

Laboratorinė medicina.
2018, t. 20, Nr. 2(78), p. 123–129.

Naujos kartos sekoskaitos duomenų automatinio analizės algoritmo bei skirtingų praturtinimo sistemų įvertinimas naudojant *Sanger* sekoskaitą

Gabrielė Žukauskaitė
Laima Ambrozaitytė
Tautvydas Rančelis
Ingrida Kavaliauskienė
Vaidutis Kučinskas

Santrauka

Darbo tikslas. Įvertinti naujos kartos sekoskaitos *SOLiD* platformos automatinį analizės algoritmą, apibrėžiant metodo tikslumo, jautrumo ir specifiškumo reikšmes, bei įvertinti šiai platformai naudojamas taikinių praturtinimo sistemas (*TargetSeq* ir *SureSelect*) *Sanger* sekoskaitos metodu.

Tiriamieji ir metodai. Tiriamųjų grupę sudarė 96 asmenys. Buvo vertinama 60 genomo variantų, identifikuotų naujos kartos sekoskaitos metodu – atliekant *Sanger* sekoskaitą. Taikinių praturtinimo sistemoms palyginti naudotos naujos kartos sekoskaitos kokybės duomenyse nurodytos padengimų reikšmės.

Tyrimo rezultatai. Šeši variantai buvo klaidingai identifikuoti arba negalėjo būti nustatomi dėl *Sanger* sekoskaitos metodo ribotumo ar klaidų. Du variantai buvo klaidingai įvardyti naujos kartos sekoskaitos *SOLiD* platformos automatinio analizės algoritmo.

Bendras vidutinio padengimo vidurkis naudojant *SureSelect* ir *TargetSeq* praturtinimo sistemas yra atitinkamai 32,77 ir 31,58. Atsižvelgiant į šias reikšmes, *SureSelect* praturtinimo sistemos bendras padengimas yra 3,65 % didesnis. Daugiau variantų buvo nustatyta naudojant *SureSelect* praturtinimo sistemą, tačiau *TargetSeq* sistema buvo nustatyti šiai sistemai unikalūs variantai.

Taip pat buvo apskaičiuotas automatinio analizės algoritmo tikslumas, jautrumas ir specifiškumas – atitinkamai 99,66 %, 99,22 %, 99,7 %.

Išvados. Naujos kartos sekoskaitos *SOLiD* platformos automatinio analizės algoritmo tikslumo, jautrumo ir specifiškumo reikšmės leidžia teigti, kad naujos kartos sekoskaitos duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, nebūtina tikrinti *Sanger* metodu.

Lyginant *SureSelect* ir *TargetSeq* padengimo reikšmes naujos kartos sekoskaitos duomenyse paaiškėjo, kad skirtingų praturtinimo sistemų padengimo vidurkiai statistškai reikšmingai nesiskiria. Tačiau kiekviena sistema buvo nustatyti tai sistemai unikalūs variantai, todėl efektyviausia naudoti abi praturtinimo sistemas, kad variantai būtų nustatyti tiksliausiai.

Reikšminiai žodžiai: naujos kartos sekoskaita, NKS *SOLiD* platforma, praturtinimo sistemos, jautrumas, specifiškumas.

Vilniaus universiteto Medicinos fakulteto Biomedicinos mokslų instituto Žmogaus ir medicininės genetikos katedra

Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University

El. paštas: gabriele.zukauskaite@mf.vu.lt

IVADAS

Vienas iš 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 sekvenuoti koduojančią genomo dalį (egzomą) ar net visą genomą. Tačiau genomo variantams nustatyti dažnai vis dar yra taikoma *Sanger* sekoskaita. Be to, šis metodas iki šiol naudojamas NKS būdu nustatytiems genomo variantams tvirtinti dėl būtinybės įvertinti gautų rezultatų patikimumą bei paties metodo jautrumą ir specifiskumą. Šią būtinybę lemia NKS klaidos, dėl kurių sunku nustatyti variantus. Pavyzdžiui, mažą padengimą turintys genomo regionai gali lemti, kad kritinėse genomo srityse bus praleisti reikšmingi variantai [1].

Įtakos sekoskaitos rezultatams turi ir beveik kiekviename NKS etape daromos klaidos, pradedant sekoskaitos technologijos pasirinkimu, automatinio analizės algoritmo klaidomis, baigiant *in silico* analizės bei duomenų interpretavimo klaidomis. Labai svarbus žingsnis yra ir praturtinimo sistemos pasirinkimas, nes skirtingomis praturtinimo sistemomis nustatomi skirtingi variantai skirtinguose genomo regionuose.

