VILNIAUS UNIVERSITETAS

TAUTVYDAS RANČELIS

PATOGENINIŲ GENOMO VARIANTŲ IR JŲ GENŲ, LEMIANČIŲ AUTOSOMINES RECESYVIĄSIAS LIGAS, ĮVAIROVĖS ANALIZĖ, PANAUDOJANT VISO EGZOMO SEKOSKAITĄ

Daktaro disertacija

Biomedicinos mokslai, medicina (06 B)

Vilnius, 2016 metai

Disertacija rengta Vilniaus universiteto Medicinos fakulteto Žmogaus ir medicininės genetikos katedroje 2012– 2016 metais

Mokslinis vadovas – **prof. habil. dr. Vaidutis Kučinskas** (Vilniaus universitetas, biomedicinos mokslai, biologija – 01 B)

Mokslinis konsultantas – **prof. dr. Loreta Cimbalistienė** (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B)

TURINYS

Santrumpų sąrašas

Terminų santrumpos:

1000G – 1000 genomų projektas (angl. *1000 Genomes Project*)

A – Aukštaitija

AD – paveldimas autosominiu dominantiniu būdu

AR – paveldimas autosominiu recesyviuoju būdu

bp – bazių pora (angl. *base pair*)

ClinVar – viešas duomenų archyvas, kuriame kaupiamos žmogaus genomo pakaitos, ir jų įtaka fenotipui

dAMP – deoksiadenozinomonofosfatas (angl. *deoxyadenosine monophosphate*)

dATP – deoksiadenozinotrifosfatas (angl. *deoxyadenosine triphosphate*)

db – duomenų bazė (angl. *database*)

ddH_2O – didejonizuotas vanduo (angl. *dideionized water*)

ddNTP – dideoksiribonukleotidas (angl. *deoxyribonucleotide*)

DNR – deoksiribonukleorūgštis (angl. *deoxyribonucleic acid*)

dNTP – deoksiribonukleotidas (angl. *deoxyribonucleotide*)

ExAC – egzomų kaupimo ir saugojimo konsorciumas (angl. *The Exome Aggregation Consortium*)

KSP – kopijų skaičiaus persitvarkymai (angl. *copy number variation*)

LITGEN – Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis – projektas

MAF – retojo alelio dažnis (angl. *Minor Allele Frequency*)

NKS – naujos kartos sekoskaita (angl. *Next-generation Sequencing*)

PA – Pietų Aukštaitija

PGR – polimerazės grandininė reakcija (angl. *Polymerase Chain Reaction*)

PŽ – Pietų Žemaitija

RA – Rytų Aukštaitija

RNR – ribonukleorūgštis (angl. *ribonucleic acid*)

SOLiD – sekoskaita, atliekant oligonukleotidų priligavimą ir nustatymą (angl. *Sequencing by Oligonucleotide Ligation and Detection*)

SV – struktūros variantas (angl. *structural variant*)

ŠŽ – Šiaurės Žemaitija

VA – Vakarų Aukštaitija

VNP – vieno nukleotido polimorfizmas (angl. *single nucleotide polymorphism*)

VNP_A – visuotinė naujagimių patikra (angl. *newborn screening*)

VNV – vieno nukleotido variantas (angl. *single nucleotide variant*)

VU MF ŽMGK – Vilniaus universiteto Medicinos fakulteto Žmogaus ir medicininės genetikos katedra

VŽ – Vakarų Žemaitija

Ž – Žemaitija

Programų ir duomenų formatų santraukos:

BAM duomenų formatas – binarinis naujos kartos sekoskaitos duomenų dokumento formatas, kuris gaunamas prilygiavus pirminius sekoskaitos duomenis prie referentinio genomo (angl. *Binary Alignment/Map format)*

BFAST programa – greitai trumpus fragmentus prie referentinio genomo prilygiuojantis įrankis (angl. *Blat-like Fast Accurate Search Tool*)

BWA programa – DNR fragmentus prie referentinio genomo prilygiuojanti programa, naudojanti *Burrows-Wheeler* algoritmą (angl. *Burrows-Wheeler Aligner*)

CASAVA programa – *Illuminos* platformai sukurta programa, galinti atskirti sujungtus mėginius, prilygiuoti tiriamus duomenis prie referentinio genomo ir atlikti teisingų variantų atranką (angl. *Consensus Assessment of Sequence And Variation*)

FASTQ duomenų formatas – *Illumina* pirminiai sekoskaitos duomenys su kokybės vertėmis (angl. *Fasta Format with Quality Score*)

GATK programa – genomo analizės paketas (*angl. Genome Analysis Toolkit*)

IGV programa – integruota genomo vaizduoklė (angl. *The Integrative Genomics Viewer*)

MAQ programa – prilygiavimo prie referentinio genomo ir genomo surinkimo įrankis su kokybės vertėmis (angl. *Mapping and Assembly with Qualities*)

SAM duomenų formatas – naujos kartos sekoskaitos duomenų dokumento formatas, kuris gaunamas prilygiavus pirminius sekoskaitos duomenis prie referentinio genomo (angl. *Sequence Alignment/Map Format)*

SHRiMP programa – trumpų DNR fragmentų prilygiavimo prie referentinio genomo paketas (angl. *SHort Read Mapping Package*)

VCF duomenų formatas – naujos kartos sekoskaitos duomenų dokumento formatas, kai yra teisingai atrinkti genomo variantai (angl. *Variant Call Format*)

XSQ duomenų formatas – *SOLiD* pirminiai sekoskaitos duomenys su kokybės vertėmis (angl. *The Extensible Sequence with Quality Score)*

Genų santrumpos:

ABCA4 – ATP sujungiantis blokas, ketvirtas narys, priklausantis pošeimui A (angl. *ATP Binding Cassette Subfamily A Member 4*)

ABCG8 – ATP sujungiantis blokas, aštuntas narys, priklausantis pošeimui G (angl. *ATP Binding Cassette Subfamily G Member 8)*

ACAD9 – acil-CoAdehidrogenazės šeimos genas, 9 narys (angl. *Acyl-CoA Dehydrogenase Family Member 9*)

AMPD 1 – adenozino monofosfato deaminazės 1 genas (angl. *Adenosine Monophosphate Deaminase 1*)

COL18A1 – kolageno 8 tipo alfa 1 genas (angl. *Collagen Type XVIII Alpha 1*)

CPT1A – karnitino palmitoiltransferazės 1A genas (angl. *Carnitine Palmitoyltransferase 1A*)

DCLRE1C – DNR kryžminės sąveikos reparacijos 1C genas (angl. *DNA Cross-Link Repair 1C*).

DDX11 – DEAD/H bloko (asparto rūgštis – glutamo rūgštis – alaninas – asparto rūgštis) 11 helikazės genas (angl. *DEAD/H – Asp-Glu-Ala-Asp/His Box Helicase 11*)

EPB4.2 – eritrocitų membranos baltymų juostos 4.2 genas (angl. *Erythrocyte Membrane Protein Band 4.2*)

FECH – ferochelatazės genas (angl. *Ferrochelatase*)

FMO3 – flaviną turinčios monooksigenazės 3 genas (angl. *Flavin Containing Monooxygenase 3*)

GLI3 – GLI šeimos „cinko pirštelių“ 3 genas (angl. *GLI Family Zinc Finger 3*)

LIPC – lipazės C genas (kepenyse) (angl. *Lipase C, Hepatic Type*)

MYO7A – miozino 7A genas (angl. *Myosin VIIA*)

OR2J3 – uoslės receptorių genas, 3 narys, priklausantis antros šeimos J pošeimiui (angl. *Olfactory Receptor Family 2 Subfamily J Member 3*)

PDE11A – fosfodiesterazės 11A genas (angl. *Phosphodiesterase 11A*)

PHYKPL–5-fosfohidroksi-L-lizino fosfoliazės genas (angl. *5-Phosphohydroxy-L-Lysine Phospho-Lyase*)

POLG – mitochondrijų DNR polimerazės katalizuojančio subvieneto genas (angl. *Mitochondrial DNA Polymerase Catalytic Subunit*)
RAPSN – baltymo, susijusio su sinapsės receptoriumi, genas (angl. *Receptor Associated Protein of the Synapse*)
RET – pirminis onkogenas (angl. *Proto-Oncogene*)

Pagrindinių vartotų terminų paaiškinimai:

Anotavimas – metaduomenų prijungimas prie tiriamų genomo variantų, gautų po teisingų genomo variantų atrankos etapo NKS analizės algoritme.

Indeksavimas – NKS analizės algoritmo etapas, kuris palengvina analizę – leidžia programoms neskaityti visos DNR sekos iki dominančios pozicijos, o atsidurti toje pozicijoje iš karto.

Nesinoniminė pakaita – egzone įvykusi nukleotido pakaita, kai pakeičiama koduojama aminorūgštis (angl. *missens*) arba virstama į baigties kodoną (angl. *nonsens*).

NKS analizės algoritmas – seka veiksmų, kurie gali apimti kelių skirtingų programų darbą tam, kad iš pirminių sekoskaitos duomenų būtų gaunami teisingi genomo variantai, kurie skiriasi nuo referentinio genomo.

Padengimas – vykdant NKS yra sekvenuojama ne tik viena genomo seka, bet ir kelios jo kopijos. Analizės algoritmo prilygiavimo prie referentinio genomo etape DNR fragmentai susikloja, dėl to kiekvienas atskiras nukleotidas nuskaitomas kelis kartus.

Patogeninis genomo variantas – tai genomo variantas, kuris turi neigiamą funkcinę reikšmę organizmui tiesiogiai lemiant ligos pasireiškimo riziką.

Prilygiavimas prie referentinio genomo – NKS analizės algoritmo etapas, kai sekoskaitos metu gauti DNR fragmentai yra išdėliojami pagal referentinio genomo seką, taip nustatant DNR fragmentų poziciją tiriamo organizmo genome.

Referentinis genomas – visuotinai taikomas apibendrintas statistinis genomas, kuris naudojamas palyginti su mokslininkų tiriamaisiais duomenimis.

Sekoskaita – nukleotidų pirminės sekos nustatymas DNR molekulėje (heterociklinių bazių – adenino, timino, guanino, citozino – nustatymas DNR molekulėje).

Sinoniminė pakaita – tai egzone įvykusi nukleotido pakaita, kai nepakeičiama koduojama aminorūgštis (angl. *silent*).

Struktūros variantai – kai skirtumas tarp DNR sekų būna didesnis nei 50 bp. Tokie skirtumai varijuoja nuo 50 bp iki kelių milijonų bp.

Teisingi genomo variantai – genomo varantai, kurie gaunami genomo analizės algoritmo metu. Tai genomo variantai, kurie ne tik skiriasi nuo referentinio genomo, bet ir juos analizės algoritmas yra įvertinęs kaip patikimus ir atskyręs nuo klaidingai gautų genomo variantų.

Trumpos iškritos ir intarpai – tai nuo 1 iki 50 bp dydžio DNR sekos iškritos arba intarpai. Disertacijoje analizuojamos iškritos yra iki 14 nukleotidų, o intarpai iki 4 nukleotidų.

Vieno nukleotido polimorfizmas – genomo variantas, kurio retojo alelio dažnis populiacijoje yra didesnis nei 1 % (MAF > 1%).

Vieno nukleotido variantas – taškinė mutacija, kai vienas nukleotidas yra pakeičiamas kitu nukleotidu.

VNP genotipavimas – DNR sekų nustatymas panaudojant genotipavimo lustus. Į lustų sudėtį įtraukiami žinomi vieno nukleotido polimorfizmai. Dažniausiai tai genomo variantai, kurių retojo alelio dažnis yra 5 % ir didesnis.

ĮVADAS

Darbo tematika

Disertacija „Patogeninių genomo variantų ir jų genų, lemiančių autosomines recesyviąsias ligas, įvairovės analizė, panaudojant viso egzomo sekoskaitą“ susijusi su keletu šiuo metu žmogaus genetikai aktualių sričių:

1) naujos kartos sekoskaita (NKS) ir jos naudojimu žmogaus genetikos tyrimuose ir medicinos praktikoje,

2) populiacinio pobūdžio tyrimais, kuriais nustatomas paveldimas ligas lemiančių genomo variantų dažnis sveikuose asmenyse,

3) sveikų asmenų genetinių duomenų naudojimu genetinių ligų diagnostikai pagerinti,

4) bioinformacinių metodų taikymu šiuolaikinėje žmogaus genetikoje.

Tiriant žmogaus genetines ligas pasitelkus NKS, gali būti tiriami sergantys asmenys – nustatomi genomo variantai, lemiantys paveldimas ligas. Tačiau dėl genetinių ligų gali būti tiriami ne tik sergantys, bet ir sveiki asmenys, nes kiekvienas sveiku laikomas asmuo gali būti iki 400 patogeninių genomo variantų nešiotojas (Xue et al., 2012; MacArthur et al., 2012; Lazarin et al., 2013). Patogeninių genomo variantų tyrimą sveikuose asmenyse palengvina kitų tyrėjų paskelbti darbai, kuriuose yra nurodyti konkretūs patogeniniai genomo variantai.

Šioje disertacijoje, panaudojant NKS ir specializuotą duomenų bazę, kurioje saugoma iki šiol mokslininkų sukaupta informacija apie patogeninius genomo variantus, ieškota, ar sveiki lietuvių populiacijos asmenys turi patogeninių genomo variantų, atliktas jų dažnio palyginamas su kitų populiacijų duomenimis.

Nors didelės apimties sekoskaitos nauda neabejotina, sekoskaita reikalauja tinkamai atlikti analizės algoritmą, todėl disertacijoje bioinformacinėms problemoms spręsti yra skiriama itin daug dėmesio.

Darbo naujumas, aktualumas ir reikšmė

Tyrimas, ar sveiki asmenys turi patogeninių genomo variantų, panaudojant naujos kartos sekoskaitą, yra pirmas tokio pobūdžio Lietuvoje.

Disertacinis darbas yra aktualus Lietuvos mastu, nes lietuvių egzomų duomenys pirmą kartą naudojami tokio pobūdžio tyrimuose ir suteikiama naujų žinių apie patogeninių variantų paplitimą tarp sveikų lietuvių.

Nors šioje srityje pažanga pasaulyje vyksta itin sparčiai, atlikto darbo pobūdis yra aktualus ir pasauliniu mastu, nes, remiantis tyrimo duomenimis, įvertinama *ClinVar* duomenų bazė.

Darbe analizuojami asmenų iš bendros lietuvių populiacijos genomo duomenys atskleidžia, kurie patogeniniai genomo variantai statistiškai patikimai dažniau pasitaiko lietuvių populiacijoje, palyginti su kitų populiacijų duomenimis iš 1000 genomų ir ExAC duomenų bazių. Tai leidžia tokius tyrimo duomenis panaudoti diagnostikai – sudaryti specialią lietuvių populiacijai skirtą patogeninių variantų lentelę ir gali paskatinti lietuvių populiacijoje dažnesnę paveldimą ligą įtraukti į visuotinę naujagimių patikrą.

Tyrimo tikslas – nustatyti ir įvertinti lietuvių populiacijoje esančius patogeninius genomo variantus panaudojant viso egzomo sekoskaitą.

Tyrimo tikslui įgyvendinti iškelti uždaviniai:

1. Parinkti ir atlikti NKS analizės algoritmą, kuris būtų tinkamiausias analizuojant turimus pirminius *SOLiD* sekoskaitos duomenis.
2. Nustatyti lietuvių populiacijoje esančius genomo variantus, kurie yra žinomi kaip patogeniniai.

3. Kaip patogeninių variantų šaltinį naudoti *ClinVar* duomenų bazę.
4. Nustatyti patogeninių genomo variantų dažnių skirtumus, lyginant lietuvių populiaciją su kitomis populiacijomis.
5. Atlikti vidupopuliacinį patogeninių genomo variantų palyginimą pagal lietuvių etnolingvistines grupes.

Ginamieji teiginiai

1. Lietuvių populiacijoje paplitusių patogeninių genomo variantų dažnis skiriasi nuo patogeninių genomo variantų dažnio kitose populiacijose.
2. Patogeninių genomo variantų dažnių skirtumas yra ir vidupopuliaciniu lygmeniu – yra tik aukštaičiams ir žemaičiams būdingų patogeninių genomo variantų.
3. Tirtoje lietuvių populiacijoje didžioji dalis identifikuotų patogeninių genomo variantų yra susiję su metabolizmo sutrikimais.
4. *ClinVar* duomenų bazėje yra daug patogeniniais įvardytų genomo variantų, kurie nėra patogeniški.

Darbo aprobacija

Darbo rezultatai paskelbti dviejuose periodiniuose Lietuvos mokslo žurnaluose, dviejuose užsienio leidiniuose – vienas jų yra ISI sąrašuose ir vienas turi ISI citavimo indeksą; pristatyti dviejose tarptautinėse mokslinėse konferencijose ir dvejose Lietuvoje vykusiose konferencijose.

Disertacijos tema paskelbtų straipsnių sąrašas

1. **Tautvydas Rančelis**, Loreta Cimbalistienė, Vaidutis Kučinskas. Next-generation whole-exome sequencing contribution to identification of rare autosomal recessive diseases. Acta Medica Lituanica, 2013, Vol. 20, N. 1. p. 43–51.
2. **Tautvydas Rančelis**, Erinija Pranckevičienė, Vaidutis Kučinskas. Anotaciniai įrankiai ir kompiuterinės programos genomo / egzomo

duomenų analizei. Laboratorinė medicina, 2013, T. 15, N. 4 (60), p. 206–212.

3. Erinija Pranckevičienė, **Tautvydas Rančelis**, Aidas Pranculis, Vaidutis Kučinskas. Challenges in exome analysis by LifeScope and its alternative computational pipelines. BMC Research Notes, 2015, 8:1.
4. Violeta Mikštienė, Audronė Jakaitienė, Jekaterina Byčkova, Eglė Gradauskienė, Eglė Preikšaitienė, Birutė Burnytė, Birutė Tumienė, Aušra Matulevičienė, Laima Ambrozaitytė, Ingrida Uktverytė, Ingrida Domarkienė, **Tautvydas Rančelis**, Loreta Cimbalistienė, Eugenijus Lesinskas, Vaidutis Kučinskas, Algirdas Utkus. The high frequency of *GJB2* gene mutation c.313_326del14 suggests its possible origin in ancestors of Lithuanian population. BMC Genetics, 2016, 19;17 (1):45. Epub 2016, doi: 10.1186/s12863-016-0354-9.

Stendiniai pranešimai tarptautinėse konferencijose tezės:

1. **T. Rančelis**, E. Pranckevičienė, A. Pranculis, V. Kučinskas, Comparison of *SOLiD* sequencing data analysis pipelines. In European Journal of Human Genetics. London, Nature Publishing Group. ISSN 1018-4813. Vol. 23, Supplement 1. 2015. p. 314–314.

2. V. Kučinskas, **T. Rančelis**, I. Domarkienė, E. Pranckevičienė, I. Uktverytė, L. Ambrozaitytė. Profile of pathogenic alleles in healthy Lithuanian population / 65th Annual Meeting of The American Society of Human Genetics, October 6–10, 2015, Baltimore MD: poster abstracts. Baltimore, The American Society of Human Genetics, 2015, p. 483.

1. LITERATŪROS APŽVALGA

1.1. Sekoskaita ir jos naudojimas medicinoje

DNR sekoskaita jau tapo vienu iš pagrindinių ir įprastų tyrimų atliekant diagnostiką. Naudojant *Sanger* sekoskaitą efektyviai nustatomos paveldimos ligos ir išaiškinami jas lemiantys genomo variantai.

Itin sparčiai tobulėjančios sekoskaitos technologijos lėmė naujos kartos sekoskaitos (NKS) atsiradimą. Ją naudodamas tyrėjas gali gauti informaciją ne tik apie pavienį geną, bet ir apie visą koduojančią genomo dalį – egzomą ar net apie visą genomą.

NKS per trumpą laikotarpį nuo jos naudojimo žmogaus tyrimuose pradžios tapo vienu pagrindinių iki tol nežinomų paveldimų ligų priežasčių identifikavimo metodų. Spartus šių technologijų taikymo atpigimas leido NKS naudoti ne tik moksliniuose tyrimuose, bet ir kasdienėje medicinos praktikoje.

1.1.1. Naujos kartos sekoskaitos (NKS) metodų raida

DNR sekoskaitos pradžia buvo 1971 metais, kai *Ray Wu* ir *Ellen Taylor* sukūrė „plius-minus“ metodą ir juo nustatė 12 bakteriofago lambda nukleotidų. Tai buvo fermentinė reakcija naudojant 1, 2 ar 3 dNTP. Nors „plius-minus“ metodas nebuvo tikslus ir reikalavo labai daug laiko, jis įnešė savo dalį į sekvenavimo istoriją ir tolesnių sekvenavimo metodų kūrimą, nes, taikant šį metodą, pirmą kartą nukleotidų sekai nustatyti buvo panaudotas specifinis pradmuo ir jį pratęsianti DNR polimerazė, poliakrilamido gelis bei grandinės terminacija ties specifine baze (Wu and Taylor, 1971).

1977 metais nepriklausomai vienas nuo kito sukurti du sekoskaitos metodai – *Allan Maxam* ir *Walter Gilbert* – aprašė cheminio sekvenavimo metodą, kai naudojami cheminiai reagentai, specifiškai nuskeliantys tik tam tikrą bazę, o *Frederick Sanger* aprašė fermentinį sekoskaitos metodą, kai naudojama DNR grandinės sintezės reakcija ir dideoksiribonukleotidai

(ddNTP), užbaigiantys sintezę ties tam tikra baze (Sanger et al., 1973; Maxam and Gilbert, 1977).

1986-ais metais Kalifornijos technologijos institute buvo sukurta pusiau automatinė fermentinė sekoskaita, o po metų atsirado visiškai automatizuota fermentinė sekoskaita (Smith et al., 1986; Prober et al., 1987).

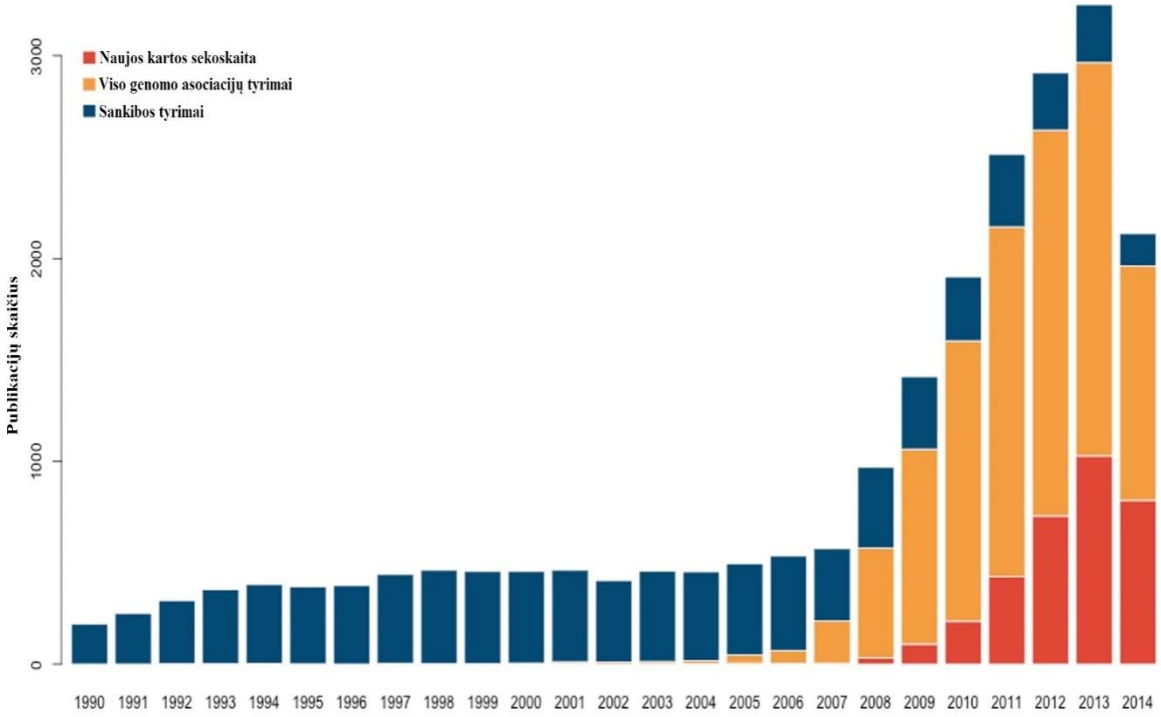
1.1.2. Naujos kartos sekoskaita

Paskelbus Žmogaus genomo projekto pabaigą, 2003 metais prasidėjo nauja sekoskaitos era, leidusi atlikti plataus masto sekoskaitą ir suteikusi daug didesnes galimybes nei kapiliarinė sekoskaita. 2003 metais pasirodė pirmosios komercinės NKS sistemos (Voelkerding et al., 2009).

Skirtingos bendrovės pateikė skirtingus sekoskaitos metodo principus:

1. 454 sistema (Roche Applied Science, Basel, Switzerland) – **pirosekoskaita**, sekoskaita pagal pirofosfatų susidarymą (Kling, 2003; Siqueira et al., 2012).
2. *Illumina/Solexa* genomo analizatorius (Illumina, San Diego, CA, USA) – **sekoskaita sintezės būdu** (Bennett, 2004; Bentley et al., 2008).
3. *SOLiD* sistema (Applied Biosystems, Foster City, CA, USA) – **sekoskaita ligavimo būdu** (McKernan et al., 2009).
4. *HeliScope* (Helicos BioSciences, Cambridge, MA, USA) – **vienos molekulės sekoskaita**, sekoskaita, atliekama be amplifikacijos etapo (Thompson and Steinmann, 2010).
5. *Ion Torrent* (Thermo Fisher Scientific, Inc., USA) – **pH nustatanti sekoskaita** (Malapelle et al., 2015).
6. MinION™ (Oxford Nanopore Technologies, Oxford Science Park, Oxford, UK) – **sekoskaita pasitelkus nanoporas** (Tarraga et al., 2016).

Atsiradus naujos kartos sekoskaitai ir jai sparčiai pingant, ji tapo viena svarbiausių paveldimų ligų tyrimo dalių (1 pav.).



1 pav. Publikacijų, kuriose pateikiami genetiniai žmonių tyrimo rezultatai (1990–2014), skaičius. Į sąrašą įtrauktos tik tos publikacijos, kurios yra MEDLINE duomenų bazėje. Diagramoje pavaizduotas padidėjęs NKS naudojimas žmonių genetiniuose tyrimuose per pastaruosius kelerius metus (Wang et al., 2015).

1.1.2.1. Sekoskaitos ligavimo būdu metodo principai

Pateiktame darbe naudota *SOLiD* sistema egzomų sekoskaitai. Sekoskaita ligavimo būdu (angl. *Sequencing by Oligonucleotide Ligation and Detection*) yra naujos kartos sekoskaitos technologija, sukurta „Applied Biosystems“ bendrovės 2007 metais.

Esminiai *SOLiD* sekoskaitos bruožai – DNR fragmentams pagausinti naudojama emulsinė PGR ir spalvinis binukleotidinių zondų kodas.

SOLiD sekoskaita, kaip ir su daugeliu kitų sekoskaitos sistemų, galima atlikti trijų tipų DNR bibliotekų sekoskaitą: vienos DNR sekos (angl. *fragment library*), abiejų DNR sekų (angl. *pair-end library*) ir suporuotų dviejų fragmentų

(angl. *mate-paired library*). Šiame darbe *SOLiD* sistema sekvenuoti trumpi DNR fragmentai – iki 75 bp tiesioginėje grandinėje ir iki 35 bp atvirkštinėje grandinėje.

SOLiD sistemos sekoskaitos strategija paremta ne DNR sinteze, bet DNR ligavimu ir naudojamu spalviniu kodavimu. Keturių spalvų fluorescenciniai dažai yra naudojami užkoduoti šešiolika dviejų nukleotidų derinių. Toks dinukleotidinis kodas labai sumažina sekoskaitos klaidų skaičių, nes kiekvienas nukleotidas yra įvertinamas du kartus (Mardis, 2008; McKernan et al., 2009; Ku and Roukos, 2013).

1.1.3. Egzomų sekoskaita

Egzomas – tai visa koduojančioji genomo dalis (visi žinomų genų egzonai). Jis sudaro apie 1 % genomo. Nors egzomas yra nedidelė genomo dalis, jo funkcinė svarba organizmui yra itin didelė, nes lemia apie 85 % žinomų genetinių mendelinių ligų (Majewski et al., 2011; Gilissen et al., 2012).

Egzomo tyrimas, palyginti su viso genomo sekoskaita, turi daug pranašumų: 1) egzomas yra esminė genomo dalis funkciniu požiūriu, o tyrimas sukoncentruotas tik į nedidelę genomo dalį, ir tai leidžia gerokai sumažinti tyrimo kainą; 2) tyrimo kokybei labai svarbus yra DNR fragmentų padengimas. Kuo padengimas yra didesnis, tuo daugiau toje pačioje genomo pozicijoje yra nuskaitoma nukleotidų, tarp jų ir tų, kurie skiriasi nuo referentinio genomo. Konkrečioje pozicijoje atsikartojančios pakaitos NKS analizėje yra laikomos rodikliu, leidžiančiu atskirti realias pakaitas nuo sekoskaitos klaidų. Puikų padengimą daug pigiau ir paprasčiau galima gauti vykdant egzomų sekoskaitą; 3) lengvesnė viso egzomo sekoskaitos duomenų analizė. Kiekvieno asmens egzomo sekoskaitos rezultatas yra 30–50 tūkstančių genomo variantų, o genomo sekoskaita gaunamas genomo variantų skaičius siekia iki trijų milijonų. Ieškant vienos konkrečios genetinės ligos priežasties, genomo sekoskaita identifikuotas didelis genomo variantų skaičius pailgina ir pasunkina analizę. Negeninės

genomo sritys taip pat turi gerokai mažesnį spektrą anotacinių įrankių, kurie skirti tolesniam jų įvertinimui; 4) egzomų sekoskaita reikalauja gerokai mažiau kompiuterinių pajėgumų ir laiko duomenims išanalizuoti (Teer and Mullikin, 2010; Gilissen et al., 2012).

