koespetsiifiline annoteerimine
	SIFT	Jah	Jah	dbNSFP kaudu
	PolyPhen-2	Jah	Jah	dbNSFP kaudu
	Filtreerimine	Jah, Perli skript		Jah, Käsurea käsud

Piirkonnapõhine annotatsioon annab informatsiooni selle kohta, kas variatsioonid kattuvad huvi pakkuvate piirkondadega, nagu näiteks konserveerunud genoomsed elemendid, microRNA sihtmärkalad või DNA Elementsi (ENCODE) poolt annoteeritud regioonidega.

Filtripõhise annoteerimise puhul on võimalik filtreerida annoteeritavate variatsioonide hulgast kindlatele kriteeriumtele vastavad variatsioonid, lisada 1000 Genomes Projecti andmete põhjal variatsioonide alleelisagedused, leida variatsioonidele SIFTi ja PolyPhen-2 skoorid või identifitseerida variatsioonid dbSNP andmebaasis. [13]

SnpEff on Pablo Cingolani (McGilli Ülikool, Quebeck, Kanada) ja tema meeskonna poolt ainult käsureal töötav annotatsioonitööriist. Lisaks SNPdele analüüsib SnpEff ka insertsioone ja deletsioone (Cingolani *et al*, 2012).

Üheks SnpEffi eeliseks on tema kiirus, mille tagab andmete, nii referentsandmebaasi kui ka sisendandmete, töötlemine intervallimetsa meetodil. SnpEff on integreeritav mitme teise tööriistaga, näiteks Galaxy serverisse või GATKsse. SnpEffi on pikemaks perioodiks kasutusele võtnud ka näiteks Illumina Inc., Sangeri Instituut ja AstraZeneca [14].

Sisendfailiks vajab SnpEff VCF või muud tabuleeritud tekstifaili, BED faili või SAMtoolsi mpileup faile. Väljundiks on modifitseeritud VCF või tekstifail. Sisendfail peab kindlasti sisaldama variatsiooni koordinaate, variatsiooni ID-d, referents- ja muutunud järjestust, soovitavalt ka kvaliteediskoori ning kvaliteedifiltri läbimise informatsiooni (kas kvaliteediskoor läbis filtri või mitte) ning soovi korral muud lisainformatsiooni, mida lisatakse variatsioonide tuvastamise käigus. Väljendfailiks on VCF, kuhu lisatakse iga variatsiooni kohta geneetiline informatsioon (geeni ID, nimi, transkripti ID, eksoni ID jmt) ja informatsioon variatsiooni mõju kohta (mõju tüüp, aminohappe muutus, koodoni muutus, koodoni number Consensus CDSi projekti baasil (Cingolani *et al*, 2012).

1.4.2. Järjestuse konserveeritusel ja homoloogial põhinevad tööriistad

DNA ja valkude järjestuste põhjal konserveerituse hindamist kasutavad annoteerimismeetodina mitmed tööriistad, näiteks „Functional Analysis through Hidden Markov Models“ ehk FATHMM (Shihab *et al.*, 2013), SIFT (Ng ja Henikoff, 2003) ja „Protein Variation Effect Analyzer“ ehk PROVEAN (Choi *et al.*, 2012). Nende ühiseks jooneks on skoori, mis sisuliselt väljendab konserveeritusse astet, arvutamine ning selle põhjal variatsioonide liigitamine tolereeritavateks või kahjulikeks (Ritchie ja Flicek, 2014). Kokkuvõtlik ülevaade enimlevinud konserveeritusel ja homoloogial põhinevatest annoteerimistööriistadest on toodud tabelis 2.

SIFT on algoritm, mis ennustab SNPdest ja indelitest tulevate aminohappeasenduste potentsiaalset mõju valkude funktsioonile. SIFT on loodud 2001. aastal Pauline Ng meeskonna poolt (Fred Hutchinson Cancer Research Center, Seattle, USA), 2008. aastal viidi teenus üle J. Craig Venteri Instituuti (California ja Maryland, USA) ning alates 2010. aastast Singapuri Genoomiinstituuti [15].

SIFTi tööriistadest on suurem osa kohandatud inimese SNVde analüüsi jaoks. SIFT dbSNP 138 andmebaas sisaldab ette arvutatud ennustuste väärtusi inimese ja 28 teise liigi SNPde kohta, mis on omakorda pärit NCBI dbSNP (versioon 138) andmebaasist [15]. SIFT 4G („SIFT Databases for Genomes“) sisaldab SNVsid üle 200 liigi, sealhulgas inimese, jaoks [16].

SIFT on treenitud analüüsima lisaks SNVdele ka lühikesi, kuni 20bp pikki ideleid. Indelite analüüsimiseks on arendatud SIFT Indel tööriist *(Sim et al*, 2012).

SIFT analüüsib uuritavat SNVd sisaldava geeni poolt kodeeritavat aminohappejärjestust. Antud järjestusele otsitakse kasutaja valikul üles Uniprot SwissProt, Uniprot TrEMBL või NCBI NR („NCBI non-redundant proteiin database“) valgu andmebaasidest PSI-BLASTi (Position-Specific Iterated BLAST) otsinguga sarnased järjestused (Kumar *et al*, 2009).

Asenduse mõju skoor arvutatakse SIFTis lähtuvalt uuritava SNP tagajärjel tekkinud aminohappeasenduse esinemise sagedusest antud valgu ortoloogide seas. Lõplik aminohappe esinemise tõenäosus uuritavas positsioonis on aminohappe kaalutud esinemissagedus joondatud järjestuste uuritavates positsioonides ning Dirichleti hinnangu kaalutud keskmine [15].

SIFT jagab mutatsioonide mõju valkude funktsioonile kaheks: kas kahjulikuks või tolereeritavaks. Normaliseeritud skoori korral on katseliselt selgeks tehtud, et asendused skooriga alla 0,05 on kahjulikud, 0,05-st võrdse või suurema skooriga asendused aga tolereeritavad. SIFTi normaliseeritud skoori piirväärtusteks on 0…1.

Lisaks normaliseeritud skoorile hinnatakse ka asenduste konserveeritusse väärtust. Konserveerituse väärtus ulatub 0-st, kui kõik 20 aminohapet on antud positsioonil esindatud, kuni $\log_2 20$ (=4,32), mille puhul on antud positsioonis esindatud ainult üks aminohape (Sim *et al*, 2012). Selle lõplik mediaanväärtus üle kõikide väärtuste peaks olema ligikaudu 3,0. Kui väärtus ületab 3,25, on see liiga konserveerunud ja antakse madala usaldusväärtuse hoiatus, kuna järjestused on mediaaniga võrreldes vähem mitmekesised – analüüsitavate järjestuste seas

võib olla liiga sarnaseid järjestusi, põhjustades kõrgema valepositiivse vea (Kumar *et al*, 2009) [15].

Veebitööriistade puhul on sisendiks kas FASTA formaadis järjestuse lõik uuritava valgu identifitseerimiskoodi ja asendust sisaldava järjestusega, kasutatava andmebaasi (RefSeq, NCBI) valgu identifitseerja koos aminohappeasenduste nimekirjaga või SNP ID dbSNP andmebaasis (Kumar *et al*, 2009) [15].

Käsurea- ja veebitööriista jaoks on võimalik konverteerida VCF, mpileup, Maq, SOAP ja CASSAVA failid SIFT-i formaadi failideks. SIFTi väljundiks on tekstifail, kus on ära toodud analüüsitav SNV, muutunud koodon, toimunud asendus, asenduse regioon ja tüüp ning asenduse skoor koos ennustusega. Kasutajapoolselt on võimalik lasta lisada erinevat lisainformatsiooni (Ensembl ID, OMIM haigus, alleelisagedused HapMap ja 1000 *Genome* järgi) sisaldavad tulbad. [15]

PROVEAN on annotatsioonitööriist, mis ennustab, kas aminohappe asendus, insertsioon või deletsioon muudab valgu bioloogilist funktisoooni. PROVEAN on arendatud J. Craig Venteri Instituudis (California ja Maryland, USA) Agnes P. Changi meeskonna poolt.

Sarnaselt SIFTile hindab PROVEAN asenduste mõju valgu funktsioonile vastavalt valgujärjestuse konserveerumisastmele, kuid erinevalt SIFTist kasutab PROVEAN järjestuste homoloogia hindamiseks skoori paarisjärjestuse joondusest, mis näitab joonduste sarnasust.

Lisaks SNVde mõju hindamisele on PROVEANi võimalik kasutada ka insertsioonide ja deletsioonide mõju hindamiseks (Cho*i et al*, 2012).

Analüüsiks vajalike homoloogsete ja madalama sarnasusega järjestusi otsitakse NCBI NR valguandmebaasist BLASTP tööriistaga ning valitakse välja kõik järjestused, mille E-väärtus on 0,1 või väiksem [17] [18]. Järgnevalt paigutab CD-HIT järjestused klastritesse, kuhu kuuluvad kõik vähemalt 75% sarnasusega järjestused. 30 päringujärjestusele kõige sarnasemat klastrit moodustavad *supporting sequence set*-i [17]. CD-HIT on valgu- või nukleotiidijärjestuste klasterdamise ehk sarnasuse alusel grupeerimise ja võrdlemise tööriist [19].

Delta joondusskoor arvutatakse välja igale järjestusele. Seejärel keskmistatakse skoorid klastri sees ja kasutatakse neid PROVEANi skoori arvutamiseks. Delta joonduse skoor on $\Delta(Q, v, S) = A(Q', S) - A(Q, S)$, kus Q – päringujärjestus, v – variatsioon, Q' – Q v poolt põhjustatud variantjärjestus, S – järjestus ning A – ülekattega joondusskoor kahe

valgujärjestuse vahel, mis on arvutatud kindla aminohapete asendusmaatriksi (nt BLOSUM62) peal.

Delta skoor on erinevus kahe võrreldava joonduse vahel enne ja pärast variatsiooni (SNP, indel) sisseviimist. Kui variatsiooni sisseviimise järel on päringujärjestus ning andmebaasist leitud homoloogne vaste vähemsarnased, on alust eeldada, et variatsioon on kahjulik.

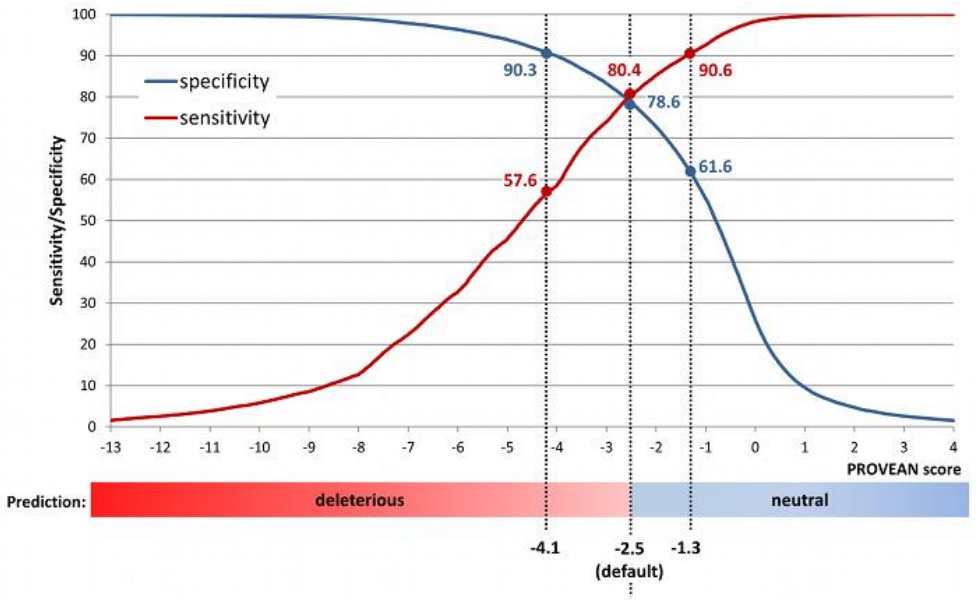
PROVEANi skoor on arvutatav järgmiselt: $PROVEANi\ skoor = \frac{1}{N}\sum_{i=1}^{Nc}\Delta c,i$, kus N – klastrite arv *supporting set*-is, Nc – järjestuste arv klastris c, Dc,i – järjestuse i delta skoor klastris c.

Kui PROVEANi skoor on väiksem või võrdne eelseadistatud lävendiga (vaikimisi -2,5), loetakse mutatsioon kahjulikuks. Kui skoor on üle -2,5, loetakse mutatsioon neutraalseks.

Lävend valitakse lähtuvalt tasakaalustatud täpsusest (tundlikkuse ja spetsiifilisuse keskmine) ja tasakaalustatud eraldatusest (maksimeerib tundlikkuse ja spetsiifilisuse miinimumi). Valiku põhimõtet illustreerib joonis 1.

Skoori väärtused sõltuvad analüüsitavatest järjestustest. Inimese valguvariatsioonide skoori piirväärtused jäävad vahemikku ~ -38,5…+11,5 (Choi *et al*, 2012).

Töörista sisendiks on FASTA fail, kus on kirjeldatud valgujärjestus järgnevalt: <positsioon>,<referentsaminohape>,<asendusaminohape> või HGVSi (*Human Genome Variation Society*) formaadis. Genoomivariantide puhul on sisendiks komaga eraldatud väärtused formaadis <kromosoom>,<positsioon>,<referentsalleel>,<variandialleel>, <märkused (vajadusel)>. Väljundiks on tabuleeritud tekstifail, kus tulpadeks on informatsioon analüüsitud variandi kohta, PROVEANi skoori info ja ennustus [17].



Joonis 1. PROVEANi skooridele vastavad otsuse tundlikkus (*sensitivity*) ja spetsiifilisus (*specificity*) koos vaikesättena välja toodud skoori otsustusläve väärtusega (Choi et al, 2012). Joonis illustreerib, et otsustuslävest sõltuvalt väheneb tundlikkuse tõstmisel spetsiifilisus ja vastupidi.

FATHMM on kodeerivate ja mitte-kodeerivate SNPde mõju ennustamiseks loodud tööriist. See on arendatud professorite Julian Gough ja Tom R. Gaunti meeskondade poolt (Bristoli Ülikool, Suurbritannia).

FATHMM kasutab ennustusskooride arvutamiseks peidetud Markovi mudelit (Shihab *et al*, 2013).

FATHMMi veebitööriist võimaldab valida kodeerivate (keskendude pärilike haiguste, kasvajate ja haigus-spetsiifilise SNVde annotatsioonile) ja mitte-kodeerivate SNVde mõju ennustamise vahel [20].

FATHMMi ennustus põhineb arvutusliku skoori tõlgendamisel. Analüüsida on võimalik nii liigispetsiifilisi kui ka liigist sõltumatuid andmeid, kasutades vastavalt kas kaalutud või kaalumata skoore.

Kaalutud skoor arvutatakse valemiga $kaalutud\ skoor = ln\frac{(1{,}0-Pw)(Wn+1{,}0)}{(1{,}0-Pm)(Wd+1{,}0)}$, kus P_w ja P_m tähistavad referentsi ja mutantse aminohappe esinemise tõenäosust, W_d ja W_n tähistavad relatiivseid haigusseoseliste ja funktsionaalselt neutraalsete aminohappeasenduste peidetud Markovi mudelile kaardistamise sagedusi.

Tabel 3. Aminohappelise järjestuse konserveerituse astmest lähtuvate annotatsiooniööriistade omadused. Tabelis on toodud välja veebiaadress, kust pääseb ligi

antud tööriistadele ning sisendi tüübid. Lisaks on toodud välja tööriistade poolt arvutatavate skooride meetodid, väärtused ning nende tõlgendamine. Samuti on toodud välja põhilised andmebaasid, mille baasil tööriistad enda andmestikke koostavad.

	SIFT (Ng ja Henikoff, 2003; Sim *et al.*, 2012)	PROVEAN (Choi *et al.*, 2012)	FATHMM (Shihab *et al.*, 2013)
Viide	sift.jcvi.org/	provean.jcvi.org/	fathmm.biocompute.org.uk
Veeb	Jah	Jah	Jah
Käsurida	Jah	Jah	Jah
Sisendi järjestus	SNV, indel	Valgujärjestus, SNV, indel	SNP või valgujärjestus
Skoori arvutamise meetod	Dirichleti segu	Delta joonduse skoor	HMM
Skoori piirväärtused	Konserveerituse väärtus: Minimaalne: 0 maksimaalne: $\log_2 20$ (= 4,32) Normaliseeritud tõenäosus:minimaalne: 0, maksimaalne: 1	Skoori piirväärtused sõltuvad analüüsitavatest järjestustest.	
Skoori tõlgendus	Normaliseeritud tõenäosus: Madalam kui 0,05 – kahjulik variatsioon 0,05 või kõrgem – tolereeritav variatsioon	Skoor lävendist madalam või võrdne – kahjulik variatsioon Skoor lävendist kõrgem – neutraalne variatsioon Vaikimisi on lävendiks 2,5.	Skoor madalam kui 0 – kahjulik variatsioon Skoor väärtusega 0 – neutraalne variatsioon Skoor kõrgem kui 0 – soodne variatsioon Empiiriliste andmete põhjal soovitatud otsustuskünnis väärtusega -0,75.
Põhilised andmebaasid	RefSeq, UCSC, CCDS and Ensembl gene annotations	NCBI NR *protein database*, UniProt	NCBI NR *protein database*, UniProt, HGMD, SwissVar, VariBnc

Kaalumata skoor arvutatakse valemiga $kaalumata\ skoor = \ln \frac{Pm/(1,0 - Pm)}{Pw/(1,0-Pw)}$, kus P_w ja P_m tähistavad metsiktüüpi ja mutantse aminohappe esinemise tõenäosust.

Teoreetiliselt tähendab FATHMM skoor alla 0, et tegu on ebasoodsa asendusega, 0 tähistab neutraalset otsustuskünnis ja skoori väärtus üle 0 soodsat asendust. Empiiriliselt on testmisel tuvastatud, et skoori väärtus 0,75 oleks otsustuskünnisena täpsem lähtepunkt (Shih*ab et al*, 2013).

