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

Kelli Grand

Funktsioonikaoga mutatsioonide analüüs 2300 inimese genoomi ja terviseandmete põhjal

Magistritöö
Geenitehnoloogia eriala, 30 EAP

Juhendajad: Lili Milani, Ph.D
Prof. Pärt Peterson, Ph.D

TARTU 2016

INFOLEHT

Funktsioonikaoga mutatsioonide analüüs 2300 inimese genoomi ja terviseandmete põhjal

Lühikokkuvõte: Alates esimesest täisgenoomi sekveneerimisest on järjestatud mitmeid tuhandeid genoome üle terve maailma, mis on andnud uusi teadmisi genoomi mitmekesisusest. Struktuursete ja koopiaarvu variatsioonide ning ühenukleotiidsete polümorfismide efekt fenotüübile on suurem kui siiani arvati. Käesoleva magistritöö eesmärgiks oli anda ülevaade seni populatsioonide täisgenoomide järjestamise projektide kohta, võrrelda üldtulemusi Tartu Ülikooli Eesti Geenivaramu geenidoonori sekveneerimisandmetega ning iseloomustada funktsioonikaoga mutatsioone Eesti populatsioonis. Veel üheks eesmärgiks oli hinnata biopanga andmete kasutamisvõimalust geneetiliste seoste leidmisel analüüsides väljavalitud immuungeenidest saadud funktsioonikaoga mutatsioone ja geenidoonorite terviseandmeid. Töö tulemusena identifitseerisime seitse potentsiaalset geneetilist varianti edasiseks analüüsiks.

Märksõnad: Täisgenoomide sekveneerimine, funktsioonikaoga mutatsioon, immuunsüsteem, elektrooniline terviseandmestik, populatsioonigenoomika
CERCS: B220 Geneetika, tütogeneetika

Analysing loss-of-function mutations by pairing 2300 whole genomes with electronic health records

Abstract: Sequencing genetic material has become accessible for large-scale population analysis. Genomic data and its grown volume has illustrated that the effect of different structural and copy number variations has been underestimated. The purpose of this study was to give an overview of population based whole-genome sequencing studies done so far, compare the results with data from the Estonian Genome Center of the University of Tartu, and to characterize loss-of-function (LoF) mutations in the Estonian population. Another aim of the study was to evaluate the effect of the identified rare variants by investigating the electronic heath records of the individuals. As a first study, we focused on genes related to the immune system and highlight seven genes with potential LoF variants for further analysis.

Keywords: Whole-genome sequencing, loss-of-function mutation, immune system, electronic health records, population genomics
CERCS: B220 Genetics, cytogenetics

SISUKORD

KASUTATUD LÜHENDID

AC – alternatiivse alleeli koguarv, inglise keeles *alleel count*

AF – alleeli sagedus, inglise keeles *alleel frequency*

Alt – alternatiivne järjestus

AN – alleeli üldarv

BGI – Beijing genoomika instituut, inglise keeles *Beijing Genomics Institute*

BioVu – Vanderbilt DNA andmepank

CNV – koopiaarvu muutus, inglise keeles *copy-number variant*

dbSNP – ühenukleotiidsete polümorfismide andmebaas, inglise keeles *SNP database*

EHR – elektroonilised terviseandmed, inglise keeles *electronical health rekord*

eMERGE – elektroonilise terviseandmete ja genoomika võrgusti, inglise keeles *Electronic Medical Records and Genomics Network*)

ExAC – eksoomivariantide andmebaas, inglise keeles *the Exome Aggregation Consortium*

GATK – genoomi analüüsi tööriist, inglise keeles *Genome Analysis Toolkit*

GoNL – Hollandi täisgenoomi projekt, inglise keeles *Genome of the Netherlands*

GWAS – ülegenoomsed assotsiatsiooniuuringud, inglise keeles *genoome-wide association studies*

Het – heterosügoot

HGMD – inimese geenimutatsiooni andmebaasis, inglise keeles *the Human Gene Mutation Database*

HIV – inimese immuunpuudulikkuse viirus, inglise keeles *Human immunodeficiency virus*

Hom – homosügoot

ICD-10 – 10. haiguste ja terviseprobleemide rahvusvahelise statistilised klassifikatsioonid, inglise keeles *International Statistical Classification of Diseases and Related Health Problems*

IFN – interferoon

LoF – funktsioonikaoga mutatsioon, inglise keeles *loss-of-function*

MAF – minoorse alleeli sagedus, inglise keeles *minor alleel frequency*

MHC – koesobivuskompleks, inglise keeles *major histocompatibility complex*)

OMIM – mendeliaarsete haiguste andmebaas, inglise keeles *online Mendelian Inheritance in Man*

Ref – referentsjärjestus

SISu – Soomlaste täisgenoomi projekt, inglise keeles *the Sequencing Initiative Suomi*

SNP – ühenukleotiidiline polümorfism, inglise keeles *single nucleotide polymorfism*

SNV – ühenukleotiidiline variatsioon, inglise keeles *single nucleotide variation*

UK10K – Suurbritannia ja Põhja-Iiri Ühendkuningriigi 10 000 genoomi projekt

VEP – variantide efekti ennustustööriist, inglise keeles *Variant Effect Predictor*

SISSEJUHATUS

Alates 2008. aastast kui sekveneeriti esimesed personaalsed genoomid Illumina tehnoloogiaga (Bentley jt., 2008), on DNA järjestamine muutunud laialdaselt kättesaadavamaks, võimaldades võrrelda sadade ja tuhandete inimeste täisgenoome. Indiviidi või populatsiooni tasandil leitavate genoomi variatsioonide iseloomustamine annab meile informatsiooni inimkonna ajaloo ja struktuuri kohta, kirjeldab erinevates genoomipiirkondades toimuvat looduslikku valikut ning võimaldab uurida, kuidas mõjutavad geneetilised variatsioonid inimese organismi.

Haruldase ja harvaesinevate variantide panus fenotüübile on suuresti teadmata. Sellised variandid on alaesindatud seni enim kasutust leidnud ülegenoomsetes assotsiatsiooniuuringutes. Olles enamasti populatsiooni- või indiviidpõhised, on vaja harvade variantide avastamiseks suure hulga inimeste DNA sekveneerimist. Haruldaste leidude bioloogiliste protsessidega seostamine võib olla väga keerukas. Mutatsiooni efekti meditsiinilisel interpreteerimisel tuginetakse mutatsioonide esinemiste alleelisagedustele, tõeseks hindamiseks on tarvis geneetilisi variatsioone laiaulatuslikult sisaldavat referentsjärjestust.

Aastakümneid on teadlased uurinud geenide talitlust neid loommudelites inaktiveerides. Suurte valimite sekveneerimine on näidanud, et iga inimese genoomis leidub sadakond geeni funktsiooni rikkuvat varianti ehk looduses naturaalselt esinevaid inaktiivseid geeninokaut mudeleid. Sellised leiud on väärtuslik informatsiooni geeni bioloogilise funktsiooni kohta. Populatsioonide analüüs võimaldab koguda informatsiooni selliste inimnokautide kohta eri maailma paigus ning kõrvutades massilised sekveneerimisandmed varieeruvate tervisenäitajatega, saame hinnata geeni mittefunktsioneerimise bioloogilist tagajärge.

Käesoleva magistritöö eesmärgiks on anda ülevaade seni läbiviidud populatsiooni täisgenoomide projektide kohta, võrrelda neid üldisemalt Eesti Geenivaramu sekveneerimisprojektiga ning täpsemalt iseloomustada funktsioonikaoga mutatsioone Eesti populatsioonis. Lisaks hinnata biopanga andmete kasutamisvõimalust geneetiliste seoste leidmisel, kõrvutades Eesti Haigekassa terviseandmeid immuungeenidest saadud variantidega.

1 KIRJANDUSE ÜLEVAADE

Populatsioonide sarnasuste ja erinevuste uurimine on alguse saanud juba Darwinist. Esimesed meetodid geneetilise varieeruvuse uurimiseks olid geel-elektroforees ning restriktaasiga lõikamine, mida kasutatakse ka tänapäeval väiksemamahuliste analüüside korral. Viimaste aastate nii bio- kui ka infotehnoloogilised uuendused on aga viinud bioloogilise revolutsioonini – võimalus suuremahuliselt lugeda ükskõik milliste organismide DNA järjestust ning analüüsida nende bioloogilist tähendust. Teise põlvkonna sekveneerimine avas inimgenoomide sekveneerimise ajastu.

Maailma kõige esimene inimese täisgenoom avaldati 2001. aastal (Lander jt., 2001). Üle kümne aasta kestnud projekt kaardistas küll oodatust vähem geene aga tunduvalt rohkem geneetilist variatsiooni. Täisgenoomide andmete analüüs leiab veel avastamata seoseid genotüübi ja fenotüübi vahel, aitab mõista bioloogilisi protsesse ning efektiivsemalt kaardistada haiguspõhjuslikke seoseid.

1.1 Täisgenoomide analüüs

Iga inimese genoomist või leida miljoneid üksiknukleotiidseid polümorfisme (SNP) (“The International HapMap Project,” 2003), insertsioon- ja deletsioonmutatsioone (Mills, 2006) ning erinevates pikkustes koopiaarvu muutuseid (CNV) (Redon jt., 2006). Kõik eelpool väljatoodud mutatsioonid ja variatsioonid osalevad inimestevahelise geneetilise mitmekesisuse tagamisel. Genoomsete variatsioonide uurimine võimaldab inimkonnal mõista erinevaid bioloogilisi radu, ravimiresistentsusi ning haigusseoseliste mutatsioonide efekti. Inimese genoomis leiduvate sagedaste geneetiliste variantide kaardistamiseks on loodud rahvusvaheline HapMap projekt (ingl. *The International HapMap Project*) (“The International HapMap Project,” 2003).

Kõige sagedasem geneetiline variatsioon inimese genoomis on SNP – DNA üksiknukleotiidsest muutusest põhjustatud polümorfism. SNP-de sagedus on keskmiselt üks polümorfism 200-300 aluspaari kohta, mis teeb 10 miljonit ühes genoomis (Heinaru, 2012). Lühikeste geneetiliste variatsiooni andmebaas (dbSNP) on avalikkusele kättesaadav lühikeste geneetiliste järjestuste kogum, mille alleelisagedus on piisavalt kõrge, et arvata neid

polümorfseks (minoorse alleeli sagedus ehk MAF >1%) (Sherry jt., 2001). Andmebaasis nüüdseks üle 100 miljoni leiu (tabel 1). Kodeerivas alas võib SNP ümber kujundada transleeritava valgu aminohappelist järjestust (mittesünonüümne SNP) ning tagajärjeks põhjustada terve valgu struktuurilist muutust, kuid nukleotiidi muutus ei pruugi alati põhjustada asendust aminohappe tasandil (sünonüümne SNP).

Tabel 1. dbSNP andmebaasi andmestik, versioon 146 (november 2015)

Organism	**dbSNP versioon**	**Genoomi versioon**	**Kõik sissekanded**	**RefSNP (valideeritud)**	**RefSNP geenides**
Homo sapiens	146	GRCh38.p2	538 341 120	150 482 731 (100 135 281)	87 339 846

Indel (tuletatud inglisekeelsetest sõnadest *INsertion/DELetion*) tähistab lühikeste, vähemalt ühe aluspaari pikkuste insertsioonide või deletsioonide klassi. Indelite sagedus ei ole nii kõrge kui SNP-del, kuid neid leiab üle terve genoomi. 1000 genoomi projekti raames leiti 15 miljonit geneetilist varianti, millest indelid moodustasid ühe viiendiku (Durbin jt., 2010). Kui SNP mutatsioonid ei pruugi alati muuta aminohappelist järjestust, siis indelite korral on mõjutused genoomile suuremad, põhjustades tihtipeale raaminihkemutatsioone. Indelite pikkused võivad olla varieeruvad, mikroindeliteks peetakse 1-50 bp pikkuseid järjestusi (Gonzalez jt., 2007).

Inimese genoomis on geneetilised modifikatsioonid mitmel kujul. Deletsioonid, duplikatsioonid, insertsioonid ning translokatsioonid kõik võivad põhjustada koopiaarvu muutuseid. CNV-d võivad olla mõne kuni mitmekümne tuhande aluspaari pikkused ning seega mõjutada terveid DNA segmente (Redon jt., 2006). Ühes genoomis võib leiduda üle 1000 CNV, mille kogupikkus võib ulatuda 1%-ni genoomi suurusest ning mõjutada keskmiselt 73-87 geeni koopiaarvu (Alkan jt., 2009; Conrad jt., 2010).

Geneetiliste variatsioonide leidmine on mitmete teadlaste huviorbiidis, et mõista inimeste fenotüübilisi tunnuseid ning seeläbi ka leida haigusseoselisi variante. Täisgenoomide uurimine annab teadlastele arusaamise kuivõrd tolerantne on inimese genoom erinevatele deleterioossetele mutatsioonidele. Siinkohal on oluline ka haruldaste, populatsioonipõhiste variantide iseloomustamine, mis võivad vastutada harvaesinevate haiguste ning populatsioonivaheliste fenotüübiliste erinevuste eest.

1.1.1 1000 genoomi projekt

1000 genoomi projekt (aastatel 2008–2015) oli esimene suuremahuline inimgenoomide sekveneerimisprojekt, mille eesmärk oli kaardistada suurel hulgal geneetilisi variatsioone. Nende siht oli koostada detailne inimese geneetilise variatsiooni kataloog, mida laiem teadlaskond assotsiatsiooniuuringutel kasutada saaks. Teiseks suureks eesmärgiks oli uuendada ning parandada inimese referentsjärjestust. Valim koosnes viiest (aafrika, ameerika, ida-aasia, euroopa, lõuna-aasia) suuremast maailmajaost pärit indiviididest ning kokku analüüsiti mitme etapi vältel 2500 inimese geneetilist varieeruvust (Durbin jt., 2010). Projekti võib jagada kahte faasi. Pilootuuringud võrdlesid potentsiaalseid sekveneerimis- ja analüüsivõimalusi. Esiteks sekveneeriti kahe perekonna triod, teiseks analüüsiti 179 indiviidi täisgenoomid ja kolmandaks vaadati 697 inimese tuhatkond juhuslikult valitud kodeerivat geeni (kokku 1,4 Mb genoomist). Pilootprojektiks kasutatud proovid saadi HapMap kollektsioonist. Täisprojekti läbiviimiseks kasutati vabatahtlikke doonoreid ning lõplik andmebaas sisaldab 2504 indiviidi täisgenoomi andmeid (Sudmant jt., 2015).

Projekti tulemusena kirjeldatakse 68 818 struktuurset varianti ja luuakse seni suurim avaliku variantide ning genotüübi informatsiooni kataloog (Sudmant jt., 2015). Andmestikku kasutatakse igapäevaselt referentsjärjestusena, näiteks ülegenoomsetes assotsiatsiooniuuringutes puuduolevate markerite imputeerimisel. 1000 genoomi projekti võib pidada teerajajaks inimese genotüübi suuremahulistele uurimisprojektidele

1.1.2 Populatsiooni genoomika

1.1.2.1 GoNL

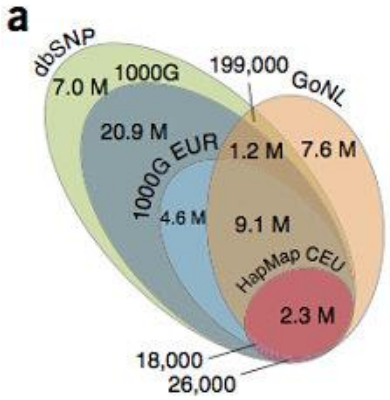
(ingl. *Genome of the Netherlands*)

Kõige esimene populatsioonipõhine täisgenoomide järjestamise projekt teostati hollandlaste poolt eesmärgiga iseloomustada rahvuse geneetilist varieeruvust. Viie erineva biopanga koostöös sekveneeriti 769 indiviidi kogugenoom. Suurem osa valimist moodustasid isa-ema-laps triod (n=231). Valimis olid ka 19 ühe – või kahemuna kaksikutega peret. Kõik indiviidid olid täiskasvanud vanuses 19-87 aastat (keskmine vanus 53 aastat). Sekveneerimiseks vajalik geneetiline materjal eraldati verest, sekveenerimine Illumina HiSeq 2000 platvormil (Illumina Inc., San Diego, CA, USA) teostati Beijing genoomika instituudis (BGI). Analüüsimiseks kasutati 1000 genoomi projekti parimat praktikat koostöös Broad instituudiga (Broad Institute

of MIT and Harvard, Cambridge, MA, USA) (tabel 2). Järjestuste joondamiseks kasutati Burrows-Wheeler Aligner algoritmi (BWA) (H. Li ja Durbin, 2009) ning mrsFAST tööriista (Hach jt., 2014). (Francioli jt., 2014)

Perekondliku uuringu disain võimaldas grupil analüüsida ka *de novo* ehk uusi mitte vanematelt päritud mutatsioone. Tuvastati 11 020 sellist mutatsiooni. Neist uutest mutatsioonidest 74% on põhjustatud isapoolsest geneetilisest materjalist ning korreleeruvad isa vanusega. Ainult kolm protsenti *de novo* leidudest on somaatilised. Uuringu läbiviijad eeldavad, et madala sekveneerimiskattuvuse tõttu jäi oluline osa mutatsioone avastamata. (Francioli jt., 2014)

Hollandlaste uuring lisas dbSNPi andmebaasi juba teadaolevatele mutatsioonidele juurde 7,6 miljonit ühenukleotiidset variatsiooni (joonis 1), millest 75% esinesid valimis vaid korra. Antud mutatsioonide hulk iseloomustab populatsioonipõhiste uuringute tähtsust mõistmaks harvade variantide geneetilist mõju fenotüübile. (Francioli jt., 2014)



Joonis 1. GoNL projekti leiud ning kattuvus teiste andmebaasidega nagu dbSNP, 1000 genoomi projekt (1000 G), 1000 genoomi projekti eurooplased (1000G EUR) ning HapMap Euroopa pärituoluga ameeriklased (HapMap CEU) (allikas: Francioli jt., 2014).