Egzomų sekoskaita paveldimoms ligoms nustatyti pradėta naudoti 2009 metais (tais metais pasirodė elektroninė publikacijos versija, o popierinė – 2010 metais), kai *Sarah Ng,* naudodamasi egzomu, nustatė *Miller* sindromo genetinę priežastį (Ng et al., 2010). Nuo to laiko egzomų sekoskaita vaidina itin svarbų vaidmenį nustatant naujus genomo variantus, susijusius su ligomis. Štargarto geltonosios dėmės distrofija, pigmentinis retinitas, *Usher* sindromas, *Perrault* sindromas, nebaigtinė osteogenezė, *Joubert* sindromas – tai tik kelios ligos, kurių genominės priežastys identifikuotos šiuo sekoskaitos metodu (Pierce et al., 2010; Becker et al., 2011; Zuchner et al., 2011; Cabral et al., 2012; Eisenberger et al., 2012; Strom et al., 2012; Rančelis et al., 2013; Srour et al., 2015).

1.2. Žmogaus genomo įvairovės samprata

Genomo įvairovė – tai skirtumai DNR sekoje, kurie matomi palyginus dvi ir daugiau giminingų DNR sekų. Pati variacija yra skirstoma pagal jos dydį, dažnį bei pakaitos pobūdį (Alkan et al., 2011; Haraksingh and Snyder, 2013; Tattini et al., 2015).

Vieno nukleotido variantas (VNV) (angl. *single nucleotide variant – SNV*) yra vieno nukleotido pakaita.

Vieno nukleotido polimorfizmas (VNP) (angl. *single nucleotide polymorphism – SNP*) yra genomo variantas, kurio retasis alelis populiacijoje pasireiškia didesniu dažniu nei 1 % (MAF >1%).

Trumpos iškritos/intarpai (angl. *small insertions or deletions – INDEL*) – nuo 1 iki 50 bp dydžio DNR sekos iškritos arba intarpai.

Struktūros variantai (angl. *structural variant – SV*) yra, kai skirtumas tarp DNR sekų būna didesnis nei 50 bp. Pakaita būna nuo 50 bp iki kelių milijonų

bp. Tai gali būti iškritos, intarpai, inversijos, duplikacijos bei translokacijos. Didelė dalis struktūros variantų vadinami **kopijų skaičiaus persitvarkymais (KSP)** (angl.*copy-number variation – CNV*). KSP apima iškritas, intarpus bei duplikacijas.

Struktūros variantai (SV) gali užimti net apie 1,2 % genomo, o VNV tesudaro 0,1 % (Pang et al., 2010). Tačiau, nors VNV sudaro gerokai mažesnę genomo dalį, jų yra gerokai daugiau. Yra nustatyta, kad bet kurių dviejų asmenų DNR sekos skiriasi viena nuo kitos viena nukleotido baze kas 1 000 nukleotidų; tad kai žmogaus genomas turi apie tris milijardus bazių porų, du individai tarpusavyje skiriasi maždaug trimis milijonais VNV (Genomes Project et al., 2010).

Baigus vykdyti Žmogaus genomo projektą, prasidėjus naujai genominei erai ir nusekvenavus didesnę imtį žmonių genomų tarptautiniuose projektuose, iš kurių *HapMap* pojektas tai atliko anksčiausiai, atsirado dar vienas, tikslesnis genomo variantų skirstymas pagal jų dažnį populiacijoje (International HapMap, 2003; Manolio et al., 2009; International HapMap et al., 2010):

Dažni genomo variantai (angl. *common genome variants*) – tai variantai, kurių dažnis populiacijoje yra 5 % ir didesnis.

Mažo dažnio genomo variantai (angl. *genome variants low frequency*) – tai variantai, kurių MAF populiacijoje yra tarp 1 % ir 5 %.

Reti genomo variantai (angl. *rare variants*) – tai variantai, kurių MAF populiacijoje yra mažesnis nei 1 %.

1.2.1. Patogeniškumo samprata

Prieš žmogaus genomo sekos paskelbimą 2003 metais buvo paplitę du terminai – polimorfizmas ir mutacija, kurių nedetalus apibrėžimas publikuotas 1993 metais ir patikslintas 1999 metais. Polimorfizmu laikytas genomo variantas, kurio MAF populiacijoje yra 1 % ir didesnis, o mutacija laikyta reta DNR sekos pakaita, kuri dažniausiai, bet ne visada, lemia ligą (Beaudet and Tsui, 1993; Brookes, 1999).

Prasidėjus naujai genomikos erai ir pradėjus sekvenuoti egzomus ir genomus, šie plačiai vartojami terminai dažnai lėmė neteisingą interpretavimą, ar genomo variantai yra patogeniški. Didelis sekoskaitos duomenų kiekis pakeitė iki tol buvusią nuomonę apie retus variantus bei retas ligas sukeliančius variantus. Iki tol laikyti retais variantai pasirodė nesantys reti (Karki et al., 2015). Naujausi tyrimai taip pat atskleidė, kad yra ir apsauginių genomo variantų (angl. *protective*), kurie lemia normalų fenotipą, nors tas asmuo ir turi genomo variantą, lemiantį ligą (Chen et al., 2016). Tai dar labiau pasunkina genomo varianto kaip patogeninio įvertinimą.

Iki tol vartoti terminai, skirti patogeniškumui nusakyti, tapo pasenę ir atsirado naujų (MacArthur et al., 2014), kaip antai:

- **patogeninis** (angl. *pathogenic*) – turintis tiesioginę įtaką ligos pasireiškimui, bet nebūtinai yra visiškai išreikštas, penetrantiškas;
- **turintis asociaciją su liga** (angl. *associated*) – gerokai dažniau pasireiškiantis sergančių, nei sveikų asmenų genome;
- **kenksmingas** (angl. *damaging*) – pakeičiantis geno produkto funkciją;
- **žalingas** (angl. *deleterious*) – natūralios atrankos pasirinktas pašalinti taikinys.

Vartojamas terminas **sukeliantis ligą variantas** (angl. *disease-causing variant*), tačiau dauguma tyrėjų vengia vartoti šį terminą dėl neužtikrintos įtakos ligai ir vietoj jo renkasi terminą „patogeninis“.

Atlikti publikuotų ligas sukeliančių variantų tyrimai parodė, kad net 10–27 % ir daugiau laikytų patogeniniais variantų yra visiškai nežalingi (Bell et al., 2011). Toks klaidingas patogeniškumo įvertinimas gali turėti rimtų pasekmių pacientams. Tarp tų priežasčių yra nevienodas terminų vartojimas ir patogeninių genomo variantų paskelbimas nesilaikant sisteminio variantų įvertinimo.

Poreikį suvienodinti patogeniškumo terminus 2015 metais paskelbė Amerikos medicinos genetikos ir genomikos kolegija (angl. *American College of Medical Genetics and Genomics – ACMG*). Ji pateikė penkias patogeniškumo kategorijas: 1) **patogeninis** (angl. *pathogenic*); 2) **tikėtinai patogeninis** (angl. *likely pathogenic*); 3) **neaiškios klinikinės reikšmės** (angl. *uncertain significance*); 4) **tikėtinai nepatogeninis** (angl. *likely benign*); 5) **nežalingas** (angl. *benign*), ir kriterijus, pagal kuriuos genomo variantai turėtų būti priskirti vienai ar kitai kategorijai (Richards et al., 2015). Paskelbus rekomendacijas *Jeffey Kleinberger* ir jo kolegos pateikė genomo variantų interpretavimo įrankį (angl. *Genetic Variant Interpretation Tool*), skirtą tyrėjams, nurodantį, kaip pateikti savo nustatytus genomo variantus pagal Amerikos medicinos genetikos ir genomikos kolegijos rekomendacijas (Kleinberger et al., 2016). *Gail P. Jarvik* ir *Brian L. Browning* pasiūlė Amerikos medicinos genetikos ir genomikos kolegijos rekomendacijas papildyti kiekybiniais kriterijais (Jarvik and Browning, 2016).

Ši klasifikacija taikoma viešai prieinamoje patogeninių variantų duomenų bazėje *ClinVar*, taip pat populiariausiose genomo naršyklėse, tokiose kaip *Ensembl*. Amerikos medicinos genetikos ir genomikos kolegijos siūloma patogeninių variantų klasifikacija taikoma ir šioje disertacijoje.

1.3. Sveikų asmenų tyrimai naudojant NKS

Kai kalbama apie sveikų asmenų populiacinius tyrimus, žodis „sveikas“ reiškia bendrąją populiacijos imtį, kurioje esantys asmenys neinformuoja, kad serga konkrečiomis ligomis ir laiko save sveikais. Tokie sveiki asmenys, nors ir neserga, gali būti tam tikrų paveldimų ligų nešiotojai, taip pat gali sirgti tam tikra liga to nežinodami.

1.3.1. Populiaciniai projektai, kuriuose naudota NKS

Sveikus asmenis tirti naudojant NKS pradėta pasibaigus Žmogaus genomo projektui (Lander et al., 2001). Įvykdžius šį projektą, prasidėjo didelio masto sveikų asmenų projektai, kurių pagrindinis tikslas buvo išsiaiškinti žmonių genomo variantų įvairovę, nes, nors bet kurių dviejų asmenų genomai yra daugiau nei 99 % vienodi, likusioje genomo įvairovėje gali slypėti svarbi informacija apie konkrečią tiriamą populiaciją, ligų priežastis, skirtingą vaistų poveikį asmenims.

Pirmasis didelio masto sveikų asmenų *HapMap* projektas (angl. *HapMap Project*) prasidėjo 2003 metais. Projekto metu tirta 270 asmenų iš Nigerijos, Kinijos, Japonijos, JAV ir Europos. Projekto metu buvo nustatyta 3,5 milijono vieno nukleotido polimorfizmų (*International HapMap*, 2003).

Pasibaigus šiam projektui, 2008 metais pradėtas kitas labai svarbus tarptautinis projektas – 1000 genomų projektas, kurį inicijavo *Richard Durbin* iš Jungtinės Karalystės Sengerio instituto. Pagrindiniai projekto partneriai buvo JAV Nacionalinis žmogaus genomo tyrimo institutas (angl. *National Human Genome Research Institute*), Jungtinės Karalystės Sengerio institutas (angl. *Wellcome Trust Sanger Institute*) ir Pekino genomo institutas (angl. *Beijing Genomics Institute*), prie projekto prisidėjo ir daugiau mokslininkų iš įvairių pasaulio valstybių, taigi buvo atliekamas sveikų žmonių iš viso pasaulio tyrimas. Projekto tikslas buvo nusekvenuoti daugiau nei 2 500 genomų iš viso pasaulio ir apibūdinti daugiau nei 95 % genomo variantų, kurių dažnis yra 1 % ar didesnis (Genomes Project et al., 2010).

Kitas didelis projektas buvo 6500 egzomų projektas, kuriame buvo sekvenuotas didelis JAV asmenų egzomų skaičius, siekiant palyginti Amerikoje gyvenančius asmenis, kurių protėviai buvo kaukazoidų kilmės, su tais, kurių protėviai kilę iš Afrikos (Fu et al., 2013).

Tiek 1000 genomų, tiek 6500 egzomų projektai davė labai didelę naudą ir duomenų gausą, o šių projektų duomenis tyrėjai iki šiol naudoja savo tiriamų asmenų genomo alelių dažniams palyginti su pasauliniais dažniais.

Pirmiausia sveikų asmenų genomų tyrimo kryptis buvo nustatyti genomo variantų įvairovę ir jų dažnį didelėse populiacijose. Sukaupti duomenys tapo etalonu, kuriuo naudodamiesi tyrėjai gali palyginti savo tyrimo duomenis, ir tai atvėrė galimybę tirti mažesnių populiacijų genetinę įvairovę.

Didžiųjų projektų duomenys parodė, kad didelių populiacijų, tokių kaip Amerikos, Afrikos, Europos, Azijos, genomo variantų dažnis skiriasi, todėl kita tyrimo kryptis buvo detali nedidelių populiacijų genomo variantų analizė.

Pačios įdomiausios populiacijos tyrėjams yra tos, kurios buvo izoliuotos nuo kitų populiacijų ir kuriose pasireiškė pradininko efektas arba nebuvo migracijos. Tokiose populiacijose galimas didžiulis dažnių skirtumas nuo pasaulinių dažnių ir, tikėtina, jose galima atrasti daug naujų dalykų. Pradininko efekto pavyzdys yra suomių populiacija, kur ryškus pradininko efektas; nedidelė islandų populiacija, kuri izoliuota saloje nuo žemyno ir kurioje beveik nevyko migracija, taip pat kitų izoliuotų salų populiacijos, pavyzdžiui, Kroatijai priklausanti izoliuota Vis salos populiacija (Lim et al., 2014; Gudbjartsson et al., 2015; Jeroncic et al., 2016).

Ypač įdomus yra Islandijos populiacijos tyrimas. Šioje izoliuotoje valstybėje gyvena 325 tūkstančiai gyventojų. Joje vykdytas deCODE projektas, kurio metu sekvenuoti net 2 636 islandų genomai ir gauti 104 000 islandų genotipavimo, taikant VNP lyginamosios genomo hibridizacijos metodą, duomenys. Tai šiuo metu viena iš geriausiai ištirtų populiacijų pasaulyje. Joje nustatyta 20 milijonų unikalių VNP ir 1,5 milijono trumpų iškritų ir intarpų.

Tyrimas padėjo atskleisti retas ligas, kurios gerokai dažniau pasireiškia šioje populiacijoje (Gudbjartsson et al., 2015).

Taip pat įdomus yra 50 egzomų tyrimas iš mažos Tibeto gyvenvietės kalnuose. Tyrimo metu gauti svarbūs rezultatai – nustatyti genomo variantai, kurie leidžia geriau adaptuotis būnant dideliame aukštyje (Yi et al., 2010).

Kiekvienai valstybei, nors ir labai neizoliuotai nuo kitų populiacijų, svarbu žinoti joje paplitusius genomo variantus. Tad sveikų asmenų tyrimai paplito po atskiras valstybes, pavyzdžiui, korėjiečių genomų tyrimai, 200 danų egzomų analizė ar didelės apimties Didžiosios Britanijos populiacijos tyrimas UK10K projekte, kuriame tiriami 10-ies tūkstančių asmenų genomo duomenys (Kim et al., 2009; Li et al., 2010; Ju et al., 2011; Consortium et al., 2015).

Olandijoje buvo vykdomas GoNL projektas, kurio metu atlikta 250 asmenų viso genomo sekoskaita ir gautas haplotipų genolapis, sudarytas iš 20,4 milijono VNV ir 1,2 milijono intarpų ir iškritų (Francioli et al., 2014).

Lietuvoje taip pat vykdytas populiacinis tyrimas LITGEN projekte, kurio metu atlikta lietuvių kilmės asmenų 144 egzomų analizė siekiant įvertinti lietuvių populiacijos genetinės įvairovės ir genetinės struktūros ypatumus (Uktverytė et al., 2013).

Sveikų asmenų tyrimai svarbūs ne tik todėl, kad parodo genetinę įvairovę ir jos dažnį tam tikroje populiacijoje. Vis geriau suprantant genomą ir jo sudėtį, kur yra koduojanti dalis, atskiriant, kur yra nesinoniminės pakaitos, o kur yra funkcijos praradimo mutacijos (angl. *loss of function mutations – LoF*) ir atsiradus bioinformacinių įrankių, kuriais galima įvertinti genomo variantų patogeniškumą, pastebėta, kad sveiki asmenys turi daug genomo pakaitų, kurios gali daryti įtaką sveikatai. 2012 metais *Yali Xue* ir jo kolegos paskelbė, kad kiekvienas sveiku laikomas asmuo gali turėti iki 400 žalingų patogeninių variantų (Xue et al., 2012; Lazarin et al., 2013).

Minėti sveikų asmenų genetinės įvairovės tyrimai taip pat suteikė svarbios informacijos apie konkrečiose populiacijose paplitusius patogeninius genomo variantus ir jų sąsajas su ligomis. Islandų populiacijoje rasta 300 tūkstančių funkcijos praradimo mutacijų, nustatyta penkis kartus didesnė rizika susirgti bet kuriuo vėžiu, taip pat nustatyti genomo variantai, kurie padvigubina tikimybę susirgti Alzheimerio liga (Gudbjartsson et al., 2015). Suomių populiacijoje aptiktos 83 funkcijos praradimo mutacijos, kurių MAF tarp suomių yra du kartus didesnis nei likusioje Europos populiacijoje (Lim et al., 2014).

Tokių funkcijos praradimo mutacijų tyrimų būta ir daugiau, pavyzdžiui, *Daniel G. MacArthur* su bendraautoriais, tirdami 185 sveikus asmenis, nustatė, kad kiekvieno asmens genome yra per 100 funkcijos praradimo mutacijų ir apie 20 genų, kurie visiškai inaktyvinti (MacArthur and Tyler-Smith, 2010; MacArthur et al., 2012).

Atsiradus NKS, paveldimų genomo ligų priežasčių tyrimai vykdyti itin sparčiai, rasta daugybė patogeninių genomo variantų, tiesiogiai ar netiesiogiai lemiančių ligą. Duomenų bazės, kuriose kaupiama informacija apie patogeninius genomo variantus, yra šios: žmogaus genų mutacijų duomenų bazė *HGMD®* (angl. *Human Gene Mutation Database*) ir viešai pasiekiama patogeninių genomo variantų duomenų bazė *ClinVar* (Krawczak et al., 2000; Landrum et al., 2014).

Didelė žinių apie patogeninius variantus gausa leidžia analizuoti žinomus patogeninius variantus sveikuose asmenyse. Tokių tyrimų nebuvo vykdyta labai daug, nors pradinis tokio pobūdžio tyrimas atliktas *Callum J. Bell* ir jo kolegų 2011 metais (Bell et al., 2011).

1.3.2. NKS naudojimas visuotinėje naujagimių patikroje

Visuotinė naujagimių patikra (VNP_A) – tai didelės imties paveldimų ligų identifikavimas naujagimiams. Laiku identifikavus šias ligas galima pagerinti sergančio vaiko būklę.

VNP_A pradžia galima laikyti 1950 metus, kai buvo sukurtas veiksmingas paveldimos medžiagų apykaitos ligos fenilketonurijos gydymas. Pats naujagimių tikrinimas nebuvo labai efektyvus – taikytas geležies trichlorido testas, kuris leido identifikuoti ligą tik jau esant didelei fenilalanino koncentracijai kraujyje. 1963 metais *Robert Guthrie* sukūrė naują kraujo paėmimo būdą, kuris leido išdžiovinti naujagimio kraują ant filtrinio popieriaus, ir naują fenilalanino koncentracijos kraujyje nustatymo būdą taikant bakterijų augimo inhibicijos metodą (Guthrie and Susi, 1963).

Patogi *Robert Guthrie* metodika paskatino daugelį valstybių pradėti visuotinę naujagimių patikrą. Prasidėjus VNP_A., sąrašas ligų, dėl kurių tikrinami naujagimiai, sparčiai ilgėjo. Tai daugiausia paveldimos medžiagų apykaitos ligos, kurias nustačius ir pritaikius specialią dietą, labai pagerėja paciento gyvenimo kokybė, nes ankstyvas šių ligų nustatymas palengvina sergančiųjų jomis gyvenimą. Kiekviena valstybė turi savo patvirtintą ligų, kurios tiriamos visuotinės naujagimių patikros metu, sąrašą.

Lietuvoje visuotinė naujagimių patikra pradėta 1975 metais – tikrinta dėl fenilketonurijos, nuo 1993 metų – dėl įgimtos hipotirozės, nuo 2015 metų – dėl galaktozemijos ir dėl įgimtos antinksčių hiperplazijos (Kučinskas et al., 1996; Cimbalistienė, 2008).

Jungtinėse Amerikos Valstijose VNP_A ligų sarašas nuolat atnaujinamas. Šiuo metu visose valstijose VNP_A vykdoma 31 įgimtam sutrikimui nustatyti. Į pagrindinį 31 ligos sąrašą pirmą kartą įtraukta ir imuninė liga – sunkaus kombinuoto imunodeficito liga. Be šių 31 ligos, kiekvienoje atskiroje valstijoje visuotinės naujagimių patikros metu nustatomų ligų skaičius varijuoja ir siekia

iki 56 ligų (Calonge et al., 2010; Meade and Bonhomme, 2014; Kwan and Puck, 2015).

2008 m. Viskonsino valstija (JAV) buvo pirmoji pasaulyje, kuri atliko visų naujagimių patikrą sunkaus kombinuoto imunodeficito ligai nustatyti. Taip pat tai buvo pirmas kartas, kai VNP_A analizei molekuliniai genetiniai tyrimai buvo panaudoti kaip pirminiai, o ne antriniai tyrimai (Baker et al., 2010).

Atsiradus naujos kartos sekoskaitai ir jai labai atpigus, ji buvo pradėta naudoti VNP_A. 2013 metais JAV Nacionalinis žmogaus genomo tyrimo institutas (angl. *The National Human Genome Research Institute – NHGRI*) (JAV) skyrė 25 milijonus JAV dolerių programai, kuri padėtų išaiškinti NKS panaudojimo VNP_A galimybes ir problemas (Knoppers et al., 2014). Šios programos remiami projektai tebevyksta iki šiol, tačiau jau galima išskirti viso egzomo sekoskaitos ar viso genomo sekoskaitos problemas.

Viena iš didžiausių NKS panaudojimo VNP_A problemų yra lėšų, skiriamų šiems tyrimams, kiekis. Šiuo metu molekuliniai tyrimai dažniausiai atliekami kaip antriniai tyrimai, prieš tai įvykdžius metabolitų tyrimus, kurie yra kur kas pigesni, o NKS, nors ir labai atpigus, visgi yra daugumai valstybių, įskaitant ir JAV, per brangi, kad būtų panaudota VNP_A. Pavyzdžiui, VNP_A procedūra 2011 metais Serbijoje kainavo tik 0,46 euro vienam naujagimiui tiriant du sutrikimus, Olandijoje – 43,24 euro tiriant 17 sutrikimų. Tai gerokai mažesnės išlaidos nei molekulinių genetinių tyrimų konkrečiame gene, o juo labiau pigiau nei NKS tyrimai. Tačiau net turtingai šaliai, tokiai kaip JAV, kur kasmet gimsta 4 milijonai naujagimių, NKS yra per brangi (Burgard et al., 2014; Howard et al., 2015).

Kita problema yra gaunamas didelis NKS duomenų kiekis. Didelis informacijos kiekis apsunkina duomenų analizės etapą, reikalauja daug kompiuterinių išteklių analizei ir duomenų saugojimui.

Nors NKS nėra plačiai taikoma VNP_A ir vyksta tik pavieniai populiaciniai sveikų naujagimių tyrimai, pati NKS jau vaidina svarbų vaidmenį klinikinėje

diagnostikoje. Tokie tyrimai atliekami tiek suaugusiems asmenims, tiek naujagimiams, kai aptinkamas nežinomas, galintis būti įgimtas sutrikimas. Čia kyla etinių klausimų dėl vadinamųjų antrinių rezultatų (angl. *secondary findings*), kurie būna nesusiję su pagrindine tyrimo priežastimi, bet gali būti svarbūs tiriamajam ir jo sveikatos būklei ateityje. Svarstoma, ar derėtų pacientui skelbti antrinius rezultatus, ar ne (Gambin et al., 2015).

Europos žmogaus genetikos draugijos rekomendacija yra naudoti VNP genotipavimą konkrečiam genų skaičiui ir geriau nenaudoti viso egzomo sekoskaitos ar viso genomo sekoskaitos, taip išvengiant antrinių rezultatų (Howard et al., 2015).

Amerikos medicinos genetikos ir genomikos kolegijos (angl. *The American College of Medical Genetics and Genomics – ACMG*) rekomendacijos yra nuolaidesnės NKS ir antrinių rezultatų atžvilgiu. Rekomendacijose siūloma pateikti antrinius rezultatus iš rekomenduojamų 56 genų, kuriuose esantys pakitimai gali turėti didelę įtaką konkretiems sutrikimams (Green et al., 2013).

1.4. Bioinformacinė NKS duomenų analizė

Sekoskaitos prietaisams tobulėjant ir pingant, proporcingai turi tobulėti ir bioinformaciniai įrankiai, kuriais galima atlikti sekoskaitos duomenų analizę.

Atliekant *Sanger* sekoskaitą ir analizuojant pavienius genų egzonus (kai vidutinis egzono ilgis yra apie 150 nukleotidų), tyrėjui nereikia naudoti sudėtingo analizės algoritmo. Visai kitokia situacija yra atliekant viso egzomo analizę, kai yra 50 milijonų nukleotidų, ar atliekant viso genomo analizę, kai yra 3,2 milijardo nukleotidų. Tad žmogaus genetikas, norintis neatsilikti nuo žmogaus genetikos tyrimų pažangos, turi išmanyti NKS analizės algoritmus ir gebėti apdoroti gautus duomenis.

1.4.1. NKS analizės algoritmas

NKS duomenų analizė prasideda gavus pirminius sekoskaitos duomenis (angl. *raw data*). Pradiniai NKS analizės duomenys – tai daugybė trumpų DNR fragmentų (angl. *reads*) kartu su kiekvieno nukleotido kokybės vertėmis.

Pirminių duomenų analizės formatai skiriasi – priklauso nuo sekoskaitos sistemos gamintojo. Labiausiai paplitęs pradinis analizės formatas yra FASTQ, kuris automatiškai suformuojamas iš .bcl failų *CASAVA* programa. Tyrėjų šis formatas yra naudojamas dažniausiai, nes jis yra *Illumina* gaminio pradinis formatas, o *Illumina* bendrovė kontroliuoja daugiau nei 70 % visos NKS rinkos (Thayer, 2014). Kitų gamintojų formatai yra panašūs į FASTQ formatą arba lengvai į jį transformuojami, išskyrus *Applied Biosystems SOLiD* sekoskaitos sistemą, kurioje pradiniai duomenys yra XSQ formatu, o jis skiriasi nuo FASTQ, nes faile yra užkoduotas spalvinis kodas.

Kad galėtų pradėti analizuoti NKS duomenis, tyrėjas pirmiausia turi atlikti bioinformacinę analizę, kuri susideda iš NKS analizės algoritmo (angl. *pipeline*) ir NKS duomenų anotacijos (angl. *annotation*).

NKS analizės algoritmas – tai seka bioinformacinio pobūdžio komandų, kurios sujungia kelių programų darbą tam, kad būtų gaunami tiriamų duomenų genomo variantai, kurie skiriasi nuo referentinio genomo. Kad ir kas būtų sekoskaitos gamintojas, NKS analizės algoritmas susideda iš dviejų pagrindinių etapų: tiriamų duomenų prilygiavimo prie referentinio genomo ir teisingų genomo variantų atrankos.

NKS analizės programinius paketus pateikia patys sekoskaitos gamintojai, taip pat galima naudoti trečiųjų šalių programas. Dirbant su *Illumina* sekoskaitos duomenimis, jau yra įsigalėjęs analizės algoritmo „pavyzdinis standartas“, kurį sudaro nuskaitytų pirminių duomenų sulygiavimas su referentiniu genomu, panaudojant *Burrows-Wheeler* prilygiavimo programą *BWA* (angl. *Burrows-Wheeler Aligner*) ir teisingų variantų atranką genomo analizės paketu *GATK* (angl. *The Genome Analysis Toolkit*) (Li and Durbin, 2010; DePristo et al.,

2011). *SOLiD* pirminių duomenų, kurie gerokai skiriasi nuo FASTQ pirminio formato, analizės algoritmas neturi plačiai taikomo trečiųjų šalių programų pavyzdinio standarto.

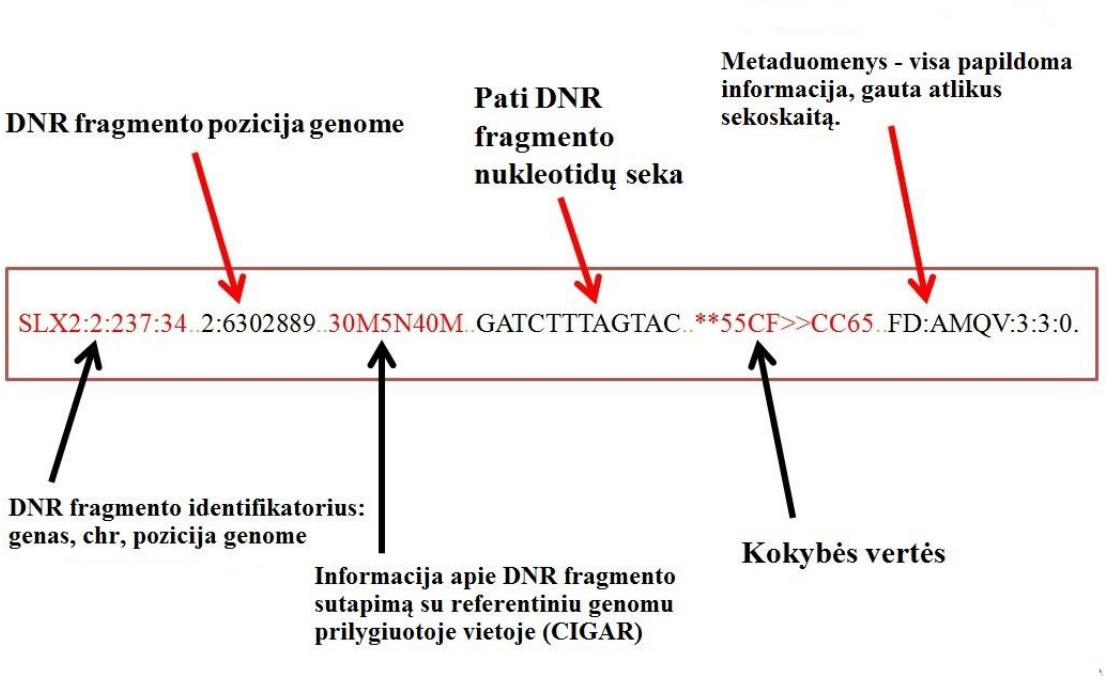
Pirmame NKS analizės algoritmo etape pirminiai sekoskaitos duomenys yra sulygiuojami su referentiniu genomu.

Oficialiai žmogaus genomo projektas buvo baigtas 2003 metais, tačiau net dabar, 2016 metais, dar visiškai nežinoma visa žmogaus genomo seka, nes dar yra nesekvenuotų vietų.

Nežinomi genominės sekos tarpai (angl. *gap*) dažniausiai yra sudaryti iš didelio nukleotidų pasikartojimo. Tobulėjant sekoskaitai, vis daugiau sekvenuojama sekų, kurių iki to laiko nebuvo galima sekvenuoti. Tad nuo 2003 metų buvo sukurta daugybė referentinio genomo versijų, kurias leidžia naudoti žmogaus referentinio genomo konsorciumas (angl. *Human Genome Reference Consortium*). Žinoti, kokia referentinio genomo versija buvo naudota NKS analizės algoritme, yra labai svarbu atliekant tolesnę NKS duomenų analizę, nes skirtingose referentinėse genomo versijose gali smarkiai skirtis pakaitos genominė pozicija.

