TARTU ÜLIKOOL

LOODUS- JA TÄPPISTEADUSTE VALDKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Inimese eksoomis leiduvate valgujärjestust muutvate teadaolevate SNVde võimalik tuvastamine ja genotüübi määramine sekveneerimisandmetest *k*-meeride abil

Bakalaureusetöö

12 EAP

Marlen Timm

Juhendaja PhD Age Brauer

Tartu 2018

Infoleht

Inimese eksoomis leiduvate valgujärjestust muutvate teadaolevate SNVde võimalik tuvastamine ja genotüübi määramine sekveneerimisandmetest *k*-meeride abil

Inimese genoomis leidub suurel hulgal üksiknukleotiidseid variatsioone. Osad nendest variatsioonidest paiknevad eksoomis ja muudavad esialgse valgu funktsiooni, mis võib põhjustada fenotüübilist muutust. Antud bakalaureusetöö eesmärgiks on tutvustada inimese genoomis leiduvate SNVde tuvastamise viise ning analüüsida eksoomis leiduvate uute SNVde rakendamist joondusvabas SNV genotüübi määramise tarkvaras.

Märksõnad: SNV, sekveneerimine, FastGT, *k*-meer

CERCS kood: B110 (bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika)

Calling known protein sequence altering SNVs and genotypes from human exome sequencing data by using *k*-mers

A large number of single nucleotide variants can be found in the human genome. Some of these variants occur in the exome, altering the function of a protein which can lead to a change in phenotype. The purpose of this thesis is to introduce the methods of calling known SNVs in the human genome and to analyze applying known new SNVs in the human exome to an alignment-free SNV genotype calling software.

Key words: SNV, sequencing, FastGT, *k*-mer

CERCS code: B110 (bioinformatics, medical informatics, biomathematics, biometrics)

Sisukord

Kasutatud lühendid

HGP – Inimese genoomi projekt (*The Human Genome Project*)

IHGSC – Rahvusvaheline Inimese Genoomi Sekveneerimiskonsortsium (*The International Human Genome Sequencing Consortium*)

SNV – üksiknukleotiidne variatsioon (*single nucleotide variant*)

SNP – üksiknukleotiidi polümorfism (*single nucleotide polymorphism*)

LD – ahelduse tasakaalustamatus (*linkage disequilibrium*)

DNP – kahe nukleotiidi polümorfism (*dinucleotide polymorphism*)

TNP – kolme nukleotiidi polümorfism (*trinucleotide polymorphism*)

FS – raaminihet põhjustav mutatsioon (*frameshift mutation*)

NFS – raaminihet mitte põhjustav mutatsioon (*non-frameshift mutation*)

LCR – lookuse kontrollregioon (*the locus control region*)

SV – struktuurne variatsioon (*structural variant*)

CNV – koopiaarvu variatsioon (*copy-number variant*)

LoF – funktsioonikaoga mutatsioon (*loss-of-function mutation*)

ddNTP – didesoksünukleotiid (*dideoxynucleotide*)

NGS – teise põlvkonna sekveneerimine (*next-generation sequencing*)

WGS – täisgenoomi sekveneerimine (*whole-genome sequencing*)

WES – täiseksoomi sekveneerimine (*whole-exome sequencing*)

VCF – *Variant Call Format*

Sissejuhatus

Inimese genoom varieerub kahe mittesuguluses oleva indiviidi vahel umbes 0,5%, andes suure panuse inimese fenotüübi kujunemisele. See erinevus indiviidide vahel tuleneb üksiknukleotiidide, indelite ja struktuursete variatsioonide esinemisest. Kõige rohkem leidub genoomis just üksiknukleotiidseid variatsioone, mis võivad geenipiirkondades põhjustada valgu funktsiooni kadu ja mille mõju inimese fenotüübile võib varieeruda.

Teise põlvkonna sekveneerimismeetoditega suudetakse toota suurel hulgal andmeid, mida kasutatakse muuhulgas juba teadaolevate variatsioonide tuvastamiseks. Hetkel laialdaselt kasutuses olevad variatsioonide ja genotüüpide tuvastajad vajavad variatsiooni kindlaks määramiseks lugemi paigutamist referentsgenoomile, mis tõstab variatsiooni tuvastamise protsessi ajakulu. Alternatiiviks on välja töötatud tuvastamise meetodid, mis lugemi paigutamise etappi ei vaja. Üks nendest on FastGT, mis vajab üksiknukleotiidsete variatsioonide tuvastamiseks eelnevalt kokkupandud variatsioonide ja neile vastavate *k*-meeride andmebaasi.

Tuvastatud üksiknukleotiidsete variatsioone hulk suureneb pidevalt. Kuna FastGT suudab tuvastada ainult tarkvarale teadaolevaid variatsioone, tuleks FastGT andmebaasi täiendada. Üks variatsioone sisaldav andmebaas on 60706 inimese eksoomis esinevate indelite ja üksiknukleotiidsete variatsioonide andmebaas ExAC, kus suur osa variatsioonidest esinevad madala sagedusega ning ei pruugi teistes andmebaasides esineda.

Antud bakalaurusetöö eesmärgiks on anda ülevaade inimese genoomis esinevatest variatsioonidest, nende variatsioonide tuvastamise viisidest ning analüüsida ExACi andmebaasis olevate variatsioonide võimalikku rakendamist FastGT tarkvaras.

1. Inimese genoomi varieeruvus

Inimese genoomi varieeruvuse uurimisega on tegeldud mitmeid aastaid, mille käigus leitud info on võimaldanud inimgenoomi süstemaatiliselt tundma õppida ning seeläbi lubanud hinnata ka geneetilise variatsiooni mõju ja osakaalu genoomis. 2001. aastal teatas Rahvusvaheline Inimese Genoomi Sekveneerimiskonsortsium (inglise keeles *The International Human Genome Sequencing Consortium* või IHGSC) Inimese Genoomi projekti raames inimgenoomi esialgsest ülevaatest, mille täiendatud järjestus ehk referentsgenoom, avaldati kolm aastat hiljem, 2004. aastal. Kasutades olemasolevat informatsiooni, jõuti selle projekti käigus näiteks järeldusele, et inimeses leidub umbes 20000-25000 valku kodeerivat geeni (International Human Genome Sequencing Consortium, 2004). Kodeerivad geenid on genoomi osad, millelt transkribeeritakse mRNA ning millelt transleeritakse omakorda valk. Praeguseks teadaolevate valku kodeerivate geenide arv jääb eelnevalt mainitud vahemikku - üles on märgitud 20376 valku kodeerivat geeni [1]. Lisaks suudeti Inimese Genoomi projekti käigus tuvastada erinevaid genoomseid variatsioone, nagu 1,4 miljonit üksiknukleotiidset polümorfismi, ning kirjeldada kordusjärjestusi ja segmentaalseid duplikatsioone (International Human Genome Sequencing Consortium, 2001).

Kuid 2001. aastal IHGSC poolt avaldatud ja 2004. aastal lõpule viidud inimese genoom pandi kokku mitme eri indiviidi järjestustest (International Human Genome Sequencing Consortium, 2001, 2004). Genoomse info kasutamiseks näiteks personaalses meditsiinis on oluline mõista aga üksikinimese tervet diploidset geneetilist koostist, kaasates nukleotiidse järjestuse kõik geneetilised variatsioonid, kuna mõju võivad avaldada alleelide ja alleelipaaride vastastiktoimed, mis mängivad rolli mendeliaalsetes ja keerukates haigustes. Näiteks epistaasi tagajärjel võivad kahe või enama geeni alleelide interaktsioon luua uue fenotüübi, varjutada või muuta alleeli efekti ühes või mitmes lookuses [2]. Seega oli oluline esimene individuaalse inimese diploidne sekveneeritud genoom (Levy *et al*., 2007) ning esimene NGS tehnoloogiaga sekveneeritud inimese genoom (Wheeler *et al*., 2008), mida võib pidada suureks sammuks inimese genoomi resekveneerimises. Just personaalse genoomi põhjal on võimalik uurida alleelide ja geneetiliste variatsioonide omavahelist mõju indiviidile ning seostada seda mõju mõne

haiguse või teise fenotüübilise omadusega. Selleks, et luua seoseid haiguste ja variatsioonide vahel, on eelnevalt vaja uurida inimeste genoomset varieeruvust, õppida neid variatsioone tundma ning tuvastada uusi. Seda kõike on võimaldanud hiljutised edusammud kiires ja järjest kättesaadavamas sekveneerimises (Metzker, 2010), millest räägin lähemalt teises peatükis.

Praeguseks on välja selgitatud, et inimese genoom koosneb umbes kolmest miljardist aluspaarist, millest rohkem kui 99,5% on kahe mitte suguluses oleva inimese vahel identne. Ülejäänu 0,5% DNA järjetusest moodustabki geneetiline variatsioon, mida nimetatakse geneetiliseks varieerumiseks [3]. Geneetiline varieerumine tuleneb nukleotiidsete järjestuste suhtelisest erinevusest genoomide vahel. Lisaks võib varieeruda ka nukleotiidsete järjestusblokkide korraldus (Haraksingh ja Snyder, 2013).

Variatsioonid genoomis esinevad igal indiviidil ja need annavad suure panuse inimese fenotüübilisse erinevusesse. Suur osa variatsioone on healoomulised ja võivad panustada näiteks inimese juuksevärvi (Söchtig *et al*., 2015). Geneetiline variatsioon võib olla aluseks ka kohanemisvõimelistele omadustele. Samuti võib variatsioon muutuda levinuks populatsioonides, kus nad annavad valikulise eelise. Näiteks esineb Aafrika piirkondades, kus malaaria epideemiliselt levib, sirprakulist aneemiat tekitavat alleeli (HbS) rohkem, kuna HbSi esinemine koos normaalse (HbA) alleeliga annab teatud kaitse malaaria vastu (Piel *et al*., 2010; Serjeant, 2013). Sellisel juhul esineb variatsioon geenipiirkonnas. Vastupidiseks eelnevale, kus geneetiline variatsioon toob kasu, võib variatsiooni genotüüp tuua ka kahju: kui indiviidis on kaks HbS alleeli, põhjustab see sirprakulist aneemiat. Lisaks sellele, kas variatsioon määrab indiviidil haiguse tekkimise, võib variatsioon muuta ka ravimi metabolismi, mille tõttu muutub ravimi efekt inimesele. Näiteks tekitab podagra ravis kasutatav allopurinol HLA-B*58:01 genotüübiga inimestel tõsist nahalöövet (Hershfield *et al*., 2013).

Nagu eelnevalt näha, võib geenipiirkondades asuvate variatsioonide mõju tuleneda variatsiooni genotüübist - kas uuritav variatsioon on homosügootses või heterosügootses olekus. Variatsiooni homosügootse genotüübi puhul on kahe alleeli nukleotiidid identsed, kuid heterosügootse genotüübi puhul esinevad alleelidel alternatiivid. Sellest tulenevalt ei pruugi kahjulik või kasutoov alleel mingit mõju avaldada või avaldub mõju väiksemal määral (Serjeant, 2013).

See, milline mõju on geenipiirkondades asuvatel variatsioonidel, võib tuleneda suuresti ka sellest, kui suurel määral muudetakse geeni valgutootmisvõimet või kui palju erineb muudetud valk esialgsest. Variatsioon võib muuta aminohappelist järjestust, põhjustades liigse või liiga vähese valgutootmise võrreldes normaalse geenivariandiga. Kõrvalekaldeid normaalsest geenieksпressiooni tasemest võivad põhjustada geenikoopia tavapärasest madalam või kõrgem arv, mille tagajärjel võivad tekkida haigused (McCarthy ja Mendelsohn, 2016). Näiteks PRSS1 geeni sisaldav umbes 605 kb segmendi kolmekordistamise puhul on täheldatud päriliku pankreatiidi esinemist (Maréchal *et al*., 2006).

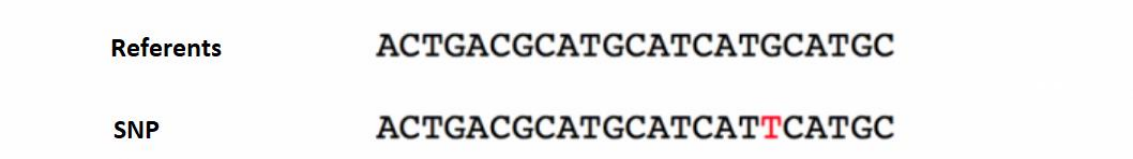
1.2 Geneetiliste variatsioonide tüübid

Inimese geneetiline varieeruvus koosneb nukleotiidsetest muutustest, mille alla kuuluvad üksik- ja mitmenukleotiidsed variatsioonid, lühikesed insertsioonid ja deletsioonid. Lisaks kuuluvad variatsioonide hulka struktuursed muutused: suuremad koopiaarvu variatsioonid ning sarnaste suurustega koopia-neutraalsed inversioonid ja translokatsioonid. Neid geneetilisi variatsioone võib klassifitseerida varieeruvate DNA segmentide olemuse, kaasuvate sündmuste või suuruse alusel, mida mõõdetakse aluspaarides. Segmentide olemuse alusel jagades peegeldab variatsiooni tüüp seda, kas DNA materjal asendati, duplitseeriti, inserteeriti, deleteeriti või paigutati ümber (inversioon, translokatsioon) (Haraksingh ja Snyder, 2013).