Mokslo pasaulyje vyksta diskusijos, ar verta NKS gautus duomenis ir nustatytus variantus visuomet tikrinti *Sanger* sekoskaita. Naujausių tyrimų autoriai atsisako tikrinti NKS duomenis *Sanger* metodu dėl gaunamų aukštų kokybės parametrų (jautrumo, specifiskumo) reikšmių. Pavyzdžiui, W. Mu su kolegomis, atlikę NKS 20-iai tūkstančių paveldimo vėžio genų rinkinių ir *Sanger* metodu patikrinę 7 845 nepolimorfinius variantus, nu-

statė, kad 98,7 % rezultatų sutapo su NKS duomenimis. Kita vertus, 1,3 % variantų buvo klaidingai teigiami. Šie variantai nustatyti probleminiuose genomo regionuose – AT ar GC gausiose srityse, homologiiniuose, pseudogenų regionuose, ir galėjo būti atmesti kaip klaidingai teigiami tik juos patikrinus *Sanger* metodu [2]. Remiantis šiuo pavyzdžiu teigtina, kad nors NKS nustatytų variantų tvirtinimas *Sanger* metodu turi akivaizdžią naudą, tačiau dalis laboratorijų visgi vengia variantų tikrinimo. Taip yra dėl to, kad pailgėja tyrimo laikas ir padidėja jo kaina pakartotinai tikrinant duomenis *Sanger* metodu.

Šiame darbe siekiama įvertinti NKS duomenų automatinę analizės algoritmą, apibrėžiant metodo tikslumą, jautrumą ir specifiskumą reikšmes bei diskutuojant, ar verta *Sanger* metodu tikrinti NKS duomenis remiantis šio tyrimo rezultatais. Taip pat lyginamos dvi skirtingos praturtinimo sistemos siekiant įvertinti, kuri iš jų yra pranašesnė.

DARBO TIKSLAS

Įvertinti naujos kartos sekoskaitos *SOLiD* platformos automatinę analizės algoritmą bei šiai platformai pritaikytas taikinių praturtinimo sistemos naudojant *Sanger* sekoskaitos metodą bei statistinius ir bioinformacinius įrankius.

TYRIMO MEDŽIAGA IR METODAI

Darbe buvo panaudoti Vilniaus universiteto Medicinos fakulteto Biomedicinos mokslų instituto Žmogaus ir medicininės genetikos katedroje (VU MF BMI ŽMGK) vykdyto projekto „Lietuvos populiacijos genetinė įvairo-

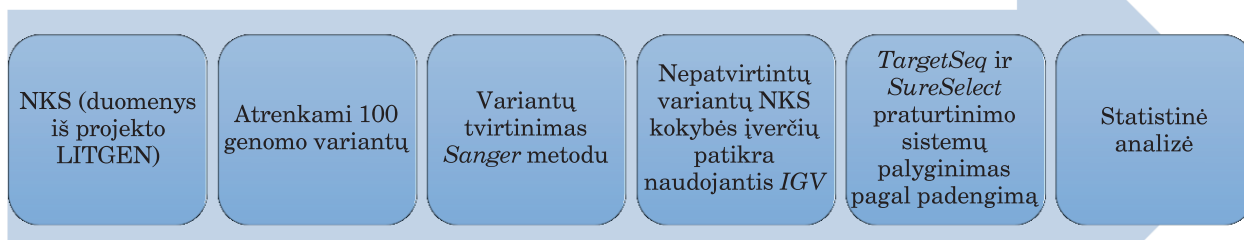
vė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis“ (LITGEN) viso egzomo tiriamųjų duomenys (projekto vadovas akad. prof. habil. dr. Vaidutis Kučinskas). Tiriamųjų grupę sudarė 96 asmenys. Apibendrinta tyrimo strategijos schema pateikiama 1 paveiksle.

Iš projekto duomenų buvo atrinkta 100 genomo variantų pagal šiuos kriterijus:

1. Nustatymas praturtinimo rinkiniais – *TargetSeq* (*Life technologies*, JAV), *SureSelect* (*Agilent*, JAV).
2. Vieno nukleotido polimorfizmo ir trumpų tarpų ir iškritų santykis – 80 % ir 20 %.
3. Dažnis tarp tiriamųjų – variantas galėjo būti nustatomas tik vienam tiriamajam.
4. Pagal lytį – nustatyti variantai lygiai pasiskirstę tarp lyčių.
5. Heterozigotinių ir homozigotinių pakaitų santykis – lygiomis dalimis.
6. Genomo variantų pasiskirstymas genome – tolygus.
7. Variantų parinkimas atsižvelgiant į geno sandarą – 85 % egzoniiniai, 10 % NTR, 5 % introniniai.
8. Pakaitos pobūdis – įtraukiamos ir sinoniminės, ir nesinoniminės pakaitos.

Darbe buvo optimizuotos polimerazės grandininės reakcijos (PGR) sąlygos 60-iai genų egzomų ar jų fragmentų. Jiems visiems buvo atlikta *Sanger* sekoskaita siekiant įsitikinti, kad NKS būdu nustatyti variantai yra nustatomi ir *Sanger* metodu.

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



1 pav. Tyrimo strategijos schema
Fig. 1. The scheme of research strategy

IGV vaizduoklėje buvo tikrinami pagrindiniai BAM faile, gautame NKS SOLiD platforma, esantys parametrai:

- padengimas (angl. *coverage*) >6;
- nukleotido kokybės įvertis (angl. *base quality score*) >20;
- DNR fragmento kokybės įvertis (angl. *mapping quality score*) >30;
- CIGAR įvertis;
- DNR fragmentų, turinčių alternatyvų alelių, skaičius >2.

