VILNIAUS UNIVERSITETO
MEDICINOS FAKULTETO
ŽMOGAUS IR MEDICININĖS GENETIKOS KATEDRA

MAGISTRO BAIGIAMASIS DARBAS

Naujos kartos sekoskaitos duomenų automatinio analizės algoritmo bei skirtingų praturtinimo sistemų įvertinimas, panaudojant Sanger sekoskaitą

Magistrantė
GABRIELĖ ŽUKAUSKAITĖ

Darbo vadovė
dr. LAIMA AMBROZAITYTĖ

Darbo konsultantas
dr. TAUTVYDAS RANČELIS

VU MF Žmogaus ir medicininės
genetikos katedros vedėjas
Prof. (HP) dr. ALGIRDAS UTKUS

leidžiama ginti

Darbo įteikimo data 2017 05 15
Registracijos Nr.

2017

Vilniaus universiteto studijuojančiojo, teikiančio baigiamąjį darbą,
GARANTIJA

Garantuoju, kad mano baigiamasis darbas yra parengtas sąžiningai ir savarankiškai, kitų asmenų indėlio į parengtą darbą nėra. Jokių įstatymų nenumatytų mokėjimų už šį darbą niekam nesu mokėjęs.
Šiame darbe tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos yra pažymėtos literatūros nuorodose.

Darbo autorė Gabrielė Žukauskaitė _____

(parašas)

Darbo vadovė dr. Laima Ambrozaitytė _____

(parašas)

Darbo konsultantas dr. Tautvydas Rančelis _____

(parašas)

TURINYS

SANTRUMPŲ SĄRAŠAS

ATP – adenozinotrifosfatas (angl. *adenosine triphosphate*)

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

ddNTP – dideoksiribonukleotidas (angl. *deoxyribonucleotide*)

DMSO – dimetilsulfoksidas (angl. *dimethyl sulfoxide*)

DNR – deoksiribonukleorūgštis (angl. *deoxyribonucleic acid*)

dNTP – deoksiribonukleotidas (angl. *deoxyribonucleotide*)

DP – genomo varianto padengimas (angl. *depth of coverage*)

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

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

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

RNR – ribonukleorūgštis (angl. *ribonucleic acid*)

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

VNP – vieno nukleotido polimorfizmas (angl. *Single Nucleotide Polymorphism*)

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

Genų santrumpos

ABR (angl. *active BCR-related*)

ANKRD36 (angl. *ankyrin repeat domain 36*)

ANKRD62 (angl. *ankyrin repeat domain 62*)

ARSD (angl. *arylsulfatase D*)

B4GALNT4 (angl. *beta-1,4-N-acetyl-galactosaminyl transferase 4*)

BAGE2 (angl. *B melanoma antigen family, member 2*)

BMS1 (angl. *BMS1 ribosome biogenesis factor*)

C1orf167 (angl. *chromosome 1 open reading frame 167*)

CCDC107 (angl. *coiled-coil domain containing 107*)

CFP (angl. *complement factor properdin*)

CIDEC (angl. *cell death-inducing DFFA-like effector c*)

CRACR2A (angl. *calcium release activated channel regulator 2A*)

CYFIP1 (angl. *cytoplasmic FMR1 interacting protein 1*)

DAP (angl. *death-associated protein*)

DHTKD1 (angl. *dehydrogenase E1 and transketolase domain containing 1*)

DISP2 (angl. *dispatched homolog 2*)

DMKN (angl. *dermokine*)

DSP (angl. *desmoplakin*)

EDDM3A (angl. *epididymal protein 3A*)

EFCAB5 (angl. *EF-hand calcium binding domain 5*)

EMILIN2 (angl. *elastin microfibril interfacer 2*)

ERCC6L (angl. *excision repair cross-complementation group 6-like*)

FAM105A (angl. *family with sequence similarity 105, member A*)

FAM200B (angl. *family with sequence similarity 200, member B*)

FGF23 (angl. *fibroblast growth factor 23*)

GAB4 (angl. *GRB2-associated binding protein family, member 4*)

GNG7 (angl. *guanine nucleotide binding protein (G protein), gamma 7*)

GPR146 (angl. *G protein-coupled receptor 146*)

GPRC5D (angl. *G protein-coupled receptor, class C, group 5, member D*)

HNRNPC (angl. heterogeneous nuclear ribonucleoprotein C)

KRT31 (angl. *keratin 31, type I*)

LATS (angl. *large tumor suppressor kinase 2*)

MDGA2 (angl. *MAM domain containing glycosylphosphatidylinositol anchor 2*)

MICAL3 (angl. *microtubule associated monooxygenase, calponin and LIM domain containing 3*)

MIPEP (angl. *mitochondrial intermediate peptidase*)

MYT1L (angl. *myelin transcription factor 1-like*)

NLRP6 (angl. *NLR family, pyrin domain containing 6*)

PERM1 (angl. *PPARGC1 and ESRR induced regulator, muscle 1*)

PHRF1 (angl. *PHD and ring finger domains 1*)

PLEKHG4B (angl. *pleckstrin homology domain containing, family G (with RhoGef domain) member 4B*)

POLRMT (angl. *polymerase (RNA) mitochondrial*)

RP1L1 (angl. *retinitis pigmentosa 1-like 1*)

SGK223 (angl. *Tyrosine-protein kinase SgK223*)

SIRPB1 (angl. *signal-regulatory protein beta 1*)

SKA3 (angl. *spindle and kinetochore associated complex subunit 3*)

SMARCA2 (angl. *SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2*)

SNRNP48 (angl. *small nuclear ribonucleoprotein 48kDa*)

SPATA13 (angl. *spermatogenesis associated 13*)

SPATA7 (angl. *spermatogenesis associated 7*)

SYCP2L (angl. *synaptonemal complex protein 2-like*)

TFF3 (angl. *trefoil factor 3 (intestinal)*)

TMEM184A (angl. *transmembrane protein 184A*)

TMPRSS15 (angl. *transmembrane protease, serine 15*)

UTY (angl. *ubiquitously transcribed tetratricopeptide repeat containing, Y-linked*)

WFDC10B (angl. *WAP four-disulfide core domain 10B*)

WHSC1 (angl. *Wolf-Hirschhorn syndrome candidate*)

XPC (angl. *xeroderma pigmentosum, complementation group C*)

ZCCHC7 (angl. *zinc finger, CCHC domain containing 7*)

ZNF469 (angl. *zinc finger protein 469*)

ZNF721 (angl. *zinc finger protein 721*)

ĮVADAS

Genomas – tai visas viengubas genetinės informacijos deoksiribonukleorūgšties (DNR) ar ribonukleorūgšties (RNR) molekulėje rinkinys, esantis vienoje organizmo ląstelėje (1).

Genome esantys pokyčiai sukuria genetinę įvairovę. Tokie pokyčiai gali neturėti įtakos organizmo sveikatos būklei ir būti paplitę daugiau kaip 1 % dažniu visame genome – jie vadinami polimorfizmais. Tačiau genomo variantai taip pat gali turėti įtakos sveikatai ar netgi būti patogeniški. Todėl genomo variantų tyrimai yra svarbūs tiek aiškinantis tarpopuliacinius skirtumus, tiek ieškant asociacijų su ligomis.

Vienas pagrindinių metodų, leidžiančių nustatyti genomo variantus bei jų įvairovę, yra DNR sekoskaita. Įvairiose gyvybės mokslų srityse vis dažniau naudojamas naujos kartos sekoskaitos (NKS) metodas, kuriuo galima nusekvenuoti koduojančią genomo dalį (egzomą) ar net visą genomą. Tačiau, genomo variantų nustatymui dažnai vis dar yra taikoma *Sanger* sekoskaita. Be to, šis metodas vis dar taikomas NKS metu nustatytų genomo variantų tvirtinimui, dėl būtinybės įvertinti gautų rezultatų patikimumą, bei paties metodo jautrumą ir specifiškumą. Ši būtinybė atsiranda dėl NKS klaidų, kurios apsunkina variantų nustatymą, pavyzdžiui, dėl prasto genomo padengimo. Mažą padengimą turintys genomo regionai gali lemti reikšmingų variantų praleidimą kritinėse genomo srityse (2).

Įtakos sekoskaitos rezultatams turi ir beveik kiekviename NKS etape padaromos klaidos, pradedant sekoskaitos technologijos pasirinkimu, automatinio analizės algoritmo klaidomis, baigiant *in silico* analizės bei duomenų interpretacijos klaidomis. Pavyzdžiui, *Ion Torrent* technologija yra jautri padengimo klaidoms, kurios atsiranda po emulsinės polimerazės grandininės reakcijos (PGR), ruošiant bibliotekas (3). Labai svarbus žingsnis yra ir praturtinimo sistemos pasirinkimas, nes naudojant skirtingas praturtinimo sistemas nustatomi skirtingi variantai skirtinguose genomo regionuose.

Šiuo metu mokslo pasaulyje vyksta debatai dėl to, ar verta NKS gautus duomenis ir nustatytus variantus visuomet tikrinti *Sanger* sekoskaita, turint omenyje tai, kad daugybėje laboratorijų jau yra gaunamas šimtaprocentinis NKS specifiškumas. *Wenbo* su kolegomis atlikto mokslinio tyrimo metu, kuomet buvo atlikta NKS 20-čiai tūkstančių paveldimo vėžio genų rinkinių ir *Sanger* metodu buvo patikrinti 7845 nepolimorfiniai variantai, buvo nustatyta, kad 98,7 % rezultatų sutapo su NKS duomenimis. Kita vertus, 1,3 % variantų buvo nustatyti kaip klaidingai teigiami. Šie variantai nustatyti probleminiuose genomo regionuose – AT ar GC gausiose srityse, homologiniuose, pseudogenų regionuose ir galėjo būti atmesti kaip klaidingai teigiami tik juos patikrinus *Sanger* metodu (4). Remiantis šiuo pavyzdžiu, nors ir NKS nustatytų variantų tvirtinimas *Sanger* metodu turi akivaizdžią

naudą, tačiau dalis laboratorijų vis tik vengia variantų tikrinimo. Taip yra dėl to, kad ši strategija padidina tyrimo kainą ir laiką.

Šiame darbe bus nagrinėjamos pastarosios problemos – bus siekiama įvertinti NKS duomenų automatinį analizės algoritmą, apibrėžiant metodo tikslumo, jautrumo ir specifiškumo reikšmes bei atsakant į klausimą, ar verta *Sanger* metodu tikrinti NKS duomenis, remiantis šio tyrimo rezultatais. Taip pat bus palyginamos dvi skirtingos praturtinimo sistemos siekiant įvertinti, kuri iš jų yra pranašesnė.

Darbo tikslas – įvertinti naujos kartos sekoskaitos *SOLiD* platformos automatinį analizės algoritmą bei šiai platformai naudojamas taikinių praturtinimo sistemas panaudojant *Sanger* sekoskaitos metodą bei statistinius ir bioinformacinius įrankius.

Darbo uždaviniai:

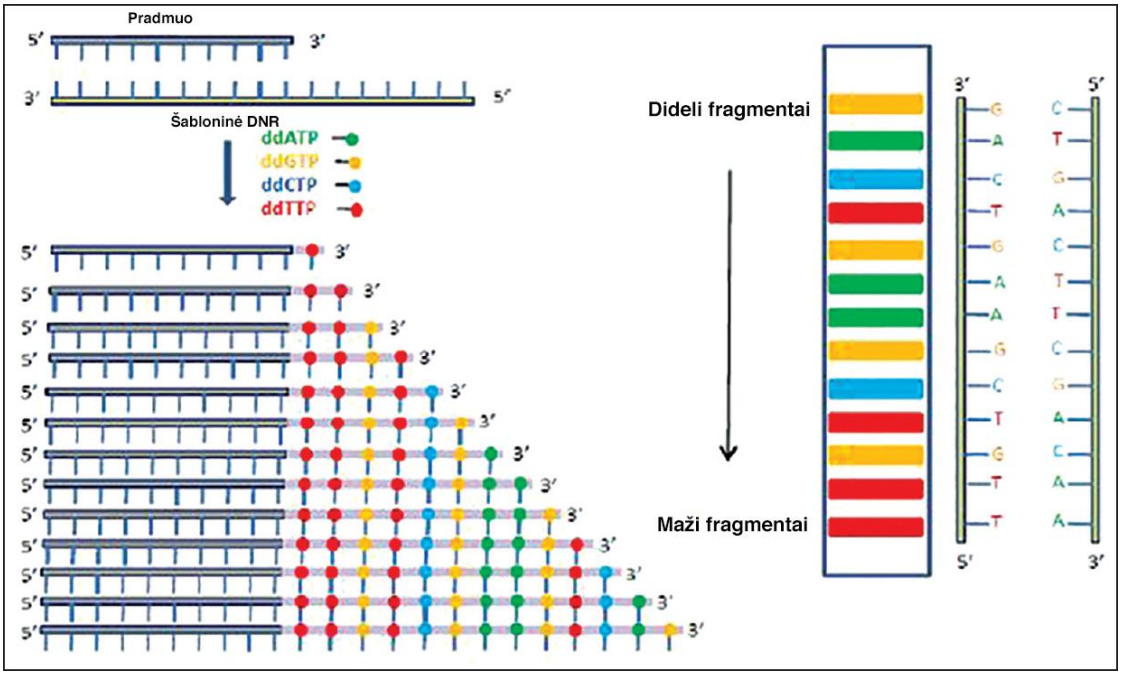
1. Optimizuoti sąlygas genomo variantų nustatymui *Sanger* metodu.
2. Iš parinktų variantų sąrašo, kurie buvo nustatyti NKS metodu ir patikrinti *Sanger* metodu, atrinkti nesutampančius variantus ir juos vizualizuoti integruota genomo vaizduokle (IGV).
3. Palyginti skirtingas, *SOLiD* platformai naudojamas taikinių praturtinimo sistemas *TargetSeq* ir *SureSelect*.
4. Įvertinti darbe naudotos *SOLiD* sekoskaitos platformos automatinio analizės algoritmo tikslumą, jautrumą ir specifiškumą.

1. LITERATŪROS APŽVALGA

1.1 *Sanger* sekoskaitos metodas

Sanger sekoskaitos metodas, dar kitaip vadinamas pirmos kartos sekoskaita dėl to, nes visi kiti naujos kartos metodai yra pastarojo tam tikros modifikacijos. Šis metodas buvo pristatytas 1977, o 1980 *Fred Sanger* už šį metodą gavo Nobelio premiją.

Metodo esmė – DNR grandinės terminacijos reakcija. Pradžioje, reakcijos mišinyje yra pradmenys, komplementarūs vienagrandei DNR molekulei, kurie inicijuoja DNR sintezę. Sintezės metu deoksiribonukleotidai (dNTP) komplementariai prisijungia prie DNR šablono, sudarydami fosfodiesterinį ryšį su naujai įsijungiančiais nukleotidais. Taip pat, reakcijos mišinyje yra dideoksiribonukleotidų (ddNTP), kurie neturi 3' hidroksilo grupės, todėl tokiam nukleotidui įsiterpus į ilgėjančią grandinę įvyksta grandinės terminacija. Kiekvienos reakcijos metu bus sudaromi skirtingo ilgio DNR fragmentai (1 pav). Taip pat šie nukleotidai turi fluorescentinę žymę, todėl įvykus grandinės terminacijai, paskutinis jos nukleotido fluoroforas, sužadinus jį genetiniame analizatoriuje, švies tam tikra spalva. Šis fluorescentinės žymės prijungimas suteikė galimybę automatizuoti *Sanger* metodą.