Lisaks eelnevale saab geneetilisi variatsioone grupeerida variatsiooni leviku alusel inimpopulatsioonis. Levinud variatsioonid ehk polümorfismid on geneetilised variatsioonid, mille alleeli sagedus populatsioonis on vähemalt 1%. Iga tuvastatud levinud (inglise keeles *common*) või haruldane variatsioon on unikaalselt kirjeldatud selle alusel, milline on geenivariandi järjestus ja kus see vastavalt referentsjärjestusele asub (Frazer *et al*., 2009; Haraksingh ja Snyder, 2013). Geneetilised variatsioonid jagatakse peamiselt kolme suuremasse rühma: üksiknukleotiidsed variatsioonid, väiksemad indelid ning struktuursed variatsioonid [4].

1.2.1 Üksiknukleotiidsed variatsioonid

Üksiknukleotiidsed variatsioonid ehk SNVd (inglise keeles *single nucleotide variation*) (joonis 1), mis kujutavad endast ühe aluspaari muutusi (näiteks C nukleotiid vahetakse samas positsioonis T vastu), on genoomis rohkelt levinud. Haruldasi üksiknukleotiidseid variatsioone on rohkesti, esinedes mõnel juhul ainult tuumperekonnas või üksikus indiviidis. SNVsid, mis esinevad rohkem kui 1%-l näidispopulatsiooni indiviididest, nimetatakse vastavalt üksiknukleotiidi polümorfismiks ehk SNPks (Frazer *et al*., 2009). Tuvastatud on ka kahe nukleotiidi polümorfismid (DNP) ja kolme nukleotiidi polümorfismid (TNP), mille esinemisel on järjestuses muutunud vastavalt kaks või kolm nukleotiidi võrreldes referentsgenoomiga (Rosenfeld *et al*., 2010).



Joonis 1. Üksiknukleotiidi polümorfism. Punasega on märgitud järjestuses muutunud nukleotiid võrreldes referentsgenoomiga (kohandatud joonis [4]).

Osade SNPde pärandumine võib olla teise SNPga korreleerunud ning seda nähtust nimetatakse ahelduse tasakaalustamatuseks (inglise keeles *linkage disequilibrium* ehk LD). See tähendab, et kaks lähestikku asetsevat SNPi päranduvad koos edasi, kui nende kahe variatsiooni vahelisel alal ei toimu rekombinatsiooni, ning ajapikku levivad need alleelid populatsioonis koos (Dudley, 2013).

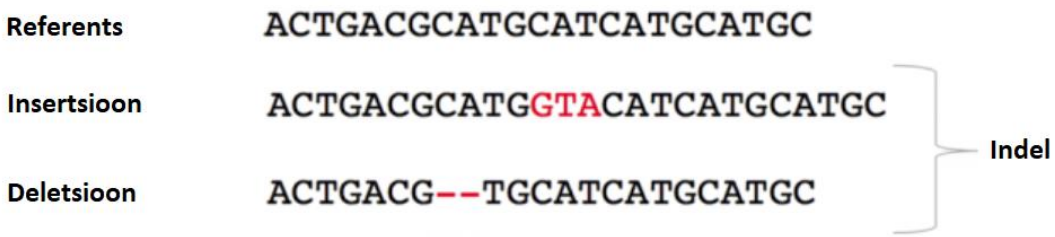
1.2.2 Indelid

Indelid on nukleotiidsed insertsioonid ja deletsioonid (joonis 2), mille pikkused võivad varieeruda olenevalt sellest, kuidas neid klassifitseeritakse, kuid hilisemad uuringud loevad indeliteks pigem kuni 50 aluspaari pikkuseid DNA järjestusi (Lin *et al*., 2017; MacDonald *et al*., 2014; Zarrei *et al*., 2015). Kui indel asub kodeerivas regioonis, toob see kaasa kahte tüüpi muutusi: raaminihet põhjustav (inglise keeles *frameshift* ehk FS) ja raaminihet mitte põhjustav (inglise keeles *non-frameshift* ehk NFS) mutatsioon. NFS indelid koosnevad ühest või mitmest trinukleotiidist, mille inserteerumise või

deleteerumise tulemusena lisatakse või eemaldatakse trinukleotiidi(de) poolt kodeeritud üks või enam aminohapet. Kuna selle käigus lugemisraamis nihet ei toimunud, jääb ülejäänud valgujärjestus muutmata. FS indelite nukleotiidide koguarv ei jagu kolmega ja seega muudavad need indelid alates insertsioonist või deletsioonist lugemisraami, mis võib viia valgujärjestuse muutumiseni või põhjustada enneaegset valgusünteesi terminatsiooni (Lin *et al*., 2017).

Indelid võivad kaasa tuua fenotüübilisi muutusi, põhjustades näiteks värvipimedust. Geenid OPN1LW ja OPNL1MW asuvad tandeemselt X-kromosoomis ning kodeerivad värvide nägemiseks vastavalt punaseid ja rohelisi kolvikesi. Punast ja rohelist pigmente tootvate geenide ekspressiooniks on vajalik LCR regioon (inglise keeles *the locus control region*), mis asub geenide 5’ otsa 3,1 ja 3,7 kb vahel, kuid indiviididel, kellel on selles regioonis deletsioon, puuduvad punased ja rohelised kolvikesed ning nad ei näe seega päevavalguses värve (Deeb, 2005; Nathans *et al*., 1986, 1989).

Variatsioonide tuvastamisega tegelevad mitmed erinevad projektid ning nende raames saadud tulemustest on osa sisse kantud dbSNP andmebaasi, milles on info nii indelite kui eelnevas peatükis räägitud SNPde kohta, lisaks sisaldab see ka lühikesi tandeemseid järjestusi. Praeguse seisuga on dbSNP andmebaasis (versioon 151) kokku nende variatsioonide 660773127 RefSNP klastrit [5].



Joonis 2. Lühikesed indelid. Punasega on märgitud insertsioon ja deletsioon võrreldes referentsgenoomiga (kohandatud joonis [4]).

1.2.3 Struktuursed variatsioonid

Struktuurseid variatsioone (SV) defineeritakse kui genoomseid muutusi, mille DNA segmendi pikkus ületab 50 aluspaari (MacDonald *et al*., 2014). Levinumad struktuursed variatsioonid hõlmavad endas suuremaid insertsioone, deletsioone, DNA järjestuse

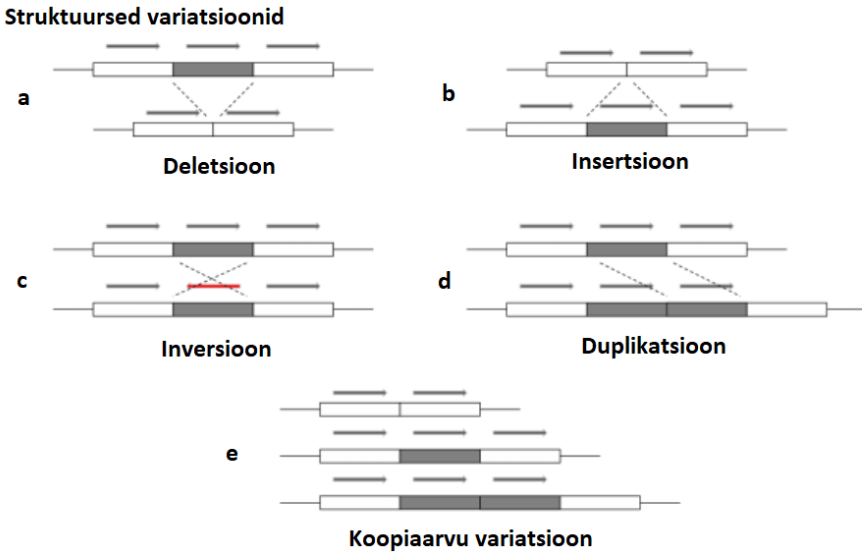
inversioone, duplikatsioone ja koopiaarvu erinevusi [4] (joonis 3). Neid variatsioone on dbVar andmebaasis 34870865 [6].

Duplikatsioonid on DNA segmendid, mille inserteerumisel lisandub järjestusse esialgse järjestusega vähemalt 90% ulatuses identne segment. Duplikatsioonide puhul varieerub tihti järjestuse koopiaarv, mistõttu võivad duplikatsioonid olla ka koopiaarvu variatsioonid (CNV) [7].

Koopiaarvu variatsioonid (inglise keeles *copy number variations*) on genoomi regioonid, mis on kromosoomis duplitseerunud või deleteerunud. Olenevalt genoomist, võib CNVde protsentuaalne osakaal olla 4,8-9,7% (Zarrei *et al*., 2015).

Erinevalt teistest struktuursetest variatsioonidest, on inversioonid sellised ümberkorraldused, mis muudavad DNA segmendi orientatsiooni, kuid ei kaota ega lisa selle käigus geneetilist materjali (Haraksingh ja Snyder, 2013).

Struktuursete variatsioonide pikkused võivad ulatuda ka saja tuhande ja isegi miljoni aluspaarini. Nende hulka kuuluvad muuhulgas ka liikuvad elemendid, mis pärinevad vanadest transponeeruvatest elementidest ja mis püsivad genoomis (Alu elemendid ja LINE'id) (Haraksingh ja Snyder, 2013).



Joonis 3. Sturktuursed variatsioonid. Hallide kastidega on märgitud järjestuses muutunud segmendid (kohandatud joonis [3]).

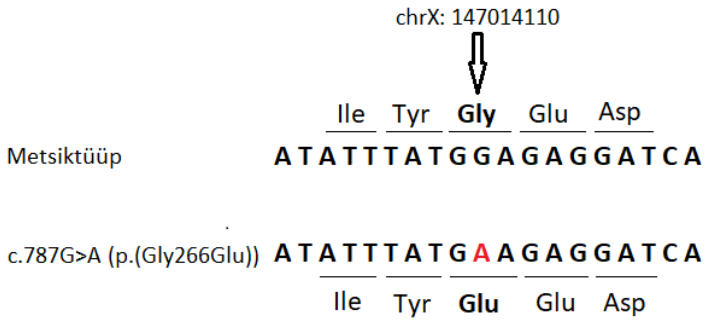
1.3 Funktsioonikaoga mutatsioonid

Variatsioonid võivad esineda genoomis nii kodeerivates kui mittekodeerivates alades. Palju on uuritud just kodeerivates alades asuvaid variatsioone, kuna need võivad muuta esialgse geeni funktsiooni. Just need geneetilised variatsioonid, mis häirivad inimese valku kodeerivate geenide funktsiooni, vähendavad geeniprodukti ehk valgu aktiivsust või hulka ning neid variatsioone nimetatakse funktsioonikaoga mutatsioonideks (inglise keeles *loss-of-function*) ehk LoFideks (Mikelsaar *et al*., 2010). Selliste geneetiliste variatsioonide poolt muudetud transkriptide valgu või mRNA produktid ei suuda täita oma esialgset funktsiooni, mis võib viia produkti lagundamiseni ja see omakorda haigusliku seisundini (Pagel *et al*., 2017). See protsess viib aga haplopuudulikkuseni. See esineb, kui rakus ekspresseeritakse ainult 50% normaalsest aktiivsest valgust. Suur osa valke kodeeritakse autosomaalsete geenide poolt, kus esineb geenil kaks koopiat või teatud alleeli. Kui ühes neis koopias esineb deletsioon või mutatsioon, siis edasist valku ei ekspresseerita või ekspresseeritakse esialgsest valgust erineva funktsiooniga valk (Torgerson ja Ochs, 2014).