1.1.2.2 SISu

(ingl. *The Sequencing Initiative Suomi*)

Meie põhjanaabrite sekveneerimisprojekti SISu esimene publikatsioon võrdleb 3000 soomlase eksoomijärjestusi teiste euroopa rahvastega (samuti 3000 indiviidi) ning uurib isoleeritud populatsioonis haruldaste variantide mõju kompleksfenotüübile (Lim jt., 2014).

Eksoomisekveneerimine ning järjestuste analüüs toimus koostöös Broad instituudiga, täpsustused on ära toodud tabelis 2. Uurigu tulemusena leidsid nad soomlaste seas harva esinevate (MAF 0.5-5%) variantide kõrgema esinemise. Sagedased variandid uuritavate populatsioonide vahel olid sarnaste esinemismustritega. Samuti leidsid nad, et soome päritolu indiviididel on pea neli korda vähem *singleton*-e ehk ainult ühel indiviidil leiduvaid variante. Lisades geneetilistele andmetele juurde verebiokeemilised näitajad nagu vererõhk ja lipiidide tase jõudsid nad järeldusele, et mutatsioonid lipoproteiin A kodeerivas geenis (*LPA*) toimivad kaitsvalt kardiovaskulaarsete haiguste suhtes, vähendades veres tsirkuleerivaid lipoproteiine. Antud leidu valideeriti ning kinnitati kolmes erinevas kontrollvalimis, muuhulgas ka Eesti Geenivaramu doonorite seas. (Lim jt., 2014)

1.1.2.3 Islandi projekt

Islandlaste biopanga ajalugu ulatub juba 1990-ndate algusaastatesse, kus käid välja maailma esimese populatsioonipõhise biopanga projekti idee (Greely, 2000). Täisgenoomide projektini jõudsid nad umbes kümmekond aastat hiljem. Iga sajanda islandlase (N=2636) DNA eraldati valgetest vererakkudest ning sekveneeriti Illumina GAIIx või HiSeq platvormil (tabel 2). (Gudbjartsson jt., 2015)

Isoleeritud populatsioonina näitavad nad homosügootsuse ning harvade variantide sagedasemat esinemist võrreldes mõne teise euroopa populatsiooniga. Lisaks koguti 104 220 indiviidi (1/3 kogu populatsiooni rahvaarvust) genotüüpide info ning tänu sekveneerimisandmetele oli võimalik määrata suurema hulga inimeste haplotüübid ja puuduolevad variandid imputeerida. Andmekogumist leidsid nad mitmeid uusi korrelatsioone nagu näiteks *MYL4* geeni raaminihkemutatsiooni põhjuslikku seost südamehaigustega (kodade virvendus ja laperdus), maksahaiguste riski tõusmist *ABCB4* geeni mutatsioonide

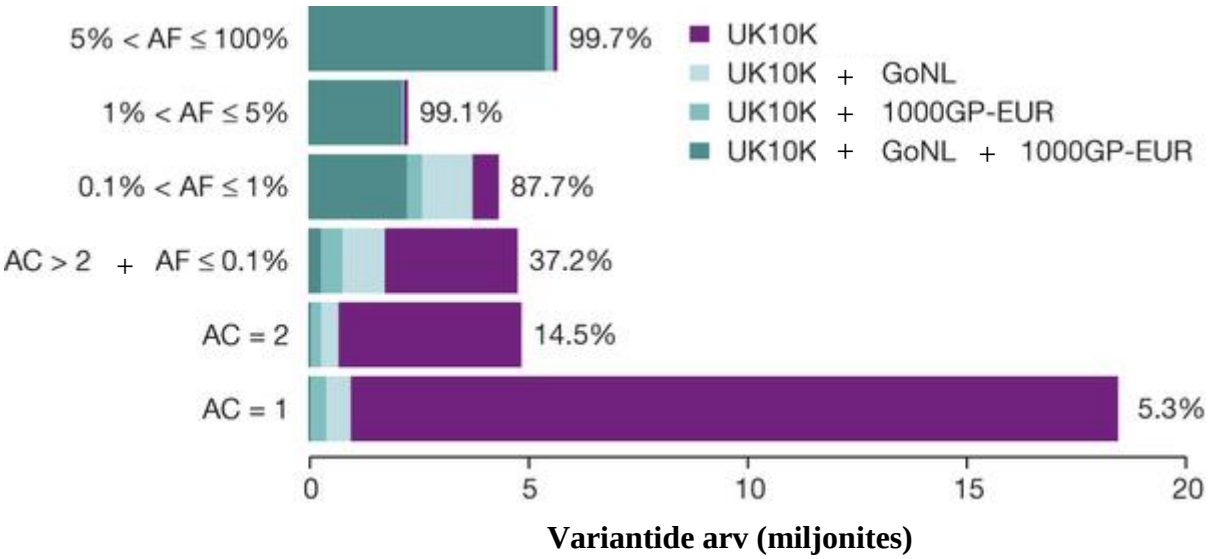
ning Alzheimer haiguse saamisriski kahekordistumist *ABCA7* geeni mutatsiooni tõttu. (Gudbjartsson jt., 2015)

1.1.2.4 UK10K

Suurbritannia ja Põhja-Iiri Ühendkuningriik (UK) on esimene populatsioon, mis küündib oma uuringu valimi suurusel juba 10 000 lähedale. Projekt disainiti iseloomustamaks haruldasi ning harvasid variante UK populatsioonis, nende seost varieeruvate bioloogiliste fenotüüpide ning geneetiliste haigustega. Kokku leiti üle 42 miljoni ühenukleotiidse variatsiooni ning 3,5 miljonit indelit (joonis 2). (Walter jt., 2015)

Uuringu ülesehitust võib jagada kaheks – esiteks sekveneeriti 3781 terve indiviidi lümfoblastoidsetest rakkudest eraldatud genoomne DNA madala kattuvusega (keskmiselt 7X). Katsed viidi läbi Wellcome Trust Sanger instituudis (Wellcome Trust Sanger Institute, Hinxton, UK) ning BGI-s (tabel 2). (Walter jt., 2015)

Otsiti geneetilist variatsiooni 64 erineva fenotüübi tunnusele nagu ülekaalulisus, diabeet, verebiokeemilised näitajad, vererõhk, südame ja maksa talitlus jpt. Sekveneerimisandmetest leiti mitmeid seoseid haruldaste ning harva esinevate variantide ja eelpoolt mainitud fenotüübiliste tunnustega. Näiteks haruldane mutatsioon *APOC3* geeni intronis mõjutab vereplasma triglütseriidide taset, mis omakorda vähendab kardiovaskulaarse haiguse saamise riski (Timpson jt., 2014, p. 3). Teiseks analüüsiti kolmest erinevast kohordist (harvad haigused, raskeloomuline ülekaalulisus ning neuraalarenguga seotud haigused) pärinevaid 5182 indiviidi eksoome. DNA eraldati samuti lümfoblastoidsetest rakkudest ning sekveneetiti kõrge kattuvusega (keskmiselt 80X) Illumina platvormil Wellcome Trust Sanger instituudis. Teadlased leidsid 25 täiesti uut põhjuslikku varianti viiele haruldasele geneetilisele haigusele. (Walter jt., 2015)



Joonis 2. UK10K, GoNL ja 1000 genoomi projektides (arvestatud ainult euroopa populatsiooni tulemused) leitud variandid alleelisageduste kaupa (Walter jt., 2015, kohandatud).

Suuremahulised populatsioonide uuringud analüüsides kogugenoomi järjestusi annavad võimaluste mõista geneetiliste variatsioonide mõju inimeste mitmekesisusele.

Tabel 2. Senised suuremad populatsioonipõhised sekveneerimistööd

Projekt	***GoNL (Francioli jt., 2014)***	***SISu (Lim jt., 2014)***	***Island (Gudbjartsson jt., 2015)***	***UK10K (Walter jt., 2015)***
Valim	769 (WGS)	3000 (WES)	2636 (WGS)	3781 (WGS) + 5182 (WES)
DNA allikas	veri	veri	veri	veri
Sekveneerimise platvorm	Illumina	Illumina	Illumina	Illumina
Järjestuste joondamine	BWA v0.5.9-r16	NA	BWA v0.5.7-0.5.9	BWA v0.5.9-r16
Variantide analüüs	GATK v1.6 jt*	GATK, VEP v2.5	GATK v2.3.9, VEP v2.8	GATK v1.1, VEP v77

Sekveneerimise kattuvus	13X	NA	10-30X (Me =20)	7X, 80X
Leiud				
SNV (miljonit)	20,4	NA	19,7	42
Indel (miljonit)	1,2 (<20bp)	NA	1,4	3,5
Suured deletsioonid	27 500 (>20bp)	NA	NA	18 739 (Me=3,7kb)

WGS=täisgenoomi sekveneerimine
WES=eksoomi sekveneerimine
Me = mediaan
BWA = Burrows-Wheeler Aligner (Li ja Durbin, 2009)
mrsFAST tööriist (Hach jt., 2014)
GATK = Genome Analysis Toolkit (McKenna jt., 2010).
VEP = Variant Effect Predictor (McLaren jt., 2010).
*Pindel (Ye, Schulz, Long, Apweiler, ja Ning, 2009), 1-2-3SV, Breakdancer (Chen jt., 2009), DWAC, CNVnator (Abyzov, Urban, Snyder, ja Gerstein, 2011), FACADE (Coe, Chari, MacAulay, ja Lam, 2010), MATE-CLEVER (Marschall, Hajirasouliha, ja Schönhuth, 2013), GenomeSTRiP (Handsaker, Korn, Nemesh, ja McCarroll, 2011) ja SOAPdenovo (R. Li jt., 2010)

1.1.2.5 ExAC

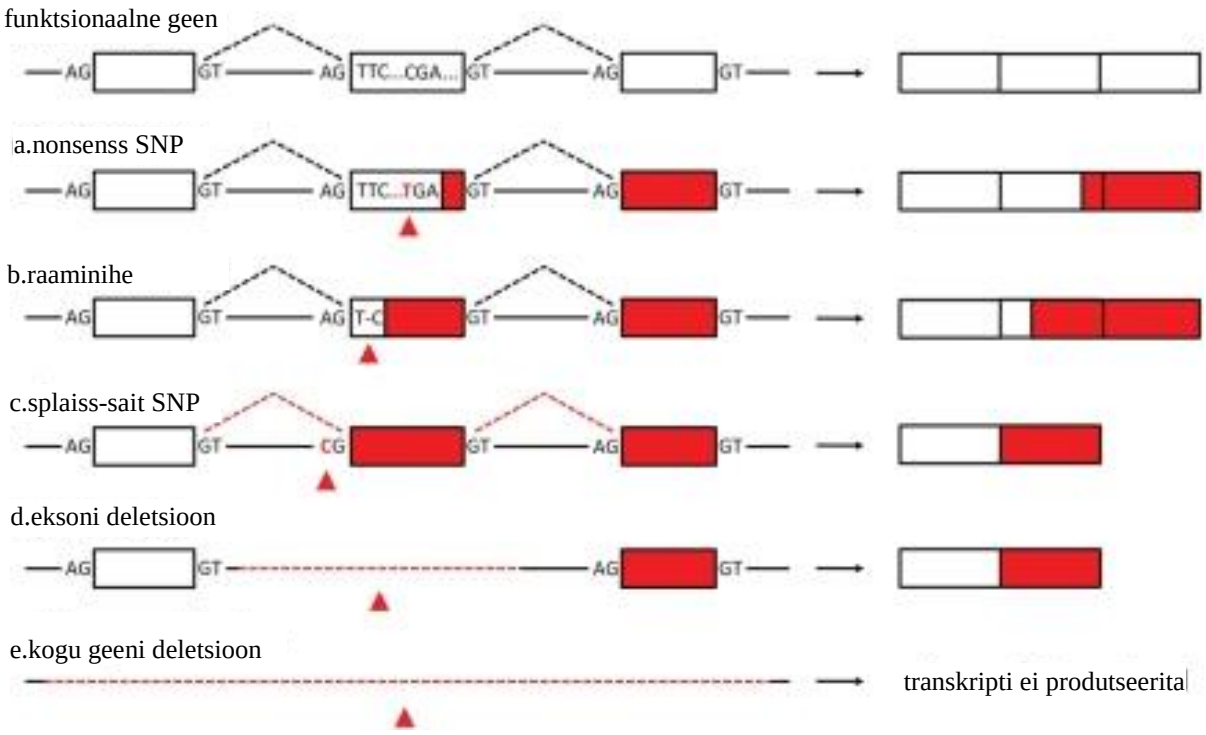
Eksoomi liitkonsortsium (lühend ExAC, ingl. *Exome Aggregation Consortium*; exac.broadinstitute.org) on pea kahekümne teadusasutuse koostöö projekt, millega püütakse ühendada mitmete suuremahuliste sekveneerimisprojekti andmed kokkuvõtvaks andmebaasiks laiemale teadlaskonnale. Hetkel on ExAC-i koondatud 60 706 erineva etnilise taustaga indiviidide eksoomianalüüsi tulemused nii populatsiooni kui ka haigusspetsiifilistest teadusuuringutest.

ExAC-i esimeses võrgus avaldatud publikatsioonis kirjeldavad nad 7 404 909 kõrge kvaliteediga varianti, keskmiselt üks mutatsioon iga kaheksa aluspaari järel. Enamik leitud variantidest on haruldased – pea 99% on alleelisagedusega vähem kui üks protsent ja 54% kõigist variantidest on *singleton*-id. ExAC on seni suurim geeni funktsiooni mõjutavate mutatsioonide andmebaas, sisaldades 179 774 valgu funktsioonile võimalikku mõju omavat varianti. Nad tõstavad esile 3230 potentsiaalset geeni, mis ei tolereeri geeni avaldumise efektiivsust mõjutavaid mutatsioone. Rohkem kui kolmveerandil nendest geenidest puudub kirjeldatud (haiguslik) fenotüüp tuntud andmebaasides (OMIM ja ClinVar). (Lek jt., 2015)

Konsortsiumi idee on kirjeldada inimese geneetilise variatsiooni mustreid, eriti mendeliaarsete haigustega seotud geenides, ning leida nokautmutatsioonidega inimesi. Nende kataloog võimaldab teadlastel sekveneeritud andmestikke uurides filtreerida potentsiaalseid haigusseoselisi variante, mida käesolevas töös ka kasutatakse.

1.2 Funktsioonikaoga mutatsioonid

Funktsioonikaoga mutatsioon (lühend LoF, inglise keeles *loss-of-function*) on muutus DNA järjestuses, mis nõrgendab või kõrvaldab geeni avaldumise või geeni produkti funktsiooni (Heinaru, 2012). LoF mutatsioonid võivad potentsiaalselt rikkuda erinevate geneetiliste elementide funktsionaalsust nii kodeerivas kui ka mittekodeerivas alas, mille tagajärjel on häiritud normaalse valgu süntees või geenide ekspressioon. LoF mutatsioonid võivad olla nii ühe nukleotiidsed muutused kui ka väiksemad ja suuremad struktuursed variatsioonid. Valkukodeerivate LoF mutatsioonid võivad olla mitut tüüpi (joonis 3):



Joonis 3. LoF mutatsioonide näiteid. Joonise üleval on näidatud funktsioonaalne geen ning alumised pildid iseloomustavad võimalikke LoF mutatsioonide efekti geeni funktsioonile. Punasega on märgitud muutused geenis ja ka valgu tasandil (D. G. MacArthur ja Tyler-Smith, 2010).