Sisendiks võib olla SwissProt/TrEMBL, RefSeq ja Ensembli valgujärjestuse identifitseerija koos välja toodud asendusega või dbSNPis esindatud SNV unikaalne identifitseerimiskood. Samuti on võimalik sisendina kasutada VEPi annotatsiooni sisaldavaid VCF faile, mida tuleb konverteerida õige formaadi saamiseks parseVCF.py skripti [21] abil.

Väljundiks on tekstifail, kus on välja toodud dbSNP ID, valgu ID, asendus, skoor ning skoori tõlgendus [20].

1.4.3. Masinõppel põhinevad tööriistad

Tuntuim masinõppel töötav annoteerimistööriist on PolyPhen ning selle välja vahetanud PolyPhen-2. PolyPhen-2 lähenemine ühendab mitmese joondamise (sarnase SIFTile) saadaval olevate valgu struktuuriandmete, Pfam domeenide ja teiste andmetega. PolyPheni on treenitud tegema vahet polümorfsete ning UniProtis haigusseoselise annoteeringuga järjestuste vahel (Ritchie ja Flicek, 2014).

Suunatud masinõpe tähendab, et et arvuti peab treeningandmestikust tuletama reegli, mida saaks rakendada treeningandmestiku-välistele andmetele. Treeningandmestik kooseb näidisandmetest, mis hõlmavad sisendandmeid ning neile vastavaid soovitud tulemeid. Eesmärgiks on arendada treeningandmestiku põhjal selline algoritm, mis määraks korrektselt sisendandmete tagajärgi siis, kui andmed ei ole kuulunud treeningandmestikku. (Aggarwal, 2015)

Teistest tööriistadest on saadaval PolyPheniga sarnast algoritmi kasutav MutationTaster, SNAP, PhD-SNP (Ritchie ja Flicek, 2014)

PolyPhen ja PolyPhen-2 on Heidelbergi The European Molecular Biology Laboratorys (EMBL) ja Harvard Medical Schoolis Shaml Sunyaevi ja kolleegide poolt arendatud annotatsioonitööriistad, mis lisaks järjestusepõhisele konserveeritusse analüüsile rakendavad ka valgu struktuuriparameetrite-põhist analüüsi (Ramensky *et al*, 2002) (Adzhubei *et al.*, 2010).

PolyPheni on võimalik kasutada vaid käsureatööriistana, PolyPhen-2 puhul on saadaval nii käsurida- kui ka veebitööriist Aastal 2012 asendati Polyphen PolyPhen-2-ga [22].

PolyPhen-2 lähtub SNVde annoteerimisel erinevatest järjestuse- ja struktuuri-põhistest andmetest, mida tõlgenab tõenäosuslik klassifikaator.

Esimese sammuna määrab PolyPhen-2, millises valgu piirkonnas on aminohappe asendus toimunud. Selleks leitakse uuritavale valgule vastava valgu UniProtKV/Swiss-Prot andmebaasist. Sobiva valgu leidmisel kontrollitakse, kas aminohappe asendus võib tekitada ruumilisi konflikte teiste aminohapetega. Samuti kontrollitakse, kas asendus toimus valgu transmembraanses piirkonnas, sellisel juhul kasutatakse PHAT transmembraanset maatriksit, et hinnata annoteeritavate SNVde võimalikku mõju.

Järgnevalt otsitakse BLASTi abil UniRef100 andmebaasist annoteeritavale järjestusele sarnaseid homolooge. Edasiseks analüüsiks jäetakse alles järjestused, mille identsus uuritava järjestusega on vahemikus 30%...94% ning joondatul peab päringjärjestusega olema vähemalt 75 ühist aminohapet.

Saadud mitmikjoondust kasutab integreeritud PSIC (Position-Specific Independent Counts) tarkvara ning arvutab välja profiilimaatriksi. Maatriksi elemendiks ehk profiiliskooriks on logaritmiline suhe aminohappe esinemise tõenäosusest antud kindlal positsioonil/aminohappe esinemise tõenäosus igas positsioonis. PolyPhen-2 arvutab välja vahe polümorfse positsiooni mõlema alleeli kahe profiiliskoori vahel.

Täiendav valgu tertsiaarstruktuuri kasutamine variatsioonide mõju ennustamisel aitab määrata, kas asendus lõhub tõenäoliselt valgu hüdrofoobse tuuma, elektrostaatilisi interaktsioone, interaktsioone ligandidega ja teisi olulisi parameetreid. Kui uuritavat valgustruktuuri andmebaasis pole, võib kasutada homoloogsete valkude struktuure.

tertsiaarstruktuuri andmete saamiseks otsib PolyPhen-2 BLASTi abil Protein Structure Database'ist (PDB) analüüsitavale järjestusele vasteid järgnevate kriteeriumite järgi: järjestuse identsuse lävi uuritava järjestusega on 50%, mis tagab valgu põhiosade struktuuride konserveerituse, minimaalne joondatud aminohapete arv peab olema 100 ning maksimaalne lünkade arv joonduses on 20.

PolyPhen-2 kasutab Dictionary of Secondary Structure in Proteins (DSSP) andmebaasi, kust pärinevad järgnevad struktuuriandmed: valgu sekundaarstruktuur vastavalt DSSP nomenklatuurile, lahustile ligipääsetav ala ehk valgu üldpindala ning phi-psi dihedraalnurgad.

Struktuuriandmete analüüsil on võimalik tuvastada, kas aminohappe asendus muudab märgatavalt valgu funktsiooni, puutudes näiteks kokku teiste aminohapetega või vähendades seostumispindala teise valguga, seega kahjustades valgu funktsiooni.

Alleeliasenduse funktsionaalsuse hindamiseks kasutab PolyPhen-2 Naiivse Bayesi klassifikaatori, mida on treenitud suunatud masinõppega.

PolyPhen-2-te on treenitud kahe andmekoguga, HumDiviga ja HumVariga. HumDiv koosneb kõigist kahjulikest alleelidest, mille puhul on teada variatsiooni molekulaarbioloogilised mõjud, mis põhjustab Mendeliaalseid haigusi. Andmed on võetud UniProtKB andmebaasist ning lisaks inimese-spetsiifiliste valkude sisaldab see treeningkogumik inimesele evolutsiooniliselt lähedaste imetajate homolooge, mis eeldatavalt ei ole kahjulikud.

HumVar koosneb kõigist inimese haigusseoselistest mutatsioonidest UniProtKB andmebaasist, lisaks sagedastest (MAF > 1%) mitte-sünonüümsetest SNPdest, millel puudub annotatsiooni järgi seos haigustega ning mida on käsitletud mitte-kahjulikena.

Mendeliaalsete haiguste diagnostika eeldab vahet tegemist olulise mõjuga variatsioonide ning kõikide teiste, sealhulgas kergel kahjuliku variatsioonide vahel. Selleks on sobilik HumVari mudel. Haruldaste kompleksete fenotüüpidega seotud lookuste määramisel on kasutusel HumDiv. Samuti on HumDiv kasutusel tiheda kaardistamisega regioonide või uuritavate geenide evolutsioneerumise uurimisel, mis eeldavad kergelt kahjulike alleelide käsitlemist kahjulikena [23]. Vaikimisi kasutatab näiteks VEP HumVar andmekogusse, kuid vajaduse korral on võimalik seda asendada HumDiv admekoguga [12].

Iga variatsiooni jaoks arvutab PolyPhen-2 välja tõenäosuse, et see variatsioon on kahjulik ning annab hinnangulise valepositiivse (VP) ja valenegatiivse (VN) määra. Variatsioonid jaotatakse ka ühte kolmest kategooriast - tolereeritavad, võimalikult kahjulikud ning tõenäoliselt kahjulikud, mida tehakse VP määra baasil. Mõlema puhul on skoori piirväärtuseks 0 (healoomulik)…1 (kahjulik).

PolyPhen-2 versiooni 2.1.0 puhul määratakse variatsioonid kahjulikkuse järgi kategooriasse Tabelis 4 toodud reeglite põhjal. Kui pole piisavalt andmeid ennustuse tegemiseks, märgitakse variatsiooni kohta raportisse *unknown* ehk teadmata [23].

Tabel 4. Variatsioonide kahjulikkuse määramine valepositiivse määra järgi. HumDiv ja HumVar on PolyPhen-2 poolt kasutatavad treeningpaketid.

	HumDiv	HumVar
Tolereeritav	< 5%	< 10%
Võimalikult kahjulikud	5%...10%	10%...20%
Tõenäoliselt kahjulikud	> 10%	> 20%

1.5. Annoteerimisel kasutatavad transkriptid, andmebaasid ja terminloogia

Annoteerimistööriistad sõluvad palju andmestikest, mille alusel või millega variatsioone võrreldes algoritmid oma otsuseid teevad. Informatsiooni hulk täieneb pidevalt ning inimese genoom on tänaseks veel täielikult annoteerimata. Sealhulgas on andmebaasidest, näiteks inimese genoomi annoteerimise referentsis GENCODE, puudu paljud transkriptid ning olemas olevad transkriptid ei ole täielikult annoteeritud (Mudge *et al*, 2013). Transkriptide komplekti kuuluvad lisaks transkriptide järjestustele muuhulgas informatsioon genoomi ja transkriptide struktuuri (järjestus, eksonite, intronite, UTRide ja regulatoorsete piirkondade asukohad) kohta ehk informatsioon, mis on vajalik variatsiooni tagajärje ennustamiseks (McCarthy *et al.*, 2014).

Annoteerimisel kasutatavad transkriptide kogud, neist levinuimad on RefSeq, GENCODE ja Ensembl, ei sisalda ainult cDNA ja mRNA püüdmisel ning RNA sekveneerimise saadud järjestusi, vaid ka nende annotatsioone. GENCODEi andmebaasi annotatsioon lähtub referentsgenoomist, mitte tavapärasest transkriptoomikast. See ühendab Human and Vertebrate Analysis and Annotation grupi (HAVANA) poolt käsitsi annoteeritud informatsiooni Ensembli arvutuslike mudelitega. RefSeq ühendab samuti käsitsi annoteerimise arvutuslike protsessidega, kuid inimese annoteerimise aluseks on cDNAd, informatsioon viiakse kromosoomikoordinaadistikuga kokku alles pärast annoteerimist (Mudge *et al*, 2013).

Transkriptide komplekti valiku olulisust näitab 2014. aastal Davis McCarthy poolt läbi viidud võrdlus annotatsioonide vahel kasutades RefSeqi ja Ensembli (HAVANA ja GENCODE) transkriptikomplekte. RefSeq andmebaasi 57. versiooni 105258st transkriptist kasutas ANNOVAR annoteerimisel 41 501 transkripti. Sarnaselt kasutas ANNOVAR Ensebli andmebaasi 69. versioonist 208 677st transkriptist 115 901.

Võrdluseks annoteeriti sama variatsioonide kogu ANNOVARiga, kasutades kahte trankriptikomplekti ning valiti analüüsiks välja potentsiaalsed *loss-of-funcion* (raaminihke deletsioonid ja insertsioonid, stopkoodoni tekkimine või kadumine) ning suurem osa splaissingukohti mõjutavaid variatsioone. Kui üle 80 miljoni variatsiooni annoteerimisel oli üleüldine kokkulangevus umbes 85%, oli see võimalikke LoF variatsioonide annotatsioon kahe erineva transkriptikomplektiga vaid 44%. Seejuures määrass ANNOVAR Ensembli transkriptidest lähtuvalt eksoni piirkonna variatsiooniks tuhandeid variatsioone rohkem kui RefSeqi transkripte kasutades. Samuti annoteeris tarkvara Ensembli transkriptide baasil üle 2000 raaminihke indeli ja üle 1000 stopkoodoni tekke või kadumise rohkem. Selline

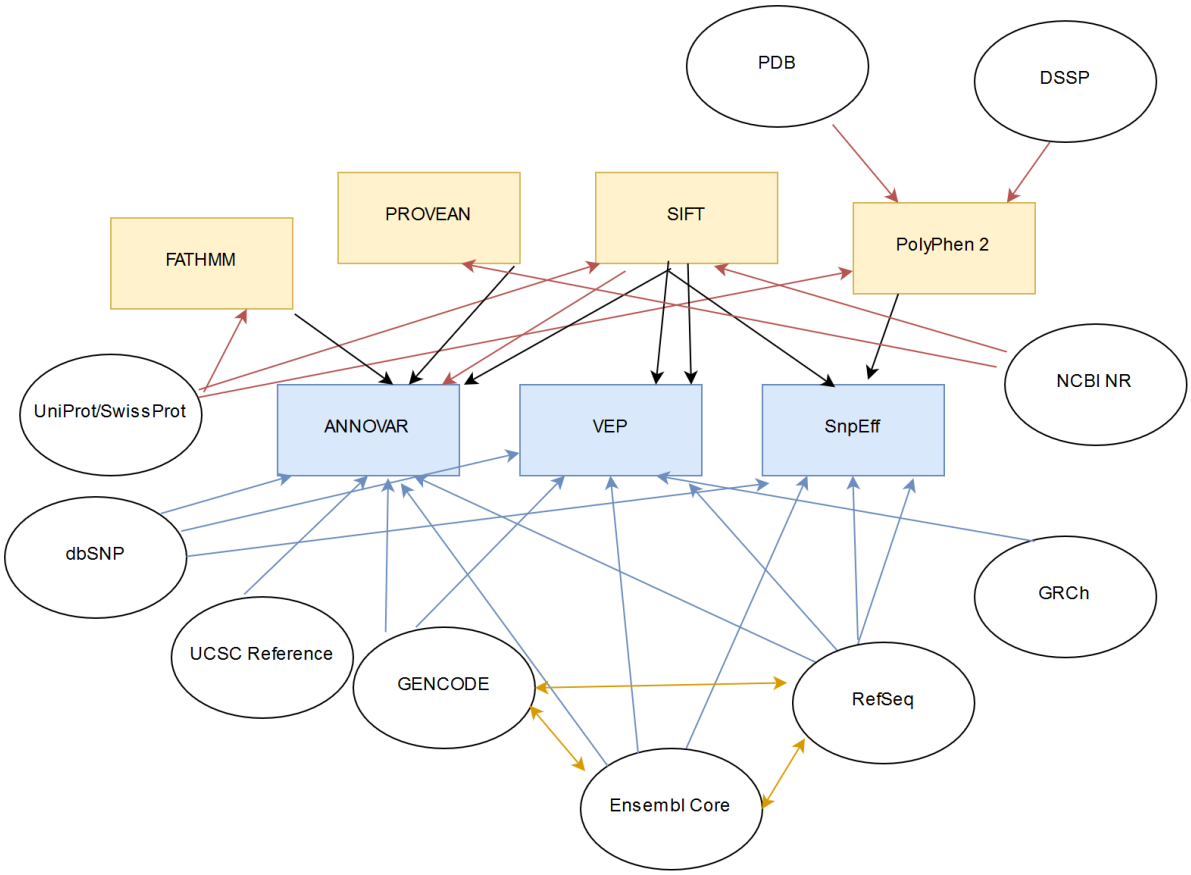
asümmeetria on põhjustatud kahe transkriptikomplekti erinevast sisust – RefSeq sisaldab 105 258 transkripti ehk selle valke kodeerivad järjestused katavad umbes 1,07% genoomist. Ensembl aga sisaldab 208 677 transkripti, mis katab umbes 28% genoomist, sealhulgas introneid, ning valke selle kodeerivad transkriptid katavad umbes 1,12% genoomist.

Antud uuring juhib tähelepanu õigete transkriptide valiku olulisusele. Võimaluse korral tuleb eelistada transkriptikomplekti, milles olvad transkriptid on ekspresseeritud proovi päritolukoes ning arvestada transkriptide annoteerimiskvaliteedi ja genoomi katvusega.

Samuti põhjustab palju erinevusi annoteerimistööriistade poolt kasutatav terminoloogia. Kahe annotatsioonitööriista, ANNOVARi ja VEPi annotatsioonide võrdluses selgus, et paljud lahknevused tööriistade tööriistade poolt antud hinnangute vahel olid näilised – tööriistad võivad nimetada samu tagajärgi erinevate terminitega, tuua välja erinevaid tagajärgi (nt ANNOVARil on ainult *splice site variant*, VEPil *splice acceptor, splice donor* ja *splice variant*) või ainult ühte kõige olulisemat tagajärge määrates määravad tööriistad ühele ja samale variatsioonile erinevaid tagajärgi (vaata ka lisa 1). Seega tuleb annoteerimistööriistu, eriti paralleelselt mitut erinevat kasutades jälgida terminoloogiat ning vajadusel seda ühtlustada (McCarthy *et al*, 2014).

Terminoloogia ühtlustamseks on loodud SO Sequence Ontology Browser. See sisaldab variatsioonide võimalikke tagajärgi koos definitsioonidega ning SO terminoloogia on võetud kasutusele Ensembli ja VEPi poolt (Eilbeck *et al*, 2005).

Homo- või ortoloogiliste järjestuste otsimisel ja valkude struktuuriparameetrite saamisel kasutatakse erinevaid geeni- ja valgujärjestuse andmebaase. Ülevaade enimkasutatud andmebaasidest on toodud lisas 1. Põhilised seosed andmebaaside ja annoteerimistööriistade vahel on toodud järgnevas joonisel 2.



Joonis 2. Põhilised seosed andmebaaside ja annotatsoonitööriistade vahel. Töö autori poolt koostatud ülevaatlikul joonisel on toodud välja sinise taustaga kastides kolm antud töös esitatud reeglite-põhist annoteerimistööriista, kollase taustaga kastides arvutuslikel ennustusmeetoditel töötavad annoteerimistööriistad ning valgetes ovaalides põhilised andmebaasid, mida tööriistad oma tööks vajavad. Noole suund näitab, millisele tööriistale informatsioon edastatakse või millisesse annotatsoonitarkvarasse on tööriist integreeritud. Kahepoolsed nooled näitavad võimalikke andmebaaside omavahelist annoteeritud ja/või kureeritud informatsiooni vahetamist ja integreerimist. Punaste nooltega on tähistatud valkude järjestute või struktuuri puudutava informatsiooni liikumine, sinised nooled vastavad DNA või RNA järjestusi puudutava informatsiooni liikumist.

2. EKSPERIMENTAALOSA

2.1. Töö eesmärgid

Antud töö eesmärgiks on anda ülevaate SNVde annoteerimiseks kasutatavatest tööriistadest ning selgitada nende tööpõhimõtteid.