Trumpi DNR fragmentai, naudojant prilygiavimo programas, tokias kaip *BWA* ar *Lifescope*, yra prilygiuojami prie referentinio genomo sekos. DNR fragmentų yra labai didelis skaičius, ir jie tarpusavyje susikloja, taip sukurdami labai svarbų parametrą – susiklojimą (angl. *coverage*).

Prilygiavimo programos sukuria binarį BAM formatą, kuris yra vienodas visoms sekoskaitos sistemoms ir kuriame yra saugoma informacija apie DNR fragmentus ir jų prilygiavimo prie referentinio genomo rezultatus (2 pav.).



2 pav. Prilygiavimo prie referentinio genomo BAM failo struktūra.

Be persidengiančių DNR fragmentų, dar yra ir tų pačių DNR fragmentų kopijų – duplikacijų. Jos duomenų analizėje yra nepageidautinos, nes gali iškraipyti tyrimo rezultatus. Jei DNR fragmente, kuris turi daug kopijų, sekoskaitos metu įvyksta klaida ir nuskaitomas neteisingas nukleotidas, dėl duplikacijos teisingų variantų atrankos etape toks nukleotidas būtų nuskaitytas kaip tikrai besiskiriantis nuo referentinio genomo, ir tai galėtų lemti klaidingus rezultatus. Tad DNR fragmentų duplikacijos po prilygiavimo prie referentinio genomo etapo yra pašalinamos. Tai atlieka BAM formatą sugebančios tvarkyti programos, tokios kaip *SAMtools* (Li, 2011).

SAMtools programa ne tik pašalina DNR fragmentų duplikacijas, bet ir atlieka kitas svarbias analizės algoritmui funkcijas, kaip antai duomenų failo indeksavimą (angl. *index*) ir išrūšiavimą (angl. *sorting*). Indeksavimas leidžia tolesniuose etapuose programoms neskaityti visos DNR sekos iki dominančios vietos, o atsidurti dominančiose vietose akimirksniu, tad pagelbėja kaip knygose esanti rodyklė. Kadangi NKS metu gaunamas didelis duomenų kiekis, toks indeksavimas, atsižvelgiant į tyrimo tikslą, gali sutaupyti labai daug analizės

laiko. Išrūšiavimas – tai duomenų išrikiavimas pagal tam tikrą parametrą, analizės algoritme tyrimo duomenys būna išrikiuojami pagal chromosominę poziciją, t. y. nuo pirmosios iki 22 chromosomos.

BAM formatu yra saugomi DNR fragmentai, bet patį tyrėją domina ne šie DNR fragmentai, o genomo variantai, kurie skiriasi nuo referentinio genomo. Pačiuose DNR fragmentuose ne visi nuo referentinio genomo besiskiriantys nukleotidai yra teisingi ir gali būti atsiradę dėl įvairių klaidų.

Tokių neteisingų nukleotidų yra dauguma, tad po prilygiavimo prie referentinio genomo etapo yra vykdomas sudėtingas etapas – teisingų genomo variantų atranka. Ji atliekama naudojant specializuotas programas, kaip antai *GATK*. Nuo naudojamos programos priklauso, ar variantų atranka turi mažiau, ar daugiau etapų, pavyzdžiui, *GATK* turi papildomos kokybės kontrolės kalibravimo, duomenų sulyginimo su populiaciniais duomenimis etapus.

Tačiau, kad ir kokia būtų programa, pagrindinis teisingų variantų atrankos darbas yra remtis iki šio etapo gautais parametrais ir atrinkti teisingus genomo variantus. Šiam uždaviniui išspręsti naudojami keli pagrindiniai parametrai:

1) Padengimas (angl. *coverage*) – DNR fragmentai susikloja, tad kiekvienas atskiras nukleotidas nuskaitomas kelis kartus. Laikytina, kad kuo daugiau kartų tas pats nukleotidas pasikartoja, tuo geresni yra rezultatai, išskyrus tuos atvejus, kai didelis nukleotido pasikartojimų skaičius sukuriamas dirbtinai (pvz., dėl duplikacijų).

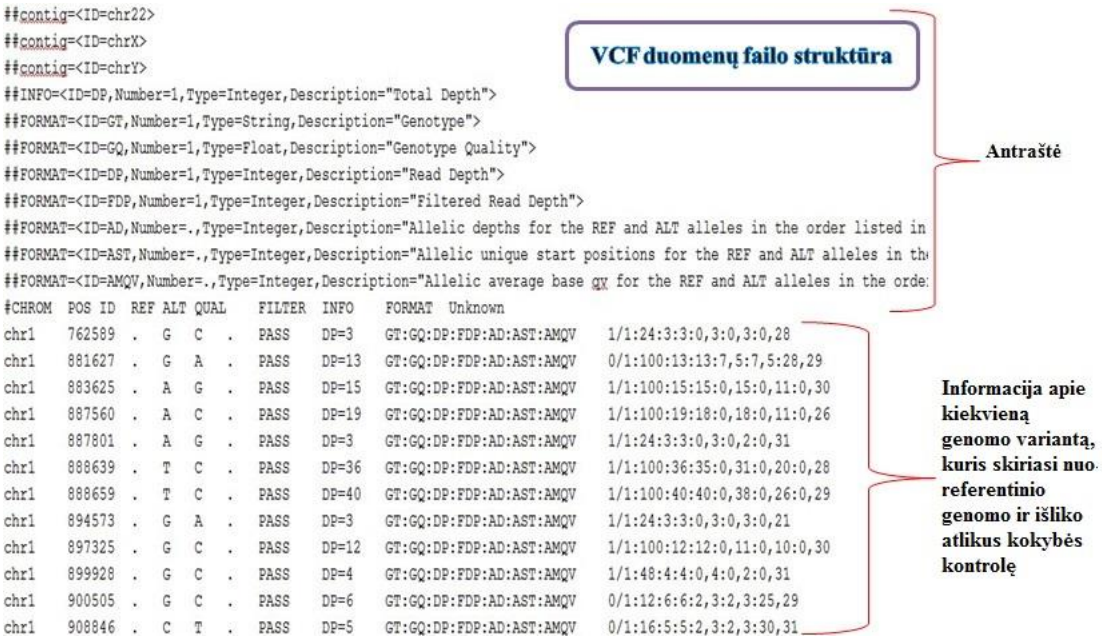
2) Kiekvieno nukleotido kokybės vertė (angl. *base quality score*) – tai vienintelis vertinimo parametras, kuris suformuojamas paties sekoskaitos prietaiso. Pateikiama *Phred* tikimybinė vertė, kuri parodo neteisingai identifikuoto nukleotido tikimybę. Tai įvertis, kuris įvertina kiekvieną atskirą nukleotidą.

3) DNR fragmento kokybės vertė (angl. *mapping quality score*) – tai įvertis, kuris įvertina kiekvieną atskirą DNR fragmentą. Labiausiai šis kokybės įvertis priklauso nuo žmogaus genomo sekos homozigotiškumo.

Vienodų sekos fragmentų gali būti skirtingose chromosomose, pavyzdžiui, pirmoje, penktoje ir septynioliktoje. Disertacijoje naudotos *SOLiD* sistemos sekvenuoti DNR fragmentai yra trumpi, iki 75 bp ilgio, tad, vykdant prilygiavimo prie referentinio genomo etapą, programa dažnai susiduria su uždaviniu, kuriai genominei vietai priskirti DNR fragmentą, kai seka yra identiška keliose vietose. Programa tokiam DNR fragmentui suteikia mažą DNR fragmento kokybės vertę, taip informuodama tyrėją ir kitas programas, kad gali būti padaryta klaida.

4) CIGAR – tai parametras, kuris parodo, kaip tiksliai DNR fragmentas sutapo su referentiniu genomu. Tarkime, DNR fragmentas yra 75 nukleotidų ilgio, ir jei visas fragmentas sutampa su referentiniu genomu, tai yra ideali situacija. Tačiau kartais DNR fragmentas panašus į referentinę seką, bet turi mažų skirtumų, pavyzdžiui, dalis nukleotidų sutampa, tada yra keturių nukleotidų tarpas ir tada vėl seka sutampa – visą tai aprašo CIGAR.

Teisingų variantų atrankos etapas atsižvelgia į šiuos parametrus, atskiria genomo variantus, kurie tikrai skiriasi nuo referentinės sekos, ir pateikia rezultatus VCF formatu (3 pav.).




```
##contig=<ID=chr22>
##contig=<ID=chrX>
##contig=<ID=chrY>
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=FDP,Number=1,Type=Integer,Description="Filtered Read Depth">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the REF and ALT alleles in the order listed in
##FORMAT=<ID=AST,Number=.,Type=Integer,Description="Allelic unique start positions for the REF and ALT alleles in th
##FORMAT=<ID=AMQV,Number=.,Type=Integer,Description="Allelic average base qv for the REF and ALT alleles in the orde:
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  Unknown
chr1    762589  .   G   C   .   PASS    DP=3    GT:GQ:DP:FDP:AD:AST:AMQV    1/1:24:3:3:0,3:0,3:0,28
chr1    881627  .   G   A   .   PASS    DP=13   GT:GQ:DP:FDP:AD:AST:AMQV    0/1:100:13:13:7,5:7,5:28,29
chr1    883625  .   A   G   .   PASS    DP=15   GT:GQ:DP:FDP:AD:AST:AMQV    1/1:100:15:15:0,15:0,11:0,30
chr1    887560  .   A   C   .   PASS    DP=19   GT:GQ:DP:FDP:AD:AST:AMQV    1/1:100:19:18:0,18:0,11:0,26
chr1    887801  .   A   G   .   PASS    DP=3    GT:GQ:DP:FDP:AD:AST:AMQV    1/1:24:3:3:0,3:0,2:0,31
chr1    888639  .   T   C   .   PASS    DP=36   GT:GQ:DP:FDP:AD:AST:AMQV    1/1:100:36:35:0,31:0,20:0,28
chr1    888659  .   T   C   .   PASS    DP=40   GT:GQ:DP:FDP:AD:AST:AMQV    1/1:100:40:40:0,38:0,26:0,29
chr1    894573  .   G   A   .   PASS    DP=3    GT:GQ:DP:FDP:AD:AST:AMQV    1/1:24:3:3:0,3:0,3:0,21
chr1    897325  .   G   C   .   PASS    DP=12   GT:GQ:DP:FDP:AD:AST:AMQV    1/1:100:12:12:0,11:0,10:0,30
chr1    899928  .   G   C   .   PASS    DP=4    GT:GQ:DP:FDP:AD:AST:AMQV    1/1:48:4:4:0,4:0,2:0,31
chr1    900505  .   G   C   .   PASS    DP=6    GT:GQ:DP:FDP:AD:AST:AMQV    0/1:12:6:6:2,3:2,3:25,29
chr1    908846  .   C   T   .   PASS    DP=5    GT:GQ:DP:FDP:AD:AST:AMQV    0/1:16:5:5:2,3:2,3:30,31
```

3 pav. Teisingų genomo variantų VCF failo struktūra.

VCF faile yra nurodomi visi genomo variantai, kurie skiriasi nuo referentinio (ALT), nurodoma jų chromosoma (#CHOM), genominė pozicija (POS ID), kokie nukleotidai yra referentiniame genome (REF), kiekvieno genomo varianto padengimas (INFO:DP), homozigotiškumo būklė (FORMAT:GT) ir kiti parametrai, svarbūs tolesnei analizei. VCF struktūros sudėtis detaliau aprašyta 1 priede „VCF failų pagrindiniai komponentai“.

1.4.2. Anotavimas

Atlikus NKS analizės algoritmą, gaunami genomo variantai, prie kurių yra prijungiami anotaciniai įrankiai – įvairios duomenų bazės ir tikimybiniai algoritmai. Egzomo sekoskaitos atveju gaunama apie 30–50 tūkstančių genomo variantų, kurie skiriasi nuo referentinio genomo. Anotuojant prie kiekvieno iš genomo variantų yra prijungiama anotacinių įrankių informacija. Tai leidžia tyrėjui atlikti genomo variantų atranką jį dominančiais aspektais (Rančelis et al., 2013). Įrankius galima suskirstyti į kelias grupes. Toks anotacinių įrankių sugrupavimas pateikiamas 1 lentelėje.

1 lentelė. Anotacinių įrankių grupės

Anotacijų grupės	Anotaciniai įrankiai
Anotacijos, kuriomis pagal pakaitos poziciją gaunama visa informacija apie geną, kuriame yra pakaita	*Refseq*, *Ensembl*, *UCSC*
Anotacijos, rodančios pakaitos įtaką baltymui	*SIFT*, *PolyPhen*, *MutationTaster*, *MutationAssessor*, *LRT*, *FATHMM*
Anotacijos, nurodančios pakaitos dažnį įvairiose žmonių populiacijose	1000 genomų projektas, 6500 egzomų projektas (ESP)
Anotacijos konservatyvumui nustatyti	*GERP++*, *SiPhy*, *phyloP*, *phastCons*
Anotacijos, kuriomis pagal pakaitos poziciją pateikiama informacija apie su šia pakaita siejamas ligas	*OMIM*, *ClinVar*, *COSMIC*, *ORPHANET*
Anotacijos, jungiančios kitus anotacinius įrankius ir duomenų bazes	*dbSNP*, *CADD*

1.4.2.1. Anotacijos, tinkančios populiaciniams tyrimams

Tai informacinės anotacijos, kurios suteikia informaciją apie iki tol sukauptus duomenis, kurie susiję su konkrečiu genomo variantu.

Refseq, *Ensembl*, *UCSC* duomenų bazių anotacijos gali suteikti informaciją apie tai, kur ta pakaita įvyko – negeninėje ar geninėje dalyje, egzone ar introne. Jei pakaita yra koduojančioje dalyje, šios anotacijos nurodys, kokiame gene ji yra. Taip pat šios anotacijos nurodo pakaitos tipą. Jeigu pakaita yra egzone, yra nurodoma, ar ji yra sinoniminė (angl. *silent*), ar nesinoniminė (angl. *missens* ir *nonsens*) pakaita (Hubbard et al., 2002; Karolchik et al., 2003; Pruitt et al., 2007).

Baigus Žmogaus genomo projektą, buvo vykdyti keli dideli populiaciniai projektai. 1.3 skyrelyje minimų 1000 genomų, 6500 egzomų projektų duomenys yra vieši ir naudojami anotacijose palyginti tiriamų asmenų genomo variantų dažnį su kitų populiacijų dažniais.

Taip pat yra duomenų bazių, kuriose kaupiama informacija apie žinomas pakaitas ir jų įtaką ligoms. Geriausi tokių duomenų bazių pavyzdžiai yra *OMIM* (angl. *Online Mendelian Inheritance in Man*) – mendelinio paveldėjimo duomenų bazė, kurioje saugoma informacija apie visas žinomas mendelines ligas, ir *ClinVar*, kurioje saugoma informacija apie genomo variantus, turinčius įtakos žmogaus sveikatai (Hamosh et al., 2005; Landrum et al., 2014).

1.4.2.2. Anotacijos, skirtos tik paveldimų ligų priežastims nustatyti

Išskirtinai ligų priežastims nustatyti skirti anotaciniai įrankiai yra algoritminiai įrankiai, kurie tik apskaičiuoja tikimybę, ar tam tikras genomo variantas yra funkciškai svarbus ar, tikėtina, yra patogeninis.

Visų pirma, tai anotacinių įrankių grupė, kuri nurodo genomo varianto ar regiono, kuriame yra genomo variantas, konservatyvumą. Kuo genomo regionas yra konservatyvesnis ir kuo labiau jame nelinkę būti pokyčių, tuo didesnė tikimybė, kad tas regionas yra funkciškai svarbus ir jame atsiradusios pakaitos turės funkcinę įtaką. Konservatyvumo įrankių spektras gana platus, paminėtini *PhastCons* – regionų, o *phyloP*, *GERP++* – pavienių nukleotidų konservatyvumui parodyti (Siepel et al., 2005; Garber et al., 2009; Davydov et al., 2010; Pollard et al., 2010).

Taip pat yra anotacinių įrankių, kurie nurodo galimą genomo varianto įtaką koduojamam baltymui. Populiariausi įrankiai, priklausantys šiai grupei, yra *PolyPhen* ir *SIFT* (Adzhubei et al., 2010 ; Sim et al., 2012).

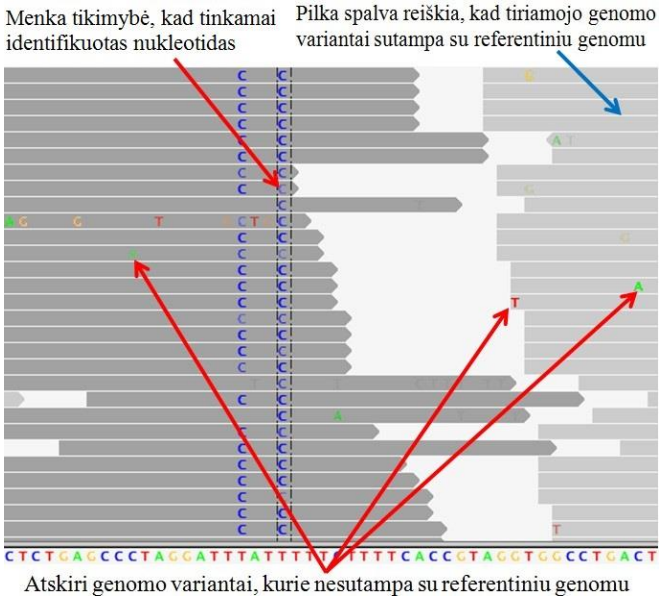
Yra anotacinių įrankių, kurie sugeba į visumą sujungti daugumą grupių, paminėtų 1 lentelėje, ir pateikti tikimybinę vertę, kad genomo variantas yra funkciškai svarbus. Paminėtinas šios grupės anotacinis įrankis yra *CADD* (angl. *Combined Annotation Dependent Depletion*). Šis įrankis turi pranašumą prieš daugumą kitų anotacijų, nes gali įvertinti genomo variantus, esančius ir negeninėje genomo dalyje (Kircheret al., 2014).

Anotacinius įrankius į visumą sujungia specializuotos genomo variantų anotavimo programos. Į galinę anotaciją įtraukiama daugiau ar mažiau anotacinių įrankių (tai priklauso nuo programos), gana nemažas spektras programų, tiek komercinių, kaip antai *Geneious*, *Ingenuity Variant Analysis*, *GoldenHelixSNP&Variation Suite* (SVS), kurios, be anotacijų, turi integruotus ir vizualizacijos įrankius, galimybę palyginti triadas, tiek nemokamų (Buckingham et al., 2008, Howard et al., 2014).

Dauguma tyrėjų naudoja *ANNOVAR*, kuri nuolat atnaujinama, turi platų spektrą anotacinių įrankių ir yra universali visiems organizmams (Yang and Wang, 2015).

Kai tyrėjas atsirenka jį dominančius genomo variantus, dažniausiai įvertinama, kaip genomo variantai atrodo DNR fragmentuose. Tam naudojamas BAM failas ir vizualizacijos programos, kaip antai integratyvi genomo vaizduoklė (angl. *Integrative Genomics Viewer – IGV*) (Thorvaldsdottiret al., 2013), (4 pav.).

4 pav. Genomo vizualizacija *IGV* programa.



2. TYRIMO METODAI

2.1. Tiriamieji

Disertacijos tikslui ir uždaviniams įgyvendinti buvo panaudoti VU MF Žmogaus ir medicininės genetikos katedroje 2011–2015 metais vykdyto projekto LITGEN „Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis“, duomenys (projektas finansuotas Europos socialinio fondo lėšomis; projekto vadovas LMA tikrasis narys prof. habil. dr. Vaidutis Kučinskas).

Tyrimams yra gautas Vilniaus regioninio biomedicininių tyrimų etikos komiteto leidimas (Nr. 158200-05-329-79; data: 2011-05-03), gauti tiriamų asmenų pasirašyti sutikimai dėl dalyvavimo šiame projekte. Kiekvieno asmens mėginiui ir duomenims buvo suteiktas specialus kodas, kad būtų apsaugotas tiriamų asmenų privatumas ir užtikrinta jų duomenų sauga.

Mėginius ėmę mokslo darbuotojai: dr. Neringa Burokienė, Raimonda Meškienė, Aidas Pranculis, dr. Laima Ambrozaitytė, dr. Ingrida Domarkienė, dr. Ingrida Kavaliauskienė.

Vykdant LITGEN projektą, buvo imtas veninis kraujas, kraujo plazma ir užpildyti 1 000 tiriamų asmenų (279 triadų) anketiniai duomenys. Visi tiriamieji buvo lietuviai, kurių bent trys kartos yra gimusios Lietuvoje. Tiriamų asmenų duomenys suskirstyti į šešias Lietuvos etnolingvistines grupes – Pietų Aukštaitiją (PA), Rytų Aukštaitiją (RA), Vakarų Aukštaitiją (VA), Pietų Žemaitiją (PŽ), Šiaurės Žemaitiją (ŠŽ), Vakarų Žemaitiją (VŽ) (Uktverytė et al., 2013; Uktverytė, 2014).

Disertacijoje aprašomame tyrime panaudoti LITGEN projekte sekvenuoti viso egzomo duomenys. Iš viso projekte buvo sekvenuoti 144 triadų (abu tėvai ir probandas) egzomai. Disertacijoje buvo panaudoti 96 asmenų duomenys, neįtraukiant probandų duomenų (2 lentelė).

2 lentelė. Tiriamųjų, kuriems atlikta egzomo sekoskaita, atsižvelgiant į lytį ir etnolingvistines grupes, skaičius

Tiriamųjų imtis							
Lytis		Etnolingvistinės grupės					
Vyrai	Moterys	PA	RA	VA	PŽ	ŠŽ	VŽ
48	48	16	16	16	16	16	16

Probandų duomenys neįtraukti, nes tai dirbtinai padidintų patogeninių variantų dažnius, todėl populiacinio pobūdžio tyrimuose tiriami tik negimininingi asmenys. Tirtas vienodas vyrų ir moterų skaičius, po 48 asmenis.

2.2. Tyrimo strategija

Rengiant disertaciją vykdytų tyrimų strategija rėmėsi trimis faktais:

1. Plataus masto sekoskaita per trumpą laikotarpį turėjo didelę įtaką genetinių ligų priežasčių nustatymui.
2. Pasaulyje jau yra atlikta daugybė genetinių ligų tyrimų, sukaupta nemažai informacijos apie paveldimų ligų priežastis, atsirado sistematizuotų ir informatyvių duomenų bazių, kuriose saugoma informacija apie žmogaus patogeninius genomo variantus.
3. Padaugėjo žmonių populiacinių tyrimų, kuriais nustatyta, kad patogeninių genomo variantų turi ir sveiki asmenys.

Disertacijoje siekta sujungti visus šiuos tris faktus, panaudojant LITGEN projekto egzomų sekoskaitos duomenis ir žinomą informaciją apie patogeninius genomo variantus, siekant identifikuoti asmenų iš bendrosios lietuvių populiacijos turimus patogeninius genomo variantus.

LITGEN projekto metu pirmą kartą Lietuvoje atlikta žmogaus plataus masto sekoskaita. Atliekant tokius NKS tyrimus, susiduriama su dideliu duomenų kiekiu, kuris reikalauja teisingos bioinformacinės analizės, todėl numatytoje strategijoje vienas iš pagrindinių uždavinių buvo sukurti tinkamas

bioinformacines sąlygas ir parametrus lietuvių populiacijos NKS duomenims analizuoti. Dėl pasirinktos tyrimo strategijos buvo būtina tinkamai parinkti NKS analizės algoritmą ir atlikti anotaciją, parenkant tinkamas sekoskaitai programas ir parametrus.

Disertacijoje taip pat siekta patikrinti, ar gauti sekoskaitos duomenys nėra klaidingi, atsižvelgiant į tranzicijų ir transversijų pakaitų santykį, homozigotinių ir heterozigotinių pakaitų santykį tyrimo duomenyse, įvertinant DNR fragmentų padengimo lygį.

Gavus galutinius anotuotus tyrimo duomenis, buvo siekiama panaudoti iki šiol sukauptą informaciją apie patogeninius genomo variantus ir nustatyti, ar jų turi sveiki asmenys.

Tyrimo strategija įpareigojo palyginti Lietuvos asmenų duomenis su kitų populiacijų duomenimis ir pasitelkti statistinius modulius, siekiant pateikti statistiškai patikimus skirtumus.

Vykdant disertacinį darbą taip pat buvo keliamas uždavinys įvertinti naudotas žmogaus patogeninių genomo variantų duomenų bazes, ieškant paskelbtų patogeninių genomų variantų, kurių priskyrimas prie patogeninių yra labai abejotinas.

2.3. Tyrimo eiga ir metodai

Tyrimo eigą galima suskirstyti į keturis etapus: darbas pagal laboratorinį protokolą, kuris buvo atliktas VU MF ŽMGK darbuotojų, NKS analizės algoritmo, anotacijos ir galutinio duomenų vertinimo, kuriuos atliko pats doktorantas.

2.3.1. Laboratorinis protokolas: DNR išskyrimas ir sekoskaita

Disertacijoje naudojamų duomenų laboratorinį protokolą vykdę asmenys:

1. DNR išskyrimą atliko Daiva Kazlauskaitė ir Dalytė Pliaugo.
2. Optimizuotą sekoskaitos protokolą įdiegė ir įvykdė dr. Ingrida Kavaliauskienė, dr. Laima Ambrozaitytė, dr. Ingrida Domarkienė.
3. Patogeninių genomo variantų egzomų sekoskaitą, tikrinimą *Sanger* sekoskaita atliko VU MF medicininės genetikos studijų programos magistrantė Kristina Aleknavičienė.

2.3.1.1. DNR išskyrimas

DNR išskyrimas iš veninio kraujo atliktas dviem būdais – taikant fenolio ir chloroformo metodą bei automatizuota TECAN Freedom EVO®200 sistema (gamintojas *Tecan Group Ltd.*, Šveicarija), kuri remiasi magnetinių dalelių prijungimo prie DNR principu.

DNR koncentracija bei švarumas pamatuoti NanoDrop® 1000 spektrofotometru (gamintojas *Thermo Fisher Scientific*, Vilmingtonas, JAV).

2.3.1.2. Sekoskaita ir jos etapai

Tyrimo metu buvo naudoti sekoskaitos duomenys, gauti paruošus pavienių DNR fragmentų biblioteką (angl. *fragment library*).

Biblioteka buvo pradėta ruošti nuo genominės DNR fragmentavimo ultragarsu naudojant Covaris®S220 sistemą (gamintojas *Covaris, Inc.*, Masačusetsas, JAV) etapo. Prietaisas fragmentuoja genominę DNR į įvairaus ilgio DNR fragmentus, kurie dažniausiai būna apie 160 bp ilgio.

Po DNR suskaldymo DNR fragmentai su lipniais 5` ir 3` DNR galais paversti į DNR fragmentus, turinčius bukus galus, pasitelkiant fermentus, turinčius 5`→3`

polimerazinius ir 3`→ 5` egzonukleazinius aktyvumus. Naudojant kinazes, prie DNR fragmentų 5`- galo prijungtas fosfatas. Toliau atliktas valymas naudojant *The Agencourt AMPure XP®* sistemą (gamintojas *Beckman Coulter Life Sciences*, Indiana, JAV). Ši sistema remiasi magnetinių dalelių panaudojimu ir atskiria DNR fragmentus, kurių ilgis yra 100 bp ir didesnis. Tuomet prie DNR fragmentų 3`- galų prijungti dATP, kad prie jų galėtų prisijungti adapteriai. Naudojant *5500 SOLiD™* fragmentų bibliotekos standartinį adapterių rinkinį (angl. *5500 SOLiD™ Fragment Library Standard Adaptor Kit*; gamintojas *Thermo Fisher Scientific*, Vilmingtonas, JAV), kuriame yra oligonukleotidiniai adapteriai (P1-T) ir PGR pradmenys fragmento bibliotekai sukurti, buvo prijungti adapteriai.

Toliau atliktas DNR fragmentų pagausinimas PGR reakcija, vykdyta hibridizacija su RNR zondais, žymėtais biotinu. DNR taikinio regionai (egzonai) atrinkti panaudojus magnetines streptavidino daleles. Tam naudoti du skirtingi rinkiniai: *TargetSeq™ Exome kit* (gamintojas *Thermo Fisher Scientific*, Vilmingtonas, JAV) bei *Agilent SureSelect Exome kit* (gamintojas *Agilent Technologies*, Kalifornija, JAV).

Kitame etape vykdytai sekoskaitai ligavimo būdu taikyta emulsinė PGR, kurios metu daugybėje mikroreaktorių prie atskirų dalelių buvo pagausinti DNR fragmentai. Išvalytos dalelės suleistos į sekoskaitos tėkmės lustą, kuris buvo dedamas į naujos kartos genetinį analizatorių *5500 SOLiD™ Sequencer* (gamintojas *Hitachi High-Technologies Corp*., Japonija), kuriame vykdyta sekoskaita.

2.3.1.3. NKS patogeninių genomo variantų patikrinimas *Sanger* sekoskaita

NKS metu gautų patogeninių genomo variantų specifiškumui ir jautrumui įvertinti buvo atlikta pakartotinė asmenų, turinčių patogeninių genomo variantų, sekoskaita. Tyrimui atlikti naudota *Sanger* sekoskaita.

Pasirinktiems sekos fragmentams, kuriuose yra patogeninių genomo variantų, buvo sukurti oligonukleotidiniai pradmenys *Primer3 v.4.0.0* programa. Genomo variantus apimančių fragmentų pagausinimas atliktas PGR, prieš tai optimizavus jos sąlygas.

Baigus PGR produktų valymą atlikta sekoskaitos PGR, kuri pagrįsta fermentinės terminacijos reakcijos principu.

Kapiliarinė elektroforezė vykdyta naudojant genetinį analizatorių *ABI PRISM 3130xl Genetic Analyzer*.

2.3.2. NKS duomenų analizė

Baigus *SOLiD* sekoskaitą, pirminiai duomenys buvo generuojami binukleotidiniu spalviniu kodu, kuris pateikiamas XSQ formatu. Pirminių duomenų analizė atlikta XSQ formatui pritaikytu analizės algoritmu, vėliau vykdytas duomenų anotavimas.

Prie NKS duomenų algoritmo ir jo optimizavimo prisidėję asmenys yra dr.Erinija Pranckevičienė, tyrėjas Aidas Pranculis.

2.3.2.1. Tyrimo duomenų NKS analizės algoritmas

NKS analizės algoritmui buvo panaudota *SOLiD* gamintojo pateikiama *LifeScope™ Genomic Analysis Software 2.5.1* programa. XSQ duomenų failai, kuriuose yra informacija apie genetiniu analizatoriumi sekvenuotus 75 bp ilgio fragmentus, *LifeScope™* programa buvo sulygiuojami su žmogaus referentiniu genomu – taip gauti BAM formato duomenų failai. Darbe buvo naudotas hg19 referentinis genomas.