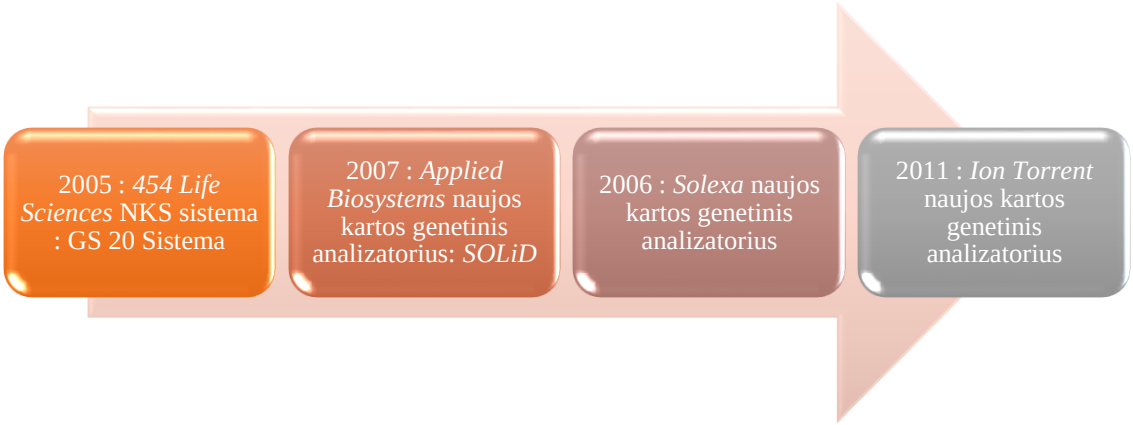
1 pav. Sanger metodo principas (pagal Aliyu S. Bacterial whole genome sequencing: The future of clinical bacteriology. Ann Nigerian Med. 2014; 8:51-7)

Nors *Sanger* metodas yra tikslus, tačiau jis yra šiuolaikinėje genetikoje nebenaudojamas viso genomo sekoskaitai. Taip yra dėl naujų – antros kartos, arba kitaip – naujos kartos sekoskaitos metodų, kurie leidžia atlikti milijonus sekoskaitos reakcijų vienu metu, bei tirti visą egzomą ar genomą per trumpesnį laiką. Nors *Sanger* metodas daugelio vis dar yra laikomas "auksiniu standartu" (5, 6, 7, 8), vis dažniau pasirodo mokslinių publikacijų apie NKS tikslumą ir *Sanger* naudojimo nebūtinumą NKS gautų rezultatų tvirtinimui. *Neveling* su kolegomis atlikto tyrimo metu nustatė, kad egzomo sekoskaita, atlikta *SOLiD* platformos pagalba, turėjo žymiai didesnį diagnostinį našumą kurtumo, aklumo, mitochondrinių ligų nustatyme, lyginant su *Sanger* metodu (9).

Taip pat jau egzistuoja ir trečios kartos sekoskaitos technologijos. Šiuo metu klinikinėje praktikoje dažniausiai naudojami antros kartos sekoskaitos įrenginiai, o dauguma trečios kartos įrenginių išlieka vis dar tobulinimų ir mokslinių tyrimų stadijoje.

1.2 Naujos kartos sekoskaitos metodai

Šiame darbe buvo naudota naujos kartos sekoskaitos *SOLiD* platforma egzomo duomenims gauti, todėl toliau bus aptarti antros kartos sekoskaitos metodai. Antrame paveiksle yra pateikiami minėtieji metodai, bei jų išsidėstymas laiko juostoje pagal pasirodymą rinkoje (6).



2 pav. Antros kartos sekoskaitos metodų laiko juosta

Antros kartos sekoskaita, nepriklausomai nuo platformos, turi kelis esminius etapus, kuriuos būtina įvykdyti siekiant gauti duomenis:

1. Išskiriama genominė DNR;
2. DNR sufragmentuojama iki mažų fragmentų (apie 200 bp);
3. Ruošiamos DNR fragmentų bibliotekos, kurios gali būti tolygiai ir tiksliai sekvenuojamos milijonuose paralelių reakcijų;

4. Naujai iš bazių porų sudarytų grandinių fragmentai yra sugrupuojami pagal žinomą referentinę seką; sugrupuoti fragmentai sulyginami ir pagal jų persidengimus yra nustatoma visa seka kiekvienos genominės DNR mėginyje.

Tačiau tai yra tik apibendrinti metodo etapai. Be to, jie šiek tiek skiriasi, jei yra pasirenkama sekvenuoti ne visą genomą, o tik egzomą ar genomo dalį. Viso egzomo sekokaitos atveju yra sekvenuojami tik egzonai – baltymus koduojantys regionai, kurie sudaro apie 1 % viso genomo (10). Didžioji dalis klinikinį pasireiškimą turinčių ligų priklauso būtent nuo šių regionų, tokiu būdu yra sutaupoma kaštų ir laiko ne tik atlikti tyrimui, bet ir bioinformacinei analizei.

Prieš atliekant egzomo sekoskaitą, turi būti tinkamai išskirta genominė DNR. Ypač svarbus veiksnys yra DNR grynumas, nes nereikalingi komponentai reakcijos mišinyje gali inhibuoti fermentines reakcijas, o RNR gausa gali turėti įtakos DNR fragmentacijai.

Kitame etape yra ruošiama DNR fragmentų biblioteka. Šis etapas susideda iš smulkesnių žingsnių, tokių kaip fragmentavimo ir adapterių prijungimo. DNR fragmentavimas gali būti atliekamas fiziniu metodu, pavyzdžiui, *Covaris* sistema, kuomet aukšto dažnio akustinė energija yra nukreipiama DNR struktūros link, taip ją suskaldant į 100 bp-5 kbp dydžio fragmentus. Taip pat gali būti atlikta ir fermentinė fragmentacija, naudojant restrikcijos endonukleazes (11) bei naudojant kitus metodus.

Po fragmentavimo yra atliekamas adapterių prijungimas, kurio metu yra nubukinami DNR fragmentų galai, fosforilinamas fragmento 5’ galas ir prikabinamas adeninų sekos fragmentas 3’ gale. Adapteriai reikalingi fragmentų prisijungimui prie plokštelės paviršiaus kuris yra išklotas komplementariais oligonukleotidų fragmentais. Jie taip pat reikalingi fragmentų identifikavimui esant skirtingiems mėginiams vienoje reakcijoje. Tai vadinama unikalios DNR sekos adapterio prijungimu: esant daug skirtingų mėginių, yra galimybė kiekvieno mėginio seką pažymėti tam tikra žyme – unikaliu DNR sekos adapteriu ir dėka to, skirtingus mėginius tirti vienu metu.

Po šių žingsnių vykdomas svarbus egzomo sekoskaitai mėginio praturtinimo etapas. Jo metu yra atrenkami tam tikrų sekų DNR fragmentai. Šiuo atveju, norint tirti egzomą, turi būti atrenkami egzonai. Tam dažniausiai pasitelkiami komerciniai taikinio praturtinimo rinkiniai. Mėginį galima praturtinti naudojant mikrogardelių (*Roche*/*NimbleGen EZCap*) ar tirpale esančių, hibridizacijos principu veikiančių dalelių technologijas. Dėl savo lengvesnio automatizavimo, lankstumo, patogesnio naudojimo pastaroji yra žymiai populiaresnė. Vieni populiariausių praturtinimo rinkinių yra *TargetSeq* ir *SureSelect*. Šie rinkiniai bus aptarti plačiau kitame skyriuje.

Po aptartųjų etapų vykdomas DNR fragmentų gausinimas naudojant mikrogardeles arba emulsinę PGR (priklausomai nuo naudojamos platformos) bei mėginio praturtinimas, siekiant pašalinti

reagentų perteklių, taip pat, naudojant emulsinę PGR, užtikrinti, kad viename mikroreaktoriuje būtų tam tikras tinkamas komponentų skaičius. Tokios „nepraturtintos“ mikrodalelės yra atskiriamos nuo praturtintų, ant kurių DNR buvo pagausinta. Kai mėginys yra centrifuguojamas, praturtintos mikrodalelės lieka viršutiniame sluoksnyje, o nepraturtintos nusėda apačioje. Viršutinis sluoksnis yra surenkamas ir mikrodalelės yra paskirstomos ir pritvirtinamos kovalentiškai ant stiklinės plokštelės paviršiaus. Kiekvienoje dalelėje vyks atskira sekvenavimo reakcija (12).

Iki šio etapo, skirtingi sekoskaitos metodai beveik nesiskiria tarpusavyje. Esminis skirtumas yra pačiame sekoskaitos reakcijos principe, kuris skiriasi tarp platformų.

Roche 454 sistema

2005-iais metais *Roche 454* tapo pirmąja komerciškai prieinama platforma, kuri naudojo paralelios sekoskaitos metodą. Šis metodas yra paremtas pirosekoskaitos technologija ir šiuo metu gali produkuoti apie 500 Mb informacijos (13). Pirosekoskaita yra paremta neorganinių fosfatų nustatymu sekoskaitos reakcijos metu, kurių skaičius yra proporcingas komplementarių nukleotidų įsiterpimų skaičiui.

DNR šablonai pirosekoskaitai yra paruošiami taip: netaisyklingai sufragmentuoti DNR fragmentai yra prijungiami ant specialaus objektinio stiklelio paviršiaus, ant kurių yra dalelės su trumpais adapteriais. Prie šių adapterių DNR fragmentai ir prisijungia. Tuomet yra atliekama emulsinė PGR, pagausinami fragmentai. Mikrodalelės su fragmentais yra perkeliamos į šulinėlius. Taip pat į šulinėlius dedamos ir mažesnės dalelės, kuriose yra fermentai sulfurilazė ir luciferazė. Tuomet nuosekliai prie šabloninės DNR grandinės pradedami jungti dNTP. Kuomet pridedamas naujas komplementarus dNTP, yra inicijuojama DNR sintezė ir yra išskirimas pirofosfatas, kuris yra paverčiamas adenozinotrifosfatu (ATP) dėka ATP sulfurilazės. Tuomet ATP yra panaudojama luciferino pakeitimui į oksiluciferiną (veikiantis fermentas – luciferazė). Dėl to atsiranda nustatomos šviesos pliūpsnis (14). Šie pliūpsniai yra fiksuojami – laikoma, jog vienas pliūpsnis atitinka vieną komplementariai įsijungusį dNTP. Su pagrindiniu šio metodo trūkumu susiduriama sekvenuojant homogeniškus regionus. Sistema negali atskirti, ar tas pats nukleotidas pasikartoja 10 ar 20 kartų. Tuo tarpu, pagrindinis privalumas – ilgų fragmentų ir aukštos kokybės dėka šia technologija gautus duomenis galima analizuoti esant mažesniam padengimui lyginant su *SOLiD* ar *Illumina* platformomis.

Solexa/Illumina sistema

Illumina yra NKS lyderė, atsakinga už 70 % platformų, įsigijamų ir naudojamų pasaulyje (15), bei turinti platformų, gebančių sugeneruoti itin didelius duomenų kiekius. Pavyzdžiui, *Highseq3000/4000* gali sugeneruoti net iki 1500 Gb duomenų per eksperimentą.

Illumina metodas yra paremtas sekoskaita sintezės metu – nukleotidai yra nustatomi DNR grandinės sintezės metu. DNR fragmentai yra pritvirtinami prie plokštelės paviršiaus ir yra vykdomas tiltelinis gausinimas. Tokiu būdu yra pagaminami klasteriai – vienos DNR molekulės kopijos tam tikrame plokštelės plote. Viso to eigoje yra pagaminami 100-200 milijonų klasterių. Tuomet į reakcijos mišinį yra įdedami visi 4 skirtingai fluorescentiškai žymėti dNTP su terminalinėmis grupėmis, kurios vėliau gali būti pašalinamos – taip gaunamas fluorescentinis signalas, o grandinės sintezė gali būti vykdoma toliau (16).

Šios technologijos dėka buvo sumažintas tyrimo laikas ir kaina – *HiSeq2000* platforma 2014-iais metais buvo pripažinta pigiausia (0,02 JAV dolerio už milijoną nukleotidų) (17).

***SOLiD* sistema**

2008-iais *Applied Biosystems* (dabar *Life Technologies*) sukūrė NKS platformą, kurios veikimas paremtas emulsine PGR ir sekoskaitos ligavimo principu metodika. Panašiai kaip ir *Roche 454* sistema, DNR šablonai yra paruošiami juos gausinant emulsinės PGR metu, tačiau skirtingai nuo *Roche 454*, *SOLiD* naudoja mažesnes daleles, dėl to gaunamas tankesnis dalelių ant plokštelės pasiskirstymas. Ši sistema geba sugeneruoti aukštos kokybės duomenis (>60 Gb) gavus didelį kiekį trumpų (50 bp) fragmentų. Šia savybe *SOLiD* platforma yra panaši į *Illumina*. Tačiau pagrindinis trūkumas, kuris būdingas ir *Illumina* sistemai – šie trumpi fragmentai yra netinkami siekiant gero padengimo pasikartojančiuose regionuose (12).

Šios sistemos veikimo principas visiškai kitoks, lyginant su aptartais metodais. DNR sekai nustatyti yra pasitelkiami zondai. Kiekvienas zondas turi vieną iš 16 dinukleotidų 1-2 pozicijose, kurie vėliau komplementariai jungsis prie DNR grandinės. Kiekvienas zondas taip pat turi degeneruotas bazes 3-5 pozicijose, o bazės 6-8 pozicijose laiko vieną iš keturių fluorescuojančių dažų. Kadangi yra 16 skirtingų dinukleotidų kombinacijų, iš viso yra 16 skirtingų zondų, kurie fluorescuoja 4 spalvomis.

Visų pirma, prie DNR fragmento adapterio prisijungia pradmuo. Tada ligazės pagalba prie grandinės yra prijungiamas pirmasis zondas. Kaip jau minėta, kiekvienas zondas yra žymėtas vienu iš keturių fluoroforų. Kai jis randa komplementarią seką ir prie jos prisijungia, įvyksta fermentinė reakcija, kurios metu fluoroforas atskyla nuo zondo, o išspinduliuota spalva yra užfiksuojama kameros. Po 5-7 ciklų naujai susidariusi grandinė disocijuoja, o prie šablono prisijungia naujas n-1 pradmuo.

Procesas pakartojamas su n-1 pradmeniu, po to su n-2, n-3 ir n-4 pradmenimis. Taip yra sugeneruojami persidengiantys duomenys. 5-7 ligacijos reakcijos su 5 pradmenų pakeitimais sugeneruoja duomenis ~35 bazių sekai nustatyti. Pagrindinis šios technikos privalumas yra tai, kad kiekvienas nukleotidas yra nusekvenuojamas po du kartus, taip sumažinama klaidos tikimybė. Be to, metodas leidžia atskirti vieno nukleotido pakitimus nuo įvairių sistemos ar žmogaus klaidų – vieno nukleotido polimorfizmų (VNP) atveju yra galimos tik trys spalvų kombinacijos; visos kitos kombinacijos reiškia ne VNP, o sistemos klaidą. Tačiau šiuo atveju duomenų interpretavimas yra sudėtingesnis dėl naudojamo dvispalvio kodo lyginant su metodais, kurie naudoja sintezės technologiją (17).

Ion Torrent sistema

Ion Torrent yra technologija, kuri priklauso tiek antros, tiek trečios kartos sekoskaitos metodų kategorijoms. Metodas pagrįstas puslaidininkių technologija, kuomet specialiuose šulinėliuose, tankiai išsidėsčiusiuose ant plokštelės, yra integruoti specifiniai jutikliai, kurie registruoja vandenilio jonų atsipalaidavimą, kuomet inkorporuojamos naujos bazės. Šio metodo dėka nereikalingos sudėtingos šviesos registravimo, skenavimo sistemos, dėl to sekoskaitos procesas yra žymiai greitesnis, o kaina mažesnė. Ši metodika, kaip ir visos antros kartos sekoskaitos metodikos pagrįsta „plovimo ir registravimo“ principu, reikalaujančiu ir DNR bibliotekų gausinimo etapo PGR metodu, o taip pat ir terminaciniais įvykiais, kurie stabdo sekoskaitos procesą, todėl ši technologija nėra galutinai priskiriama trečios kartos sekoskaitos technologijų grupei (18). Skirtingų aptartų sekoskaitos metodų palyginimas pateikiamas pirmoje lentelėje.

1 lentelė. Skirtingų antros ir (ar) trečios kartos sekoskaitos metodų privalumai ir trūkumai (pagal Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour (2011); 11 759-69.)

Sistema	Privalumai	Trūkumai
Roche 454	Didelis DNR fragmentų ilgis	Didelė Mb kaina
Illumina	Žemiausia Mb kaina, Gb duomenų gaunama per dieną, sugeneruoja daugiausiai DNR fragmentų; pakankamai paprastas ir greitas DNR fragmentų bibliotekų paruošimas, kai nevykdoma emulsinė PGR	Brangi įranga, išlaikymas

Sistema	Privalumai	Trūkumai
SOLiD	Nedidelė Gb kaina, didelis tikslumas	Palyginus trumpi DNR fragmentai, surenkant fragmentus į genomą lieka daugiau tarpų, lyginant su *Illumina*. Senesnės kartos prietaisai naudoja sudėtingą spalvinį kodą
Ion Torrent	Nesudėtingas įrenginys, nesunku naudotis	Mb kaina panaši į *Roche 454*; sunkumai sekvenuojant homogeniškus regionus; neįmanoma tirti viso egzomo

1.3 Naujos kartos sekoskaitos ir *Sanger* metodų sąsaja

NKS evoliucionavo į labai vertingą technologiją, kuri yra pritaikoma diagnostikoje. Tačiau, atsižvelgiant į problemas, kylančias dėl sudėtingo analizės algoritmo, kuris eikvoja tyrėjo laiką ir reikalauja bioinformatikos žinių, diagnostikos centrams šią technologiją naudoti kaip rutininį metodą, atitinkantį visus kokybės standartus, gali būti sudėtinga. Be to, kyla tam tikrų problemų dėl spartaus šių technologijų vystymosi. Pavyzdžiui, šiuo metu yra sukurtos kelios skirtingos praturtinimo sistemos, kurios visos turi savo privalumų ir trūkumų, o nuo to iš dalies priklauso metodo specifiškumas ir jautrumas. Todėl dažnai skirtingi mokslinių publikacijų autoriai gauna skirtingus rodiklius ar negali priimti vieningos nuomonės rūpimais klausimais, o tai trukdo šio metodo rutinizacijai.

