VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

MATEMATINĖS INFORMATIKOS KATEDRA

Mantas Bernotas

Bioinformatikos studijų programa

Molekulinės biologijos eksperimentinių duomenų apdorojimas

R pagalba

Bakalauro baigiamasis darbas

Vadovas: lekt. I. Grinis

Vilnius 2015

Turinys

Santrumpos 3
Įvadas 4
1. Literatūros apžvalga 5
1.1. R programavimo aplinka 5
1.2. Bioconductor programinė įranga 5
1.3. Bioconductor bibliotekos 6
1.3.1. Affy 6
1.3.2. AffyExpress 7
1.3.3. SimpleAffy 7
1.3.4. Limma 8
1.4. Genų raiškos profiliavimas 8
1.5. „Cel“ failų formatas 8
1.6. Plaučių vėžys 9
1.6.1. Nesmulkialąstelinis plaučių vėžys (NSLPV) 11
1.6.2. NSLPV genai 11
2. Metodinis skyrius 12
2.1. Eksperimentiniai duomenys 12
2.2. Duomenų išankstinis apdorojimas 13
2.2.1. Foninis taisymas 13
2.2.2. Duomenų normalizavimas 13
2.2.3. Mikrogardelės kokybės įvertinimas 14
2.2.4. Genų pavadinimų generavimas 14
2.3. Reikšmingai skirtingos raiškos genų paieška 15
2.3.1. Tiesinis modelis 15
2.3.2. Empirinis Bajeso metodas 15
2.3.3. Vulkano tipo grafikas 16
3. Rezultatai 17
3.1. Programa 17
3.2. Duomenų apdorojimo rezultatai 17
3.2.1. Duomenų normalizavimo rezultatai 17
3.2.2. Mikrogardelių kokybės įvertinimo rezultatai 19
3.3. Reikšmingai skirtingos raiškos genai pateikti tekstiniu formatu 20
3.4. Reikšmingai skirtingos raiškos genai pateikti grafiniu formatu 21
3.5. Unikalūs reikšmingai skirtingos raiškos genai 25
Išvados 28
Santrauka 29
Summary 30
Literatūra 31

Santrumpos

AC - adenokarcinoma - piktybinis navikas, išsivystęs iš liaukinio audinio epitelinių ląstelių.

SCC – plokščiųjų ląstelių karcinoma.

ACs I - adenokarcinoma 1 stadijos.

ACs II - adenokarcinoma 2 stadijos.

SCCs I - plokščiųjų ląstelių karcinoma 1 stadijos.

SCCs II - plokščiųjų ląstelių karcinoma 2 stadijos.

Affymetrix – Komercinė Affymetrix pagaminta GeneChip DNR mikrogardelė skirta įvairioms visuminėms analizėms. Kartais taip vadinamas pats eksperimentas, kuriame atliekama visuminė analizė naudojant Affymetrix technologijas.

DNR mikrogardelės zondas (toliau zondas) – 25 nukleotidų ilgio oligonukleotidas, pritvirtintas ant mikrogardelės, skirtas nukleorūgčių iš tiriamojo mėginio prisijungimui.

NSLPV– nesmulkialąstelinis plaučių vėžys.

SLPV – smulkių ląstelių plaučių vėžys.

Pokyčio reikšmingumo slenkstis – santykis tarp kontrolinės ir tiriamos reikšmes, pavyzdžiui santykis tarp geno raiškos reikšmės iš ACs I ir SCCs I grupės to paties geno raiškos reikšmės, kurį pasirenkama laikyti svarbiu.

Įvadas

Bakalauro baigiamojo darbo objektas yra reikšmingai skirtingos raiškos genai tarp skirtingų plaučių vėžio tipų. Pagrindinis darbo tikslas yra R ir Bioconductor pagalba sukurti programinę įrangą plaučių vėžio visuminių tyrimų duomenų analizei. Trumpai aprašoma kodėl pasirinktas toks darbo tikslas bei suformuluojami svarbiausi uždaviniai sprendžiami šiame darbe.

Plaučių vėžys pasaulyje yra antra dažniausiai tiek moterims tiek vyrams diagnozuojama vėžio forma. Lietuvoje kasmet nustatoma apie 1200 naujų plaučių vėžio atvejų. Smulkialąstelinis plaučių vėžys yra labai greitai progresuojantis. Dauguma plaučių vėžio simptomų išryškėja ligai pasiekus vėlyvas stadijas. Deja, dalis net ir ankstyvoje ligos stadijoje diagnozuotų vėžio atvejų progresuoja ypač greitai. Siekiant, kad plaučių vėžiu sergančių pacientų gydymas būtų kuo veiksmingesnis būtina sukurti metodus greitai progresuojančių plaučių vėžio atvejų identifikavimui (plaučių vėžio atvejų klasifikavimo metodai) dar iki progresavimo. Vienas iš tokių metodų kūrimo kelių – visuminių tyrimų duomenų analizė.