- Nonsensssmutatsioon

Geeni DNA-järjestuse muutus, mille tõttu tekib enneaegne stoppkoodon ning funktsionaalset geeniprodukti ei sünteesita (joonis 3a) (Heinaru, 2012). Eukarüootsetes organismides leidub rakusisene kvaliteedikontrolli mehhanism (*nonsense mediated mRNA decay)*, mis degradeerib enneaegse stoppkoodoniga mRNA transkriptid. On teada, et ~5%-25% ulatuses võib vigast mRNA-s siiski alles jääda ning sellistele transkriptidele vastavad valgud võivad olla vigase või ka täiesti puuduliku funktsiooniga (Danckwardt, 2002; Isken ja Maquat, 2007). Seega ei pruugi olla nonsenssmutatsioonidel otsene seos geeni inaktiveerimisega. Nonsenss LoF mutatsioone on seostatud ~15%-30% monogeneetiliste haigustega nagu hemofiilia, Duchenne'i lihasdüstroofia jt (Mort, Ivanov, Cooper, ja Chuzhanova, 2008).

- Raaminihkemutatsioon

Mutatsioon, mis muudab mRNA lugemisraami nukleotiidide lisandumise või väljalangemise tõttu (joonis 3b) (Heinaru, 2012). Raaminihet põhjustavad indelid, mille nukleotiidne pikkus ei jagu arvuga kolm. Selline mutatsioon omab drastilist efekti polüpeptiidile, muutes mitte ainult ühe lüli ahelast, vaid kogu mutatsioonile järgneva aminohappelise järjestuse. Raaminihkemutatsioon põhjustab raskekujulisi geneetilisi haiguseid nagu Tay-Sachs, Crohn-i tõbi, tsüstiline fibroos jt (Myerowitz ja Costigan, 1988; Ogura jt., 2001, p. 2; White jt., 1990).

- SNP splaiss-saidis

Mutatsioonid geeni splaiss-saitides (joonis 3c) ning nende potentsiaalsed tagajärjed saavad olla väga erinevad. Muutused kanoonilises splaiss-saidis võivad mõjutada splaissimist ning selle tagajärjel inaktiveerida geenifunktsiooni (Baralle ja Baralle, 2005; Krawczak, Reiss, ja Cooper, 1992). SNP splaiss-saidis võib põhjustada splaissimismeetodis vigu, mille tulemusena valmis mRNA sisaldab intronit või on ekslikult ekson vahelejäetud ning puudub lõpp-produktis (Aoshima jt., 1996). Samuti võib SNP luua *de novo* splaiss-saidi ning tulemuseks on täiesti uue struktuuriga mRNA.

- Kogu geeni deletsioon

Suuremate genoomsete ümberkorralduste tõttu võib puududa mingi jupp (joonis 3d) või terve geeni järjestus ja seega ka geeniprodukt (joonis 3e).

LoF mutatsioonide tagajärjed fenotüübile võivad olla väga erinevad. Eelkõige on geenifunktsiooni lõhkuvaid funktsioone seostatud raskeloomuliste mendeliaarsete haigustega, kuid viimaste aastate suuremahulised sekveneerimisprojektid on näidanud, et ka tervetel inimestel leidub suures hulgas LoF mutatsioone. LoF mutatsioonide rohkus ning kõrge sagedus viitab nende pigem neutraalsele või isegi healoomulisele mõjule, paljud LoF variandid omavad kohasusele väikest efekti. Kõige enam leidub neutraalseid, inimese ellujäämisvõimalust mitte mõjutavaid variante inimesele eluks mitte esmavajalikes geenides, näiteks haistmismeeltega seotud geenides. Samuti on LoF mutatsioonid sagedamini leitavad erinevate veregruppide ning metabolismi protsessidega seotud geenides (Calafell jt., 2008; Cohen jt., 2005; Cvejic jt., 2013; Farris jt., 2004).

Esimesi tõendeid soodsatest LoF mutatsioonist leiti juba 20-nda sajandi alguses ABO veregrupi avastusega – ühe aluspaari pikkune deletsioon tekitab O veregrupi alleeli (Calafell jt., 2008). Kõige suurem ravimite metabolismis osalevate ensüümide geeniperekond on *CYP* (McLean jt., 2005), vastutades 75% kogu ravimite metabolismi eest (Guengerich, 2008). Erinevused ravimi vastustes tulenevad mitmetest LoF mutatsioonidest *CYP*-geenide perekonnas, mõjutades ravimite metabolismi kiirust ning võimekust (de Morais jt., 1994; Gaedigk, Blum, Gaedigk, Eichelbaum, ja Meyer, 1991).

LoF mutatsioonid võivad olla ka positiivse selektsiooni all (hüpotees „vähem on rohkem“)– fenotüüpi kaitsvad ja kasulikud variandid (Olson, 1999). Kõige tuntum näide on *PCSK9* geenis leiduvad mutatsioonid, mille tulemusena on veres ringleva LDL kolesterooli tase madalam ja kardiovaskulaarsete haiguste risk väiksem (Cohen jt., 2005). Teine tuntud näide on 32 aluspaari pikkune deletsioon *CCR5* geenis, mis põhjustab valgete vererakkude pinnal oleva retseptori mittefunktsionaalsuse ning takistab HI-viirusel (inimese immuunpuudulikkuse viirus ehk HIV) rakku pääseda ning seda nakatada (Samson jt., 1996). Nonsenssmutatsioon *CASP12* geenis, mis on ühtlasi ka üks sagedasemaid mutatsioone (alleelisagedus Euraasia populatsioonides pea 100% (Xue jt., 2006)) vähendab riski haigestuda kogu keha põletikku ehk sepsisesse ning seega tagab paremad võimalused haiglakeskkonnas hakkama saada (Saleh jt., 2004). Sarnaseid näiteid leiab kirjandusest veelgi ning need positiivse efektiga mutatsioonid on sihtmärgiks ravimitööstuses (tabel 3).

Tabel 3. Positiivse efektiga LoF mutatsioonide näiteid

Geen	Valgu funktsioon	Efekt
CCR5 *(Samson jt., 1996)*	Raku pinnaretseptor	Kaitse HIV viiruse eest
PCSK9 *(Cohen jt., 2005)*	Seondub LDL kolesterooli retseptoriga	Madalam LDL kolesterooli tase veres ning väiksem kardiovaskulaarsete haiguste risk
SLC30A8 *(Flannick jt., 2014)*	tsink transporter	Vähendab tüüp 2 diabeedi haigestumisriski
ACTN3 *(Daniel G. MacArthur jt., 2007)*	Valk lihastes	Madalam võimekus äkilises lihastöös (näiteks sprintimine)
CASP12 *(Saleh jt., 2004)*	Immuunvastus bakteritele	Vähendab riski haigestuda sepsisesse
SCN9A *(Weiss jt., 2011)*	Naatriumkanalites	Väike valutundlikkuse

Esimene suurejoonelisem töö kaardistamaks inimese LoF mutatsioone tehti 2012. aastal Daniel. G. MacArthur ja tema töögrupi poolt. Nad analüüsisid 185 inimese genoomi (1000 genoomi pilootprojekti raames) ning leidsid, et igal tervel indiviidil leidub genoomis keskmiselt 100 LoF mutatsiooni ning umbes viiendik neist põhjustab geeni täielikku nokauti (MacArthur jt., 2012). Nende 2951-st kandidaat LoF variandist (tervetes indiviidides) 26 olid teadaolevad haigusseoselised variandid. MacArthur jt. leidsid oma uurimistöös ka vähemalt ühe indiviidi, kellel oli kokku 253 geeni mõlemad koopiad mittefunktsionaalsed. Ei ole veel kindlalt teada kui palju genoom suudab sääraseid mutatsioone taluda. Homosügootsete LoF mutatsioonidega inimesed on justkui nokaut mudelid – võtmeisikud mõistmaks paremini geenide funktsiooni ning leidmaks kasulikke ja kaitsvaid mutatsioone, mida kasutada geeniteraapias ja ravimistööstuses.

1.3 Elektroonilised terviseandmed ja genoomika

Teaduslike uuringute ning avastusportsesside meetodid on pidevalt muutumises. Üha enam kasutatakse uurimustöödeks juba olemasolevaid vahendeid nagu näiteks elektroonilised terviseandmed (EHR, ingl *electronic health record*). Elektroonilised terviseandmed on eelkõige mõeldud kliinilisteks uuringuteks, kuid võimaldavad teadlastel väljaspool haiglat olemasolevate indiviidide pealt analüüsida näiteks haiguste kulgu, ravimite efekti ja vastust.

Biopangad on muutunud genoomika uurimise lahutamatuks osaks. Genoomika valdkonna edasijõudmised ja EHR andmete kasvava kasutamisega on andmete kõrvutamisel potentsiaali

mõista paremini geneetilist komponenti inimese tervises ning seeläbi parandada tervisesüsteemi terviklikult.

1.3.1 Elektroonilised terviseandmed teadusuuringutes

Elektroonilised terviseandmed on süstematiseeritult kogutud digitaalne patsiendi tervise informatsioon. EHR koosneb nii struktureeritud andmetest (laboratoorsed mõõtmistulemused, diagnoosi koodid, ravimiretseptid jms) kui ka vabatekstist (arsti märkmed vms). Lisaks võib patsiendi terviselugu sisaldada ka erinevaid dokumente või analüüsipilte.

Terviseandmete digitaalset andmebaasi on lihtsasti võimalik, vastavate kooskõlastuste olemasolul, jagada erinevate teadusasutuste või biopankadega uuringute läbiviimiseks. Biopanganduses on väga olulisel kohal liitujate informeeritud nõusolek teadusuuringuteks ja parimal juhul ka nõusolek andmete täiendamiseks riiklikest registritest ning andmebaasidest. Teadlaste jaoks selline informatsioonipagas võimaldab jälgida suurel hulgal kindla fenotüübiga indiviide pika aja vältel, mis parandab oluliselt tõeste positiivsete leidude eristamise valepositiivsetest. Lisaks on võimalik biopangal koguda geneetilist materjali enne inimese haigestumist. Sellised juhtumid on head võimalused haiguseid ennustavate mudelite väljatöötamiseks.

Teadustööd on tänaseks näidanud, et kõrvutades tervisandmeid vaadeldavate indiviidide geneetilise infoga on võimalik leida haruldasi, kuid suure efektiga haigusseoselisi variante (Crosby jt., 2014; Flannick jt., 2014; Holm jt., 2011). 2007. aastal loodi Elektroonilise terviseandmete ja Genoomika võrgustik (lühend eMERGE, ingl. *Electronic Medical Records and Genomics Network*) eesmärgiga analüüsida EHR andmetest võetud fenotüübilise info kasutamist (ka eetilisi ja juriidilisi aspekte) ülegenoomsetes teadusuuringutes. Tänaseks päevaks kuulub eMERGE gruppi üle kümne teadus- või terviseasutuse nii bioloogilise kui fenotüübilise andmekoguga, kokku üle 350 000 indiviidi (Gottesman jt., 2013; A. N. Kho jt., 2011). Viimase kuue aasta jooksul on eMERGE arvukad ülegenoomsed assotsiatsiooniuuringud (lühend GWAS, ingl. *genome-wide association studies*) näidanud EHR andmete kasutamise tulemuslikkust ning replitseerinud mitmeid juba teadaolevaid genotüübi-fenotüübi seoseid (Crosslin jt., 2012; J. C. Denny, Ritchie, Basford, jt., 2010; J. C. Denny, Ritchie, Crawford, jt., 2010; Joshua C. Denny jt., 2011; Abel N Kho jt., 2012; Kullo jt., 2011; Kullo, Ding, Jouni, Smith, ja Chute, 2010; Ritchie jt., 2010).

Üks 12-st eMERGE võrgustikku kuuluvast asutusest on Vanderbilt DNA andmepank (BioVU). BioVU kogub patsientide haiguslugude kõrvale ka bioloogilist materjali eraldades DNA rutiinsest vereanalüüsist järgi jäänud (äraviskamisele kuuluvast) verest (Roden jt., 2008). Andmepank on kasutanud doonorite geneetilist infot hindamaks, kas EHR andmeid kasutades on võimalik leida juba teadaolevaid genotüüp-fenotüüp seoseid. Ritchie jt. analüüsisid viit komplekshaigust – tüüp II diabeet, reumatoidartriit, Crohn-i tõbi, polüskleroos ja kodade virvendus ning nendega seostatud ja replitseeritud 21 geneetilist markerit (Barrett jt., 2008; Gudbjartsson jt., 2007; Hafler jt., 2007, 2007; Parkes jt., 2007; Ritchie jt., 2010). Iga haiguse kohta koostati lingitud EHR andmete alusel haigete ja tervete valimid, kust valiti koguvalimisse välja 9483 indiviidi, kusjuures, iga valimi liige oli ühe haiguse suhtes juhtum ning kõigi teiste kontrollisikuks. Kõigi viie uuritud haiguse korral leiti assotsiatsioon vähemalt ühe teadaoleva SNP-ga. Kõigi 21 väljavalitud markeri seoste suund tunnustega ennustati õigesti ning 18 neist suudeti antud valimis ka replitseerida (Ritchie jt., 2010). Nende tulemused toetavad tugevalt tervisesüsteemi andmete sidumist biohoidlate genotüüpidega geneetilise diagnostika arendamiseks.

Terviseandmeid kasutavad projektid on avastatud uusi potentsiaalseid variante (Crawford jt., 2014). eMERGE 2011. aasta uuring leidis euroopa pärisolu ameeriklaste seas seose nelja *FOXE1* geeni SNP-i ja hüpotüreoidismi vahel (Joshua C. Denny jt., 2011). Antud leidu replitseeriti aasta hiljem Mayo genoomi konsortsiumi (ingl. *Mayo Genome Consortia*) andmetega (Eriksson jt., 2012). Mayo kliinik on muuhulgas analüüsinud ka erütrotsüütide näitajate geneetilisi mõjutegureid ning kaardistanud neli lookust, mis mõjutavad punaste vereliblede arvu, hemoglobiinitaset jms kasutades EHR andmetest pärit näitajaid (Kullo jt., 2010).

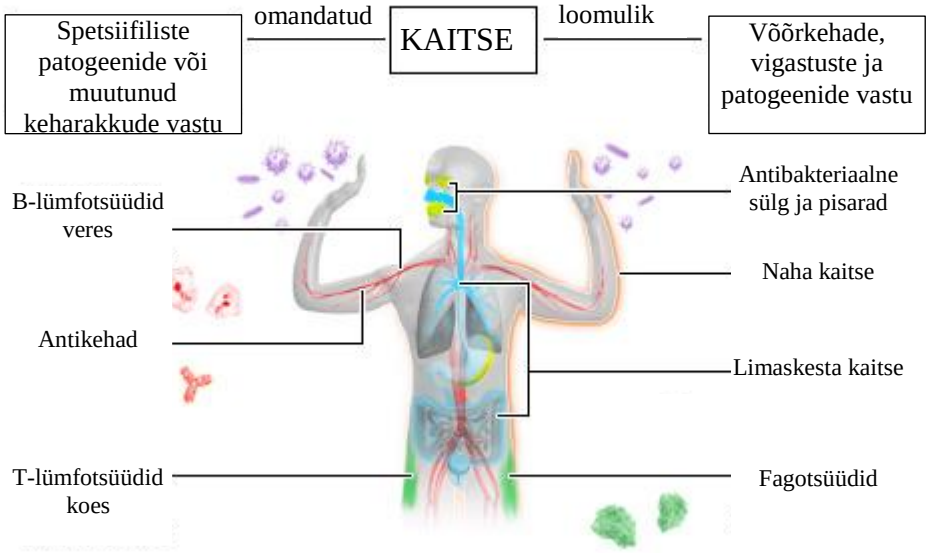
EHR andmetest on peale diagnooside ja haiguste võimalik kasutada ka erinevaid terviseuuringute laboratoorseid mõõtmistulemusi. Seosed genoomi ja tervisenäitajate vahel võivad viia uute geneetiliste leidudeni. Verma jt. Pennsylvania osariigi ülikoolist hindasid võimalikke seoseid 21 kliinilise näitaja (verebiokeemia, kehamassiindeks jms) ning 635 525 geneetilise markeri vahel (Verma jt., 2016). Analüüsi käigus leidsid nad 286 olulist SNP-i (p-väärtus $< 1{,}37 \times 10^{-8}$), millest üle pooled olid kirjanduses väljatoodud assotsiatsioonid. Lisaks leidsid nad ka potentsiaalseid uusi variante. Antud uuring iseloomustab terviseandmete kasutamise võimaluste rohkust.

Biopankade geneetiline materjal on väärtuslikum kui on võimalik juurde kõrvutada fenotüübi informatsioon. Kindla valimiga ühekordsel andmekogumisel võivad olulised aspektid esialgu märkamata jääda. EHR andmete kättesaadavus võimaldab aga biopanga doonoreid kogudes luua arvukaid valimeid väga erinevate haiguste kohta nii planeeritud kui ka tuleviku projektideks. Selline lähenemine loob ka kliinilises praktikas võimaluse pidevalt uuendada ja täpsustada genotüüp-fenotüüp seoseid.