Kita problema – skirtingos NKS technologijos, susidurdamos su probleminiais genomo regionais (homogeniniai bei gausūs mažų iškritų bei intarpų regionai), gali turėti skirtingas jautrumo ir specifiškumo reikšmes. Todėl buvo padaryta išvada, kad nors NKS technologijos yra tinkamos tūkstančių variantų nustatymui vienu metu, siekiant sutaupyti laiko ir išlaidų, tradicinis *Sanger* sekoskaitos metodas turi būti naudojamas kaip pagalbinis metodas NKS nustatytų variantų patvirtinimui prieš įvardinant galimą funkcinę reikšmę tam tikrų ligų atsiradimui (8). *Sanger* metodas yra tikslus, tačiau jo naudojimą šiuolaikinėje klinikinėje genetikoje riboja didėjantis tiriamų genų skaičius tam tikriems daugiaveiksniams bei poligeniniams susirgimams, kaip, pavyzdžiui, kurtumas ar intelektinė negalia. Taip yra, nes *Sanger* metodas iš esmės yra netinkamas dideliam genų skaičiui sekvenuoti, o tokiu atveju tyrimo kaina ir laikas smarkiai išauga. Tačiau *Sanger* metodas puikiai tinkamas monogeninėms ligoms nustatyti ir tokiu atveju NKS naudojimas yra nebūtinas. Taip yra nurodoma ir NKS naudojimo gairėse (8). Taigi, šie metodai iš esmės papildo vienas kitą – NKS

metodai leidžia atlikti didelio masto genomo tyrimus, o *Sanger* metodas naudojamas mažesnių apimčių tyrimams bei NKS gautų rezultatų tvirtinimui. Tačiau daugėjant žinių apie žmogaus genomą, NKS technologijos taps dar svarbesniu įrankiu ne tik moksliniams, bet ir klinikiniams tyrimams. Atsiranda galimybė nustatyti itin retus variantus – ir būtent NKS metodas šių variantų paieškai tinka labiausiai. Kuomet NKS kaina kris, o bioinformaciniai metodai dar labiau patobulės, bus įmanoma šį metodą galutinai integruoti į klinikinę rutiną, atsisakant rezultatų tvirtinimo *Sanger* metodu.

1.4 Skirtingos praturtinimo sistemos

Nors viso genomo sekoskaita geba nustatyti įvairius genomo pokyčius, kaip vieno nukleotido variantai, intarpai ir iškritos, kopijų skaičiaus pokyčiai visame genome, tačiau tokių duomenų apdorojimas vis dar kelia iššūkių. Tuo tarpu viso egzomo sekoskaita, kurios metu yra vykdoma tik žinomų genomo genų egzonų sekoskaita, sugeneruoja mažiau duomenų ir taip ne tik gali būti sumažinama tyrimo kaina net iki penkių kartų, bet ir supaprastinama duomenų analizė (19).

Todėl egzomo sekoskaita yra puiki alternatyva viso genomo sekoskaitai, nes apytikriai 85 % visų genomo mutacijų, lemiančių ligas, yra nustatomos koduojančiuose genomo regionuose (20).

Siekiant atskirti egzomą nuo genomo yra naudojamos taikinių praturtinimo sistemos. Jos gali būti dviejų tipų – gardelėmis paremtas egzomo atskyrimas arba egzomo atskyrimas tirpale. Pirmiausia buvo sukurta sistema paremta gardelėmis (21), tačiau šis metodas reikalavo didelio kiekio DNR, todėl buvo greitai pakeistas atskyrimu tirpale.

Vienos populiariausių taikinių praturtinimo sistemų yra *SureSelect* bei *TargetSeq*. Abi šios sistemos veikia hibridizacijos principu. Paruošta biblioteka yra hibridizuojama su RNR zondais, kurie yra praturtinti biotinu. Po hibridizacijos, RNR zondai prisijungia prie komplementarių sekų, o iš kitos pusės – su magnetinėmis dalelėmis, kurios praturtintos streptavidinu ir sudaro ryšius su biotinu. Magneto pagalba norimi fragmentai yra atskiriami iš likusio mėginio.

Pagrindiniai sistemų skirtumai yra zondo ilgis bei tikslinis regionas, kurį apima pasirinkta sistema. Taip pat tyrėjus domina nedideli koduojančių regionų padengimo skirtumai taikant skirtingas sistemas, nes šie skirtumai tiesiogiai atspindi gebėjimą nustatyti retus variantus koduojančiuose regionuose (22).

Šie skirtumai gali daryti įtaką gaunamiems rezultatams, todėl svarbu atlikti sistemų palyginamąją analizę. Mokslininkai visame pasaulyje publikuoja gautus tokios analizės rezultatus ir taip padeda kitiems apsispręsti, kurią sistemą yra verta naudoti (23).

Mokslinėse publikacijose minima, jog *SureSelect* sistema yra tikslesnė variantų nustatymo atžvilgiu, ypač analizuojant trumpus intarpus bei iškritas. *Zhang* su kolegomis atlikto tyrimo metu buvo nustatyta, kad 70 % variantų buvo bendrai nustatyti abejomis sistemomis, tuo tarpu 30 % buvo platformai specifiški. Tiriant praturtinimo sistemai specifinius VNP, *Sanger* metodu buvo patvirtinti 88,3 % *SureSelect* sistemai specifiniai ir tik 60 % *TargetSeq* sistemai specifiniai VNP. Dar labiau stebina trumpų intarpų ir iškritų patvirtinimo rezultatai – *SureSelect* sistemai specifiniai trumpi intarpai ir iškritos buvo patvirtinti 89,6 % atvejų, ir tik 15,8 % atvejų *TargetSeq* specifiniams variantams. Įdomu tai, kad tikrinant persidengiančius variantus, nustatytus abejomis sistemomis, *Sanger* metodu, buvo gauti geresni rezultatai – VNP buvo patvirtinti 91,5 %, o trumpų intarpų bei iškritų – 100 % visų atvejų (24).

Shigemizu su kolegomis savo mokslinio tyrimo metu padarė panašią išvadą – *SureSelect* pasiekia geriausią taikinių praturtinimo efektyvumą, geriausią sekos padengimą koduojančiuose regionuose bei geriausią vieno nukleotido variantų ir trumpų intarpų bei iškritų nustatymo jautrumą (22).

SureSelect sistema pripažinta geresne ir *Park* bei jo kolegų tyrime. Pastarojo darbo metu buvo nustatyta, kad geresnis regiono padengimas pasiekiamas būtent su šia sistema (25). Be to, remiantis abiejų gamintojų pateikiama informacija, naudojant *TargetSeq* praturtinimo sistemą padengimas yra mažesnis (26, 27).

Tuo tarpu, *Londin* su kolegomis nagrinėjo farmakogenomikoje svarbius variantus ir aiškinosi egzomo sekoskaitos naudą šių variantų nustatymui. Buvo iškeltas klausimas, ar naudojant egzomo sekoskaitą bei skirtingas taikinių praturtinimo sistemas gali būti pasiektas geras padengimas svarbiose genomo pozicijose. Idomu tai, kad šių autorių duomenimis mažiausiai variantų egzonuose buvo nustatyta naudojant *SureSelect* sistemą. Tuo tarpu naudojant *TargetSeq* sistemą variantų buvo nustatyta apie 10 % daugiau (28).

Tačiau yra mokslinių publikacijų, kuriose lyginami keli skirtingi praturtinimo sistemų gamintojai ir gauti rezultatai rodo, kad visos lyginamos technologijos pasiekė panašų tikslumą nustatant VNP (29). Taigi, skirtingose mokslinėse publikacijose daromos skirtingos išvados atskleidžia šios temos nevienalypiškumą ir tolimesnį tyrimų aktualumą.

1.5 Naujos kartos sekoskaitos automatinis analizės alogritmas, jo parametrai

Sparčiai tobulėjant NKS technologijoms ir didėjant gaunamų duomenų kiekiui, gyvybės ir gamtos mokslų srityse dirbantiems specialistams bioinformacinės analizės tema darosi vis aktualesnė. Be to, ką tik po NKS gautuose failuose esanti informacija yra neinformatyvi tyrėjui, todėl ją reikia specialiai

apdoroti. Todėl yra itin svarbu išmanyti ne tik pačios NKS principą, bet ir mokėti analizuoti bei tvarkyti gautus duomenis taikant bioinformacinius įrankius. Sujungiančių kelių programų darbą, bioinformacinio pobūdžio komandų naudojimo seka, kurios pagalba yra gaunama informacija apie tiriamojo asmens genomo variantus, yra vadinama NKS duomenų automatiniu analizės algoritmu. Yra išskiriami pagrindiniai analizės algoritmo etapai:

1. Sekos prilygiavimas prie referentinio genomo;
2. Teisingų genomo variantų atranka;
3. Anotavimas.

Pirmajame etape vykdomas NKS metu gautų DNR fragmentų prilygiavimas prie referentinio genomo. Čia yra svarbios dvi sąvokos: DNR fragmentas ir referentinis genomas. DNR fragmentas apibrėžiamas kaip galutinis NKS produktas, nusekvenuota tam tikrų nukleotidų atkarpa. Referentinis genomas – visuotinai taikomas genomas, sudarytas iš unikalių ir anotuotų DNR sekų rinkinio ir naudojamas mokslininkų gautų duomenų palyginimui.

Priklausomai nuo NKS platformos, po sekoskaitos gali būti gaunami dviejų skirtingų formatų failai – XSQ, jei sekoskaita buvo atlikta *SOLiD* platforma, arba BCL (kuri automatiškai verčiama į FASTQ), jei buvo naudota *Illumina* ar kita sekoskaitos platforma. *SOLiD* formatas skiriasi nuo kitų sekoskaitos platformų dėl specifinių, sekoskaitoje naudojamų zondų ir jų dvispalvio kodo.

Svarbu paminėti, kad tobulėjant sekoskaitos technologijoms, tobulėja ir prilygiavimo programos, jų daugėja, todėl renkantis prilygiavimo programą, reikėtų atkreipti dėmesį į naudotą NKS platformą, programos tikslumą bei kitus parametrus, nes tai gali turėti įtakos galutiniams rezultatams. Vieni populiariausių bei geriausių prilygiavimo rezultatų pasiekiantys bioinformaciniai įrankiai šiame etape yra *BWA, Bowtie, SHRiMP2* (30). Be to, visas šias programas galima naudoti nepriklausomai nuo sekoskaitos platformos.

Prilygiavus DNR fragmentus prie referentinio genomo, pradinis failo formatas yra pakeičiamas į BAM formatą, kuris yra universalus visoms sekoskaitos platformoms. Šiame faile yra užkoduota esminė informacija apie DNR fragmentų sekas, pozicijas genome, bei jų prilygiavimo kokybę. Šio failo vizualizavimui dažnai yra naudojama integruota genomo vaizduoklė (IGV).

Panaudojant prilygiavimo programas, gautuose duomenyse yra matomas labai svarbus parametras NKS duomenų analizėje – padengimas (angl. *coverage*). Padengimas parodo, kiek DNR fragmentų nukleotidų persidengia tam tikroje genomo pozicijoje prilygiavus juos prie referentinio genomo. Kuo daugiau kartų tas pats nukleotidas pasikartoja, tuo geresni yra rezultatai, išskyrus tuos atvejus, kai

didelis nukleotido pasikartojimų skaičius sukuriamas dirbtinai, pavyzdžiui, dėl DNR fragmentų duplikacijų (31).

DNR fragmentų duplikacijomis laikomos DNR fragmentų kopijos, kurios gali turėti įtakos galutiniams rezultatams. Jei viename DNR fragmente būtų klaidingai nustatytas pokytis, tuomet to fragmento duplikacijose šis pokytis taip pat būtų. Po prilygiavimo, šios fragmento duplikacijos dirbtinai padidintų padengimo reikšmę. Joms pašalinti naudojamos programos, tokios kaip *SAMtools* (32). Tačiau yra pasirodę mokslinių publikacijų, kuriose teigiama, kad šis etapas gali būti praleistas, o fragmentų duplikacijų pašalinimas turi nedidelę įtaką genomo variantų atrankos tikslumui (33).

Kita su padengimu susijusi problema yra netolygus padengimas. Jis gali atsirasti dėl to, kad ne visų sekų fragmentų sekoskaitos našumas yra vienodas. Pavyzdžiui, sunkiau yra sekvenuoti sekas, linkusias sudaryti antrines struktūras; dažnai pasikartojančias sekas; GC turtingus regionus – tokie regionai dažnai pasižymi nedideliu padengimu (34).

BAM faile taip pat pateikiama informacija apie kiekvieno nukleotido kokybės įverčius (angl. *base quality score*). Šis įvertis nurodo, kaip tiksliai buvo įvardintas nukleotidas arba nurodo tikimybę, kad NKS metu buvo padaryta klaida. Nukleotido kokybės įvertis pateikiamas *Phred* įverčio formatu. Kuomet *Phred* įvertis lygus 10, tai reiškia, kad yra 1 iš 10 tikimybė, jog bazė identifikuota klaidingai. Tuo tarpu, jei *Phred* įvertis lygus 30, tai reiškia, kad klaidos tikimybė yra lygi 1 iš 1000. Todėl šis įvertis svarbus siekiant atskirti teisingai identifikuotus variantus nuo klaidingai teigiamų variantų (2).

Taip pat šiame faile nurodomas *CIGAR* įvertis – šis įvertis teikia informaciją apie DNR fragmento sutapimą su referentiniu genomu prilygiuotoje vietoje. Pavyzdžiui, *SOLiD* platforma gaunamų DNR fragmentų ilgis yra 75 bp. Jei fragmentas idealiai prilygiuojamas prie referentinio genomo, tuomet *CIGAR* vertė yra lygi 75 M. Tačiau, yra galimas variantas, kad DNR fragmento pirmieji 30 nukleotidų bus prilygiuoti tinkamai, toliau kiti penki nebus prilygiuoti, o likę 40 vėl atitiks referentinio genomo seką. Tokiu atveju *CIGAR* įvertis būtų lygus 30M5N40M.

Ketvirtas, labai svarbus įvertis, kuris yra nustatomas prilygiavimo prie referentinio genomo etape, yra DNR fragmento kokybės įvertis (angl. *mapping quality score*). Šis parametras labai priklauso nuo sekų homologijos. Nusekvenuoti DNR fragmentai yra trumpi, todėl jie gali atsikartoti skirtingose genomo vietose, o tai gali lemti neteisingą DNR fragmento prilygiavimą prie referentinio genomo. Todėl DNR fragmentui, kuris gali būti prilygiuotas keliose genomo pozicijose, programa suteikia žemesnį DNR fragmento kokybės įvertį.

Po prilygiavimo vykdomas antrasis – teisingų genomo variantų atrankos etapas. Šis etapas yra itin svarbus ir nuo jo priklauso, ar variantai bus identifikuoti. Teisingu variantu yra laikomas NKS analizės algoritmo identifikuotas genomo variantas, kuris ne tik skiriasi nuo referentinio genomo, bet ir turi

aukštus anksčiau aptartus kokybės įverčius, kurie leidžia teisingus variantus atskirti nuo klaidingai teigiamų variantų. Teisingų variantų atrankai naudojami parametrai, aprašyti BAM failo formate ir buvo aptarti aukščiau – tai padengimas, kiekvieno nukleotido kokybės, *CIGAR* bei DNR fragmento kokybės įverčiai. Šiam etapui įvykdyti gali būti naudojama daugybė skirtingų programų (35). Viena iš dažniausiai naudojamų programų šiame analizės etape yra *GATK* (angl. *Genome Analysis Toolkit*) programa. Ši programa naudoja BAM arba SAM formato failus, o juos apdorojus, paverčia VCF formato failais. Šiuose failuose jau yra nurodomi visi nustatyti genomo variantai, besiskiriantys nuo referentinio genomo, jų padengimas (DP), genotipas, bei kiti svarbūs rodikliai (36).