Ta pati programa naudota atliekant teisingų variantų atranką siekiant gauti VCF formato failus. *LifeScope™* programa sukurti BAM duomenų failai

buvo sutvarkomi ir sujungiami naudojant *SAMtools v1.1* programą. Plačiau NKS analizės algoritmas aprašytas skyrelyje „1.4.1. NKS analizės algoritmas“.

Lyginant su dauguma kitų sekoskaitos sistemų, pavyzdžiui, *Illumina*, NKS analizės algoritmas yra geresnis naudojant trečiųjų šalių programas, tad darbe siekta rasti alternatyvius algoritmus naudojant trečiųjų šalių programas, kad būtų sukaupta kuo daugiau informacijos apie kiekvieną tiriamą asmenį ir kad būtų galima ją palyginti su rezultatais, gautais su *LifeScope*™ programa. Šiam tikslui buvo sudarytas sąrašas programų, tinkamų naudoti su *SOLiD*™ duomenimis, kad būtų galima sulygiuoti tiriamojo duomenis su referentiniu genomu. Iš jų pasirinktos ir naudotos šios sulygiavimo programos: *MAQ* (angl. *Mapping and analysis with qualities*), *SHRiMP* (angl. *Short read interpretation and mapping*) ir *BFAST* (angl. *Blast-like fast alignment of sequences tool*). Gavus šiuos BAM failus su trečiųjų šalių programomis, tolesnė analizė, variantų identifikavimas ir filtravimas vykdyti pasitelkus plačiausiai tam taikomą genomo analizės programinį paketą *GATK* (angl. *Genome Analysis Toolkit*).

NKS analizės algoritmo duomenų patikimumas įvertintas pagal tranzicijų/tranversijų santykį panaudojant *SAMtools* versiją ir analizuojant metaduomenis, gautus atlikus analizę *LifeScope*™ programa.

2.3.2.1.1. NKS analizės algoritmo kokybės įvertinimas

Siekiant įsitikinti sekoskaitos analizės duomenų patikimumu ir kad genomo variantai nebuvo atrinkti atsitiktinai, buvo pasinaudota sekoskaitos metu gautų nukleotidų pakaitų santykiu. Yra dvi nukleotidų pakaitų rūšys – tranzicija ir transversija. Tranzicijos metu purinas pakeičiamas purinu (A→G, G→A) arba pirimidinas pirimidinu (C→T, T→C). Transversijos metu purinas pakeičiamas pirimidinu (A→T, A→C, G→T, G→C) arba pirimidinas purinu (T→A, T→G, C→A, C→G).

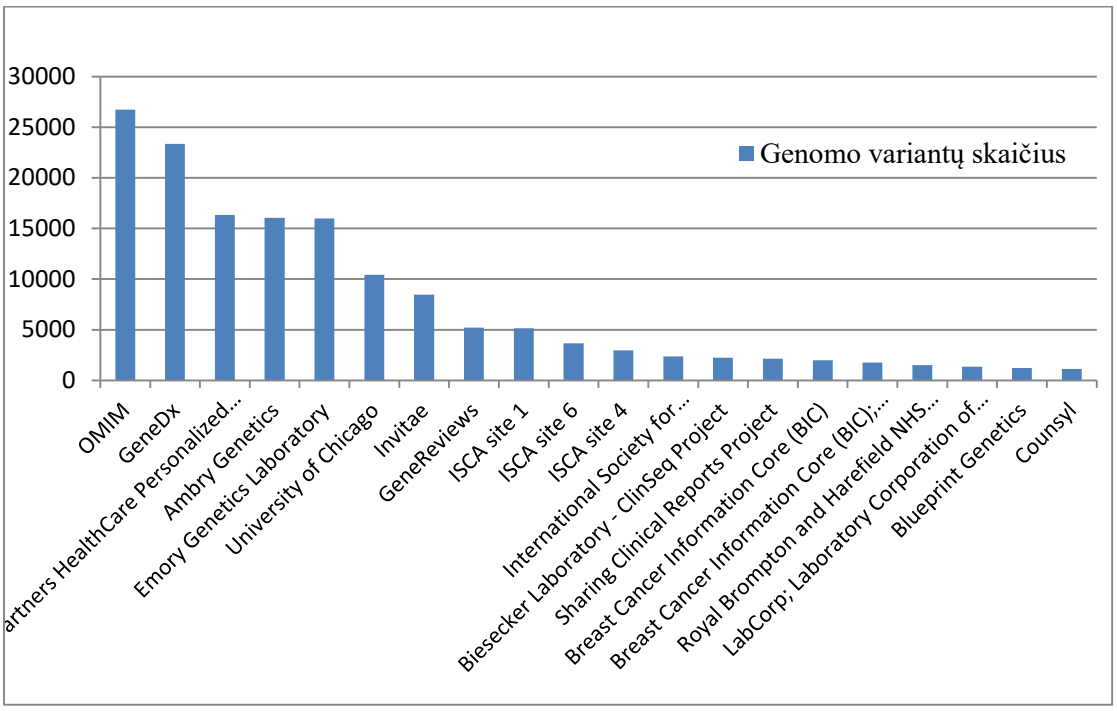
Tranzicijų ir transversijų santykis tuose pačiuose organizmuose yra maždaug vienodas, ir tai svarbus kriterijus, kuris naudojamas siekiant įsitikinti, kad gauti sekoskaitos duomenys nėra atsitiktiniai. Nuskaičius daugiau žmonių genomų, ypač 1000 genomų projekte, buvo nustatyta, kad žmogaus genomo tranzicijų/transversijų santykis yra 2,1. Jeigu tiriama vien koduojančioji genomo dalis, tas santykis yra didesnis ir siekia 2,8. Tad jei santykis yra itin mažas, tai rodo, kad gauti genomo variantai neatitinka žmogaus genomui būdingo pakaitų santykio, ir, greičiausiai, jie neteisingi ir gauti atsitiktinai.

Šiame darbe tranzicijų ir transversijų santykis nustatytas *Bedtools* programa ir iš *LifeScope* ™ programa gautų duomenų.

Itin svarbus aspektas NKS sekoskaitoje yra genomo variantų padengimas. *GATK* programos kūrėjai rekomenduoja, kad populiaciniuose tyrimuose jis turėtų būti apie 20 x 85 % sekoskaitos „taikinių“ ir apie 30x–50x klinikiniuose tyrimuose. Tokį padengimą taiko ir daugelis kitų autorių (Hedges et al., 2009; Dias et al., 2012; Karow, 2015). Disertaciniame darbe NKS sekoskaitos „taikinių“ padengimas nustatytas „Bedtools“ programa.

2.3.2.2. Tyrimo duomenų anotavimas

Darbo pradžioje buvo sudaromas žinomų patogeninių variantų, kurie yra paskelbti mokslinėse publikacijose, sąrašas. Vėliau tyrimo metu tokio sąrašo atsisakyta, nes yra vieša duomenų bazė *ClinVar*, kurioje galima rasti tūkstančius genomo variantų ir kuri pildoma pačių tyrėjų bei šaltinių, tokių kaip OMIM ar GeneDx, todėl daug tikslingiau buvo naudoti šios duomenų bazės informaciją anotavimui tiriamiems duomenims analizuoti, nei daryti savo atskirą sąrašą (5 pav.).



5 pav. Pagrindiniai *ClinVar* duomenų bazės šaltiniai pagal pateiktus genomo variantus, turinčius klinikinę interpretaciją.

Disertaciniame darbe genomo variantams egzomuose anotuoti naudota populiariausia pasaulyje nemokama anotavimo programa *Annovar v.2014nov12*. Taip pat darbe naudoti *Annovar* kūrėjo pateikiami anotaciniai įrankiai ir duomenų bazės, kurių formatas suderintas su *Annovar* programa.

Be pačios patogeninių variantų duomenų bazės, papildomai anotuoti 1000 genomų projekto ir ExAC duomenų bazės populiacijų dažniai, kad būtų galima palyginti lietuvių bendrosios populiacijos dažnius su kitų populiacijų dažniais.

Nestandartiškai tokio pobūdžio tyrimui buvo naudotas ir visas paketas algoritminių įrankių, kurie rodo galimą patogeniškumą. Tokie įrankiai naudojami ne populiaciniuose tyrimuose, bet ieškant naujų ligas lemiančių DNR sekos pokyčių, tačiau tie įrankiai buvo įtraukti ir į šį tyrimą, nes siekta įvertinti pačios naudojamos *ClinVar* duomenų bazės duomenis. Naudotas anotacinių įrankių paketas v2.6 (skirtas nesinoniminiams genomo variantams analizuoti), sudarytas iš: *SIFT v5*, *PolyPhen v2.2.2*, *MutationTaster v2*, *MutationAssessor*

v2, *LRT*, *FATHMM v2.3*, *GERP++ v2*, *SiPhy v0.5*, *PhyloP v1* algoritminių anotacinių įrankių.

Taip pat prijungtos duomenų bazės, kurios suteikia informaciją apie genomo varianto geną, ir *dbSNP* duomenų bazė, kad kiekvienam genomo variantui būtų turimas „rs“ kodas. Pastarasis yra kaip brūkšninis kodas ir labai palengvina paiešką, kai konkretus genomo variantas analizuojamas detaliau.

2.3.2.3. Galutinė duomenų analizė ir statistiniai metodai

Anotuoti genomo variantai, kurie skyrėsi nuo žmogaus referentinio genomo, galiausiai buvo analizuojami „Microsoft Office 2010“ „Excel“ programa, kuri turi patogią informacijos, pateiktos stulpeliais, atrankos funkciją.

Integruota genomo vaizduoklė (angl. *Integrative Genomics Viewer – IGV*) naudota patikrinti dominančius genomo variantus BAM failuose.

Atlikus dažnių palyginimus, jų statistinis patikimumas įvertintas naudojant Pirsono $\chi 2$ bei tikslųjį Fišerio kriterijus. Skaičiavimai atlikti naudojant „Microsoft Office 2010“ „Excel“ programos formules. Statistiškai reikšmingais buvo laikomi skirtumai, kurių $P < 0,05$.

2.3.2.4. Darbe naudotos programos ir duomenų bazės

Naudotos programos:

1. *SOLiD* sekoskaitos duomenų analizės algoritmo programa *LifeScope*™ https://www.thermofisher.com/lt/en/home/life-science/sequencing/next-generationsequencing/SOLiD-next-generation-sequencing/SOLiD-next-generation-sequencing-data-analysis-solutions/lifescope-data-analysis-SOLiD-next-generation-sequencing.html
2. Prilygiavimo prie referentinio genomo programa *MAQ* http://*MAQ*.sourceforge.net/
3. Trumpų DNR fragmentų prilygiavimo prie referentinio genomo programa *SHRiMP* (angl. *SHort Read Mapping Package*) http://compbio.cs.toronto.edu/shrimp/
4. Greitai trumpus fragmentus prie referentinio genomo prilygiuojanti programa *BFAST* (angl. *Blat-like Fast Accurate Search Tool*) https://sourceforge.net/projects/BFAST/
5. BAM failų tvarkymo programa *SAMtools* http://samtools.sourceforge.net/
6. NKS duomenų analizės programa http://bedtools.readthedocs.io/en/latest/
7. BAM failų tvarkymo programa *Picard* http://broadinstitute.github.io/picard/
8. Genomo analizės paketas (angl. *The Genome Analysis Toolkit*) http://www.broadinstitute.org/*GATK*/
9. Anotavimo programa *Annovar* http://*Annovar*.openbioinformatics.org/en/latest/
10. Integruota genomo vaizduoklė https://www.broadinstitute.org/igv/

Duomenų bazės:

1. 1000 genomų projekto duomenų bazė http://www.1000genomes.org
2. Egzomų konsorciumo duomenų bazė ExAC http://exac.broadinstitute.org/
3. Patogeninių genomo variantų duomenų bazė *ClinVar* http://www.ncbi.nlm.nih.gov/clinvar/
4. OMIM® (angl. *Online Mendelian Inheritance in Man®)* http://www.omim.org/
5. Genomo vieno nukleotido polimorfizmų duomenų bazė dbSNP http://www.ncbi.nlm.nih.gov/projects/SNP/
6. *PolyPhen*-2 (angl. *Polymorphism Phenotyping v2*) http://genetics.bwh.harvard.edu/pph2/
7. *SIFT* nurodo, ar aminorūgšties pakaita keičia baltymo funkciją http://sift.jcvi.org/

3. REZULTATAI

3.1. Tinkamiausio analizės algoritmo parinkimas

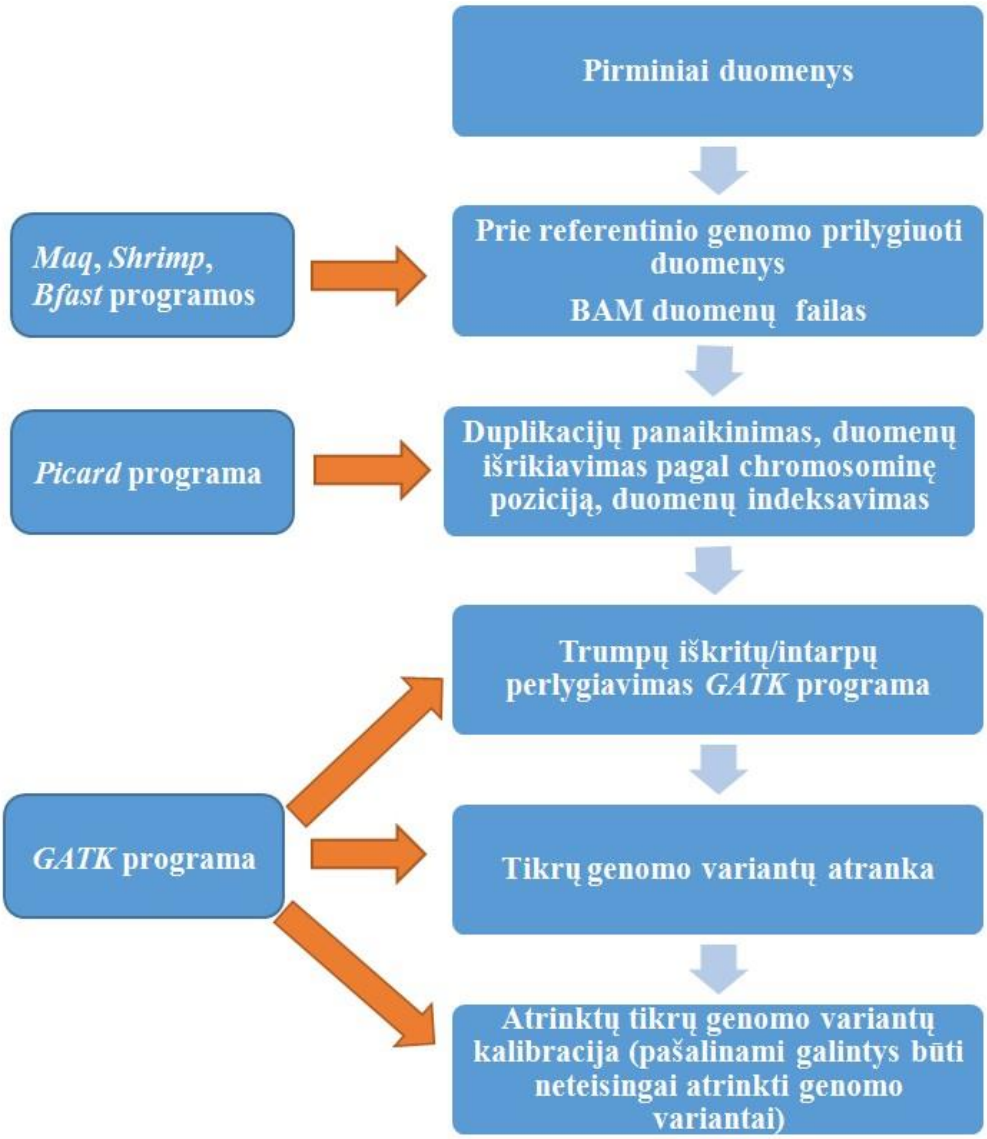
Analizuojant sekoskaitos duomenis, gautus ligavimo būdu naudojant *SOLiD* sistemą, pirminiai duomenys yra generuojami binukleotidiniu spalviniu kodu. Tokie duomenys nėra gerai pritaikyti dirbti su populiariausiomis sekoskaitos analizės programomis ir, priešingai nei *Illumina* sekoskaita, *SOLiD* neturi standartizuoto trečiųjų šalių programų analizės standarto.

Disertacijoje buvo naudota *Lifescope* programa, kurią pateikia *SOLiD* sistemos gamintojai (naudoti parametrai pateikiami 2 priede), tačiau, siekiant įsitikinti rezultatų patikimumu ir taip pat padidinti gaunamų rezultatų efektyvumą, tiems patiems duomenims pritaikytas ir alternatyvusis analizės algoritmas.

Lifescope programa atlieka visus analizės algoritmo etapus, o taikant alternatyviąsias analizės algoritmo programas, analizę sudaro du pagrindiniai etapai: sulygiavimas su referentiniu genomu ir teisingų variantų atranka.

Nestandartinis XSQ pirminių duomenų formatas įpareigojo analizės algoritmui atrinkti programas, kurios sugebėtų tinkamai analizuoti pirminius *SOLiD* sistemos duomenis. Darbui atrinktos *MAQ* (angl. *Mapping and analysis with qualities*), *SHRiMP* (angl. *Short read interpretation and mapping*) ir *BFAST* (angl. *Blast-like fast alignment of sequences tool*) programos (6 pav.).

Atlikus visų alternatyvių sulygiavimo programų ir gamintojų pateiktos *Lifescope* programos padengimo įvertinimą *Bedtools* programa, nustatyta, kad su visomis naudotomis programomis buvo padengta apie 97 % regionų, kurie buvo sekoskaitos taikiniai. Skyrėsi pats padengimo lygis – naudojant *Lifescope* ir *SHRiMP* programas, padengimas buvo gerokai didesnis nei su *BFAST* ir *MAQ* programomis. Didesnis DNR fragmentų padengimas *Lifescope* programa greičiausiai gautas todėl, kad *Lifescope* programa turi papildomą etapą, skirtą binukleotidiniam spalviniam kodui apdoroti.



6 pav. Naudotas alternatyvusis *Lifescope* programai NKS analizės algoritmas.

Naudojant *SHRiMP* programą gautuose duomenyse moterų egzome nustatytas gerokai didesnis DNR fragmentų, priskiriamų Y chromosomai, skaičius, palyginti su kitomis sulygiavimo programomis. Ir nors yra pseudoautosominiai regionai ir jų sekų homologija, gerokai didesnis Y chromosomai priskirtų DNR fragmentų skaičius moteryse gali reikšti netikslų prilygiavimą šia programa (Pranckeviciene et al., 2015).

Vykdytame populiaciniame genomo variantų tyrime padengimas buvo apie 20 kartų 80 % sekoskaitos „taikinių“, tai nedaug skiriasi nuo kitų autorių

rekomenduojamo padengimo (20 kartų padengimas 85 % sekoskaitos taikinių). Tokia rekomendacija labiau taikytina atliekant sekoskaitą *Illumina*, o sekoskaitą atliekant *SOLiD*, kuri yra tikslesnė, atitinkamai ir padengimas galėtų būti mažesnis.

Analizuotuose duomenyse tranzicijų/transversijų santykis buvo apie 2,2–2,8. Toks santykis rodo, kad gauti duomenys generuoti neatsitiktinai.

Teisingų variantų atranka vykdyta dažniausiai tam skirtu genomo analizės programiniu paketu *GATK* (angl. *Genome Analysis Toolkit*), kuris taikytas visų prilygiavimo programų rezultatams. Šiuo programiniu paketu taip pat analizuotas BAM formato duomenų failas, gautas *Lifescope* programa (žr. 3 priede – *GATK* analizės algoritmas).

Po teisingų variantų atrankos įvertinimo etapo gautas esminis *SOLiD* sistemai skirtos programos *Lifescope* ir trečiųjų šalių programų palyginimo rezultatas rodo, kad įvairių programų galutiniai rezultatai palyginti mažai skyrėsi tarpusavyje. Nepasitaikė, kad su viena iš naudotų programų būtų gaunami genomo variantai, kurie skirtųsi nuo referentinio genomo ir nuo kitomis programomis gaunamų rezultatų. Toks rezultatų atsikartojimas skirtingais algoritmais rodo, kad duomenų analizė atlikta tinkamai.

Patogeninių genomo variantų duomenų bazė *ClinVar* ir su vėžiu susijusių genomo variantų duomenų bazė *COSMIC* buvo pasitelktos siekiant išsiaiškinti, su kuria programa yra nustatomas didžiausias genomo variantų skaičius. Rezultatas ir šiuo atveju buvo geriausias su *Lifescope* programa. Tai lemia esminė priežastis – didesnis genomo variantų padengimas. Su kitomis programomis, esant itin mažam kai kurių genomo variantų padengimui, jie nebuvo laikomi tikrais genomo variantais dėl *GATK* programos atrankos parametrų.

Lyginant teisingų variantų atranką, atliktą *Lifescope* ir *GATK* programomis, abiem atvejais gauti geri rezultatai derinant du parametrus pagal duomenų specifiškumą (siekiant panaikinti kuo daugiau galimų artefaktų, tačiau

taip rizikuojant prarasti teisingus genomo variantus) ir pagal jautrumą (siekiant rasti kuo daugiau teisingų variantų, tačiau padidinant neteisingai atrinktų genomo variantų skaičių).

Naudojant *GATK* programą, yra ypač geras specifiškumo bei jautrumo balansas, tačiau *GATK* griežčiau atrinko variantus pagal padengimo kriterijų, tad *GATK* programą geriausia naudoti apdoroti duomenims, kai DNR fragmentų padengimas yra labai didelis.

Lifescope programa gerai atrenka VNV, tačiau gaunami blogesni rezultatai – yra trumpų iškritų ir intarpų, kurių specifiškumas gana mažas. Detaliau analizuojant (naudojant vizualizavimo programą) *Lifescope* programos atrinktas trumpas iškritas/intarpus, neretai susiduriama su realių genomo variantų nebuvimo problema.

Apibendrinant rezultatus, gautus palyginus alternatyvias analizės algoritmo programas su sekoskaitos gamintojo pateikta *Lifescope* programa, *Lifescope* programa gauti genomo variantai turėjo gerą specifiškumą ir jautrumą, o DNR fragmentai turėjo geriausią padengimą, todėl *Lifescope* programa buvo naudojama tolesniuose darbo etapuose, analizuojant visų 96 asmenų egzomus.

3.2. NKS duomenų patikrinimo *Sanger* sekoskaita rezultatai

Siekiant įvertinti NKS duomenų jautrumą ir specifiškumą, gauti tų pačių asmenų duomenys palyginti skirtingu metodu, naudojant *Sanger* sekoskaitą.

Darbui buvo pasirinkti 30 NKS metu patogeniniais įvardytų genomo variantų, kurie skyrėsi įvairiais NKS sekoskaitos parametrais ir DNR fragmentų padengimu (3 lentelė).

3 lentelė. *Sanger* sekoskaita patikrintų patogeninių genomo variantų *dbSNP* kodas ir su šiais genomo variantais susijusios ligos bei sutrikimai

Nr.	*dbSNP* kodas	Būklė	OMIM kodas	TLK-10-AM kodas
1	rs80338800	IIA tipo galūnių juostos raumenų distrofija	253600	G71.0
2	rs3214759	Įvairaus tipo katarakta	615188	Q12.0
3	rs1800553	Štargarto geltonosios dėmės distrofija	248200	H35.5
4	rs34526199	Adenozino monofosfato deaminazės stoka	615511	G71.3
5	rs76308115	Pirminė abipusė mazginė antinksčių hiperplazija	610475	E24.8
6	rs116100695	Anemija dėl piruvato kinazės stokos	266200	D55.2
7	rs7732671	Nutukimas		E66
8	rs726070	Įgimta ichtiozė (Arlekino vaisius)	242500	Q80.4
9	rs1130335	Pakitęs placentos šarminės fosfatazės aktyvumas	171800	R74.8
10	rs116474260	Paveldimoji hiperekpleksija	149400	G25.8
11	rs1142345	Tiopurino metiltransferazės stoka	610460	E79.8
12	rs1800460	Tiopurino metiltransferazės stoka	610460	E79.8
13	rs4151667	Senatvinė geltonosios dėmės degeneracija	615489	H35.3
14	rs16879498	Rh- hemolizinė anemija		D58.9
15	rs73715573	Cistinė fibrozė	219700	E84
16	rs45562031	Paveldimoji sferocitozė	612653	D58.0
17	rs35312232	Lamelarinė ichtiozė	242300	Q80.2
18	rs77724903	Šerdinė skydliaukės karcinoma	155240	C73
19	rs7270101	Inozino trifosfatazės stoka	613850	
20	rs139382018	II tipo paveldimoji motorinė neuropatija	613376	G12.2
21	rs111033566	Paveldimasis lėtinis pankreatitas	167800	K86.1
22	rs121918164	Fankoni anemija	609053	D61.0
23	rs121434369	I tipo gliutaro acidurija	231670	E72.3
24	rs1800450	Manozę sujungiančio baltymo stoka	614372	
25	rs193922688	Hipofizės hormonų stoka	262600	E23.0
26	rs13078881	Biotinidazės stoka	253260	E53.8
27	rs268	Šeiminė lipoproteinlipazės stoka	144250	E78.3
28	rs41297018	Sunkus kombinuotas imunodeficitas dėl DCLRE1C stokos	602450	D81.1
29	rs61747071	Jouberto sindromas	213300	Q04.3
30	rs3848519	Paveldimoji eritropoetinė porfirija	177000	E80.0

Toks NKS duomenų patikrinimas aprašomas ir kitų autorių darbuose, pavyzdžiui, *Linnea M. Baudhuin* ir kt. atlikta analizė parodė, kad visi 919 tikrinti NKS metu identifikuoti variantai buvo nustatyti ir atlikus *Sanger* sekoskaitą (Baudhuin et al., 2015). Tačiau toks 100 % rezultatas buvo gautas naudojant NKS duomenis, kai DNR fragmentų padengimas buvo per 100, o tai yra gerokai daugiau, palyginti su šio darbo duomenimis. Daug artimesnis DNR fragmentų padengimas buvo gautas *Samuel P. Strom* ir kt. darbe, kur padengimas siekė 5 kartus. Gautas taip pat puikus rezultatas – 99,1 % variantų buvo patvirtinti (Strom et al., 2014).

Atlikus *Sanger* sekoskaitą, gautas geras *Sanger* ir NKS genomo variantų nustatymo sutapimas, siekiantis 99,3 %. NKS metodo jautrumas yra 100 %, o specifiškumas – 88,24 %. Iš tikrintų patogeninių variantų NKS analizė neteisingai identifikavo tik vieną genomo variantą, kuris atsiranda dėl genomo sekų homologijos.

3.3. Gautų genomo variantų aprašomoji statistika

Kiekvieno atskiro asmens egzome vidutiniškai buvo apie 42 tūkstančiai unikalių vieno nukleotido pakaitos pokyčių ir apie 2 300 unikalių trumpų iškritų/intarpų, kurie skyrėsi nuo referentinio genomo.

Tačiau, gavus atskirų asmenų galutinius NKS algoritmo rezultatus ir jeigu analizė būtų atliekama kiekvienam asmeniui atskirai, tolesni tyrimai būtų gerokai sudėtingesni. Tad iškeltas uždavinys juos analizuoti kaip vieną bendrą imtį, panaudojant *GATK* programinio paketo programą *CombineVariants*, kuri sujungė visų tiriamų asmenų duomenis į bendrą imtį, tačiau leido matyti ir kiekvienam atskiram asmeniui priklausančius genomo variantus.

Galutiniame jungtiniame visų asmenų egzomo tyrime iš viso buvo nustatyta 243 tūkstančiai unikalių vieno nukleotido pakaitų ir apie 32 tūkstančiai unikalių trumpų iškritų/intarpų.

3.4. Gautų patogeninių variantų duomenų analizė

Kiekvienas asmuo vidutiniškai turėjo 45 vieno nukleotido genomo variantus, kurie patogeninių genomo variantų duomenų bazėje įvardyti kaip patogeniniai. Patogeniniais įvardytų iškritų/intarpų kiekvienas asmuo turėjo nuo nulio iki trijų, vidutiniškai apie vieną.

321 skirtingi VNV iš visų tirtų asmenų duomenų, buvo įvardyti kaip patogeniniai. Iš viso rasta 30 skirtingų iškritų/intarpų, kurie laikomi patogeniniais.

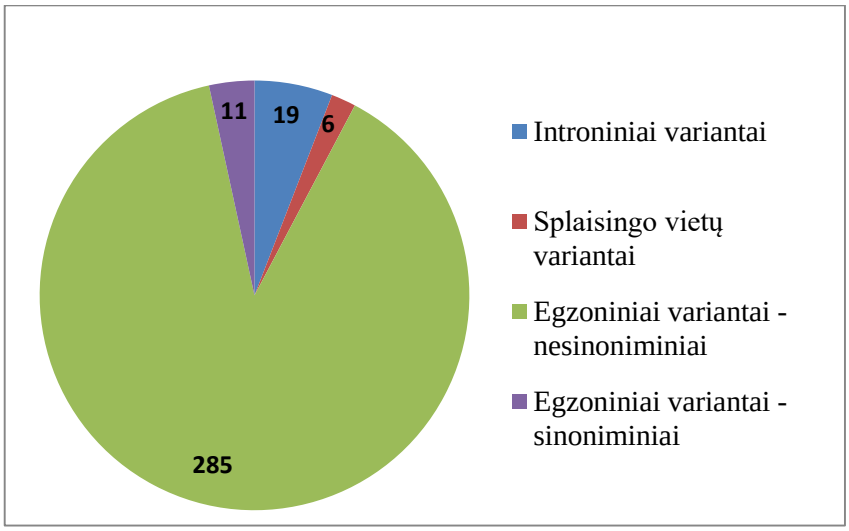
Tarp 321 VNV identifikuotų patogeninių genomo variantų 317-ai yra suteiktas dbSNP duomenų bazės „rs“ kodas, 4 jo neturi. Šiuo aspektu tarp identifikuotų patogeninių VNV ir trumpų iškritų/intarpų yra didelis skirtumas, nes iš 30 patogeninių trumpų iškritų/intarpų 17-ai yra suteiktas dbSNP duomenų bazės „rs“ kodas, o net 13 – jo neturi.