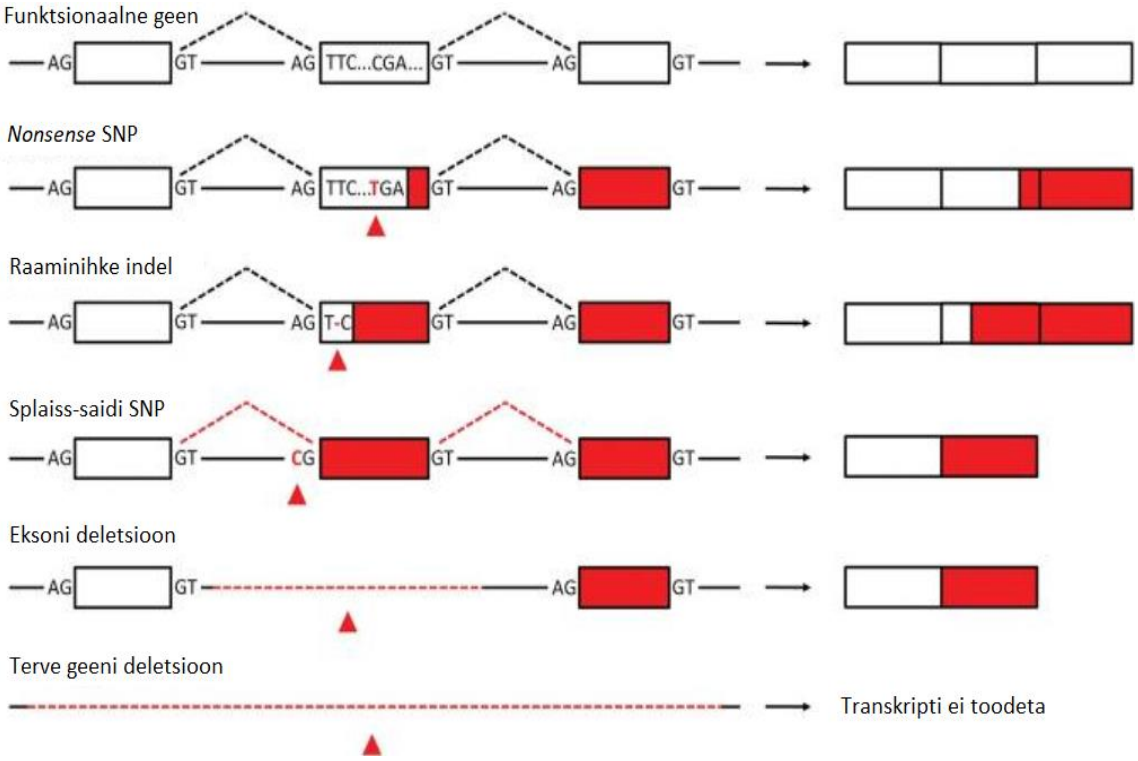
LoF variatsioone on traditsiooniliselt vaadatud kui mendeliaalsete haiguste tekitajaid, kuid tegelikkuses leidub neid ka näiliselt tervetes indiviidides. Näiteks võib LoF mutatsiooni teiste hulgas tekitada ka nonsense SNP, mille tagajärjel transkribeeritakse esialgsest lühem valk. Just seda tüüpi variatsiooni kohta viidi 1151 näiliselt terve indiviidi seas läbi uuring, mille kohaselt varieerus testitud 805-st nonsense SNPst inimeste vahel 169. Lisaks tuvastati, et keskmiselt kandsid indiviidid 14 nonsense/nonsense homosügootset SNPi ja 18 nonsense/normaalne heterosügootset SNPi. (Yngvadottir *et al*., 2009). Nonsense SNP võib samas ka kasu tuua. Näiteks tekib geenis CASP12 variatsiooni tõttu stopp-koodon, mille tagajärjel transkribeeritakse lühem valk. Seda lühemat valku seostatakse suurenenud vastupanuga tõsise sepsise välja arenemisel. Kui variatsioon on homosügootses seisundis, on efekt kõige suurem. Kui variatsiooni üks alleel tekitab lühemat ja teine alleel pikemat valku ehk variatsioon on heterosügootses seisundis, ei pruugi variatsiooni efekt üldse esile tulla või toob kaasa nõrgema efekti (Saleh *et al*., 2004). Lisaks üks levinud LoF mutatsioonidest, mis ei põhjusta haiguslikku fenotüüpi, on näiteks O alleeli esinemine ABO veregrupi antigeenide seas, mis on tekkinud üksiku nukleotiidi deleteerumise tõttu (Yamamoto *et al*., 1990).

LoF variatsiooni mõju inimese fenotüübile varieerub, sõltudes variatsiooni tüübist. Funktsioonikaoga mutatsioonid võivad tekitada tõsiseid haigusi, millest tuleb juttu allpool, kuid leidub ka kergelt kahjulikke variatsioone või nagu eespool mainitud - neutraalseid variatsioone, mis ei mõjuta hädavajalike geenide funktsiooni ega põhjusta haiguslikku fenotüüpi. LoF variandid võivad häirida iga geneetilise elemendi tööd, seda ka mittekodeerivates regulatoorsetes alades (Kleinjan ja van Heyningen, 2005), kuid antud töös keskendun LoF variatsioonidele geenipiirkondades.

LoF variatsioone saab grupeerida vastavalt sellele, millise tagajärje antud mutatsioon kaasa toob (joonis 5). Missense mutatsiooni (inglise keeles *missense mutation*) puhul toob üksiku nukleotiidi asendumine koodoni mittesünonüümses positsioonis kaasa kodeeritava aminohappe asendumise teise aminohappega, mis võib viia eristatava fenotüübini. Näiteks põhjustab üksiku nukleotiidi asendus FMR1 geeni 266. positsioonis fragiilse X sündroomi, vahetades glütsiini glutamiinhappe vastu (joonis 4) (Myrick *et al*., 2014). Samas ei pruugi aminohappe muutus mingit mõju kaasa tuua, kui üks aminohape asendub teise biokeemiliselt sarnase aminohappega (Roth, 2007). Sellisel juhul ei ole tegu LoF variatsiooniga.

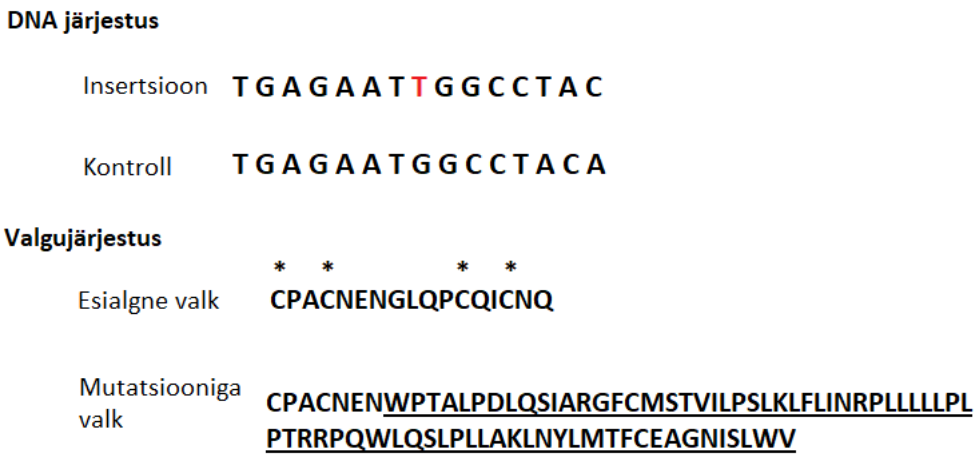


Joonis 4. FMR1 geenis paiknev mutatsioon. DNA lõik FMR1 geenist, kus punasega on märgitud fragiilse X sündroomiga patsiendi missense mutatsioon. G nukleotiid vahetub A vastu, millest tulenevalt asendub glütsiin glutamiinhappega (kohandatud joonis (Myrick *et al*., 2014)).



Joonis 5. Erinevate variatsioonide mõju valku kodeerivatele regioonidele. Mudeliks on kolmest eksonist koosnev funktsionaalne geen. Järgnevad LoF variatsioonid, kus mutatsioon on tähistatud punase kolmnurgaga. LoF variatsioonide tulemiks on valgujärjestuse muutumine variatsioonist allavoolu (joonisel märgitud punaste kastidega) (kohandatud joonis (MacArthur ja Tyler-Smith, 2010)).

Raaminihkemutatsioonid (inglise keeles *frameshift mutation*) on nukleotiidide insertsioonid ja deletsioonid, mille pikkus ei jagu kolmega, põhjustades muutuse mRNA kodeerimisraamis (joonis 5) (Pagel *et al*., 2017). Näiteks raaminihkemutatsioon 5q kromosoomis asuva geeni GRXCR2 3. eksonis põhjustab C-terminuse ebanormaalset pikenemist 63 aminohappe võrra ja seeläbi põhjustab kuulmise kadumist (joonis 6) (Imtiaz *et al*., 2014).



Joonis 6. GRXCR2 geenis paiknev raaminihkemutatsioon. DNA järjestuslõik näitab inserteerunud T-nukleotiidi, mis muudab lugemisraami. Esialgse valgujärjestuse puhul on tärnidega ära toodud konserveerunud kohad. Mutatsiooniga valgujärjestusel on alla joonitud insertsiooni tõttu toimunud valgujärjestuse muutus ja pikenemine (kohandatud joonis (Imtiaz *et al*., 2014)).

Nonsenss mutatsiooni (inglise keeles *nonsense mutation*) puhul muutub järjestuses üks nukleotiid, mis põhjustab enneaegse stoppkoodoni teket ja viib omakorda valgujärjestuse lühenemiseni (joonis 5) (Mikelsaar *et al*., 2010). Kui see mutatsioon esineb ühes geenikoopias, võib see viia haplopuudulikkuseni, mille tõttu väheneb valgu tootmine, mis võib omakorda viia haiguslikku seisundini (Yang *et al*., 2015).

Splaiss-saidi (inglise keeles *splice-site*) variatsoonide puhul esinevad mutatsioonid intronite järjestuses või eksonite kokkupuute kohas (joonis 5). Selle tulemusena võib variatsioon splaiss-doonori või -aktseptori saidi tekitada või ära kaotada (Jameson ja Kopp, 2015). Splaiss-doonori puhul muutuvad 1 või 2 nukleotiidi introni 3’ otsas ning splaiss-aktseptori puhul muutuvad 1 või 2 nukleotiidi introni 5’ otsas (Krawczak *et al*., 1992).

Erinevate variatsioonide hulk geenides on ebaühtlane ja varieerub. Eeldatakse, et tõsiselt kahjulikud mutatsioonid esinevad geenides madalama sagedusega, kuna negatiivne selektsioon eemaldab enamasti evolutsiooni käigus variatsiooni, mis inaktiveerib valku kodeeriva geeni. Samas eeldatakse, et sekveneerimisest tingitud vead on genoomis enam-vähem ühtlaselt jaotunud. Seega on leitud, et sellistes saitides, kus variatsioon võiks

tõsiselt mõjutada geeni valgutootmisvõimet, on polümorfisme vähem kui genoomis keskmiselt. Kuid tuvastatud valepositiivsete variatsioonide hulk on genoomis ühtlane, võib nendes saitides valepositiivseid variatsioone osakaalult rohkem esineda kui näiteks geenides, kus varieeruvus on suurem (MacArthur *et al*., 2012). Järgnevalt tutvustangi, kuidas jõutakse üldse variatsioonide tuvastamiseni.

2. Genoomi sekveneerimismetoodikad

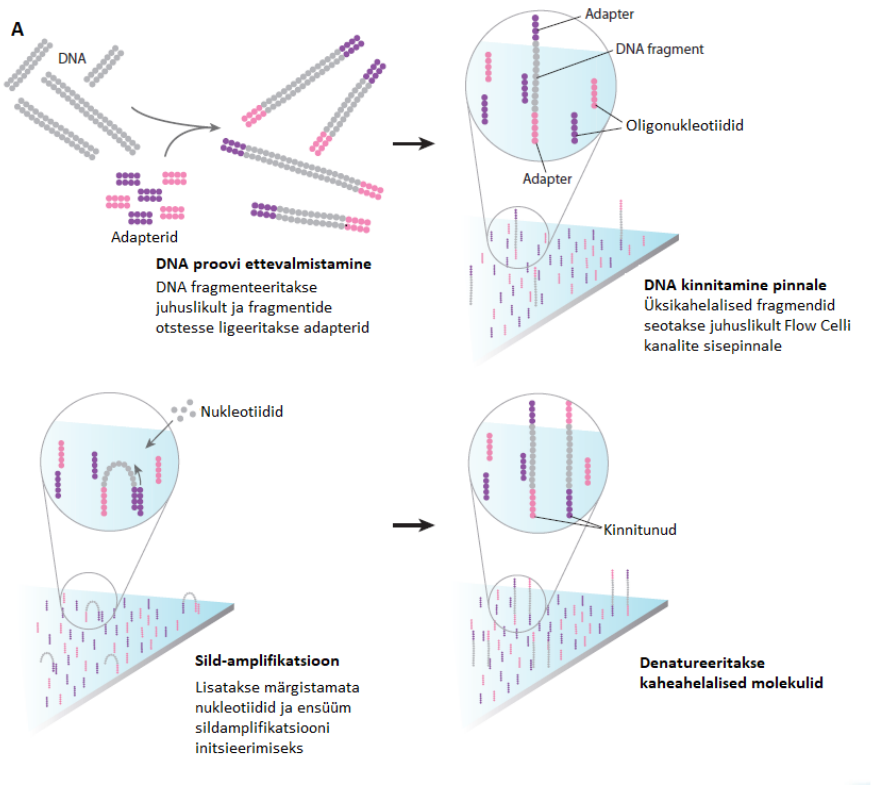
Sekveneerimine on protsess, kus järjestatakse DNA regiooni nukleotiidid. Esimeste inimgenoomide sekveneerimiseks kasutati Sangeri meetodit, mille käigus tekivad didesoksünukleotiidide tõttu erineva pikkusega fragmendid, mis lahutatakse geelelektroforeesil DNA järjestuse määramiseks (Sanger *et al*., 1977). Kuid selle meetodi puhul on puudusteks madal läbilaskevõime ja kõrge hind, mis vähendasid esialgu DNA sekveneerimise potentsiaali selle rakendamiseks näiteks personaalse genoomi sekveneerimises (Reuter *et al*., 2015).