Siekiant palyginti skirtingas praturtinimo sistemas, variantai buvo suskirstyti į dvi grupes pagal praturtinimo sistemą, jungiant skirtingų tiriamųjų genomo variantus, ir tam buvo naudota *GATK Combine Variants* (3.7 versija, Broad institutas, JAV) programa. Variantai buvo anotuojami *ANNOVAR* programiniu įrankiu [3].

Statistiniams skaičiavimams ir rezultatų vaizdinimui naudota *Microsoft Office Excel 2015* programa ir laisvos prieigos *R Commander 3.2.4* programa [4]. Statistinis patikimumas buvo įvertintas naudojant Mano, Vitnio ir Vilkoksono (*Mann-Whitney-Wilcoxon*) rangų sumų kriterijų. Statistiškai reikšmingais buvo laikomi skirtumai, kurių reikšmingumo lygmuo $p < 0,05$. Statistinės analizės schema pateikiama 2 paveiksle.

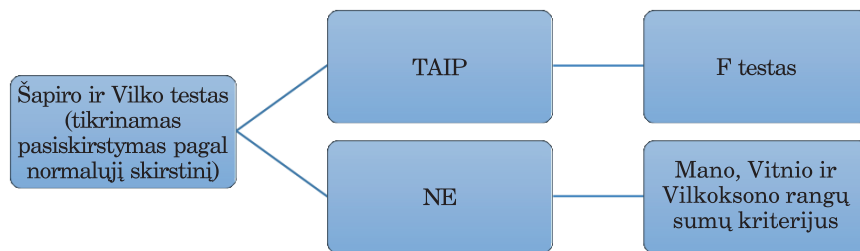
TYRIMO REZULTATAI

Aprašomoji variantų statistika

Pasitelkiant *ANNOVAR* programinį įrankį, visi tirti variantai buvo anotuoti – prie turimų duomenų buvo prijungta duomenų bazių informacija apie kiekvieną tiriamojo variantą.

Variantai darbui buvo pasirinkti atsitiktinai, tačiau taip, kad būtų pasiskirstę tolygiai visame genome – visoje chromosomose buvo tirta bent po vieną variantą. Daugiausia variantų buvo tirta 13-oje ir 14-oje chromosomose (po keturis variantus), mažiau – 16-oje ir Y chromosomose (po vieną variantą).

Didžioji dalis (85 %) tirtų genomo variantų buvo egzoniiniai, nes buvo naudojami viso egzomo sekoskaitos duomenys. Kadangi praturtinimo sistemos aprėpia ne tik egzonus, bet ir aplinkines sritis, todėl atliekant egzomo sekoskaitą buvo identifikuoti variantai, esantys ir intronuose ar ne-transliuojamose geno dalyse. Dėl šios priežasties buvo įtraukti ir tokie variantai, kurie buvo nustatyti viso egzomo sekoskaitos metodu. Tokių variantų dalis siekė 15 %.



2 pav. Statistinės analizės schema

Fig. 2. The scheme of statistical analysis

Sanger sekoskaitos duomenų analizė ir vaizdinimas

Darbe buvo optimizuotos PGR sąlygos 60-iai genų egzono ar jų fragmentų. Jiems visiems buvo atlikta *Sanger* sekoskaita, siekiant įvertinti naudoto automatinio analizės algoritmo patikimumą. Iš viso buvo sekvenuoti 747 genomo fragmentai iš 890, o 143 genomo fragmentai nesekvenuoti dėl ribotų DNR išteklių šiam moksliniam tyrimui.

Visų šių variantų DNR fragmentai iš NKS duomenų buvo patikrinti juos vaizduojančiu integruota genomo vaizduokle (IGV). Šioje programoje tikrinami pagrindiniai BAM faile esantys parametrai, gauti pasitelkiant NKS SOLiD platformą: padengimas (angl. *coverage*), nukleotido kokybės įvertis (angl. *base quality score*), DNR fragmento kokybės įvertis (angl. *mapping quality score*), CIGAR įvertis, alternatyvų alelių turinčių DNR fragmentų skaičius.

Padengimas rodo, kiek DNR fragmentų nukleotidų sutampa ir (ar) nesutampa tam tikroje genomo pozicijoje prilygiavus juos prie referentinio genomo. Kuo daugiau kartų tas pats nukleotidas pasikartoja, tuo rezultatai yra geresni, išskyrus tuos atvejus, kai didelis nukleotido pasikartojimų skaičius sukuriamas dirbtinai, pavyzdžiui, dėl DNR fragmentų duplikacijų [5]. BAM faile taip pat pateikiama informacija apie kiekvieno nukleotido kokybės įverčius (angl. *base quality score*). Šie įverčiai nurodo nukleotido įvardijimo tikslumą arba tikimybę, kad atliekant NKS buvo padaryta klaida. Nukleotido kokybės įvertis pateikiamas *Phred* įverčio formatu. Jei *Phred* įvertis lygus 10 – tai reiškia, kad yra 1 tikimybė iš 10, jog nukleotidas identifikuotas klaidingai. 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 [1].