Dauguma identifikuotų variantų yra egzoninėje geno dalyje. Tokių rezultatų buvo galima tikėtis, nes pakaitos egzoninėje dalyje dažniausiai turi gerokai didesnę reikšmę organizmo sveikatai. Taip pat reikėtų pabrėžti, kad buvo atlikta viso egzomo sekoskaita, tad introninė dalis nebuvo visiškai sekvenuota, ir realus patogeninių variantų skaičius, atlikus viso genomo sekoskaitą, turėtų būti didesnis.

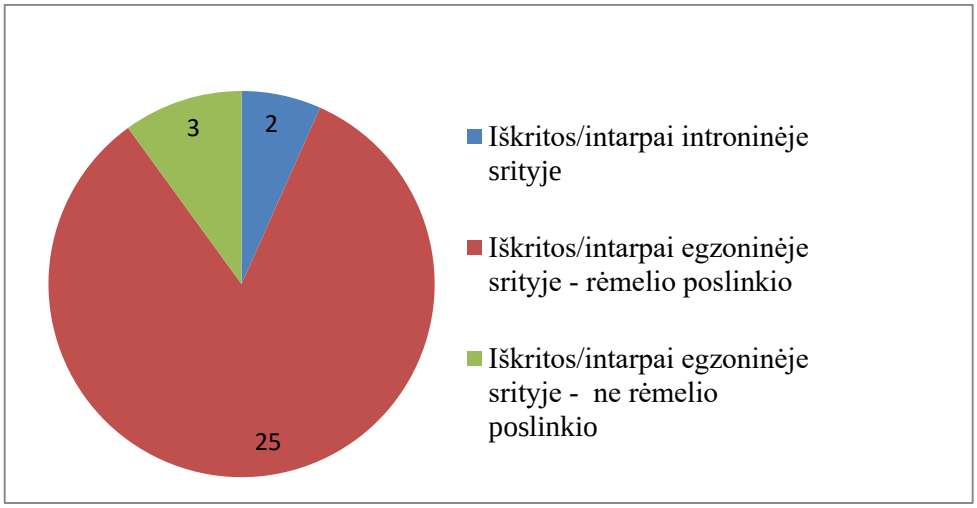
Dauguma VNV patogeninių variantų, esančių egzoninėje dalyje, yra nesinoniminės pakaitos, keičiančios aminorūgštį ar terminuojančios baltymo sintezę. Tarp trumpų iškritų/intarpų dominuoja rėmelio poslinkio pakaitos. Tokie duomenys yra labai svarbūs, nes nustatyti pakaitų tipai keičia geno koduojamą produktą, ir tai gali turėti didelę įtaką organizmo sveikatai.

Nepaisant didelės nesinoniminių pakaitų gausos, pasitaikė ir patogeniniams variantams priskirtų pakaitų, kurios yra sinoniminės ar ne rėmelio poslinkio ir nekeičia geno koduojamo produkto. Taip pat yra pakaitų, kurios yra introninėse genų srityse ar splaisingo vietose.

Visų identifikuotų patogeninių variantų (VNV ir trumpų iškritų/intarpų) dažniai pavaizduoti 7 ir 8 pav.



7 pav. VNV variantų iš bendrosios lietuvių populiacijos dažniai



8 pav. Patogeninių trumpų iškritų/intarpų bendrojoje lietuvių populiacijoje dažniai

Prieš atliekant galutinę patogeninių variantų analizę, buvo sumažintas jų skaičius. Pagrindinė tokio mažinimo priežastis – didelis įvardytų patogeninių variantų dažnis tiek 1000G, ExAC projektų duomenyse, tiek tarp tirtų lietuvių. Nors pats žodis „patogeninis“ nereiškia, kad genomo variantas tiesiogiai sukels ligą, o tik gali padidinti riziką susirgti ar lemti sunkesnius ligos simptomus,

tačiau jei patogeninio varianto dažnis bendrojoje populiacijoje didesnis nei 50 %, – laikyti jį patogeniniu vargu ar reikėtų. Plačiau apie tai rašoma 4.2. skyrelyje „Naudotos patogeninių genomo variantų duomenų bazės įvertinimas“.

Dėl per didelio dažnio iš 321 VNV tolesnei analizei buvo paliktas 301 genomo variantas, šių variantų dažnis mažesnis nei 50 %, taip pat atskirti genomo variantai, kurių dažnis bendrojoje populiacijoje yra mažesnis nei 1 %. Genomo variantai, kurių dažnis bendrojoje populiacijoje yra 1 % ir mažesnis, yra laikytini retais variantais. Tokie variantai gali būti tiesiogiai lemiantys retas ligas. Tyrime nustatyti 124 patogeniniai genomo variantai, kurių dažnis yra 1 % ir mažesnis. Patogeninių VNV ir trumpų iškritų/intarpų dažniai, atsižvelgiant į dažnį kitose populiacijose pateikti 4 lentelėje.

Iš viso buvo identifikuoti 52 patogeniniais įvardyti genomo variantai, kurių nebuvo 1000G projekto duomenyse, 9-ių genomo variantų nebuvo ExAC duomenų bazėje, 6-ių genomo variantų nebuvo nei 1000G projekto duomenyse, nei ExAC duomenų bazėje.

Tarp genomo variantų, kurių dažnis yra mažesnis nei 50 % ExAC ir 1000G duomenų bazėje, buvo 234 patogeniniais įvardyti genomo variantai, kurių genotipas buvo vien heterozigotinis, ir nebuvo nė vieno homozigotinio genotipo pagal patogeninį alelį. Iš 96 tirtų asmenų atskirus genomo variantus turėjo nuo 1 iki 38 asmenų.

Tarp genomo variantų, kurių dažnis yra mažesnis nei 1 %, ExAC ir 1000G duomenų bazėje buvo 134 patogeniniais įvardyti genomo variantai, kurių genotipas buvo vien heterozigotinis, ir nebuvo nė vieno homozigotinio genotipo pagal patogeninį alelį. Iš 96 tirtų asmenų atskirus genomo variantus turėjo nuo 1 iki 13 asmenų.

4 lentelė. Patogeninių variantų bendrojoje lietuvių populiacijoje statistika atsižvelgiant į jų dažnumą

Patogeniniai variantai bendrojoje lietuvių populiacijoje	VNV	Iškritos / intarpai
Variantai, kurių dažnis 1000 genomų projekte ir ExAC duomenų bazėje yra mažesnis nei 50 %	301	30
Introniniai variantai	13	2
Splaisingo vietų variantai	6	–
Egzoniniai variantai	282	28
Nesinoniminiai, rėmelio poslinkio	268	25
Sinoniminiai, kurie nėra rėmelio poslinkio	14	3
Variantai, kurių dažnis 1000 genomų projekte ir ExAC duomenų bazėje yra mažesnis nei 1 %	150	24
Introniniai variantai	2	0
Splaisingo vietų variantai	6	–
Egzoniniai variantai	141	24
Nesinoniminiai, rėmelio poslinkio	138	23
Sinoniminiai, kurie nėra rėmelio poslinkio	3	1

3.5. Vidupopuliacinis patogeninių genomo variantų palyginimas

Identifikuotas 321 patogeninis genomo variantas tarp tiriamų etnolingvistinių grupių pasiskirstė maždaug vienodai. Daugiausiai, 168 iš 321 galimų patogeninių variantų, buvo Šiaurės Žemaitijoje, mažiausiai – 148 iš 321 galimų patogeninių variantų, buvo identifikuoti Rytų Aukštaitijos grupėje. Kiekvienoje etnolingvistinėje grupėje buvo vidutiniškai po 150 patogeninių genomo variantų. Kiekvieno atskiro genomo patogeninių variantų skaičius kiekvienoje etnolingvistinėje grupėje svyravo nuo 0 iki maksimalaus 16. Toks didelis asmenų, turinčių atskirą patogeninį variantą, skaičius susidarė dėl trijų patogeniniais įvardijamų genomo variantų, kurių turėjo visi 96 asmenys.

Vykdant atranką pagal dažnumą bendrojoje pasaulio populiacijoje (remiantis 1000G ir ExAC duomenimis), patogeninių genomo variantų pasiskirstymas tarp etnolingvistinių grupių yra labai panašus. Detalus patogeninių genomo variantų dažnis kiekvienoje etnolingvistinėje grupėje, atsižvelgiant į dažnumą bendrojoje populiacijoje, pateiktas 5 lentelėje.

5 lentelė. Patogeninių genomo variantų skaičius kiekvienoje etnolingvistinėje grupėje, patogeninius variantus atrenkant pagal skirtingą dažnumą, gautą iš 1000G, ExAC duomenų

PA	**RA**	**VA**	**PŽ**	**ŠŽ**	**VŽ**	**Iš viso genomo variantų**
Bendras identifikuotų patogeninių genomo variantų skaičius						
158 (49,22 %)*	**148** (46,11 %)	**158** (49,22 %)	**158** (49,22 %)	**168** (52,34 %)	**155** (48,29 %)	**321**
Kai genomo varianto dažnis1000G, ExAC duomenyse yra mažesnis nei 50 %						
140 (46,51 %)	**131** (43,52 %)	**141** (46,84%)	**141** (49,22 %)	**150** (49,83 %)	**138** (45,85 %)	**301**
Kai genomo varianto dažnis 1000G, ExAC duomenyse yra mažesnis nei 25 %						
119 (42,5 %)	**111** (39,64 %)	**122** (43,57 %)	**121** (46,84 %)	**130** (46,43 %)	**118** (42,14 %)	**280**
Kai genomo varianto dažnis 1000G, ExAC duomenyse yra mažesnis nei 1 %						
33 (22 %)	**28** (18,67 %)	**38** (25,33 %)	**38** (25,33 %)	**47** (31,33 %)	**34** (22, 67 %)	**150**

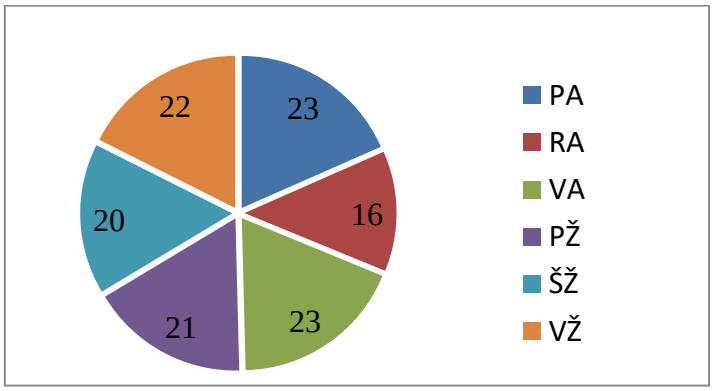
*Skliausteliuose nurodytas santykinis dažnis.

Analizuojant tik retus genomo variantus, kurių dažnis bendrojoje populiacijoje yra 1 % ir mažesnis, toks rezultatas atsispindi ir tyrimų duomenyse, kur kiekvieno atskiro varianto skaičius kiekvienoje etnolingvistinėje grupėje svyruoja nuo 0 iki 6 iš 16 galimų.

Tarp moterų identifikuotų patogeninių variantų pasitaikė šiek tiek daugiau negu tarp vyrų. Santykinis visų patogeninių genomo variantų dažnis moterų yra apie 80 %, o vyrų – apie 60 %.

Identifikuoti du patogeniniai genomo variantai, kurie buvo tik Žemaitijos grupėse, ir jų nebuvo aptikta nė vienoje aukštaičių grupėje. Išskirtinai aukštaičiams buvo būdingas vienas patogeninis genomo variantas.

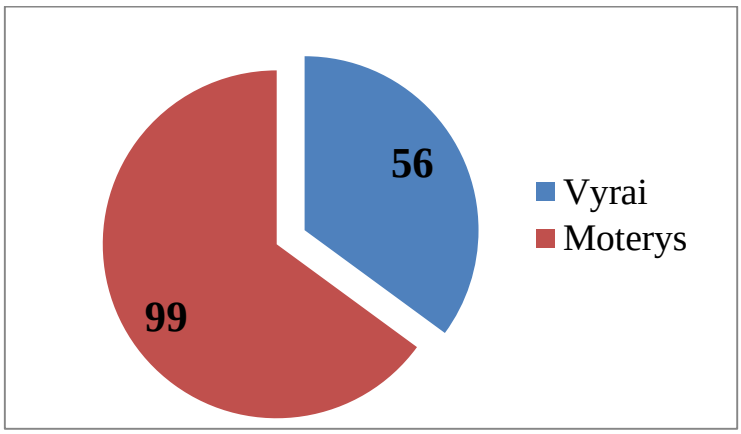
Kiekvienos etnolingvistinės grupės tiriami asmenys turėjo apie 20 patogeninių variantų, kurių buvo tik vienoje konkrečioje grupėje. Unikalių (būdingų tik konkrečiai etnolingvistinei grupei) etnolingvistinėms grupėms patogeninių genomo variantų skaičius skirtingose grupėse yra apylygis ir varijuoja nuo 16 iki 23 (9 pav.).



9 pav. Unikalių, tik tam tikroje etnolingvistinėje grupėje identifikuotų patogeninių genomo variantų dažnis

Dauguma unikalių etnolingvistinėms grupėms patogeninių genomo variantų yra pavieniai, turimi vieno ar dviejų asmenų, ir tai yra natūralu patogeniniams genomo variantams, kurie populiacijoje yra reti. Tačiau yra ir patogeninių variantų, kurių dažnis didesnis, o tai rodo, kad tie patogeniniai variantai labai paplitę konkrečioje etnolingvistinėje grupėje. Šiuo atžvilgiu galima išskirti Rytų Aukštaitiją ir Šiaurės Žemaitiją. Plačiau apie tai rezultatų aptarime 4.5. „Patogeninių recesyviųjų genomo variantų įvertinimas etnolingvistinių grupių atžvilgiu“.

Unikalių patogeninių variantų skaičius lyties atžvilgiu gerokai skiriasi. Identifikuoti 99 moterų turimi patogeniniai genomo variantai, kurie nebuvo būdingi vyrams. Būdingų tik vyrams patogeninių variantų skaičius tesiekė 56 (10 pav.).



10 pav. Unikalių, tik tam tikrai lyčiai būdingų, patogeninių genomo variantų dažnis

3.6. Tarppopuliacinis patogeninių genomo variantų palyginimas

Siekiant išryškinti lietuvių populiacijos išskirtinumą patogeninių genomo variantų atžvilgiu, palyginti su kitomis populiacijomis, pirmiausiai lietuviams buvo įvestas apribojimas pagal patogeninių alelių skaičių.

Kai kurie patogeniniai genomo variantai yra itin reti, ir nors ExAC duomenų bazėje bendra alelių imtis yra per 120 tūkstančių, patogeninių alelių gali būti tik keli. Vykdytame tyrime buvo ištirti 96 asmenys, tad vieno patogeninio genomo varianto maksimali tirtų alelių suma sudarė 192, o tai yra gerokai mažiau nei dideliuose ExAC duomenyse.

Dėl patogeninių alelių retumo, t. y. dėl per mažo jų skaičiaus, atsiranda problema atlikti dažnių palyginimą. Ir nors pavieniai patogeniniai variantai tiriamoje imtyje gali būti ir labai reikšmingi, jie netinka analizuojant dažnius, nes gali būti atsitiktiniai. Tad, siekiant analizuojamus duomenis su pasauliniais duomenimis (iš ExAC ir 1000G) palyginti teisingai, lietuviams taikytas kriterijus, kad tą patį patogeninį variantą turėtų mažiausiai trys skirtingi asmenys ir į analizę įtraukiamų alelių skaičius būtų trys ir didesnis, taip pat naudoti Pirsono $\chi 2$ bei tikslusis Fišerio kriterijai norint įsitikinti, kad skirtumas tarp lyginamų grupių yra statistiškai reikšmingas.

6 lentelė. Patogeninių ir bendras alelių skaičius lietuvių bendrojoje populiacijoje bei alelių skaičius pagal 1000 genomų projekto ir ExAC bazės duomenis

Genomo variantas. dbSNP 142 kodas	ExAC PA	ExAC VA	1000G PA	1000G VA	1000G EPA	1000G EVA	L PA	L VA
rs1800553	613	121 302	16	4 992	4	1 002	5	192
rs76308115	359	118 214	9	4 999	5	1 001	4	192
rs115532916	2 213	121 370	39	4 969	22	984	14	192
rs142181517	421	121 046	4	5 004	4	1 002	7	192
rs587780273	0	117 448	0	0	0	0	6	192
rs121917710	693	121 412	13	4 995	7	999	7	192
rs41297018	1 342	120 654	31	4 977	22	984	10	192
rs77724903	216	119 802	1	5 007	1	1 005	4	192
rs41298135	413	116 632	7	5 001	4	1 002	4	192
rs201968272	551	119 750	0	0	0	0	4	192
rs45495503	135	118 688	2	5 006	2	1 004	4	192
rs113298164	335	121 400	6	5 002	5	1 001	4	192
rs41549716	762	121 386	11	4 997	7	999	6	192
rs104895094	668	121 410	9	4 999	7	999	13	192
rs28935490	275	87 762	8	3 767	4	762	7	192
rs34526199	3 425	121 378	55	4 953	37	969	14	192
rs1736557	9 769	121 328	489	4 519	63	943	5	192
rs11887534	1 105	19 676	302	4 705	80	926	3	192
rs28757581	17 421	120 442	961	4 047	120	886	13	192
rs3749977	35 483	119 374	1861	3 147	230	776	31	192
rs45547231	13 149	115 732	383	4 625	151	855	8	192
rs2229738	7855	121 408	119	4 889	87	919	28	192

6 lentelės tęsinys

Genomo variantas. dbSNP 142 kodas	ExAC PA	ExAC VA	1000G PA	1000G VA	1000G EPA	1000G EVA	L PA	L VA
rs11539445	15 214	120 880	634	4 374	163	843	11	192
rs2269219	29 792	120 766	1676	3 332	202	804	29	192
rs12483377	5 540	49 258	222	4 786	84	922	7	192

***ExAC PA** – ExAC patogeniniai aleliai; **ExAC VA – ExAC –** visi aleliai; **1000G PA** – iš visų 1000G duomenų patogeniniai aleliai, **1000G VA** – iš visų 1000G duomenų visi aleliai, **1000G EPA** – iš 1000G europiečių duomenų patogeniniai aleliai, **1000G EVA** – iš 1000G europiečių duomenų visi aleliai, **L PA** – Lietuvos patogeniniai aleliai, **L VA** – Lietuvos visi aleliai.

Anotacija dažnį pateikia procentine išraiška, o norint gauti rezultatus pagal Pirsono $\chi 2$ bei tikslųjį Fišerio kriterijus, skaičiavimams reikia naudoti pačių alelių dažnį. Tad iš ExAC ar 1000 genomų projekto duomenų bazių atskirai buvo surinkta informacija apie visų tirtų patogeninių alelių bei konkretaus patogeninio varianto visų alelių skaičius. 6 lentelėje pateikiami pagrindinių atrinktų patogeninių variantų alelių skaičiai.

Palyginus tyrimo duomenis su ExAC pasauliniais duomenimis, gauti net 103 statistiškai patikimai besiskiriantys patogeniniai variantai. Toks didžiulis skaičius nestebina, nes kitų autorių darbuose ir pačiose ExAC ar 1000 genomų projekto duomenų bazėse pasitaiko gana didelis dažnių skirtumas tarp didesnių subpopuliacijų, tokių kaip Europos, Azijos ar Afrikos.

Tad, siekiant atlikti tikslesnį palyginimą, lietuvių duomenys buvo palyginti su 1000 genomų projekto europiečių duomenimis. Informacija apie statistiškai visiškai patikimai (nustatyti pagal Pirsono $\chi 2$ ir tikslųjį Fišerio kriterijus) nuo europiečių ir pasaulio populiacijos besiskiriančius patogeninius genomo variantus pateikta 5-o priedo lentelėje.

Nuo europiečių dažnių statistiškai patikimai besiskiriantys patogeniniai genomo variantai, su jais susijusios ligos ir identifikaciniai ligų kodai pateikti 7 lentelėje.

7 lentelė. Statistiškai reikšmingai nuo europiečių genomo variantų dažnių besiskiriantys tirti patogeniniai variantai, jų sąlygotos ligos bei šių ligų OMIM ir TLK-10 kodai

Genomo variantas. dbSNP 142 kodas	Genas	Sutrikimas	OMIM	TLK kodas
rs1800553:G>A	*ABCA4*	Štargarto geltonosios dėmės distrofija, paveldima tinklainės distrofija	601691	H35.5
rs76308115:C>T	*PDE11A*	Pirminė abipusė mazginė antinksčių hiperplazija.	604961	E24.8
rs115532916:G>A	*ACAD9*	Acil-kofermento A dehidrogenazės stoka	611103	E71.3
rs142181517:T>A	*PHYKPL*	Fosfohidroksilizi-nurija	614683	-
rs121917710:G>A	*GLI3*	Polidaktilija	165240	Q69
rs41297018:G>A	*DCLRE1C*	Sunkus kombinuotas imunodeficitas	605988	D81
rs77724903:A>T	*RET*	Šeiminė medulinė skydliaukės karcinoma	164761	E31.2
rs41298135:G>A	*MYO7A*	*Usher* sindromas	276903	H35.5
rs201968272:G>A	*DDX11*	*Warsaw* sindromas	601150	-
rs45495503:C>T	*EPB42*	Paveldėtoji sferocitozė	612690	D58.0
rs113298164:C>T	*LIPC*	Kepenų lipazės stoka	614025	E78.4
rs41549716:A>G	*POLG*	Progresuojanti išorinė oftalmoplegija	157640	H49.4
rs34526199:T>A	*AMPD1*	Mitochondrinė miopatija	615511	G71.3
rs2229738:G>A	*CPT1A*	Adrenoleukodistrofija	255120	E71.3
rs1736557:G>A	*FMO3*	Trimetilaminurija	136132	E88.8
rs11887534:G>C	*ABCG8*	Tulžies pūslės liga	605460	K80
rs28757581:A>G	*OR2J3*	Galimybė užuosti žolės kvapą	615082	-
rs3749977:C>T	*OR2J3*	Galimybė užuosti žolės kvapą	615082	-
rs45547231:C>T	*RAPSN*	Įgimtas miastenijos sindromas	616326	G70.2
rs2269219:C>T	*FECH*	Paveldėtoji eritropoezinė porfirija	177000	E80.0
rs12483377:G>A	*COL18A1*	*Knobloch* sindromas	267750	Q15.8

Iš 30 unikalių patogeninių trumpų iškritų ir intarpų 25 neatitiko pasirinkto kriterijaus, kad būtų daugiau nei trijuose aleliuose, o tarp likusių 5 neatsirado patogeninių pakaitų, kurios reikšmingai skirtųsi nuo europiečių duomenų, todėl tarppopuliaciniuose palyginimuose duomenys yra tik VNV pakaitų pabūdžio.

4. TYRIMO REZULTATŲ APTARIMAS

Pagrindinis disertacinio darbo tikslas buvo, naudojant viso egzomo sekoskaitą, nustatyti lietuvių populiacijoje paplitusius patogeninius genomo variantus. Tokia egzomų sekoskaita buvo atlikta 96 asmenims. Šiuo metu populiariausios patogeninių genomo variantų duomenų bazės *ClinVar* duomenimis, tiriami asmenys turėjo 321 VNV variantus ir 30 trumpų iškritų/intarpų, įvardytų kaip patogeniniai.

Apibendrinant tyrimų rezultatus, bus aptartas šių patogeninių variantų pobūdis, patogeninių variantų lemiamos ligos, lietuvių turimų patogeninių variantų išskirtinumai bei patogeninių variantų ir pačios patogeninių genomo variantų duomenų bazės *ClinVar* patikimumas.

4.1. Patogeninių genomo variantų sukeliamos ligos ir sutrikimai

Egzomo sekoskaita nustatytų patogeninių genomo variantų lemiamoms ligoms suklasifikuoti buvo panaudotas Tarptautinės statistinės ligų ir sveikatos sutrikimų klasifikacijos dešimtasis leidimas (TLK-10-AM).

Galutinė anotacija pateikė ligos ar sutrikimo pavadinimą, kuris buvo įvedamas į *Orphanet* duomenų bazę, taip gautas TLK-10-AM kodas, kuris buvo tikrinamas TLK-10-AM sisteminiame ligų sąraše (8 lentelė).

Daugiausia patogeninių variantų buvo susiję su endokrininių, mitybos ir medžiagų apykaitos ligomis bei sutrikimais. Jie sudarė net 30,5 %. Tokią didelę imtį galima paaiškinti tuo, kad paveldimų mitybos ligų genetinės priežastys yra ištirtos geriausiai. Daug tyrimų yra orientuoti į medžiagų apykaitos ligų gydymą, nes, gydant tokias ligas, galima pritaikyti specialią dietą. Didelė tokių ligų tyrimų gausa lemia didelį patogeninių genomo variantų, kurie yra paviešinti patogeninių genomo variantų duomenų bazėje, skaičių.

8 lentelė. Ligų grupės, kurioms priklauso patogeniniai genomo variantai, nustatyti 96 sveikiems asmenims iš bendrosios lietuvių populiacijos

Ligų klasės pagal TLK-10	TLK-10 grupė	Dažnis, %
Endokrininės, mitybos ir medžiagų apykaitos ligos	E	30,5
Kraujo ir kraujodaros organų ligos bei tam tikri sutrikimai, susiję su imuniniais mechanizmais	D	15,2
Įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos	Q	14,3
Akies ir jos priedinių organų ligos	H	7,6
Nervų sistemos ligos	G	6,7
Kraujotakos sistemos ligos	I	6,7
Virškinimo sistemos ligos	K	4,8
Navikai	C	2,9
Simptomai, požymiai ir nenormalūs klinikiniai bei laboratoriniai radiniai, neklasifikuojami kitur	R	2,9
Kai kurios infekcinės ir parazitinės ligos	B	1,8
Jungiamojo audinio ir raumenų bei skeleto ligos	M	1,8
Urogenitalinės sistemos ligos	N	1,8
Ausies ir speninės ataugos ligos	H	1,0
Kvėpavimo sistemos ligos	J	1,0

Kraujo ir kraujodaros organų ligos ir tam tikri sutrikimai, susiję su imuniniais mechanizmais, bei įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos sudaro apie 15 % visų identifikuotų patogeninių genomo variantų. Šiai ligų grupei priklauso įvairios anemijos bei kraujo krešėjimo sutrikimai, kuriems paveldėjimo faktorius turi didelę įtaką, tad žinomas platus spektras patogeninių variantų, kurie susiję su šia klase.

Įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos – tai klasė ligų, kuri yra išskirtinai genetinio pobūdžio, tad gana didelis su šia klase susijusių patogeninių variantų skaičius nestebina.

Po 7 % visų patogeninių variantų buvo susiję su akių, nervų, kraujotakos sistemų ligomis. Su virškinimo ligomis siejama apie 5 % visų patogeninių variantų. Šioms klasėms priklausančioms ligoms egzomų sekoskaita buvo itin naudinga, ir ja nustatytas nemažas skaičius su šiomis ligomis siejamų genomo variantų.

Likusių ligų, susijusių su nustatytais patogeniniais genomo variantais, imtis yra nedidelė. Įdomu, kad tik mažas patogeninių variantų skaičius sukelia ausies ir speninės ataugos klasės ligas. Su klausa susiję paveldimi sutrikimai yra gerai ištirti, tad tikėtasi didesnio patogeninių šiai klasei priklausančių variantų skaičiaus.

Lietuvos higienos institutas pateikia įvairią statistiką, kartu ir sergamumo bei mirtingumo, kuri taip pat pateikiama klasifikuojant pagal TLK-10. Tad tyrimo duomenis yra lengva palyginti su realiu Lietuvos gyventojų sergamumu (9 lentelė).

9 lentelė. Lietuvos gyventojų 2014 metų sergamumo dažnis, tenkantis 100 gyventojų, Lietuvos higienos instituto Sveikatos informacijos centro duomenimis

Ligų klasės pagal TLK-10-AM	**TLK-10-AM grupė**	**Sergamumo dažnis, %**
Klasės, su kuriomis susiję patogeniniai variantai aptikti egzomo duomenyse		
Kraujotakos sistemos ligos	I	29,3
Kvėpavimo sistemos ligos	J	20,6
Jungiamojo audinio ir skeleto-raumenų sistemos ligos	M	19,0
Virškinimo sistemos ligos	K	18,2
Urogenitalinės sistemos ligos	N	15,2
Endokrininės, mitybos ir medžiagų apykaitos ligos	E	14,3
Akies ir jos priedinių organų ligos	H	13,6
Nervų sistemos ligos	G	12,0
Navikai	C	7,6
Simptomai, požymiai ir nenormalūs klinikiniai bei laboratoriniai radiniai, neklasifikuojami kitur	R	6,7
Ausies ir speninės ataugos ligos	H	5,5
Kai kurios infekcinės ir parazitinės ligos	B	4,7
Kraujo ir kraujodaros organų ligos bei tam tikri sutrikimai, susiję su imuniniais mechanizmais	D	2,3
Įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos	Q	0,5
Klasės, su kuriomis susijusių patogeninių variantų nerasta egzomo duomenyse		
Psichikos ir elgesio sutrikimai	F	7,3
Odos ir poodžio ligos	L	8,2
Nėštumas, gimdymas ir laikotarpis po gimdymo	O	1,9
Traumos, apsinuodijimai ir kiti išorinių priežasčių padariniai	S-T	13,6

(http://www.hi.lt/news/904/789/Isleistas-leidinys-Lietuvos-gyventoju-sergamumas-apskrityse-ir-savivaldybese-2014-m.html)

Palyginus mūsų nustatytus ir Lietuvos higienos instituto duomenis, matyti didelis skirtumas tarp su ligomis susijusių įvardytų patogeninių variantų ir Lietuvos gyventojų sergamumo. Tai nestebina, nes visų pirma didžiausią dalį sudaro kraujotakos, kvėpavimo sistemų ligos, kurioms labai didelę įtaką turi gyvenimo būdas. Tas pat galioja ir jungiamojo audinio ir skeleto-raumenų sistemos ligų klasei, ypač turint omenyje, kad duomenys yra vien tik suaugusių asmenų. Didelę dalį sudaro ir traumos, kurių, žinoma, nėra tarp paveldimų ligų. Įgimtos formavimosi ydos, deformacijos ir chromosomų anomalijos – specifinio genetinio pobūdžio ligos – tesudaro 0,5 %.

Nepaisant didelių skirtumų, kurie atsiranda dėl aplinkos poveikio asmenų sveikatai, galima padaryti keletą išvadų. Pirmiausia, genetiniai veiksniai turi nedidelę įtaką lietuvių sergamumui, ir tinkamas piliečių gyvenimo būdas gerokai sumažintų Lietuvos gyventojų sergamumą.