Paskutinysis analizės algoritmo etapas – anotavimas. Tai gali būti atliekama tokiomis programomis kaip *ANNOVAR*. Jo metu prie nustatytų variantų yra prijungiama informacija iš įvairių duomenų bazių apie variantų tipą, patogeniškumą, paplitimą populiacijose ir kt. Patogu tai, kad tyrėjas pats pasirenka duomenų bazes, kuriose sukaupti duomenys aprašo nustatytus variantus ir taip leidžia juos analizuoti tam tikrame kontekste.

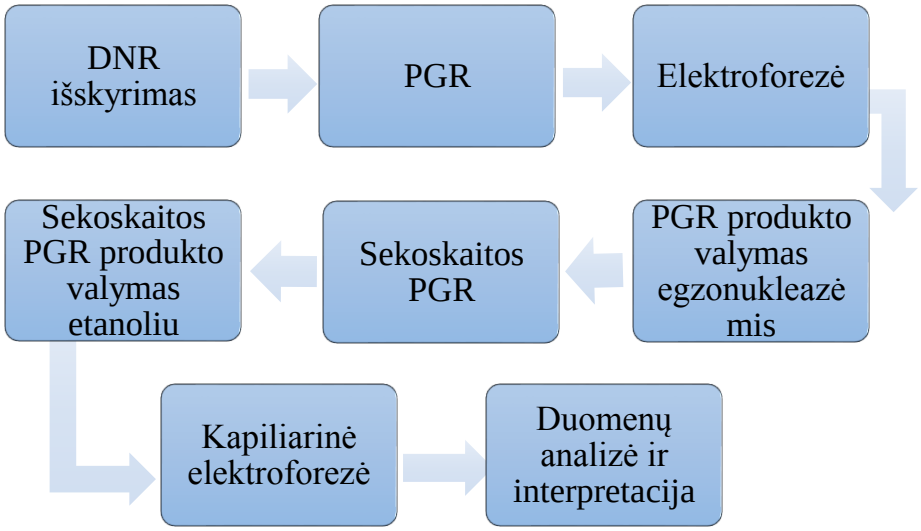
Šiuo metu jau egzistuoja tokie bioinformaciniai įrankiai, kurie apjungia visus šiuos etapus ir visi veiksmai šių įrankių pagalba yra atliekami vienoje platformoje. Tokių įrankių pavyzdys yra *Galaxy* platforma, kuri yra lengvai prieinama žiniatinklyje ir padeda tiriant genomą (37). Be to, mokslininkai kuria vis daugiau tokio tipo platformų, kaip *MutAid* platforma, kuri jau leidžia apjungti NKS ir *Sanger* sekoskaitos duomenis, tokiu būdu supaprastinant NKS duomenų analizę ir mutacijų tvirtinimą *Sanger* metodu (38).

2. TYRIMO METODAI

2.1 Tiriamųjų grupė ir tyrimo eiga

Darbe buvo panaudoti VU MF Žmogaus ir medicininės genetikos katedroje vykdyto projekto „Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis“ (LITGEN) viso egzomo tiriamųjų duomenys (projektas finansuotas Europos socialinio fondo lėšomis; projekto vadovas LMA tikrasis narys akad. prof. habil. dr. Vaidutis Kučinskas). Viso tiriamųjų grupę sudarė 96 tiriamieji.

Darbo metu buvo optimizuotos PGR sąlygos 60-čiai genų egzonų ar jų fragmentų. Jiems visiems buvo atlikta *Sanger* sekoskaita, siekiant įsitikinti, kad NKS būdu nustatyti variantai yra nustatomi ir *Sanger* metodu. Apibendrinta tyrimo darbo schema pateikiama trečiame paveiksle.



3 pav. Apibendrinta tyrimo darbo schema

2.2 DNR išskyrimas

DNR buvo išskirta VU MF ŽMGK darbuotojų fenolio-chloroformo metodu pagal laboratorijoje patvirtintą metodiką arba naudojant automatizuotą *TECAN Freedom EVO®200* sistemą (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 Pradmenų kūrimas

Siekiant pagausinti koduojančias 60-ies genų sritis buvo pasitelkta polimerazės grandininė reakcija. Pradmenų kūrimui buvo naudojama *Ensembl* duomenų bazė ir *Primer3* programa, kuri yra prieinama internete (http://bioinfo.ut.ee/primer3-0.4.0). Buvo sukurti pradmenys, komplementarūs 37 skirtingiems egzonams ar jų fragmentams, likusiems 23 genų egzonams pradmenys buvo sukurti anksčiau VU MF ŽMGK. Kuriant pradmenis buvo laikomasi pagrindinių pradmenų kūrimo kriterijų – pradmenų ilgis 20-22bp, pradmenų lydymosi temperatūra 58-60°C, G/C nukleotidų kiekis tarp 40-60 %.

2.4 Polimerazės grandininė reakcija

Polimerazės grandininė reakcija (PGR) – tai nukleorūgščių sintezės *in vitro* metodas, kuriuo mėgintuvėlyje gali būti specifiškai pagausinti atskiri DNR fragmentai. Dėl to, kad kiekvienam egzonui

specifiškos pradmenų prisijungimo temperatūros skyrėsi, iš pradžių buvo būtinas PGR sąlygų optimizavimas. Pagal teorines pradmenų prisijugimo temperatūras, buvo pritaikytas eksperimentinis PGR sąlygų optimizavimas bei gradientinė PGR. Tolimesniam darbui buvo parinktos reakcijos sąlygos, nurodytos antroje lentelėje.

2 lentelė. Egzonų, jų fragmentų bei šalia esančių genomo sričių gausinimo PGR metodu sąlygos

Ciklas	Temperatūra (°C)	Trukmė	Ciklų skaičius
Pradinė denatūracija	95	2 min	1
Denatūracija	95	30 s	30
Pradmens prisijungimas	56 (33 egzonams), 58 (23 egzonams), 60 (4 egzonams)	30 s	
Sintezė	72	1 min	
Inkubacija	72	10 min	1

PGR metodo atlikimo eiga:

1. Mėginiai ruošiami traukos spintoje, prieš tai sterilizavus aplinką naudojant UV lempą 20 min. bei kruopščiai nuvalius paviršių 70 % etanoliu.

2. Steriliame 0,5 ml talpos mėgintuvėlyje ruošiamas reakcijos mišinys. Mišiniui naudotų medžiagų kiekis nurodytas trečioje lentelėje. Be įprastų PGR komponentų, į reakcijos mišinį dedamas ir dimetilsulfoksidas (DMSO), kuris apsaugo nuo nespecifinio pradmenų jungimosi– pakeisdamas GC turtingų regionų konformaciją sumažina pradmenų jungimosi temperatūrą. Visų medžiagų mišinys gerai supurtomas ir nucentrifuguojamas.

3 lentelė. PGR mišiniui naudotas medžiagų kiekis

Mišinio komponentai	Medžiagų tūris vienam mėginiui (µL)
*PCR Master Mix (*Thermo Fisher Scientific Baltics, Lietuva)	7,5
H2O	5,1
*DMSO (*Thermo Fisher Scientific Baltics, Lietuva)	0,6

3. Į sterilius 0,2 ml talpos mėgintuvėlius įpilama po 13,2 µl mišinio.

4. Į kiekvieną mėgintuvėlį su mišiniu pridedama po 0,6 µL genominės DNR tirpalo (50-100 ng/µL). Viename mėgintuvėlyje vykdoma neigiama kontrolė – vietoje DNR pridedamas 0,6 µL vandens, siekiant patikrinti ar nėra vandens bei reagentų užterštumo.
5. Į mėgintuvėlius pridedama po 0,6 µL tiesioginio ir atvirkštinio pradmens. Pradmuo parenkamas pagal tai, kurį fragmentą norima gausinti.
6. Mėgintuvėliai supurtomi, nucentrifuguojami ir sudedami į termociklerį. Skirtingiems egzonams reikalingos skirtingos reakcijos sąlygos, todėl taikant eksperimentinį PGR sąlygų optimizavimą, buvo nustatytos dvi pradmenų jungimosi temperatūros, prie kurių buvo gauti specifiški PGR produktai. Tuomet buvo sukurtos dvi skirtingos programos termocikleryje. Programą sudaro penki etapai: pradinė DNR grandinių denatūracija, DNR grandinių denatūracija, pradmenų jungimasis, sintezė ir galutinė sintezė. 2-4 etapai kartojami 30 kartų. Ciklų temperatūros ir laikas nurodyti antroje lentelėje.

2.5 Elektroforezė

Elektroforezė – tai metodas, naudojamas įvertinti PGR produkto pagausinimą. DNR molekulė turi neigiamą krūvį, todėl patalpinta elektriniame lauke, judės link teigiamo elektrodo. Ilgesni fragmentai gelyje judės lėčiau, trumpesni – greičiau, todėl šis metodas naudojamas DNR ar kitų molekulių bei cheminių junginių frakcionavimui. Šio tyrimo metu, norint nustatyti PGR produkto buvimą, ilgį, specifiškumą bei reagentų ir reakcijos mišinio užterštumą buvo naudojama horizontali elektroforezė 1,5 % agarozės gelyje. Elektroforezės metu naudojamos medžiagos ir jų kiekiai nurodyti ketvirtoje lentelėje.

Metodo atlikimo eiga:
1. Paruošiama elektroforezės forma, įdedamos šukutės takeliams formuoti.
2. Paruošiamas 1,5 % agarozės tirpalas 1xTBE buferyje. Agarozė ištirpinama, kaitinant mikrobangų krosnelėje. Tirpalas atvėsinamas iki 60°C temperatūros.
3. Į agarozės tirpalą įlašinami 2µl etidžio bromido tirpalo (10 mg/ml).
4. Agarozė įpilama į plokštelę ir paliekama polimerizuotis kambario temperatūroje 30 min.
5. Atsargiai ištraukiamos šukutės.
6. Gelis perkeliamas į elektroforezės aparatą, kurio kamera užpildoma 1xTBE buferiu.
7. 2 µl produkto sumaišomi su 1 µl 3x įvedimo dažo (bromfenolio mėliu) ir įnešami į agarozės gelio šulinėlius. Į pirmąjį šulinėlį, norint nustatyti PGR produkto ilgį, įleidžiamas molekulinio ilgio

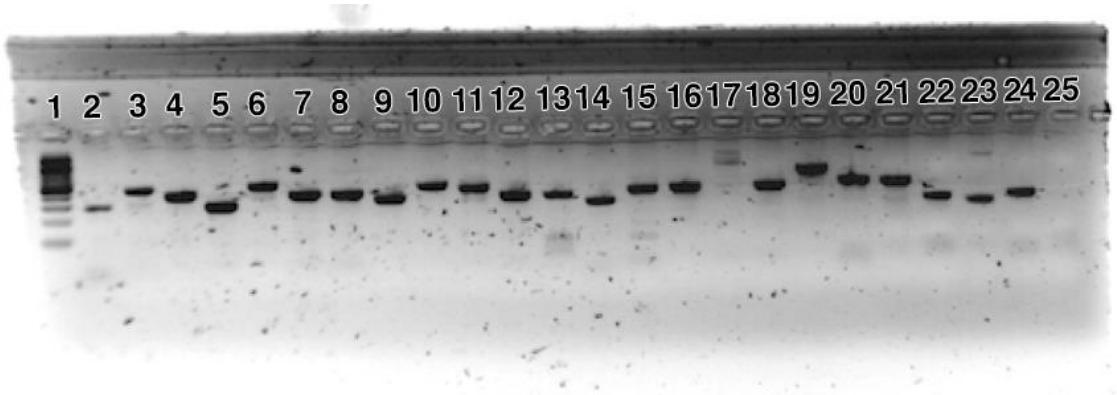
standartas (100bp), o į paskutinį – neigiamas kontrolinis mėginys (K) su įvedimo dažu, siekiant nustatyti ar nėra šalutinių PGR produktų bei reagentų ir mišinio reagentų užterštumo.

8. Elektroforezės aparatas uždengiamas, prijungiamas srovės šaltinis. Elektroforezė vykdoma esant 125V įtampai 15 min.

9. Pasibaigus elektroforezei, gelis fotografuojamas transiluminatoriuje ultravioletinėje šviesoje. Naudojant kompiuterinę *Easy Win 32* programą vizualizuojami ir įvertinami gauti rezultatai (4 pav.).

4 lentelė. Elektroforezės metu naudojamos medžiagos

Medžiagos	**Kiekis**
Agarozės gelio paruošimui naudojamos medžiagos	
SeaKem® LE Agarozė, Lonza, JAV	1,5 g
TBE 1X buferis, Carl Roth GmbH+ Co.KG, Vokietija	100 ml
Etidžio bromidas, Carl Roth GmbH+ Co.KG, Vokietija	2 µl
Fragmentų identifikavimui naudojamos medžiagos	
Bromfenolio mėlis (3X DNA loading dye), Thermo Fisher Scientific Baltics, Lietuva	1 µl
Molekulinis ilgio standartas (GeneRuler™ 100bp DNA Ladder), Thermo Fisher Scientific Baltics, Lietuva	1,5µl



4 pav. Agarozės gelio nuotrauka po elektroforezės (sąlygų optimizavimas 23-ims egzonams ar jų fragmentams ir kontrolei; pradmenų jungimosi temperatūra – 58 °C). Pirmajame takelyje matomas 100bp molekulinio ilgio standartas; 2-24 takeliuose matomi skirtingo ilgio produktai po PGR; 25-ame takelyje – neigiama kontrolė.

2.6 PGR produkto valymas fermentais (*EXOSAP*)

Siekiant atlikti sekoskaitą, būtina mėginius tinkamai paruošti. Tam naudojamos egzonukleazės (angl. *exonuclease, EXO*). Tai fermentai, kurių dėka yra degraduojami po PGR mišinyje likę nepanaudoti vienagrandžiai DNR fragmentai (pradmenys). Kitas fermentas, naudojamas šiame etape yra krevečių šarminė fosfatazė (angl. *Shrimp Alkaline Phosphatase, SAP*). Ji naudojama nepanaudotų dNTP defosforilinimui.

Metodo atlikimo eiga:

1. Į 0,5 ml tūrio mėgintuvėlius paruošiamas valymui skirtas mišinys. Mišiniui naudotas medžiagų kiekis nurodytas penktoje lentelėje. Visų medžiagų mišinys yra gerai supurtomas ir nucentrifuguojamas.

5 lentelė. EXOSAP mišiniui naudotas medžiagų kiekis

Mišinio komponentai	**Mišinio komponentų kiekis vienam mėginiui (µL)**
TE buferis (10 mMTris-HCl, 1 mM EDTA, pH 8), AppliChem GmbH, Vokietija	1,8
Krevečių šarminė fosfatazė (angl. *shrimp alkaline phosphatase, SAP*), Thermo Fisher Scientific Baltics,Lietuva	0,14
Egzonukleazė I (angl. *exonuklease I*), E.coli, Thermo Fisher Scientific Baltics, Lietuva	0,06

2. Į 0,2 ml tūrio mėgintuvėlius įpilama po 2µL mišinio.

3. Į mėgintuvėlius su mišiniu pridedama 5µL PGR produkto. Turinys gerai supurtomas ir nucentrifuguojamas.

4. Mėgintuvėliai sudedami į *Eppendorf Mastercycler (Eppendorf AG,* Vokietija*)* termociklerį vienai valandai, pasirenkant *EXOSAP* programą (40min 37°C, siekiant hidrolizuoti vienagrandžius DNR fragmentus ir dNTP likučius, ir 20min 80°C fermentų inaktyvacijai).

2.7 Sekoskaitos PGR

Pilnam DNR charakterizavimui būtina nustatyti jos nukleotidų seką. *Sanger* metodas dar vadinamas dideoksi sekoskaitos metodu, nes yra pagrįstas fluoroforais žymėtų dideoksiribonukleotidų (ddNTP) naudojimu. Šiame metode vienagrandė DNR naudojama kaip šablonas, pagal kurį *in vitro* sąlygomis bus sintetinama komplementari DNR grandinė. Sekoskaitos metu yra atliekamos reakcijos, kurių metu yra maža, skirtingomis spalvomis žymėtų ddNTP ir daug didesnė deoksiribonukleotidų (dNTP) koncentracija. Reakcijos metu ddNTP atsitiktinai prijungiamas komplementarioje dNTP vietoje. Dideoksiribonukleotidai 3' padėtyje vietoje hidroksilo grupių turi prijungtą vandenilį, todėl jam įsikorporavus grandinės sintezė yra terminuojama. Terminacija įvyksta dėl to, nes trečioje padėtyje nesant hidroksilo grupės, kitas nukleotidas nebegali susijungti fosfodiesteriniu ryšiu. Taip gaunami tos pačios sekos skirtingo ilgio fragmentai su prisijungusiu ddNTP sekos gale, kuriuos galima atskirti kapiliarinės elektroforezės metodo pagalba ir taip nustatyti visą tiesioginę nukleotidų seką.