Nende puuduste ületamiseks töötati alternatiiviks välja teise põlvkonna sekveneerimise (NGS) tehnoloogiad, mis on võimaldanud laiaulatuslikku inimese genoomijärjestuse uurimist, sealhulgas variatsioonide ja mutatsioonide tundma õppimist. Suurimad eelised, mida NGS tehnoloogiad kaasa on toonud, seisnevad võimes produtseerida andmeid odavalt ja suurel hulgal, tänu millele saab NGS tehnoloogiaid kasutada genoomide resekveneerimises. Seda saab omakorda kasutada selleks, et parandada meie teadmisi, kuidas geneetiline varieeruvus mõjutab inimeste tervist ja haigusi (Metzker, 2010).

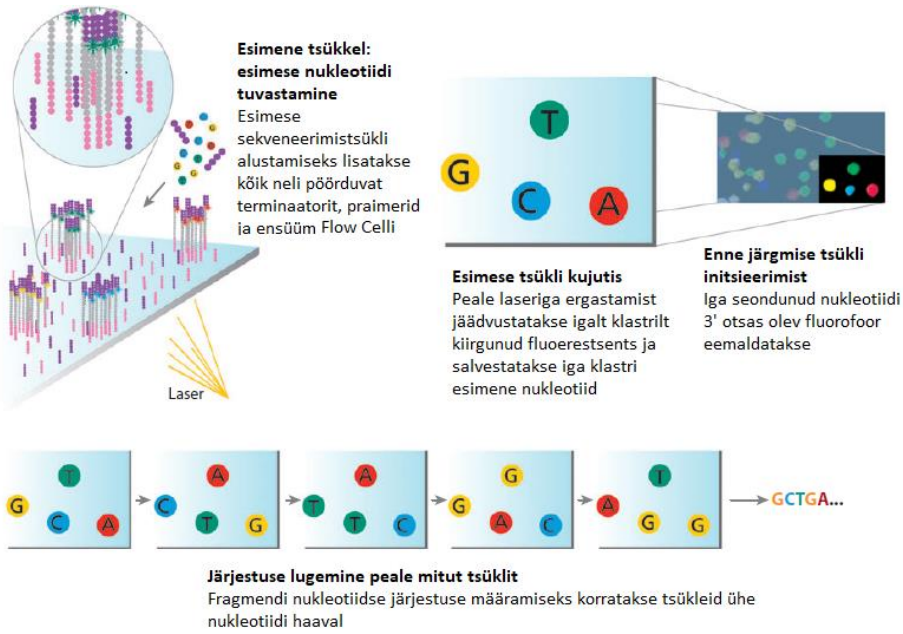
Teise põlvkonna sekveneerimise platvormidest on laialdasemalt kasutuses näiteks Illumina/Solexa, Ion Torrent ja Roche/454, millest igaüks kasutab sekveneerimiseks erinevat meetodit. Sekveneerimises läbitavad etapid võib kõigi platvormide puhul laialdaselt jagada järgnevalt: raamatukogu ettevalmistamine, sekveneerimine ja andmete analüüs (Pfeifer, 2017). Kuigi sekveneerimisplatvorme on mitmeid, domineerib Illumina hetkel kõrge läbilaskvusega järjestamise turgu (Shendure *et al*., 2017).

Illumina platvorm kasutab sünteesi teel sekveneerimise (inglise keeles *sequencing-by-synthesis*) meetodit, mis põhineb pöörduva terminaatori kasutamisel (inglise keeles *reversible terminator chemistry).* Selle puhul immobiliseeritakse

matriitsid adapteritega tahkele kandjale ehk reaktsioonikambri pinnale, milleks Illuminal on Flow Cell. Sellel paiknevad komplementaarsed oligonukleotiidid, millele liidetakse DNA genoomsete fragmentide ükskikahelalised järjestused. Sildamplifikatsiooni tulemusena tekivad Flow Celli põhja DNA molekuli identsete koopiate klastrid (joonis 7). Nende miljonite klastrite järjestamiseks kasutatakse fluorestseeruva märgistusega terminaatornukleotiide, mis võimaldavad pöörduvalt peatada polümeraasreaktsiooni. Iga tsükli jooksul lisandub üks selline terminaatornukleotiid, mille märgistus ära loetakse, seejärel eemaldatakse, et saaks toimuda järgmise nukleotiidi lisamine (joonis 8). Selliseid tsükleid korratakse, kuni matriitsi kõik nukleotiidid on üks haaval fluorestseeruvate terminaatornukleotiididega järjestatud (Bentley *et al.*, 2008). Inimese genoomi mutatsioonide usaldusväärseks tuvastamiseks soovitatakse ühte nukleotiidi sekveneerida 10- kuni 30-kordse katvusega [8]. See tähendab, et ühte nukleotiidi katab 10 kuni 30 fragementi.



Joonis 7. Illumina sekveneerimismeetod. DNA fragmendid seotakse adapterite kaudu Flow Celli pinnale, millele tekivad sildamplifikatsiooni tulemusena klastrid (kohandatud joonis (Mardis, 2008)).



Joonis 8. Järjestamise tsükkel. DNA polümeraas pikendab klastrite ahelaid ükshaaval spetsiaalsete nukleotiididega. Kasutamata nukleotiidid pestakse ära, lisatakse skanneerimiseks vajalik puhver. Peale nukleotiidide lugemist eemaldatakse fluorestseeruv märgis ja lisatakse nukleotiidid uue tsükli käivitamiseks (kohandatud joonis (Mardis, 2008)).

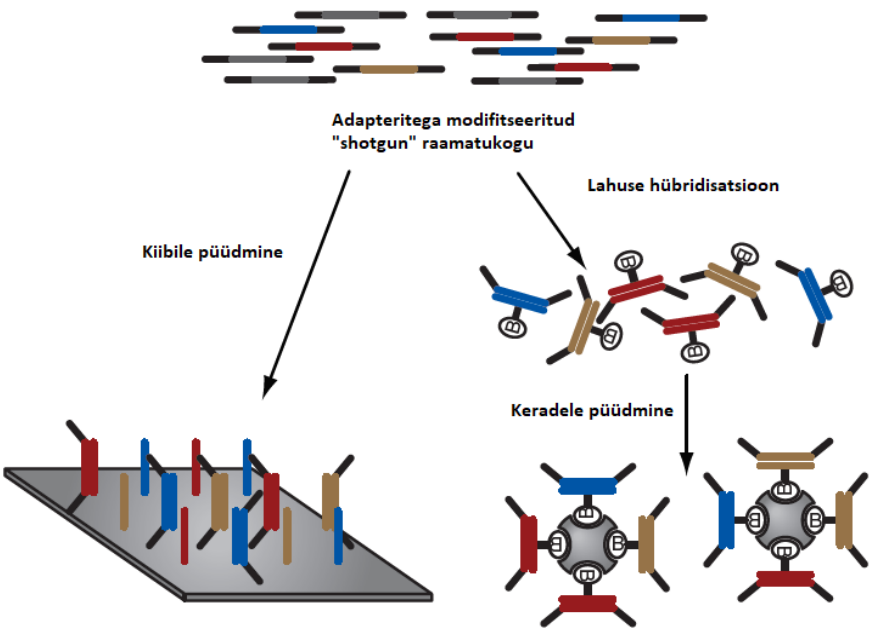
Lisaks Illuminale, kasutab eelnevalt mainitud platvormidest Roche/454 FLX pürosekvenaator emulsioon-PCRi, millele järgneb lutsiferaasi ja ATP sulfurülaasi poolt kiirgunud valguse detekteerimine pikotiiterplaadi aukudes (Margulies *et al*., 2005). Ion Torrent tehnoloogia kasutab sekveneerimiseks pooljuhist kiipi, mis tõlgib geneetilise koodi (A, C, G, T) digitaalseks infoks (0, 1) [9].

2.1 Eksoomi sekveneerimine

Igal platvormil on sekveneerimiseks erinevad tehnoloogiad, mida rakendatakse terve genoomi või teatud genoomsete regioonide sekveneerimiseks. Terve genoomi sekveneerimise (inglise keeles *whole-genome sequencing*) ehk WGSi käigus pannakse paika terve genoomi nukleotiidne järjestus (Warr *et al*., 2015). Täiseksoomi sekveneerimine (inglise keeles *whole-exome sequencing*) ehk WES seisneb valku kodeerivate regioonide ehk eksoomi nukleotiidide järjestamises. Eksoom koosneb valku

kodeerivate geenide kõikidest eksonitest ning katab genoomist vähem kui 2% [10]. Kuigi sekveneerimistehnoloogiad on läbi teinud jõudsa arengu, on täisgenoomi sekveneerimine võrreldes suunatud sekveneerimisega kallis ning seepärast kasutatakse fenotüüpi mõjutavate variatsioonide tuvastamiseks tihti just täiseksoomi sekveneerimist. Selle eeliseks on väiksem hind proovi kohta, suurem katvuse sügavus, väiksemad hoiustamisnõuded ning lihtsam andmete analüüsi teostus (Warr *et al*., 2015).

Eksoomi järjestamisel on kaks põhilist viisi: lahusepõhine ja kiibipõhine meetod (joonis 9). Lahusepõhise meetodi puhul sünteesitakse huvipakkuvale märklaudregioonile vastav spetsiifiline oligonukleotiid. Oligonukleotiidid seotakse helmestega ning mittehuvipakkuvad genoomi regioonid saab lahusest välja pesta. Seejärel kasutatakse polümeraasi ahelraktsiooni (PCR), et huvipakkuv regioon üles amplifitseerida, millele järgneb selle regiooni sekveneerimine. Kiibipõhine meetod on sarnane, aga selle puhul on spetsiifilise oligonukleotiididis seotud kiibi pinnale. Nüüdseks kasutatakse rohkem lahusepõhist meetodit, kuna see vajab vähem DNA algmaterjali (Warr *et al*., 2015).



Joonis 9. Eksoomi sekveneerimise viisid (kohandatud joonis (Mamanova *et al*., 2010).

Rohkem kui 60 000 inimese eksoomi variatsioonide uurimiseks loodud Exome Aggregation Consortiumi (ExAC) projektis, millest räägin lähemalt 4. peatükis, kasutati eksoomide sekveneerimisel Illumina ja Agilent tehnoloogiaid (Lek *et al*., 2016).

3. Sekveneerimise toorandmete analüüs

Variatsioonide tuvastamiseks ei piisa ainult teise põlvkonna sekveneerimistehnoloogiate poolt nukleotiidsete järjestuste kindlaks määramisest. Nende sekveneerimismeetoditega fragmenteeritakse terve genoom või genoomi huvipakkuvad regioonid juhuslikult väikesteks juppideks, mille nukleotiidide järjestamise tulemusena saadakse suur kogus toorandmeid. Selleks, et neid andmeid saaks variatsioonide tuvastamiseks kasutada, tuleb andmeid töödelda. Töötlemise etapid saab jagada kui toorandmete kvaliteedi hindamine, lugemite joondamine referentsgenoomile, variatsioonide tuvastamine ja variatsioonide annoteerimine. (Pabinger *et al*., 2014). On loodud ka metoodikaid, mille puhul eelnevalt teadaolevate variatsioonide tuvastamiseks pole lugemite referentsgenoomile joondamise etapp vajalik. Nendest meetoditest räägin lähemalt 5. peatükis.

3.1 Toorandmete kvaliteedi hindamine

Sekveneerimise järel saadud toorandmetes esinevad tihti vead, mis võivad mõjutada täpset lugemi paigutamist referentsgenoomile, mis omakorda mõjutab variatsiooni ja genotüübi detekteerimise täpsust. Selleks, et suurendada järgnevate analüüside usaldusväärsust, on vaja toorandmeid töödelda. Toorandmetes määratakse igale nukleotiidile Phred skoor, mis näitab vea esinemise tõenäosust antud nukleotiidi puhul (Pfeifer, 2017).

$$Q_{\text{Phred}} = -10 \log_{10} P$$

Kui Phred skoori väärtus on 20, viitab see 1%-lisele veamäärale (Nielsen *et al*., 2011).

Selle skoori määramisega tehakse kindlaks, et sekveneerimisandmetes esineks võimalikult vähe müra. Lisaks vaadatakse, et nukleotiidi katvus oleks piisav. Madala kvaliteediga nukleotiidide puhul rakendatakse kahte meetodit: vea parandamine või madala kvaliteediga regioonide eemaldamine (Pfeifer, 2017). Kui kvaliteediskoor on 30,

tähendab see, et antud nukleotiidi määramise täpsus on 99,9%. Just seda väärtust peetakse NGSi puhul etaloniks, mille alusel määratakse kas nukleotiid on kvaliteetne või ei [11]. Selle kvaliteediskoori määramine vähendab hilisemaid variatsioonide tuvastamise ja genotüüpide määramise vigu (Nielsen *et al*., 2011).

3.2 Lugemite joondamine referentsgenoomile

Variatsioonide ja genotüüpide tuvastamiseks on välja töötatud nii joondusel põhinevaid kui ka joonduse vabasid metoodikaid. Joondusel põhinevate meetodite puhul paigutatakse lugem referentsgenoomis kohale, kust lugem on kõige tõenäolisemalt pärit (inglise keeles *mapping*) ning seejärel tuvastab tarkvara varieeruva koha ja selle genotüübi.