Taip pat šiame faile nurodomas CIGAR įvertis, kuris 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 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, nustatomas prilygiavimo prie referentinio genomo etape, yra DNR fragmento kokybės įvertis (angl. *mapping quality score*). Šis parametras labai priklauso nuo sekų homologijos. Sekvenuoti DNR fragmentai yra trumpi, todėl jie gali atsikartoti skirtingose genomo vietose, o tai gali lemti, kad DNR fragmentas bus neteisingai prilygiuotas prie referentinio genomo. Todėl DNR fragmentui, kuris gali būti prilygiuotas keliose genomo pozicijose, programa suteikia žemesnį DNR fragmento kokybės įvertį.

Šiame tyrime etape nustatyta, kad šeši variantai (esantys *SKA3* rs11147976, *CRACR2A* rs11062745, *ANKRD62* rs201537483, *BMS1* rs12764004, *BAGE2* rs4913758 ir *CFP* rs2235182 genuose) buvo klaidingai identifikuoti arba negalėjo būti nustatomi dėl *Sanger* metodo ribotumo ar klaidų.

Du variantai (*TMPRSS15* rs2824804 ir *FAM105A* rs16903574 genuose; po vieną tiriamąjį) buvo klaidingai įvardyti NKS SOLiD platformos automatinio analizės algoritmo.

Likusių patvirtintų variantų skaičius buvo lygus 100 %, t. y. visiems tiriamiesiems pokytis buvo patvirtintas ir *Sanger* metodu.

Praturtinimo sistemų palyginimas

Siekiant palyginti praturtinimo sistemas, naudojant *GATK Combine Vari-*

ants programą buvo sujungti skirtingų tiriamųjų genomo variantai. Vėliau NKS duomenys buvo suskirstyti į dvi grupes – variantai, kurie buvo nustatyti *TargetSeq* sistema, ir variantai, kurie buvo nustatyti *SureSelect* sistema. *TargetSeq* sistema buvo nustatyta 17-a variantų, *SureSelect* – 22, o 21-as variantas buvo nustatytas abiejomis praturtinimo sistemomis.

Praturtinimo sistemos buvo lyginamos pagal vieną pagrindinių NKS parametrų – padengimą. Naudojantis NKS kokybės duomenimis buvo apskaičiuoti kiekvieno varianto padengimo vidurkiai ir palyginti abiem sistemomis nustatyti bendri variantai.

Remiantis gautais bendrais vidurkiais (1 lentelė) galima teigti, kad *SureSelect* praturtinimo sistemos bendras padengimas yra šiek tiek (3,65 %) didesnis.

Siekiant patikrinti, ar šis skirtumas yra statistiškai reikšmingas, buvo taikytas Mano, Vitnio ir Vilko (Shapiro–Wilk) testu bei tolesniems statistiniams vertinimams, nes tokiu atveju testas netenka galios atmesti nulinę hipotezę ir nedidelės imtys dirbtinai patenkina normalųjį dėsni.

Iš visų tiriamųjų dviem LITGEN projekto tyrėjai atliko viso egzomo sekoskaitą abiem praturtinimo sistemomis. Gauti duomenys leido tiksliau palyginti šias sistemas. Iš analizuojamų variantų daugiau buvo nustatyta *SureSelect* praturtinimo rinkiniu (šeši variantai vienam tiriamajam, aštuoni kitam). Svarbu pabrėžti, kad nors *TargetSeq* sistema buvo identifiukuota mažiau variantų (keturi pirmam tiriamajam ir trys antram tiriamajam), tačiau buvo nustatyta tokių, kurie nebuvo identifiukuoti *SureSelect* sistema.

Remiantis šiais duomenimis buvo atliktas ir abiem sistemomis nustatytų bendrų variantų padengimo vidurkių palyginimas. Bendras padengimo vidurkis su *TargetSeq* ir *SureSelect* – atitinkamai 21,5 ir 33,5. *SureSelect* vidutinis padengimas daug didesnis (35,82 %). Tačiau svar-

1 lentelė. Bendrų variantų vidutinio padengimo reikšmių palyginimas
Table 1. Comparison of mean coverage values between common variants

Genas	Pokytis pagal HGVS	SU DP vidurkis	TS DP vidurkis
SKA3	c.1157A>G	27,33	47,67
HNRNPC	c.429G>A	33,25	39,60
ZNF721	c.389G>A	65,00	51,00
FAM105A	c.957C>G	50,09	19,36
DSP	c.5213G>A	102,31	34,71
TMEM184A	c.53C>T	6,55	7,00
SMARCA2	c.280C>G	38,09	49,50
NLRP6	c.2229G>A	9,70	10,78
FGF23	c.716C>T	13,93	12,50
SPATA13	c.719G>A	51,38	4,30
EDDM3A	c.184G>T	46,43	49,00
CYFIP1	c.1165G>A	42,70	13,38
ABR	c.1304A>G	35,15	15,17
POLRMT	c.3243C>G	12,60	32,67
SIRPB1	c.316A>C	75,89	162,33
TMPRSS15	c.95A>G	33,58	36,00
GAB4	c.487G>T	9,44	17,33
ARSD	c.671C>A	11,59	18,08
EMILIN2	c.29T>C	11,75	15,57
GNG7	c.-30A>G	4,00	9,32
UTY	c.1029+33C>A	7,50	17,89
Bendras vidurkis		32,774	31,579

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

bu paminėti, kad tokių bendrų variantų buvo tik trys, todėl imties nepakanka normalumui įvertinti Šapiro ir Vilko (*Shapiro–Wilk*) testu bei tolesniems statistiniams vertinimams, nes tokiu atveju testas netenka galios atmesti nulinę hipotezę ir nedidelės imtys dirbtinai patenkina normalųjį dėsni.