Darbo tikslui pasiekti, suformuluoti šie uždaviniai:

- Apžvelgi Bioconductor bibliotekas;
- Atrinkti Bioconductor bibliotekas, naudojamas visuminių tyrimų duomenų analizei;
- Trumpai apžvelgti plaučių vėžio tipus ir plaučių vėžio duomenų analizavimo svarbą;
- Sukurti programą, kuri randa reikšmingai skirtingos raiškos genus tarp skirtingų plaučių vėžio tipų.

Bakalaurinis darbas suskirstytas į tris pagrindinius skyrius: literatūros apžvalgos, metodinį ir rezultatų. Kiekvienas skyrius turi mažesnius poskyrius, kuriuose išdėstyti jiems svarbiausi aspektai.

1. Literatūros apžvalga

Šiame skyriuje trumpai apžvelgiama mokslinė literatūra, kuri buvo naudojama rašant darbą. Pristatoma R programinė aplinka, Bioconductor programinė įranga. Taip pat aprašomos Bioconductor bibliotekos, skirtos mikrogardelių eksperimentinių duomenų apdorojimui ir analizei. Pateikiama trumpa apžvalga apie plaučių vėžį.

1.1. R programavimo aplinka

R yra programavimo kalba ir programavimo aplinka skirta statistiniams skaičiavimams ir grafiniam duomenų vizualizavimui. R kalba yra labai populiari tarp statistikų dėl statistinių programų kūrimo ir patogios duomenų analizės [1]. Pastaruoju metu R kalbos populiarumas vis labiau auga [2]. Dėl gebėjimų valdyti didelius duomenų kiekius ir kalbos lankstumo ji tapo vienu plačiausiai naudojamu programiniu įrankiu bioinformatikoje [3].

1.2. Bioconductor programinė įranga

Bioconductor yra projektas, skatinantis bendradarbiavimą kuriant programinę įrangą bioinformatikai ir skaičiuojamajai biologijai [4]. Bioconductor pagrindas yra R programavimo kalba. Biologijoje ir molekulinėje biologijoje vyksta dvi susijusios transformacijos:

- Gerėja supratimas apie skaičiavimo pobūdžius daugelyje biologinių procesų ir skaičiavimo bei statistiniai modeliai gali būti naudingiau panaudojami;
- Pokyčiai didelio našumo duomenų gavime nustato reikalavimus skaičiavimams ir statistinėms žinioms kiekviename biologinio tyrimo etape.

Pagrindinis Bioconductor projekto tikslas yra sukurti stabilią ir lanksčią programų kūrimo ir priežiūros aplinką, kuri atitinka naujus skaičiavimų iššūkius. Taip pat, siekiama sumažinti kliūtis, norint patekti į mokslinių tyrimų rinką bioinformatikos ir biologijos srityse. Siekiama supaprastinti procesus, kurių pagalba statistikai gali tikslingai naudoti ir tirti duomenų išteklius bei algoritmus, ir kurių pagalba biologai gali gauti prieigą naudotis statistiniais metodais, kad gautų tikslias išvadas.

Tarp daugelio iššūkių, kurie kyla tiek statistikams, tiek biologams, yra tokie uždaviniai, kaip:

- duomenų gavimo,

- duomenų valdymo,
- duomenų transformavimo,
- duomenų modeliavimo,

apjungiant skirtingus duomenų šaltinius, kuriant naujas modeliavimo strategijas tinkančias bioinformatikai ir biologijai. Šiems iššūkiams Bioconductor kaip atsaką pateikia:

- aiškumą;
- atkuriamumą;
- plėtros efektyvumą.

Esminis visų šių užduočių bruožas – programinės įrangos poreikis, vien idėjos negali išspręsti kylančių problemų. Bioconductor siūlo būtent tai, ko reikia – lengvai prieinamą ir patogią programinę įrangą.

1.3. Bioconductor bibliotekos

1.3.1. Affy

Affy biblioteka yra dalis Bioconductor projekto. Patogi ir interaktyvi aplinka Affymetrix oligonukleotidų mikrogardelių zondų lygmens duomenų analizei ir tyrimui.

Affy įrankiai, aprūpinti Affymetrix programinės įrangos paketais, apibendrina zondų rinkinių intensyvumą ir suformuoja vieną raiškos vertę kiekvienam genui. Genų raiškos vertės yra duomenys tinkami analizei. Tačiau, daug galima sužinoti, tiriant ir atskirus zondų intensyvumus, atliekant zondų lygmens duomenų analizę [5]. Į biblioteką įeina:

- grafikų braižymo funkcijos, kurios reikalingos mikrogardelių kokybės tikrinimui;
- RNR degradacijos vertinimas;
- skirtingo lygmens zondų duomenų normalizacijos;
- foninių pataisymų procedūros;
- lanksčios funkcijos, kurios leidžia vartotojui konvertuoti zondų lygmens duomenis į genų raiškos vertes.