Kõige laialdasemalt kasutatakse spetsiaalselt NGSi jaoks mõeldud programme, mille puhul paigutatakse huvipakkuv järjestus referentsgenoomile (Pfeifer, 2017). Inimese genoomi referentsina kasutatakse Genome Reference Consortiumi poolt assambleeritud referentsgenoomi. Hetkel viimane kokkupandud versioon on GRCh38.p12, mis avaldati 2017. aasta lõpus [12].

Lugemite asukoht leitakse referentsjärjestuse ja lugemi vahel sarnasusi otsides, lubades samal ajal kahe järjestuse vahel üksikuid valestipaardumisi ja tühikuid. Joondamisprogrammid eeldavad valestipaardumiste arvu vastavalt liigi polümorfismide määrale ja tehnoloogia tõttu esinevate vigadele (Reinert *et al*., 2015). Joondamise täpsus mängib andmete edasises kasutamises olulist rolli. Valesti joondatud lugemid võivad viia variatsiooni tuvastamise vigadeni, seega on oluline, et joondamine oleks võimalikult täpne (Pfeifer, 2017).

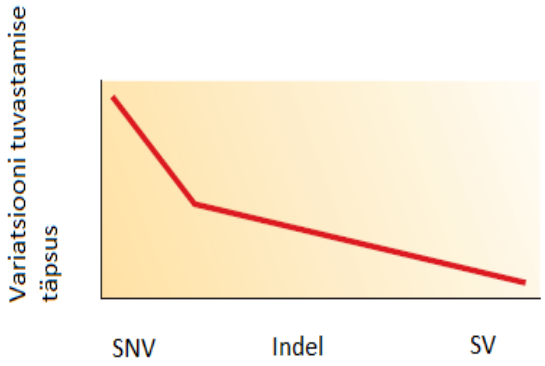
Joondamiseks võib kasutada *seed-and-extend* paradigmat. Selle meetodi puhul tehakse iga lugemi kohta lugemist lühem *k*-meer, mida nimetatakse *seed'*iks. Kui *k*-meer vastab täpselt referentsile, pikendatakse *seed'*i nii paremalt kui vasakult, võttes samas arvesse lubatud maksimaalset valestipaardumiste arvu ja indelite pikkust. Valestipaardumiste lubamiseks kasutatakse *spaced seed'*e. *Seed'*idel on nö mitte-spetsiifiline positsioon „x“, kus algoritm ei kontrolli nukleotiidi. Näiteks *spaced seed* ACGxACG on võimeline sobituma nii ACGAACG kui ACGCACG-ga. See meetod tõstab ka kalkulatsiooni aega.

Selle algoritmi puuduseks on vajadus lugemite lühemate alamsõnede järele, kui tõsta lubatud vigade arvu. Väikesed alamsõned, näiteks 4 nukleotiidi pikad, ei ole aga *seed'*ile vastavuse otsimise faasis efektiivsed, kuna võivad sobida genoomis mitmele erinevale kohale. Sellepärast kasutatakse harva *seed'*e, mis on alla 10 nukleotiidi pikad (Schbath *et al*., 2012).

Teised algoritmid joondavad terveid lugemeid referentsgenoomi põhjal tehtud alamsõnede ehk sõne lühema järjestuse vastu. Selle meetodiga suudetakse läbi viia kiiret lugemi otsingut, hoides referentsgenoomi järjestuste sufikseid sufikspuu kujul (Keel ja Snelling, 2018). Sõne sufiks on alamsõne, mis algab sõne suvalises kohas ja lõpeb sõne lõpuga. Näiteks TTACA on GATTACA sufiks, samas kui TTAC ei ole. Sufikspuu puhul moodustavad kõik sõne sufiksid raja juurest lehtedeni. Sufiksipuu on puu, milles on üks-ühele vastavused juurest lehtedeni ja sufiksid, mis eksisteerivad sõnes, teisisõnu - stringi kõik sufiksid on nagu rada juurest lehtedeni (Schbath *et al*., 2012).

3.3 Variatsioonide tuvastamine

Lugemite referentsgenoomile paigutamisele järgneb varieeruva saidi ehk variatsiooni tuvastamine ja varieeruva saidi alleelide määramine ehk genotüübi tuvastamine (Pfeifer, 2017). Nii täiseksoomi kui täisgenoomi sekveneerimisega on võimalik tuvastada nii SNVsid, indeleid kui struktuurseid variatsioone, kuid tuvastamise täpsus eksoomi sekveneerimise puhul sõltub sellest, millisel määral on avatud lugemisraam häiritud (joonis 10) (Ashley, 2016).



Joonis 10. Variatsiooni tuvastamise täpsuse vähenemine vastavalt avatud lugemisraami muutuse suurusele. (kohandatud joonis (Ashley, 2016)).

On arendatud palju bioinformaatilisi tööriistu, et hõlbustada variatsioonide avastamist NGS andmetest. Varasemad meetodid toimisid lihtsal *cut-off* reeglil: kõrge kvaliteediga alleelid loeti igast saidist üle ning nukleotiidid, mille Phred skoori väärtus oli alla 20, eemaldatati. Kasutades keskmise ja madala katvusega genoomide puhul genotüübi tuvastamist fikseeritud *cut – off'*idel, kaob osa informatsioonist, kuna heterosügootseid genotüüpe ei suudeta piisavalt tuvastada (Nielsen *et al*., 2011). Seega on arendatud tõenäosuslikud meetodid, mis kasutavad kvaliteediskoore, et hinnata iga genotüübi esinemise tõenäosust.

Tõenäosuslike meetodite puhul eeldatakse, et suudetakse arvutada genotüübi tõenäosus p(X | G), genotüübile G. Sümbol X tähistab kõigi lugemite andmeid konkreetse indiviidi kohta konkreetses saidis. Tänu sellele suudetakse genotüüpi määrata suurema täpsusega. Eelmainitud meetod ühendab nukleotiidi tuvastamise, lugemi referentsgenoomile paigutamise ja järjestuste kokku panemise käigus tekkinud vigade info muude andmetega nagu näiteks alleelide esinemissagedused ja aheldustasakaalutuse mustrid (Nielsen *et al*., 2011).

Variatsioonide tuvastamistööriistade väljundiks on VCF fail (inglise keeles *Variant Call Format*), milles hoitakse infot tuvastatud variatsioonide kohta. VCF fail sisaldab päist, mis algab „##“ tähistega ja sisaldab metainformatsiooni, mis seletavad andmetesektsioonis esinevaid lühendeid. Andmetesektsiooniks on teksti formaadis tabel, kus iga veerg kirjeldab antud variatsioonile iseloomulikke omadusi. Iga variatsioon on

kirjeldatud kromosoomiga (CHROM), positsiooniga kromosoomis (POS), unikaalse identifitseerijaga (ID), referentsalleeliga antud positsioonis (REF), alternatiivse alleeliga (ALT), PHRED kvaliteetskooriga (QUAL), saidi filtreerimisinfoga (FILTER) ja semikooloniga eraldatud lisainfoga antud variatsiooni kohta (INFO) (Danecek *et al*., 2011).

3.4 Variatsioonide annoteerimine

Pärast variatsiooni tuvastamist toimub selle annoteerimine, mis võimaldab variatsiooni kasutada edaspidistes uuringutes või haiguse kindlaks määramises. Variatsiooni annoteerimine on protsess, mille käigus kirjeldatakse variatsiooni loomust ja variatsiooni poolt tekitatud DNA muutusega kaasnevat efekti (Eilbeck *et al*., 2017). Kasutades automatiseeritud programme, on NGSi eksperimentide poolt toodetud andmetega võimalik ennustada variatsioonide funktsionaalset mõju. Arvutipõhine annoteerimine võimaldab haigust põhjustavaid mutatsioone filtreerida ja edaspidisteks analüüsideks prioriteediks seada. Suur osa annoteerimistööriistadest keskenduvad just SNPde annoteerimisele (Pabinger *et al*., 2014). Üks sellistest tööriistadest on näiteks VEP, mis ennustab kiirelt ja täpselt variatsioonide mõju Ensemblis olevatele transkriptidele (McLaren *et al*., 2010). Seda tööriista kasutati ka ExACi projekti käigus tuvastatud SNPde annoteerimiseks ning selle käigus saadud info lisati VCF faili (Exome Aggregation Consortium *et al*., 2016).

4. Inimese genoomi variatsioonide andmebaasid

Variatsioonide andmebaasid pannakse kokku erinevate projektide käigus tuvastatud variatsioonide põhjal. Neid andmebaase on mitmeid ja enamasti on igal andmebaasil oma kindel väljund. Näiteks on dbVar andmebaasis info erinevate struktuursete variatsioonide kohta [13], OMIM andmebaas seostab inimeste geneetilisi variatsioone fenotüüpidega [14] samas kui ClinVar andmebaas on keskendunud seoste esitamisele ainult struktuursete variatsioonide ja inimese tervise vahel [15]. Käesolevas töös tutvustan lähemalt kahte andmebaasi – dbSNP ja ExAC andmebaas.

4.1 DbSNP andmebaas

DbSNP andmebaas [16] on välja töötatud National Center of Biotechnology Information (NCBI) poolt koostöös National Human Genome Research Institute'iga (NHGRI). DbSNP sisaldab üksiku nukleotiidi asendusi, lühikesi deletsiooni ja insertsiooni polümorfisme, mikrosatelliit markereid ja polümorfseid insertsioonilisi elemente nagu retrotransposoonid.

Need variatsioonid kogutakse dbSNP andmebaasi erinevatest allikatest, kaasates andmeid individuaalsetest laboritest, suuremahulistest genoomi sekveneerimiskeskustest ja eratööstustest ning samuti sisaldab andmebaas erinevate koostööde käigus avastatud variatsioone.

Variatsioonide jaotus üle genoomi on ebaühtlane - eeldatav minimaalne tihedus on 1/3000 aluspaari kohta juhuslikes genoomijärjestuste regioonides ning kõrgem hästi kirjeldatud geenide ümber (Sherry *et al*., 1999).

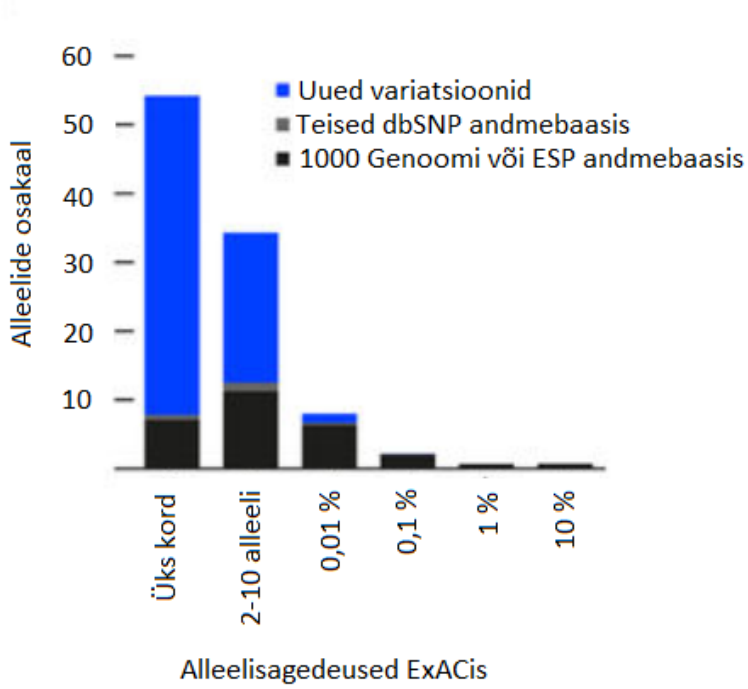
DbSNP andmebaasis on kahte tüüpi variatsioonide esitusi: ss loend iga originaalse sisestuse jaoks ja referents SNP (rs) loend, mis on konstrueeritud algoritmi abil. 22. märtsil avaldati dbSNP versioon 151, mille kohaselt on andmebaasi sisestatud 660773127 RefSNP klastrit, millest on valideeritud 113862023. Geenides on leitud rs-e 381785470 [17].

4.2 ExACi andmebaas

The Exome Aggregation Consortium (ExAC) on loonud andmebaasi, mis sisaldab keskmiselt ühte varianti iga eksoomi kaheksa nukleotiidi kohta ja annab tunnistust suurehulgalisele mutatsioonide olemasolule. Nende variatsioonide tuvastamiseks kasutati toorandmeid 91796 inimese individuaalsetest eksoomidest, mis koguti peamiselt haigustele keskendunud konsortsiumitelt. Sellele esialgsele toorandmete hulgale teostati filtreerimine, mis vähendas andmebaasi loomiseks kasutatavaid eksoome. ExACi lõplik andmebaas sisaldab infot 60706 inimese eksoomidest, mis võimaldab madala sagedusega geneetiliste variatsioonide analüüsil suuremat eraldusvõimet kui varasemad andmebaasid: Exome Variant Server sisaldab informatsiooni 6503 eksoomist (Fu *et al*.,

2013) ja 1000 Genomes Project koosneb 2504 indiviidi täisgenoomi ja eksoomijärjestuse andmetest (Consortium, 2015). ExACi andmed avaldati kahes formaadis: veebibrauseris kasutatava [18] ja allalaaditava andmefailina. Andmebaasile on loodud veebibrauseris kasutajaliides, kust leiavad infot ka need, kellel pole bioinformaatilist tausta. Nagu mainitud, on kogu toorandmestik allalaetav ka VCF failina.