Kitas dalykas yra tas, kad, žvelgiant į skirtumus, galima nuspėti, apie kurių klasių ligas lemiančius patogeninius genomo variantus žinių trūksta, ir, pasirinkus būtent šių ligų tyrimus, galima labiau tikėtis rasti iki šiol neįvardytus patogeninius genomo variantus.

Lyginant realaus ligotumo ir patogeninių genomo variantų lemiamas ligas įdomu, kad nėra patogeninių genomo variantų, susijusių su odos ir poodžio ligų klase. Derėtų pabrėžti, kad vis dėlto įvardytų patogeninių genomo variantų, susijusių su odos ligomis, būta, tačiau šie genomo variantai neįtraukti į tolesnę analizę. Taip yra dėl to, kad šie genomo variantai lemia šviesią odos spalvą. Šviesi oda yra odos vėžio rizikos veiksnys, tačiau šiuos variantus turi visa baltaodžių rasė, tad abejotina, ar tokius genomo variantus reikėtų analizuoti kaip patogeninius.

4.2. Naudotos patogeninių genomo variantų duomenų bazės įvertinimas

Vieša patogeninių genomo variantų duomenų bazė *ClinVar*, paskelbta 2013 metais, itin sparčiai didėja ir yra naudojama praktiškai atliekant diagnostinio pobūdžio tyrimus visame pasaulyje. Dauguma patogeninių variantų yra patalpinti bendradarbiaujant su kitomis didelėmis duomenų bazėmis, kaip antai OMIM, ar privačiomis genetinių tyrimų bendrovėmis, tokiomis kaip *GeneDx*, taip pat didelė dalis patogeninių variantų patalpinti didelių genetinių tyrimų laborotorijų iš viso pasaulio, daugiausia iš JAV (pagrindiniai *ClinVar* šaltiniai yra pateikti 4 priede). Šiuo metu iš viso yra daugiau kaip 524 skirtingi *ClinVar* duomenų šaltiniai. Pateikti savo patogeninį genomo variantą gali ir pavienis tyrėjas, tokiu atveju ekspertai peržiūri tyrėjo pateikiamą variantą. Pati duomenų bazė turi priskirtas penkias patogeniškumo kategorijas ir rangų sistemą, rodančią duomenų patikimumą.

Ši duomenų bazė yra vienas geriausių įrankių, ypač tiriant patogeninius genomo variantus, turimus sveikų asmenų. Tačiau, rengiant disertacinį darbą, prireikus pasinaudoti šia duomenų baze, iškilo klausimų dėl kai kurių duomenų patikimumo. Toks klausimas iškilo įvertinus patogeniniais laikytinų variantų dažnį. Kai kurių patogeniniais laikytinų variantų dažnis buvo netgi 100 %. Keletas *ClinVar* duomenų bazėje patogeniniais laikomų variantų, kurių dažnis bendrojoje populiacijoje yra itin didelis, pavyzdžių pateikiama 10 lentelėje.

10 lentelė. Tyrimo metu aptiktų genomo variantų, kurie vertinami kaip patogeniniai, tačiau yra labai dažni bendrojoje populiacijoje, pavyzdžiai

Genas	*dbSNP*	1 000G pasaulio pop	1000G Europos pop	ESP Europos	LITGEN
DPYD	rs1801265	73,98 %	78,53 %	76,52 %	72,45 %
DBT	rs12021720	89,18 %	92,15 %	91,38 %	92,86 %
NOS3	rs1799983	82,37 %	65,61 %	75,30 %	70,41 %
HPD	rs1154510	87,64 %	87,18 %	85,02 %	84,18 %
BBS2	rs4784677	99,64 %	99,11 %	99,38 %	100 %
PRODH	rs450046	90,56 %	92,84 %	91,71 %	63,78 %

Nors patogeninis genomo variantas gali tik padidinti riziką susirgti kokia nors liga, tačiau kai bendrojoje populiacijoje atsiranda patogeninių variantų, kurių dažnis 100 %, tai yra signalas, kad genomo variantas priskiriamas prie patogeninių neteisingai. Tai teigia ir kiti autoriai, pradedant *Callum J. Bell* ir kt., – pastarieji 2011 metais iškėlė esamų duomenų bazių patikimumo problemą ir nurodė, kad bent 10 % skelbiamų patogeninių variantų nėra teisingi (Bell et al., 2011).

Kyla klausimas, kodėl viena plačiausiai naudojamų duomenų bazių, turinti ekspertų grupę, įvertinančią pateiktus patogeninius genomo variantus, leidžia egzistuoti patogeniniams variantams, kurių dažnis bendrojoje populiacijoje yra labai didelis.

Kaip papildantis veiksnys nemažai patogeniniais įvardytų variantų buvo homozigotiniai pagal alternatyvųjį (patogeninį) alelį, tai turėtų reikšti, kad didelei jų daliai turėtų būti būdingas vienas ar kitas sutrikimas. Vienas iš kai kurių patogeniniais laikytinų ir didelį dažnį turinčių genomo variantų paaiškinimų galėtų būti, kad kai kurios žmonių rasės turi pranašumą prieš kitą rasę. Tokio atvejo pavyzdys yra tamsi ar tamsesnė oda – šviesi baltaodžių oda, mažas pigmento kiekis padidina tikimybę susirgti odos vėžiu.

Tiriamajame darbe 90 laikytinų patogeniniais variantų, t. y. 27,86 % visų identifikuotų patogeninių variantų, nebuvo įtraukti į analizę suabejojus jų patogeniškumu.

Apdorojus *ClinVar* bazės pateikiamus duomenis, padaryta išvada, kad ateityje ši duomenų bazė taps asmeninės medicinos pagrindu, tačiau iki to laiko ši duomenų bazė turi tobulėti tiek patogeninių variantų skaičiaus, tiek jų kokybės požiūriu. Dirbantys su dabartine duomenų baze tyrėjai *ClinVar* duomenų bazėje pateikiamus genomo variantus turėtų vertinti atsargiai.

4.2.1. Patogeniniai genomo variantai, paveldimi autosominiu dominantiniu būdu

Daugumai identifikuotų patogeninių genomo variantų buvo būdingas autosominis recesyvusis paveldėjimo tipas. Tokio rezultato ir tikėtasi, nes buvo analizuoti asmenys iš bendrosios lietuvių populiacijos, kurie save laikė sveikais. Tyrime buvo numatyta nustatyti autosominiu recesyviuoju būdu paveldimų ligų nešiotojus ir patogeninių variantų dažnį.

Tačiau, nors daugumai patogeninių variantų ir būdingas autosominis recesyvusis paveldėjimas, visgi kai kuriems asmenims buvo būdingas autosominis dominantinis paveldėjimas. Tokiems asmenims turėtų pasireikšti ligos simptomatika ir teturint tik vieną patogeninį alelį, bet patogeninis variantas gali turėti ir labai nedidelį paveikį organizmui, kuris gali būti nepastebimas. Tačiau tarp patogeninių variantų pasitaikė ir tokių, kurie turėtų labai rimtų pasekmių organizmui, todėl kyla pačios duomenų bazės patikimumo klausimas.

Tai ypač akivaizdu analizuojant LITGEN projekto atskiros sutrikimų grupės egzomų duomenis, t. y. patogeninius genomo variantus, susijusius su kurtumu (Mikštienė et al., 2016). Nors buvo tiriami sveiki asmenys, net trys genomo variantai buvo priskirti prie sukeliančių šiuos sutrikimus, kurie turi labai rimtų klinikinių pasekmių, o paveldėjimo tipas yra autosominis. Tai du variantai, esantys *MYO1A* gene: rs33962952 G>A ir rs55679042 G>A, bei *MYH14* gene

esanti rs119103280 G>T pakaita. *MYO1A* gene: rs33962952 G>A (c.1150G>T (p.Gly384Cys)) ir rs55679042 G>A (c.1985G>A (p.Gly662Glu)) į *ClinVar* duomenų bazę pateko iš OMIM, labai svarbios mendelinių ligų duomenų bazės, o į pastarąją iš poros publikacijų (Donaudy et al., 2003; Eisenberger et al., 2014).

Detalesni šių genomo variantų tyrimai parodė, kad šie variantai yra nepatogeniniai, tačiau iš *ClinVar* duomenų bazės tokie neteisingai patogeniniais laikomi genomo variantai nėra pašalinti. Taip pat yra ir su *MYH14* gene esančia rs119103280G>T (c.1150G>T (p.Gly384Cys)) pakaita. Tokių netikrų genomo variantų saugojimas duomenų bazėje pasunkina anotaciją ir galutinę analizę.

GDF3 gene esanti pakaita rs140926412 C>T (c.796C>T (p.Arg266Cys)) lemia retą *Klippel-Feil* sindromą. Tai sunki kaulų liga, turinti aiškius fenotipinius požymius, todėl irgi labai mažai tikėtina, kad tai bus realus variantas tarp sveikų asmenų (Ye et al., 2010).

Itin įdomus genomo variantas, kuris įvardijimas kaip paveldimas autosominiu dominantiniu būdu, yra *PDE11A* gene esantis rs76308115 C>T(c.169C>T (p.Arg57Ter)). Jis lemia pirminę abipusę mazginę antinksčių hiperplaziją, sukeliančią Kušingo sindromui būdingus antinksčių pokyčius (Horvath et al., 2006). Šis variantas yra retas pasaulyje, bet Lietuvoje jo dažnis yra apie 2 %. Šis variantas statistiškai patikimai skiriasi ne tik nuo bendros populiacijos, bet ir nuo europiečių populiacijos. Variantas tampa dar įdomesnis, kai analizuojamas etnolingvistiniu požiūriu – jį turi tik asmenys, esantys iš Šiaurės Žemaitijos (net 4 asmenys iš 16 galimų).

Vienam tiriamam asmeniui pasitaikė labai retas patogeniniu įvardytas genomo variantas, neturintis žinomo dažnio 1000 genomų projekto ir ExAC duomenyse. Tačiau toks dominantį paveldėjimą turintis variantas, lemiantis metabolinę hipokalcemijos ligą, gali būti ir tikras, nes neturi itin gerai išreikšto fenotipo, o šios ligos požymis yra nedidelis kalcio kiekis kraujyje.

Kad asmenys iš bendrosios populiacijos yra ne tik ligų ir sutrikimų nešiotojai, bet ir patys turi nedidelių sutrikimų, rodo tiek dominantiniai variantai, tiek recesyviuoju būdu paveldimi genomo variantai, kurie yra homozigotiniai pagal abu patogeninius alelius.

Tarp tirtų patogeninių variantų, kurių paveldėjimo tipas yra autosominis dominantinis, pasitaikė variantų, kurių dažnis, palyginti su europiečiams būdingu šių variantų dažniu, buvo statistiškai reikšmingai didesnis (11 lentelė).

11 lentelė. Patogeniniai genomo variantai, kurie lemia autosominius dominantinius sutrikimus ir turi statistiškai reikšmingai didesnius dažnius lietuvių bendrojoje populiacijoje, palyginti su 1000G projekto europiečiams būdingais šių variantų dažniais. Pateikiami patogeninių variantų kodai dbSNP duomenų bazėje ir statistiniai įverčiai

Genomo variantas. dbSNP 142 kodas	**Genas**	**Sutrikimas**	**1000G EUR, %**	**LT, %**	**1000G EUR chi**	**1000G EUR Fisher**
rs76308115	*PDE11A*	Pirminė abipusė mazginė antinksčių hiperplazija	0,50	2,04	0,0224	0,0447
rs121917710	*GLI3*	Polidaktilija	0,70	3,57	6,4 x 10-4	3,53 x 10-3
rs77724903	*RET*	Šeiminė medulinė skydliaukės karcinoma	0,10	2,04	1,1281 x 10-4	3,0128 x 10-3
rs41549716	*POLG*	Progresuojanti išorinė oftalmoplegija	0,70	3,06	0,0036	0,0113

Tarp tirtų duomenų buvo trys genomo variantai, kurie paveldimi autosominiu dominantiniu būdu ir kurių dažniai buvo statistiškai reikšmingai mažesni, palyginti su europiečiams būdingais šių variantų dažniais (12 lentelė).

12 lentelė. Patogeniniai genomo variantai, kurie lemia autosominius dominantinius sutrikimus ir turi statistiškai reikšmingai mažesnius dažnius lietuvių bendrojoje populiacijoje, palyginti su 1000G projekto europiečiams būdingais šių variantų dažniais. Pateikiami patogeninių variantų kodai dbSNP duomenų bazėje ir statistiniai įverčiai

Genomo variantas. *dbSNP* 142 kodas	**Genas**	**Sutrikimas**	**1000G EUR, %**	**LT, %**	**1000G EUR chi**	**1000G EUR Fisher**
rs1805124	*SCN5A*	Atrioventrikulinė disociacija	21,67	13,78	$1,4802 \times 10^{-5}$	$7,5158 \times 10^{-6}$
rs28757581	*OR2J3*	Galimybė užuosti žolės kvapą	11,93	6,63	$7,662 \times 10^{-3}$	$7,765 \times 10^{-3}$
rs3749977	*OR2J3*	Galimybė užuosti žolės kvapą	22,86	17,86	$9,5588 \times 10^{-5}$	$6,6306 \times 10^{-5}$

Patogeninių variantų buvimas tarp save įvardinančių sveikais asmenų yra visiškai realus, nes nemažai patogeninių variantų neturi aiškiai apibrėžto fenotipinio pasireiškimo – nutukimas, nejutimas tam tikrų kvapų, didesnė rizika susirgti, nedidelę įtaką turinti fermentų stoka ir t. t.

4.3. Naudojamo referentinio genomo sąlygoti netikslumai

Iš 321 įvardyto patogeninio varianto pasitaikė iš viso trys genomo variantai, kuriuos turėjo visi tirti asmenys, o šių variantų dažnis bendrojoje populiacijoje siekė apie 100 %. Tai klaidingas tiriamų duomenų anotavimas, bet šiuo atveju ne dėl patogeninių duomenų bazės netikslumų kaltės. Referentinis genomas nėra nei vieno konkretaus asmens genomas, nei iš labai didelės žmonių populiacijos duomenų sudarytas genomas. Didžioji referentinio genomo dalis

sudaryta iš keleto asmenų, tad jeigu šiems asmenims būtų būdingas koks nors patogeninis genomo variantas, tai gali labai pasunkinti sąlygas tyrėjams, kurie analizuos savo tyrimo duomenis, lygindami su tokiu referentiniu genomu.

Šiame darbe trys patogeniniais įvardyti genomo variantai, kurių dažnis yra apie 100 %, atsiradę dėl to, kad pačiame referentiniame genome yra ne patogeninis, o itin dažnas ir neturintis neigiamo klinikinio poveikio genomo variantas. Analizės algoritmo teisingų genomo variantų atrankos etape programų tikslas tėra rasti patikimą skirtumą tarp tiriamų duomenų ir referentinio genomo, tad jei tiriami asmenys turi normalų genomo variantą, o referentiniame genome yra patogeninis variantas, – toks variantas fiksuotas kaip tikras, o vėliau anotuojamas, taip tyrėjams pasunkinant galutinę analizę.

Darbe buvo naudotas hg19 referentinis genomas, tad dėl jo trys genomo variantai įvardyti kaip patogeniniai. Jie pateikiami 13 lentelėje.

13 lentelė. Variantai, klaidingai įvardyti kaip patogeniniai referentiniame genome hg19

Chromo-soma	Pozicija	REF	Alt	Genas	*dbSNP* 142 kodas	1000G pasaulinis dažnis	*Clinvar*
1	100672060	T	C	DBT	rs12021720	0,89	Klevų sirupo šlapimo liga
1	169519049	T	C	F5	rs6025	0,99	Trombofilija dėl Leideno 5 faktoriaus
16	56548501	C	T	BBS2	rs4784677	0,99	*Bardet-Biedl* sindromas

4.4. Patogeninių genomo variantų ir jų genų, lemiančių recesyviąsias ligas lietuvių asmenų grupėje, įvertinimas

Kuriant šio tyrimo strategiją – analizuoti sveikų asmenų patogeninius genomo variantus – svarbiausiu tikslu buvo iškeltas recesyviojo paveldėjimo tyrimas. Siekta pažinti retų recesyviųjų ligų, kuriomis serga lietuviai, įvairovę,

nustatyti ligų nešiotojus ir ligas, kurių dažnis tarp tirtų asmenų yra didesnis nei palyginimui numatytose populiacijose.

Kiekvieno asmens egzomų sekoskaita leido identifikuoti apie 44 tūkstančius genomo variantų (įskaičiuojant tiek VNV, tiek trumpas iškritas ir intarpus), kurie skyrėsi nuo referentinio genomo. Nustatyta, kad visi 96 asmenys turi per 275 tūkstančius pakaitų. Įvardytas 321 patogeninis genomo variantas, 9 jų buvo paveldimi dominantiniu būdu, o daugumai likusiųjų, kaip ir buvo tikėtasi, buvo būdingas recesyvusis paveldėjimas. Net 103 patogeninių genomo variantų dažniai statistiškai reikšmingai skyrėsi nuo pasaulinių dažnių. Palyginimas su europiečių duomenimis padėjo išryškinti savitus lietuvių genomo variantus.

14 lentelėje pateikti recesyviai paveldimi genomo variantai, kurie turėjo statistiškai reikšmingai didesnį dažnį nei jis buvo europiečių populiacijose.

Itin įdomus patogeninis genomo variantas yra susijęs su sunkiu kombinuotu imunodeficitu dėl *DCLRE1C* geno produkto stokos. Jį turėjo net 10 iš 96 asmenų ir visur jis buvo heterozigotinis patogeninio alelio atžvilgiu.

Atliktas *Sanger* sekoskaitos tikrinimas įrodė, kad visiems asmenims šis variantas NKS buvo sekvenuotas teisingai. Šis genomo variantas svarbus ir tuo, kad pradėtas tirti atliekant visuotinę naujagimių patikrą. Kadangi lietuvių populiacijoje šio genomo varianto dažnis didesnis už pasaulinį ir europiečių dažnį, toks variantas būtų puikus kandidatas į Lietuvoje esantį naujagimių patikros sąrašą.

14 lentelė. Patogeniniai genomo variantai, kurie lemia autosomines recesyviąsias ligas ir turi statistiškai reikšmingai didesnius dažnius lietuvių bendrojoje populiacijoje, palyginti su 1000G projekto europiečiams būdingais šių variantų dažniais. Pateikiami tų patogeninių variantų kodai dbSNP duomenų bazėje ir statistiniai įverčiai

Genomo variantas. dbSNP 142 kodas	**Genas**	**Sutrikimas**	**1000G EUR, %**	**LT,%**	**1000G EUR chi**	**1000G EUR Fisher**
rs1800553	*ABCA4*	Štargarto geltonosios dėmės distofija	0,40	2,55	$1,419 \times 10^{-3}$	$8,006 \times 10^{-3}$
rs115532916	*ACAD9*	Acil-kofermento A dehidrogenazės stoka	2,19	7,14	$2,6454 \times 10^{-4}$	$9,75 \times 10^{-4}$
rs142181517	*PHYKPL*	Fosfohidroksilizi-nurija	0,40	3,57	$2,0602 \times 10^{-5}$	$5,2093 \times 10^{-4}$
rs41297018	*DCLRE1C*	Sunkus kombinuotas imunodeficitas	2,19	5,10	0,0241	0,04919
rs41298135	*MYO7A*	*Usher* sindromas	0,40	2,04	$9,8591 \times 10^{-3}$	0,0283
rs45495503	*EPB42*	Paveldėtoji sferocitozė	0,20	2,04	$8,2718 \times 10^{-4}$	$7,9094 \times 10^{-3}$
rs113298164	*LIPC*	Kepenų lipazės stoka	0,50	2,04	0,0224	0,0447
rs34526199	*AMPD1*	Mitochondrinė miopatija	3,68	7,14	0,0380	0,053
rs2229738	*CPT1A*	Adrenoleukodis-trofija	8,65	16,33	0,044	0,0519
rs104895094	*MEFV*	Šeiminė Vidužemio jūros karštligė	0,55	6,63	$3,2363 \times 10^{-9}$	$1,1564 \times 10^{-6}$

Penki genomo variantai, paveldimi autosominiu recesyviuoju būdu, turėjo gerokai mažesnius dažnius, palyginti su europiečiams būdingais šių variantų dažniais (15 lentelė). Tai duoda naujų įžvalgų, kaip antai, ar nereikėtų šių variantų įtraukti į platesnius tyrimus Lietuvoje.

15 lentelė. Patogeniniai genomo variantai, kurie lemia autosomines recesyviąsias ligas ir turi statistiškai reikšmingai mažesnius dažnius lietuvių bendrojoje populiacijoje, palyginti su 1000G projekto europiečiams būdingais šių variantų dažniais. Pateikiami tų patogeninių variantų kodai dbSNP duomenų bazėje ir statistiniai įverčiai

Genomo variantas. dbSNP 142 kodas	**Genas**	**Sutrikimas**	**1000G EUR, %**	**LT, %**	**1000G EUR chi**	**1000G EUR Fisher**
rs1736557	*FMO3*	Trimetilaminurija	6,26	2,55	0,0264	0,0296
rs11887534	*ABCG8*	Tulžies pūslės liga	7,95	2,04	0,0006	0,0001
rs45547231	*RAPSN*	Įgimtas miastenijos sindromas	15,01	5,10	$1,7077 \times 10^{-6}$	$1,2823 \times 10^{-7}$
rs2269219	*FECH*	Paveldėtoji eritropoezinė porfirija	20,08	17,35	$2,096 \times 10^{-3}$	$1,773 \times 10^{-3}$
rs12483377	*COL18A1*	*Knobloch* sindromas	8,35	4,59	0,01	0,009

Kai kurie patogeniniais įvardijami sutrikimai turi nedidelę fenotipinę reikšmę. Tokių mažą fenotipinę įtaką turinčių autosominiu recesyviuoju būdu paveldimų patogeninių genomo variantų yra nemažai – įvairūs nedideli fermentų trūkumai, rizikos veiksniai ar galimos problemos dėl nutukimo.

Nemažai identifikuotų genomo variantų buvo susiję su metabolizmo ir medžiagų apykaitos ligomis. Tikėtina, kad tai rodo didesnį šių ligų ištirtumą, nes šias genetines ligas galima gydyti.

4.5. Patogeninių recesyviųjų genomo variantų įvertinimas etnolingvistinių grupių atžvilgiu

Patogeniniai genomo variantai yra retesni nei normalūs. Kai kurie patogeniniai genomo variantai yra itin reti, tad ir visa tirta 96 asmenų imtis laikytina nedidele. Atsižvelgus į etnolingvistines grupes, ir taip maža imtis gerokai sumažėja ir skyla į grupeles po 16 asmenų. Nepaisant šio fakto, gauta gana įdomių rezultatų, visų pirma išskirti variantai, kurie būdingi tik žemaičiams ar tik aukštaičiams.

Tik tarp žemaičių egzomų aptikta autosominiu recesyviuoju būdu paveldima rs142181517 A>T (c.1310A>T (p.Glu437Val)) pakaita. Šis patogeninis variantas lemia fosfohidroksilizinuriją – tai metabolizmo sutrikimas, kai yra padidėjusi fosfohidroksilizino koncentracija šlapime. Iš visų tirtų 96 asmenų septyni yra šio sutrikimo turėtojai: 4 yra iš Vakarų Žemaitijos, 2 – iš Šiaurės Žemaitijos, 1 – iš Pietų Žemaitijos. Nė vieno nėra iš Aukštaitijos. Šį patogeninį genomo variantą nustatė *Maria Veiga-da-Cunha* ir kiti bei atliko patikrinimą, naudodami *E. coli* rekombinantus (Veiga et al., 2013). Šis patogeninis genomo variantas pasitaiko 1 iš 16 000 asmenų, o tarp žemaičių – gerokai dažniau. Asmenų, turinčių šį sutrikimą, šlapime nustatoma apie 5 mmol fosfohidroksilizino koncentracija, o kontrolinės grupės asmenų (neturinčių šio sutrikimo) šlapime fosfohidroksilizino neaptinkama. Asmenims, turintiems šį sutrikimą, yra būdingas per didelis jungiamojo audinio (sąnarių) lankstumas.

Tik žemaičiams buvo nustatytas kitas autosominiu recesyviuoju būdu paveldimas genomo variantas, esantis *BCHE* gene rs1799807 A>G (c.293A>G (p.Asp98Gly)). Tai metabolinis sutrikimas, paveikiantis butirilcholinesterazės fermento struktūrą ir veiklą. Į *ClinVar* programą šis patogeninis genomo variantas yra perkeltas iš OMIM duomenų bazės. Fermentas butirilcholinesterazė apsaugo acetilcholinesterazę nuo cholinesterazės inhibitorių, taip užtikrindamas sklandžią acetilcholino neurotransmisiją. Šis variantas labai svarbus farmakologiniu požiūriu, nes asmeniui, turinčiam tokį

patogeninį genomo variantą, gali pasireikšti apnėjos forma, vadinama poanestezine apnėja. Vykdant nuskausminimą asmenys yra paveikiami tam tikrais raumenis atpalaiduojančiais vaistais, žinomais cholino esterių vardu, – tai ir sukelia poanestezinę apnėją.

Tik tarp aukštaičių buvo nustatyta autosominiu recesyviuoju būdu paveldima rs1800562 G>A (c.845G>A (p.Cys282Tyr)) pakaita. Šis patogeninis variantas lemia hemochromatozę – tai metabolizmo sutrikimas, kai yra padidėjusi geležies koncentracija kraujyje. Patogeninis genomo variantas yra būdingas išskirtinai Europai, kur nešiotojų dažnis yra 1 iš 200, o šiaurės Europoje net 1 iš 10. Ankstyvas nustatymas gali padėti apsaugoti nuo dėl geležies pertekliaus atsirandančių sutrikimų, tokių kaip pažeistos kepenys, cukrinis diabetas. Sutrikimo pasireiškimas priklauso ir nuo lyties: 30 % vyrų, turinčių šį sutrikimą, buvo kilę rimtesnių komplikacijų, ir tik 1 % moterų, nes geležies perteklius pašalinamas menstruacijų metu. *Pale Pedersen* ir *Nils Milman* aprašo šį sutrikimą 6 000 Danijos vyrų (Pedersen and Milman, 2009). Iš lietuvių šis sutrikimas buvo būdingas visoms aukštaičių etnolingvistinėms grupėms, bet nepasitaikė nė vienoje žemaičių etnolingvistinėje grupėje.

Neįprastai didelis dažnis autosominiu recesyviuoju būdu paveldimos pakaitos rs104895094 A>G (c.2084 A>G (p. Lys695Arg)), kuri yra *MEFV* gene, buvo būdingas rytų ir pietų aukštaičiams. Ši pakaita lemia šeiminę Vidužemio jūros karštligę, kuri dar vadinama armėnų liga. Didelis šios pakaitos dažnis tarp rytų ir pietų aukštaičių yra įdomus ir tuo, kad šis patogeninis variantas yra paplitęs Vidužemio jūros regione ir yra būdingas žydams, arabams, graikams, italams, armėnams ir turkams. Tirtų asmenų bent trys kartos buvo lietuvių kilmės, ir tokio varianto paplitimas gali rodyti praeityje įvykusią migraciją iš Vidužemio jūros regiono. Tarp lietuvių pagal šį patogeninį genomo variantą buvo 13 heterozigotų ir 0 homozigotų, 9 iš jų yra iš Pietų ir Rytų Aukštaitijos (3 PA+6 RA). Į *ClinVar* duomenų bazę šis patogeninis genomo variantas yra įkeltas iš keleto šaltinių (Bernot et al., 1998).

Jei analizuotume kiekvieną etnolingvistinę grupę atskirai, unikalių joms patogeninių genomo variantų kiekis būtų panašus (16 lentelė).

16 lentelė. Unikalių pakaitų skaičius kiekvieme regione priklausomai nuo pakaitų dažnumo bendrojoje populiacijoje

Regionas	**Dažnis**	
	50 %	1 %
PA	23	16
RA	16	14
VA	23	21
PŽ	20	20
ŠŽ	20	18
VŽ	22	20

Tačiau grupės po 16 asmenų yra per mažos retiems variantams analizuoti. Visose grupėse asmenų, turinčių toms grupėms būdingų pakaitų, tebuvo po 1 ar 2, todėl tų pakaitų negalime atskirti nuo atsitiktinių. Vienintelėje Šiaurės Žemaitijoje vieną pakaitą turėjo net keturi asmenys. Kiti etnolingvistiniai regionai jos neturėjo. Tai įdomus variantas, tačiau jo paveldėjimo pobūdis yra autosominis dominantinis, jis paminėtas 4.2.1 skyrelyje.

IŠVADOS

1. Atliktas NKS analizės algoritmo programų lyginimas parodė, kad *Lifescope*^{TM} programa geriausiai tinka *SOLiD* sekoskaitos duomenims analizuoti.
2. Kiekvienas asmuo vidutiniškai turėjo 45 VNV, kurie įvardyti kaip patogeniniai. Patogeniniais įvardytų iškritų/intarpų pasitaikydavo nuo nulio iki trijų vienam asmeniui. Tarp 96 tirtų asmenų patogeniniais įvardytas 321 unikalus VNV ir 30 unikalių iškritų/ intarpų. Tirtoje lietuvių populiacijoje didžioji dalis identifikuotų patogeninių genomo variantų buvo susiję su metabolizmo sutrikimais.
3. *ClinVar* patogeninių genomo variantų duomenų bazėje aptikta patogeniniais įvardijamų genomo variantų, kurių dažnis itin didelis Lietuvos ir pasauliniuose duomenyse.
4. Lyginant tirtų asmenų patogeninių variantų dažnius su europiečių genomo variantų dažniais, identifikuota 15 patogeninių genomo variantų, kurie paveldimi autosominiu recesyviuoju būdu ir kurie statistiškai patikimai skyrėsi nuo kitų europiečių populiacijų – 10 iš jų turėjo didesnį dažnį, o 5 turėjo mažesnį dažnį nei europiečių dažniai. Nustatyti 103 patogeniniai variantai, kurie statistiškai patikimai skyrėsi nuo pasaulinių žmogaus genomo variantų dažnių (lyginant su 1000G, ExAC duomenimis).
5. Tarp etnolingvistinių grupių 321 patogeninis VNV pasiskirstė maždaug vienodai, vidutiniškai po 150 patogeninių genomo variantų kiekvienoje etnolingvistinėje grupėje. Identifikuoti du patogeniniai genomo variantai – rs142181517 A>T ir rs1799807 A>G, kurie buvo būdingi visoms Žemaitijos etnolingvistinėms grupėms ir nė vienai aukštaičių lingvistinei grupei. Išskirtinai aukštaičiams buvo būdingas vienas patogeninis genomo variantas rs1800562 G>A.