Sekveneerimisel saadud andmetest variatsioonide tuvastamine ja annoteerimine on mitmeastmeline protsess, mis eeldab mitmete bioinformaatiliste tööriistade kasutamist. Käesolev töö annab ülevaate antud protsessist ning selgitab täpsemalt annoteerimistööriistade ja andmebaaside sobiliku valiku olulisust.

Töö oluliseks osaks on erinevate enimkasutatud annoteerimistööriistade võimalike kitsaskohtadele ja piirangutele tähelepanu juhtimine, mis aitaks edaspidi genoomi variatsioonide annoteerimist nõudvate tööde planeerimisel valida õigeid tööriistu või andmebaase.

Eksperimentaalosa ülesandeks oli annoteerida erinevate joondusalgoritmidega joondatud lugemitest tuvastatud valepositiivseid SNVsid. See võimaldas annoteerimistulemuste tõlgendamise abil määrata, kas võimalikud valepositiivselt määratavad proovid võivad osutuda reaalsete proovide analüüsimisel komistuskiviks variatsioonide ja fenotüübi seoste määramisel.

2.2. Materjalid ja metoodika

2.2.1. Annoteeritavate SNVde saamine ja ülevaade

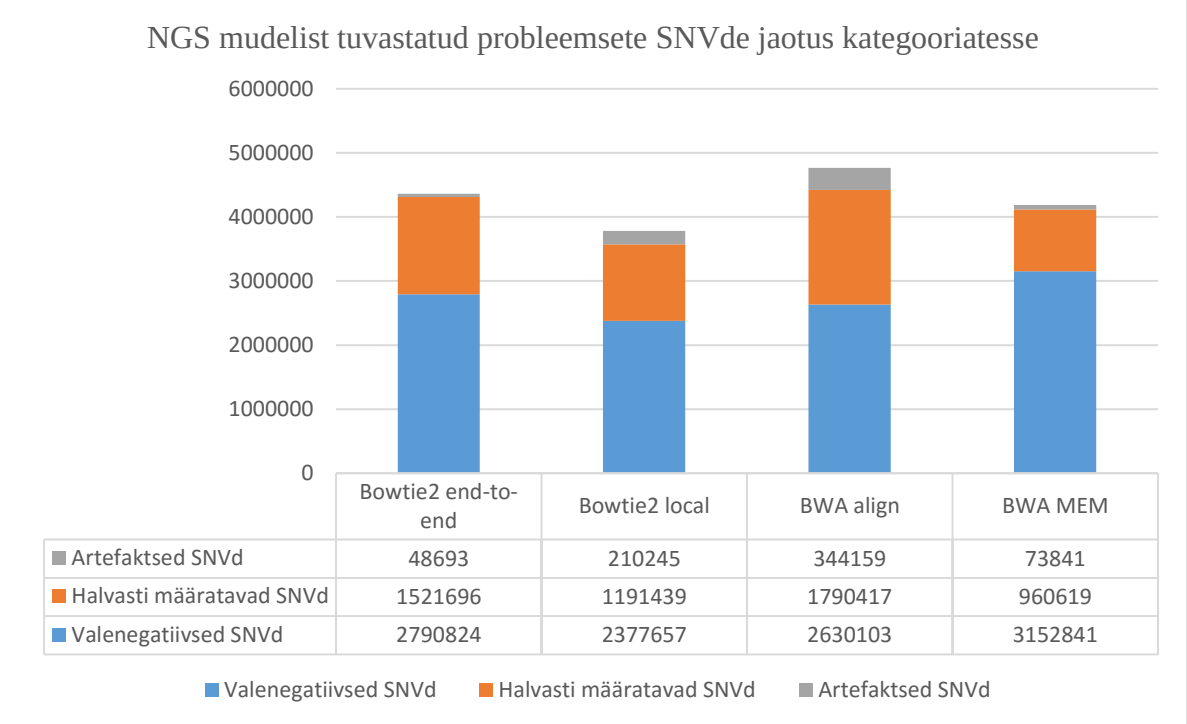
Töö lähtub teoreetilisest inimese genoomi sekveneerimine mudelist, mis on koostatud virtuaalsest kõrgkvaliteedilistest teise põlvkonna sekveneerimise lugemitest. See on loodud Genome Reference Consortiumi hallatava inimese referentsgenoomi versiooni 37.p13 (GRCh37.p13) ning dbSNP andmebaasi versioonis 135 sisalduvate SNVde põhjal. Kasutatud oli neljale erinevale DNA pikkusele vastavaid virtuaalsete paarislugemite (2 x 102 nukleotiidi) raamatukogusid, mis kas sisaldasid variatsioonidega lugemeid või olid ainult referentsgenoomist tuletatud täielikult joonduvad lugemid ja nende kombinatsioone Lugemeid joondati inimese referentsgenoomile (GRCh27.p13), kasutades BWA-MEM, BWA *align* ning Bowtie2 *local* ja *end-to-end* joondusalgoritme. Joondustest leiti SNVd SAMtools paketiga (versioon 1.1.18).

Võrreldes tulemusi mudelisse sisestatud SNVde loendiga, sai SNVd jagada järgnevatesse kategooriatesse:

- universaalselt sõltumata raamatukogu tüübistõigesti tuvastatavad SNVd,
- artefaktsed SNVd,
- valenegatiivsed SNVd,
- sõltuvalt raamatukogu valikust erinevalt määratavad SNVd.

Mudeli koostamise ja SNVde tuvastamise viisid läbi Tartu Ülikooli molekulaar-ja rakubioloogia instituudi bioinformaatika õppetooli teadur Ulvi Gerst Talas ja programmeerija Mikk Eelmets.

Kokku annoteeriti 45 651 821 Bowtie2 *end-to-end* poolt joondatud järjestustest saadud SNVd, 45 813 368 Bowtie2 *local* poolt joondatud järjestustest saadud SNVd, 45 947 281 BWA *aligni* poolt joondatud järjestustest saadud SNVd ning 45 676 964 BWA-MEM poolt joondatud järjestustest saadud SNVd. Probleemseks osutunud SNVde jaotus kasutatud joondusalgoritmi ja määratavuse kategooria järgi on toodud joonisel 3.



	Bowtie2 end-to-end	Bowtie2 local	BWA align	BWA MEM
Artefaktsed SNVd	48693	210245	344159	73841
Halvasti määratavad SNVd	1521696	1191439	1790417	960619
Valenegatiivsed SNVd	2790824	2377657	2630103	3152841

Joonis 3. Mudeli põhjal probleemseteks osutunud SNVde jaotus määratavate kategooriate järgi. Tulpades on toodud välja neli joondusalgoritmi, mille lugemite joondusest SNVsid määrati. Jooniselt on välja jäetud alati korrektselt määratavad SNVd. SNVsid on lähtuvalt tuvastamise õigsusest võimalik jaotada järgnevatesse kategooriatesse universaalselt õigesti tuvastatavad SNVd, artefaktsed SNVd, valenegatiivsed SNVd ja sõltuvalt raamatukogu valikust erinevalt määratavad SNVd.

Üldjoontes võimaldavad erinevate joondusalgoritmide poolt joondatud lugemid tuvastada suurema osa SNVsid korrektselt: Bowtie2 *end-to-end* poolt joondatud lugemitest tuvastati õigesti 90,4%, Bowtie2 *local* poolt joondatud lugemitest tuvastati õigesti 95,5%, BWA-*align* poolt joondatud lugemitest 89,6% ning BWA-MEM poolt joondatud lugemitest 90,8%.

2.2.2. Annoteerimine Variant Effect Predictoriga

Saadud SNVde annoteerimiseks rakendati Ensembl'i tööriista Variant Effect Predictor ehk VEP. Veebitööriist ning käsureaprogramm on saadaval aadressil http://www.ensembl.org/info/docs/tools/vep/index.html.

Annoteerimiseks kasutati VEPi versiooni 81 käsureaprogrammina. Annoteerimisel kasutati järgnevate andmebaaside informatsiooni:

- referentsgenoomiks oli GRCh37.p13,
- dbSNP andmebaasi versiooni 142,
- NHLBI Exome Sequencing Project andmebaasi seisuga 03. november 2014,
- Catalogue Of Somatic Mutations In Cancer (COSMIC) versiooni 71,
- GENCODE andmebaasi versiooni 19,
- Human Gene Mutation Database (HGMD-public) andmebaasi seisuga aprill 2014,
- ClinVar andmebaasi seisuga jaanuar 2016.

Samuti lisati igale SNVle VEPi poolt, kui vastavad andmed olid saadaval, eelarvutatud PolyPhen-2 (versioon 2.2.2) ning SIFTi (versioon 5.2.2) mõju ennustuse skoorid.

Annoteerimisel kasutati võrreldes vaikeseadistustega mitmeid täiendfunktsioone, mis lisasid annotatsioonile informatsiooni variatsiooni geeni, transkripti, esinemissageduse ja variatsiooni tagajärgede kohta. Annoteerimisel kasutatud käsurida koos lisafunktsioonidega ning nende võrdlus veebiversiooniga on toodud lisas 2-is.

Sisendiks kasutati NGS Illumina inimese virtuaalses sekveneerimismudelis tuvastatud valepositiivseid ja valenegatiivseid tulemusi (SNVsid). Sisend- ja väljundfaili formaadiks oli kasutusel VCF formaat. Väljundfaili näidis on toodud lisas 3.

Käsureapõhine annoteerimine viidi läbi Tartu Ülikooli molekulaar- ja rakubioloogiainstituudi bioinformaatika õppetooli serveris.

Väljundfailide loetavamaks muutmiseks need tabuleeriti. Väljundfailidest huvi pakkuvate variatsioonide sorteerimiseks kasutati awk ja UNIXi *shell*i käske. Andmete lõplikuks

sorteerimiseks ning visualiseerimiseks kasutati Michrosoft Office'i tarkvarapaketi programmi Excel.

Variatsioonide sorteerimine toimus nende võimalke füsioloogiliselt oluliste mõjude alusel, hinnates eelkõige nende asukohta geenis, geeni tüüpi (jättes välja pseudogeenid ja RNAd) ning variatsioonide võimalike tagajärgede tüüpe.

2.3. Ülevaade artefaktselt määratavatest variatsioonidest nende füsioloogilise olulisuse seisukohast

Kuigi suuremat osa SNVdest määratakse joondatud algoritmidest korrektselt, jääb alles hulk SNVsid, mida tuvastatakse valesti. Eksperimentaalse töö käigus on annoteeritud artefaktsed või sageli artefaktsetena tuvastatud SNVd ehk need SNVd, mida teoreetilise mudeli puhul sisendjärjestusse disainitud ei olnud.

Tulemusena näitas artefaktsete SNVde annotatsioon, et valdav osa artefaktseid SNVsid jääb väljapoole valke kodeerivaid alasid. Täpsemalt jääb Bowtie2 *end-to-end* joonduse lugemitest tuvastatud vale-SNVdest 80% kas intronite või geenidevahelisse alasse, Bowtie2 *local* joondustest tuvastatud vale-SNVdest jääb 75% intronite või geenidevahelisse alasse, BWA-MEM joondustest tuvastatud vale-SNVdest jääb 84% intronite või geenidevahelisse alasse ning BWA *align* joondustest valesti tuvastatud SNVde puhul on antud osakaaluks 75%. See võimaldab jätta ebaolulisena välja need SNVd, millel pole annotatsiooni kohaselt ei teadaolevat ega ennustatavat funktsionaalset mõju. Valesti tuvastatud SNVde näiliste tagajärgede jaotus on toodud tabelis 5.

Et hinnata täpsemalt, kui kaalukat eksitavat informatsiooni võib artefaktsete SNVde mõjul annoteerimisel saada, sorteeriti välja annotatsiooni järgi võimalikku füsioloogiliselt olulist mõju omavad variatsioonid ja neile vastavad geenid. Selleks sorteeriti välja geenidega seotud variatsioonid, mis asuvad valke kodeerivates alades. Seejärel hinnati omakorda veel tõenäoliselt üliolulist mõju, nagu näiteks enneaegse stopkoodoni tekkimine või alternatiivse splaissingu-koha loomine, omavate variatsioonide osakaalu. Ülevaade erinevate joondusalgoritmide poolt joondatud lugemitest saadud artefaktsetest variatsioonidest, mis võiksid omada näitlikult väga olulist molekularbioloogilist mõju, on toodud tabelis 6.

Tabel 5. Tabelis on välja toodud valepositiivselt tuvastatud SNVde arv vastavalt variatsioonide kõige tõsisemale tagajärjele. Variatsioonide arvud on toodud eraldi välja iga joondusalgoritmi poolt joondatud lugemitest määratud SNVde hulga kohta. Eraldi on rõhutatud variatsioone, mille kohta annab annoteerimine piisavalt palju informatsiooi ning mille omadused on sellised, et võiksid eksitada füsioloogiliselt oluliste variatsioonide otsinguil.

	Bowtie2 *end-to-end*	Bowtie2 *local*	BWA *align*	BWA MEM
UTR-3'-Järjestus	430	2030	3547	891
UTR-5'-järjestus	83	524	985	203
Allavoolu variant	2776	13874	24118	4831
Variatsioon geenidevahelises piirkonnas	18067	78911	112746	25862
Variatsioon intronis	21419	80732	143808	29280
miRNA	2	10	21	0
Missenssvariatsioon	509	3494	6204	1695
Variatsioon mitte-kodeerivas transkriptis	1197	9154	16436	3448
Splaissimiskoha aktseptor	5	52	100	19
Splaissimise doonor	9	67	123	27
Splaissimispiirkonna variatsioon	77	420	882	162
Startkoodoni kadu	2	9	14	3
Stopkoodoni teke	26	147	251	72
Stopkoodoni kadu	0	7	15	3
Stopkoodonit mittemuutev variatsioon	1	2	5	3
Sünonüümne variatsioon	237	1675	3013	755
Ülesvoolumutatsioon	3853	19137	31891	6587
Kokku	48693	210245	344159	73841

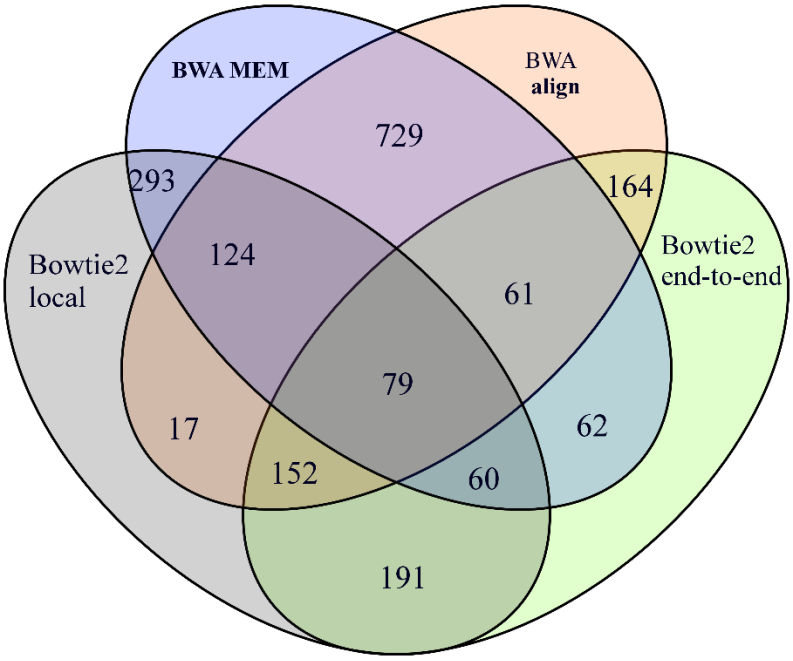
Tabel 6. Ülevaade erinevate joondusalgoritmide poolt joondatud lugemitest saadud artefaktsetest variatsioonidest, mis võiksid omada näitlikult väga olulist molekularbioloogilist mõju.

	Bowtie2 *end-to-end*	Bowtie2 **local**	BWA MEM	BWA *align*
Variatsioon splaissimiskohas	27	38	78	18
Variatsioon splaissimiskohas ja mitte-kodeerivas geenis	1	47	81	16
Stopkoodoni teke	44	113	218	63
Stopkoodoni teke ja splaissimissait		2	3	
Stopkoodoni kadu		4	6	1
Stopkoodoni kadu ja splaissimissait		1	1	
Kokku	72	205	387	98

Lisaks splaissingut ja stopkoodoneid mõjutavatele variatsioonidele on olulised ka mitte-sünonüümsed variatsioonid. Petlike, näiliselt olulistena näivate artefaktsete SNVde välja filtreerimisel valiti välja lisaks splassingut ja stop- ning startkoodoneid mõjutavatele variatsioonidele sellised variatsioonid, mis asuvad valke kodeerivates piirkondades, omavad seost kindla geeniga ning mille SIFTi ja PolyPhen2 skoori väärtused näitasid, et tegu ei ole tolereeritava asendusvariatsiooniga. Näiliselt oluliste artefaktsete variatsioonide kokkuvõte on esitatud tabeli kujul lisas 4, nimekirja võimalikest näiliselt olulistest artefaktsetest variatsioonidest on võimalik töö autorilt saada digitaalselt.

Näiliselt olulisteks variatsioonideks liigitatud SNVde abil on võimalik illustreerida, kuidas joondusalgoritmid eksivad nii kokkulangevalt kui ka erinevalt individuaalselt. Joonisel 4 on näidatud, kuidas erinevate joondusalgoritmide poolt saadud lugemite joondustest on määratud nii samu kui ka erinevaid näiliselt olulisi artefaktseid variatsioone. See tähendab, et osa

lugemite joondamisel eksivad kõik joondusalgoritmid ühemoodi, samas mõne lugemi puhul joondab mõni algoritm lugemi korrektselt ning teine eksib.



Joonis 4. **Erinevate joondusalgoritmide abil loodud joondustest tuvastatud oluliste artefaktsete SNVde jaotumine algoritmi-põhiste joonduste vahel.** Iga algoritmiga joondatud virtuaalsete raamatukogude komplekt on tähistatud ühe värviringiga. Tumedama värviga alades on mitme andmestiku ülekattekohad ehk vale-SNVde arv, mida tuvastati ühtselt mitme joondusalgoritmide poolt joondatud lugemitest.