SOLiD metodo tikslumo, jautrumo ir specifiškumo nustatymas

Atlikus NKS metodu nustatytų variantų tvirtinimą *Sanger* sekoskaita, buvo apskaičiuotas darbe vykdyto NKS SOLiD metodo automatinio analizės algoritmo tikslumas, jautrumas ir specifiškumas. Rezultatai pateikiami 2 lentelėje.

REZULTATŲ APTARIMAS

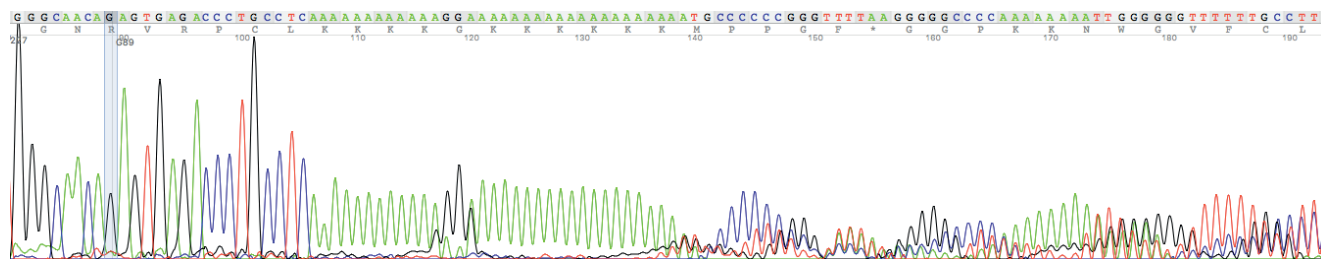
Sanger sekoskaitos duomenų palyginimas su NKS duomenimis

Šiame tyrimo etape patikrinus variantą SKA3 (rs11147976) gene IGV vaizdukle nustatyta, kad DNR fragmentų kokybės įverčiai yra aukšti. Didžiosios dalies DNR fragmentų CIGAR įvertis 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 metodu buvo tiksliai identifiukuotas esamas

2 lentelė. NKS SOLiD metodo automatinio analizės algoritmo tikslumo, jautrumo ir specifiškumo apibrėžimai bei gauti įverčiai

Table 2. Definitions and values of automated NGS SOLiD data analysis pipeline's accuracy, sensitivity and specificity

Savoka	Apibrėžimas	Darbe nustatytas SOLiD platformos įvertis
Tikslumas	Matavimo rezultato ir priimtos pamatinės vertės atitikimo artumas	99,66 %
Jautrumas	Gebėjimas tiksliai nustatyti genomo variantus, kurie iš tikrųjų yra	99,22 %
Specifiškumas	Teisingai nenustatytų variantų, kurių ir neturėtų būti toje genomo pozicijoje, dalis	99,7 %



3 pav. *CFP* geno egzono fragmento sekvenograma
Fig. 3. The chromatogram of *CFP* gene exon

variantas. 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 matomos dvigubos smailės. Patikrinus šį variantą *IGV* pastebėta, kad nukleotidų kokybės įverčiai yra aukšti (daugumai fragmentų lygūs 41), vidutinis padengimas lygus 44, o tai turėtų sufleruoti apie gerą atskiros nukleotido kokybę ir šių parametrų vertės gali būti įrodytas, kad variantas tiriamiesiems yra nustatomas. Tačiau didelis padengimas ir prastas DNR fragmento kokybės įvertis (apie 8) rodo, kad DNR fragmentas yra homologiškas keliuose genomo vietose, todėl sukurti pradmenys pagausina DNR fragmentus iš skirtingų genomo vietų. Dėl tos priežasties *Sanger* sekoskaitos metodu pokytis tироje genomo srityje nefiksuojaamas, o NKS metodu jis fiksuojamas. Be to, dėl homologiinių sričių genome naujos kartos sekoskaitos metodu gauti fragmentai gali būti netiksliai prilygiuojami prie referentinio genomo. Panaši situacija buvo nustatyta ir vizualizavus variantą *BAGE2* (rs4913758) gene. Visi šie gauti rezultatai leidžia daryti išvadą, kad atliekant NKS duomenų analizę būtina atsižvelgti į sekoskaitos ir duomenų analizės kokybės įverčius.

Analizuojant *Sanger* sekoskaitos duomenis, buvo susidurta ir su keblesniais variantais. 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-oje chromosomoje, todėl visiems tiriamiesiems nuo egzono vidurio iki galo buvo matomos dvigubos smailės, kurios iš pirmo žvilgsnio buvo panašios į pokytį, sukeliantį rėmelio poslinkį. Daliai tiriamųjų dėl minėtų dvigubų, o kartais ir trigubų smailių buvo sudėtinga nustatyti tikrąjį varianto tipą.