ExACi andmebaas sisaldab pea 7,5 miljonit variatsiooni, millest suur osa esinevad väga madala sagedusega. 99%-l juhtudest on variatsioonide sagedus andmete hulgas väiksem kui 1% ning 54%-i variatsioonidest nähti terve andmete hulga peale ainult üks kord. Kuid nagu eelnevalt mainitud, esinevad variatsioonid kollektiivselt kõrge sagedusega. Suur osa ExACi andmebaasis olevaid geneetilisi variatsioone leiti olevat haruldased ja uued: need puudusid eelnevatest geneetiliste variatsioonide andmebaasidest, nagu näiteks dbSNP (joonis 11).



Joonis 11. Alleelide osakaal ExACi andmebaasis. Sinisega on märgitud variatsioonid, mida teistes andmebaasides ei tuvastatud; halliga variatsioonid, mis esinesid ainult dbSNP andmebaasis ning mustaga variatsioonid, mis esinesid 1000 Genoomi või ESP andmebaasis (kohandatud joonis (Lek *et al.*, 2016)).

Lisaks, tänu ExACi andmebaasi kodeerivates regioonides esinevatele variatsioonide suurele hulgale, tulevad esile inimese geneetilise variatsiooni omadused, mis väiksemates andmekogudes võivad märkamata jääda. Näiteks ExACis tuvastatud kõrge-kvaliteediga saitidest on 7,9% multialleelsed, mis on tunduvalt kõrgem protsent kui eelnevates andmetes leitud: 1000 Genoomi andmebaasis oli see vastavalt 0,48% (Lek *et al*., 2016).

ExACi andmebaasis on välja toodud kõrge-kvaliteediga variatsioonid, mis on kõikide variatsioonide seast teatud kriteeriumite alusel välja valitud. Kõrge-kvaliteediga variatsioonid on kõigepealt läbinud rekalibreerimise ja filtreerimise etapi. Teiseks tehti kindlaks, et nukleotiidi katvuse sügavus on vähemalt 80%-l indiviididest 10 või rohkem ning genotüübi kvaliteediks 20 või rohkem. Seejärel vaadati, et vähemalt ühel indiviidil oleks antud saidis leitud eelnevale punktile vastav alternatiivne alleel ning viimaks ei tohtinud variatsioon asuda nendes genoomi piirkondades, kus on tuvastatud kõrgeim multialleelsus. See tähendab, et lisaks muule, pole kõrge-kvaliteediga variantide hulka arvatud sugukromosoome, välja arvatud pseudoautosomaalsed regioonid, mis päranduvad nagu autosoomsed geenid. Just nendele kriteeriumitele vastavad ExACi andmebaasi 7,4 miljonit kõrge-kvaliteediga variatsiooni, mille hulgast tuvastati 221860 valku lühendavat variatsiooni (inglise keeles *protein-truncating variant* ehk PTV). Selle alla kuuluvad raaminihkemutatsioonid, splaiss-doonor ja splaiss–aktseptor mutatsioonid ning nonsenss mutatsioonid. Filtreerimise käigus eemaldati 42186 variatsiooni, mille tulemusena jäi järgi 179774 variatsiooni, mida peetakse kõrge usaldusväärsusega PTVdeks. Nendest 121309 variatsiooni nähti andmete jooksul kusjuures ainult üks kord. Keskmiselt on ExACi andmetel igal inimesel 85,1 heterosügootses olekus ja 34,2 homosügootses olekus PTVd (Exome Aggregation Consortium *et al*., 2016).

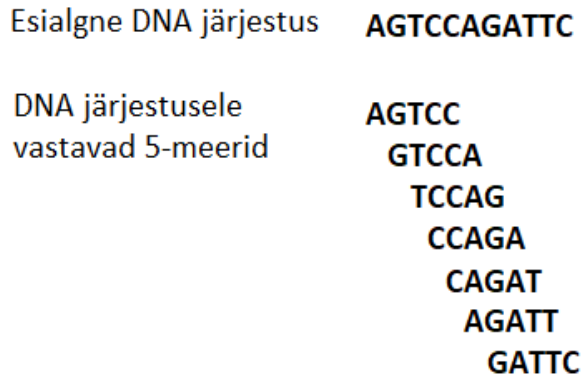
Lisaks eelnevale uuriti ka geenide tundlikkust variatsioonidele, mille käigus leiti 3230 geeni, mis olid eelnevalt mainitud 179774 kõrge-usaldusväärsusega variatsioonile kõrgelt tundlikud. Geenide tundlikkuse hindamiseks jagati geenid kõigepealt kolme kategooriasse: null-kategooria, kus PTVsid tolereeritakse nii heterosügootses kui homosügootses seisundis, retsessiivne-kategooria, kus tolereeritakse ainult heterosügootses seisundis PTVsid, ning haplopuudulikkuse-kategooria, kus ei tolereerita heterosügootseid ega ka homosügootseid PTVsid. Neis 3230 kõrgelt tundlikkus geenis esinevatele 179774 variatsioonile pole 72%-le omistatud haigust tekitavat fenotüüpi OMIM andmebaasis. Lisaks pole neid variatsioone täheldatud inimese geneetiliste

mutatsioonide hulgas ClinVari andmebaasis (Exome Aggregation Consortium *et al*., 2016).

Kõik tuvastatud variatsioonid esitati ka VCF faili kujul, kus on lisaks funktsionaalsetele üksiku nukleotiidi muutustele ära toodud ka need variatsioonid, mille tulemusena on tekkinud LoF mutatsioon. ExAC loeb LoFideks nonsenss („stop_gained“), splaiss-aktseptor („splice_acceptor_variant“) ja splaiss-doonor („splice_donor_variant“) ühenukleotiidsed variatsioonid (Exome Aggregation Consortium *et al*., 2016).

5. Teadaolevate variatsioonide tuvastamine joondusvabade meetoditega

Kuigi joonduspõhised meetodid on variatsioonide tuvastamiseks kõige rohkem kasutatud, on välja töötatud ka alternatiivsed joondusvabad meetodid, mida saab kasutada eelnevalt teadaolevate variatsioonide kiiremaks tuvastamiseks. Üheks alternatiivseks meetodiks on FastGT, mis loeb FASTQ-formaadis NGS andmetelt unikaalsete *k*-meeride sagedusi ja kasutab seda informatsiooni, et määrata teadaolevate varieeruvate saitide genotüüpi (Pajuste *et al*., 2017). FastGTs kasutatatavd *k*-meerid on k nukleotiidi pikkused oligomeerid (joonis 12).



Joonis 12. Näitlik joonis 11 nukleotiidi pikkusele DNA järjestusele vastavatest 5-meerist.

Inimese genoomis leidub kohti, mille puhul on variatsioonide tuvastamine raskendatud, sellisteks kohtadeks on näiteks genoomsed kordusjärjestused. FastGT ei määra hetkel variatsioone nendest keerulistest kohtadest. See joondusvaba meetod on umbes 1-2 suurusjärku kiirem kui traditsioonilisel joondamisel põhinevad genotüüpide detekteerimise meetodid (Pajuste *et al*., 2017).

Variatsioonide tuvastamine FastGTga toimub *k*-meeride kattumisel variatsiooniga ja on võimalik tänu eelnevalt kokku pandud SNVde andmebaasile ning neile vastavatele *k*-meeride paaridele. Iga bi-alleelne SNV positsioon on kaetud *k k*-meeri paariga, kus paar moodustub kahele alternatiivsele alleelile vastavast *k*-meerist (Pajuste *et al*., 2017). Kuna meetod ei vaja lugemite paigutamist referentsgenoomile, on see traditsioonilistest meetoditest 1-2 suurusjärku kiirem. FastGT on hetkel orienteeritud ainult SNVde tuvastamisele ning suudab praegu tuvastada 30238283 SNVd. Kasutusel olev *k*-meeride list sisaldab *k*-meere valideeritud levinuid (*common*) bi-alleelsetele SNVdele dbSNP andmebaasist. Lisaks on FastGTl olemas eraldi *k*-meeride andmestik ka eksoomi puudutavate SNVde kohta, kuid see on dbSNPi alamhulk ning ei hõlma hetkel kõiki (kliiniliselt) olulisi SNVsid. Praegu on saadaval dbSNP versioon 151, kuid FastGT *k*-meerid on disainitud SNVdele, mis pärinevad andmebaasi versioonist 146. Selles andmebaasis on üles märgitud 46954719 valideeritud ja levinud bi-alleelset SNPi, mida kasutati FastGT andmebaasi loomiseks (Pajuste *et al*., 2017). Antud versiooni põhjal loodud FastGT andmebaas ei sisalda aga kõiki eelnevalt mainitud ExACi andmebaasis olevaid variatsioone. Võrreldes ExACi ja FastGT andmebaase, tõin välja andmebaasides kattuvad ja FastGTs puuduolevad varitasioonid (tabel 1).

Tabel 1. Variatsioonide kattuvus ExACi ja FastGT andmebaasides. Tabelis on ära toodud erinevate variatsioonide hulgad ExACi andmebaasis ning esitatud variatsioonide arv, mis kahes andmebaasis kattuvad ning milline hulk on FastGT andmebaasist ExACi põhjal puudu.

Variatsioon	ExACis	Olemas nii FastGTs kui ExACis	Puuduolevad
Missense	3564452	198932	3365520
Raaminihke	164343	1525	162818
Stopp	115756	4406	111350
Aktseptor	33668	1285	32383
Doonor	38917	1171	37746
Kokku	3917136	207319	3709817

FastGT tuvastab variatsioone *k*-meeridega ning toetub eeldusele, et vähemalt mingi arv nendest SNVdele vastavatest *k*-meeri paaridest on unikaalsed ja esinevad ainult selles genoomi kohas, seega võib unikaalsete *k*-meeride paaride esinemiste arvu sekveneerimisandmetes kasutada inimese variatsiooni genotüübi määramiseks (Pajuste *et al*., 2017). Selleks, et kasutusel olevad *k*-meeride paarid oleks võimelised SNVsid tuvastama, filtreeriti dbSNP andmebaasi SNVsid. Eemaldati üksteisele liiga lähedal asuvad SNVd, kuna ühe *k*-meeri pikkuse kohta võib esineda ainult üks SNV. Variatsiooni tuvastamiseks võib sellele vastav *k*-meer esineda andmebaasis ainult üks kord, seega eemaldati SNVd, mille puhul eelnev täidetud polnud. Viimaks jäeti välja ka need SNVd, mille puhul täheldati ebanormaalseid tulemusi (Pajuste *et al*., 2017). Võrdlesin antud filtreerimisetappide variatsioone ExACis olemasolevate variatsioonidega ning tõin tabelis 2 arvuliselt välja erinevates etappides välja pruugitud ja alles jäänud variatsioonide

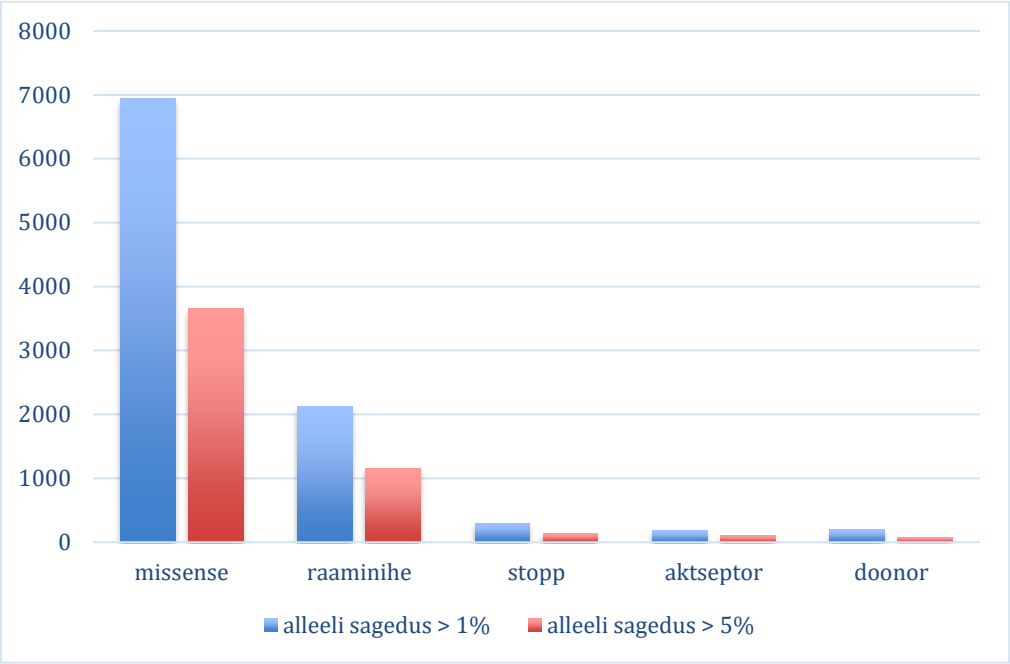
hulgad. Hetkel on ExACis olemasolevate variatsioonidega võrdlemata ja tabelis välja toomata esimeses filtreerimisetapis välja jäänud ehk üksteisele liiga lähedal asetsevate SNVde arv.