Kokkuvõttes saab öelda, et kuigi artefaktselt määratakse vaid väikest osa kõikidest SNVdest, on annoteerimisel võimalik ekslikult selekteerida olulisena näivate variatsioonide sekka ka artefaktseid variatsioone.

2.4. Näited artefaktsete variatsioonide näilise panuse kohta

Illustreerimaks variatsioonide õige tuvastamise ja korrektse annotatsiooni olulisust, rõhutaksin kahte näitejuhtu – variatsioone rinnavähigeenina tuntud BRCA2 geenis ning 11-beeta-hüdrolaasi kodeerivas CYP11B geenis.

BRCA2 on DNA reparatsioonis osalev geen, mida ekspresseeritakse nii rinna- kui ka teistes kudedes. Tegemist on proto-onkogeeniga, kuna kindlate mutatsioonide olemasolul võib geen tõsta rinna- või munasarjavähi tekke ohtu [24] Tegemist on ühega geenidest, mida kasutatakse

personaalmeditsiinis või kommertsiaalsetes testides, et ennustada võimalikku kasvaja tekke riski [25].

Kliiniliselt potentsiaalselt olulistest ehk splaiss-saiti ja enneaegseid stopkoodonieid mõjutavatest SNVdest, mis esinesid dbSNP andmebaasis ning omasid VEPi kliinilise olulisuse hinnangul patogeense või potentsiaalselt patogeense mõju, tuvastati joondatud lugemitest kokkuvõttes üle 160 variatsiooni valesti või jäid tuvastamata. Täpsemalt tuvastati BWA-MEM poolt joondatud lugemitest sõltuvalt raamatukogu tüübist valesti 22 variatsiooni ning 21 variatsiooni jäid tuvastamata. BWA *align* algoritmil joondatud lugemitest jäi tuvastamata 161 unikaalset tõenäoliselt patogeenset variatsiooni. Bowtie2 *local* poolt joondatud lugemitest jäi määramata 9 ning sõltuvalt raamatukogu parameetritest oli raske määrata 16 variatsiooni. Bowtie2 *end-to-end* poolt joondatud lugemitest jäi määramata100 ning sõltuvalt raamatukogu parameetritest oli raske määrata 71 variatsiooni.

Kahel juhul jäi lugemitest tuvastamata suur hulk kliiniliselt olulisi variatsioone, mistõttu on oluline endale teadvustada, et lisaks võimalusele annoteerida oluliseks artefaktsed variatsioonid, võivad kliiniliselt olulist rolli omavad variatsioonid jääda tuvastamata.

CYP11b1 on geen, mis kodeerib 11-beeta hüdrolaasi. 11-beeta-hüdrolaasil on oluline roll neerupealistes, kus ta osaleb kortisooli ja kortikosterooni regulatsioonis. 11-beeta-hüdrolaasi puudusel tekib neerude hüperplaasia ning sellega on paeguseks seostatud üle 80 variatsiooni [26].

BWA-MEM joondustest tuvastatud artefaktsete variatsioonide analüüsil leiti 10 CYP11B1 geeni variatsiooni, mis nii PolyPhen-2 kui SIFTi skoori baasil võiksid olla patogeensed. Seejuures on neli võimalikku variatsiooni rohkem cDNAs järjestuse algusosas ning omavad SIFTi skooride väärtusi 0 (kus 0 on kõige kahjulikum ja 1 kõige tolereeritavam) ning PolyPhen-2 skooride väärtusi üle 0,9 (kus 1 on kõige kahjulikum ja 1 kõige tolereeritavam). Antud variatsioonid on toodud välja tabelis 7.

Antud geenivariatsioonid võiksid jääda annoteerimisel sõelale, kuna sisaldavad variatsioone evolutsiooniliselt kõrgelt konserveerunud järjestustes, mis võib viidata potentsiaalselt oluliste tagajärgedega variatsioonidele. Siinkohal tuleb rõhutada, et tegemist on valepositiivsete SNVdega, mida selle koha peal tegelikult uuritavas mudelis ei olnud. Peame endale teadvustama, et huvi pakkuvate variatsioonide filtreerimisel annoteeritute hulgast võivad sõelale jääda ka selliseid variatsioone, mis on näiliselt paljutähenduslikud, aga tegelikult olemas ei ole.

Tabel 7 Geeni CYP11B1 valitud artefaktsete variatsioonide kokkuvõte. Välja on toodud variatsiooni asukoht genoomis, referents- ja alternatiivne alleel, variatsiooni poolt põhjustatud muutus ja aminohappeline asendus, geeni identifitseerimisnumber ning variatsiooni asukoht cDNA järjestuses ning SIFTi ja PolyPhen-2 ennustusskooride väärtused.

Chr	Positsioon	Ref	Alt	Tagajärg	Geeni ID	Positsioon cDNAs	Aminohappeline muutus	SIFT	PolyPhen-2
8	143958480	G	A	NON_SYNONYMOUS_CODING	CYP11B1	587	T/I	deleterious(0)	probably_damaging(0.988)
8	143958493	G	A	NON_SYNONYMOUS_CODING	CYP11B1	574	R/W	deleterious(0)	probably_damaging(0.946)
8	143996558	C	T	NON_SYNONYMOUS_CODING	CYP11B2	502	D/N	deleterious(0)	probably_damaging(0.99)
8	143996658	A	T	NON_SYNONYMOUS_CODING	CYP11B2	402	N/K	deleterious(0)	probably_damaging(0.986)

Lisas 4 on toodud ülevaatlik tabel artefaktsete SNVde annotatsiooni kohta mõjutatud geenide kaupa. Tabel on loodud juhtimaks tähelepanu geenidele, millega seotud SNVde tuvastamisel võib tekkida, lähtuvalt lugemite joondamiseks kasutatud algoritmi,st artefaktseid variatsioone

2.5. Arutelu

On näidatud, et erinevad joondusalgoritmid, isegi kui nad kasutavad tööks samu üldpõhimõtteid, saavad joondamisel erinevaid tulemusi, ei ole uuemate joondusalgoritmide puhul täpselt kindlaks määratud, kui palju erinevad joondusalgoritmid eksivad.

Samuti on oluline küsimus, kas joondusalgoritmide vigadest tulenevad SNVde valesti tuvastamised võivad osutuda komistuskiviks hilisemale SNVde võimalike mõjude uurimisel. Kuna eksimused on paratamatud, on oluline kindlaks määrata, kas valesti tuvastatavad SNVd võivad annoteerimise ja tõlgendamise järel näiliselt omada olulist füsioloogilist mõju ning

Selleks on oluline kindlaks määrata, kas eksimused joondamistel leiavad aset nende genoomipositsioonide suhtes, mille funktsionaalsust ei osata hinnata ning mille annoteerimisel ei saaks olulisena näivat valeinformatsiooni või toimuvad vead selliste genoomipiirkondade osas, mis võivad viia ekslike järelduste tegemiseni vigase informatsiooni põhjal.

Käesoleva töö eksperimentaalne osa näitab, et erinevate joondusalgritmide poolt joondatud lugemitest tuvastatud artefaktsed SNVd ei lange annotatsiooni lisamise järel näiliselt mitte ainult intronite ning geenidevahelistesse piirkondadesse, vaid ligikaudu 20% artefaktselt tuvastatud SNVdest asuvad ka valke kodeerivates alades ning teistes potentsiaalselt olulistes genoomipiirkondades. Samuti näitab töö, et artefaktsed SNVd võivad annotatsiooni põhjal näiliselt põhjustada üliolulisi tagajärgi, nagu näiteks splaissimiskoha muutust või enneaegse stopkoodoni tekkimist. Sellised variatsioonid võivad jääda potentsiaalselt huvipakkuvatena sõelale teadusuuringutes ning osutuda potentsiaalseks komistuskiviks SNVde ja fenotüübiliste seoste uurimisel.

Samuti on oluline teada, et lisaks artefaktsetele variatsioonidele jäävad paljud dbSNP andmebaasis kirjeldatud variatsioonid NGS andmetest tuvastamata ehk saame sisuliselt valenegatiivseid tulemusi. See tähendab, et joondusalgoritmide eripärade või muude põhjuste tõttu jäävad meile „nähtamatuks“ variatsioonid, mis on genoomis tegelikult olemas, sealhulgas funktsionaalselt olulised variatsioonid. Nii teaduslikes uuringutes kui ka kliinilises praktikas oleks vajalik valenegatiivseid tulemusi minimaliseerida. Kui valepositiivsed ehk artefaktsed tulemused saavad välistatud valideerimise või korduvdiagnostika käigus, siis valenegatiivsed tulemused toovad kaasa selle, et oluline seos või haiguse põhjus jääb avastamata.

Kokkuvõttes on oluline teadvustada, kust tulevad võimalikud vead variatsioonide tuvastamisel ja kallutatus nende annoteerimisel ning leida parima lahenduse valepositiivsete ja –negatiivsete tulemuste vähendamiseks.

Käesoleval hetkel puudub täpne avalikult ligipääsetav ülevaade ebausaldusväärselt määratavatest SNVdest lähtuvalt joondusalgoritmist, millega joondatud lugemitest SNVd määrati. Antud töö on osa suuremast projektist, mille üks kaugele ulatuvatest eesmärkidest on luua andmebaas, mille abil oleks võimalik märgistada ebausaldusväärselt määratavad genoomsed positsioonid lähtuvalt NGS andmete joondamiseks kasutatud joondusalgoritmidest. Selline andmebaas aitaks vältida võimalikke probleeme, mis võivad tekkida valesti tuvastatud SNVde uurimisel.

KOKKUVÕTE

SNVde õige tuvastamine ja annoteerimine on väljakutse, mille käigus on oluline minimeerida võimalikke tekkivaid vigu. Minimaalne vigade arv variatsioonide analüüsil võimaldab saada täpsemaid tulemusi NGS andmete analüüsil nii teaduslikes uuringutes kui ka kliinilises praktikas..

Variatsioonide annoteerimisel on oluline lähtuda õigetest algandmetest. On näidatud, et sobiva andmebaasi valik mõjutab oluliselt saadavaid tulemusi, mille tõttu tuleb annoteerimisel kasutatavate andmebaaside puhul lähtuda võimalikult palju analüüsitavate proovide eripärast. Võimaluse korral tuleb annoteerimiseks kasutada transkriptide komplekti, mis ekspresseerub just uuritavas koes. Samuti peab hindama, kas annotatsiooniandmete saamiseks võib, lähtuvalt uuringu spetsiifikast, kasutada automaatselt koostatud annotatsioonidega andmebaasi, näiteks Ensembli andmebaasi, või vajab uuring kureeritud andmebaasist, näiteks GENCODEist, pärit andmestikku.

Samuti on oluline tähele panna, et erinevad annoteerimistööriistad kasutavad annoteerimiseks erinevaid lähenemisi. Enimkasutatud annoteerimistööriistad, nagu näiteks VEP ja SnpEff, lähtuvad annoteerimisel juba teadaolevast informatsioonist variatsioonidega seotud geenide, transkriptide ja geenielementide funktsioonide kohta, millele lisatakse bioloogilistel reeglitel põhineva varianti tagajärje ennustamise. Sellistesse annotatsioonitööriistadesse on reeglina võimalik integreerida mitmeid lisainformatsiooniallikaid, näiteks informatsiooni variatsioonide esinemise ja mõju kohta erinevates kliinilise suunitlusega andmebaasides. See teeb annoteerimise täpsemaks ning aitab filtreerida välja huvipakkuvaid variatsioone.

Alternatiivina on loodud mitmeid masinõppel või arvutuslikel meetoditel põhinevaid annoterimistööriistu. Meetodid, mis põhinevad evolutsioonilise konserveeruvuse hindamisel või masinõppe abil algoritmide otsimisel variatsioonidega treeningmudelite abil, annavad hinnangu, kui tõenäoline on, et annoteeritav variatsioon võib olla kahjulik. Kuigi sellised meetodid on ennustustes piisavalt täpsed, peab selliste tööriistade annotatsiooni tõlgendamisel olema ettevaatlik. Esiteks, selliste tööriistade otsustus põhineb skoori arvutamisel ning võrdlemisel eelseadistatud lävega. Kuna lävendid on tavaliselt kompromiss algoritmi tundlikkuse ja spetsiifilisuse vahel, tuleb skoori tõlgendamisel olla ettevaatlik. Samuti vajavad arvususlikel meetoditel huvipakkuvateks valitud SNVd kindlasti valideerimist, kuna arvutuslikud algoritmid võivad eksida, näiteks kasutades ebapiisaval hulgal või evolutsiooniliselt liiga kaugete geenijärjestusi.

Töö eksperimentaalses osas näidati, et NGS andmete analüüsiprotsessis kasutatavad erinevad joondusalgoritmid võivad oluliselt mõjutada SNVde tuvastamist. Kasutades annoteerimistööriista VEP näidati, et osa artefaktselt tuvastatud SNVdest võivad annoteerimise tulemuste tõlgendamisel omada näiliselt olulisi funktsionaalseid mõjusid. Näiliselt olulise mõjuga artefaktsed SNVd võivad osutuda komistuskiviks teaduslikes uuringutes või kliinilises praktikas, mille tõttu on oluline teadvustada taoliste vigade tekkimise võimalust.

Kokkuvõtteks saab öelda, et variatsioonide tuvastamine ja annoteerimine on veel uus ning väljakutseid täis ala, mis peab tegelema nii teadmistepõhiste anotatsioonivahendite täpsemaks ja informatiivsemaks muutmiseks kui ka samal ajal arvesse võtma juba praegu teadaolevaid vigu variatsioonide tuvastamisel ning annoteerimisel.

SUMMARY

Correct SNV calling and annotation is a multi-step procedure, where it is important to minimise possible errors that can lead to misjudgement in further analysis of variation-phenotype associations.

In variant annotation process, the applied methods and databases play an essential part. It is highly recommendable to use datasets, which describe the source of sequenced genomes as accurately as possible. It is possible to choose transcripts for specific tissue samples. Whole-exome and whole-genome sequences have separate transcript sets and there are several datasets designed for disease-related (for example ANNOVAR) or novel variant annotations. It is also shown that annotation based on high-quality transcripts can lead to different results compared to annotation based on wider, but computionally annotated transcript set.

It is also important to notice that different annotation tools use different approaches to annotation. While most commonly used annotation tools, like VEP or SnpEff, encorporate known variant information, like position in gene and transcripts, with rule-based (impact of variation on transcription or translation process) annotation, there are several different approaches to annotation. Supervised machine learning or conservation-based approaches will try to predict, how tolerated is variation based on comparison to related sequences and give the prediction as a score value. While analysing information obtained from annotation, it is important to emphasize the source of annotation. For example the effect predictions from PolyPhen-2 (supervised computer learning algorithm) needs additional validation because it only outputs theoretically computed predictions of possible importance of the variants.

The experimental part of the thesis showed that different alignment algorithms could play an important role in outcomes of SNV calling. Due to misaligned reads, SNV callers can call artefact SNVs and miss existent SNVs in the aligned sequence. Using Variant Effect Predictor tool for annotation of artefact SNVs, the experiments showed, that miscalled SNVs could be seen important physiologically. The research concludes that it is important to acknowledge possible mistakes coming from alignment procedure and with that avoid analysing possibly miscalled SNVs.

KIRJANDUSE LOETELU

Adams, M.D., Veigl, M.L., Wang, Z., Molyneux, N., Sun, S., Guda, K., Yu, X., Markowitz, S.D. and Willis, J. (2012). Global mutational profiling of formalin-fixed human colon cancers from a pathology archive. Mod. Pathol. 25: 1599–1608.

Adzhubei, I. a, Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods 7: 248–249.

Aggarwal, C.C. (2015). Data Classification: Algorithms and Applications (CRC Press).

Aubourg, S. and Rouzé, P. (2001). Genome annotation. Plant Physiol. Biochem. 39: 181–193.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L. and Abel, L. (2015). Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl. Acad. Sci. U. S. A. 112: 5473–5478.

Burrows, M. and Wheeler, D.J. (1995). A Block-Sorting Lossless Data Compression Algorithm. 1.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. and Chan, A.P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 7.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 6: 80–92.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38: 1767–1771.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15: 901–913.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., *et al.* (2011). The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–498.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 6: R44.

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. 390.

Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014). Adaptable probabilistic mapping of short reads using position specific scoring matrices. BMC Bioinformatics 15: 100.

Kumar, P., Henikoff, S. and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4: 1073–1081.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25: 1966–1967.

Manber, U. and Myers, G. (1990). Suffix arrays: a new method for on-line string searches. 319–327.

McCarthy, D.J., Humburg, P., Kanapin, A., Rivas, M.A., Gaulton, K., Cazier, J.-B. and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. Genome Med. 6: 26.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069–2070.

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

Mudge, J.M., Frankish, A. and Harrow, J. (2013). Functional transcriptomics in the post-ENCODE era. Genome Res. 23: 1961–1973.

Ng, P.C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 31: 3812–3814.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–451.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., *et al.* (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5: 28.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J. and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. Brief. Bioinform. 15: 256–278.

Ramensky, V., Bork, P. and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 30: 3894–3900.

Reinert, K., Langmead, B., Weese, D. and Evers, D.J. (2015). Alignment of Next-Generation Sequencing Reads. Annu. Rev. Genomics Hum. Genet. 16: 133–151.

Ritchie, G.R. and Flicek, P. (2014). Computational approaches to interpreting genomic sequence variation. Genome Med. 6: 87.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., *et al.* (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036–1040.

Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E.A., Liu, Y., Weinstock, G.M., Wheeler, D.A., Gibbs, R.A., *et al.* (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 20: 273–280.

Shihab, H. a., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L. a, Edwards, K.J., Day, I.N.M. and Gaunt, T.R. (2013). Predicting the Functional, Molecular, and Phenotypic

Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum. Mutat. 34: 57–65.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res. 40: W452–W457.

Trapnell, C. and Salzberg, S.L. (2009). How to map billions of short reads onto genomes. Nat. Biotechnol. 27: 455–457.

Wang, K., Li, M. and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38: e164.

Yu, X. and Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14: 274.

KASUTATUD VEEBIAADRESSID

[1] EMBL-EBI Train online: What is Next Generation DNA Sequencing. https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna- Kasutatud 02.05.2016.