1.4 Immuunsüsteem

Immuunsus on organismi vastupanu haigustekitajate suhtes ja seda tagab immuunsüsteem (joonis 4). Immuunsüsteem koosneb mitmetest olulistest elunditest nagu põrn, harkelund, lümfisõlmed, nahk, luuüdi ning mitmesugustest rakkudest ja biomolekulidest. Seda koordineeritud süsteemi võib jagada kaheks – loomulik ja omandatud immuunsüsteem. Kui võõraine või patogeen tungib organismi, siis kõigepealt käivitub evolutsiooniliselt vanem loomulik immuunsus, näiteks toimub nahas ja limaskestadel mikroorganismide mittespetsiifiline hävitamine. Haigustekitajate jäämisel organismi pikemaks ajaks kutsutakse esile omandatud immuunreaktsioon. Omandatud immuunreaktsioon on kõrge spetsiifikaga immuunvastus, mis on vahendatud kahte tüüpi immuunvastusena – humoraalse (antikehad) ja rakulise (T-lümfotsüüdid) immuunreaktsioonina. (Uibo, Kisand, Peretson, ja Reimand, 2015; Velbri, 1982)

Immuunsüsteem on keeruline võrgustik. Immuunsüsteemi funktsiooni kahjustused võivad põhjustada autoimmuunsust, põletikulisi protsesse ning isegi vähki.



Joonis 4. Immuunsüsteem (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0025680/, kohandatud)

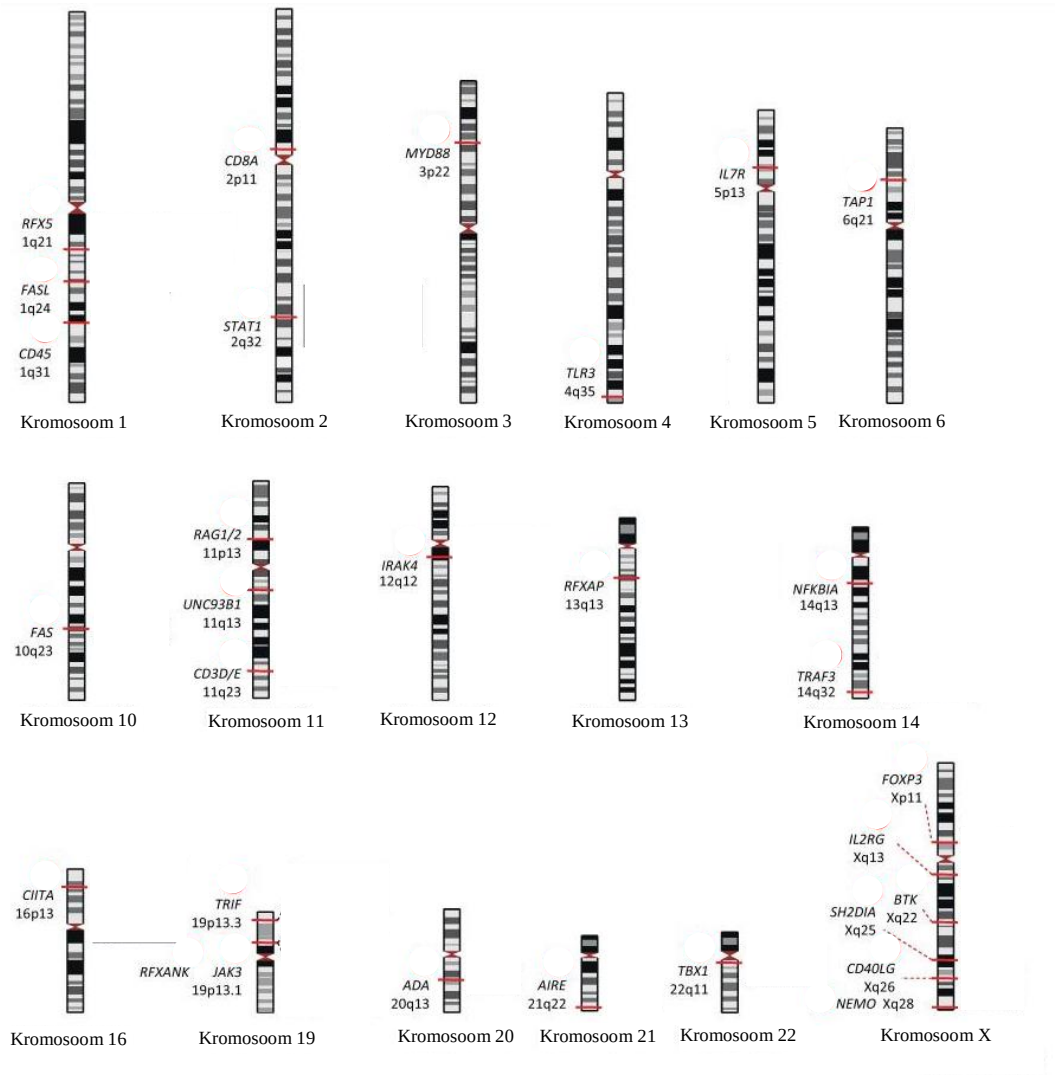
1.4.1 Immuunsüsteemi geneetiline struktuur ja haigused

Arusaamine immuunsusest ning selle geneetilisest taustast on siiani sageli toimunud pärandunud mutatsioonide analüüsimisel. Nii suuremahuline genotüpiseerimine kui täisgenoomide analüüsid täidavad klassikalise geneetika puudujääke, võimaldades leida harvaesinevate haiguste geneetilisi põhjuseid ning selgitada sagedasemate immuunsüsteemiga seotud haiguste tagamaad. Millised variandid on põhjuslikud ja millist rolli immuunsüsteemi haigustes mängib geneetiline komponent, on veel vähe teada.

Immuunsüsteemi õige funktsioneerimine on vajalik terve organismi toimimiseks. Ülegenoomsed assotsiatsiooniuuringud on seostanud seni teadmata lookuseid ja geene erinevate immuunprotsesside ja haigustega (joonis 5). Näiteks, immuunpuudulikkuse geneetiline kirjeldus hõlmab üle 200 iseloomustatud geeni (Picard jt., 2015). Sardiinia teadlased sõelusid täisgenoomi andmetest 89 geeni, mis osalevad immuunsüsteemi rakkude tootmise regulatsioonis (Sidore jt., 2015). Peamine koesobivuskompleks ehk MHC regioon (ingl. *major histocompatibility complex*) kuuendas kromosoomis omab kõige tugevamat geneetilist efekti ja on seotud mitmete erinevate immuunsüsteemiga seotud häiretega (Hosomichi, Shiina, Tajima, ja Inoue, 2015).

Immuunsüsteemi ülesanne on organismi kaitsta väliskeskkonna kahjulike tegurite eest. Kui immuunsüsteemi häirumise tulemusena tekib organismi enda tervete kudede vastu immuunreaktsioonid, siis on tegemist autoimmuunsusega. Immuunsüsteemi funktsioneerimise eest vastutavatest geenidest on suur osa seotud ka autoimmuunhaiguste tekkimisega. Näiteks kõige enam levinud immuunglobuliin A puudulikkuse korral esineb ka autoimmuunhaigusi sagedamini (Uibo jt., 2015). Tänu GWAS uuringutele on autoimmuunhaigustele nagu Crohn-i tõbi, reumatoidartriit või psoriaas leitud genoomseid haiguspõhjuslikke variante (Festen jt., 2011; Franke jt., 2010). A.P Gregory jt läbiviidud GWAS-i signaal *TNFRSF1A* geenis leidus polüskleroosi (ingl. *multiple sclerosis*) haigete seas (Gregory jt., 2012). Edasine analüüs antud leiuga viis alternatiivselt splaissitud *TNFR1* geeni transkriptini, mis ühtlasi mängib rolli polüskleroosi ravi mittetoimimisel. Tänaseks on autoimmuunhaigustega seostatud üle 200 lookuse (Feero, Guttmacher, Cho, ja Gregersen, 2011).

Analüüsides immuunhaigustega seotud GWAS leide, siis enamik korrelatsioone ei seleta ära täielikult haiguse geneetilist komponenti ning tihti on kaardistatud lookuse efektisuurus väike (Park jt., 2010). Näiteks on praeguseks teadaolevate variantidega äraseletatud 50% polüskleroosi geneetilisest komponendist (Sawcer jt., 2011). Oletatakse, et viimastel aastatel aina uuritum mikrobioom ning selle geneetika on samuti inimese immuunsüsteemi suureks modulaatoriks (Grice ja Segre, 2012).



Joonis 5. Immuunhaigustega seotud mendeliaarseid variante üle genoomi. Koos on näidatud nii geen kui genoomne asukoht (Knight, 2013, kohandatud).

Immuunhaigused on väga heterogeenne haiguste rühm, mistõttu klassifikatsioonides esineb kohati liiga laiapõhjaline haiguste määramine. Immuunsüsteemi kujundavate geenide kaardistamine ning nende efektide mõistmine võimaldaks selgemat diagnoosimist ja immuunhaiguste ravivõimaluste väljatöötamist.

2 EKSPERIMENTAALOSA

2.1 Töö eesmärgid

- Anda ülevaade seni tehtud populatsiooni täisgenoomide projektide kohta ning võrrelda üldtulemusi Eesti Geenivaramu 2240 geenidoonori sekveneerimisandmetega
- Iseloomustada funktsioonikaoga mutatsioone Eesti populatsioonis
- Hinnata biopanga andmete kasutamisvõimalust geneetiliste seoste leidmisel, analüüsides immuungeenidest saadud funktsioonikaoga mutatsioone ja geenidoonorite terviseandmeid
- Prioritiseerida immuungeenidest leitud variandid edasisteks geneetilisteks ja funktsionaalseteks analüüsideks

2.2 Materjal ja metoodika

2.2.1 Valim

Täisgenoomide valim koosneb 2240 Tartu Ülikooli Eesti Geenivaramu (TÜ EGV) geenidoonorist, kes jagunevad kolme alamvalimisse: siirdegenoomika valim (N=977), kaks suurt perekonda (N=21), juhuvalim sünnikoha järgi (N=1242). Siirdegenoomika valimisse kuuluvad eestlased vanuses 18-74 aastat ning fenotüübipõhised lähisugulased on välja filtreeritud. Koguvalimi eesmärk oli katta ära Eesti riigi inimasustus võimalikult laiaulatuslikult ning seeläbi saada kätte võimalikult suure osa eestlaste geneetilisest varieeruvusest. Valimi sooline ja vanuseline jaotus on äratoodud tabelis 4.

Tabel 4. Täisgenoomide valimi sooline (tabel 4a) ja vanuseline (tabel 4b) jaotus sünniaastate järgi

a.

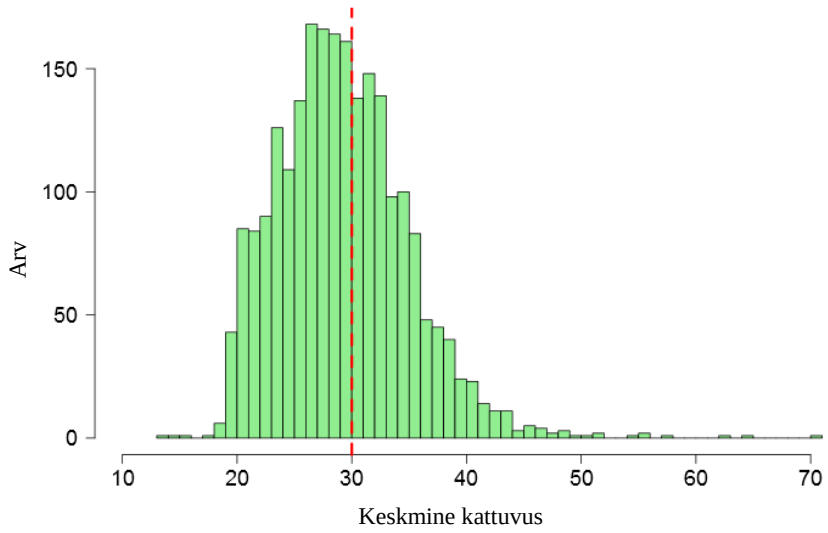
	Arv	Jaotuvus
Mees	1138	50,80%
Naine	1102	49,20%
Kokku	2240	100%

b.

	Sünniaasta
Keskmine	1964
Noorim	1995
Vanim	1920

2.2.2 Täisgenoomide sekveneerimine ja järjestuste joondamine

Kogu sekveneerimine ning esialgne bioinformaatiline analüüs viidi läbi Broad instituudis. Proovid valmistati Illumina PCR-vaba raamatukogu kitti (TruSeq DNA PCR-Free Library Preparation Kit) kasutades ning sekveneeriti Illumina HiSeq X Ten platvormil. Sekveneeriti 150 aluspaari pikkused *paired-end* järjestused, keskmine kattuvus 30x (joonis 6).



Joonis 6. Täisgenoomide kattuvuse jaotus (autor: Mart Kals)

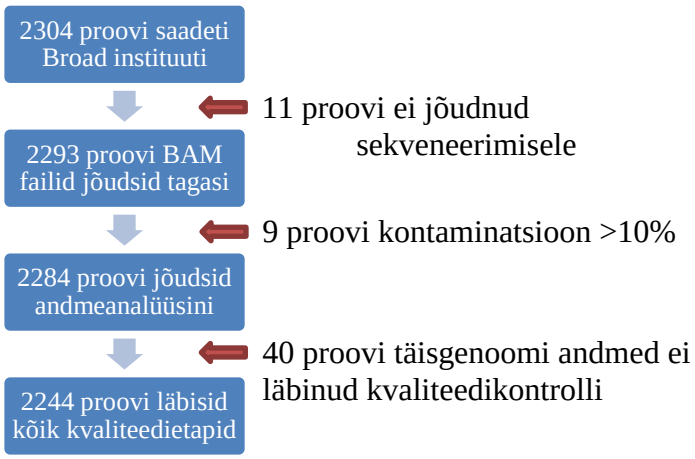
Saadud DNA järjestused paigutati inimese referentsgenoomile versioon GRCh37 (hg19), mis sisaldab ka Epstein-Barr viiruse (kontiig NC_007605) DNA järjestust (http://www.broadinstitute.org/ftp/pub/seq/references/Homo_sapiens_assembly19.fasta). Joondamiseks kasutati Burrows-Wheeler Aligner algoritmi (v0.77) (H. Li ja Durbin, 2009). Saadud SAM failide binaarsesse formaati konventeerimiseks kasutati samtools (v0.1.19)(H. Li jt., 2009) ja Picard (v1.136) tööriistu.

2.2.3 Variantide analüüs

Variantide määramiseks kasutati Broad instituudis tarkvara Genome Analysis Toolkit (lühend GATK, v3.4-46) (McKenna jt., 2010). Broad instituudist saadeti TÜ EGV teadlastele geenidoonorite täisgenoomide andmed Variant Call Format failidena, mille edasist variantide annotatsiooni teostas TÜ EGV spetsialist Mart Kals. Variantide efekti geenifunktsiooni ja valku struktuurile hinnati Ensembl Variant Effect Predictor (lühend VEP, v84) tööriistaga (McLaren jt., 2010).

2.2.4 Proovide ja variantide kvaliteedikontroll

Proovide ja variantide kvaliteedietapid on ära toodud joonisel 7. Esialgsest 2304 proovist seitsme puhul ei saadud kvaliteetset raamatukogu. Sekveneeritud proovidest üheksal juhul oli kontaminatsiooni aste kõrgem lubatud 10%, mistõttu need eemaldati analüüsist. 2284 proovi sekventsi analüüsiti GATK tarkvaraga, kuid 40 proovi ei vastanud kõigile kvaliteedinäitajatele ning jäid lõppvalimist välja. Neli proovi on väljapool geenidoonorite valimit ning neid antud töös ei vaadeldud.



Joonis 7. Proovide ning esmaste variantide kvaliteedikontrolli etapid

2.2.5 Funktsioonikaoga mutatsioonide analüüs

LoF mutatsioonide annotatsioon toimus VEP tööriista LOFTEE[1] (ingl. *Loss-of-function Transcript Effect Estimator*) pluginit kasutades, mis hindas variantide mõju valgu funktsioonile. LOFTEE hindab ainult neid variante, mis põhjustavad genoomis:

- enneaegset stoppkoodonit
- splaiss-saidi rikkumist
- raaminihet

LOFTEE plugin filtreerib välja

1. enneaegse stoppkoodoni ja raaminihkemutatsiooni variandid kui variant:
    - asub transkripti viimase 5% sees

[1] https://github.com/konradjk/loftee

- asub eksonis, mida ümbritseb mittekanooniline splaiss-sait

2. splaiss-sait variandid kui mutatsioon:
    - asub ≤ 15bp pikkuses intronis
    - asub mitte-kanoonilise splaiss-saidiga intronis
3. kõikide variantide seast kui variandi:
    - LoF alleel on eellasalleel (primaatide seas)

Edasised filtreerinud ning andmete interpretatsiooni viis läbi töö autor, võttes aluseks MacArthuri tööprotsessi (MacArthur jt., 2012). Sorteeriti välja kõik variandid, mis läbisid GATK Variant Quality Score Recalibration filtreeringu. Lisaks rakendati kolme filtrit, võttes välja mutatsioonid, mis asusid madala komplekssusega alades, segmentaalsetes duplikatsioonides ja/või tandeemsetes järjestustes. Antud andmehulgast võeti edasiseks analüüsiks ainult LoF variandid.