Metodo atlikimo eiga:

1. Į 0,5 ml tūrio mėgintuvėlius paruošiamas sekoskaitos PGR skirtas mišinys. Šiame mišinyje esminis vaidmuo atitenka *BigDye Terminator* mišiniui, kuriame yra skirtingais fluoroforais žymėti ddNTP. Mišiniui naudotas medžiagų kiekis nurodytas šeštoje lentelėje. Visų medžiagų mišinys yra gerai supurtomas ir nucentrifuguojamas.

6 lentelė. Sekoskaitos PGR mišiniui naudotas medžiagų kiekis.

Mišinio komponentai	**Mišinio komponentų kiekis vienam mėginiui (µL)**
Dejonizuotas vanduo	3,15
Sekoskaitos buferis (*Big Dye® terminator, 5* Sequencing Buffer*), *Applied Biosystems*™, Didžioji Britanija	1,4
Sekoskaitos mišinys (*Big Dye® terminator Mix*), *Applied Biosystems*™, JAV	0,25

2. Į 0,5 ml tūrio mėgintuvėlius įpilama po 4,8 µL mišinio.
3. Į mėgintuvėlius su mišiniu pridedama 2µL išvalyto PGR produkto.

4. Į mėgintuvėlius pridedama po 0,2 µL pradmens (tiesioginis (angl. *forward*)- naudojamas tik vienas iš pradmenų, siekiant amplifikuoti tik vienagrandį fragmentą). Turinys gerai supurtomas ir nucentrifuguojamas.

5. Mėgintuvėliai sudedami į *Eppendorf Mastercycler (Eppendorf AG,* Vokietija*)* termocikleri dviem valandoms. Termociklerio programos nustatymai pateikiami septintoje lentelėje.

7 lentelė. Sekoskaitos PGR programos etapai ir sąlygos

Ciklas	Temperatūra (°C)	Trukmė (s)	Ciklų skaičius
Pradinė denatūracija	96	60	1
Denatūracija	96	11	23
Pradmens prisijungimas	52	11	
Sintezė	60	240	
Inkubacija	4	∞	1

2.8 PGR produkto valymas etanoliu

Šis etapas reikalingas perteklinių nukleotidų, ddNTP bei pradmenų pašalinimui iš mišinio po sekoskaitos polimerazės grandininės reakcijos.

Metodo atlikimo eiga:

1. Į 0,5 ml tūrio mėgintuvėlius paruošiamas antrajam valymui skirtas mišinys. Mišiniui naudotas medžiagų kiekis nurodytas aštuntoje lentelėje. Visų medžiagų mišinys yra gerai supurtomas ir nucentrifuguojamas.

8 lentelė. PGR produkto valymui etanoliu naudotas medžiagų kiekis

Mišinio komponentai	Mišinio komponentų kiekis vienam mėginiui (µL)
Etanolis (96 %; AB „Vilniaus degtinė“, Lietuva)	62,5
Dejonizuotas vanduo	14,5
Natrio acetatas (Applichem GmbH, Vokietija)	3

2. Į mėgintuvėlius su PGR produktu pridedama 80µL mišinio.
3. Inkubuojama 5 minutes tamsoje, siekiant išsaugoti sekoskaitos PGR metu naudotų ddNTP fluoroforų aktyvumą.
4. Centrifuguojama 13 000 aps/min greičiu 5 minutes. Dekantuojama.
5. Į dekantuotus mėgintuvėlius pridedama 300 µL 70° etanolio. Supurtoma 1 minutę.
6. Centrifuguojama 13 000 aps/min greičiu 5 minutes. Dekantuojama. Mėgintuvėliai patalpinami į termociklerį 5 minutėms 37°C temperatūroje, siekiant išgarinti likusį etanolį.

2.9 Kapiliarinė elektroforezė ir sekų analizė

Tiesioginės vienagrandžių DNR fragmentų sekos yra nustatomos kapiliarinės elektroforezės pagalba, kuomet detektorius fiksuoja žymėto ir sužadinto genetiniame analizatoriuje ddNTP skleidžiamą šviesos bangos ilgį.

Metodo atlikimo eiga:
1. Į kiekvieną mėgintuvėlį su paruošta sekoskaitai DNR įnešama po 8 µl Hi-DiTM formamido (*Applied Biosystems™,* Didžioji Britanija).
2. Visas turinys supipetuojamas ir įnešamas į sekoskaitos plokštelę. Centrifuguojama 1 min. 2000 aps/min greičiu.
3. Sekoskaitos plokštelė su mėginiais dedama į genetinį analizatorių *3130xl Genetic Analyzer (Applied biosystems™, Life technologies,* JAV*)*. Atliekama kapiliarinė elektroforezė.
Duomenys yra gaunami sekvenogramos pavidalu ir yra analizuojami specialiomis programomis. Pirminė duomenų analizė vykdoma naudojantis *Sequencing Analysis 5.2 (Applied biosystems™, Life technologies, JAV*) programa. Toliau naudojama internete prieinama *Chromas lite 2.1* (*Technelysium Pty Ltd)* programa sekvenogramos vertinimui ir pokyčių nustatymui. Sekvenogramos lyginamos su referentinėmis DNR sekomis pateiktomis *Ensembl (Ensembl release 78 - December 2014 ©WTSI*) duomenų bazėje. Kai kuriais atvejais, gautos sekos buvo prilygiuojamos prie referentinio genomo naudojantis internete prieinamu *BLAST* prilygiavimo įrankiu Analizuojant informaciją, esančią duomenų bazėse yra identifikuojami pokyčiai.

2.10 Bioinformacinė ir statistinė analizė, naudotos programos

Variantai, kurie nebuvo patvirtinti *Sanger* metodu bei variantai, kurie buvo nustatyti *Sanger* metodu, bet nebuvo žinoma informacija apie juos NKS duomenyse, buvo patikrinti NKS duomenų BAM failus vizualizuojant integruota genomo vaizduokle *IGV (2.3 versija).*

Siekiant palyginti skirtingas praturtinimo sistemas, variantų suskirstymui į dvi grupes pagal praturtinimo sistemą, apjungiant skirtingų tiriamųjų genomo variantus, buvo naudota *GATK CombineVariants (3.7 versija)* programa.

Variantų anotavimui buvo naudotas *ANNOVAR* programinis įrankis.

Statistiniams skaičiavimams ir rezultatų vizualizavimui naudota *Microsoft Office Excel 2015* programa ir laisvos prieigos *R Commander 3.2.4* programa. Statistinis patikimumas buvo įvertintas panaudojant Vilkoksono kriterijų. Statistiškai reikšmingais buvo laikomi skirtumai, kurių $p < 0,05$.

3. REZULTATAI IR JŲ APTARIMAS

3.1 *Sanger* sekoskaitos duomenų analizė ir vizualizavimas

Darbo metu buvo optimizuotos PGR sąlygos 60-čiai genų egzonų ar jų fragmentų. Jiems visiems buvo atlikta *Sanger* sekoskaita, siekiant įvertinti naudoto automatinio analizės algoritmo patikimumą. Iš viso buvo nusekvenuota 747 genomo fragmentų iš 890. Nenusekvenuota genomo fragmentų dalis (143 fragmentai) yra dėl ribotų DNR resursų šiam moksliniam tyrimui.

Trys genomo variantai, kurie buvo identifikuoti NKS metu, *Sanger* sekoskaita nebuvo nustatyti nei viename tiriamajame. Pastarųjų genomo variantų identifikaciniai numeriai dbSNP duomenų bazėje yra šie: rs11147976 (*SKA3*), rs12764004 (*BMS1*) ir rs4913758 (*BAGE2*).

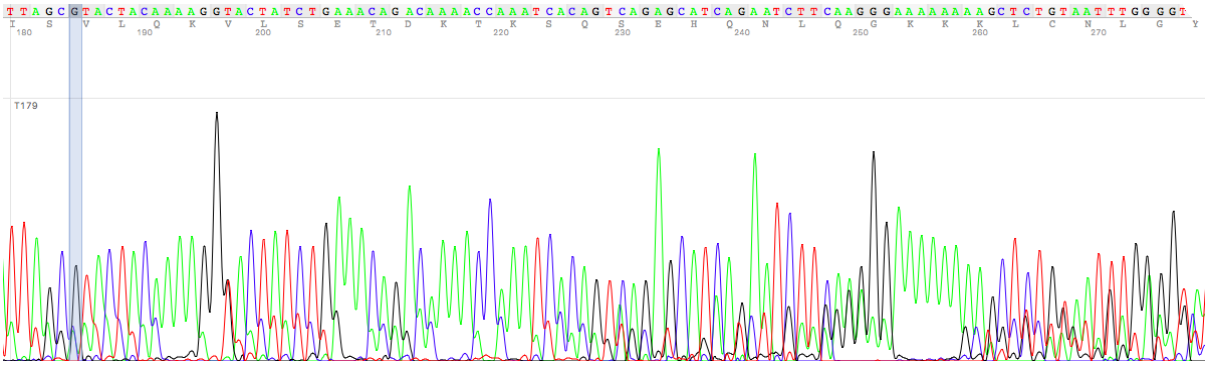
Visų šių variantų DNR fragmentai iš NKS duomenų buvo patikrinti juos vizualizuojant integruota genomo vaizduokle (IGV). Šioje programoje tikrinami pagrindiniai BAM faile esantys parametrai (*CIGAR*, kiekvieno nukleotido ir DNR fragmento kokybės įverčiai bei padengimas), gauti NKS *SOLiD* platformos pagalba.

Patikrinus variantą *SKA3* (rs11147976) gene nustatyta, kad DNR fragmentų kokybės įverčiai yra aukšti. Didžiosios dalies DNR fragmentų *CIGAR* buvo lygus 75M, o tai yra geriausias galimas variantas su šiais duomenimis. DNR fragmento kokybės įvertis lygus 95 iš 100, nukleotido kokybės įvertis lygus 41 iš 41, o padengimas geras (vidutinis padengimas lygus 38). Atsižvelgus į kokybės įverčius, galima daryti išvadą, kad NKS metodas tiksliai įvardino esamą variantą. Labiausiai tikėtina, kad *Sanger* metodu šis variantas buvo nenustatytas dėl metodo klaidos.

BMS1 (rs12764004) geno variantas taip pat nebuvo nustatytas *Sanger* metodu. Sekų kokybė buvo prasta, egzono pabaigoje stebimi dvigubi pikai. Patikrinus šį variantą IGV, pastebėta, kad nukleotidų kokybės įverčiai yra aukšti (daugumai fragmentų lygus 41), vidutinis padengimas lygus 44, o tai turėtų indikuoti apie gerą atskiro nukleotido kokybę ir šių parametrų vertės gali būti įrodymu, kad variantas tiriamiesiems yra nustatomas. Tačiau, didelis padengimas ir prastas DNR fragmento kokybės įvertis (apie 8) parodo, kad DNR fragmentas yra homologiškas keliose genomo vietose, todėl sukurti pradmenys pagausina DNR fragmentus iš skirtingų genomo vietų. Todėl *Sanger* sekoskaitos metodu matome, kad tirtoje genomo srityje pokyčio nėra, o NKS metodu pokytis fiksuojamas. Be to, dėl homologinių sričių genome, vykstant naujos kartos sekoskaitos metu gautų fragmentų prilygiavimui prie referentinio genomo, fragmentai gali būti prilygiuojami netiksliai.

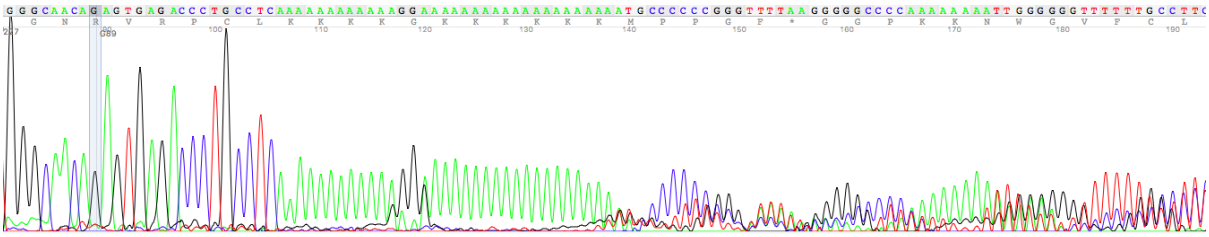
Panaši situacija buvo nustatyta ir vizualizavus variantą *BAGE2* (rs4913758) gene. Buvo pastebėta, kad kokybės įverčiai prasti, ypač žemas DNR fragmentų kokybės įvertis (kai kurių fragmentų įverčiai siekė tik 4). Tai byloja apie DNR fragmentų homologiškumą genome. *BAGE2* genas, esantis 21 chromosomoje, turi homologišką geną 22 chromosomoje (*BAGE5*). Taip pat yra ir kitų homologinių genų šioje B melanomos antigenų šeimoje. Visi šie gauti rezultatai leidžia daryti išvadą, kad atliekant NKS duomenų analizę yra būtina atsižvelgti į sekoskaitos ir duomenų analizės kokybės įverčius.

Taip pat, buvo nustatyta pora įdomesnių atvejų, analizuojant *Sanger* sekoskaitos duomenis. Pirmasis yra susijęs su *ANKRD62* gene esančiu variantu (rs201537483). Jis buvo nustatytas 19 asmenų iš 21, tačiau išsiaiškinta, kad šis genas turi 93 % analogišką pseudogeną (*ANKRD62P1*) 22 chromosomoje, todėl visiems tiriamiesiems nuo egzono vidurio iki galo buvo stebėti dvigubi pikai, kurie iš pirmo žvilgsnio buvo panašūs į pokytį, sukeliantį rėmelio poslinkį (5 pav). Daliai tiriamųjų dėl minėtų dvigubų, o kartais ir trigubų pikų buvo sudėtinga nustatyti tikrąjį varianto tipą.



5 pav. ANKRD62 pokyčio sekvenograma

Kitas variantas (rs2235182), kurį analizuojant kilo sunkumų, buvo *CFP* gene. Nusekvenavus *Sanger* metodu šio geno egzono ir šalia esančio introno dalį buvo pastebėta, kad sekos kokybė suprastėja po poliA regiono, matomi dvigubi pikai (6 pav). Šiuo atveju, tai *Sanger* metodo trūkumas sekvenuojant homogeniškus regionus – patiriamas DNR sekos praslydimas (angl. *DNA Sequence Slippages*). Patikrinus IGV programoje, nustatyta, kad nors padengimas ir mažas (vidutinis padengimas lygus 6), tačiau visi likusieji kokybės įverčiai rodo, kad pokytis yra. Deja, jis nebuvo patvirtintas *Sanger* metodu dėl minėto ribotumo.



6 pav. CFP geno egzono fragmento sekvenograma

Be to, vykdant *Sanger* sekoskaitą, buvo nustatyti trys variantai, kurių pavienių tiriamųjų rezultatai nesutapo su NKS duomenimis. Šie variantai yra: rs2824804 (*TMPRSS15*), rs16903574 (*FAM105A*), rs11062745 (*CRACR2A*).