[2] Illumina: Sequencing Systems. http://www.illumina.com/systems/sequencing.html Kasutatud 02.05.2016.

[3] Illumina: Next-generetion sequencing. http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html Kasutatud 02.05.2016.

[4] Illumina: Sequencing coverage. http://www.illumina.com/science/education/sequencing-coverage.html Kasutatud 02.05.2016.

[5] Rererence Genome Consortium. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human Kasutatud 03.05.2016.

[6] NCBI Genome Remapping Service. http://www.ncbi.nlm.nih.gov/genome/tools/remap Kasutatud 03.05.2016.

[7] Burrows-Wheeler Aligner. http://bio-bwa.sourceforge.net/ Kasutatud 20.04.2016.

[8] GATK Main Page https://www.broadinstitute.org/gatk . Kasutatud 01.05.2016.

[9] Bowtie2: Manual http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml Kasutatud 01.05.2016:

[10] TopHat. https://ccb.jhu.edu/software/tophat/index.shtml Kasutatud 18.05.2016.

[11] Sequence Alignment/Map Format Specification. https://samtools.github.io/hts-specs/SAMv1.pdf Kasutatud 15.04.2016.

[12] Ensembl Variant Effect Predictor. http://www.ensembl.org/info/docs/tools/vep/index.html Kasutatud 13.05.2016.

[13] ANNOVAR Documentation. http://annovar.openbioinformatics.org/en/latest/ Kasutatud 20.05.2016.

[14] SnpEff and SnpSift. http://snpeff.sourceforge.net/SnpEff.html Kasutatud 13.05.2016.

[15] SIF Help. Available: http://sift.bii.a-star.edu.sg/www/SIFT_help.html Kasutatud 12.05.2016

[16] SIFT4g. http://sift.bii.a-star.edu.sg/sift4g/ Kasutatud 28.05.2016.

[17] PROVEAN Home. Available: http://provean.jcvi.org/about.php Kasutatud 28.04.2016.

[18] Protein BLAST: search protein databases using a protein query. http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins Kasutatud 19.04.2015.

[19] CD-HIT Official Website. Kasutatud 19.04.2016. http://weizhongli-lab.org/cd-hit/.

[20] fathmm - Home. Kasutatud 07.05.2016. http://fathmm.biocompute.org.uk/index.html.

[21] FATHMM parseVCF.py. http://fathmm.biocompute.org.uk/parseVCF.py Kasutatud 07.05.2016

[22] PolyPhen-2 prediction of functional effects of human nsSNPs. http://genetics.bwh.harvard.edu/pph2/index.shtml Kasutatud 15.05.2015.

[23] PolyPhen-2 Wiki. http://genetics.bwh.harvard.edu/pph2/dokuwiki/overview Kasutatud 14.05.2016.

[24] OMIM Entry - *600185 BRCA2 GENE; BRCA2. http://www.omim.org/entry/ 600185 Kasutatud 15.05.2015.

[25] Geenome | How Personalized Medicine is Changing Breast Cancer http://genomemag.com/how-personalized-medicine-is-changing-breast-cancer/#.Vz68EL6omXc

[26] CYP11B1 - Genetics Home Reference. https://ghr.nlm.nih.gov/gene/CYP11B1 Kasutatud 11.05.2016.

[27] VEP script. Available: http://www.ensembl.org/info/docs/tools/vep/script/index.html Kasutatud 10.05.2016.

[28] Running the script. http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html Kasutatud 16.05.2016.

TÄNUAVALDUSED

Soovin tänada eelkõige oma juhendajat, teadur Ulvi Gerts Talast, kes oli nii nõu kui jõuga toeks töö valmimisel. Samuti tänan sellise huvitava ning arendava teemavaliku eest.

Soovin tänada Eliisa Univeri ja Kaja Karlsonit keelekorrektuuride ning õigeaegsete motivatsioonisõnade eest.

LISAD

LISA 1 Reeglite-põhise annotatsiooni üldised reeglid ja võimalikud „tagajärgede“ definitsioonid populaarsemate annoteerimistööriistade baasil.

Tabel 8 Reeglite-põhise annotatsiooni üldised reeglid ja tagajärgede“ definitsioonid. Tabeli tulpades on toodud enimkasutatud annotatsioonitööriistade terminite kasutus vastavalt variatsiooni mõju ennustamise reeglile või mõju tähendusele. „Ei määra“ näitab, et vastava definitsiooniga efekti antud tööriist ei määra. Samas on oluline tähele panna, et paljud efektid on sarnaste definitsioonidega ning võivad teatud variatsioonide efektide puhul olla käsitletavad sünonüümsetena.

** SO - VEPi poolt kasutatav Sequence Onthology andmebaas*

Reegel/tähendus	SnpEff (Cingolani et al. 2012)	VEP/SO* [27]	ANNOVAR [13]
Elemendi eemaldus, kui eemaldatud piirkond sisaldab transkripti	*ei määra*	transcript_ablation	*ei määra*
Splaissimisvariant, kus muutuvad 2 nukleotiidi introni 3' otsas	SPLICE_SITE_ACCEPTOR	splice_acceptor_variant	*ei määra*
Splaissimisvariant, kus muutuvad 2 nukleotiidi introni 5' otsas	SPLICE_SITE_DONOR	splice_donor_variant	*ei määra*
Järjestuse variant, mille puhul muutub üks koodonluues varajase stopkoodoni ja lühema transkripti	STOP_GAINED	stop_gained	Stopgain
Järjestuse variant, mis tekitab häireid translatsiooni lugemisraamis, kuna insertsiooni või deletsiooni pikkus ei ole kolme kordne	FRAME_SHIFT	frameshift_variant	frameshift block substitution
Järjestuse variant, kus vähemalt üks terminaatorikoodon on muudetud ja tulemiks on pikenenud transkript	STOP_LOST	stop_lost	Stoploss
Reegel/tähendus	SnpEff	VEP/SO*	ANNOVAR

Vähemalt ühe kanoonilise startkoodoni nukleotiidi muutus	START_LOST	start_lost	*ei määra*
Transkripti sisaldava elemendi aplifikatsioon	*ei määra*	transcript_amplification	*ei määra*
Raami-sisene mitte-sünonüümne insertsioon kodeerivasse järjestusse	CODON_INSERTION	inframe_insertion	nonframeshift insertion
Raami-sisene mitte-sünonüümne deletsioon kodeerivas järjestusse	*ei määra*	inframe_deletion	nonframeshift deletion
Järjestuse variant, kus muutub üks või mitu nukleotiidi, millega kaasneb aminohappe muutus, kuid järjestuse kogupikkus säilib	*ei määra*	missense_variant	nonsynonymous SNV
Järjestuse variant, kus muudab järjestuse poolt kodeeritavat valku	*ei määra*	protein_altering_variant	*ei määra*
Splaissimiskoha variant, kus muutus toimus 1…3 nukleotiidis eksonis või 3…8 nukleotiidis intronis	*ei määra*	splice_region_variant	splicing
Järjestuse variant, kus vähemalt üks lõpukoodoni nukleotiidest osaliselt annoteeritud transkriptis muutub	*ei määra*	incomplete_terminal_codon_variant	*ei määra*
Järjestuse variant, kus vähemalt üks terminaatori koodonitest muutub, kuid terminaator säilib	SYNONYMOUS_STOP	stop_retained_variant	*ei määra*
Järjestuse variant, kus ei toimu muutust kodeeritud aminohapetes	SYNONYMOUS_CODING	synonymous_variant	synonymous SNV
Järjestuse variant, mis muudab kodeeritavad järjestust	CDS	coding_sequence_variant	*ei määra*

Transkripti vairant, mis asub mature miRNA järjestuses	*ei määra*	mature_miRNA_variant	*ei määra*
Reegel/tähendus	SnpEff	VEP/SO*	ANNOVAR
Mittetransleeritava järjestuse variant, mis asub 5'-otsas	UTR_5_PRIME	5_prime_UTR_variant	UTR5
Mittetransleeritava järjestuse variant, mis asub 3'-otsas	UTR_3_PRIME	3_prime_UTR_variant	UTR3
Järjestuse variant, mis muudab mittekodeeriva eksoni järjestuse mittekodeerivaks transkriptiks	*ei määra*	non_coding_transcript_exon_variant	*ei määra*
Transkripti variant introni sees	INTRON	intron_variant	intronic
Variant transkriptis, mis on *nonsense-mediated decay* sihtmärgiks	*ei määra*	NMD_transcript_variant	*ei määra*
Mitte-kodeeriva RNA geeni järjestuse variant	*ei määra*	non_coding_transcript_variant	ncRNA
Geeni 5'-osas asuv variant	UPSTREAM	upstream_gene_variant	upstream
Geeni 3'-osas asuv variant	DOWNSTREAM	downstream_gene_variant	downstream
Elemendi kadumine, kus deleteeritud piirkond sisaldas transkriptsiooni sidumise piirkonda	*ei määra*	TFBS_ablation	*ei määra*
Piirkonna kordus, kus asub transkriptsioonifaktorite sidumisala	*ei määra*	TFBS_amplification	*ei määra*
Järjestuse variant, mis asub transktriptsioonifaktori sidumise alas	*ei määra*	TF_binding_site_variant	*ei määra*

Elemendi kadumine, kus piirkonnas oli regulatoorne ala	*ei määra*	regulatory_region_ablation	*ei määra*
Elemendi piirkonnda kordus, mis sisaldab regulatoorset piirkonda	*ei määra*	regulatory_region_amplification	*ei määra*
Reegel/tähendus	SnpEff	VEP/SO*	ANNOVAR
Järjestuse variant, mis asub regulatoorses alas	*ei määra*	feature_elongation	*ei määra*
Järjestuse variant, mis asub regulatoorses alas	*ei määra*	regulatory_region_variant	*ei määra*
Järjestuse variant, mis toob kaasa genoomse elemendi lühenemise võrreldes referentsjärjestusega	*ei määra*	feature_truncation	*ei määra*
Järjestuse variant, mis asub geenidevahelises piirkonnas	INTERGENIC	intergenic_variant	intergenic
Variant deleteerib eksoni, mis on transkripti 5'-UTR piirkond	UTR_5_DELETED	*ei määra*	*ei määra*
5'-UTR piirkonna variant, milles on kolmenukleotiidiline järjestus, mis võib olla startkoodon	START_GAINED	*ei määra*	*ei määra*
Variatsioon muudab ühe startkoodoni teiseks startkoodoniks	SYNONYMOUS_START	*ei määra*	*ei määra*
Variatsioon asub geenis	GENE	*ei määra*	*ei määra*
Variant asub transkritptis	TRANSCRIPT	*ei määra*	*ei määra*
Variatsioon asub eksonis	EXON	*ei määra*	*ei määra*

Deletsioon eemaldab terve eksoni	EXON_DELETED	*ei määra*	*ei määra*
Ühe või mitu koodonit muudetakse	CODON_CHANGE	*ei määra*	*ei määra*
Variant on kõrgelt konserveerunud introni piirkonnas	INTRON_CONSERVED	*ei määra*	*ei määra*
Reegel/tähendus	SnpEff	VEP/SO*	ANNOVAR
Variant on kõrgelt konserveerunud geenidevahelises piirkonnas	INTERGENIC_CONSERVED	*ei määra*	*ei määra*
Koodoni muutus ja koodoni deletsioon	CODON_CHANGE_PLUS_CODON_DELETION	*ei määra*	*ei määra*
Variant asub eksonis	*ei määra*	*ei määra*	exonic
Järjestuse variant, mis tekitab häireid translatsiooni lugemisraamis, kuna insertsiooni pikkus ei ole kolme-kordne	*ei määra*	*ei määra*	frameshift insertion
Raaminihet mitte põhjustav järjestuse variant	*ei määra*	*ei määra*	nonframeshift block substitution
Teadmata funktsiooniga järjestuse variant	*ei määra*	*ei määra*	unknown

LISA 2 Annoteerimistööriista VEPi käsureapõhine kasutamine: käsurida ning selle võrdlus veebiversiooniga.

Annoteerimisel kasutatud käsurida on järgnev:

zcat /SISENDFAIL/ | perl /usr/local/ensembl-tools-release-81/scripts/variant_effect_predictor/variant_effect_predictor.pl --pick --merged --assembly GRCh37 --vcf --html --sift b --polyphen b --regulatory --total_length --terms ensembl --canonical --biotype --maf_1kg --maf_esp --fork 8 --offline --force_overwrite --cache --dir_cache /usr/local/ensembl-tools-release-81/scripts/variant_effect_predictor/cache_database/ -o /VÄLJUND/ > log.out &

Tabel 9. Annoteerimisel kasutatud käsud, nende tähendus ning veebitööriista analoogid

Käsurea tähis	Veebiversiooni analoog [12]	Tähendus [28]
--pick	Restrict results: Show one selected consequence	Valib ühe tagajärje variatsiooni kohta alustades kõige kahjulikumast.
--merged	Transcript database to use: Ensembl and RefSeq transcripts	Kasutatakse ühendatud Ensembl'i ja RefSeqi transkriptide andmebaasi.
--assembly GRCh37	Vastava referentsgenoomi-põhine tööriist on saadaval aadressil http://grch37.ensembl.org/Homo_sapiens/Tools/VEP	Referetnsiks kasutatakse inimese referentsgenoomi versiooni 37.
--html	Puudub	Genereeritakse väljundile lisaks html-fail, mis sisaldab koondtulemusi ning hüperlinke Ensembli ja teistesse andmebaasidesse
--sift b	SIFT: Prediction and score	Lisatakse SIFTi ennustusskoor ning tõlgendus

Käsurea tähis	Veebiversiooni analoog	Tähendus
--polyphen b	PolyPhen: Prediction and score	Lisatakse PolyPhen-2 ennustusskoor ning tõlgendus, kasutati HumVar treeningmudelit
--regulatory	Get regulatory region consequences:	Katvuste otsimine regulatoorsete piirkondadega
--total_length	*Vaikimisi*	Antakse variatsiooni positsiooni cDNAs, CDSis ja valgus formaadis positsiooni/kogupikkus.
--terms ensembl	Puudub	Tagajärgede kirjeldamise terminoloogia valimine
--canonical	*Identify canonical transcripts*	Lisab tähise, et transkript on uuritava geeni kanooniline transkript.
--biotype	*Transcript biotype*	Lisatakse transkripti biotüüp
--maf_1kg	*Vaikimisi aktiivne*	Lisatakse mandipopulatsioonides (Aafrika, Ameerika, Aasia, Euroopa) esinevad alleelisagedused lähtuvalt 1000 Genoomi projektist.
--maf_esp	*Vaikimisi aktiivne*	Lisatakse NHLBL-ESP populatsiooni esinevad alleelisagedused lähtuvalt 1000 Genoomi projektist.
--fork 8	Puudub	Lubab jaotada tööprotsessi mitme tuuma vahel. Kasutati protsessi jaotamist kaheksaks.
--offline	Puudub	Lubab töötada võrguühenduseta kasutades andmebaaside (varu) koopiaid
--cache	Puudub	*Lubab cache kasutamist.*
--dir_cache	Puudub	Täpssutab *cache* asukohta.

LISA 3 Annoteerimistööriistade poolt kasutatavad andmebaasid

Tabel 10. Annoteerimistööriistade poolt kasutatavad andmebaasid. Tabelis on toodud annoteerimistööriistade poolt kasutatavate andmeaaside nimetused, andmebaaside asukohad, haldajad või arendajad ning informatsioon, mida tööriistad andmebaasidest võtavad.

Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
Transkriptid				
Ensembl (Core)		http://www.ensembl.org/info/docs/api/core/index.html	European Bioinformatics Institute (EBI), Wellcome Trust Sanger Institute (WTSI)	Automaatselt annoteeritud transkriptid
GENCODE	GENCODE Project: Encyclopædia of genes and gene variants	http://www.gencodegenes.org/	National Human Genome Research Institute (NHGRI), WTSI ja tesed	Kontrollitud valke kodeerivate lookuste annotatsioon
RefSeq	NCBI Reference Sequence Database	http://www.ncbi.nlm.nih.gov/refseq/	The National Center for Biotechnology Information (NCBI)	Annoteeritud genoomse DNA, transkirptide ja valkude järjestused, sh käsitsi kureeritud jäjrestused

Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
INSDC	International Nucleotide Sequence Database Collaboration	http://www.insdc.org/	DNA Data Bank of Japan (DDBJ), NCBI, EBI	NGS toorandmed, annotatsioonid, proovide ja katsete informatsioon
HAVANA	Human and Vertebrate Analysis and Annotation Project		WTSI	Käsitsi kureeritud transkriptid
UCSC Known Genes	University of California Santa Cruz Known Genes Dataset	http://genome.ucsc.edu	University of California Santa Cruz (UCSC)	Tuntud valke kodeerivad geenid, automaatne uuendus
VEGA	Vertebrate and Genome Annotation	http://vega.sanger.ac.uk/info/about/vega_proj.html	WTSI	Kõrge kvaliteediga geenimudelid, põhinevad genoomide käsitsi annoteerimisel
Valkude järjestus, struktuur				
neXtProt		http://www.nextprot.org/	SIB, Geneva Bioinformatics SA (GeneBio)	Kõrge kvaliteediga informatsioon valkude funktsiooni, asukoha, ekspressiooni ja interaktsioonide kohta

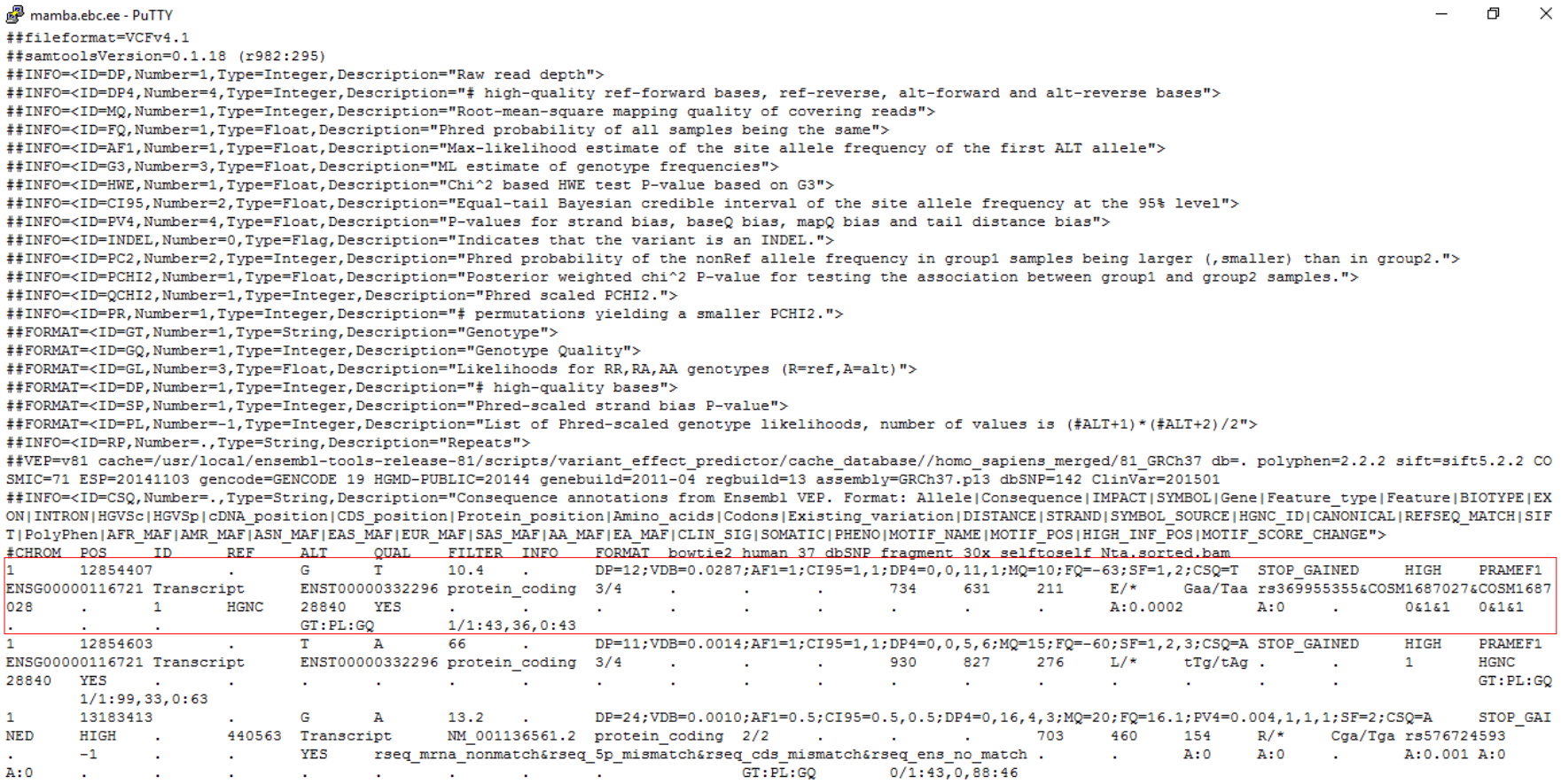
Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
Uniprot Swiss-Prot	Universal Protein resource Swiss-Prot	http://www.uniprot.org/	EBI, Swiss Institute of Bioinformatics (SIB), the Protein Information Resource (PIR)	Käsitsi annoteeritud valkude järjestused ja funktsioonid
Uniprot TrEMBL	Uniprot Translated EMBL Nucleotide Sequence Data Library			Automaatselt annoteeritud valkude järjestused ja funktsioonid
Uniref90	The UniProt Reference Clusters 90	http://www.uniprot.org/uniref/		UniProt Knowledgebase'i ja valitud UniParci järjestuste klastrid
SWALL	Non-Redundant Protein Sequence Database incorporating SWISSPROT, TREMBL, TREMBLNEW			
PDB	Protein Data Bank	http://www.rcsb.org/pdb/home/home.do	Research Collaboratory for Structural Bioinformatics kaks liiget Rutgers University ja University of California San Diego	Kristallograafiliselt, krüoelektronmikroskoopiaga ja tuumamagnetresonantsiga kinnitatud struktuuridega valgud

Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
DSSP	Define Secondary Structure of Proteins	http://swift.cmbi.ru.nl/gv/dssp/	Centre for Molecular and Biomolecular Informatics (CMBI), Radboudi Ülikool	Valkude sekundaarstruktuurid
NCBI NR	NCBI non-redundant proteiin database	ftp://ftp.ncbi.nih.gov/blast/db/	NCBI	Valkude järjestused
Nomenklatuur				
SO	Sequence Ontology	http://www.sequenceontology.org/index.html	Gene Ontology Consortium, EBI, WTSI	Järjestuste elementide terminoloogia
HGNC	HUGO Gene Nomenclature Committee	http://www.genenames.org/	Human Genome Organisation (HUGO), NHGRI, WTSI	Inimese geenide nimetused
HGVS	Sequence Variant Nomenclature	http://varnomen.hgvs.org/	Human Genome Variation Society (HGVS), Human Variome Project (HVP), HUGO	Variatsioonide nomenklatuur
Geenide piirkonnad, funktsioonid, variatsioonid, epigenoomika				
CCDS	Consensus CDS project	https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi	NCBI	Valke kodeerivad piikonnad ja nende kõrgkvaliteetne annotatsioon

Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
dbSNP	Single Nucleotide Polymorphism Database	http://www.ncbi.nlm.nih.gov/SNP/	NCBI	SNVde, indelite ja teiste variatsioonide arhiiv
dbNSFP	Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations	https://sites.google.com/site/jpopgen/dbNSFP	Xiaoming Liu, Ph.D ja University of Texas Healt Science Centre	SNVde annotatsioon, sealhulgas mõju ennustajate skoorid
Epigenome Roadmap	NIH Roadmap Epigenomics Mapping Consortium	http://www.roadmapepigenomics.org/	National Institute of Health (NIH)	Informatsioon DNA histoonide modifikatsiooni ja teiste epigeneetiliste nähtuste kohta
NHLBI ESP	NHLBI GO Exome Sequencing Project	http://evs.gs.washington.edu/EVS/	National Heart, Lung, and Blood Institute (NHLBI) ja teised	Südame-veresoonkonna, kopsude ja vereloomeelundkonnaga seotud geenide eksoomide annotatsioon
1000 Genomes Project		http://www.1000genomes.org/	EBI	Inimeste geneetilise varieeruvuse informatsioon

Andmebaas	Andmebaasi täisnimi	Viide	Haldaja(d), arendajad	Sisaldus
ExAC	Exome Aggregation Consortium	http://exac.broadinstitute.org/	Broad Institute	Eksoomide sekveneerimise andmed
HRC	Haplotype Reference Consortium	http://www.haplotype-reference-consortium.org/	Prof Jonathan Marchini (Oxfordi Ülikool), prof. Goncalo Abecasis (Michigani Ülikool), prof. Richard Durbin (WTSI)	Inimese haplotüübid
BLUEPRINT	BLUEPRINT Consortium	http://www.blueprint-epigenome.eu/	Radboudi Ülikool ja arvukalt väikeettevõtteid	Referents-epigenoomid
Haigus-seoselised andmebaasid				
COSMIC	Catalogue of somatic mutations in cancer	http://cancer.sanger.ac.uk/cosmic/	WTSI	Kasvajates leiduvad mutatsioonid, sh käsitsi kureeritud andmed
ClinVar		http://www.ncbi.nlm.nih.gov/clinvar/	NCBI	Genoomsete variatsioonide seos terviseinformatsiooniga

LISA 3 – VEPi väljundfaili näidis



```
mamba.ebc.ee - PuTTY
##fileformat=VCFv4.1
##samtoolsVersion=0.1.18 (r982:295)
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the site allele frequency of the first ALT allele">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=CI95,Number=2,Type=Float,Description="Equal-tail Bayesian credible interval of the site allele frequency at the 95% level">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples being larger (,smaller) than in group2.">
##INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between group1 and group2 samples.">
##INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
##INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=-1,Type=Integer,Description="List of Phred-scaled genotype likelihoods, number of values is (#ALT+1)*(#ALT+2)/2">
##INFO=<ID=RP,Number=.,Type=String,Description="Repeats">
##VEP=v81 cache=/usr/local/ensembl-tools-release-81/scripts/variant_effect_predictor/cache_database//homo_sapiens_merged/81_GRCh37 db=. polyphen=2.2.2 sift=sift5.2.2 CO
SMIC=71 ESP=20141103 gencode=GENCODE 19 HGMD-PUBLIC=20144 genebuild=2011-04 regbuild=13 assembly=GRCh37.p13 dbSNP=142 ClinVar=201501
##INFO=<ID=CSQ,Number=.,Type=String,Description="Consequence annotations from Ensembl VEP. Format: Allele|Consequence|IMPACT|SYMBOL|Gene|Feature_type|Feature|BIOTYPE|EX
ON|INTRON|HGVSc|HGVSp|cDNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|DISTANCE|STRAND|SYMBOL_SOURCE|HGNC_ID|CANONICAL|REFSEQ_MATCH|SIF
T|PolyPhen|AFR_MAF|AMR_MAF|ASN_MAF|EAS_MAF|EUR_MAF|SAS_MAF|AA_MAF|EA_MAF|CLIN_SIG|SOMATIC|PHENO|MOTIF_NAME|MOTIF_POS|HIGH_INF_POS|MOTIF_SCORE_CHANGE">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  bowtie2_human_37_dbSNP_fragment_30x_selftoself_Nta.sorted.bam
1       12854407        .       G       T       10.4    .       DP=12;VDB=0.0287;AF1=1;CI95=1,1;DP4=0,0,11,1;MQ=10;FQ=-63;SF=1,2;CSQ=T  STOP_GAINED     HIGH    PRAMEF1
ENSG00000116721 Transcript      ENST00000332296 protein_coding  3/4     .       .       .       734     631     211     E/*     Gaa/Taa rs369955355&COSM1687027&COSM1687
028     .       1       HGNC    28840   YES     .       .       .       .       .       .       A:0.0002        A:0     .       0&1&1   0&1&1
.       .       .               GT:PL:GQ        1/1:43,36,0:43
1       12854603        .       T       A       66      .       DP=11;VDB=0.0014;AF1=1;CI95=1,1;DP4=0,0,5,6;MQ=15;FQ=-60;SF=1,2,3;CSQ=A STOP_GAINED     HIGH    PRAMEF1
ENSG00000116721 Transcript      ENST00000332296 protein_coding  3/4     .       .       .       930     827     276     L/*     tTg/tAg .       .       1       HGNC
28840   YES     .       .       .       .       .       .       .       .       .       .       .       .       .       .       .               GT:PL:GQ
        1/1:99,33,0:63
1       13183413        .       G       A       13.2    .       DP=24;VDB=0.0010;AF1=0.5;CI95=0.5,0.5;DP4=0,16,4,3;MQ=20;FQ=16.1;PV4=0.004,1,1,1;SF=2;CSQ=A     STOP_GAI
NED     HIGH    .       440563  Transcript      NM_001136561.2  protein_coding  2/2     .       .       .       703     460     154     R/*     Cga/Tga rs576724593
.       -1      .       .       YES     rseq_mrna_nonmatch&rseq_5p_mismatch&rseq_cds_mismatch&rseq_ens_no_match .       .       A:0     A:0     .       A:0.001 A:0
A:0     .       .       .       .       .       .       .       .               GT:PL:GQ        0/1:43,0,88:46
```

Üks mähitud rida.

Joonis 5. VEPi väljundfaili näidis. Fail algab päisega ning järgneb iga variatsiooni kirjeldusega ühe rea kohta. Paremaks lugemiseks on antud näidisfaili tabuleeritud. Esimene variatsioon on PRAMEF1 geenis asuv võib põhjustada stopkoodoni teket. Välja on toodud selle variatsiooni ID dbSNP ja COSMIC andmebaasides. Variatsioon asub kolmes selle geeni transkriptis neljast ning paikneb valku kodeerivad piirkonnas.

LISA 4 – Võimalikku olulist mõju omavate artefaktsete variatsioonide kokkuvõte geenide tasandil.

Tabel 11. Võimalikku olulist mõju omavate artefaktsete variatsioonide kokkuvõte geenide tasandil. Tabeli mugavamaks lugemiseks on toodud välja geenid, mille kohta oli artefaktsete SNVde seas kas splaissingut või stop- ja startkoodoneid mõjutavad variatsioonid. Välja on jäetud ainult mitte-sünonüümseid asendusi põhjustavad variatsioonid sõltumata nende olulisusest. Täismahus tabel on saadaval elektrooniliselt töö autorilt. Tabeli koostamisel on kasutatud VEPi terminiloogiat, kus SYMBOL tähistab geeni sümbolit, CHR, START ja STOP tähistavad geeni asukohta genoomis. „Tagajärgede" lühendid on järgnevad: ESS (essential slice site) - splaissimiskoha variatsioon, NSC (non-synonomous coding) – mittesünonüümne variatsioon, SG (stop-gained) – stopkoodoni lisandumine, SGjaSS (stop-gained&splice site) – stopkoodonit tekitav ja splaissimiskohas olev variatsioon, SL (stop-lost) – stopkoodoni kadumine.

BWA align joondamisalgoritmi poolt joondatud lugemitest saadud artefaktsete SNVde ülevaade olulisuse ja geenide järgi								
GEEN	CHR	START	STOP	ESS	NSC	SG	SGjaSS	
ACSM2A	16	20462897	20498991	0	12	1	0	
ACSM2B	16	20547547	20587689	0	12	5	0	
ADAM30	1	120436155	120439118	0	1	1	0	
AKR1B15	7	134233887	134264627	0	3	1	0	
AKR1C1	10	5005444	5025475	1	12	0	0	
AKR1C2	10	5029966	5046215	1	6	2	0	
ALG1L	3	125648117	125655882	0	0	1	0	
AMY1B	1	104230039	104238912	1	1	0	0	
AMY2B	1	104097321	104122151	0	11	1	0	
ANKRD30A	10	37414784	37521495	1	0	0	0	
ANKRD30B	18	14748238	14852479	1	7	2	0	
ANTXRL	10	47658233	47701443	1	0	0	0	
ANXA8L1	10	47157988	47174040	1	0	0	0	

APOBEC3F	22	39436608	39449915	0	2	1	0	
ARHGAP21	10	24872537	25012597	0	5	2	0	
ARHGAP5	14	32546494	32628934	0	3	1	0	
BCLAF1	6	136578000	136610989	0	13	1	0	
BTN3A2	6	26365458	26378546	1	0	0	0	
C5orf60	5	179068544	179072047	0	1	1	0	
C9orf57	9	74666291	74675521	0	0	1	0	
CATSPER2	15	43922871	43941024	2	2	1	0	
CBWD2	2	114195267	114253766	0	6	2	0	
CBWD3	9	70856396	70914929	0	0	1	0	
CCDC144A	17	16593574	16707767	1	6	2	0	
CCDC74B	2	130896859	130902707	0	7	2	0	
CD8B	2	87042679	87089038	1	3	0	0	
CD97	19	14492255	14519533	0	5	1	0	
CDRT1	17	15491976	15522826	0	12	1	0	
CELA3A	1	22328148	22339015	1	6	0	0	
CELA3B	1	22303513	22315837	0	8	1	0	
CEP170	1	243287729	243418352	1	8	0	0	
CFH	1	196621007	196716634	2	3	0	0	
CFHR1	1	196788886	196801319	0	9	2	0	
CFHR4	1	196857143	196887843	0	2	1	0	
CKMT1B	15	43885251	43891604	1	4	2	0	
CRYBB2	22	25615488	25627836	1	1	0	0	
CSH2	17	61949371	61951126	0	5	1	0	
CST4	20	23666276	23669677	0	3	1	0	
CXorf40A	X	148622186	148628855	1	6	0	0	
CYP11B1	8	143954669	143961262	0	4	1	0	
CYP2D6	22	42522500	42526908	1	10	1	0	
CYP4A11	1	47394848	47407137	1	8	0	0	
CYP4A22	1	47603126	47615413	0	10	1	0	
DGCR6	22	18893735	18899601	0	2	1	0	
DGCR6L	22	20301798	20307603	0	4	1	0	
DHRS4L2	14	24458030	24475617	0	3	1	0	

EEF1A1	6	74225472	74230741	0	1	1	0	
EIF3CL	16	28390904	28415165	1	0	0	0	
FAM153B	5	175490539	175541801	1	7	0	0	
FAM185A	7	102389654	102448849	0	1	2	0	
FAM230A	22	20692437	20716938	1	5	0	0	
FAM35A	10	88854952	88951220	0	0	1	0	
FAM47B	X	34960912	34963034	0	1	1	0	
FAM86C1	11	71498559	71512282	2	1	0	0	
FCGBP	19	40353962	40440533	0	38	1	0	
FCGR1A	1	149754245	149764074	0	3	1	0	
FLG2	1	152321210	152332482	0	9	2	0	
FOLH1	11	49168186	49230222	2	9	1	0	
FOXD4L5	9	70175706	70178815	0	3	1	0	
FRG1B	20	29611856	29634010	2	9	0	0	
FUT3	19	5842901	5851482	0	6	1	0	
FUT5	19	5866181	5903798	0	4	1	0	
GAGE12C	X	49296813	49304141	1	0	0	0	
GOLGA6B	15	72947078	72958735	0	0	1	0	
GOLGA6D	15	75575175	75586816	0	1	1	0	
GOLGA8O	15	32737306	32747835	0	1	1	0	
GOLGA8R	15	30695942	30706365	0	1	1	0	
GRM8	7	126078651	126893348	0	0	1	0	
GSTM2	1	110210687	110217908	1	0	0	0	
GTF2IRD2B	7	74508386	74565623	0	4	1	0	
HERC2	15	28356185	28567298	1	16	0	0	
IFITM1	11	313852	315272	0	0	1	0	
IFNA10	9	21206179	21207142	0	6	1	0	
IFNA17	9	21227241	21228221	0	6	3	0	
IGFN1	1	201159952	201198080	0	7	1	0	
IGSF3	1	117117030	117210314	0	8	4	0	
KLRC3	12	10568182	10573149	0	1	1	0	
KMT2C	7	151832009	152133090	1	8	2	0	
KRT6A	12	52880957	52887041	1	9	3	0	
KRT6B	12	52840434	52845910	1	12	0	0	