2.2.5.1 Immuungeenid

Immuunsüsteemi ja selle kõikvõimalike protsessidega seotud geenid said allalaaditud USA Rahvusliku Terviseinstituudi (NIH) Immunoloogilise Andmebaasi ja Analüüsi Portaalist (ImmPort[2], The Immunology Database and Analysis Portal) (Heng jt., 2008), vaadeldavaid geene oli kokku 6540. LoF variantide seast sorteeriti immuungeenides leiduvad variandid kasutades selleks koostatud skripti. Saadud variantide hulgast kasutati edasiseks analüüsiks madalama alleelisagedusega (MAF<2%) homosügootseid leide.

2.3 Tulemused ja arutelu

Võrreldes ühe inimese DNA järjestust referentsgenoomiga, võib leida kuni 3 miljonit erinevust. Enamik neist on sagedasti esinevad ning populatsioonide vahel jagatavad variandid, kuid on ka suur hulk harvaesinevaid ning haruldasi mutatsioone. Laialdaselt kasutusel olevate ülegenoomsete assotsiatsiooniuuringute markerite seas on haruldaste variantide osa väheesindatud. Viimaste aastate täisgenoomide sekveneerimisprojektid on aga näidanud kui ulatuslik on inimestevaheline geneetiline varieeruvus, ka eripopulatsioonide vahel.

[2] https://immport.niaid.nih.gov/

Kokku leidsime EGV doonorite täisgenoomides 28 815 114 SNV-d ja 2 801 178 indelit. Võrreldes seniste avaldatud populatsioonigenoomika tulemustega oleme leidude arvult sarnases vahemikus (tabel 5). 72% kaardistatud leidudest oli haruldased variandid (MAF<0,5%), 15 385 146 leidub vaid ühel indiviidil kogu valimist (AC=1 või AC=2).

Tabel 5. Filtreeritud variandid Eesti populatsioonis võrreldes teiste populatsioonide täisgenoomidega.

	GoNL	Island	UK10K	**TÜ EGV**
SNV	20 400 000	19 689 642	42 000 000	**28 815 114**
Indel	1 200 000	1 441 572	3 500 000	**2 801 178**

Täisgenoome on rohkearvuliselt analüüsitud üle kogu maailma, kuid süstemaatilist funktsioonikaoga variantide kirjeldamist võib leida alles viimase paari aasta publikatsioonidest.

2.3.1 Funktsioonikaoga mutatsioonid Eesti populatsioonis

Eesti populatsiooni analüüsist leidsime kokku 14 438 potentsiaalset funktsioonikaoga mutatsiooni, variandid asusid 7826 erinevas geenis. SNV-sid oli kokku 7146 ja indeleid 7292, vastavalt 5004 ning 4726 erinevas geenis. Üle poolte (57%) leidudest esinesid vaadeldud populatsioonis vaid korra (AC=1). Kõigist funktsionaalsetest mutatsioonidest esineb harvade variantide seas kõige enam LoF-e. EGV LoF leidudest 92% olid haruldased variandid (MAF<0,5%). Sarnased tulemused on ka teistel projektidel – Islandi analüüsis leitud LoF mutatsioonidest olid haruldased 85% (Sulem jt., 2015). Tabel 6 iseloomustab kõigi LoF variantide jaotust alleelisageduste kaupa ning. Pigem madalamad/harvaesinevad alleelisagedused viitavad, et tegemist on deleterioosse mutatsiooniga ning seega võivad olla haigususseoselised. Kokkuvõttev tabel LoF arvude kohta kromosoomi kaupa on lisas 1 (lehekülg 56).

Tabel 6. Funktsioonikaoga mutatsioonide jaotus alleelisageduste kaupa. Tabeli allosas on väljatoodud variantide hulk, kus alleeli esineb kogu populatsioonis kaks või vähem korda

MAF	**SNV**	**Indel**
<0,5%	6630	6710
0,5-2%	270	267
2-5%	84	97
>5%	207	367
	7191	**7441**
AC=1	4155	4106
AC=2	879	1009

MAF = minoorse alleeli sagedus
AC = alternatiivse alleeli koguarv

Funktsioonikaoga mutatsioonide koguarv kirjanduses on väga varieeruv, kõikudest mõnesajast leiust mitmekümne tuhandeni (tabel 7). Siinkohal tuleb arvesse võtta valimi suurust, sekveneerimise sügavust (kattuvust), filtreeringute rangust ning populatsioonide geneetilist erinevust. MacArthur jt kasutasid oma uuringus 185 indiviidi 1000 genoomi projekti andmeid, nende LoF mutatsioonide koguarv oli 2951 (D. G. MacArthur jt., 2012). UK10K projektis vaadeldi kordades rohkem indiviide (N=3781) ning nende täisgenoomide analüüsist saadi kokku 14 516 funktsioonikaoga mutatsiooni (Walter jt., 2015), mis on väga lähedane ka EGV leidude koguarvule.

Tabel 7. LoF variantide esinemine avaldatud täisgenoomide uuringus võrreldes Eesti Geenivaramu kohordiga.

	MacArthur	**Holland**	**Island**	**UK10K**	**EXAC**	**EGV**
Kokku LoF mutatsioone	2951	NA	6795*	14 516	179 774	14 438
Geene	NA	NA	4924	NA	NA	7826
Täielikult nokaut geene	NA	NA	1171	576	3230	596
Indiviidi kaupa						
Hom	18	NA	21	NA	35	20
Het	79	NA	128	NA	85	106
Kokku	**97**	**144**	**149**	**NA**	**120**	**126**

*Islandi projektis vaadeldi autosomaalseid variante

Homosügootseid LoF mutatsioone leidus EGV populatsioonis 726 varianti 596 erinevas geenis. Veerand variantidest (N=167) olid harvaesinevad (MAF <1%), sealjuures 21% esinesid valmis vaid ühel indiviidil. Kõige enam oli esindatud raaminihet põhjustavad mutatsioonid (N=327). Selliste mutatsioonide efekt geeniproduktile võib olla drastiline, kuid leidub [LM1]ka päästvaid variante – näiteks samal kromosoomil LoF lähedal paiknev alternatiivne mutatsioon võib geeni funktsiooni päästa. Geeniekspressiooni mõõtmine oleks üks lahendus homosügootsete leidude tegeliku mõju hindamisel.

Kui mutatsioonide koguhulk võib varieeruda sõltuvalt analüüsi käigust ja valimist, siis kõikide tabelis 8 väljatoodud projektide LoF arvud indiviidi kohta on aga küllaltki sarnased. EGV tulemuste põhjal võib öelda, et igal indiviidil on keskmiselt 20 täielikku geeni nokautmutatsiooni ja 106 heterosügootset leidu. Maksimaalne ühe indiviidi funktsioonikaoga mutatsioone arv tuli 159, minimaalne jäi alla 100.

Tabel 8. Funktsioonikaoga mutatsioonide arv indiviidi kohta, homosügootsed (HOM) ning heterosügootsed (HET) leiud on eraldi väljatoodud.

	HOM	HET	KOKKU
MAX	40	139	159
MIN	5	75	96
KESKMINE	20,32	106,39	126,71

Max= maksimaalne
Min = minimaalne

Tänaseks on kirjeldatud üle 6000 mendeliaarse haiguse (OMIM[3], ingl. *online Mendelian Inheritance in Man*) ning inimese geenimutatsiooni andmebaasist (HGMD[4], ingl. *the Human Gene Mutation Database*) leiab rohkem kui 150 000 haigusseoselist varianti (Stenson jt., 2014). Puudub aga andmebaas, mis kirjeldaks haigusseoselisi geene koos (valideeritud) põhjuslike mutatsioonidega sama hästi kui kliinilised andmed. Tabel 9 iseloomustab leidude esinemist enim kasutamist leidvates andmebaasides nagu dbSNP, HGMD, ClinVar (Landrum jt., 2014) ja ExAC. 14 438 variandist 80% on juba esindatud dbSNP andmestikus, ExAC-s leidus vaid 61% EGV variantidest. 726-st geeninokautmutatsioonist esines HGMD ja ClinVar andmebaasides leidudest vastavalt 10% ja 1,7%.

[3] http://www.omim.org/
[4] http://www.hgmd.cf.ac.uk/ac/hahaha.php

Tabel 9. Eesti populatsiooni 14 438 LoF mutatsiooni leidumine teistes andmebaasides

	Puudub (N=14438)	**AC=1** (N=8261)	**< 0,5%** (N=13 340)	**0,5 – 2%** (N=537)	**> 5%** (N=561)	**Hom** (N=726)
dbSNP	11653	859	2016	317	320	551
HGMD	13977	186	370	45	22	66
ClinVar	14243	93	175	9	8	12
ExAC	8802	2679	5415	445	325	605

Eripopulatsioonide sekveneerimine annab informatsiooni inimese geneetilise varieeruvuse kohta väga erinevates keskkondades. Populatsiooni pudelikaela efekt põhjustab harvaesinevate variantide rikastumist teatud aja jooksul (Lim jt., 2014; Sidore jt., 2015). Hinnanguliselt esineb enamik Euroopa populatsioonidest homogeensemas soomlaste populatsioonis rohkem madala esinemissagedusega funktsioonikaoga mutatsioone. Lim jt. toovad 2014. aasta artiklis välja 83 varianti, mis väidetavalt esinevad soomlaste seas sagedamini (Lim jt., 2014). Nendest leidusest kaheksa tükki esinesid EGV andmestikus sarnase või kõrgema alleelisagedusega (tabel 10).

Tabel 10. Soome populatsioonis rikastatud 83 leiust kaheksa LoF varianti, mille alleelisagedus eestlaste seas on kõrgem

Krom	**Positsioon**	**Ref**	**Alt**	**Tüüp**	**Geen**	**SiSu AF**	**Eesti AF**
1	89729566	GC	G	frameshift	*GBP5*	0,0101	0,0102
4	76521525	C	T	splice	*CDKL2*	0,0195	0,0221
7	143048771	C	T	stop	*CLCN1*	0,0154	0,0205
9	21304913	G	A	stop	*IFNA5*	0,0166	0,0272
14	94754643	C	T	stop	*SERPINA10*	0,0166	0,0207
14	94756669	G	A	stop	*SERPINA10*	0,0008	0,0036
16	47495300	G	A	stop	*PHKB*	0,0144	0,0149
17	28268857	G	A	splice	*EFCAB5*	0,0160	0,0553

Krom = kromosoom
Ref = referentsjärjestus
Alt= alternatiivne järjestus
AF = alleelisagedus

EGV täisgenoomides leiduvad funktsioonikaoga mutatsioonid on uudne andmestik Eesti populatsiooni haruldastest variantidest, mis omab suurt potentsiaali eestlaste geneetiliste variantide efektide interpretatsioonil, kuna meil on neile kõrvutada rikkalikke terviseandmeid.

2.3.2 Funktsioonikaoga mutatsioonid immuungeenides

Immuunsüsteemiga seotud geenides esines 1920-s (29%-l) vähemalt üks funktsioonikaoga mutatsioon. Kokku oli 3623 varianti, millest üle 90% oli madala alleelisagedusega (<1%) (tabel 11). Geeni potentsiaalseid nokautmutatsioone ehk homosügootseid leide oli immuungeenides kokku 184 (128 erinevas geenis). 81 homosügootset varianti esinesid valimis viiel või vähemal indiviidil.

Tabel 11. Funktsioonikaoga mutatsioonide jaotus immuungeenides alleelisageduste kaupa. Tabeli allosas on väljatoodud variantide hulk, kus alleeli esineb kogu populatsioonis kaks või vähem korda.

MAF	**SNV**	**Indel**
<0,5%	1681	1659
0,5-2%	66	70
2-5%	21	34
>5%	33	59
	1801	**1822**
AC =1	1057	1064
AC=2	198	235

MAF = minoorse alleeli sagedus
AC = alternatiivse alleeli koguarv

Käesolevas töös oli põhifookuses homosügootsete leidude potentsiaalsete seoste analüüs. Kõigist leidudest vaatlesime homosügootseid variante, mille MAF jäi alla 2% (N=60). Hinnates geenide bioloogilist mõju, kirjandusest juba kirjeldatud mutatsioone ja seoseid ning variantide esinemissagedust teistes andmebaasides (kõige enam tuginedes ExAC-ile), jäid sõelale 16 potentsiaalselt huvitavat leidu. Puuduva (või vähese) fenotüübi andmete, ebaselgete seoste ning valepositiivsete tulemuste tõttu eemaldasime 16-st üheksa. Antud töös käsitlen järelejäänud seitsme funktsioonikaoga mutatsiooni võimalikke bioloogilisi efekte. Leitud mutatsioonid paiknevad järgnevates geenides: *SMIM1*, *IFNE NOD2*, *NCR3LG1*, *C2*, *CXCR3*, ja *IL8* .

2.3.2.1 *SMIM1*

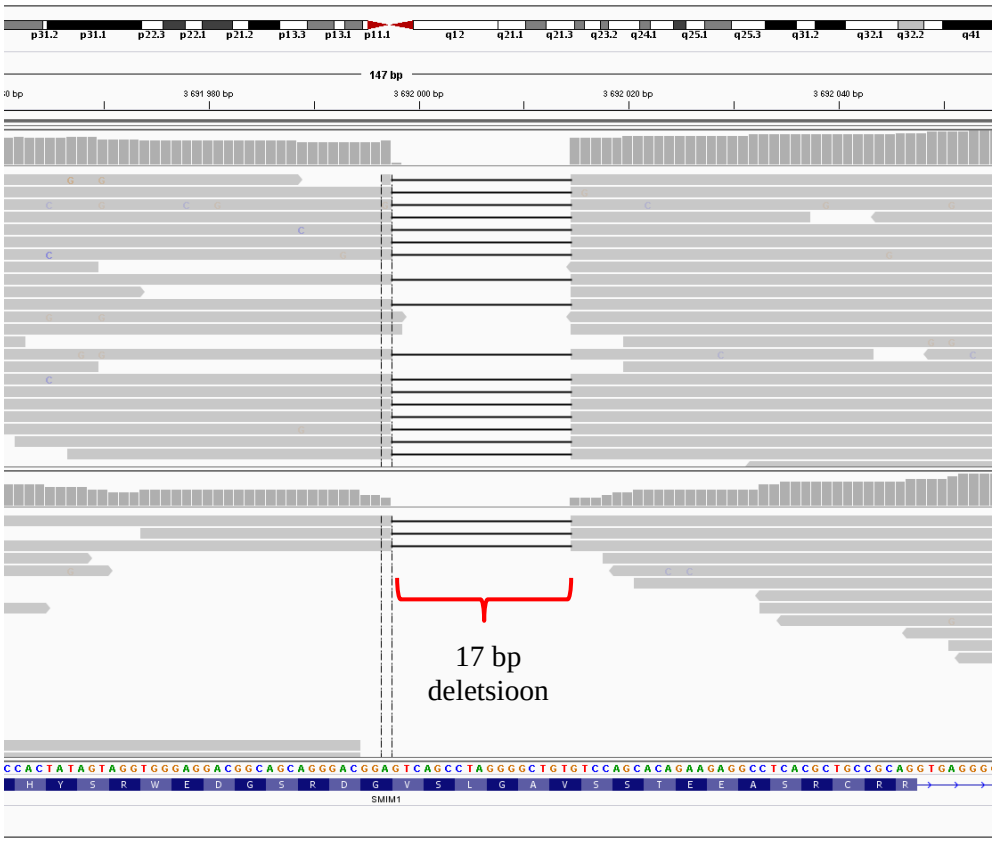
Alleelide variatsioonist tingitud erinev valkude olemasolu erütrotsüütide pinnal määrab antigeenide olemasolu ning seeläbi veregruppi kuuluvuse. Haruldane veregrupp Vel avastati juba 1950-ndatel mitmete vereülekannete ebaõnnestumiste analüüsimisel (Sussman ja Miller, 1952). Vel veregrupi esinemissagedus eurooplaste seas on mõnede allikate andmeil 1:4000 (Daniels, 2002) ja tunduvalt kõrgem skandinaavlastel (1:1200) (Cvejic jt., 2013; Jill R Storry jt., 2013). 2013. aastal avaldatud artiklis kirjeldati esmakordselt Vel veregrupiga seotud lookust ja negatiivse veregrupi põhjuslikku varianti (Jill R Storry jt., 2013).