6. Patogeninių variantų, būdingų sveikiems asmenims, tyrimai gali turėti didelę praktinę naudą ligų diagnostikai ir prognozavimui. Tarp Lietuvos asmenims identifikuotų patogeninių variantų, turinčių statistiškai patikimai didesnį dažnį, palyginti su pasaulio arba Europos šių variantų dažniais, pasitaikė ir variantas rs41297018:G>A *DCLRE1C* gene, lemiantis sunkų kombinuotą imunodeficitą, kitų šalių praktikoje tiriamą visuotinės naujagimių patikros metu.

LITERATŪRA

1. Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. 2010. 'A method and server for predicting damaging missense mutations', *Nat Methods*, 7: 248–9.
2. Alkan, C., B.P. Coe, and E. E. Eichler. 2011.' Genome structural variation discovery and genotyping,' *Nature Rev Genet*, 12: 363-76
3. Baker, M. W., R. H. Laessig, M. L. Katcher, J. M. Routes, W. J. Grossman, J. Verbsky, D. F. Kurtycz, and C. D. Brokopp. 2010. 'Implementing routine testing for severe combined immunodeficiency within Wisconsin's newborn screening program', *Public Health Rep*, 125, Suppl 2: 88–95.
4. Baudhuin L. M, S. A. Lagerstedt, E. W. Klee, N. Fadra, D. Oglesbee, and M. J. Ferber. 2015.'Confirming Variants in next-generation sequencing panel testing by Sanger sequencing', *J Mol Diagn*, 17: 456-61.
5. Beaudet, A. L., and L. C. Tsui. 1993. 'A suggested nomenclature for designating mutations', *Hum Mutat*, 2: 245–8.
6. Becker, J., O. Semler, C. Gilissen, Y. Li, H. J. Bolz, C. Giunta, C. Bergmann, M. Rohrbach, F. Koerber, K. Zimmermann, P. de Vries, B. Wirth, E. Schoenau, B. Wollnik, J. A. Veltman, A. Hoischen, and C. Netzer. 2011. 'Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta', *Am J Hum Genet*, 88: 362–71.
7. Bell, C. J., D. L. Dinwiddie, N. A. Miller, S. L. Hateley, E. E. Ganusova, J. Mudge, R. J. Langley, L. Zhang, C. C. Lee, F. D. Schilkey, V. Sheth, J. E. Woodward, H. E. Peckham, G. P. Schroth, R. W. Kim, and S. F. Kingsmore. 2011. 'Carrier testing for severe childhood recessive diseases by next-generation sequencing', *Sci Transl Med*, 3: 65ra4.
8. Bennett, S. 2004. 'Solexa Ltd', *Pharmacogenomics*, 5: 433– 8.
9. Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo,

E. Catenazzi M. Chiara, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G. D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. 2008. 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature*, 456: 53–9.

10. Bernot, A., C. da Silva, J. L. Petit, C. Cruaud, C. Caloustian, V. Castet, M. Ahmed-Arab, C. Dross, M. Dupont, D. Cattan, N. Smaoui, C. Dodé, C. Pêcheux, B. Nédelec, J. Medaxian, M. Rozenbaum, I. Rosner, M. Delpech, G. Grateau, J. Demaille, J. Weissenbach, and I. Touitou. 1998. 'Non-founder mutations in the MEFV gene establish this gene as the cause of familial Mediterranean fever (FMF)', *Hum Mol Genet.*,7:1317–25.

11. Biesecker, L. G., and R. C. Green. 2014. 'Diagnostic clinical genome and exome sequencing', *N Engl J Med*, 370:2418-25.

12. Bredenoord, A. L., M. C. de Vries, and J. J. van Delden. 2013. 'Next-generation sequencing: does the next generation still have a right to an open future?', *Nat Rev Genet*, 14: 306.
13. Brookes, A. J. 1999. 'The essence of SNPs', *Gene*, 234: 177–86.
14. Buckingham, S. D. 2008. ‘Scientific software: seeing the SNPs between us,’ *Nature Methods*, 5: 903 – 908.

15. Burgard S.A. and P.V. Chen. 2014. ‘Challenges of health measurement in studies of health disparities’. *Soc Sci Medicine*, 106: 143-50**.**
16. Cabral, R. M., M. Kurban, M. Wajid, Y. Shimomura, L. Petukhova, and A. M. Christiano. 2012. 'Whole-exome sequencing in a single proband reveals a mutation in the CHST8 gene in autosomal recessive peeling skin syndrome', *Genomics*, 99: 202–8.
17. Caggana, M., E. A. Jones, S. I. Shahied, S. Tanksley, C. A. Hermerath, and I. M. Lubin. 2013. 'Newborn screening: from Guthrie to whole genome sequencing', *Public Health Rep*, 128 Suppl 2: 14–9.
18. Calonge, N., N. S. Green, P. Rinaldo, M. Lloyd-Puryear, D. Dougherty, C. Boyle, M. Watson, T. Trotter, S. F. Terry, R. R. Howell and Newborns Advisory Committee on Heritable Disorders in, and Children. 2010. 'Committee report: Method for evaluating conditions nominated for population-based screening of newborns and children', *Genet Med*, 12: 153–9.
19. Chen, R., L. Shi, J. Hakenberg, B. Naughton, P. Sklar, J. Zhang, H. Zhou, L. Tian, O. Prakash, M. Lemire, P. Sleiman, W. Y. Cheng, W. Chen, H. Shah, Y. Shen, M. Fromer, L. Omberg, M. A. Deardorff, E. Zackai, J. R. Bobe, E. Levin, T. J. Hudson, L. Groop, J. Wang, H. Hakonarson, A. Wojcicki, G. A. Diaz, L. Edelmann, E. E. Schadt, and S. H. Friend. 2016. 'Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases', *Nat Biotechnol*, 34: 531–8.
20. Clarke, A. J. 2014. 'Managing the ethical challenges of next-generation sequencing in genomic medicine', *Br Med Bull*, 111: 17–30.
21. Cimbalistienė, L. 2008. ,,Paveldimos medžiagų apykaitos ligos‘‘. Vilnius: Vilniaus un-to leidykla, 272 p.
22. Consortium, Uk K., K. Walter, J. L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J. R. Perry, C. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A. E. Hendricks, P. Danecek, R. Li, J. Floyd, L. V. Wain, I. Barroso, S. E. Humphries, M. E. Hurles, E. Zeggini, J. C. Barrett, V. Plagnol, J. B. Richards, C. M. Greenwood, N. J. Timpson, R. Durbin, and N. Soranzo. 2015. 'The UK10K project identifies rare variants in health and disease', *Nature*, 526: 82–90.
23. Day, I. N. 2010. 'dbSNP in the detail and copy number complexities', *Hum Mutat*, 31: 2–4.
24. Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. 2010. 'Identifying a high fraction of the human genome to be under selective constraint using GERP++', *PLoS Comput Biol*, 6: e1001025.
25. DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J.

Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nat Genet*, 43: 491–8.

26. Dias C., M. Sincan, P. F. Cherukuri, R. Rupps, Y. Huang, H. Briemberg, K.Selby, J. C. Mullikin, T. C. Markello, D. R. Adams, W. A. Gahl, and C. F. Boerkoel. 2012. 'An analysis of exome sequencing for diagnostic testing of the genes associated with muscle disease and spastic paraple,' *Hum Mutat*, 33: 614–26.
27. Donaudy, F., A. Ferrara, L. Esposito, R. Hertzano, O. Ben-David, R. E. Bell, S. Melchionda, L. Zelante, K. B. Avraham, and P. Gasparini. 2003. Multiple mutations of MYO1A, a cochlear-expressed gene, in sensorineural hearing loss, *Am. J Hum Genet.*, 72: 1571–7.
28. Eisenberger, T., R. Slim, A. Mansour, M. Nauck, G. Nurnberg, P. Nurnberg, C. Decker, C. Dafinger, I. Ebermann, C. Bergmann, and H. J. Bolz. 2012. 'Targeted next-generation sequencing identifies a homozygous nonsense mutation in ABHD12, the gene underlying PHARC, in a family clinically diagnosed with Usher syndrome type 3', *Orphanet J Rare Dis*, 7: 59.
29. Eisenberger, T., N. Di Donato, S. M. Baig, C. Neuhaus, A. Beyer, E. Decker, D. Mürbe, C. Decker, C. Bergmann, and H. J. Bolz. 2014. 'Targeted and genomewide NGS data disqualify mutations in MYO1A, the "DFNA48 gene", as a cause of deafness', *Hum Mutat.*, 35: 565–70.
30. Francioli, L. C. , A. Menelaou, S. L. Pulit, F. van Dijk, P.F. Palamara, C. C. Elbers, P. B. T. Neerincx, K. Ye, V. Guryev, W. P. Kloosterman, W. P. Kloosterman, P.Deelen, A. Abdellaoui, E. M. van Leeuwen, M. van Oven, M. Vermaat, M. Li, J. F. J. Laros, L. C. Karssen, A. Kanterakis, N. Amin, J. J. Hottenga, E.-W. Lameijer, M. Kattenberg, M. Dijkstra, and H. Byelas. 2014. 'Whole-genome sequence variation, population structure and demographic history of the Dutch population', Nature Genet, 46: 818–25.
31. Fu, W., T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, Nhlbi Exome Sequencing Project, and J. M. Akey. 2013. 'Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants', *Nature*, 493: 216–20.
32. Gambin, T., S. N. Jhangiani, J. E. Below, I. M. Campbell, W. Wiszniewski, D. M. Muzny, J. Staples, A. C. Morrison, M. N. Bainbridge, S. Penney, A. L. McGuire, R. A. Gibbs, J. R. Lupski, and E. Boerwinkle. 2015. 'Secondary

findings and carrier test frequencies in a large multiethnic sample', *Genome Med*, 7: 54.

33. Garber, M., M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie. 2009. 'Identifying novel constrained elements by exploiting biased substitution patterns', *Bioinformatics*, 25: i54–62.
34. Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. 'A map of human genome variation from population-scale sequencing', *Nature*, 467: 1061–73.
35. Gilissen, C., A. Hoischen, H. G. Brunner, and J. A. Veltman. 2012. 'Disease gene identification strategies for exome sequencing', *Eur J Hum Genet*, 20: 490–7.
36. Green, R. C., J. S. Berg, W. W. Grody, S. S. Kalia, B. R. Korf, C. L. Martin, A. L. McGuire, R. L. Nussbaum, J. M. O'Daniel, K. E. Ormond, H. L. Rehm, M. S. Watson, M. S. Williams, L. G. Biesecker, and Genetics American College of Medical, and Genomics. 2013. 'ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing', *Genet Med*, 15: 565–74.
37. Gudbjartsson, D. F., H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason, S. Besenbacher, G. Magnusson, B. V. Halldorsson, E. Hjartarson, G. T. Sigurdsson, S. N. Stacey, M. L. Frigge, H. Holm, J. Saemundsdottir, H. T. Helgadottir, H. Johannsdottir, G. Sigfusson, G. Thorgeirsson, J. T. Sverrisson, S. Gretarsdottir, G. B. Walters, T. Rafnar, B. Thjodleifsson, E. S. Bjornsson, S. Olafsson, H. Thorarinsdottir, T. Steingrimsdottir, T. S. Gudmundsdottir, A. Theodors, J. G. Jonasson, A. Sigurdsson, G. Bjornsdottir, J. J. Jonsson, O. Thorarensen, P. Ludvigsson, H. Gudbjartsson, G. I. Eyjolfsson, O. Sigurdardottir, I. Olafsson, D. O. Arnar, O. T. Magnusson, A. Kong, G. Masson, U. Thorsteinsdottir, A. Helgason, P. Sulem, and K. Stefansson. 2015. 'Large-scale whole-genome sequencing of the Icelandic population', *Nat Genet*, 47: 435–44.
38. Guthrie, R, and Susi A. 1963. A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics*, 32:33843
39. Hamosh, A., A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. 2005. 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res*, 33: D514–7.
40. Haraksingh, R.R. and M.P. Snyder. 2013. 'Impacts of variation in the human genome on gene regulation,' *J Mol Biol*, 425: 3970–77.
41. Hedges, D., D. Burges, E. Powell, C. Almonte, J. Huang, S. Young, B. Boese, M. Schmidt, M. A. Pericak-Vance, E. Martin, X. Zhang, T. T. Harkins, and S.

Zuchner. 2009.'Exome sequencing of a multigenerational human pedigree', *PLoS ONE*, 4(12): e8232.

42. Horvath, A., S. Boikos, C. Giatzakis, A. Robinson-White, L. Groussin, K. J. Griffin, E. Stein, E. Levine, G. Delimpasi, H. P. Hsiao, M. Keil, S. Heyerdahl, L. Matyakhina, R. Libè, A. Fratticci, L. S. Kirschner, K. Cramer, R. C. Gaillard, X. Bertagna, J. A. Carney, J. Bertherat, I. Bossis, and C. A. Stratakis. 2006. 'A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11A4 (PDE11A) in individuals with adrenocortical hyperplasia', *Nat Genet.*, 38(7): 794–800.
43. Howard, H. C., B. M. Knoppers, M. C. Cornel, E. Wright Clayton, K. Senecal, P. Borry, Genetics European Society of Human, P. G. International Paediatric Platform, Organisation Human Genome, and P. H. G. Foundation. 2015. 'Whole-genome sequencing in newborn screening? A statement on the continued importance of targeted approaches in newborn screening programmes', *Eur J Hum Genet*, 23: 1593–600.
44. Howard, M. F., Y. Murakami, A. T. Pagnamenta, C. Daumer-Haas, B. Fischer, J. Hecht, D. A. Keays, S. J. Knight, U. Kolsch, U. Kruger, S. Leiz, Y. Maeda, D. Mitchell, S. Mundlos, J. A. Phillips, 3rd, P. N. Robinson, U. Kini, J. C. Taylor, D. Horn, T. Kinoshita, and P. M. Krawitz. 2014. 'Mutations in PGAP3 impair GPI-anchor maturation, causing a subtype of hyperphosphatasia with mental retardation', *Am J Hum Genet*, 94: 278–87.
45. Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. 'The Ensembl genome database project', *Nucleic Acids Res*, 30: 38–41.
46. International HapMap, Consortium. 2003. 'The International HapMap Project', *Nature*, 426: 789–96.
47. International HapMap, Consortium, D. M. Altshuler, R. A. Gibbs, L. Peltonen, P. E. Bonnen, P. I. de Bakker, P. Deloukas, S. B. Gabriel, R. Gwilliam, S. Hunt, M. Inouye, X. Jia, A. Palotie, M. Parkin, P. Whittaker, F. Yu, K. Chang, A. Hawes, L. R. Lewis, Y. Ren, D. Wheeler, R. A. Gibbs, D. M. Muzny, C. Barnes, K. Darvishi, M. Hurles, J. M. Korn, K. Kristiansson, C. Lee, S. A. McCarrol, J. Nemesh, E. Dermitzakis, A. Keinan, S. B. Montgomery, S. Pollack, A. L. Price, N. Soranzo, P. E. Bonnen, R. A. Gibbs, C. Gonzaga-Jauregui, A. Keinan, A. L. Price, F. Yu, V. Anttila, W. Brodeur, M. J. Daly, S. Leslie, G. McVean, L. Moutsianas, H. Nguyen, S. F. Schaffner, Q. Zhang, M. J. Ghori, R. McGinnis,

W. McLaren, S. Pollack, A. L. Price, S. F. Schaffner, F. Takeuchi, S. R. Grossman, I. Shlyakhter, E. B. Hostetter, P. C. Sabeti, C. A. Adebamowo, M. W. Foster, D. R. Gordon, J. Licinio, M. C. Manca, P. A. Marshall, I. Matsuda, D. Ngare, V. O. Wang, D. Reddy, C. N. Rotimi, C. D. Royal, R. R. Sharp, C. Zeng, L. D. Brooks, and J. E. McEwen. 2010. 'Integrating common and rare genetic variation in diverse human populations', *Nature*, 467: 52–8.

48. Jarvik G. P., and B. L Browning. 2016. 'Consideration of cosegregation in the pathogenicity classification of genomic variants', *Am J Hum Genet.,* 98(6): 1077–81.
49. Jeroncic, A., Y. Memari, G. R. Ritchie, A. E. Hendricks, A. Kolb-Kokocinski, A. Matchan, V. Vitart, C. Hayward, I. Kolcic, D. Glodzik, A. F. Wright, I. Rudan, H. Campbell, R. Durbin, O. Polasek, E. Zeggini, and V. Boraska Perica. 2016. 'Whole-exome sequencing in an isolated population from the Dalmatian island of Vis', *Eur J Hum Genet*. doi: 10.1038/ejhg 2016.23.
50. Ju, Y. S., J. I. Kim, S. Kim, D. Hong, H. Park, J. Y. Shin, S. Lee, W. C. Lee, S. Kim, S. B. Yu, S. S. Park, S. H. Seo, J. Y. Yun, H. J. Kim, D. S. Lee, M. Yavartanoo, H. P. Kang, O. Gokcumen, D. R. Govindaraju, J. H. Jung, H. Chong, K. S. Yang, H. Kim, C. Lee, and J. S. Seo. 2011. 'Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals', *Nat Genet*, 43: 745–52.
51. Karki, R., D. Pandya, R. C. Elston, and C. Ferlini. 2015. 'Defining "mutation" and "polymorphism" in the era of personal genomics', *BMC Med Genomics*, 8: 37.
52. Karow, J. 2015.’NIPT outperforms standard screening for T21 but false positives call for caution, NEJM studies find,’ *Genome Web*, 27 Apr. 2015.
53. Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, W. J. Kent, and Cruz University of California Santa. 2003. 'The UCSC Genome Browser Database', *Nucleic Acids Res*, 31: 51–4.
54. Kim, J. I., Y. S. Ju, H. Park, S. Kim, S. Lee, J. H. Yi, J. Mudge, N. A. Miller, D. Hong, C. J. Bell, H. S. Kim, I. S. Chung, W. C. Lee, J. S. Lee, S. H. Seo, J. Y. Yun, H. N. Woo, H. Lee, D. Suh, S. Lee, H. J. Kim, M. Yavartanoo, M. Kwak, Y. Zheng, M. K. Lee, H. Park, J. Y. Kim, O. Gokcumen, R. E. Mills, A. W. Zaranek, J. Thakuria, X. Wu, R. W. Kim, J. J. Huntley, S. Luo, G. P. Schroth, T. D. Wu, H. Kim, K. S. Yang, W. Y. Park, H. Kim, G. M. Church, C. Lee, S. F. Kingsmore, and J. S. Seo. 2009. 'A highly annotated whole-genome sequence of a Korean individual', *Nature*, 460: 1011–5.

55. Kircher M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J.Shendure. 2014. 'A general framework for estimating the relative pathogenicity of human genetic variants,' *Nature Genet*, 46, 310–15.

56. Kleinberger, J., K. A. Maloney, T. I. Pollin, and L. J. Jeng. 2016. 'An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants', *Genet Med*.
57. Kling, J. 2003. 'Ultrafast DNA sequencing', *Nat Biotechnol*, 21: 1425–7.
58. Knoppers, B. M., and K. Senecal, P. Borry, and D. Avard. 2014. 'Whole-genome sequencing in newborn screening programs', *Sci Transl Med*, 6: 229cm2.
59. Krawczak, M., E. V. Ball, I. Fenton, P. D. Stenson, S. Abeysinghe, N. Thomas, and D. N. Cooper. 2000. 'Human gene mutation database-a biomedical information and research resource', *Hum Mutat*, 15: 45–51.
60. Ku, C. S., and D. H. Roukos. 2013. 'From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine', *Expert Rev Med Devices*, 10: 1–6.
61. Kucinskas V., V. Jurgelevicius, L. Cimbalistiene, D. Jusciene, M. Smirnova, and D. Zamkauskiene. 1996. Management and results of mass neonatal screening in Lithuania, *Neonatal and perinatal screening: theasian pacific perspectives*. *Conference: 2nd Asian pacific regional meeting of the international-society-for-neonatal-screning*, 47–51
62. Kwan, A., and J. M. Puck. 2015. 'History and current status of newborn screening for severe combined immunodeficiency', *Semin Perinatol*, 39: 194–205.
63. Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher,

M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and Consortium International Human Genome Sequencing. 2001. 'Initial sequencing and analysis of the human genome', *Nature*, 409: 860–921.

64. Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. 2014. 'ClinVar: public archive of relationships among sequence variation and human phenotype', *Nucleic Acids Res*, 42: D980–5.

65. Lazarin, G. A., I. S. Haque, S. Nazareth, K. Iori, A. S. Patterson, J. L. Jacobson, J. R. Marshall, W. K. Seltzer, P. Patrizio, E. A. Evans, and B. S. Srinivasan. 2013. 'An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals', *Genet Med*, 15: 178–86.

66. Li, H. 2011. 'A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics*, 27: 2987–93.
67. Li, H., and R. Durbin. 2010. 'Fast and accurate long-read alignment with Burrows-Wheeler transform', *Bioinformatics*, 26: 589–95.
68. Li, Y., N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, H. Cao, T. Korneliussen, N. Grarup, Y. Guo, I. Hellman, X. Jin, Q. Li, J. Liu, X. Liu, T. Sparso, M. Tang, H. Wu, R. Wu, C. Yu, H. Zheng, A. Astrup, L. Bolund, J. Holmkvist, T. Jorgensen, K. Kristiansen, O. Schmitz, T. W. Schwartz, X. Zhang, R. Li, H. Yang, J. Wang, T. Hansen, O. Pedersen, R. Nielsen, and J. Wang. 2010. 'Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants', *Nat Genet*, 42: 969–72.
69. Lim, E. T., P. Wurtz, A. S. Havulinna, P. Palta, T. Tukiainen, K. Rehnstrom, T. Esko, R. Magi, M. Inouye, T. Lappalainen, Y. Chan, R. M. Salem, M. Lek, J. Flannick, X. Sim, A. Manning, C. Ladenvall, S. Bumpstead, E. Hamalainen, K. Aalto, M. Maksimow, M. Salmi, S. Blankenberg, D. Ardissino, S. Shah, B. Horne, R. McPherson, G. K. Hovingh, M. P. Reilly, H. Watkins, A. Goel, M. Farrall, D. Girelli, A. P. Reiner, N. O. Stitziel, S. Kathiresan, S. Gabriel, J. C. Barrett, T. Lehtimaki, M. Laakso, L. Groop, J. Kaprio, M. Perola, M. I. McCarthy, M. Boehnke, D. M. Altshuler, C. M. Lindgren, J. N. Hirschhorn, A. Metspalu, N. B. Freimer, T. Zeller, S. Jalkanen, S. Koskinen, O. Raitakari, R. Durbin, D. G. MacArthur, V. Salomaa, S. Ripatti, M. J. Daly, A. Palotie, and Project Sequencing Initiative Suomi. 2014. 'Distribution and medical impact of loss-of-function variants in the Finnish founder population', *PLoS Genet*, 10: e1004494.
70. MacArthur, D. G., and C. Tyler-Smith. 2010. 'Loss-of-function variants in the genomes of healthy humans', *Hum Mol Genet*, 19: R125–30.
71. MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, Consortium Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. 2012. 'A systematic survey of loss-of-function variants in human protein-coding genes', *Science*, 335: 823–8.

72. MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. 2014. 'Guidelines for investigating causality of sequence variants in human disease', *Nature*, 508: 469–76.
73. Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit, and N. Jabado. 2011. 'What can exome sequencing do for you?', *J Med Genet*, 48: 580–9.
74. Malapelle, U., E. Vigliar, R. Sgariglia, C. Bellevicine, L. Colarossi, D. Vitale, P. Pallante, and G. Troncone. 2015. 'Ion Torrent next-generation sequencing for routine identification of clinically relevant mutations in colorectal cancer patients', *J Clin Pathol*, 68: 64–8.
75. Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. 'Finding the missing heritability of complex diseases', *Nature*, 461: 747–53.
76. Mardis, E. R. 2008. 'Next-generation DNA sequencing methods', *Annu Rev Genomics Hum Genet*, 9: 387–402.
77. Maxam, A. M., and W. Gilbert. 1977. 'A new method for sequencing DNA', *Proc Natl Acad Sci U S A*, 74: 560–4.
78. McKernan, K. J., H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, J. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De La Vega, and A. P. Blanchard. 2009. 'Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding', *Genome Res*, 19: 1527–41.
79. Meade, C., and N. F. Bonhomme. 2014. 'Newborn screening: adapting to advancements in whole-genome sequencing', *Genet Test Mol Biomarkers*, 18: 597–8.
80. Mikštienė, V., A. Jakaitienė, J. Byčkova, E. Gradauskienė, E. Preikšaitienė, B. Burnytė, B. Tumienė, A. Matulevičienė, L. Ambrozaitytė, I. Uktverytė, I.

Domarkienė, T. Rančelis, L. Cimbalistienė, E. Lesinskas, V. Kučinskas, and A. Utkus. 2016.'The high frequency of *GJB2* gene mutation c.313_326del14 suggests its possible origin in ancestors of Lithuanian population'. *BMC Genetics*, 19, 17(1): 45.

81. Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad. 2010. 'Exome sequencing identifies the cause of a mendelian disorder', *Nat Genet*, 42: 30–5.
82. Pang, A. W., J. R. MacDonald, D. Pinto, J. Wei, M. A. Rafiq, D. F. Conrad, H. Park, M. E. Hurles, C. Lee, J. C. Venter, E. F. Kirkness, S. Levy, L. Feuk, and S. W. Scherer. 2010. 'Towards a comprehensive structural variation map of an individual human genome', *Genome Biol*, 11: R52.
83. Pedersen, P., and N. Milman. 2009. 'Genetic screening for HFE hemochromatosis in 6,020 Danish men: penetrance of C282Y, H63D, and S65C variants', *Ann Hematol.*, 88:775–84.
84. Pierce, S. B., T. Walsh, K. M. Chisholm, M. K. Lee, A. M. Thornton, A. Fiumara, J. M. Opitz, E. Levy-Lahad, R. E. Klevit, and M. C. King. 2010. 'Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome', *Am J Hum Genet*, 87: 282–8.
85. Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. 2010. 'Detection of nonneutral substitution rates on mammalian phylogenies', *Genome Res*, 20: 110–21.
86. Pranckeviciene, E., T. Rancelis, A. Pranculis, and V. Kucinskas. 2015. 'Challenges in exome analysis by LifeScope and its alternative computational pipelines', *BMC Res Notes*, 8: 421.
87. Prober, J. M., G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister. 1987. 'A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides', *Science*, 238: 336–41.
88. Pruitt, K. D., T. Tatusova, and D. R. Maglott. 2007. 'NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins', *Nucleic Acids Res*, 35: D61–5.
89. Rančelis,T., E. Pranckevičienė, and V. Kučinskas. Next-generation whole-exome sequencing contribution to identification of rare autosomal recessive diseases. 2013. Acta Medica Lituanica, 20 (1):43–51.
90. Rančelis,T., E. Pranckevičienė, and V. Kučinskas. „Anotaciniai įrankiai ir kompiuterinės programos genomo/egzomo duomenų analizei“. 2013. *Laboratorinė medicina*, 15: 206–12.

91. Reinstein, E. 2015. 'Challenges of using next generation sequencing in newborn screening', *Genet Res (Camb)*, 97: e21.
92. Richards, S., N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, and Acmg Laboratory Quality Assurance Committee. 2015. 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genet Med*, 17: 405–24.
93. Sanger, F., J. E. Donelson, A. R. Coulson, H. Kossel, and D. Fischer. 1973. 'Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA', *Proc Natl Acad Sci U S A*, 70: 1209–13.
94. Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. 2005. 'Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes', *Genome Res*, 15: 1034–50.
95. Sim, N. L., P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng. 2012. 'SIFT web server: predicting effects of amino acid substitutions on proteins', *Nucleic Acids Res*, 40: W452–7.
96. Siqueira, J. F., Jr., A. F. Fouad, and I. N. Rocas. 2012. 'Pyrosequencing as a tool for better understanding of human microbiomes', *J Oral Microbiol*, 4.
97. Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood. 1986. 'Fluorescence detection in automated DNA sequence analysis', *Nature*, 321: 674–9.
98. Srour, M., F. F. Hamdan, D. McKnight, E. Davis, H. Mandel, J. Schwartzentruber, B. Martin, L. Patry, C. Nassif, A. Dionne-Laporte, L. H. Ospina, E. Lemyre, C. Massicotte, R. Laframboise, B. Maranda, D. Labuda, J. C. Decarie, F. Rypens, D. Goldsher, C. Fallet-Bianco, J. F. Soucy, A. M. Laberge, C. Maftei, Consortium Care4Rare Canada, K. Boycott, B. Brais, R. M. Boucher, G. A. Rouleau, N. Katsanis, J. Majewski, O. Elpeleg, M. K. Kukolich, S. Shalev, and J. L. Michaud. 2015. 'Joubert Syndrome in French Canadians and Identification of Mutations in CEP104', *Am J Hum Genet*, 97: 744–53.
99. Strom, S. P., Y. Q. Gao, A. Martinez, C. Ortube, Z. Chen, S. F. Nelson, S. Nusinowitz, D. B. Farber, and M. B. Gorin. 2012. 'Molecular diagnosis of putative Stargardt Disease probands by exome sequencing', *BMC Med Genet*, 13: 67.
100. Tattini, L., R.D'Aurizio, and A. Magi. 2015. 'Detection of genomic structural variants from next-generation sequencing data', *Front Bioengineering Biotechn*, 3: 92.