Tabel 2. ExACi ja FastGT andmebaaside kattuvate variatsioonide välja jäämine erinevates etappides. Tabelis on välja toodud FastGT andmebaasi kokku panemisel erinevate filtreerimissammude läbimisel alles jäänud variatsioonid, mis on olemas ka ExACi andmebaasis. Teises tulbas on kattuvate variatsioonide alamhulgad, millest erinevate filtreerimissammude käigus eemaldati variatsioonid, mida *k*-meeridega polnud võimalik tuvastada.

Variatsioon	dbSNP	Unikaalsed *k*-meerid SNVdele	Genotüüpide põhjal filtreeritud
Missense	198932	87319	67355
Raaminihke	1525	294	219
Stopp	4406	1597	1285
Aktseptor	1285	423	319
Doonor	1171	600	451
Kokku:	207319	90233	69629

FastGT andmebaasi loomiseks kasutati dbSNP andmebaasi annoteeritud ja levinud SNVsid, kuid dbSNP andmebaasi lisatakse annoteeritud SNVsid järjest juurde. Vastupidiselt dbSNPle on ExACi andmebaasis välja toodud ka väga haruldased variatsioonid, millest osasid nähti terve andmestiku jooksul ainult ühel korral ja mida mujal andmebaasides ei leidu. Võrreldes ExACi ja FastGT andmebaase tuvastasin, et

ExACis leidub ka variatsioone, mille sagedus oli andmestiku hulgas suurem kui 1% ning mida ei leidu FastGT andmebaasi loomiseks kasutatud dbSNP andmebaasi versioonis 146 või mis ei olnud selleks hetkeks annoteeritud (joonis 12).



Joonis 12. FastGTs puuduvate alleelide sagedused ExACi andmebaasis. Sinine tulp kujutab erinevate variatsioonide hulka, kui alleelisagedus on suurem kui 1% ning punane tulp variatsioonide hulka, kui alleelisagedus on suurem kui 5%.

Lisaks FastGTle on välja töötatud ka teisi joondusvabasid meetodeid SNVde tuvastamiseks. Kimura ja Koike poolt välja töötatud meetod tuvastab SNVde genotüübi. Selle puhul teisendatakse lühikese lugemi andmed sõnastikuks, võimaldades lugemi fragmente samaaegselt töödelda (Kimura ja Koike, 2015).

Kolmas kiire joondusvaba meetod teadaolevate SNPde genotüübi tuvastamiseks *k*-meeride abil kannab nime LAVA. See meetod tuvastab etteantud SNPde kogumi hulgast, kas SNP on metsiktüüpi või mutantne sobitades 32-meerid kahepoolselt järjestusele (Shajii *et al*., 2016).

6. Arutelu

Inimese genoomis leidub erinevaid variatsioone, mis annavad panuse fenotüübilisse varieerumisse. Nii SNVd, indelid kui ka struktuursed variatsioonid võivad põhjustada haigusi või teisi fenotüübilisi erinevusi, seega on oluline kõikide variatsioonide tuvastamist ja uurimist jätkata. Suurt tähelepanu on pööratud aga just SNVdele, kuna nende tuvastamine indiviidides on täpsem kui suurtemate variatsioonide puhul, neid on palju annoteeritud ja neile on omistatud erinevaid fenotüübilisi mõjusid. Lisaks võib andmebaasides leida kõige rohkem infot just SNPde kohta. Variatsioone on võimalik tuvastada nii täisgenoomi kui täiseksoomi sekveneerimisega, kuid eksoomi sekveneerimise madalama hinna ja väiksemate hoiustamisnõuete tõttu kasutatakse tuvastamiseks tihti just täiseksoomi sekveneerimist.

Variatsiooni genotüübi tuvastamise traditsioonilised meetodid põhinevad joondusel. Nii neil, traditsioonilistel, kui joondusvabadel tarkvaradel on omad plussid ja miinused. Teadaolevalt ei ole läbi viidud võrdlust nende kahe meetodi vahel, mis võimaldaks selgelt eristada nende tugevaid ja nõrku kohti - see võiks tulevikus olla kindlasti üks uurimisobjektidest. Kuid on kindel, et FastGT eeliseks traditsiooniliste meetodite ees on selle kiirus, olles joondusel põhinevatest meetoditest 1-2 suurusjärku kiirem.

FastGT suudab määrata ainult eelnevalt loodud andmebaasis olemasolevaid SNVde genotüüpe. Selle tarkvara andmebaasi kokku panemiseks kasutati dbSNP andmebaasi versiooni 146, milles oli 46954719 valideeritud ja levinud bi-alleelset SNPi, millest loodi *k*-meerid 30238283 SNVle. Kuid SNVsid tuvastatakse järjest enam, mille tõttu tuleks andmebaasi täiendada, et ka FastGT oleks võimeline määrama võimalikult paljude erinevate kliiniliselt oluliste variatsioonide genotüüpe. Täiendamiseks võiks kasutada ExACi andmebaasi, mis sisaldab 60706 inimese eksoomidest tuvastatud SNVsid ja indeleid, kuna just eksoomides paiknevad variatsioonid põhjustavad tihti kliiniliselt olulisi fenotüüpe.

Suur osa variatsioonidest on madala alleelisagedusega, mille tõttu ei pruugi neid variatsioone dbSNPs esineda. Sellest tulenevalt tuleks luua ExACi andmebaasi SNVdele *k*-meerid, et neid saaks rakendada FastGT tarkvaras SNV genotüübi tuvastamiseks. On võimalik, et kliiniliselt olulisemad ja LoFi põhjustavad ExACi SNVd on FastGT

k-meeride andmebaasi loomiseks kasutatud dbSNPs juba olemas. Seega tuleks kõigepealt kontrollida, kas ja kui paljudele ExACi andmebaasi SNVdele on FastGT *k*-meeride andmebaasis *k*-meeride paarid juba olemas.

Kuna suur osa ExACi variatsioonidest esinevad andmete vältel väiksema sagedusega kui 1% või tuvastati variatsioon andmete hulgast ainult ühel korral, tuleks andmete hulga rohkuse tõttu variatsioone esialgu prioritiseerida. Välja võiks valida variatsioonid, millel on tuvastatud teatud mõju. Näiteks mendeliaalsete haiguste kontekstis on olulised just funktsioonikaoga mutatsioonid, seega võiks LoF variatsioonidele keskendumine olla üheks prioritiseerimise meetodiks. ExAC andmebaasis defineeritakse funktsioonikaoga mutatsioonidena ühenukleotiidsed nonsense, splaiss-aktseptor ja splaiss-doonor variatsioonid.

Ka raaminihke mutatsioonide tulemuseks võib olla valgu funktsiooni kadu. Raaminihke mutatsioone leidus VCF failis 182855. Kuigi FastGT tarkvara oleks tõenäoliselt võimeline tuvastama ka indeleid, kui neile disainida sobivad *k*-meeride paarid, ei sisalda FastGT *k*-meeride andmebaas hetkel indelite jaoks loodud *k*-meere. Indelitele *k*-meeride loomine oleks üheks perspektiiviks, kuid esialgu tasub protsessi lihtsustamise mõttes keskenduda SNVdele. Lisaks eelnevatele mutatsioonidele võivad esialgse valgu funktsiooni muuta ka missense mutatsioonid, mille puhul toimub valgujärjestuses aminohappe muutus. Kuid aminohappe vahetumine näiteks teise biokeemiliselt sarnase aminohappega ei pruugi fenotüüpi mõjutada. Seega kaasates valimisse ka missense mutatsioonid, tuleks variatsioone filtreerida veel kliinilise olulisuse põhjal, et tuvastada millistel variatsioonidel on suurem kliiniline tähtsus ja millised variatsioonid nii olulised ei ole. Kuid selleks tuleks esmalt nendele variatsioonide puhul kindlaks teha, kas ja kui palju antud variatsioon üldse valku muudab.

Filtreerides välja SNVd selle alusel, millise mõju SNV kaasa toob, jääb 7,4 miljonist variatsioonist alles üle 3 miljoni, millest enamikku nähti terve andmestiku jooksul ainult ühel korral. Seega võiks järgmiseks välja filtreerida variatsioonid, mille alleelisagedus on ExACi andmetel üle 1%. Selliseid missense, nonsense, raaminihke ning splaiss-doonor ja –aktseptor mutatsioone leidus ExACis 9761.

k-meerid võiks luua seega ühenukleotiidsetele variatsioonidele, mida FastGT veel tuvastada ei suuda ja mis muudavad valgu esialgset funktsiooni. Seejärel tuleb

analoogselt dbSNP andmebaasi SNVdele ka ExACist valitud SNVdele rakendada samad filtreerimisetapid. Välja tuleb valida SNVd, mis asuvad üksteisest piisaval kaugusel, igale SNVle peab olemas olema unikaalne *k*-meeri paar ning SNV ei tohi anda autosoomides haploidsed ning mehe X ja Y kromosoomis diploidset genotüüpi.

FastGT ei vaja variatsiooni tuvastamiseks lugemi paigutamist referentsgenoomile, tänu millele on FastGT eeliseks traditsiooniliste meetodite ees selle kiirus, olles joondusel põhinevatest meetoditest 1-2 suurusjärku kiirem. Seega lisades FastGTsse eksoomis leiduvad eelkõige kliiniliselt olulised SNVd, võimaldab see meetod meil kiiresti määrata variatsioonide genotüüpe ja nende poolt võimalikke tekitatud haigusi. Kui neile variatsioonidele on omistatud fenotüübilised omadused, on võimalik läbi viia kliinilisi uuringuid.

FastGTga on võimalik määrata mendeliaalseid haigusi põhjustavaid SNVde genotüüpe juhul, kui on tehtud kindlaks, milline variatsioon mingit haigust põhjustab. Samuti leiaks variatsiooni genotüübi tuvastamine rakendust sarnaste sümptomitega haiguste kindlaks määramisel. Kiiret genotüübi määramist saab kasutada haiguste tuvastamiseks ka lootediagnostika käigus.

Lisaks kasutatakse hetkel ühe haiguse raviks kõigi indiviidide puhul enamasti sama ravimit, kuid kõikides inimestes esinevad variatsioonid, mille genotüüp võib määrata kui efektiivselt ja kas üldse ravim indiviidile mõjub. Rakendades kliiniliselt oluliste eksoomis leiduvate SNVde tuvastamist personaalses meditsiinis, saab genoomsest variatsioonist lähtudes inimesele määrata võimalikult efektiivse ravimi ja selle doosi. Selleks tuleb tuvastada eelnevalt annoteeritud ravimit mõjutava variatsiooni genotüüp.

Nagu mainitud - et eelnevat tulevikus kliinilistes uuringutes rakendada, peab olema tuvastatud variatsiooni mõju inimesele. Hetkel on fenotüübiline info puudu 72% ExACi variatsioonidest.

Kokkuvõte

Inimese geneetiline varieeruvus koosneb üksiknukleotiidsetest variatsioonidest, indelitest ja struktuursetest variatsioonidest, mille hulgast on palju uuritud just neid üksiknukleotiidseid variatsioone, mis esinevad kodeerivates järjestustes ning mis muudavad esialgse valgu funktsiooni. Valgu funktsiooni kadu võivad SNVdest põhjustada missense, nonsense, raaminihke, splaiss-aktseptor ja –doonor mutatsioonid, mis võivad indiviidis kaasa tuua fenotüübilise muutuse või haigusliku seisundi. Nii neid kui teisi variatsioone genoomis on võimalik tuvastada joondusel põhinevate kui ka joondusvabade meetoditega. Viimase alla kuulub joondusel põhinevatest meetoditest 1-2 suurusjärku kiirem SNV genotüübi tuvastamise tarkvara FastGT. Seda saab kasutada esmase analüüsi meetodina, kuna suudab variatsiooni genotüüpi kiiremini tuvastada. Kasutades seda tarkvara paralleelselt traditsiooniliste meetoditega, võimaldaks see tõenäoliselt genotüübi määramise usaldusväärsuse tõusu, kuid selle taseme hindamiseks tuleks läbi viia täiendavaid analüüse.

Variatsioonide genotüübi määramisel on oluline roll kliinilistes uuringutes, tänu millele on võimalik tuvastada variatsioonide poolt tekitatud haigusi. Lisaks saab variatsioonide tuvastamist rakendada lootediagnostikas ning ravimi mõju hindamisel inimesele, kui variatsioonidele suudetakse määrata fenotüübilised mõjud.