Kitas variantas (rs2235182), sukėlęs analizės sunkumų, buvo *CFP* gene. Sekvenavus *Sanger* metodu šio geno egzono ir šalia esančio introno dalį pa-

stebėta, kad sekos kokybė suprastėja po poliA regiono, matomos dvigubos smailės (3 pav.). Šiuo atveju tai homogeniškų regionų sekvenavimo *Sanger* metodu trūkumas – DNR sekos prasydimas (angl. *DNA sequence slippages*). Patikrinus *IGV* programa 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.

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*). Patikrinus variantus *IGV* programa, *TMPRSS15* ir *FAM105A* genuose šie variantai nebuvo nustatyti. Tai reiškia, kad analizės algoritmas teisingų variantų atrankos etape klaidingai priskyrė šiuos variantus ir tai yra NKS algoritmo klaida. *CRACR2A* gene esantis variantas taip pat nebuvo nustatytas *Sanger* metodu, jo NKS duomenys buvo vaizdinami *IGV* programa. 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 laikomas aukso vertės standartu [6–8], tačiau gauti rezultatai prieštarauja šiam teiginiui ir parodo, kad metodas turi ir trūkumų. *Sanger* metodas yra tikslus, tačiau jo naudojimą šiuolaikinėje klinikinėje genetikoje riboja didėjantis tiriamųjų genų skaičius tam tikroms daugiaveiksnėms bei poligeninėms ligoms nustatyti, pavyzdžiui, kurtumui ar intelektinei negaliai. *Sanger* metodas yra netinkamas dideliu genų skaičiui sekvenuoti – tyrimo kaina ir laikas smarkiai išauga. Tokiu atveju naudojamos NKS technologijos, kurios yra tinkamos tūkstančiams variantų nustatyti vienu metu. Tačiau skirtingos NKS technologijos, susidurdamos su problemineis genomo regio-

nais (homogeniniai bei gausūs trumpų intarpų ir iškritų regionai), gali turėti skirtingas jautrumo ir specifiškumo reikšmes. Todėl buvo padaryta išvada, kad tradicinis *Sanger* sekoskaitos metodas turi būti naudojamas kaip pagalbinis NKS nustatytiesiems variantams patvirtinti prieš įvardijant galimą funkcinę reikšmę susirgti tam tikromis ligomis. Be to, *Sanger* metodas puikiai tinka monogeninėms ligoms nustatyti, todėl NKS naudoti nebūtina. Taip yra nurodoma ir NKS naudojimo klinikinėje genetikoje gairėse [9]. Taigi, šie metodai iš esmės papildo vienas kitą – NKS metodai leidžia atlikti didelio masto genomo tyrimus, o *Sanger* metodas naudojamas mažesnėms imtims tirti bei NKS gautiems rezultatams tvirtinti.

Praturtinimo sistemų palyginimas

Šiuo metu yra sukurta keletas skirtingų taikinių praturtinimo sistemų bei nemažai skirtingų kiekvienos sistemos naujinių, siekiant užtikrinti kuo efektyvesnį egzomo ar konkrečių genomo sričių praturtinimą. Naujausiose mokslinėse publikacijose minima, jog *SureSelect* sistema leidžia tiksliau nustatyti variantus, ypač analizuojant trumpus interpus bei iškritas. Be to, geresnis koduojančių regionų padengimas pasiekiamas būtent šia sistema [10–12].

Kita praturtinimo sistema – *TargetSeq*. Nors ji 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 [11]. E. R London su kolegomis nagrinėjo farmakogenomikai svarbius variantus ir aiškinois egzomo sekoskaitos naudą nustatant šiuos variantus. Įdomu tai, kad, šių autorių duomenimis, mažiausiai variantų egzonuose buvo nustatyta naudojant *SureSelect* sistemą, o naudojant *TargetSeq* sistemą variantų buvo nustatyta apie 10 % daugiau [13].

Tačiau yra ir tokių mokslinių publikacijų, kuriose lyginami keli skirtingi praturtinimo sistemų gamintojai. Gauti rezultatai rodo, kad visomis lyginamomis technologijomis pasiektas panašus tikslumas nustatant vieno nukleotido polimorfizmus [14]. Taigi, skirtingose mokslinėse publikacijose daromos skirtingos išvados atskleidžia šios temos sudėtingumą ir tolesnių tyrimų aktualumą.

Šiame darbe NKS duomenys apie tiriamus variantus buvo gauti naudojant dvi skirtingas taikinių praturtinimo sistemas – *SureSelect* ir *TargetSeq*. Buvo siekiama nustatyti, kuri sistema yra pranašesnė. Pagal gautus rezultatus, pateiktus 1 lentelėje, mažiausias vidutinis padengimas (DP = 4) buvo nustatytas variantui, esančiam *GNG7* geno 5' netransliuojamame regione. Šiuo atveju naudota *SureSelect* praturtinimo sistema. Tai galima paaiškinti remiantis varianto genomine pozicija – ši praturtinimo sistema naudoja genomo koduojančių dalių analizei, todėl geriausias padengimas bus koduojančiose srityse, o nekoduojančioms genomo dalims padengimas yra daug mažesnis.