101. Tarraga, J., A. Gallego, V. Arnau, I. Medina, and J. Dopazo. 2016. 'HPG pore: an efficient and scalable framework for nanopore sequencing data', *BMC Bioinformatics*, 17: 107.
102. Teer, J. K., and J. C. Mullikin. 2010. 'Exome sequencing: the sweet spot before whole genomes', *Hum Mol Genet*, 19: R145–51.
103. Thayer, A. M. 2014. 'Next-Gen Sequencing Is A Numbers Game',*Chemical and Engineering News* 33:11–15http://cen.acs.org/articles/92/i33/Next-Gen-Sequencing-Numbers-Game.html
104. Thompson, J. F., and K. E. Steinmann. 2010. 'Single molecule sequencing with a HeliScope genetic analysis system', *Curr Protoc Mol Biol*, Chapter 7: Unit7 10.
105. Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov. 2013. 'Integrative Genomics Viewer (*IGV*): high-performance genomics data visualization and exploration', *Brief Bioinform*, 14: 178–92.
106. Uktverytė, I., R. Meškienė, L. Ambrozaitytė, I. Domarkienė, A. Pranculis, N. Burokienė, A. Coj, A. Mažeikienė, V. Kasiulevičius, Z. A. Kučinskienė, and V. Kučinskas. 2013. LITGEN – revealing genetic structure of the population of Lithuania. European Journal of Human Genetics: European Human Genetics Conference 2013, Paris, France, June 8–11.
107. Uktverytė, I. 2014. „Lietuvos etnolingvistinių grupių genetinės struktūros analizė remiantis informatyviais genomo žymenimis“: Daktaro dis. biomedicinos mokslai: medicina (06 B).
108. Veiga-da-Cunha, M., N. M. Verhoeven-Duif, T. J. de Koning, M. Duran, B. Dorland, E. and Van Schaftingen 2013. 'Mutations in the AGXT2L2 gene cause phosphohydroxylysinuria', *J Inherit Metab Dis.*, 36: 961–6.
109. Voelkerding K. V., S. A. Dames, and J. D. Durtschi. 2009. 'Next-generation sequencing: from basic research to diagnostics', *Clin Chem.*, 55 (4):641–58.
110. Wang, Q., Q. Lu, and H. Zhao. 2015. 'A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing', *Front Genet*, 6: 149
111. Wu, R., and E. Taylor. 1971. 'Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA', *J Mol Biol*, 57: 491–511.
112. Xue, Y., Y. Chen, Q. Ayub, N. Huang, E. V. Ball, M. Mort, A. D. Phillips, K. Shaw, P. D. Stenson, D. N. Cooper, C. Tyler-Smith, and Consortium Genomes Project. 2012. 'Deleterious- and disease-allele prevalence in healthy individuals:

insights from current predictions, mutation databases, and population-scale resequencing', *Am J Hum Genet*, 91: 1022–32.

113. Yang, H., and K. Wang. 2015. 'Genomic variant annotation and prioritization with *ANNOVAR* and w*ANNOVAR*', *Nat Protoc*, 10: 1556–66.

114. Ye, M., K. M. Berry-Wynne, M. Asai-Coakwell, P. Sundaresan, T. Footz, C. R. French, M. Abitbol, V. C. Fleisch, N. Corbett, W. T. Allison, G. Drummond, M. A. Walter, T. M. Underhill, A. J. Waskiewicz, and O. J. Lehmann. 2010. 'Mutation of the bone morphogenetic protein GDF3 causes ocular and skeletal anomalies', *Hum Mol Genet*, 19:287–98.

115. Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. Cuo, J. E. Pool, X. Xu, H. Jiang, N. Vinckenbosch, T. S. Korneliussen, H. Zheng, T. Liu, W. He, K. Li, R. Luo, X. Nie, H. Wu, M. Zhao, H. Cao, J. Zou, Y. Shan, S. Li, Q. Yang, Asan, P. Ni, G. Tian, J. Xu, X. Liu, T. Jiang, R. Wu, G. Zhou, M. Tang, J. Qin, T. Wang, S. Feng, G. Li, Huasang, J. Luosang, W. Wang, F. Chen, Y. Wang, X. Zheng, Z. Li, Z. Bianba, G. Yang, X. Wang, S. Tang, G. Gao, Y. Chen, Z. Luo, L. Gusang, Z. Cao, Q. Zhang, W. Ouyang, X. Ren, H. Liang, H. Zheng, Y. Huang, J. Li, L. Bolund, K. Kristiansen, Y. Li, Y. Zhang, X. Zhang, R. Li, S. Li, H. Yang, R. Nielsen, J. Wang, and J. Wang. 2010. 'Sequencing of 50 human exomes reveals adaptation to high altitude', *Science*, 329: 75–8.

116. Zuchner, S., J. Dallman, R. Wen, G. Beecham, A. Naj, A. Farooq, M. A. Kohli, P. L. Whitehead, W. Hulme, I. Konidari, Y. J. Edwards, G. Cai, I. Peter, D. Seo, J. D. Buxbaum, J. L. Haines, S. Blanton, J. Young, E. Alfonso, J. M. Vance, B. L. Lam, and M. A. Pericak-Vance. 2011. 'Whole-exome sequencing links a variant in DHDDS to retinitis pigmentosa', *Am J Hum Genet*, 88: 201–6.

PADĖKA

Visų pirma norėčiau padėkoti savo moksliniam vadovui **akad. prof. habil. dr. Vaidučiui Kučinskui** už suteiktą galimybę atlikti šį darbą, pastabas, patarimus, kantrybę, rūpestį, už suteiktą galimybę stažuotis Estijoje ir Šveicarijoje ir raginimą laiku pabaigti rašyti disertaciją. Tai pat norėčiau padėkoti savo moksliniam konsultantui **prof. (HP) dr. Loretai Cimbalistienei** už pastabas ir konsultacijas klinikiniais klausimais**.**

Prie šios disertacijos atsiradimo prisidėjo daug kolegų, kuriems esu labai dėkingas. Ypač norėčiau padėkoti LITGEN projekto, kurio duomenys panaudoti šioje disertacijoje, vadovui **akad. prof. habil. dr. Vaidučiui Kučinskui** ir visam vykdytojų kolektyvui. Labai ačiū **dr. Laimai Ambrozaitytei**, **dr. Ingridai Domarkienei**, **dr. Ingridai Kavaliauskienei**, **dr. Neringai Burokienei**, **Raimondai Meškienei**, **Aidui Pranculiui** už darbe naudotų mėginių surinkimą.

Darbo pagrindas lietuvių egzomų duomenys, kurių nebūtų be tinkamai atlikto laboratorinio darbo – DNR išskyrimo, sekoskaitos. Tad norėčiau padėkoti **Daivai Kazlauskaitei**, **Dalytei Pliaugo**, **dr. Ingridai Kavaliauskienei**, **dr. Laimai Ambrozaitytei**, **dr. Ingridai Domarkienei ir Kristinai Aleknavičienei**.

Svarbi darbo dalis yra naujos kartos sekoskaitos duomenų analizės algoritmas. Nuoširdžiai dėkoju **dr. Erinijai Pranckevičienei** už pamokymus, suteiktas žinias, bendrą darbą ir tai, kad mane, genetiką, sugebėjo sudominti bioinformatika, programavimu. Dėkoju **Aidui Pranculiui** už bendrą darbą prižiūrint ŽMGK turimą serverį ir kartu vykdytą analizę. Dėkoju **Justui Arasimavičiui** už patarimus statistiniais klausimais. Taip pat ačiū **Violetai Mikštienei** už bendrą darbą ir suteiktą galimybę giliau išanalizuoti atskirą negalią tarp lietuvių.

Už pagalbą, vertingus patarimus, pastabas rašant disertaciją norėčiau padėkoti **akad**. **prof. habil. dr. Zitai Aušrelei Kučinskienei**, **prof. dr. Audronei Jakaitienei**, **dr. Aušrai Morkūnienei**, **dr. Aušrai Matulevičienei, dr. Almai Molytei**, **dr. Laimai Ambrozaitytei, ir prof. (HP) dr. Algirdui Utkui**.

Galiausiai noriu padėkoti visam VU MF ŽMGK kolektyvui ir artimiesiems už palaikymą ir pagalbą.

PRIEDAI

1 priedas. VCF failų pagrindiniai komponentai

- **REF** –variantas, esantis referentiniame genome.
- **ALT** – variantas, kuris skiriasi nuo esančio referentiniame genome.
- **QUAL –** *Phred* tikimybinė vertė, parodanti, kad tam tikra pakaita realiai egzistuoja tam tikroje genomo vietoje. Kadangi *Phred* įvertis yra –10 * log(1-p), tai 10 vertė reiškia, kad klaidos tikimybė yra 1 iš 10, o 30 vertė rodo, kad klaidos tikimybė yra 1 iš 1000.
- **FILTER –** remiasi statistiškai nepriklausomais įverčiais siekiant nustatyti, kurie besiskiriantys nuo referentinio genomo variantai yra sekoskaitos klaidos, o kurie yra realūs VNP. „PASS“ reiškia, kad variantas perėjo kokybės patikrinimą.
- **GT – tiriamojo genotipas**.

  0/0 – homozigotinis kaip referentiniame genome

  0/1 – heterozigotinis REF/ALT

  1/1 – homozigotinis kaip pakaita

- **GQ ir PL –** genotipo kokybinė vertė. Pateikiami *Phred* tikėtumo įverčiai, kurie nurodo paklaidos tikimybės lygį ir tai, kad genotipas, pateiktas „GT“ duomenyse, yra teisingas.
- **AD –** nurodo DNR fragmentų skaičių. Pateikiamos dvi vertės: kiek iš viso yra DNR fragmentų, kuriuose genomo variantas yra toks pat kaip ir referentiniame genome, ir kiek iš viso yra DNR fragmentų, kuriuose genomo variantas skiriasi nuo referentinio genomo.
- **DP –** nurodo besiskiriančio nuo referentinio genomo padengimą.

2 priedas. *Lifescope* programos parametrai

Parametrai	Įverčiai
SAET	
Update Quality Values	TRUE
Trusted Quality Values	25
Support Votes	3
Maximum Correction per Read	0
K-mer size	0
Genome Length	30 000 000
Position of Error Inflation Point	0
Disable Random Sampling for Large Data	FALSE
Trusted Frequency	0
On Target Ratio	0,5
Number of Recursive Runs	1
Fragment Mapping	
Mapping QV Threshold	0
Create Unmapped BAM Files	FALSE
Reference Weight	8
Second Map Gapped Algorithm	GLOBAL
Base Quality Filter Threshold	10
Map in Base Space	FALSE
Add Color Sequence	TRUE
BAM soft clip	FALSE
BAM Stats	
Whether to combine Data from Both the Strands for Coverage in WIG format	1
Bin Size for Coverage in WIG file format	100
Maximum insert size	100 000
Insert size bin	100
Primary Alignments only for Coverage in WIG file format	1
Maximum Coverage	500 000
Small Indel	
Detail Level	3
Number Alignment per Pileup	1 000
Random Seed	94 404
Min Num Evid	2
Max Num Evid	–1
ConsGroup	1
Max Reported Alignments	–1

2 priedo tęsinys

Min Mapping Quality (MAPQ)	8
Min Best Mapping Quality	10
Min Anchor Mapping Quality	–1
SNP Finding	
Call Stringency	Medium
Skip High Coverage Positions (Het)	FALSE
Minimum Mapping QV	8
Detect Adjacent SNPs	FALSE
Polymorphism rate	0,001
Include Reads with unmapped Mate	FALSE
Exclude Reads with Indels	TRUE
Require Only Uniquely Mapped Reads	FALSE
Minimum Ratio of the Filtered Reads and Raw reads	0
Require alleles to be present in both strands	FALSE
Minimum Base QV a Read for a Position	28
Minimum Color QV of a Read for a Position	7
Min Base QV of the non-reference allele of the position	28
Minimum Unique Start Positions of Less Common Allele	0
The Less Common Allele on Both Strands	FALSE
Minimum Allele Ratio (Het)	0,15
Minimum Coverage (Het)	2
Minimum Unique Start Positions (Het)	2
Minimum non-Reference Color QV (Het)	7
Minimum non-Reference Base QV (Het)	28
Minimum Ratio of Valid Reads (Het)	0,65
Minimum Valid Tricolor Count (Het)	2
Minimum Coverage (Hom)	1

3 priedas. Analizės algoritmo komandos naudojant *GATK* programinį paketą

Komandos uždavinys	Naudota komanda
Duplikuotų DNR fragmentų pažymėjimas, kad analizuojant būtų vertinamas tik vienas iš jų	java -jar MarkDuplicates.jar INPUT = your_bam_file OUTPUT = step1.bam METRICS_FILE = Fmetrics_step1.bam ASSUME_SORTED = true
Metainformacija pakeičiama naujais įrašais, kurie reikalingi dirbant su *GATK* programa	java -jar AddOrReplaceReadGroups.jar INPUT= step1.bam OUTPUT = step2.bam RGID= Read_Group ID RGLB = Read_Group_Library RGPL= platform RGPU = platform_unit RGSM= sample_name RGDS = Read_Group_Description RGDT = Read_Group_Run_Date
Etapas, skirtas suvienodinti referentinių genomų skirtumus	java -jar ReorderSam.jar INPUT = step2.bam OUTPUT = step3.bam REFERENCE = ucsc.hg19.fasta
Duomenų rūšiavimas pagal chromosominę poziciją	java -jar SortSam.jar INPUT = step3.bam OUTPUT = step4.bam SORT_ORDER = coordinate
Indeksų sukūrimas genominėms pozicijoms	java -jar BuildBamIndex.jar INPUT= step4.bam
Iškritų / intarpų regionams atliekamas perlygiavimas su referentiniu genomu. I etapas. Sudaromas sąrašas regionų, kuriuos reikia perlygiuoti.	java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R ucsc.hg19.fasta -S STRICT -I step4.bam -o indels.intervals -allowPotentiallyMisencodedQuals
Iškritų / intarpų regionams atliekamas perlygiavimas su referentiniu genomu. II etapas. Atliekamas pats perlygiavimo procesas	java -jar GenomeAnalysisTK.jar -T IndelRealigner -R ucsc.hg19.fasta -S STRICT -I step4.bam -targetIntervals indels.intervals -o step5.bam -known Mills_and_1000G_gold_standard.indels.hg19.vcf -known 1000G_phase1.indels.hg19.vcf -allowPotentiallyMisencodedQuals
Duomenų rūšiavimas pagal chromosominę poziciją	java -jar SortSam.jar INPUT = step5.bam OUTPUT = step6.bam SORT_ORDER = coordinate
Indeksų sukūrimas genominėms pozicijoms	java -jar BuildBamIndex.jar INPUT = step6.bam
Nukleotidų kokybės įverčio kalibracija. I etapas. Prie tyrimo duomenų pritaikomi kokybės įverčiai, naudojami *GATK* programoje	java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -I step6.bam -R ucsc.hg19.fasta -S STRICT -knownSites dbsnp_138.hg19.vcf -o recal.grp –covariate QualityScoreCovariate –covariate ReadGroupCovariate –covariate ContextCovariate –covariate CycleCovariate –*SOLiD*_nocall_strategy PURGE_READ –*SOLiD*_recal_mode SET_Q_ZERO_BASE_N -allowPotentiallyMisencodedQuals

3 priedo tęsinys

Nukleotidų kokybės įverčio kalibracija. II etapas. Nauji kokybės įverčiai pridedami prie tyrimo duomenų	java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -I step6.bam -R ucsc.hg19.fasta -S STRICT -knownSites dbsnp_138.hg19.vcf -o recal.grp –covariate QualityScoreCovariate –covariate ReadGroupCovariate –covariate ContextCovariate –covariate CycleCovariate –*SOLiD*_nocall_strategy PURGE_READ –*SOLiD*_recal_mode SET_Q_ZERO_BASE_N -allowPotentiallyMisencodedQuals
Teisingų genomo variantų atranka	java -jar GenomeAnalysisTK.jar -R ucsc.hg19.fasta -T HaplotypeCaller -I step7.bam -S STRICT –dbsnp dbsnp_138.hg19.vcf -minPruning 3 -o step8.vcf -stand_call_conf 50 -stand_emit_conf 30
Iš rezultatų atrenkami VNP	java -jar GenomeAnalysisTK.jar -R ucsc.hg19.fasta -T SelectVariants –variant step8.vcf -o step9_SNP.vcf -selectType SNP -S STRICT
Sukuriamas VNP kalibracijos modelis, kuris skirtas pagerinti tinkamai atrinktų variantų rezultatus	java -jar GenomeAnalysisTK.jar -T VariantRecalibrator –input step9_SNP.vcf -R ucsc.hg19.fasta -S STRICT -resource:1000G,known = false,training = true,truth = false,prior = 10 1000G_phase1.snps.high_confidence.hg19.vcf -resource:hapmap, known =f alse, training = true, truth = true, prior = 15.0 hapmap_3.3.hg19.vcf -resource:omni, known=false, training = true, truth = true, prior = 12.0 1000G_omni2.5.hg19.vcf -resource:dbsnp, known = true, training = false, truth = false, prior = 2.0 dbsnp_138.hg19.vcf -an QD -an MQRankSum -an ReadPosRankSum -an FS -an MQ –maxGaussians 4 -mode SNP -recalFile recal -tranchesFile tranches
VNP kalibracijos modelis pritaikomas tiriamojo duomenims	java -jar GenomeAnalysisTK.jar -R ucsc.hg19.fasta -T ApplyRecalibration -S STRICT –input step9_SNP.vcf -ts_filter_level 99.5 -mode SNP -tranchesFile tranches -recalFile recal -o step10_final.vcf

4 priedas. *ClinVar* šaltiniai, kurie į duomenų bazę pateikė daugiau nei 30 genomo variantų, turinčių klinikinę interpretaciją

Pateikėjas	Iš viso pateikta genomo variantų	Iš viso pateikta genomo variantų, turinčių klinikinę interpretaciją	Genų skaičius pateiktuose genomo variantuose
OMIM; Johns Hopkins University	26742	26741	4006
GeneDx	24693	23357	671
Laboratory for Molecular Medicine; Partners HealthCare Personalized Medicine	16430	16339	518
Ambry Genetics	16035	16035	317
Emory Genetics Laboratory; Emory University	16047	15982	1299
Genetic Services Laboratory; University of Chicago	10413	10413	818
Invitae	13161	8472	344
GeneReviews	5212	5196	626
ISCA site 1	5157	5157	24869
ISCA site 6	3673	3673	21226
ISCA site 4	2952	2952	21156
International Society for Gastrointestinal Hereditary Tumours	2368	2368	9
Biesecker Laboratory - ClinSeq Project; National Insittutes of Health, National Human Genome Research Institute	2221	2221	108
Sharing Clinical Reports Project	2145	2144	2
Breast Cancer Information Core (BIC)	2001	1976	1
Breast Cancer Information Core (BIC); NIH	1792	1758	1
Cardiovascular Biomedical Research Unit; Royal Brompton and Harefield NHS Foundation Trust	1521	1520	14
LabCorp; Laboratory Corporation of America	1391	1363	162
Blueprint Genetics	1219	1219	151

Counsy	1136	1136	107
Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA); c/o QIMR Berghofer Medical Research Institute	1030	1030	3
Systems Biology Platform; Zhejiang California International NanoSystems Institute	1024	1024	1
RettBASE	1097	973	5
Genetics Diagnostic Laboratory; Children's Hospital of Eastern Ontario	958	958	21
University of Washington Department of Laboratory Medicine; University of Washington	921	921	56
Division of Genomic Diagnostics; The Children's Hospital of Philadelphia	860	860	490
Juha Muilu Group; Institute for Molecular Medicine Finland (FIMM)	840	840	43
CSER _CC_NCGL	699	699	91
ALPORT Syndrome and COL4A5; ARUP Laboratories	629	629	2
ISCA site 8	537	537	10616
Stanford Center for Inherited Cardiovascular Disease; Stanford University	534	534	87
Developmental Genetics Unit; King Faisal Specialist Hospital & Research Centre	533	533	391
Collagen Diagnostic Laboratory; University of Washington Medical Center	411	411	2
Science for Life laboratory; Karolinska Institutet	407	407	423
Mayo Clinic Genetic Testing Laboratories; Mayo Clinic	403	403	6
Richard Lifton Laboratory; Yale University	400	390	285

4 priedo 2 tęsinys

ISCA site 17	359	359	11091
ISCA site 7	326	326	497
Michigan Medical Genetics Laboratories; University of Michigan	321	321	2
Clinical Biochemistry Laboratory; UCL Hospitals	318	318	30
GenMed Metabolism Lab	317	317	1
Pharmacogenetics and Pharmacogenomics Knowledge Base; Stanford University	301	301	110
Baylor Miraca Genetics Laboratories; Baylor College of Medicine	293	293	175
UW Hindbrain Malformation Research Program; University of Washington	282	282	25
Athena Diagnostics Inc; Quest Diagnostics	281	281	27
Galactose-1-Phosphate Uridyl Transferase (GALT); ARUP Laboratories	264	264	1
Next Generation Diagnostics; Novartis Institutes for BioMedical Research, Inc.	257	257	2281
CFTR2; Johns Hopkins University	250	250	1
ISCA site 2	248	248	2980
ISCA site 13	247	247	10793
PALB2 database	242	242	2
Lupski Lab, Baylor-Hopkins CMG; Baylor College of Medicine	241	241	98
Pathway Genomics	238	238	18
Martin Pollak; Beth Israel Deaconess Medical Center	234	234	41
ISCA site 15	229	229	5908
Center for Bioinformatics; Peking University	223	223	2
UCLA Clinical Genomics Center; UCLA	211	211	203
ISCA site 14	181	181	5584

4 priedo 3 tęsinys

Department of Ophthalmology and Visual Sciences; Kyoto University	192	171	76
Biotinidase Deficiency and BTD; ARUP Laboratories	168	168	2
Multiple Endocrine Neoplasia type 2 (MEN2) and RET; ARUP Laboratories	155	155	1
Cardiovascular Genetics Laboratory; PathWest Laboratory Medicine WA	153	153	1
ARUP Laboratories, Molecular Genetics	150	150	5
ClinVar; National Center for Biotechnology Information	222	141	64
Institute of Molecular, Cell and Systems Biology; University of Glasgow	137	137	1
Laboratory for Medical Science Mathematics; RIKEN	121	121	100
Vantari Genetics	116	116	40
HudsonAlpha Institute for Biotechnology	114	114	86
Génétique et pathophysiologie de maladies neurodéveloppementales et épileptogènes; Institut de génétique et de biologie moléculaire et cellulaire (IGBMC)	109	109	79
Mendelics; Mendelics Analise Genomica	106	106	99
Claritas Genomics	99	99	7
Cancer Genetics Laboratory; Peter MacCallum Cancer Centre	96	96	1
Albrecht-Kossel-Institute; Medical University Rostock	92	92	2
Laboratory of Genetics and Molecular Cardiology; University of São Paulo	91	91	21

4 priedo 4 tęsinys

Medical Genomics Laboratory; Department of Genetics, University of Alabama at Birmingham	88	88	4
Juvenile Polyposis Syndrome and SMAD4; ARUP Laboratories	88	88	1
Greenwood Genetic Center Diagnostic Laboratories; Greenwood Genetic Center	81	81	20
Division of Human Genetics; Children's Hospital of Philadelphia	80	79	62
SNPedia; River Road Bio	113	74	22
Agnes Ginges Centre for Molecular Cardiology; Centenary Institute	70	70	16
Genetic Diagnostic Laboratory; University of Pennsylvania School of Medicine	68	68	1
Cytogenetics and Genomics Laboratory; Univeristy of Washington	67	67	1243
Paul Sabatier University EA-4555; Paul Sabatier University	63	63	50
Prostate Cancer Research Center; Institute of Biosciences and Medical Technology, University of Tampere, Finland	61	61	63
Department of Zoology; Govt. MVM College	58	58	3
Molecular Genetics Laboratory; Children's Mercy Hospital and Clinics	57	57	2
Eye Genetics Research Group; Children's Medical Research Institute	56	56	27
Genomic Research Center; Shahid Beheshti University of Medical Sciences	97	50	35

4 priedo 5 tęsinys

Inserm U 954 Nutrition-Génétique-Exposition aux risques environnementaux	71	48	17
Evolutionary and Medical Genetics Lab; Centre for Cellular and Molecular Biology	46	46	3
Division of Human Genetics Innsbruck; Medical University Innsbruck	44	44	2
University Children's Hospital; University of Zurich	42	42	1
Baylor-Hopkins Center for Mendelian Genomics; Johns Hopkins University	42	42	32
Courtagen Diagnostics Laboratory; Courtagen Life Sciences	42	42	34
Molecular Diagnostics Lab; Nemours A. I. duPont Hospital for Children	42	42	9
Knight Diagnostic Laboratories; Oregon Health and Sciences University	41	41	41
Centre for Genomic Medicine, Manchester; Central Manchester University Hospitals	39	39	20
Northcott Neuroscience Laboratory; ANZAC Research Institute	37	37	15
University of Washington Center for Mendelian Genomics; University of Washington	35	35	22
Immunobiology Lab; University of Kashmir	35	35	3
Research Lab; National Institute of Public Health	34	34	3
Willoughby Group	33	33	1
Center of Medical Genetics; Sir Ganga Ram Hospital	32	32	8
LISIN; Facultad de Ciencias Exactas, Universidad Nacional de La Plata	32	32	1

5 priedas. Fišerio tiksliojo ir Pirsono chi kvadratų kriterijų P vertės, nurodančios palyginimo tarp lietuvių bendrosios populiacijos ir 1000G projekto bei ExAC duomenų bazės duomenų reikšmingumą

Liga	ExAC ALL chi	ExAC ALL Fisher	1000G ALL chi	1000G ALL Fisher	1000G EUR chi	1000G EUR Fisher
Štargarto geltonosios dėmės distrofija, paveldima tinklainės distrofija	5,76 x 10^{-5}	3,462237 x 10^{-3}	1,4045 x 10^{-6}	9,075 x 10^{-4}	1,419 x 10^{-3}	8,006 x 10^{-3}
Pirminė abipusė mazginė antinksčių hiperplazija.	1,11 x 10^{-5}	3,263835 x 10^{-3}	3,1391 x 10^{-7}	0,0011	0,0224	0,0447
Acil-kofermento A dehidrogenazės stoka	2,87 x 10^{-8}	1,83305 x 10^{-5}	4,5823 x 10^{-18}	5,2791 x 10^{-9}	2,6454 x 10^{-4}	9,75 x 10^{-4}
Fosfohidroksilizinurija	2,89 x 10^{-14}	7,25637 x 10^{-6}	1,6865 x 10^{-25}	2,8235 x 10^{-8}	2,0602 x 10^{-5}	5,2093 x 10^{-4}
Polidaktilija	2,88 x 10^{-8}	1,57364 x 10^{-4}	2,1348 x 10^{-13}	5,0178 x 10^{-6}	6,4 x 10^{-4}	3,53 x 10^{-3}
Sunkaus kombinuoto imunodeficito liga	1,12 x 10^{-7}	8,34154 x 10^{-5}	4,0168 x 10^{-12}	1,9486 x 10^{-6}	0,0241	0,04919
Šeiminė medulinė skydliaukės karcinoma	1,17 x 10^{-9}	4,97744 x 10^{-4}	3,3295 x 10^{-19}	9,486 x 10^{-6}	1,1281 x 10^{-4}	3,0128 x 10^{-3}
***Usher* sindromas**	7,61 x 10^{-5}	5,571966 x 10^{-3}	1,3451 x 10^{-8}	5,2565 x 10^{-4}	9,8591 x 10^{-3}	0,0283
***Warsaw* sindromas**	0,0011	0,0135	9,0479 x 10^{-22}	2,9464 x 10^{-6}	1,0296 x 10^{-5}	8,3404 x 10^{-4}
Paveldėtoji sferocitozė	3,07 x 10^{-15}	8,9652 x 10^{-5}	5,6834 x 10^{-16}	2,7639 x 10^{-5}	8,2718 x 10^{-4}	7,9094 x 10^{-3}
Kepenų lipazės stoka	2,84 x 10^{-6}	2,325035 x 10^{-3}	1,7415 x 10^{-9}	3,4437 x 10^{-4}	0,0224	0,0447
Progresuojanti išorinė oftalmoplegija	1,73 x 10^{-5}	1,649749 x 10^{-3}	8,459 x 10^{-12}	2,3405 x 10^{-5}	0,0036	0,0113
Trimetilaminurija	4,6632 x 10^{-3}	2,2471 x 10^{-3}	2,1473 x 10^{-4}	3,7525 x 10^{-5}	0,0264	0,0296
Tulžies pūslės liga	0,0131	0,0074	5,3798 x 10^{-3}	2,3019 x 10^{-3}	0,0006	0,0001
Galimybė užuosti žolės kvapą	0,0018	0,001	2,6382 x 10^{-8}	8,93 x 10^{-10}	7,662 x 10^{-3}	7,765 x 10^{-3}
Įgimtas miastenijos sindromas	0,0013	6,201 x 10^{-4}	0,0349	0,0321	1,7077 x 10^{-6}	1,2823 x 10^{-7}
Adrenoleukodis-trofija	8,9747 x 10^{-6}	1,0222 x 10^{-4}	2,7274 x 10^{-22}	7,357 x 10^{-13}	0,044	0,0519
Paveldėtoji eritropoezinė porfirija	1,3527 x 10^{-3}	8,8063 x 10^{-4}	4,1846 x 10^{-22}	4,6123 x 10^{-24}	2,096 x 10^{-3}	1,773 x 10^{-3}
***Knobloch* sindromas**	6,7777 x 10^{-4}	2,329 x 10^{-4}	0,4844	0,6023	0,01	0.009
Šeiminė Viduržemio jūros karštligė	4,09 x 10^{-30}	1,2572 x 10^{-10}	2,1232 x 10-42	7,8166 x 10^{-14}	3,2363 x 10^{-9}	1,1564 x 10^{-6}

APIE AUTORIŲ

Vardas ir pavardė	Tautvydas Rančelis
Adresas	Žmogaus ir medicininės genetikos katedra Santariškių g. 2, LT-08661, Vilnius, Lietuva
Kontaktai	+37060176123, tautvydas.rancelis@mf.vu.lt
Gimimo data ir vieta	1987 m. lapkričio mėn. 23 d., Vilnius

DARBO PATIRTIS

Nuo 2012.06.27 Vilniaus universiteto Žmogaus ir medicininės genetikos jaunesnysis mokslo darbuotojas

IŠSILAVINIMAS

2006 baigė Vilnius Žemynos gimnaziją

2006 – 2010 Biologijos bakalauras, Vilniaus universitetas

2010 – 2012 Genetikos magistras, Vilniaus universitetas

2012 – 2016 Vilniaus universiteto Žmogaus ir medicininės genetikos katedros doktorantas

MOKSLINIAI INTERESAI

Žmogaus genetika, naujos kartos sekoskaita, didelės apimties duomenų analizė, anotacija, retos ligos

NARYSTĖ PROFESINĖJE ORGANIZACIJOJE

Europos Žmogaus genetikos draugijos narys

VYKDYTI PROJEKTAI

LITGEN (Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis) projekto vykdytojas

UNIGENE (Unikalūs genomo persitvarkymai esant įgimtiems nervų sistemos raidos sutrikimams: kilmė, genominiai mechanizmai, funkcinės ir klinikinės pasekmės) projekto vykdytojas

STAŽUOTĖS SUSIJUSIOS SU DISERTACIJOS TEMATIKA

Stažuotė į Estijos genomo centrą, Tartu universitete (kursai susiję su viso egzomo sekoskaita)

Stažuotė į Šveicarijos integratyvios genomikos centrą, Lozanos universitete (pirminė sekoskaitos duomenų analizė)