SMIM1 geen asub 1. kromosoomi lühemas õlas (1p36.32) ning koosneb neljast eksonist. Geen vastutab 78 aminohappe pikkuse transmembraanse valgu kodeerimise eest (B2RUZ4, UniProt ID). *SMIM1* geeni ekspressioon on spetsiifiline luuüdi koele, vähem leidub seda mitte-hematopoeetilistes kudedes (Jill R Storry jt., 2013). Storry jt kirjeldavad 17 nukleotiidi pikkust deletsiooni *SMIM1* geeni kolmandas eksonis, mis põhjustab raaminihke valgu transmembraanset domääni kodeerivas alas (Jill R Storry jt., 2013). Geenis on kaardistatud ka ühenukleotiidseid muutusi, kuid Storry jt. kirjeldatud deletsioon on ainuke funktsionaalselt valideeritud Vel-negatiivse veregrupi põhjuslik variant. Leiu dbSNP-i rs number on rs566629828 ning mutatsiooniga seotud fenotüüp on kirjeldatud ka OMIM andmebaasis (615264).

Haruldane Vel-negatiivne fenotüüp ei näita kõigil isikutel kliinilisi ilminguid. Vel-positiivse vere ülekandel produtseeritakse anti-Vel antikehad, mis ründavad organismi erütrotsüüte ja põhjustavad hemolüütilisi reaktsioone (Jill R Storry jt., 2013). Sellised immuunreaktsioonid võivad viia tõsiste tagajärgedeni nagu organite (eriti neerude) puudulikkuse ja surmani (Daniels, 2002; J. R. Storry ja Mallory, 1994).

Analüüsides immuungeenide funktsioonikaoga mutatsioone leidsime EGV doonorite hulgast kaks indiviidi (doonor 1 ja doonor 2), kelle *SMIM1* geeni mõlemas alleelis esineb eespool kirjeldatud mutatsioon (joonis 8). Selle variandi alleelisageduseks tuleb 0,0116, mis on tunduvalt sagedasem kirjandusest leitavatest numbritest. ExAC andmebaasis antud mutatsiooni suhtes homosügoote ei leidu, kuid alleeli esinemine (AF=0,0107) ei ole oluliselt madalama sagedusega EGV tulemusest. Ühe mittefunktsionaalse alleeli kandjaid seostatakse madalama hemoglobiinitaseme ja erütrotsüütide arvuga (Cvejic jt., 2013). EGV valimis

leidub 48 heterosügooti. Vel-negatiivse veregrupi esindajate leidmine valimist on praktiline näide biopanga genotüübi andmete võimalikust rakendusest igapäevases kliinilises töös.



Joonis 8. *SMIM1* geeni 17 aluspaari pikkune deletsioon kolmandas eksonis. Pildil Doonor 1 ja Doonor 2 joondatud järjestused antud lookuses.

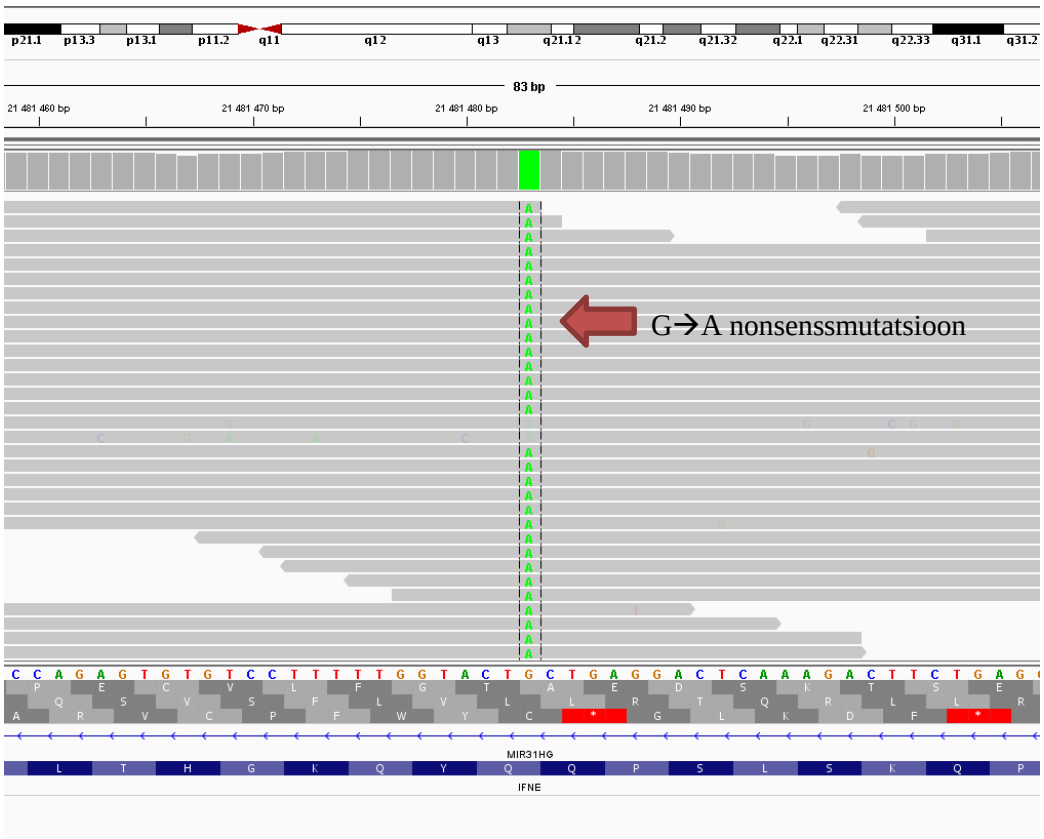
2.3.2.2 *IFNE*

Tsütokiinid on mediaatorid, mis vastutavad rakkudevahelise kontakti eest erinevate immuunreaktsioonide korral (Velbri, 2002). Tsütokiinide põhirühma kuuluvad interferoonid (IFN) on valkude grupp, mis osalevad rakkude signaaliradades, ning seda valgu perekonda saab jagada kolmeks – alfa, beeta ja gamma INF-d (De Andrea, Ravera, Gioia, Gariglio, ja Landolfo, 2002). *IFN* geeniperekonna valgud mõjutavad eelkõige viirusinfektsiooni levikut, kuid seondudes spetsiifiliste retseptoritega aktiveeritakse rida geene, mis on peale viirusinfektsiooni seotud raku jagunemise ja immuunsüsteemi aktivatsiooniga (De Andrea jt., 2002; Velbri, 2002). Inimese genoomist on identifitseeritud üle 20 IFN geeni ja neid jagatakse kolme klassi – tüüp I-III.

IFN-ε (*INFE*) kuulub *IFN* geeniperekonda ning jaotub alamklassi tüüp I. *IFNE* geenieksрessiooni võib leida mitmetes kudedes nagu aju, pärgarteri ja mikrovaskulaarsetes endoteelirakkudes (Pestka, Krause, ja Walter, 2004). *IFNE* geen indutseeritakse läbi IFN retseptorite faktorite nagu põletikutsütokiinid (TNF-α, IL-1 ja IL-6) ja viiruste poolt (Hardy, Owczarek, Jermiin, Ejdebäck, ja Hertzog, 2004). *IFNE* geen asub klasterdunult koos teiste tüüp I *IFN* geenidega üheksanda kromosoomi lühemal õlal (9p21.3). *IFNE* kodeeritav valk (Q86WN2, UniProt ID) koosneb 208 aminohappest, millest aminohapped 1-21 moodustavad N-terminaalse signaalpeptiidi ja aminohapped 22-208 moodustavad polüpeptiidse IFNE ahela (http://www.uniprot.org).

Kirjandusest teadaolevat vastutab *IFNE* aju struktuuri ja funktsiooni eest (Peng jt., 2007), on seotud vitiliigoga Korea populatsioonis (Cho jt., 2013) ning omab potentsiaalset seost ka insuldi riskiga (Kim jt., 2014). Lisaks on näidatud *IFNE* spetsiifilist geenieksрessiooni naiste sugutee epiteelrakkudes, mis on hormonaalselt reguleeritud ning eriti kõrge menstruaalse tsükli proliferatiivses faasis (Fung jt., 2013). On ka teada, et kõik tüüp I IFN-d kaitsevad Herpes simplex viirus II (HSV-2) infektsiooni eest (Conrady, Halford, ja Carr, 2011). Fung jt. näitasid oma katsetes, et emastel *IFNE* -/- nokauthiirtel tõusis võrreldes *wildtype* hiirtega vastuvõtlikkus seksuaalsel teel levivate infektsioonide suhtes (HSV-2 ja *Chlamydia muridarum*-i poolt põhjustatud infektsioonid) ning järeldasid IFNE tsütokiini antipatogeenset funktsiooni (Fung jt., 2013).

Täisgenoomide valimis leidus *IFNE* geenis varem kirjeldatud variant rs2039381 (Cho jt., 2013). Tegemist on nonsenssmutatsiooniga (joonis 9), mis põhjustab *IFNE* geeni transleerimisel eeldatava glutamiini (aminohappe positsioonis 71) asemel enneaegset stoppkoodonit (Gln→Stopp). Potentsiaalne LoF variant esines EGV valimis 51 indiviidi (AF=0.0116), ühel naissoost doonoril (Doonor 3) homosügootse leiuna. ExAC andmebaasis on andmeid Euroopa populatsioonis (va soomlased) ainult ühe homosügoodi kohta (AF=0,0038). Kokkuvõttev tabel *IFNE* variant rs2039381 alleelisageduste kohta on lisas 2 (lehekülg 57).



Doonor 3

Joonis 9. Doonor 3 *IFNE* geeni G→A muutus (rs2039381), mis põhjustab enneaegse stoppkoodoni (Gln71→Stop).

Doonor 3 Eesti Haigekassast saadud diagnooside hulgas on (märgitud ka diagnooside ICD-10 koodid) :

- Herpesviirusnakkused [herpes simplex], kood B00
- Emaka täpsustamata leiomüoom, kood D25.9
- Rinna üksiktsüst, kood N60.0
- Emaka täpsustamata põletikuline haigus, kood N71.9
- Kõhukeelme vaagnaosa endometrioos, kood N80.3
- Muud ja täpsustamata munasarjatsüstid, kood N83.2

Kõik loetletud haigused näitavad tõsiseid häireid naissuguelundite funktsioneerimisel. *IFNE* geeni spetsiifiline ekspresseerumine naise sugutee epiteelrakkudes viitab geeni olulisele rollile naise suguorganite töös. Fung jt. mudelorganismi katsete tulemused (Fung jt., 2013) ning Doonor 3 terviseandmed kinnitavad seose võimalikkust.

rs2039381 mutatsiooni kandjaid on valimis 50, nende seas naissoost doonoreid 21. Filtreerides heterosügootide diagnoose (Eesti Haigekassa andmed) leiame kõigil peale ühe doonori vähemalt ühe naissuguorganitega seotud haiguse: tupe ja häbeme põletik (N=11), emaka healoomuline kasvaja (N=7), *herpes simplex* nakkus (N=3), kubemesong (N=3), naiseinfertiilsus (N=3) jm. Lisaks esineb viiel indiviidil 21-st rinnamoodustisi (k.a rinnakasvaja ja –tsüstid). Kõrge naistehaiguste esinemine *IFNE* mutatsiooni kandjate seas viitab samuti potentsiaalsele *IFNE* geeni mittefunktsioneerimisest tingitud naissuguorganite epiteelkoega seotud patoloogiatele ning selle tagajärjelt haiguste tekkele.

Analüüsides, kas *IFNE* geenimutatsiooni kandjatel esineb suurem šanss naistehaiguste tekkeks võrreldes referents-homosügootidega, ei tulnud ühegi haigusgrupi ja mutatsiooni vahel statistiliselt olulist seost. Seega võib järeldada, et piisab ühest funktsionaalsest alleelist säilitamaks valgu normaalne talitlus. Et kinnitada heterosügootse variandi mõju fenotüübile võib käesolevas töös kasutatud valimisuurus jääda väikeseks. Antud leiu rohkemate homosügootsete indiviidide fenotüübi analüüs ning variandi suuremahulisem genotüpiseerimine võimaldaks täpsemalt hinnata *IFNE* geeni efekti fenotüübile.

Lisaks eelpool väljatoodud leidudele ilmnes immuungeenide LoF mutatsioonide homosügootseid indiviide ning nende fenotüüpe analüüsides veel potentsiaalseid variante, mille seoseid immuunsüsteemi häiretega tuleks edasi analüüsida (tabel 11).

Tabel 12. Potentsiaalsed ja võimalikud LoF leiud immuungeenides. ExAC andmed eurooplaste kohta (va soomlased).

Kr	Pos	dbSNP	Ref	Alt	AF	Het	Hom Alt	Tüüp	Geen	ExAC AF	ExAC AN;AC; Het;Hom	Potentsiaalne korrelatsioon
1	3691997	.	AGTCAGCCTAGGGGCTGT	A	0,0116	48	2	raaminihe	*SMIM1*	0,0107	7786;83; 83;0	Vel veregrupp
9	21481483	rs2039381	G	A	0,0116	50	1	enneaegne stoppkoodon	*IFNE*	0,0038	66444;254; 252;1	Naistehaigused
16	50763778	rs2066847	G	GC	0,0377	163	3	raaminihe	*NOD2*	0,0202	66698;1345; 1283;31	Allergia, astma, põletikulised soolehaigused
X	70837390	rs188959001	G	A	0,0087	21	9	enneaegne stoppkoodon	*CXCR3*	0,0025	10561;26; 20;0	Allergia, astma, vähk
4	74607285	rs188378669	G	T	0,0123	53	1	enneaegne stoppkoodon	*IL8*	0,0016	66426;104; 104;0	Seenhaigused
11	17394037	rs61406813	CTT	C	0,0363	161	1	raaminihe	*NCR3LG1*	0,0688	4566;314; 292;11	Kasvajad
6	31902065	rs9332736	ATGGTGGACAGGGTCAGGAATCAGGAGTC	A	0,0183	80	1	splaiss-sait muutus	*C2*	0,0081	66578;540; 540;0	Komplement 2 puudulikkus

Kr = kromosoom
Pos = positsioon
Ref = referentsjärjestus
Alt = alternatiivne järjestus
AF = alleelisagedus
Het = heterosügootide arv
Hom Alt = retsessiivne homosügoot
AN = alleeli koguarv
AC = alternatiivse alleeli koguarv

Elektroonilised terviseandmed kannavad endas suurt potentsiaali, kuid negatiivsest küljest on need disainitud eelkõige kliinilise ravi teostamiseks. EHR mittestandardiseeritud kuju tähendab, et erinevad testitulemused võivad paikned mitte struktureeritud vabateksti sees ning diagnooside ülesleidmine vajab tekstitöötlust. Seega on EHR andmete mõistmiseks vaja tõhusaid bioinformaatilisi algoritme (R. Cohen, Elhadad, ja Elhadad, 2013). Lähitulevikus on realistlik ootus, et miljonitel elektrooniliste terviseandmetega patsientidel on ka genotüübi informatsioon (suuresti tänu biopankadele) talletatud. Elektroonilised terviseandmed on võimas andmestik nii haruldaste kui ka sagedaste haiguste uurimiseks (Crawford jt., 2014; Hall jt., 2016). Seostades mutatsioone kindlate haigustega on kõrvale vaja uuritava indiviidi võimalikult põhjalikku fenotüübiandmestikku. Seega esindavad terviseandmed olulist komponenti geneetika uuringutes.

Haigusseoseliste geenide ja nende põhjuslike mutatsioonide kirjeldamiseks puudub hetkel mõni andmebaas, mis kirjeldaks seoseid sama hästi kui kliinilised andmed (Ginsburg, 2014). Käesolevas töös kaardistatud 14 483 LoF mutatsioonidest 8802 (60,77%) ei ole leitavad seni suurimas kodeerivate variantide ExAC andmebaasis. Küll aga esinevad ExAC andmestikus pea kõik EGV homosügootsed variandid (83,33%). Andmestikud nagu ExAC annavad võimaluse meditsiinigeneetikutel potentsiaalseid haigusseoselisi leide analüüsides kontrollida variandi olemasolu ning alleelisagedust (Quintáns, Ordóñez-Ugalde, Cacheiro, Carracedo ja Sobrido, 2014). Oluline ülesanne on koostada andmebaase, kuid sama tähtis on ka juba leitud variantide eksperimentaalne valideerimine. Üheks võimaluseks LoF mutatsioonide efekti reaalseks hindamiseks on kõrvale võtta transkriptoomi andmed ning kinnitada mutatsioonide ennustatavad tagajärjed ka RNA tasandil .