KRTAP4-9	17	39261583	39262740	0	3	2	0	
KRTAP5-10	11	71276608	71277666	0	0	1	0	
KRTAP9-2	17	39382899	39383904	0	6	1	0	
LCE1E	1	152758752	152760902	0	7	0	0	
LCE2B	1	152658598	152659877	0	3	1	0	
LILRB1	19	55141907	55148979	1	1	0	0	
LINC00999	10	38717073	38736820	1	1	0	0	
LRRC37B	17	30348160	30380523	1	4	0	0	
METTL2A	17	60501227	60527454	0	3	1	0	
MGAM	7	141695637	141806547	1	5	0	0	
MICB	6	31465891	31478901	1	2	0	0	
MRGPRX1	11	18955359	18956556	0	2	1	0	
MST1	3	49721379	49726486	1	1	1	0	
MT-CO1	MT	5902	7444	0	8	2	0	
MT-ND4	MT	10758	12136	0	6	3	0	
MT-ND6	MT	14147	14672	0	0	1	0	
MTRF1L	6	153308496	153323820	0	1	1	0	
MUC8	12	133049333	133050726	0	5	1	0	
NBPF1	1	16888813	16939982	0	18	2	0	
NBPF24	1	147575936	147599549	1	1	0	0	
NBPF9	1	144811747	144830413	0	8	1	0	
NLRP2	19	55476619	55512510	0	6	1	0	
NOTCH2	1	120454175	120612240	0	1	2	0	
NPEPPS	17	45608429	45700642	0	1	2	0	
NPIPA1	16	15031299	15045913	0	0	1	0	
NSUN5	7	72717231	72722813	0	7	0	0	
NUTM2A	10	88985204	88994735	0	2	1	0	
NUTM2E	10	81601113	81610628	0	2	1	0	
OR10G2	14	22101991	22103096	0	6	1	0	
OR10G9	11	123893719	123894655	0	9	1	0	
OR10H1	19	15917816	15918936	0	3	1	0	
OR1E2	17	3336163	3337135	0	1	1	0	
OR1S2	11	57970673	57971653	0	5	0	0	
OR2L8	1	248112159	248113098	0	4	1	0	
OR2T12	1	248457917	248458880	0	10	2	0	
OR2T29	1	248721783	248722797	0	0	1	0	

OR2T33	1	248436072	248437138	0	5	1	0	
OR2T34	1	248737021	248738080	0	10	1	0	
OR2T7	1	248604507	248605434	0	7	3	0	
OR4A47	11	48510268	48511332	0	8	1	0	
OR51A2	11	4976001	4976943	0	0	1	0	
OR51A4	11	4967354	4968356	0	11	1	0	
OR5H1	3	97851541	97852483	0	1	1	0	
OR5K1	3	98188323	98189420	0	8	1	0	
OR8B3	11	124266277	124267264	0	5	2	0	
ORM2	9	117092148	117095532	0	4	1	0	
PAGE2	X	55115440	55119275	0	2	1	0	
PCDHA8	5	140220906	140391929	0	5	1	0	
PCDHB10	5	140571941	140575215	0	9	1	0	
PCDHB14	5	140603077	140605858	0	9	1	0	
PCDHB4	5	140501580	140505201	0	3	1	0	
PCDHB8	5	140557370	140560081	0	21	1	0	
PDE4DIP	1	144851426	144995022	0	8	2	0	
PDXDC1	16	15068959	15131552	0	3	1	0	
PGM5	9	70971814	71145977	0	1	1	0	
PKD1	16	2138710	2185899	0	24	2	0	
PLEKHM1	17	43513265	43568110	1	11	0	0	
PMS2	7	6012869	6048756	1	2	0	0	
POTED	21	14982497	15013906	0	2	1	0	
POTEH	22	16256440	16287937	0	6	1	0	
POTEM	14	19983558	20020272	1	5	2	0	
PRAMEF1	1	12851545	12856777	0	17	1	0	
PRAMEF11	1	12884467	12891264	0	15	1	0	
PRAMEF26	1	13216355	13219581	0	1	1	0	
PRAMEF4	1	12939044	12946025	0	7	1	0	
PRB2	12	11544475	11548499	1	1	0	0	
PRH2	12	11081834	11084940	0	2	1	0	
PRKRIR	11	76060999	76091986	0	3	1	0	
PSG1	19	43371357	43383871	2	4	1	0	
PSG11	19	43511838	43530628	0	1	1	0	

PSG4	19	43696853	43709926	0	4	1	0	
RABL2B	22	51205933	51222070	0	3	1	0	
RAC1	7	6414169	6442151	0	2	1	0	
RASA4B	7	102123589	102158157	0	5	1	0	
RBMX	X	135955619	135962884	0	2	1	0	
RHCE	1	25688739	25747336	1	9	2	0	
RHD	1	25598883	25656936	0	10	2	0	
ROBO2	3	75955845	77696661	0	0	1	0	
ROPN1B	3	125687986	125702297	1	0	0	0	
RP11-514P8.7	7	102181947	102232891	1	2	0	0	
RSPH10B2	7	6793739	6838394	1	7	1	0	
SERPINB4	18	61304492	61311532	1	1	0	0	
SFTPA1	10	81370700	81373954	0	5	1	0	
SFTPA2	10	81315607	81320151	0	3	2	0	
SLC9B1P1	Y	13496240	13524717	0	1	1	0	
SLFN13	17	33762114	33775856	0	5	1	0	
SMN2	5	69345438	69373419	1	1	0	0	
SMPD4	2	130908981	130940323	0	8	1	0	
SORD	15	45315301	45369383	1	1	0	0	
STAT5A	17	40439564	40463961	1	0	0	0	
SULT1A2	16	28603265	28607801	1	2	1	0	
TBC1D28	17	18538318	18545737	0	1	1	0	
TCP10	6	167785664	167797954	0	2	1	0	
TMEM255B	13	114462215	114514899	0	0	1	0	
TPTE	21	10906740	10990882	1	0	0	0	
TRIM74	7	72430015	72439997	0	0	2	0	
TRIOBP	22	38093010	38172563	0	5	1	0	
TUBA3D	2	132233665	132240507	1	14	1	0	
ULBP2	6	150263135	150270371	0	2	1	0	
USP18	22	18632665	18660164	2	3	0	0	
USP32	17	58256454	58469495	1	4	0	0	
USP6	17	5019732	5078286	0	4	1	0	
XCL1	1	168545710	168551315	0	5	1	0	
ZNF286B	17	18561741	18585572	1	3	0	0	
ZNF468	19	53341260	53360872	1	0	0	0	

ZNF585A	19	37641003	37663615	0	8	1	0	
ZNF587	19	58361224	58376480	0	6	3	0	
ZNF587B	19	58341662	58357606	0	1	1	0	
ZNF705G	8	7213038	7243080	0	2	1	0	
ZNF83	19	53115629	53140339	0	2	1	0	
ZNF98	19	22573820	22605148	1	3	0	0	
ZXDB	X	57618268	57623906	0	7	1	0	
BWA MEM joondamisalgoritmi poolt joondatud lugemitest saadud artefaktsete SNVde ülevaade olulisuse ja geenide järgi								
GEEN	CHR	START	STOP	ESS	NSC	SG	SL	NCS&SS
LCE1E	1	152758752	152760902	0	5	0	1	0
ACSM2B	16	20547547	20587689	0	3	4	0	0
MT-ND4	MT	10758	12136	0	5	3	0	0
AKR1C2	10	5029966	5046215	1	5	2	0	0
CFHR1	1	196788886	196801319	0	11	2	0	0
DHRS4L2	14	24458030	24475617	0	2	2	0	0
FLG2	1	152321210	152332482		5	2	0	0
IGSF3	1	117117030	117210314	0	4	2	0	0
ARHGAP5	14	32546494	32628934	0	3	1	0	0
ARHGAP21	10	24872537	25012597	0	1	1	0	0
ARHGAP5	14	32546494	32628934	0	3	1	0	0
BCLAF1	6	136578000	136610989	0	3	1	0	0
C5orf60	5	179068544	179072047	0	0	1	0	0
CCDC144A	17	16593574	16707767	0	3	1	0	0
CCDC74B	2	130896859	130902707	0	3	1	0	0
CKMT1B	15	43885251	43891604	0	1	1	0	0
CST4	20	23666276	23669677	0	2	1	0	0
DGCR6	22	18893735	18899601	0	1	1	0	0
DGCR6L	22	20301798	20307603	0	1	1	0	0
FAM185A	7	102389654	102448849	0	1	1	0	0
GOLGA6B	15	72947078	72958735	0	0	1	0	0
IFNA17	9	21227241	21228221	0	0	1	0	0
KIR2DL1	19	55281262	55295774	0	10	1	0	0
METTL2A	17	60501227	60527454	0	0	1	0	0
MT-CO1	MT	5902	7444	0	5	1	0	0
MT-ND5	MT	12335	14147	0	10	1	0	0
ROBO2	3	75955845	77696661	0	0	1	0	0
THAP11	16	67876212	67878097	0	0	1	0	0
TRIOBP	22	38093010	38172563	0	0	1	0	0
FOLH1	11	49168186	49230222	2	1	0	0	0
BTN3A2	6	26365458	26378546	1	0	0	0	0
CELA3A	1	22328148	22339015	1	2	0	0	0
CEP170	1	243287729	243418352	1	2	0	0	0

CYP11B2	8	143991974	143999259	1	7	0	0	0
FRG1B	20	29611856	29634010	1	0	0	0	0
LILRB1	19	55141907	55148979	1	0	0	0	0
LINC00999	10	38717073	38736820	1	0	0	0	0
TPTE	21	10906740	10990882	1	0	0	0	0
NOTCH2	1	120454175	120612240	0	1	2	0	1
PCDHA8	5	140220906	140391929	0	1	2	0	1
PRH2	12	11081834	11084940	0	1	2	0	1
TBC1D28	17	18538318	18545737	0	1	2	0	1
ZNF585A	19	37641003	37663615	0	1	2	0	1
USP18	22	18632665	18660164	1	1	1	0	1
ZNF286B	17	18561741	18585572	1	1	1	0	1
MUC5B	11	1244295	1283402	0	4	1	0	1
MYH2	17	10424468	10453274	0	1	1	0	1
MZT2A	2	132241532	132249995	0	1	1	0	1
NCL	2	232318241	232329305	0	1	1	0	1
NEB	2	152341852	152591001	0	1	1	0	1
NOMO1	16	14927537	14990017	0	1	1	0	1
NSUN5	7	72717231	72722813	0	1	1	0	1
NUTM2F	9	97080477	97090926	0	1	1	0	1
OBP2A	9	138438000	138441792	0	1	1	0	1
ODF1	8	103563799	103573216	0	1	1	0	1
OR10A5	11	6866882	6867936	0	1	1	0	1
OR10G9	11	123893719	123894655	0	1	1	0	1
OR2T27	1	248813231	248814185	0	1	1	0	1
OR51A4	11	4967354	4968356	0	1	1	0	1
OR5H1	3	97851541	97852483	0	1	1	0	1
OR5H14	3	97868169	97869249	0	1	1	0	1
OR5K1	3	98188323	98189420	0	1	1	0	1
ORM2	9	117092148	117095532	0	1	1	0	1
PAGE2B	X	55101503	55105336	0	1	1	0	1
PCDH11X	X	91089658	91878226	0	1	1	0	1
PCDH11Y	Y	4924130	4972741	0	1	1	0	1
PDE4DIP	1	144851426	144995022	0	0	1	0	1
PDXDC1	16	15068959	15131552	0	1	1	0	1
PF4	4	74846793	74847841	0	1	1	0	1
PIGF	2	46808075	46844258	0	1	1	0	1
POM121	7	72349935	72421979	0	1	1	0	1
POM121C	7	75046068	75115548	0	1	1	0	1
POTEC	18	14511896	14543145	0	1	1	0	1
POTED	21	14982497	15013906	0	1	1	0	1
POTEF	2	130831107	130878182	0	1	1	0	1
POTEH	22	16256440	16287937	0	1	1	0	1
PRB2	12	11544475	11548499	0	1	1	0	1
PRKRIR	11	76060999	76091986	0	1	1	0	1

PRODH	22	18900294	18924066	0	1	1	0	1
PYROXD1	12	21590610	21623300	0	1	1	0	1
RABL2B	22	51205933	51222070	0	1	1	0	1
RBM4B	11	66432468	66445219	0	1	1	0	1
RGPD3	2	107021445	107084832	0	1	1	0	1
RLN1	9	5334968	5339873	0	1	1	0	1
SEC14L6	22	30918785	30942669	0	1	1	0	1
SERPINB3	18	61322430	61329197	0	1	1	0	1
SIRPB1	20	1544166	1600655	0	1	1	0	1
SLC9B1P1	Y	13496240	13524717	0	1	1	0	1
SMN1	5	70220856	70248839	0	1	1	0	1
SPAG11A	8	7705545	7721319	0	1	1	0	1
STEAP1	7	89783688	89794141	0	1	1	0	1
STEAP1B	7	22478037	22540117	0	1	1	0	1
SULT1A1	16	28616917	28634874	0	1	1	0	1
TBX20	7	35242041	35293758	0	1	1	0	1
TRIM49	11	89530822	89541743	0	1	1	0	1
TRPA1	8	72933485	72987852	0	1	1	0	1
TWF1	12	44189295	44200052	0	1	1	0	1
TYW1	7	66461801	66704501	0	1	1	0	1
UBC	12	125396149	125399894	0	1	1	0	1
UGT2B28	4	70146216	70160768	0	1	1	0	1
UNC93A	6	167704802	167729507	0	1	1	0	1
VCX	X	7810302	7812184	0	1	1	0	1
VN1R4	19	53769928	53770972	0	1	1	0	1
YY1AP1	1	155629245	155658260	0	1	1	0	1
ZDHHC11	5	795721	851101	0	1	1	0	1
ZNF585B	19	37672480	37701451	0	1	1	0	1
ZSCAN5B	19	56700938	56709289	0	1	1	0	1
ZXDA	X	57931863	57937067	0	1	1	0	1
OR2T12	1	248457917	248458880	0	2	3	0	2
OR8B3	11	124266277	124267264	0	2	3	0	2
PSG6	19	43407767	43421989	0	2	3	0	2
NUDT11	X	51232862	51239448	0	2	2	0	2
OR2T34	1	248737021	248738080	0	2	2	0	2
ORM1	9	117085335	117088755	0	1	2	0	2
PCDHA7	5	140213968	140391929	0	2	2	0	2
PCDHB10	5	140571941	140575215	0	2	2	0	2
PRAMEF11	1	12884467	12891264	0	2	2	0	2
PRDM9	5	23507723	23528706	0	2	2	0	2
RP11-683L23.1	18	47389	49557	0	2	2	0	2
SIRPA	20	1875941	1920543	0	2	2	0	2
SLC25A52	18	29339524	29340843	0	2	2	0	2
SLFN13	17	33762114	33775856	0	2	2	0	2
SPRR2B	1	153042716	153044084	0	2	2	0	2

SULT1A2	16	28603265	28607801	0	2	2	0	2
SVIL	10	29746276	30024730	0	2	2	0	2
TGIF2LY	Y	3447081	3448082	0	2	2	0	2
TLR6	4	38825335	38858437	0	2	2	0	2
TRIM16L	17	18601322	18639431	0	2	2	0	2
TUBA1C	12	49658689	49667109	0	2	2	0	2
TUBA3C	13	19747909	19755992	0	2	2	0	2
UBB	17	16284112	16286054	0	2	2	0	2
ZNF181	19	35225853	35233091	0	2	2	0	2
ZNF285	19	44886458	44905755	0	2	2	0	2
ZNF676	19	22361892	22379753	0	2	2	0	2
ZNF812	19	9800599	9811452	0	2	2	0	2
OBP2B	9	136080663	136084628	0	1	1	0	2
OR10H1	19	15917816	15918936	0	3	4	0	3
PSG11	19	43511838	43530628	0	3	4	0	3
SFTPA1	10	81370700	81373954	0	3	4	0	3
SFTPA2	10	81315607	81320151	0	3	4	0	3
XCL1	1	168545710	168551315	0	3	4	0	3
ZXDB	X	57618268	57623906	0	3	4	0	3
NBPF1	1	16888813	16939982	0	3	3	0	3
OR2T3	1	248636626	248637634	0	3	3	0	3
OR52I1	11	4615268	4616243	0	3	3	0	3
OR5H15	3	97887543	97888485	0	3	3	0	3
OR8B2	11	124252290	124253258	0	3	3	0	3
PCDHB12	5	140588290	140591696	0	3	3	0	3
SMPD4	2	130908981	130940323	0	3	3	0	3
SPRR1B	1	153003677	153005376	0	3	3	0	3
UBXN11	1	26608774	26644854	0	3	3	0	3
VCX3A	X	6451658	6453159	0	3	3	0	3
ZNF286A	17	15603090	15624101	0	3	3	0	3
RHCE	1	25688739	25747336	0	2	2	0	3
SPANXD	X	140785567	140786896	0	2	2	0	3
TRIM16	17	15531273	15587614	0	2	2	0	3
SMN2	5	69345438	69373419	1	1	1	0	3
RASA4B	7	102123589	102158157	0	4	5	0	4
NLRP7	19	55434882	55477611	0	4	4	0	4
OR10H5	19	15904760	15905892	0	4	4	0	4
OR1S1	11	57982216	57983195	0	4	4	0	4
OR2T7	1	248604507	248605434	0	4	4	0	4
PCDHB13	5	140593508	140596993	0	4	4	0	4
PCDHB2	5	140474226	140476962	0	4	4	0	4
PKD1	16	2138710	2185899	0	4	4	0	4
PSG1	19	43371357	43383871	0	4	4	0	4
TUBB8	10	92831	95241	0	4	4	0	4
ZNF746	7	149169886	149194898	0	4	4	0	4