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

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

Remiantis gautais bendrais vidurkiais, pateiktais 1 lentelėje, buvo nustatyta, kad bendras padengimo vidurkis tarp skirtingų praturtinimo sistemų statistiškai reikšmingai nesiskiria. Tačiau, atlikus dviejų negiminingų asmenų viso egzomo sekoskaitos abiem praturtinimo sistemomis duomenų analizę, buvo nustatyta mažiau variantų naudojant *TargetSeq* sistemą (keturi pirmam tiriamajam ir trys antram tiriamajam). Tačiau buvo rasta ir variantų, kurie nebuvo identifikuoti *SureSelect* sistema. Taigi galima daryti prielaidą, kad tiksliausiai variantus būtų galima nustatyti naudojant abi praturtinimo sistemas. Panašią išvadą padarė ir G. Zhang su kolegomis: vykdant egzomo sekoskaitą ir naudojant abi praturtinimo sistemas buvo pasiektas didesnis specifiskumas bei jautrumas NKS analizės algoritmo variantų nustatymo etape [11].

SOLiD metodo tikslumo, jautrumo ir specifiskumo nustatymas

Siekiant nustatyti, ar tyrimo metodo rezultatai yra tikslūs, laboratorijose kiekvienas tyrimo etapas yra vertinamas ir atidžiai stebimas. Svarbu įvertinti kiekvieno metodo tikslumą, jautrumą bei specifiskumą [1]. Puikus pavyzdys yra paveldimų ligų NKS genų rinkinių tyrimai: nustatyti metodo specifiskumą ir jautrumą yra itin svarbu, nes pagal tyrimo atsakymą formuluojama diagnozė ir parenkama tolesnė gydymo taktika [2]. Atlikus NKS metodu nustatytų variantų tvirtinimą *Sanger* sekoskaita, buvo apskaičiuotas darbe vykdyto NKS *SOLiD* platformos automatinio analizės algoritmo tikslumas, jautrumas ir specifiskumas.

Metodo tikslumas siekė 99,66 %, t. y. tokia dalis NKS duomenų sutapo su *Sanger* sekoskaitos metodu gautais duomenimis. Šis skaičius buvo artimas L. G. Biesecker ir kolegų gautiems duomenims – jų atliktame tyrime metodo tikslumas siekė 99,97 % [15]. O. Harismendy ir kolegų duomenimis, *SOLiD* metodo tikslumo reikšmė taip pat buvo labai artima šiame darbe gautajai reikšmei – 99,7 % [16].

Metodo jautrumas – tai gebėjimas tiksliai nustatyti genomo variantus, kurie iš tikrųjų yra. Šie variantai taip pat vadinami teisingai teigiamais (angl. *true positive*). Taigi, jautrumas apskaičiuojamas dalijant teisingai teigiamų variantų skaičių iš teisingai teigiamų ir klaidingai neigiamų variantų skaičiaus sumos [17]. Metodo jautrumas siekė 99,22 %.

Specifiskumu laikoma dalis teisingai nenustatytų variantų, kurių ir neturėtų būti toje genomo pozicijoje (angl. *true negative*). Jis apskaičiuojamas teisingai neigiamų variantų skaičių dalijant iš teisingai neigiamų ir klaidingai teigiamų variantų skaičiaus sumos. NKS *SOLiD* metodo specifiskumas siekė 99,7 %. Remiantis aukštais NKS metodo jautrumo ir specifiskumo įverčiais galima daryti išvadą, jog NKS duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, tikrinti *Sanger* metodu nebūtina. Panašią išvadą padarė ir L. M. Baudhuin su kolegomis [18].

Svarbu atkreipti dėmesį ir į tai, kad nors *Sanger* metodas yra laikomas aukso vertės standartu, tačiau būta nenustatytų variantų būtent dėl šio metodo ribotumo, o tai kelia klausimą, ar *Sanger* metodas gali būti laikomas etalonu.

O. Harismendy bei jo kolegų atlikto tyrimo duomenimis, devyni iš de-

šimties tirtų variantų, nustatytų *Sanger* metodu ir nenustatytų NKS metodu, buvo klaidingai identifikuoti būtent dėl *Sanger* metodo klaidos [16]. Negana to, kito tyrimo metu taip pat buvo nustatyta, kad 17 iš 19 neatitiktųjų tarp *Sanger* ir NKS duomenų buvo dėl *Sanger* klaidos. Remiantis šiais rezultatais padaryta išvada, kad nėra prasmės tikrinti NKS duomenimis 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 lėšų ir laiko [15].

IŠVADOS

Sanger sekoskaitos metodu nustatyti aštuoni variantai, kurie nesutapo su NKS duomenimis – šeši iš tų variantų buvo klaidingai įvardyti arba negalėjo būti nustatomi dėl *Sanger* sekoskaitos metodo ribotumo ar klaidų. Kiti du variantai buvo klaidingai įvardyti NKS *SOLiD* platformos automatinio analizės algoritmo. Šie rezultatai rodo, kad *Sanger* sekoskaitos metodo klaidų identifikuojant variantus buvo daugiau, todėl diskutuotina, ar šis metodas laikytinas aukso vertės standartu.