On keerukas ainult DNA variatsioonide põhjal väita, kas uuritav variant mõjutab geeniprodukti ja selle funktsiooni. Stopp-koodon võib küll lühendada valku, kuid lõpp-produktina võidakse sünteesida siiski töötav valk (Danckwardt, 2002; Isken ja Maquat, 2007). Splaiss-sait muutuste korral tuleb silmas pidada ka looduslikult esinevat alternatiivset splaissimist (Pan, Shai, Lee, Frey ja Blencowe, 2008). LoF mutatsioon võib küll rikkuda ära geeni mõne produkti, kuid alternatiivsete transkriptide abil võib organismis säilida valgu funktsioon.

Geeni annotatsioon ning variantide leidmine ei ole bioinformaatiliselt perfektne, vead võivad ilmneda juba esimestes analüüsietappides (Sims, Sudbery, Ilott, Heger ja Ponting, 2014). Lühikeste järjestuste joondamine võib tekitada genoomi kokkupanemisel vigu. Näiteks võivad

mõned mutatsioonid annoteerimisel jääda järjestamata või saavad määratletud ekslikult pseudogeenidena (Zheng jt., 2007).

Käesolevas töös iseloomustatud funktsioonikaoga mutatsioonide andmestik ei ole kasulik ainult populatsiooni uurimiseks, vaid võimaldab leida uusi seoseid geenide ja fenotüübi tunnuste vahel. Homosügootseid LoF variante, mida ei ole esindatud ExAC andmebaasis ja mille kohta on vähe (või mitte üldse) funktsionaalsete korrelatsioonide kirjeldust, on EGV valimis ligi sadakond. Doonorite fenotüüpe iseloomustavaid andmeid ja tervisenäitajaid kasutades võime leida uut informatsiooni nende vähe iseloomustatud geenide funktsioonide ja bioloogilise efekti kohta.

Vaadates ainult ühte tüüpi mutatsioone ning jättes kõrvale alternatiivsed genoomis leiduvad variatsioonid, võivad fenotüübi geneetilisi põhjuseid uurides jääda nägemata kombineerituna avalduvad juhud. Funktsioonikaoga mutatsioonidele lisaks oleks tuleviku analüüsis huvitav kaasata analüüsi ka mittekodeerivas alas paiknevad variandid, mis võivad mõjutada erinevaid regulatoorseid elemente ja seeläbi muuta geenieksрessiooni taset. Selliste mutatsioonide mõju ei ole üks ühele seoses fenotüübiga, samuti võib geenieksрessioon erineda kudede ja rakutüüpide vahel.

KOKKUVÕTE

Alates esimesest täisgenoomi sekveneerimisest on järjestatud mitmeid tuhandeid genoome üle terve maailma. Tehnoloogia areng on viinud meid ajastusse, kus on võimalik analüüsida korraga kõiki geene ning seeläbi mõista bioloogilise keerukuse molekulaarset tausta. Genoomiliste andmete maht on kordades kasvanud, mis omakorda on andnud uusi teadmisi genoomi mitmekesisusest. Struktuursete ja koopia arvu variatsioonide ning ühenukleotiidsete polümorfismide efekt fenotüübile on suurem kui siiani arvati.

Käesoleva magistritöö eesmärgiks oli anda ülevaade seni tehtud populatsiooni täisgenoomide järjestamise projektide kohta, võrrelda üldtulemusi Eesti Geenivaramu 2240 geenidoonori sekveneerimisandmetega ning iseloomustada funktsioonikaoga mutatsioone Eesti populatsioonis. Järjestatud eestlaste genoomidest leidsime 14 483 funktsioonikaoga mutatsiooni, millest 7146 on SNV ja 7292 indelid. Variandid paiknesid 7826 erinevas geenis ning 92% kõigist leidudest olid haruldased (MAF<0,5%). Keskmiselt on igal indiviidil 126 LoF-i, millest ~20 esineb mõlemas alleelis. 80% leidudest on juba kirjeldatud dbSNP andmebaasis.

Magistritöö veel üheks eesmärgiks oli hinnata biopanga andmete kasutamisvõimalust geneetiliste seoste leidmisel analüüsides väljavalitud immuungeenidest saadud funktsioonikaoga mutatsioone ja Eesti Haigekassa terviseandmeid. Immuungeenides (N=6540) leidus 3623 funktsioonikaoga mutatsiooni, kusjuures üle 90% neist olid harvaesinevad (MAF<1%). Antud töö fookuses olid homosügootsed LoF leiud, mida immuungeenides leidus pea 200. Hinnates variantide esinemissagedusi erinevates andmebaasides, geenide bioloogilist mõju ning kirjandusest juba kirjeldatud seoseid, on magistritöös välja toodud seitse potentsiaalset geneetilist varianti edasiseks analüüsiks. Nende seas *SMIM1* geeni kolmanda eksoni varasemalt kirjeldatud 17 bp pikkune deletsioon (Jill R Storry jt., 2013), mis põhjustab Vel-negatiivset veregruppi, ning *IFNE* geeni nonsenssmutatsiooni. Viimase puhul on kirjandusele tuginedes ja Eesti Haigekassa terviseandmeid hinnates võimalik seos naistehaigustega.

Analysing loss-of-function mutations by pairing 2300 whole genomes with electronic health records

SUMMARY

Kelli Grand

Sequencing genetic material has become accessible for large-scale population analysis. Characterising genetical elements within and between populations gives us insight of how genetic variation influences health. Genomic data and its grown volume has shown us the extent of diversity in genomes. The effect of different structural and copy number variations has been underestimated. Rare and low-frequency variants and their contribution to phenotype are largely unknown. These variants are underrepresented in common genome-wide association studies. As they are seen only in very few people, it is necessary to sequence large selection of individuals for detecting rare variants.

But it is not enough to only have the genotype data for interpreting potential biological functions of variants. Today, increasingly biobanks all around the world are starting to understand the powerful resource of electronic health records (EHR). EHR has the potential for interpret human variations producing clinically relevant phenotypic changes.

The purpose of this study was to give an overview of population based whole-genome sequencing (WGS) studies done so far (Francioli et al., 2014; Gudbjartsson et al., 2015; Lek et al., 2015; Lim et al., 2014; Sudmant et al., 2015; Walter et al., 2015), to compare the results with variants identified in the Estonian population by sequencing 2240 genomes, and to characterize loss-of-function (LoF) mutations in the Estonian population. We found 14,483 loss-of-function mutations in 7826 genes, 7146 SNV and 7292 indels in total. Most of the loss-of-function mutations were rare, with 92% having a MAF less than 0.5%. As shown in previous studies, from functional mutation classes LoF mutations have the highest fraction of rare variants (Gudbjartsson et al., 2015; D. G. MacArthur et al., 2012; Sulem et al., 2015).

It has been estimated by D. MacArthur, that every healthy person has 100 disrupted genes in the genome, approximately 20 totally inactivating the gene. These findings were confirmed in the Estonian population as well. On average, every individual from Estonian cohort carries

129 genuine LoF variants, with ~20 of them being in homozygous state. The maximum amount of LoF mutations found in one person was 159. Of all the LoF findings, 80% were already in dbSNP.

Another aim of this study was to evaluate the effect of the identified rare variants by investigating the electronic heath records of the individuals (from Estonian Health Insurance Fund). For that we analysed 6500 immune related genes and their loss-of-function mutations in combination with EHR data. All in all, there were 3623 LoF mutations in immune related genes but we focused on the mutations knocking out both alleles in one's genome. Analysing the occurrence in other widely known databases, biological functions and associations found in literature we highlight seven genes (*SMIM1*, *IFNE NOD2, NCR3LG1, C2, CXCR3*, and *IL8*) with potential LoF variants for further analysis. Among these variants, there was a known 17 nucleotides deletion in the *SMIM1* gene that cause a rare Vel-negative blood group (Cvejic et al., 2013) and a LoF mutation in the *IFNE* gene (Conrady et al., 2011) that might be potentially correlated with different disorders of the female reproductive tract.

TÄNUAVALDUSED

Kõigepealt tahaksin tänada oma juhendajaid Lili Milani ja Pärt Petersoni, kes igal hetkel nõu ja jõuga abiks olid. Teie näpunäiteid ja õpetussõnu hindan väga. Teiseks tänan väga oma armast pere, kes nii kannatlikult minu kõrval alati toeks on. Ilma teieta ei oleks ma kunagi nii kaugele jõudnud! Lõpetuseks tahaksin tänada oma kolleegi Mart Kalsi igasuguse abi ja nõu eest, kõiki reedeste WGS seminaride osalejaid edasiviiva konstruktiivse kriitika eest ning sõpru toetavate ja julgustavate sõnade eest.

KASUTATUD KIRJANDUS

Abyzov, A., Urban, A. E., Snyder, M., ja Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*, *21*(6), 974–984. http://doi.org/10.1101/gr.114876.110

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., … Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics*, *41*(10), 1061–1067. http://doi.org/10.1038/ng.437

Aoshima, M., Nunoi, H., Shimazu, M., Shimizu, S., Tatsuzawa, O., Kenney, R. T., ja Kanegasaki, S. (1996). Two-exon skipping due to a point mutation in p67-phox--deficient chronic granulomatous disease. *Blood*, *88*(5), 1841–1845.

Baralle, D., ja Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. *Journal of Medical Genetics*, *42*(10), 737–748. http://doi.org/10.1136/jmg.2004.029538

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., … Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, *40*(8), 955–962. http://doi.org/10.1038/ng.175

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. http://doi.org/10.1038/nature07517

Calafell, F., Roubinet, F., Ramírez-Soriano, A., Saitou, N., Bertranpetit, J., ja Blancher, A. (2008). Evolutionary dynamics of the human ABO gene. *Human Genetics*, *124*(2), 123–135. http://doi.org/10.1007/s00439-008-0530-8

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., … Mardis, E. R. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*, *6*(9), 677–681. http://doi.org/10.1038/nmeth.1363

Cho, H.-R., Kim, S. K., Lim, H.-K., Jeong Park, H., Chung, J.-H., ja Lee, M.-H. (2013). Association study between nonsense polymorphism (rs2039381, Gln71Stop) of interferon-ε and susceptibility to vitiligo in Korean population. *Immunological Investigations*, *42*(5), 423–430. http://doi.org/10.3109/08820139.2013.804836

Coe, B. P., Chari, R., MacAulay, C., ja Lam, W. L. (2010). FACADE: a fast and sensitive algorithm for the segmentation and calling of high resolution array CGH data. *Nucleic Acids Research*, *38*(15), e157. http://doi.org/10.1093/nar/gkq548

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., ja Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nature Genetics*, *37*(2), 161–165. http://doi.org/10.1038/ng1509

Cohen, R., Elhadad, M., ja Elhadad, N. (2013). Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, *14*(1), 10. http://doi.org/10.1186/1471-2105-14-10

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., … Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–712. http://doi.org/10.1038/nature08516

Conrady, C. D., Halford, W. P., ja Carr, D. J. J. (2011). Loss of the Type I Interferon Pathway Increases Vulnerability of Mice to Genital Herpes Simplex Virus 2 Infection. *Journal of Virology*, *85*(4), 1625–1633. http://doi.org/10.1128/JVI.01715-10

Crawford, D. C., Crosslin, D. R., Tromp, G., Kullo, I. J., Kuivaniemi, H., Hayes, M. G., … Ritchie, M. D. (2014). eMERGEing progress in genomics-the first seven years. *Frontiers in Genetics*, *5*, 184. http://doi.org/10.3389/fgene.2014.00184

Crosby, J., Peloso, G. M., Auer, P. L., Crosslin, D. R., Stitziel, N. O., Lange, L. A., … Kathiresan, S. (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *The New England Journal of Medicine*, *371*(1), 22–31. http://doi.org/10.1056/NEJMoa1307095

Crosslin, D. R., McDavid, A., Weston, N., Nelson, S. C., Zheng, X., Hart, E., … Jarvik, G. P. (2012). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Human Genetics*, *131*(4), 639–652. http://doi.org/10.1007/s00439-011-1103-9

Cvejic, A., Haer-Wigman, L., Stephens, J. C., Kostadima, M., Smethurst, P. A., Frontini, M., … Albers, C. A. (2013). SMIM1 underlies the Vel blood group and influences red blood cell traits. *Nature Genetics*, *45*(5), 542–545. http://doi.org/10.1038/ng.2603

Danckwardt, S. (2002). Abnormally spliced beta -globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. *Blood*, *99*(5), 1811–1816. http://doi.org/10.1182/blood.V99.5.1811

Daniels, G. (2002). *Human Blood Groups*. Blackwell Science, Oxford.

De Andrea, M., Ravera, R., Gioia, D., Gariglio, M., ja Landolfo, S. (2002). The interferon system: an overview. *European Journal of Paediatric Neurology: EJPN: Official Journal of the European Paediatric Neurology Society*, *6 Suppl A*, A41-46-58.

de Morais, S. M., Wilkinson, G. R., Blaisdell, J., Nakamura, K., Meyer, U. A., ja Goldstein, J. A. (1994). The major genetic defect responsible for the polymorphism of S-mephenytoin metabolism in humans. *The Journal of Biological Chemistry*, *269*(22), 15419–15422.

Denny, J. C., Crawford, D. C., Ritchie, M. D., Bielinski, S. J., Basford, M. A., Bradford, Y., … de Andrade, M. (2011). Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. *The American Journal of Human Genetics*, *89*(4), 529–542. http://doi.org/10.1016/j.ajhg.2011.09.008

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., … Crawford, D. C. (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, *26*(9), 1205–1210. http://doi.org/10.1093/bioinformatics/btq126

Denny, J. C., Ritchie, M. D., Crawford, D. C., Schildcrout, J. S., Ramirez, A. H., Pulley, J. M., … Roden, D. M. (2010). Identification of Genomic Predictors of Atrioventricular Conduction: Using Electronic Medical Records as a Tool for Genome Science. *Circulation*, *122*(20), 2016–2021. http://doi.org/10.1161/CIRCULATIONAHA.110.948828

Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., … McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. http://doi.org/10.1038/nature09534

Eriksson, N., Tung, J. Y., Kiefer, A. K., Hinds, D. A., Francke, U., Mountain, J. L., ja Do, C. B. (2012). Novel Associations for Hypothyroidism Include Known Autoimmune Risk Loci. *PLoS ONE*, *7*(4), e34442. http://doi.org/10.1371/journal.pone.0034442

Farris, W., Mansourian, S., Leissring, M. A., Eckman, E. A., Bertram, L., Eckman, C. B., … Selkoe, D. J. (2004). Partial loss-of-function mutations in insulin-degrading enzyme that induce diabetes also impair degradation of amyloid beta-protein. *The American Journal of Pathology*, *164*(4), 1425–1434.

Feero, W. G., Guttmacher, A. E., Cho, J. H., ja Gregersen, P. K. (2011). Genomics and the Multifactorial Nature of Human Autoimmune Disease. *New England Journal of Medicine*, *365*(17), 1612–1623. http://doi.org/10.1056/NEJMra1100030

Festen, E. A. M., Goyette, P., Green, T., Boucher, G., Beauchamp, C., Trynka, G., … Rioux, J. D. (2011). A Meta-Analysis of Genome-Wide Association Scans Identifies IL18RAP, PTPN2, TAGAP, and PUS10 As Shared Risk Loci for Crohn's Disease and Celiac Disease. *PLoS Genetics*, *7*(1), e1001283. http://doi.org/10.1371/journal.pgen.1001283

Flannick, J., Thorleifsson, G., Beer, N. L., Jacobs, S. B. R., Grarup, N., Burtt, N. P., … Altshuler, D. (2014). Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nature Genetics*, *46*(4), 357–363. http://doi.org/10.1038/ng.2915

Francioli, L. C., Menelaou, A., Pulit, S. L., van Dijk, F., Palamara, P. F., Elbers, C. C., … Wijmenga, C. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, *46*(8), 818–825. http://doi.org/10.1038/ng.3021

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–1125. http://doi.org/10.1038/ng.717

Fung, K. Y., Mangan, N. E., Cumming, H., Horvat, J. C., Mayall, J. R., Stifter, S. A., … Hertzog, P. J. (2013). Interferon- Protects the Female Reproductive Tract from Viral and Bacterial Infection. *Science*, *339*(6123), 1088–1092. http://doi.org/10.1126/science.1233321

Gaedigk, A., Blum, M., Gaedigk, R., Eichelbaum, M., ja Meyer, U. A. (1991). Deletion of the entire cytochrome P450 CYP2D6 gene as a cause of impaired drug metabolism in poor metabolizers of the debrisoquine/sparteine polymorphism. *American Journal of Human Genetics*, *48*(5), 943–950.

Ginsburg, G. (2014). Medical genomics: Gather and use genetic data in health care. *Nature*, *508*(7497), 451–453.