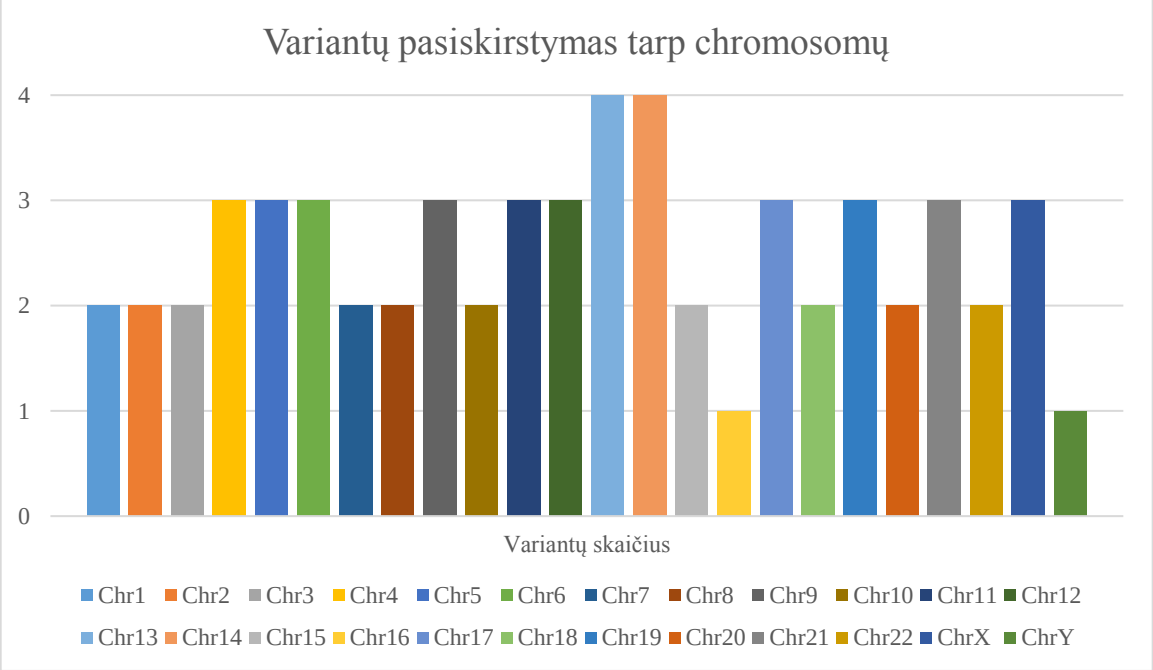
TMPRSS15 geno variantas buvo patvirtintas 13 iš 14 tiriamųjų (92,9 %). Vieno tiriamojo seka, kuriam geno variantas patvirtintas nebuvo, buvo išanalizuota IGV. Įdomu tai, kad šis variantas nebuvo nustatytas ir IGV programoje. Tai reiškia, kad analizės algoritmas teisingų variantų atrankos etape klaidingai priskyrė šį variantą ir tai yra NKS algoritmo klaida.

Variantas *FAM105A* gene nebuvo patvirtintas vienam tiriamajam (patvirtinti 28 iš 29-ių tiriamųjų, t.y. 96,67 %). Šis variantas, kaip ir variantas *TMPRSS15* gene, nebuvo nustatytas ir vizualizuojant NKS duomenis IGV programoje. Tai taip pat reiškia NKS analizės algoritmo klaidą.

CRACR2A gene esantis variantas buvo nustatytas 7 iš 8 tiriamųjų (87,5 %). Tiriamajam, kuriam nebuvo nustatytas pokytis *Sanger* metodu, NKS duomenys buvo vizualizuojami IGV. Visi kokybės įverčiai rodo, kad pokytis nurodytoje genomo pozicijoje yra. Todėl labiausiai tikėtina priežastis, dėl kurios šis variantas nenustatytas tiriamajam *Sanger* metodu – paties *Sanger* metodo klaida. Svarbu paminėti, kad dažniausiai mokslinėse publikacijose *Sanger* metodas nors ir yra laikomas "auksiniu standartu" (5, 6, 7, 8), tačiau gauti rezultatai prieštarauja šiam teiginiui ir parodo, kad metodas taip pat turi savo trūkumų. Likusių patvirtintų variantų skaičius buvo lygus 100 %, t.y. visiems tirtiems tiriamiesiems pokytis buvo patvirtintas ir *Sanger* metodu.

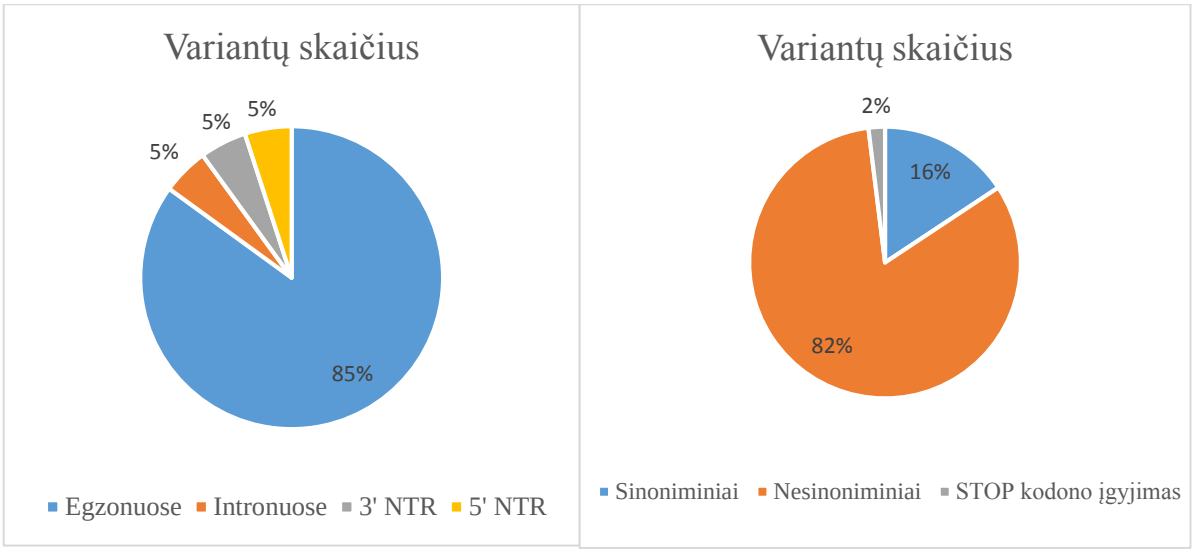
3.2 Aprašomoji variantų statistika

Naudojantis *ANNOVAR* programiniu įrankiu, visi tirti variantai buvo anotuoti – prie turimų duomenų buvo prijungta informacija iš duomenų bazių apie kiekvieną tiriamojo variantą.
Variantai darbui buvo pasirinkti atsitiktinai, tačiau taip, kad būtų pasiskirstę tolygiai visame genome – visose chromosomose buvo tirta bent po vieną variantą. Daugiausiai variantų buvo tirta 13 ir 14 chromosomose (4 variantai), mažiausiai – 16 ir Y chromosomose (1 variantas) (7 pav.).



7 pav. Variantų pasiskirstymas tarp chromosomų

Didžioji dalis genomo variantų buvo egzoniniai, nes buvo naudojami viso egzomo sekoskaitos duomenys. Kadangi naudojamos praturtinimo sistemos aprėpia ne tik egzonus, bet ir gretimas aplinkines sritis, todėl vykdant egzomo sekoskaitą identifikuoti variantai, esantys ir intronuose ar netransliuojamose geno dalyse. Todėl darbo metu buvo įtraukti ir pastarieji variantai, kurie buvo nustatyti viso egzomo sekoskaitos metu.



8 pav. Variantų skaičiaus pasiskirstymas genome (pav. kairėje) ir variantų tipai (pav. dešinėje)

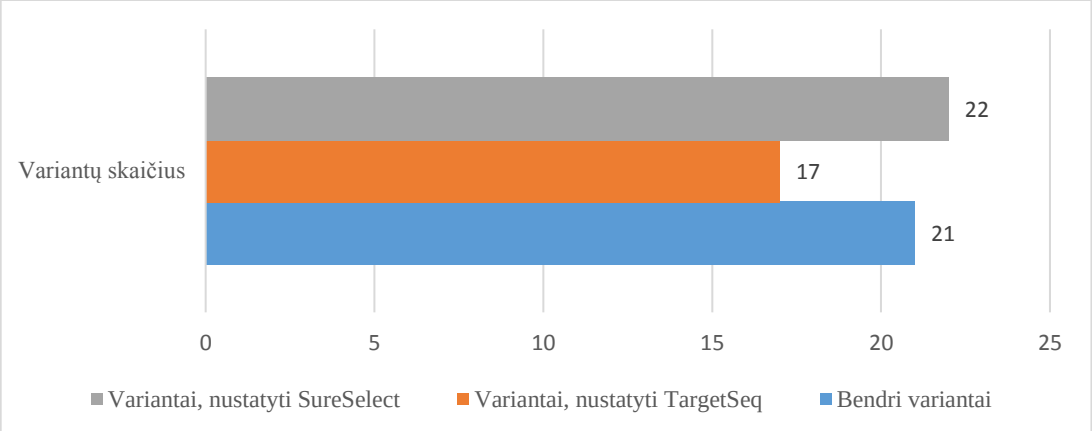
Variantų, nustatytų intronuose ir netransliuojamose geno dalyse, dalis siekė 15 %. Tuo tarpu likusią dalį sudarė variantai, esantys egzonuose (8 pav.). Iš jų, 16 % tirtų variantų nekeitė koduojamos aminorūgšties, todėl yra sinoniminiai. Tuo tarpu, didžioji dalis tirtų variantų yra nesinoniminiai, t.y. 82 % visų variantų. Vienintelis variantas iš tirtųjų lėmė STOP kodono susidarymą – tai atitiko 2 % variantų egzonuose. Šis variantas (rs28502153) yra *GAB4* gene.

3.3 Praturtinimo sistemų palyginimas

Šiuo metu yra sukurta keletas skirtingų taikinių praturtinimo sistemų bei nemažai skirtingų kiekvienos sistemos atnaujinimų, siekiant užtikrinti kuo efektyvesnį egzomo ar konkrečių genų atskyrimą nuo likusios genetinės medžiagos. Šio darbo metu NKS duomenys apie tiriamus variantus buvo gauti naudojant dvi skirtingas taikinių praturtinimo sistemas. Pirmoji – tai *SureSelect* taikinių praturtinimo sistema. Ši sistema yra laikoma geresne dėl nustatomo didesnio variantų skaičiaus bei geresnio variantų padengimo. Naujausiose mokslinėse publikacijose, kuriuose lyginamos skirtingos praturtinimo sistemos, naujausia *SureSelect* sistemos versija minima, kaip pasiekianti didžiausią koduojančių regionų padengimą (22). Antroji sistema *TargetSeq*. Nors ši sistema apima mažiau genomo regionų, tačiau yra tokių, kurie nustatomi tik šia praturtinimo sistema. Todėl mokslinėse publikacijose minima ne tik skirtingų sekoskaitos platformų, bet ir skirtingų praturtinimo sistemų naudojimo svarba, siekiant nustatyti kuo daugiau variantų (24).

Šiame darbe, siekiant palyginti praturtinimo sistemas, naudojant *GATK Combine Variants* programą buvo apjungti skirtingų tiriamųjų genomo variantai. Vėliau NKS duomenys buvo suskirstyti į dvi

grupes – variantai, kurie buvo nustatyti *TargetSeq* ir variantai, kurie buvo nustatyti *SureSelect* sistemos pagalba (9 pav.). Praturtinimo sistemos buvo lyginamos pagal vieną pagrindinių NKS parametrų – padengimą. Naudojantis NKS kokybės duomenimis, buvo apskaičiuoti kiekvieno varianto padengimo vidurkiai. Taip pat, atliktas bendrų variantų, nustatytų abejomis sistemomis, palyginimas.



9 pav. Variantų skaičius, nustatytas skirtingomis praturtinimo sistemomis

Bendrų variantų padengimo vidurkiai bei bendras visų bendrų variantų padengimo vidurkis yra pateikiamas devintoje lentelėje. Mažiausias vidutinis padengimas (DP=4) buvo nustatytas variantui, esančiame *GNG7* geno 5' netransliuojamame regione. Šiuo atveju naudota praturtinimo sistema buvo *SureSelect*. Tai galima paaiškinti remiantis varianto genomine pozicija – ši praturtinimo sistema naudojama genomo koduojančių dalių analizei, todėl geriausias padengimas bus koduojančiose srityse. Tuo tarpu, nekoduojančioms genomo dalims padengimas yra ženkliai mažesnis.

Didžiausias vidutinis padengimas (DP=162,3) buvo nustatytas naudojant *TargetSeq* praturtinimo sistemą. Šis variantas yra *SIRPB1* gene.

Didesnė dalis (57,14 %) variantų turėjo didesnį kaip 20x vidutinį padengimą naudojant *SureSelect* sistemą, kai *TargetSeq* tokį vidutinį padengimą pademonstravo tik 42,84 % variantų. Daugeliu atvejų regionai, kurie turėjo prastą padengimą, buvo gausiuose GC nukleotidų regionuose, homogeniškuose regionuose arba buvo egzono pradžioje ar pabaigoje.

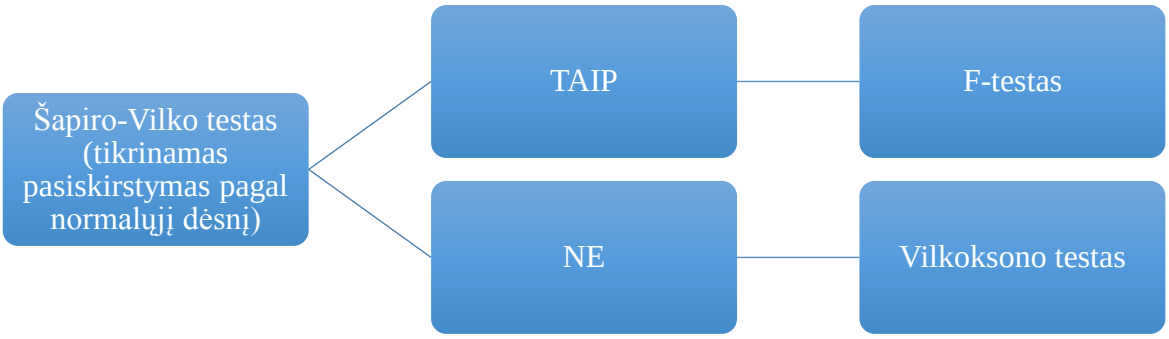
9 lentelė. Bendrų variantų vidutinio padengimo reikšmių palyginimas

Genas	Varianto dbSNP identifikacijos nr.	SU DP vidurkis	TS DP vidurkis
SKA3	rs11147976	27,33	47,67
HNRNPC	rs61745722	33,25	39,60
ZNF721	rs72501950	65,00	51,00
FAM105A	rs16903574	50,09	19,36
DSP	rs6929069	102,31	34,71
TMEM184A	rs17852421	6,55	7,00
SMARCA2	rs2296212	38,09	49,50
NLRP6	rs11246050	9,70	10,78
FGF23	rs7955866	13,93	12,50
SPATA13	rs9511156	51,38	4,30
EDDM3A	rs34552133	46,43	49,00
CYFIP1	rs7170637	42,70	13,38
ABR	rs34169260	35,15	15,17
POLRMT	rs14155	12,60	32,67
SIRPB1	rs41275422	75,89	162,33
TMPRSS15	rs2824804	33,58	36,00
GAB4	rs28502153	9,44	17,33
ARSD	rs211653	11,59	18,08
EMILIN2	rs1790994	11,75	15,57
GNG7	rs1133692	4,00	9,32
UTY	rs9341278	7,50	17,89
Bendras vidurkis		**32,774**	**31,579**

SU – SureSelect; TS – TargetSeq; DP – vidutinis padengimas

Remiantis gautais bendrais vidurkiais (9 lentelė) galima pastebėti, kad iš tiesų *SureSelect* praturtinimo sistemos bendras padengimas yra nežymiai (3,65 %) didesnis. Siekiant patikrinti, ar šis skirtumas yra statistiškai reikšmingas, buvo taikytas neparametrinis Vilkoksono testas, nes duomenys nebuvo pasiskirstę pagal normalųjį dėsnį. Normalumo patikrinimui buvo naudotas Šapiro – Vilko testas, kuris daugelio autorių yra laikomas vienu galingiausių testų (39). Remiantis Vilkoksono testu buvo gauta p

reikšmė. Ja vadiname mažiausią reikšmingumo lygmenį, kuriam teisinga nulinė hipotezė gali būti atmesta (turimiems duomenims), t.y. jei p reikšmė yra mažesnė už 0,05, tokiu atveju nulinė hipotezė yra atmetama, ir atvirkščiai. Pritaikius pasirinktą testą, p reikšmė buvo lygi 0,7079, t.y. nulinė hipotezė buvo priimta – skirtingų praturtinimo sistemų padengimo vidurkiai statistiškai reikšmingai nesiskiria. Apibendrinta statistinės analizės schema pateikiama 10-ame paveiksle.



10 pav. Apibendrinta statistinės analizės schema

Taip pat, iš visų tiriamųjų, dviems tiriamiesiems VU MF ŽMGK darbuotojai atliko viso egzomo sekoskaitą abiem praturtinimo sistemomis. Gauti duomenys leido tiksliau palyginti abi praturtinimo sistemas. Iš analizuojamų variantų, daugiau buvo nustatyta su *SureSelect* praturtinimo rinkiniu (šeši variantai nustatyti vienam tiriamajam, aštuoni kitam). Svarbu pabrėžti, kad nors buvo nustatyta mažiau variantų naudojant *TargetSeq* sistemą (keturi pirmam tiriamajam ir trys antram tiriamajam), tačiau buvo nustatyti variantai, kurie nebuvo identifikuoti *SureSelect* sistema. Tai leidžia daryti prielaidą, kad tiksliausiam variantų nustatymui efektyviausia būtų naudoti abi praturtinimo sistemas. Panašią išvadą padarė ir *Zhang* su kolegomis, savo tyrimo metu nustatę, jog vykdant egzomo sekoskaitą ir naudojant abi praturtinimo sistemas, buvo pasiektas didesnis specifiškumas bei jautrumas NKS analizės algoritmo variantų šaukimo etape (24).