ZNF799	19	12500829	12512085	0	4	4	0	4
PRAMEF1	1	12851545	12856777	0	5	7	0	5
RHD	1	25598883	25656936	0	5	6	0	5
PCDHA9	5	140227047	140391929	0	5	5	0	5
SLFN11	17	33677323	33700639	0	5	5	0	5
XCL2	1	168510002	168513235	0	4	4	0	5
NLRP2	19	55476619	55512510	0	6	7	0	6
OR4A47	11	48510268	48511332	0	6	7	0	6
PCDHA6	5	140207562	140391929	0	6	6	0	6
PLEKHM1	17	43513265	43568110	0	6	6	0	6
PSG3	19	43225793	43244721	0	6	6	0	6
SPRR2A	1	153028588	153030013	0	6	6	0	6
TUBA3D	2	132233665	132240507	0	6	6	0	6
OR13C9	9	107379528	107380485	0	7	7	0	7
TUBA3E	2	130949317	130956034	0	7	7	0	7
PCDHB8	5	140557370	140560081	0	8	8	0	8
UGT1A10	2	234545099	234681951	0	8	8	0	8
OR13C5	9	107360649	107361788	0	9	9	0	9
OR13C2	9	107366923	107367951	0	11	11	0	11
MUC6	11	1012820	1036706	0	15	15	0	15
PRAMEF2	1	12916940	12921764	0	16	16	0	16

Bowtie2 *local* joondamisalgoritmi poolt joondatud lugemitest saadud artefaktsete SNVde ülevaade olulisuse ja geenide järgi

GEEN	CHR	START	END	SG	SL	SNC	NSCjaSS	
ACSM2A	16	20462897	20498991	0	0	7	1	
ACSM2B	16	20547547	20587689	0	0	3	0	
ADAM29	4	175839508	175899331	0	0	6	0	
AKR1C2	10	5029966	5046215	1	0	4	0	
AMY1B	1	104230039	104238912	1	0	2	0	
ANKRD30A	10	37414784	37521495	1	0	0	0	
ANKRD30B	18	14748238	14852479	1	0	5	0	
ANKRD36	2	97779232	97930258	0	0	7	0	
ANXA8L1	10	47157988	47174040	1	0	0	0	
ARHGAP11B	15	30918878	30931013	0	0	2	1	
ARHGAP21	10	24872537	25012597	0	0	2	0	
ARHGAP5	14	32546494	32628934	0	0	3	0	
ARMC4	10	28101096	28287977	0	0	0	1	
BEX2	X	102564281	102565883	0	0	1	0	
BTN3A2	6	26365458	26378546	1	0	0	0	
C5orf60	5	179068544	179072047	0	0	0	0	
CACNA1C	12	2162728	2800459	0	0	0	1	
CATSPER2	15	43922871	43941024	1	0	1	0	
CBWD2	2	114195267	114253766	0	0	3	1	
CBWD6	9	69204537	69262593	0	0	0	0	
CCDC144A	17	16593574	16707767	0	0	3	0	
CCDC74B	2	130896859	130902707	0	0	4	0	

CELA3A	1	22328148	22339015	2	0	6	0	
CEP170	1	243287729	243418352	1	0	5	0	
CFH	1	196621007	196716634	1	0	0	0	
CFHR1	1	196788886	196801319	0	0	14	0	
CKMT1B	15	43885251	43891604	0	0	4	0	
CR1L	1	207818518	207897048	0	0	5	1	
CT47B1	X	120006456	120009779	0	0	1	1	
CYP11B1	8	143954669	143961262	0	0	2	1	
CYP2A7	19	41381343	41388657	0	0	7	0	
CYP2D6	22	42522500	42526908	0	0	6	0	
CYP4A11	1	47394848	47407137	1	0	9	0	
DDX11	12	31226778	31257725	0	0	4	1	
DGCR6	22	18893735	18899601	0	0	1	0	
DGCR6L	22	20301798	20307603	0	1	2	0	
DHRS4L2	14	24458030	24475617	0	0	2	0	
DPY19L2	12	63952692	64062357	0	0	1	1	
EIF3CL	16	28390904	28415165	1	0	0	0	
FAM153B	5	175490539	175541801	1	0	4	0	
FAM185A	7	102389654	102448849	0	0	1	0	
FAM27E2	9	45733558	45734896	0	0	1	0	
FAM35A	10	88854952	88951220	0	0	0	0	
FCGBP	19	40353962	40440533	0	0	10	2	
FLG	1	152274650	152297679	0	0	17	0	
FOLH1	11	49168186	49230222	1	0	6	0	
FOXD4L5	9	70175706	70178815	0	0	2	0	
FRG1	4	190861942	190884359	1	0	1	1	
FRG1B	20	29611856	29634010	2	0	2	0	
GAGE2A	X	49354131	49361430	0	0	2	1	
GOLGA6B	15	72947078	72958735	0	0	1	0	
GOLGA8H	15	30896328	30906764	0	0	0	1	
GOLGA8J	15	30375255	30385702	0	0	2	0	
GOLGA8O	15	32737306	32747835	0	0	1	0	
GSTM2	1	110210687	110217908	1	0	0	0	
HLA-DRB1	6	32546545	32557625	0	0	3	1	
HLA-DRB5	6	32485119	32498064	0	0	6	0	
HRNR	1	152184557	152196669	0	0	0	0	
IFNL2	19	39759153	39760732	0	0	4	1	
IGHV1OR15-1	15	22448381	22448819	0	0	7	0	
IGHV4-61	14	107095125	107095662	1	0	0	0	
IGKV2D-24	2	90043606	90044439	0	0	0	0	
IGKV2D-29	2	89986321	89987079	0	0	3	1	
IGLL5	22	23229959	23238005	0	1	6	0	
IGSF3	1	117117030	117210314	0	0	6	0	
KIR2DL1	19	55281262	55295774	0	0	9	0	
KLRC3	12	10568182	10573149	0	0	1	0	

KMT2C	7	151832009	152133090	1	0	0	0	
KRT6A	12	52880957	52887041	0	0	2	0	
KRTAP5-10	11	71276608	71277666	0	0	0	0	
KRTAP9-2	17	39382899	39383904	0	0	2	0	
LCE1E	1	152758752	152760902	0	1	3	0	
LILRB1	19	55141907	55148979	1	0	0	0	
LILRB4	19	55173578	55181810	1	0	2	1	
LINC00999	10	38717073	38736820	1	0	0	0	
LRRC37B	17	30348160	30380523	1	0	4	0	
METTL2A	17	60501227	60527454	0	0	1	0	
MGAM	7	141695637	141806547	1	0	1	2	
MICB	6	31465891	31478901	1	0	2	0	
MT-CO1	MT	5902	7444	0	1	6	0	
MT-ND4	MT	10758	12136	0	0	3	0	
MT-ND5	MT	12335	14147	0	0	8	0	
MTX1	1	155178489	155183614	0	0	0	1	
MUC17	7	100663352	100702020	0	0	6	0	
NBPF1	1	16888813	16939982	0	0	6	1	
NBPF9	1	144811747	144830413	0	0	3	0	
NLRP2	19	55476619	55512510	0	0	11	0	
NOTCH2	1	120454175	120612240	0	0	1	0	
NPEPPS	17	45608429	45700642	0	0	1	0	
NPIPA1	16	15031299	15045913	0	0	0	0	
NPIPA3	16	14805545	14820150	0	0	4	0	
NSUN5	7	72717231	72722813	0	0	0	1	
NUTM2A	10	88985204	88994735	0	0	0	0	
OR10H1	19	15917816	15918936	0	0	5	0	
OR2T12	1	248457917	248458880	0	0	5	0	
OR2T2	1	248616076	248617130	0	0	5	0	
OR2T33	1	248436072	248437138	0	0	2	0	
OR2T34	1	248737021	248738080	0	0	8	0	
OR2T7	1	248604507	248605434	0	0	6	0	
OR4A47	11	48510268	48511332	0	0	5	0	
OR8B3	11	124266277	124267264	0	0	4	0	
ORM1	9	117085335	117088755	0	0	2	1	
PCDHA8	5	140220906	140391929	0	0	5	0	
PCDHB13	5	140593508	140596993	0	0	8	0	
PCDHB7	5	140552242	140555957	0	0	4	0	
PDE4DIP	1	144851426	144995022	0	0	0	1	
PKD1	16	2138710	2185899	0	0	4	1	
PLEKHM1	17	43513265	43568110	0	0	8	1	
POM121C	7	75046068	75115548	0	0	5	0	
POTED	21	14982497	15013906	0	0	1	0	
POTEM	14	19983558	20020272	1	0	1	0	
PRAMEF1	1	12851545	12856777	0	0	8	0	

PRAMEF11	1	12884467	12891264	0	0	2	0	
PRAMEF26	1	13216355	13219581	0	0	0	1	
PRH2	12	11081834	11084940	0	0	1	0	
PRODH	22	18900294	18924066	0	0	1	1	
PSG1	19	43371357	43383871	1	0	8	0	
PSG11	19	43511838	43530628	0	0	4	0	
PSG6	19	43407767	43421989	0	0	5	0	
PSG9	19	43757433	43773680	0	0	2	0	
RABL2B	22	51205933	51222070	0	0	3	0	
RASA4B	7	102123589	102158157	0	0	5	0	
RHCE	1	25688739	25747336	0	0	4	1	
RHD	1	25598883	25656936	0	0	9	1	
ROBO2	3	75955845	77696661	0	0	0	0	
ROPN1B	3	125687986	125702297	1	0	0	0	
RSPH10B	7	5965776	6010314	0	0	0	1	
RSPH10B2	7	6793739	6838394	1	0	4	0	
SAA1	11	18287810	18291524	1	0	2	0	
SFTPA1	10	81370700	81373954	0	0	3	0	
SFTPA2	10	81315607	81320151	0	0	3	0	
SIGLEC11	19	50452241	50464429	0	0	0	1	
SLC9B1P1	Y	13496240	13524717	0	0	1	0	
SMN2	5	69345438	69373419	1	0	1	2	
SPAG11B	8	7308132	7321176	0	0	1	1	
SPANXD	X	140785567	140786896	0	0	3	1	
SULT1A1	16	28616917	28634874	0	0	4	1	
SUZ12	17	30264036	30328064	0	0	0	1	
TBC1D28	17	18538318	18545737	0	0	1	0	
THAP11	16	67876212	67878097	0	0	0	0	
TPTE	21	10906740	10990882	1	0	0	0	
TRIM16	17	15531273	15587614	0	0	2	1	
TRIM43	2	96257765	96265526	0	0	0	1	
TRIM49C	11	89764273	89775193	0	0	1	0	
TRIM74	7	72430015	72439997	0	0	0	1	
TRIOBP	22	38093010	38172563	0	0	2	0	
TUBA3D	2	132233665	132240507	1	0	15	1	
USP18	22	18632665	18660164	1	0	1	0	
USP32	17	58256454	58469495	1	0	3	0	
XCL1	1	168545710	168551315	0	0	3	1	
XCL2	1	168510002	168513235	0	0	4	1	
ZNF286B	17	18561741	18585572	1	0	3	0	
ZNF585A	19	37641003	37663615	0	0	4	0	
ZNF587	19	58361224	58376480	0	0	0	0	
ZNF626	19	20802866	20844399	0	0	4	0	
ZNF773	19	58011308	58019529	0	0	0	1	

Bowtie2 *end-to-end* joondamisalgoritmi poolt joondatud lugemitest saadud artefaktsete SNVde ülevaade olulisuse ja geenide järgi								
GEEN	CHR	START	STOP	ESS	NSC	NS&SS		
AC092850.1	12	47394848	47407137	0	2	0		
ACSM2B	16	196788886	196801319	0	4	0		
ACTR3B	7	25598883	25656936	0	4	0		
ADAM29	4	12851545	12856777	0	4	0		
AHNAK2	14	201159952	201198080	0	4	0		
AKR1C1	10	7844762	7905237	0	4	0		
AKR1C2	10	243287729	243418352	0	3	0		
AL390778.1	9	207669612	207812754	0	1	0		
ANAPC1	2	168510002	168513235	0	1	0		
ANKRD30B	18	120926978	120935937	0	6	0		
ANKRD36	2	152321210	152332482	0	6	0		
AP003062.1	11	207818518	207897048	0	6	0		
ARHGAP5	14	33439267	33439642	0	6	0		
ATN1	12	132285405	132291239	0	6	0		
BCLAF1	6	114256660	114258728	0	6	0		
BEX2	X	232318241	232329305	0	2	0		
C1QTNF9	13	112525216	112642267	0	7	0		
CCDC74A	2	131975881	132022965	0	7	0		
CDHR5	11	130831107	130878182	0	7	0		
CDRT15L2	17	27323465	27341971	0	1	0		
CEP170	1	108443392	108507297	0	1	0		
CFHR1	1	130908981	130940323	0	2	0		
CGREF1	2	97779232	97930258	0	3	0		
CKMT1B	15	175839508	175899331	0	3	0		
CNTNAP3	9	140552242	140555957	0	3	0		
CR1	1	179068544	179072047	0	2	0		
CR1L	1	31465891	31478901	0	4	0		
CYP11B2	8	27775898	27776429	0	4	0		
CYP2A6	19	108882068	109005971	0	1	0		
CYP2A7	19	136578000	136610989	0	3	0		
CYP4A11	1	152456833	152552463	0	3	0		
DHRS4L2	14	128116782	128142978	0	3	0		
DPY19L2	12	38388955	38389623	0	2	0		
DUOX1	15	38393315	38394118	0	2	0		
EVPL	17	75046068	75115548	0	3	0		
FAM205A	9	103563799	103573216	0	3	0		
FAM209B	20	143991974	143999259	0	3	0		
FCGBP	19	5334968	5339873	0	1	0		
FCGR1B	1	117085335	117088755	0	2	0		
FKSG48	1	39072763	39288135	0	2	0		
FLG2	1	34723051	34729464	0	2	0		
FOXD4L1	2	5299867	5304969	0	2	0		

FOXO3	6	107366923	107367951	2	0	0		
FRG1B	20	138150511	138151275	0	2	0		
HBZ	16	7584	8268	0	2	0		
HERC2	15	10758	12136	0	1	0		
HIST1H2AI	6	12335	14147	0	0	0		
IFNL2	19	14745	15886	0	2	0		
IGFN1	1	151867213	151870825	0	5	0		
IGHA1	14	140785567	140786896	0	5	0		
IGHA2	14	37430904	37536647	0	11	0		
IGHV1OR15-1	15	6451658	6453159	0	11	0		
IGHV3-66	14	102564281	102565883	0	11	0		
IGHV4-61	14	81315607	81320151	0	6	0		
KDM4E	11	29746276	30024730	0	6	0		
KIR2DL1	19	5029966	5046215	0	6	0		
KLRC2	12	5005444	5025475	0	6	0		
KRT6B	12	81370700	81373954	0	1	0		
KRT6C	12	76060999	76091986	0	6	0		
KRTAP4-1	17	616576	626078	0	6	0		
LANCL3	X	48510268	48511332	0	6	0		
LILRB3	19	94758421	94760760	0	1	0		
LRRC37A	17	66432468	66445219	0	0	0		
LRRC37A3	17	1244295	1283402	0	0	1		
LRRC37B	17	134855245	134856693	0	2	0		
MAGEA6	X	52862299	52867569	0	1	0		
METTL2B	7	52840434	52845910	0	1	0		
MICB	6	49658689	49667109	0	7	0		
MT-CO2	MT	63952692	64062357	0	7	0		
MT-CYB	MT	7033625	7051484	0	7	0		
MT-ND4	MT	10583197	10588592	0	7	0		
MT-ND5	MT	131780066	131782763	0	4	0		
MUC5B	11	114462215	114514899	0	4	0		
NCL	2	24881303	24896673	0	4	0		
NLRP2	19	105403653	105444694	0	4	0		
NOMO3	16	24458030	24475617	0	4	0		
NPEPPS	17	32546494	32628934	0	2	0		
ODF1	8	22314718	22315403	0	1	0		
OR13C2	9	106053225	106054732	0	2	0		
OR4A47	11	106173456	106175002	0	2	0		
ORM1	9	107095125	107095662	0	2	0		
PCDHB7	5	107131032	107131560	0	4	0		
PER3	1	28356185	28567298	0	4	0		
PKD1	16	43885251	43891604	1	0	0		
POM121C	7	45422130	45457774	0	2	0		
POTED	21	22448381	22448819	0	1	0		
POTEE	2	202685	204502	0	3	0		

POTEF	2	2138710	2185899	0	3	0		
PRAMEF1	1	20547547	20587689	0	8	0		
RBM4B	11	16326351	16388668	0	8	0		
RGPD4	2	58256454	58469495	0	8	0		
RHD	1	74002925	74023533	0	8	0		
RLN1	9	44372496	44415160	0	1	0		
RLN2	9	45608429	45700642	0	1	0		
SFTPA1	10	30348160	30380523	0	2	0		
SFTPA2	10	39340353	39341594	0	2	0		
SLC25A52	18	20483036	20484224	0	2	0		
SMPD4	2	62850487	62914903	0	2	0		
SPANXD	X	29339524	29340843	0	1	0		
SVIL	10	14748238	14852479	2	0	0		
TMEM255B	13	40353962	40440533	0	1	0		
TPTE	21	54720796	54726850	0	19	0		
TRAV8-2	14	41349442	41356352	0	19	0		
TRGV4	7	41381343	41388657	0	19	0		
TRGV5	7	44886458	44905755	0	19	0		
TUBA1C	12	39759153	39760732	0	19	0		
USP32	17	55281262	55295774	0	19	0		
VCX3A	X	9800599	9811452	0	19	0		
XCL2	1	55476619	55512510	0	19	0		
ZNF285	19	22940115	22966909	0	19	0		
ZNF626	19	20802866	20844399	0	19	0		
ZNF812	19	29611856	29634010	0	19	0		
ZNF99	19	55108301	55111576	0	19	0		
C5orf60	5	14982497	15013906	0	1	0		
PRKRIR	11	10906740	10990882	0	1	0		

LIHTLITSENTS

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Anna Smertina (sünnikuupäev 12.01.19911)

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

„Inimgenoomi ühenukleotiidiliste variatsioonide annotatsioon – ülevaade põhimõtetest ning teise põlvkonna sekveneerimise võimalike artefaktsete SNVde annoteerimine“,

mille juhendaja on Ulvi Gerst Talas,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu alates **01.07.2016** kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

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

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

Tartus, 24.05.2016 (*kuupäev*).