Selleks, et FastGT oleks suuteline SNVde genotüüpe määrama, tuleb eelnevalt luua neile SNVdele vastav *k*-meeride andmebaas, mille põhjal genotüüpe määratakse. FastGT andmebaasi loomiseks kasutati dbSNP andmebaasi versioon 146s leiduvaid valideeritud ja levinuid bi-alleelseid SNVsid. Kuid kuna variatsioone tuvastatakse järjest enam, oleks vaja antud andmebaasi täiendada. Selleks sobiks üle 60000 inimese eksoomidest koosnev SNVde ja indelite andmebaas ExAC, milles on tuvastatud hulgaliselt valgu funktsiooni muutvaid variatsioone. Hetkel keskendun ainult SNVde lisamisele FastGTsse, kuid üheks perspektiiviks oleks ka indelitele *k*-meeride disainimine, mis võimaldaks FastGTl tuvastada ka indeleid.

ExAC-is on tuvastatud üle 7,4 miljoni variatsiooni, millest üle 3,7 miljoni moodustavad missense, nonsense, raaminihke, splaiss-doonor ja –aktseptor variatsioonid, mida ei võetud FastGT esialgse andmebaasi loomises arvesse. Suur osa neist variatsioonidest olid

väga haruldased. Eelnevalt mainitud 3,7 miljonist variatsioonist oli >1% alleelisagedusega 9761 variatsiooni. Just nende variatsioonidega võiks andmebaasi täiendamist alustada.

Loss-of-function SNV and genotype calling from human exome sequencing data by using *k*-mers

Marlen Timm

Summary

Human genetic variation is a term which refers to single nucleotide variants, indels and structural variants. Amongst these, single nucleotide variants which occur in the coding regions and alter the function of a protein are studied a lot. Loss of protein function can be caused by missense, nonsense, frameshift, splice-acceptor and splice-donor mutations which can lead to a change in human phenotype or give way to a disease. All of these and other variations in the genome can be detected by alignement-dependant and alignement-free methods. FastGT is one of the alignement-free genotype calling programmes which is 1-2 orders of magnitude faster than traditional alignement-dependant methods. It has potential to be used as a primary analysis method because of its' ability to call genotypes faster. When using this software in parallel with traditional methods it could improve the credibility of genotype calling. But in order to confirm this statement additional test should be performed.

Genotype calling plays an important role in clinical analysis. Due to this, disease causing variations can be detected and the possible impact of a variation can be prevented. In addition, genotype calling can be applied in prenatal diagnostic testing. Furthermore, when variations are associated with certain phenotypes it is possible to evaluate the effect of a drug to an individual.

In order for FastGT to be able to call SNV genotypes a compatible *k*-mer database must be assembled. Previously assembled database is the foundation of FastGT genotype calling. For putting together this database validated and bi-allelic SNVs from dbSNP version 146 were used. But as more and more new variations are being detected there is a need to complement the FastGT database. ExAC SNV and indel database which has data from over 60000 human exomes including a large number of protein altering variations would be suitable for this. One of the perspectives is to create *k*-mers for indels which would allow FastGT to also detect indels but as for now I focus on adding only new SNVs.

ExAC has over 7.4 million variants of which over 3.7 million are missense, nonsense, frameshift, splice-donor and splice-donor variants which were not included in the creation of FastGT database. A large number of these variations were very rare. From the previously mentioned 3.7 million variations 9761 variations had an allele frequency greater than 1%. For a start these are the variations which should be used to complement the FastGT database.

Kasutatud kirjandus

Ashley, E.A. (2016). Towards precision medicine. Nat. Rev. Genet. *17*, 507–522.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., *et al*. (2008). Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature *456*, 53-59.

Consortium, T. 1000 G.P. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

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.

Deeb, S.S. (2005). The molecular basis of variation in human color vision. Clin. Genet. *67*, 369–377.

Dudley, J.T. (2013). A gentle introduction to genomics, p 19. Oxford University Press.

Eilbeck, K., Quinlan, A., ja Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. Nat. Rev. Genet. *18*, 599–612.

Exome Aggregation Consortium, Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., *et al*. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Frazer, K.A., Murray, S.S., Schork, N.J., ja Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. *10*, 241–251.

Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., *et al*. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.

Haraksingh, R.R., ja Snyder, M.P. (2013). Impacts of Variation in the Human Genome on Gene Regulation. J. Mol. Biol. *425*, 3970–3977.

Hershfield, M.S., Callaghan, J.T., Tassaneeyakul, W., Mushiroda, T., Thorn, C.F., Klein, T.E., ja Lee, M.T.M. (2013). Clinical Pharmacogenetics Implementation Consortium Guidelines for Human Leukocyte Antigen-B Genotype and Allopurinol Dosing. Clin. Pharmacol. Ther. *93*, 153–158.

Imtiaz, A., Kohrman, D.C., ja Naz, S. (2014). A frameshift mutation in GRXCR2 causes recessively inherited hearing loss. Hum. Mutat. *35*, 618–624.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jameson, J.L., ja Kopp, P. (2015). Genes, the Environment, and Disease. In Harrison's Principles of Internal Medicine, D. Kasper, A. Fauci, S. Hauser, D. Longo, J.L. Jameson, ja J. Loscalzo, eds. (New York, NY: McGraw-Hill Education), p.

Keel, B.N., ja Snelling, W.M. (2018). Comparison of Burrows-Wheeler Transform-Based Mapping Algorithms Used in High-Throughput Whole-Genome Sequencing: Application to Illumina Data for Livestock Genomes1. Front. Genet. *9*.

Kimura, K., ja Koike, A. (2015). Ultrafast SNP analysis using the Burrows–Wheeler transform of short-read data. Bioinformatics *31*, 1577–1583.

Kleinjan, D.A. ja Heyningen, V. von. (2005). Long-Range Control of Gene Expression: Emerging Mechanisms and Distruption in Disease. Am. J. Hum. Genet. 76, 8-32.

Krawczak, M., Reiss, J., ja Cooper, D.N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. Hum. Genet. *90*, 41–54.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al*. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., *et al*. (2007). The Diploid Genome Sequence of an Individual Human. PLOS Biol. *5*, e254.

Lin, M., Whitmire, S., Chen, J., Farrel, A., Shi, X., ja Guo, J. (2017). Effects of short indels on protein structure and function in human genomes. Sci. Rep. *7*, 9313.

MacArthur, D.G., ja Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Hum. Mol. Genet. *19*, R125–R130.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., *et al*. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science *335*, 823–828.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., ja Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992.

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., ja Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.

Mardis, E.R. (2008). Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet. *9*, 387–402.

Maréchal, C.L., Masson, E., Chen, J.-M., Morel, F., Ruszniewski, P., Levy, P., ja Férec, C. (2006). Hereditary pancreatitis caused by triplication of the trypsinogen locus. Nat. Genet. *38*, 1372–1374.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., *et al*. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

McCarthy, J.J. ja Mendelsohn, B.A. (2016). Appendix 1: The Human Genome and Genetic Variation. *In* Precision Medicine: A Guide to Genomics in Clinical Practice, McGraw-Hill Education, New York, NY.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., ja 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.

Mikelsaar, A-V. (2010). Pärilikkusmeditsiin, p. 66-68, AS Medicina, Tallinn.

Myrick, L.K., Nakamoto-Kinoshita, M., Lindor, N.M., Kirmani, S., Cheng, X., ja Warren, S.T. (2014). Fragile X syndrome due to a missense mutation. Eur. J. Hum. Genet. *22*, 1185–1189.

Nathans, J., Thomas, D., ja Hogness, D.S. (1986). Molecular Genetics of Human Color Vision: Science *232*, 193–202.

Nathans, J., Davenport, C., Maumenee, I., Lewis, R., Hejtmancik, J., Litt, M., Lovrien, E., Weleber, R., Bachynski, B., Zwas, F., *et al*. (1989). Molecular genetics of human blue cone monochromacy. Science *245*, 831–838.

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

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

Pagel, K.A., Pejaver, V., Lin, G.N., Nam, H.-J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., Mooney, S.D., ja Radivojac, P. (2017). When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinformatics *33*, i389–i398.

Pajuste, F.-D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., ja Remm, M. (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. Sci. Rep. *7*, 2537.

Pfeifer, S.P. (2017). From next-generation resequencing reads to a high-quality variant data set. Heredity *118*, 111–124.

Piel, F.B., Patil, A.P., Howes, R.E., Nyangiri, O.A., Gething, P.W., Williams, T.N., Weatherall, D.J., ja Hay, S.I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat. Commun. *1*, 104.

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

Reuter, J.A., Spacek, D., ja Snyder, M.P. (2015). High-Throughput Sequencing Technologies. Mol. Cell *58*, 586–597.

Rosenfeld, J.A., Malhotra, A.K., ja Lencz, T. (2010). Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing. Nucleic Acids Res. *38*, 6102–6111.

Roth, S.M. (2007). Genetics Primer for Exercise Science and Health (Human Kinetics).

Saleh, M., Vaillancourt, J.P., Graham, R.K., Huyck, M., Srinivasula, S.M., Alnemri, E.S., Steinberg, M.H., Nolan, V., Baldwin, C.T., Hotchkiss, R.S., *et al*. (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. Nature *429*, 75–79.

Sanger, F., Nicklen, S., ja Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., ja Gibrat, J.-F. (2012). Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. J. Comput. Biol. *19*, 796–813.

Serjeant, G.R. (2013). The Natural History of Sickle Cell Disease. Cold Spring Harb. Perspect. Med. *3*, a011783.

Shajii, A., Yorukoglu, D., William Yu, Y., ja Berger, B. (2016). Fast genotyping of known SNPs through approximate k-mer matching. Bioinformatics *32*, i538–i544.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., ja Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. Nature *550*, 345–353.

Sherry, S.T., Ward, M., ja Sirotkin, K. (1999). dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. *9*, 677-679.

Söchtig, J., Phillips, C., Maroñas, O., Gómez-Tato, A., Cruz, R., Alvarez-Dios, J., Cal, M.-Á.C. de, Ruiz, Y., Reich, K., Fondevila, M., *et al*. (2015). Exploration of SNP variants affecting hair colour prediction in Europeans. Int. J. Legal Med. *129*, 963–975.

Torgerson, T. ja Ochs, H. (2014). Genetics of Primary Immune Deficiencies. In: Stiehm's Immune Deficiencies 1st Edition. Academic Press. p 73-81.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., ja Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. G3 GenesGenomesGenetics *5*, 1543-1550.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., *et al*. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Yamamoto, F., Clausen, H., White, T., Marken, J., ja Hakomori, S. (1990). Molecular genetic basis of the histo-blood group ABO system. Nature *345*, 229–233.

Yang, J.–., Wu, X.–., Dou, T.–., Jiao, T., Chen, X.–., Min, M., Cai, S.–., ja Zheng, M. (2015). Haploinsufficiency caused by a nonsense mutation in NCSTN underlying hidradenitis suppurativa in a Chinese family. Clin. Exp. Dermatol. *40*, 916–919.

Yngvadottir, B., Xue, Y., Searle, S., Hunt, S., Delgado, M., Morrison, J., Whittaker, P., Deloukas, P., ja Tyler-Smith, C. (2009). A Genome-wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs. Am. J. Hum. Genet. *84*, 224-234.

Zarrei, M., MacDonald, J.R., Merico, D., ja Scherer, S.W. (2015). A copy number variation map of the human genome. Nat. Rev. Genet. *16*, 172–183.

Kasutatud veebiaadressid

[1] https://www.ensembl.org/Homo_sapiens/Info/Annotation kasutatud 09.08

[2] https://www.nature.com/scitable/topicpage/epistasis-gene-interaction-and-phenotype-effects-460 kasutatud 09.08

[3] https://www.genome.gov/10001551/genetic-variation-program/ kasutatud 09.08

[4] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/what-genetic-variation/types-genetic-variation kasutatud 09.08

[5] https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi kasutatud 10.08

[6] https://www.ncbi.nlm.nih.gov/dbvar/content/ftp_manifest/ kasutatud 10.08

[7] https://www.ncbi.nlm.nih.gov/dbvar/content/overview/#ref2 kasutatud 10.08

[8] https://emea.illumina.com/science/education/sequencing-coverage.html?langsel=/ee/ kasutatud 10.08

[9] https://www.thermofisher.com/ee/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html kasutatud 10.08

[10]https://emea.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/exome-sequencing.html?langsel=/ee/ kasutatud 10.08

[11] https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf kasutatud 10.08

[12] https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38 kasutatud 10.08

[13] https://www.ncbi.nlm.nih.gov/dbvar kasutatud 10.08

[14] https://www.omim.org/ kasutatud 10.08

[15] https://www.ncbi.nlm.nih.gov/clinvar/ kasutatud 10.08

[16] https://www.ncbi.nlm.nih.gov/SNP/ kasutatud 10.08

[17]https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=151 kasutatud 10.08

[18] http://exac.broadinstitute.org/ kasutatud 10.08

Lihtlitsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Marlen Timm (29.11.1996),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Inimese eksoomis leiduvate valgujärjestust muutvate teadaolevate SNVde võimalik tuvastamine ja genotüübi määramine sekveneerimisandmetest *k*-meeride abil“ mille juhendaja on Age Brauer

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

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

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

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

Tartus, 13.08.2018