Remiantis šiais duomenimis, taip pat buvo atliktas abejomis sistemomis nustatytų bendrų variantų padengimo vidurkių palyginimas. Bendras padengimo vidurkis su *TargetSeq* ir *SureSelect* atitinkamai yra 21,5 ir 33,5. *SureSelect* vidutinis padengimas ženkliai didesnis (35,82 %). Tačiau svarbu paminėti, kad tokių bendrų variantų buvo tik trys. Tokios imties nepakanka normalumo įvertinimui Šapiro – Vilko testu bei tolimesniems statistiniams skaičiavimams, nes tokiu atveju testas netenka galios atmesti nulinę hipotezę ir tokiu būdu nedidelės imtys dirbtinai patenkina normalųjį dėsnį. Todėl tikėtina, kad esant didesnei imčiai, t.y. išanalizavus daugiau bendrų variantų, vidutinis padengimas tarp skirtingų praturtinimo sistemų taptų panašus.

3.4 *SOLiD* metodo tikslumo, jautrumo ir specifiškumo nustatymas

Siekiant nustatyti, ar tyrimo metodo rezultatai yra tikslūs, laboratorijose kiekvienas tyrimo etapas yra vertinamas ir atidžiai stebimas. Kiekvieno tyrimo metodo kokybė turi būti vertinama – svarbu įvertinti metodo tikslumą, jautrumą bei specifiškumą (40). Puikus pavyzdys yra paveldimų ligų NKS genų rinkinių tyrimai, kuomet metodo specifiškumo ir jautrumo nustatymas yra itin svarbus, nes pagal tyrimo atsakymą suformuojama diagnozė ir tolimesnė gydymo taktika (4).

Atlikus NKS metodu nustatytų variantų tvirtinimą *Sanger* sekoskaita buvo apskaičiuotas darbe vykdyto NKS *SOLiD* metodo automatinio analizės algoritmo tikslumas. Šiame darbe visi skaičiavimai buvo atlikti, naudojantis duomenimis, esančiais žemiau pateiktose 10-oje ir 11-oje lentelėse. Metodo tikslumas siekė 99,66 %, t.y. tokia dalis NKS duomenų sutapo su *Sanger* sekoskaitos metodu gautais duomenimis. Šis skaičius buvo artimas *Biesecker* ir kolegų gautiems duomenims – jų atlikto tyrimo metu metodo tikslumas 99,97 % (41). *Harismendy* ir kolegų duomenimis, *SOLiD* metodo tikslumo reikšmė taip pat buvo labai artima šiame darbe gautajai reikšmei – ji siekė 99,7 % (42).

Svarbu paminėti, kad vertinant NKS metodo tikslumą, jautrumą ir specifiškumą buvo neįtraukti duomenys, gauti tikrinant variantą *CFP* gene. Nors šio varianto kokybės įverčiai NKS duomenyse rodo, kad šis variantas yra pateiktoje genomo pozicijoje, tačiau dėl *Sanger* metodo ribotumo jo negalima įvertinti ir jis negali būti įtrauktas į imtį. Taip pat, dėl tos pačios *Sanger* metodo ribotumo priežasties, į imtį nebuvo įtraukti ir *SKA3* bei *CRACR2A* genuose esantys variantai.

10 lentelė. Tirti genomo variantai bei jų tarpe patvirtintų ir nepatvirtintų variantų skaičius Sanger metodu

Genas	dbSNP ID	Ref	Alt	Tiriamųjų skaičius	*Sanger metodu įvertintų tiriamųjų skaičius*	Patvirtintų variantų skaičius (%)	*Sanger metodu neįvertintų tiriamųjų skaičius*	*Sanger metodu neįvertintų tiriamųjų skaičius (%)*
EFCAB5	rs9900546	G	T	8	8	8 (100%)	0	0%
SPATA7	rs4904448	G	A	3	2	2 (100%)	1	33,33%
SKA3	rs11147976	T	C	20	18	0 (0%)	2	10%
MIPEP	rs11551114	G	A	4	4	4 (100%)	0	0%
HNRNPC	rs61745722	G	A	9	7	7 (100%)	2	22,00%

Genas	dbSNP ID	Ref	Alt	Tiriamųjų skaičius	*Sanger metodu įvertintų tiriamųjų skaičius*	Patvirtintų variantų skaičius (%)	*Sanger metodu neįvertintų tiriamųjų skaičius*	*Sanger metodu neįvertintų tiriamųjų skaičius (%)*
PLEKHG4B	rs29674	A	C	13	10	10 (100%)	3	23,00%
SNRNP48	rs2757594	C	T	8	6	6 (100%)	2	25%
DMKN	rs12460932	G	T	12	11	11 (100%)	1	8,30%
PERM1	rs13303368	C	G	6	6	6 (100%)	0	0%
RP1L1	rs9657518	A	G	5	3	3 (100%)	2	40%
ZCCHC7	rs151266778	G	A	3	3	3 (100%)	0	0%
DSP	rs6929069	G	A	20	13	13 (100%)	7	35%
ZNF721	rs72501950	G	A	15	14	14 (100%)	1	6,67%
BAGE2	rs4913758	C	T	13	13	0 (0%)	0	0%
ANKRD62	rs201537483	G	A	21	21	19 (90,48%)	0	0%
CIDEC	rs456168	G	T	9	4	4 (100%)	5	55,56%
ERCC6L	rs45448501	G	A	3	3	3 (100%)	0	0%
FAM105A	rs16903574	C	G	30	29	28 (96,67%)	1	3,33%
DHTKD1	rs1279138	T	C	25	21	21 (100%)	4	16%
FAM200B	rs4235380	G	A	13	10	10 (100%)	3	23,08%
MYT1L	rs3748989	G	A	7	6	6 (100%)	1	14,29%
B4GALNT4	rs34063493	C	T	10	7	7 (100%)	3	30%
CRACR2A	rs11062745	A	G	10	8	7 (87,5%)	2	20%
TMEM184A	rs17852421	G	A	13	11	11 (100%)	2	15,40%
SMARCA2	rs2296212	C	G	19	15	15 (100%)	4	21,10%
NLRP6	rs11246050	G	A	19	18	18 (100%)	1	5,30%
FGF23	rs7955866	G	A	17	16	16 (100%)	1	5,90%

Genas	dbSNP ID	Ref	Alt	Tiriamųjų skaičius	*Sanger metodu įvertintų tiriamųjų skaičius*	Patvirtintų variantų skaičius (%)	*Sanger metodu neįvertintų tiriamųjų skaičius*	*Sanger metodu neįvertintų tiriamųjų skaičius (%)*
SPATA13	rs9511156	G	A	21	16	16 (100%)	5	23,80%
EDDM3A	rs34552133	G	T	20	13	13 (100%)	7	35,00%
CYFIP1	rs7170637	G	A	18	15	15 (100%)	3	16,70%
ABR	rs34169260	T	C	19	17	17 (100%)	2	10,50%
POLRMT	rs14155	G	C	40	34	34 (100%)	6	15,00%
SIRPB1	rs41275422	A	G	18	16	16 (100%)	2	11,10%
TMPRSS15	rs2824804	T	C	16	**14**	**13 (92,9%)**	2	12,50%
GAB4	rs28502153	C	A	20	14	14 (100%)	6	30,00%
ARSD	rs211653	G	C	51	42	42 (100%)	9	17,60%
C1orf167	rs4845880	G	A	7	7	7 (100%)	0	0,00%
ANKRD36	rs10171441	C	A	62	59	59 (100%)	3	4,80%
XPC	rs2228000	G	A	19	17	17 (100%)	2	10,50%
WHSC1	rs139753036	C	A	7	7	7 (100%)	0	0,00%
DAP	rs5745297	G	A	13	11	11(100%)	2	15,40%
SYCP2L	rs6456746	G	A	8	7	7 (100%)	1	12,50%
GPR146	rs55677825	G	A	6	5	5 (100%)	1	16,70%
SGK223	rs4840953	G	C	9	9	9 (100%)	0	0,00%
CCDC107	rs10441685	C	G	4	4	4 (100%)	0	0,00%
BMS1	rs12764004	G	A	18	17	0 (0%)	1	5,60%
PHRF1	rs35482931	C	T	4	3	3 (100%)	1	25,00%
GPRC5D	rs3741822	G	T	9	8	8 (100%)	1	11,10%
LATS2	rs2770928	C	T	12	12	12 (100%)	0	0,00%

Genas	dbSNP ID	Ref	Alt	Tiriamųjų skaičius	*Sanger* metodu *įvertintų tiriamųjų skaičius*	Patvirtintų variantų skaičius (%)	*Sanger* metodu *neįvertintų tiriamųjų skaičius*	*Sanger* metodu *neįvertintų tiriamųjų skaičius (%)*
MDGA2	rs200459170	G	A	6	6	6 (100%)	0	0,00%
DISP2	rs12443160	T	C	8	7	7 (100%)	1	12,50%
ZNF469	rs45504291	G	A	15	11	11 (100%)	4	26,70%
KRT31	rs6503629	G	T	8	8	8 (100%)	0	0,00%
EMILIN2	rs1790994	T	C	26	19	19 (100%)	7	26,90%
GNG7	rs1133692	T	C	23	13	13 (100%)	10	43,50%
WFDC10B	rs2072974	G	A	22	13	13 (100%)	9	40,90%
TFF3	rs11701143	T	C	6	6	6 (100%)	0	0,00%
MICAL3	rs5992941	T	C	9	8	8 (100%)	1	11,10%
UTY	rs9341278	G	A	22	15	15 (100%)	7	31,80%
CFP	rs2235182	A	G	9	7	?	2	22,20%
Iš viso				890	747		143	

Raudona spalva pažymėti variantai, kurie nebuvo patvirtinti *Sanger* metodu.

Metodo jautrumas – tai gebėjimas tiksliai nustatyti variantus genome, kurie iš tikrųjų yra. Šie variantai taip pat vadinami teisingai teigiamais (angl. *true positive*). Taigi, jautrumas apskaičiuojamas dalinant teisingai teigiamų variantų skaičių iš teisingai teigiamų ir klaidingai neigiamų variantų skaičiaus sumos (43).

Tam, kad būtų apskaičiuoti visi klaidingai neigiami variantai, buvo sudarytas ne tik tirtų, bet ir *Sanger* sekoskaitos duomenų analizės metu nustatytų visų likusių variantų sąrašas. Apibendrinti duomenys apie nustatytus variantus pateikiami vienuoliktoje lentelėje. Šie variantai buvo patikrinti NKS duomenyse, juos vizualizuojant IGV programos pagalba. Taip siekta įsitikinti, ar nustatyti variantai *Sanger* duomenyse taip pat yra nustatomi ir NKS metodu. NKS analizės algoritmas laiko, kad variantas yra tiriamuose duomenyse tuomet, kai bendras padengimas yra daugiau nei šeši, nukleotido kokybės įvertis ne mažiau kaip 20, o DNR fragmento kokybės įvertis ne mažiau kaip 30. Kuomet iš visų DNR fragmentų daugiau kaip du turėjo alternatyvųjį alelį, tuomet taip pat laikyta, kad pokytis

yra ir NKS duomenyse. Iš nustatytų 23 variantų 111-ai tiriamųjų, variantai atitiko NKS analizės algoritmo parametrus ir buvo patvirtinti 78-iems tiriamiesiems. Likusiems 33-ims tiriamiesiems variantai nebuvo patvirtinti, tačiau iš jų 27-iems jie nebuvo patvirtinti dėl kokybės įverčių, aptartų anksčiau, neatitikimo. Todėl šie tiriamieji buvo atmesti iš imties skaičiuojant klaidingai neigiamų variantų reikšmę. Šešiems tiriamiesiems variantai nebuvo nustatyti NKS duomenyse, tose genomo pozicijose kokybės įverčiai buvo aukšti. Šie šeši tiriamieji yra laikomi klaidingai neigiamais variantais, kuomet NKS metodu nustatytas DNR varianto referentinis, o atlikus *Sanger* sekoskaitą – alternatyvus alelis (44). Remiantis aptartomis gautomis reikšmėmis, metodo jautrumas siekė 99,22 %.

11 lentelė. Sanger sekoskaitos metu nustatyti variantai, esantys tirtųjų variantų gretimose genomo srityse bei NKS metodu patvirtintų ir nepatvirtintų variantų skaičius

Genas	**Pokyčio vieta genome**	**dbSNP ID**	**NKS metodu įvertintų tiriamųjų skaičius**	**Patvirtintų variantų skaičius**	**Nepatvirtintų variantų skaičius**	**Nepatvirtinti dėl netinkamų kokybės rodiklių**
EDDM3A	14:20747838	rs11847654	3	3	0	-
GAB4	22:16988136	rs5992604	9	7	2	2
GAB4	22:16988014	nėra	1	1	0	-
GAB4	22:16988028	rs144427521	2	2	0	-
C1ORF167	1:11768262	rs4845881	6	1	5	5
XPC	3:14158408	rs2227999	3	3	0	-
ZNF469	16:88439326	rs4782362	11	11	0	-
ZNF469	16:88439227	rs4782301	8	5	3	3
KRT31	17:41397581	rs79496913	2	1	1	-
SPATA13	13:24223774	rs1220546	6	5	1	-
CCDC107	9:35661091	rs10441686	4	4	0	-
SGK223	8:8376702-8376703	rs373458829	9	7	2	2
SGK223	8:8376561	rs4840952	7	6	1	1
PLEKG4B	5:163151	rs3810867	7	0	7	4
PLEKG4B	5:163090	rs3810870	5	1	4	3
MYT1L	2:1943142	rs3748988	6	6	0	-

Genas	Pokyčio vieta genome	dbSNP ID	NKS metodu įvertintų tiriamųjų skaičius	Patvirtintų variantų skaičius	Nepatvirtintų variantų skaičius	Nepatvirtinti dėl netinkamų kokybės rodiklių
DMKN	19:35513204	rs4806163	9	6	3	3
DMKN	19:35513386	nėra	1	0	1	1
PERM1	1:979496	rs13302983	6	4	2	2
PERM1	1:979560	rs13303033	2	1	1	1
RP1L1	8:10610117	rs112656102	1	1	0	-
RP1L1	8:10610066	rs4840499	1	1	0	-
SKA3	13:21155087	rs17345383	2	2	0	-

Kitas etapas – specifiškumo nustatymas. Specifiškumu laikoma teisingai nenustatytų variantų dalis, kurių ir neturėtų būti toje genomo pozicijoje (angl. *true negative*). Jis apskaičiuojamas teisingai neigiamų variantų skaičių dalinant iš teisingai neigiamų ir klaidingai teigiamų variantų skaičiaus sumos. Teisingai neigiamais variantais buvo laikyti visi, kurie tiek NKS, tiek *Sanger* duomenyse atitiko referentinį alelį. Tuo tarpu klaidingai teigiamais buvo laikyti variantai, kurie NKS metodu buvo nustatyti, tačiau *Sanger* metodu nebuvo patvirtinti. Tokių variantų buvo 34, todėl NKS *SOLiD* metodo specifiškumas siekė 99,7 %. Remiantis aukštais NKS metodo jautrumo ir specifiškumo įverčiais, galima daryti išvadą, jog NKS duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, tikrinti *Sanger* metodu nėra būtina. Panašią išvadą padarė ir *Baudhuin* su kolegomis (45). Svarbu atkreipti dėmesį ir į tai, kad nors *Sanger* metodas ir yra laikomas "auksiniu standartu", tačiau būta nenustatytų variantų būtent dėl *Sanger* metodo ribotumo, o tai kelia klausimą, ar *Sanger* metodas gali būti laikomas etalonu.

Harismendy bei jo kolegų tyrimas atskleidė, jog devyni iš dešimties tirtų variantų, nustatytų *Sanger* metodu ir nenustatytų NKS metodu, buvo klaidingai identifikuoti būtent dėl *Sanger* metodo klaidos (42).