SureSelect ir *TargetSeq* praturtinimo sistemos buvo lyginamos ir pagal vidutinio padengimo reikšmes. Paaiškėjo, kad *SureSelect* praturtinimo sistemos bendras padengimas yra 3,65 % didesnis. Tačiau naudojantis Mano, Vitnio ir Vilkoksono rangų sumų kriterijumi nustatyta, kad skirtingų praturtinimo sistemų padengimo vidurkiai statistiškai reikšmingai nesiskiria. Pasitelkus dviejų negiminingų asmenų, kuriems sekoskaita buvo įvykdyta abiem 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 efektyviausia taikyti abi praturtinimo sistemas, kad variantai būtų nustatyti tiksliausiai.

Darbe naudotos NKS *SOLiD* platformos automatinio analizės algoritmo tikslumas buvo 99,66 %, jautrumas – 99,22 % ir specifiskumas – 99,7 %. Remiantis šiais rezultatais galima tvirtinti, jog NKS duomenų, atitinkančių aukštus automatinio analizės algoritmo parametrus, tikrinti *Sanger* sekoskaitos metodu nebūtina. ♦

Gauta: 2018 01 02
Priimta spaudai: 2018 06 28

Summary

EVALUATION OF AUTOMATED NEXT GENERATION SEQUENCING DATA ANALYSIS PIPELINE AND DIFFERENT ENRICHMENT SYSTEMS USING SANGER SEQUENCING

Gabrielė Žukauskaitė, Laima Ambrozaitytė, Tautvydas Rančelis, Ingrida Kavaliauskienė, Vaidutis Kučinskis

Purpose. To evaluate automated next generation sequencing SOLiD platform's data analysis pipeline, defining accuracy, sensitivity and specificity values. Also, to evaluate used target enrichment systems (TargetSeq and SureSelect) using Sanger sequencing.

Participants and methods. Study included data of 96 individuals. Sixty genome variants, detected by next generation sequencing, were investigated using Sanger sequencing and other analysis tools. Coverage values of next generation sequencing data were used to compare

TargetSeq and SureSelect target enrichment systems.

Results. Six variants have been identified incorrectly or could not be identified at all because of Sanger sequencing limitations or errors. Two variants were erroneously identified by automated data analysis pipeline.

The mean coverage using SureSelect and TargetSeq enrichment systems were 32.77 and 31.58, respectively. Based on these values, the coverage of SureSelect enrichment system is 3.65% higher.

Also, more genome variants were identified using the SureSelect enrichment system. However, TargetSeq system was able to identify unique variants.

Finally, the automated next generation SOLiD data analysis pipeline's accuracy, sensitivity and specificity was estimated to be 99.66%, 99.22% and 99.7%, respectively.

Conclusions. These results indicate that there is no need to verify next generation sequencing data using Sanger sequencing when automatic analysis algorithm parameters are high.

Comparison of TargetSeq and SureSelect coverage values showed that difference between means is not statistically significant. However, both systems were able to identify unique variants, so the most effective way to accurately identify genome variants is to use both enrichment systems.

Keywords: next-generation sequencing, NGS SOLiD platform, enrichment systems, sensitivity, specificity.

LITERATŪRA

- 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.
- 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): 923–32.
- Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research* 2010; 38: e164.
- Fox J. The R Commander: a basic statistics graphical user interface to R. *Journal of Statistical Software* 2005; 14(9): 1–42.
- Rančelis T. Patogeninių genomo variantų ir jų genų, lemiančių autosominės recesyvias ligas, įvairovės analizė, panaudojant viso egzomo sekoskaitą: Doktoro disertacija. Vilniaus universitetas, 2016; 115 p.
- Johnston JJ, Rubinstein WS, Facio FM, Ng D, Singh LN, Teer JK, et al. 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.
- McCourt CM, McArt DG, Mills K, Catherwood MA, Maxwell P, Waugh DJ, et al. 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.
- Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, et al. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* 2014; 30(19): 2709–16.
- Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Brüggewirth HT, Lekanne Deprez RH, et al. 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: 1313–21.
- Shigemizu D, Momozawa Y, Abe T, Morizono T, Borojevich KA, Takata S, et al. Performance comparison of four commercial human whole-exome capture platforms. *Sci Rep* 2015; 5: 12742.
- Zhang G, Wang J, Yang J, Li W, Deng Y, Li J, et al. Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. *BMC Genomics* 2015; 16(1): 581.
- 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): 213–20.
- Londin ER, Clark P, Sponziello M, Kricka LJ, Fortina P, Park JY. Performance of exome sequencing for pharmacogenomics. *Personalized Medicine* 2014; 12(2): 109–15.
- Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, Wang J, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biology* 2011; 12(9): 95.
- Biesecker LG, Beck TF, Mullikin JC. Systematic evaluation of Sanger validation of next-generation sequencing variants. *Clinical Chemistry* 2016; 62(4): 647–54.
- Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* 2009; 10(3): R32.
- Houniet DT, Rahman TJ, Al Turki S, Hurles ME, Xu Y, Goodship J, et al. Using population data for assessing next-generation sequencing performance. *Bioinformatics* 2015; 31(1): 56–61.
- 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): 456–61.