Gonzalez, K. D., Hill, K. A., Li, K., Li, W., Scaringe, W. A., Wang, J.-C., … Sommer, S. S. (2007). Somatic microindels: analysis in mouse soma and comparison with the human germline. *Human Mutation*, *28*(1), 69–80. http://doi.org/10.1002/humu.20416

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., … Williams, M. S. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine*, *15*(10), 761–771. http://doi.org/10.1038/gim.2013.72

Greely, H. T. (2000). Iceland's plan for genomics research: facts and implications. *Jurimetrics*, *40*, 153–191.

Gregory, A. P., Dendrou, C. A., Attfield, K. E., Haghikia, A., Xifara, D. K., Butter, F., … Fugger, L. (2012). TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature*, *488*(7412), 508–511. http://doi.org/10.1038/nature11307

Grice, E. A., ja Segre, J. A. (2012). The Human Microbiome: Our Second Genome *. *Annual Review of Genomics and Human Genetics*, *13*(1), 151–170. http://doi.org/10.1146/annurev-genom-090711-163814

Gudbjartsson, D. F., Arnar, D. O., Helgadottir, A., Gretarsdottir, S., Holm, H., Sigurdsson, A., … Stefansson, K. (2007). Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature*, *448*(7151), 353–357. http://doi.org/10.1038/nature06007

Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A., Gylfason, A., … Stefansson, K. (2015). Large-scale whole-genome sequencing of the Icelandic population. *Nature Genetics*, *47*(5), 435–444. http://doi.org/10.1038/ng.3247

Guengerich, F. P. (2008). Cytochrome p450 and chemical toxicology. *Chemical Research in Toxicology*, *21*(1), 70–83. http://doi.org/10.1021/tx700079z

Hach, F., Sarrafi, I., Hormozdiari, F., Alkan, C., Eichler, E. E., ja Sahinalp, S. C. (2014). mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing

applications. *Nucleic Acids Research*, *42*(Web Server issue), W494-500. http://doi.org/10.1093/nar/gku370

Hafler, D. A., Compston, A., Sawcer, S., Lander, E. S., Daly, M. J., De Jager, P. L., … Hauser, S. L. (2007). Risk alleles for multiple sclerosis identified by a genomewide study. *The New England Journal of Medicine*, *357*(9), 851–862. http://doi.org/10.1056/NEJMoa073493

Hall, J. L., Ryan, J. J., Bray, B. E., Brown, C., Lanfear, D., Newby, L. K., … Weintraub, W. S. (2016). Merging Electronic Health Record Data and Genomics for Cardiovascular Research: A Science Advisory From the American Heart Association. *Circulation. Cardiovascular Genetics*. http://doi.org/10.1161/HCG.0000000000000029

Handsaker, R. E., Korn, J. M., Nemesh, J., ja McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nature Genetics*, *43*(3), 269–276. http://doi.org/10.1038/ng.768

Hardy, M. P., Owczarek, C. M., Jermiin, L. S., Ejdebäck, M., ja Hertzog, P. J. (2004). Characterization of the type I interferon locus and identification of novel genes☆. *Genomics*, *84*(2), 331–345. http://doi.org/10.1016/j.ygeno.2004.03.003

Heinaru, A. (2012). *Geneetika. Õpik kõrgkoolile*. Tartu Ülikooli Kirjastus.

Heng, T. S. P., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., … Kang, J. (2008). The Immunological Genome Project: networks of gene expression in immune cells. *Nature Immunology*, *9*(10), 1091–1094. http://doi.org/10.1038/ni1008-1091

Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., … Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, *43*(4), 316–320. http://doi.org/10.1038/ng.781

Hosomichi, K., Shiina, T., Tajima, A., ja Inoue, I. (2015). The impact of next-generation sequencing technologies on HLA research. *Journal of Human Genetics*, *60*(11), 665–673. http://doi.org/10.1038/jhg.2015.102

Isken, O., ja Maquat, L. E. (2007). Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes ja Development*, *21*(15), 1833–1856. http://doi.org/10.1101/gad.1566807

Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., … Lowe, W. L. (2012). Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association*, *19*(2), 212–218. http://doi.org/10.1136/amiajnl-2011-000439

Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., … Denny, J. C. (2011). Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Science Translational Medicine*, *3*(79), 79re1-79re1. http://doi.org/10.1126/scitranslmed.3001807

Kim, S. K., Park, H. J., Kim, J. W., Chung, J.-H., Yoo, S. D., Kim, D. H., … Kim, H.-S. (2014). T Allele of nonsense polymorphism (rs2039381, Gln71Stop) of interferon-ε is a risk factor for the development of intracerebral hemorrhage. *Human Immunology*, *75*(1), 88–90. http://doi.org/10.1016/j.humimm.2013.09.004

Knight, J. C. (2013). Genomic modulators of the immune response. *Trends in Genetics*, *29*(2), 74–83. http://doi.org/10.1016/j.tig.2012.10.006

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. *Human Genetics*, *90*(1–2), 41–54.

Kullo, I. J., Ding, K., Jouni, H., Smith, C. Y., ja Chute, C. G. (2010). A Genome-Wide Association Study of Red Blood Cell Traits Using the Electronic Medical Record. *PLoS ONE*, *5*(9), e13011. http://doi.org/10.1371/journal.pone.0013011

Kullo, I. J., Ding, K., Shameer, K., McCarty, C. A., Jarvik, G. P., Denny, J. C., … Chute, C. G. (2011). Complement Receptor 1 Gene Variants Are Associated with Erythrocyte Sedimentation Rate. *The American Journal of Human Genetics*, *89*(1), 131–138. http://doi.org/10.1016/j.ajhg.2011.05.019

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. http://doi.org/10.1038/35057062

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., ja Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(Database issue), D980-985. http://doi.org/10.1093/nar/gkt1113

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., … MacArthur, D. (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, 30338. http://doi.org/10.1101/030338

Li, H., ja Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. http://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. http://doi.org/10.1093/bioinformatics/btp352

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272. http://doi.org/10.1101/gr.097261.109

Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., … for the Sequencing Initiative Suomi (SISu) Project. (2014). Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLoS Genetics*, *10*(7), e1004494. http://doi.org/10.1371/journal.pgen.1004494

MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., … Tyler-Smith, C. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, *335*(6070), 823–828. http://doi.org/10.1126/science.1215040

MacArthur, D. G., Seto, J. T., Raftery, J. M., Quinlan, K. G., Huttley, G. A., Hook, J. W., … North, K. N. (2007). Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. *Nature Genetics*, *39*(10), 1261–1265. http://doi.org/10.1038/ng2122

MacArthur, D. G., ja Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*, *19*(R2), R125–R130. http://doi.org/10.1093/hmg/ddq365

Marschall, T., Hajirasouliha, I., ja Schönhuth, A. (2013). MATE-CLEVER: Mendelian-inheritance-aware discovery and genotyping of midsize and long indels. *Bioinformatics (Oxford, England)*, *29*(24), 3143–3150. http://doi.org/10.1093/bioinformatics/btt556

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. http://doi.org/10.1101/gr.107524.110

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*(16), 2069–2070. http://doi.org/10.1093/bioinformatics/btq330

McLean, K. J., Sabri, M., Marshall, K. R., Lawson, R. J., Lewis, D. G., Clift, D., … Munro, A. W. (2005). Biodiversity of cytochrome P450 redox systems. *Biochemical Society Transactions*, *33*(Pt 4), 796–801. http://doi.org/10.1042/BST0330796

Mills, R. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, *16*(9), 1182–1190. http://doi.org/10.1101/gr.4565806

Mort, M., Ivanov, D., Cooper, D. N., ja Chuzhanova, N. A. (2008). A meta-analysis of nonsense mutations causing human genetic disease. *Human Mutation*, *29*(8), 1037–1047. http://doi.org/10.1002/humu.20763

Myerowitz, R., ja Costigan, F. C. (1988). The major defect in Ashkenazi Jews with Tay-Sachs disease is an insertion in the gene for the alpha-chain of beta-hexosaminidase. *The Journal of Biological Chemistry*, *263*(35), 18587–18589.

Ogura, Y., Bonen, D. K., Inohara, N., Nicolae, D. L., Chen, F. F., Ramos, R., … Cho, J. H. (2001). A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, *411*(6837), 603–606. http://doi.org/10.1038/35079114

Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. *American Journal of Human Genetics*, *64*(1), 18–23. http://doi.org/10.1086/302219

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., ja Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. http://doi.org/10.1038/ng.259

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., ja Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, *42*(7), 570–575. http://doi.org/10.1038/ng.610

Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A., … Mathew, C. G. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics*, *39*(7), 830–832. http://doi.org/10.1038/ng2061

Peng, F.-W., Duan, Z.-J., Zheng, L.-S., Xie, Z.-P., Gao, H.-C., Zhang, H., … Hou, Y.-D. (2007). Purification of recombinant human interferon-ε and oligonucleotide microarray analysis of interferon-ε-regulated genes. *Protein Expression and Purification*, *53*(2), 356–362. http://doi.org/10.1016/j.pep.2006.12.013

Pestka, S., Krause, C. D., ja Walter, M. R. (2004). Interferons, interferon-like cytokines, and their receptors. *Immunological Reviews*, *202*(1), 8–32. http://doi.org/10.1111/j.0105-2896.2004.00204.x

Picard, C., Al-Herz, W., Bousfiha, A., Casanova, J.-L., Chatila, T., Conley, M. E., … Gaspar, H. B. (2015). Primary Immunodeficiency Diseases: an Update on the Classification from the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency 2015. *Journal of Clinical Immunology*, *35*(8), 696–726. http://doi.org/10.1007/s10875-015-0201-1

Quintáns, B., Ordóñez-Ugalde, A., Cacheiro, P., Carracedo, A., ja Sobrido, M. J. (2014). Medical genomics: The intricate path from genetic variant identification to clinical interpretation. *Applied ja Translational Genomics*, *3*(3), 60–67. http://doi.org/10.1016/j.atg.2014.06.001

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444–454. http://doi.org/10.1038/nature05329

Ritchie, M. D., Denny, J. C., Crawford, D. C., Ramirez, A. H., Weiner, J. B., Pulley, J. M., … Roden, D. M. (2010). Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *The American Journal of Human Genetics*, *86*(4), 560–572. http://doi.org/10.1016/j.ajhg.2010.03.003

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., ja Masys, D. R. (2008). Development of a large-scale de-identified DNA biobank to

enable personalized medicine. *Clinical Pharmacology and Therapeutics*, *84*(3), 362–369. http://doi.org/10.1038/clpt.2008.89

Saleh, M., Vaillancourt, J. P., Graham, R. K., Huyck, M., Srinivasula, S. M., Alnemri, E. S., … Nicholson, D. W. (2004). Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms. *Nature*, *429*(6987), 75–79. http://doi.org/10.1038/nature02451

Samson, M., Libert, F., Doranz, B. J., Rucker, J., Liesnard, C., Farber, C.-M., … Parmentier, M. (1996). Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*, *382*(6593), 722–725. http://doi.org/10.1038/382722a0

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C. A., Patsopoulos, N. A., Moutsianas, L., … Compston, A. (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, *476*(7359), 214–219. http://doi.org/10.1038/nature10251

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., ja Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., … Abecasis, G. R. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics*, *47*(11), 1272–1281. http://doi.org/10.1038/ng.3368

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., ja Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. http://doi.org/10.1038/nrg3642

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., ja Cooper, D. N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9. http://doi.org/10.1007/s00439-013-1358-4

Storry, J. R., Jöud, M., Christophersen, M. K., Thuresson, B., Åkerström, B., Sojka, B. N., … Olsson, M. L. (2013). Homozygosity for a null allele of SMIM1 defines the Vel-negative blood group phenotype. *Nature Genetics*, *45*(5), 537–541. http://doi.org/10.1038/ng.2600

Storry, J. R., ja Mallory, D. (1994). Misidentification of anti-Vel due to inappropriate use of prewarming and adsorption techniques. *Immunohematology / American Red Cross*, *10*(3), 83–86.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. http://doi.org/10.1038/nature15394

Sulem, P., Helgason, H., Oddson, A., Stefansson, H., Gudjonsson, S. A., Zink, F., … Stefansson, K. (2015). Identification of a large set of rare complete human knockouts. *Nature Genetics*, *47*(5), 448–452. http://doi.org/10.1038/ng.3243

Sussman, L. N., ja Miller, E. B. (1952). New blood factor: Vel. *Revue D'hématologie*, *7*(3), 368–371.

The International HapMap Project. (2003). *Nature*, *426*(6968), 789–796.

Timpson, N. J., Walter, K., Min, J. L., Tachmazidou, I., Malerba, G., Shin, S.-Y., … Zheng, H.-F. (2014). A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*, *5*, 4871. http://doi.org/10.1038/ncomms5871

Uibo, R., Kisand, K., Peretson, P., ja Reimand, K. (2015). *Immunoloogia. Õpik kõrgkoolidele*. Tartu Ülikooli Kirjastus.

Velbri, S. (1982). *Immunoloogia*. Tallinn: Valgus.

Velbri, S. (2002). *Immuunpuudulikkus. Diagnostika ja ravi*. Medicina.

Verma, A., Leader, J. B., Verma, S. S., Frase, A., Wallace, J., Dudek, S., … Pendergrass, S. A. (2016). INTEGRATING CLINICAL LABORATORY MEASURES AND ICD-9 CODE DIAGNOSES IN PHENOME-WIDE ASSOCIATION STUDIES. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, *21*, 168–179.

Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., … Zhang, W. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, *526*(7571), 82–90. http://doi.org/10.1038/nature14962

Weiss, J., Pyrski, M., Jacobi, E., Bufe, B., Willnecker, V., Schick, B., … Zufall, F. (2011). Loss-of-function mutations in sodium channel Nav1.7 cause anosmia. *Nature*, *472*(7342), 186–190. http://doi.org/10.1038/nature09975

White, M. B., Amos, J., Hsu, J. M., Gerrard, B., Finn, P., ja Dean, M. (1990). A frame-shift mutation in the cystic fibrosis gene. *Nature*, *344*(6267), 665–667. http://doi.org/10.1038/344665a0

Xue, Y., Daly, A., Yngvadottir, B., Liu, M., Coop, G., Kim, Y., … Tyler-Smith, C. (2006). Spread of an inactive form of caspase-12 in humans is due to recent positive selection. *American Journal of Human Genetics*, *78*(4), 659–670. http://doi.org/10.1086/503116

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., ja Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)*, *25*(21), 2865–2871. http://doi.org/10.1093/bioinformatics/btp394

Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S. W., … Gerstein, M. B. (2007). Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. *Genome Research*, *17*(6), 839–851. http://doi.org/10.1101/gr.5586307

KASUTATUD VEEBIAADRESSID

http://www.1000genomes.org/about (2.05.2016)
http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi (2.05.2016)
https://immport.niaid.nih.gov/ (2.05.2016)
http://macarthurlab.org/lof/ (10.05.2016)
http://www.ensembl.org/Homo_sapiens/Tools/VEP?db=core (12.05.2016)
https://www.immgen.org/ (16.05.2016)
https://github.com/konradjk/loftee (23.05.2016)
http://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0025680/ (23.05.2016)
http://www.uniprot.org/ (25.05.2016)
http://www.hgmd.cf.ac.uk/ac/index.php (25.05.2016)
http://exac.broadinstitute.org (26.05.2016)
http://rhk.sm.ee/ (26.05.2016)
https://github.com/Vityay/1-2-3-SV (26.05.2016)
https://github.com/Vityay/DWAC-Seq (26.05.2016)

LISAD

Lisa 1

Tabel 13. Funktsioonikaoga mutatsioonide arv kromosoomide kaupa

Kromosoom	LoF arv
1	1416
2	1085
3	874
4	654
5	657
6	776
7	757
8	541
9	557
10	542
11	925
12	813
13	300
14	496
15	478
16	628
17	792
18	204
19	916
20	355
21	165
22	337
X	170
	14438

Lisa 2

Tabel 14. *IFNE* geeni rs2039382 variandi alleelisagedused populatsioonide kaupa (ExAC andmebaas, kohandatud)

Populatsioon	**AC**	**AN**	**Homosügoote**	**AF**
Ida-aasia	1536	8574	139	0,1791
Lõuna-aasia	2323	16472	168	0,1410
Aafrika	659	10314	25	0,0639
Latiinod	515	11504	15	0,0448
Muu	23	906	1	0,0254
Euroopa (ka soomlased)	122	6610	3	0,0185
Euroopa (va soomlased)	254	66444	1	0,0038
Kokku	**5432**	**120824**	**352**	**0,0450**
Eestlased	52	4488	1	0,0012

LIHTLITSENTS

Mina, Kelli Grand
(sünnikuupäev:18.10.1990)

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

„Funktsioonikaoga mutatsioonide analüüs 2300 inimese genoomi ja terviseandmete põhjal“,

mille juhendajad on Lili Milani ja Pärt Peterson,

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, 27.05.2016