Negana to, kito tyrimo metu taip pat buvo nustatyta, kad 17 iš 19 neatitikimų tarp *Sanger* ir NKS duomenų buvo dėl *Sanger* klaidos. Buvo padaryta išvada, kad nėra prasmės tikrinti NKS duomenyse nustatytų vieno nukleotido variantų *Sanger* metodu, nes tokiu būdu gali būti atmesti ir neįvertinti reikšmingi variantai. Be to, atsisakant NKS gautų duomenų tikrinimo būtų sutaupyta finansine ir laiko prasmėmis (41).

IŠVADOS

1. Naujos kartos sekoskaitos metodu nustatytų 60-ies genomo variantų tvirtinimui buvo panaudota *Sanger* sekoskaita. Iš šių variantų, 37-iems buvo naujai sukurti pradmenys. Visoms 60-iai genominių sričių buvo optimizuotos PGR sąlygos.
2. *Sanger* metodu nustatyti aštuoni variantai, kurie nesutapo su NKS duomenimis. Nustatyta, kad šeši iš šių variantų (esantys *SKA3, CRACR2A, ANKRD62, BMS1, BAGE2* ir *CFP* genuose) buvo klaidingai įvardinti arba negalėjo būti nustatomi dėl *Sanger* metodo ribotumo ar klaidų. Du variantai (*TMPRSS15* ir *FAM105A* genuose; po vieną tiriamąjį) buvo klaidingai įvardinti NKS *SOLiD* platformos automatinio analizės algoritmo. Šie rezultatai rodo, kad *Sanger* metodas turi daugiau trūkumų identifikuojant variantus, nei NKS, todėl neturėtų būti laikomas "auksiniu standartu".
3. Bendras vidutinio padengimo vidurkis naudojant *SureSelect* ir *TargetSeq* praturtinimo sistemas atitinkamai yra 32,77 bei 31,58. Remiantis šiomis reikšmėmis, *SureSelect* praturtinimo sistemos bendras padengimas yra 3,65 % didesnis. Tačiau naudojantis neparametriniu Vilkoksono testu nustatyta, kad skirtingų praturtinimo sistemų padengimo vidurkiai statistiškai reikšmingai nesiskiria.
4. Panaudojant dviejų negimininingų asmenų, kuriems sekoskaita buvo įvykdyta abejomis praturtinimo sistemomis, duomenis buvo tiksliau palygintos abi praturtinimo sistemos. Daugiau variantų buvo nustatyta, naudojant *SureSelect* praturtinimo sistemą, tačiau *TargetSeq* sistema buvo nustatyti šiai sistemai unikalūs variantai, todėl tiksliausiam variantų nustatymui efektyviausia būtų naudoti abi praturtinimo sistemas. Lyginant su *TargetSeq,* bendras padengimo vidurkis yra 35,82 % didesnis naudojant *SureSelect* sistemą, tačiau imtis yra per maža, kad šiuos duomenis galima būtų patvirtinti statistiniais analizės metodais.
5. Darbe naudotos NKS *SOLiD* platformos automatinio analizės algoritmo tikslumas yra 99,66 %, jautrumas lygus 99,22 %, o specifiškumas – 99,7 %. Remiantis šiais rezultatais, galima daryti išvadą, jog NKS duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, tikrinti *Sanger* metodu nėra būtina.

SANTRAUKA

Vienas pagrindinių metodų, leidžiančių nustatyti genomo variantus bei jų įvairovę, yra DNR sekoskaita. Įvairiose gyvybės mokslų srityse vis dažniau naudojama naujos kartos sekoskaita (NKS) kuri leidžia nusekvenuoti visą egzomą ar netgi genomą. Tačiau, genomo variantų nustatymui dažnai vis dar yra taikoma *Sanger* sekoskaita. Šis metodas yra laikomas tikslesniu, o gauti duomenys yra lengviau analizuojami. Dėl šių priežasčių, *Sanger* metodas naudojamas NKS duomenų tvirtinimui. Tačiau šiuo metu mokslo pasaulyje vyksta debatai dėl to, ar verta NKS gautus duomenis ir nustatytus variantus visuomet tikrinti *Sanger* sekoskaita. Taip pat tyrėjai nesutaria dėl skirtingų NKS praturtinimo sistemų naudojimo efektyvumo – skirtingose mokslinėse publikacijose pateikiamos skirtingos praturtinimo sistemų rekomendacijos, o tai atskleidžia šios temos nevienalypiškumą.

Šio darbo tikslas – įvertinti NKS *SOLiD* platformos automatinį analizės algoritmą, apibrėžiant metodo tikslumo, jautrumo ir specifiškumo reikšmes, bei įvertinti šiai platformai naudojamas taikinių praturtinimo sistemas (*TargetSeq* ir *SureSelect)* panaudojant *Sanger* sekoskaitos metodą bei statistinius ir bioinformacinius įrankius.

Darbo metu buvo vertinama 60 genomo variantų, nustatytų ir identifikuotų naujos kartos sekoskaitos metodu, atliekant *Sanger* sekoskaitą. Viso tiriamųjų grupę sudarė 96 tiriamieji. Taip buvo siekta įvertinti *SOLiD* automatinio analizės algoritmo patikimumą. Nustatyta, kad šeši variantai (esantys *SKA3, CRACR2A, ANKRD62, BMS1, BAGE2* ir *CFP* genuose) buvo klaidingai identifikuoti arba negalėjo būti nustatomi dėl *Sanger* metodo ribotumo ar klaidų. Tuo tarpu, du variantai (*TMPRSS15* ir *FAM105A* genuose; po vieną tiriamąjį) buvo klaidingai įvardinti NKS *SOLiD* platformos automatinio analizės algoritmo. Šie rezultatai rodo, kad *Sanger* metodas turi daugiau trūkumų identifikuojant variantus, nei NKS, todėl neturėtų būti laikomas "auksiniu standartu".

Šio darbo metu NKS duomenys apie tiriamus variantus buvo gauti naudojant dvi skirtingas taikinių praturtinimo sistemas - *SureSelect* ir *TargetSeq*. Praturtinimo sistemos buvo lyginamos pagal vieną pagrindinių NKS parametrų – padengimą. Bendras vidutinio padengimo vidurkis naudojant *SureSelect* ir *TargetSeq* praturtinimo sistemas atitinkamai yra 32,77 bei 31,58. Remiantis šiomis reikšmėmis, *SureSelect* praturtinimo sistemos bendras padengimas yra 3,65 % didesnis. Tačiau naudojantis neparametriniu Vilkoksono testu nustatyta, kad skirtingų praturtinimo sistemų padengimo vidurkiai statistiškai reikšmingai nesiskiria. Vėliau, panaudojant dviejų negimininingų asmenų, kuriems sekoskaita buvo įvykdyta abejomis praturtinimo sistemomis, duomenis, daugiau variantų buvo nustatyta naudojant *SureSelect* praturtinimo sistemą, tačiau *TargetSeq* sistema buvo nustatyti šiai sistemai unikalūs variantai, todėl tiksliausiam variantų nustatymui efektyviausia būtų naudoti abi

praturtinimo sistemas. Lyginant su *TargetSeq*, bendras padengimo vidurkis yra 35,82 % didesnis naudojant *SureSelect* sistemą, tačiau imtis yra per maža, kad šiuos duomenis galima būtų patvirtinti statistiniais analizės metodais.

Galiausiai buvo apskaičiuotas darbe naudotos NKS *SOLiD* platformos automatinio analizės algoritmo tikslumas, jautrumas ir specifiškumas, kuris atitinkamai yra 99,66 %, 99,22 %, 99,7 %. Remiantis šiais rezultatais, galima daryti išvadą, jog NKS duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, tikrinti *Sanger* metodu nėra būtina.

Evaluation of Automated Next Generation Sequencing Data Analysis Pipeline and Different Enrichment Systems Using Sanger Sequencing

SUMMARY

DNA sequencing is one of the main methods for determining the genome variants. Next generation sequencing (NGS) is increasingly used in various life science fields due to its high efficiency, the ability to sequence exome or even whole genome. However, Sanger sequencing is still used to determine genome variants. This method is considered more accurate and the data is easier to analyze. For these reasons, Sanger sequencing method is used for NGS data validation.

Nowadays, a debate in the scientific world is held on whether NGS findings should be verified using Sanger sequencing or not. Also, researchers do not agree on the efficiency of different NGS enrichment systems – different publications present diverse point of view to enrichment systems, which reveals ambiguity of the aforementioned topic.

The aim of this work was to evaluate automated NGS SOLiD platform's data analysis pipeline, defining the accuracy, sensitivity and specificity values of the method and to evaluate used target enrichment systems (TargetSeq and SureSelect) using Sanger sequencing, statistical and bioinformatic tools.

Sixty genome variants detected using NGS were investigated using Sanger sequencing and other analysis tools. Total test group consisted of 96 subjects. The purpose of this was to evaluate reliability of the SOLiD data analysis pipeline. It was detected that six variants (in *SKA3*, *CRACR2A*, *ANKRD62*, *BMS1*, *BAGE2* and *CFP* genes) have been identified incorrectly or could not be identified at all because of Sanger sequencing limitations or errors. Meanwhile, the two variants (in *TMPRSS15* and *FAM105A* genes) were erroneously identified by automated data analysis pipeline of NGS SOLiD platform. These results indicate that Sanger method has more flaws than NGS, when identifying variants and should not be considered as the "gold standard".

NGS data was obtained using two different target enrichment systems - SureSelect and TargetSeq. They were compared by one of the main parameters of NGS – coverage. The mean coverage using SureSelect and TargetSeq enrichment systems were 32.77 and 31.58, respectively. Based on these values, SureSelect enrichment system's coverage is 3.65 % higher. However, the nonparametric Wilcoxon test showed that difference between means is not statistically significant.

Using data of two unrelated persons, NGS was performed using both enrichment systems. More genome variants were identified using the SureSelect enrichment system. However, TargetSeq system was able to identify unique variants. Consequently, the most effective way to accurately identify

genome variants is to use both enrichment systems. The mean coverage was 35.82 % higher using SureSelect system compared to TargetSeq, although the sample size is too small to confirm the data by statistical analysis.

Finally, the automated NGS SOLiD data analysis pipeline's accuracy, sensitivity and specificity was estimated to be 99.66 %, 99.22 % and 99.7 %, respectively. Based on these results, it can be concluded that there is no need to verify NGS data using Sanger sequencing when automatic analysis algorithm parameters are high.

LITERATŪROS SĄRAŠAS

1. Kučinskas V. Genetikos ir genomikos pagrindai. Vilniaus universiteto leidykla. 2012.
2. Chin E, Silva C, Hegde M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genet. 2013; 14:6.
3. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology. 2007; 8(7):R143.
4. Mu W, Lu H, Chen J, Li S, Elliott AM. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. The Journal of Molecular Diagnostics. 2016; 18(6):923–932.
5. Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, Mulkin JC, Biesecker LG. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. Am J Hum Genet. 2012; 91(1):97–108.
6. McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, Waugh DJ, Hamilton P, O'Sullivan JM, Salto-Tellez M. Validation of Next Generation Sequencing Technologies in Comparison to Current Diagnostic Gold Standards for BRAF, EGFR and KRAS Mutational Analysis. PLOS ONE. 2013; 8(7): e69604.
7. Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics. 2014; 30(19): 2709-2716.
8. Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Brüggenwirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CA, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N. Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. Human Mutation. 2013; 34:1313–1321.
9. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, Mensenkamp AR, Rodenburg RJT, Yntema HG, Spruijt L, Vermeer S, Rinne T, Gassen KL, Bodmer D, Lugtenberg D, Reuver R, Buijsman W. A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases. Wiley Online Library. 2013.
10. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799-816.

11. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. PLOS ONE. 2011; 6(11): e28240.
12. Pickrell WO, Rees MI, Chung SK. Next Generation Sequencing Methodologies - An Overview. Adv Protein Chem Struct Biol. 2012; 89:1-26.
13. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011; 11:759-69.
14. Foehlich T, et al. High-throughput nucleic acid analysis. U.S. Patent, 2010.
15. Thayer A. M. Next-Gen Sequencing Is A Numbers Game. Chemical and Engineering News. 2014; 33:11-15.
16. Liu L, Li Y, Li S, Hu N, He Y, Pong R. Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology. 2012; 2012:251364.
17. Datta S, Nettleton D. Statistical Analysis of Next Generation Sequencing Data. Springer International Publishing, Switzerland. 2014.
18. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010; 19(2):227-240.
19. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, Bakker P, Purcell SM, Sunyaev SR. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012; 44:623–30.
20. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009; 106(45):19096-101.
21. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 2007; 4: 903-905.
22. Shigemizu D, Momozawa Y, Abe T, Morizono T, Boroevich KA, Takata S, Ashikawa K, Kubo M, Tsunodaa T. Performance comparison of four commercial human whole-exome capture platforms. Sci Rep. 2015; 5:12742.
23. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. Performance comparison of exome DNA sequencing technologies. Nature biotechnology.

2011; 29(10):908-914.

24. Zhang G, Wang J, Yang J, Li W, Deng Y, Li J, Huang J, Hu S, Zhang B. Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. BMC Genomics. 2015; 16(1):581.
25. Park JY, Clark P, Londin E, Sponziello M, Kricka LJ, Fortina P. Clinical Exome Performance for Reporting Secondary Genetic Findings. Clinical chemistry. 2015;61(1):213-220.
26. TargetSeq™ Exome and Custom Enrichment for SOLiD® Next-Generation Sequencing. ThermoFisher Scientific [internetinė svetainė]. [Cituota 2017 03 15]. Adresas: http://bit.ly/2ooeziK
27. SureSelect Human All Exon V6 - Details & Specifications. AgilentTechnologies [internetinė svetainė]. [Cituota 2017 03 15]. Adresas: http://bit.ly/2p4UDnU
28. Londin ER, Clark P, Sponziello M, Kricka LJ, Fortina P, Park JY. Performance of exome sequencing for pharmacogenomics. Personalized medicine. 2014; 12(2):109-115.
29. Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, Wu M, Liu X, Tian G, Wang J, Wang J, Yang H, Zhang X. Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biology. 2011;12(9):95.
30. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. BioMed Research International. 2014; 2014:16.
31. Rančelis T. Patogeninių genomo variantų ir jų genų, lemiančių autosomines recesyviąsias ligas, įvairovės analizė, panaudojant viso egzomo sekoskaitą. Daktaro disertacija, VU. 2016.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078-2079.
33. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, Duce J, Kauwe JSK, Ridge PG. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016; 17(7):239.
34. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012; 14:72.
35. Torri F, Dinov ID, Zamanyan A, Hobel S, Genco A, Petrosyan P, Clark AP, Liu Z, Eggert P, Pierce J, Knowles JA, Ames J, Kesselman C, Toga AW, Potkin SG, Vawter MP, Macciardi F. Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. Genes. 2012; 3(3):545-575.

36. The Variant Call Format (VCF) Version 4.2 Specification. SAMtools Github. [internetinė svetainė]. [Atnaujinta: 2016 11 15; cituota: 2017 04 03]. Adresas: http://bit.ly/2oohRTa
37. Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010; 11(8):86.
38. Pandey RV, Pabinger S, Kriegner A, Weinhäusel A. MutAid: Sanger and NGS Based Integrated Pipeline for Mutation Identification, Validation and Annotation in Human Molecular Genetics. PLoS ONE. 2016; 11(2):e0147697.
39. Razali N, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics. 2011; 2(1):21–33.
40. Chin ELH, Silva C, Hegde M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genetics. 2013; 14:6.
41. Beck TF, Mullikin JC, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clinical Chemistry. 2016; 62(4):647-654.
42. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology. 2009; 10:32.
43. Houniet DT, Rahman TJ, Al Turki S, Hurles ME, Xu Y, Goodship J, Keavney B, Santibanez Koref M. Using population data for assessing next-generation sequencing performance. Bioinformatics. 2015; 31(1):56–61.
44. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med. 2011; 3: 65.
45. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing. The Journal of Molecular Diagnostics. 2015; 17(4):456 – 461.

PADĖKA

Nuoširdžiai dėkoju darbo vadovei dr. Laimai Ambrozaitytei už įgytas teorines ir praktines žinias, pastatytą mokslinės praktikos pamatą, vertingus patarimus bei pastebėjimus.

Taip pat dėkoju darbo konsultantui dr. Tautvydui Rančeliui, kurio dėka pagilinau bioinformacines žinias bei patobulinau mokslinio darbo strategijos kūrimo įgūdžius. Dėkoju už išsamias konsultacijas, vertingus patarimus ir skirtą laiką.

Galiausiai, esu ypač dėkinga akad. prof. habil. dr. Vaidučiui Kučinskui už galimybę atlikti šį darbą, bei visiems VU MF ŽMGK darbuotojams, padėjusiems darbą įgyvendinti.