Biblioteka turi įrankius, kurie apskaičiuoja genų raiškos vertes panašiai kaip:

- MAS 4.0 AvDiff [6]. Naudoja visus zondus, esančius ant mikrogardelės, gali apimti ir tikrai skirtingos raiškos genus;

- MAS 5.0 [7]. Naudoja nesutampančių zondų duomenis, kad apskaičiuotų vidurkį remiantis nesutampančių ir sutampančių zondų raiškos reikšmių skirtumu;
- DChip MBEI [5]. Naudoja tik nekintančius zondų rinkinius, išskiriami galimai skirtingos raiškos genai;
- RMA [8]. Pateikia logaritmuotas genų raiškos vertes, kurios tinkamos tolimesniems lyginimams.

1.3.2. AffyExpress

Šios bibliotekos tikslas – pateikti detalų ir lengvai naudojamą įrankį kokybės vertinimui ir skirtingos raiškos genų identifikavimui, Affymetrix genų raiškos duomenims. Funkcijos pritaikytos biologams, kurie turi mažai statistinių žinių, generuojant dizaino ir kontrasto matricas tiek paprasto, tiek sudėtingo dizaino eksperimentams. Vartotojai gali pasirinkti arba įprastą tiesinės regresijos modelį Limma, arba kombinuotus testus skirtingos raiškos genų paieškai. Skirtingos raiškos genai pateikiami lentelėje su anotacijomis susietomis su internetinėmis biologinėmis duomenų bazėmis. Wrapper funkcijos sukurtos taip, kad analizė būtų dar paprastesnė.

1.3.3. SimpleAffy

Dalis Bioconductor projekto. SimpleAffy sukurtas siekiant pateikti atspirties tašką tiriant Affymetrix duomenis ir pateikti dažnai naudojamas funkcijas, kurios kartojasi kiekvienoje užduotyje. Paremtas affy biblioteka, kuri atlieka didžiąją dalį darbo. Nors affy biblioteka turi daugybę įvairiausių funkcijų duomenų apdorojimui, tačiau kai kurioms užduotims, tokioms kaip: t-testų skaičiavimas, pokyčių tarp grupių radimas, grafikų braižymas, lentelių generavimas, reikalingas ilgas programos kodo rašymas. Taip pat, kai kurios dažnai naudojamos funkcijos galėtų būti atliekamos greičiau. Ši biblioteka siekia aprūpinti vartotoją aukšto lygio metodais, reikalingais kasdieniuose tyrimuose, o dauguma metodų perrašyti C programavimo kalba, taip išlošiant greičio.

1.3.4. Limma

Limma (Linear Models for Microarray and RNA-seq Data) yra R biblioteka, skirta genų raiškos mikrogardelių eksperimentų duomenų analizei [9]. Ypač dažnai naudojami tiesiniai modeliai, analizuojant suprojektuotus eksperimentus ir vertinant skirtingą genų raišką [10]. Limma suteikia galimybę analizuoti lyginimus tarp daugybės RNR taikinių vienu metu net ir labai sudėtinguose eksperimentuose. Bajeso metodai pateikia stabilius rezultatus, net jeigu analizuojamas mažas kiekis duomenų [10]. Normalizacijos ir duomenų analizės funkcijos yra skirtos dvispalvių mikrogardelių eksperimentų duomenų analizei. Tiesiniai modeliai ir skirtingos genų raiškos funkcijos veikia visose mikrogardelių technologijose, įskaitant Affymetrix ir kitas vienspalves oligonukleotidų platformas.

1.4. Genų raiškos profiliavimas

Žiūrint iš molekulinės biologinės pusės, genų raiškos profiliai yra tūkstančių genų veiklos matavimas tuo pat metu, kad būtų sukurtas bendras ląstelės funkcijų vaizdas. Tokie profiliai gali, pavyzdžiui, atskirti ląsteles, kurios yra aktyvaus dalijimosi arba parodyti, kaip ląstelės reaguoją į atitinkamą gydymą. Dažniausiai naudojami genų raiškos profiliavimo metodai:

- SAGE, SuperSAGE – nuoseklioji genų raiškos analizė;
- DNR mikrogardelių;
- RNA-seq – skaitmeninė mikrogardelių alternatyva.

DNR mikrogardelių metodas yra labiau patikimas ir ekonomiškesnis genų raiškos profiliavimo metodas už RNA-seq. Tačiau RNA-seq ateityje turėtų tapti dažniau atliekamu tyrimu už mikrogardelių metodą, kuris galbūt bus naudojamas tik tam tikrais atvejais [11].

1.5. „Cel“ failų formatas

„Cel“ tipo faile saugomi praktiškai visi duomenys gauti iš zondų esančių ant Affymetrix GeneChip mikrogardelės, todėl „cel failą galime laikyti pačia mikrogardele. „Cel“ failuose yra saugomi pikselių reikšmių iš „dat“ failo intensyvumo skaičiavimo rezultatai [12]. Failas apima:

- intensyvumo vertę;
- standartinį intensyvumo nuokrypį;

- pikselių skaičių, panaudotą apskaičiuojant intensyvumo vertę;
- žymenį atributui, kuris rodo, kad atitinkamas atributas apskaičiuotas automatiškai;
- vartotojo sukurtą žymenį atributui, kuris parodo, kad atributas neturėtų būti naudojamas tolimesniuose tyrimuose.

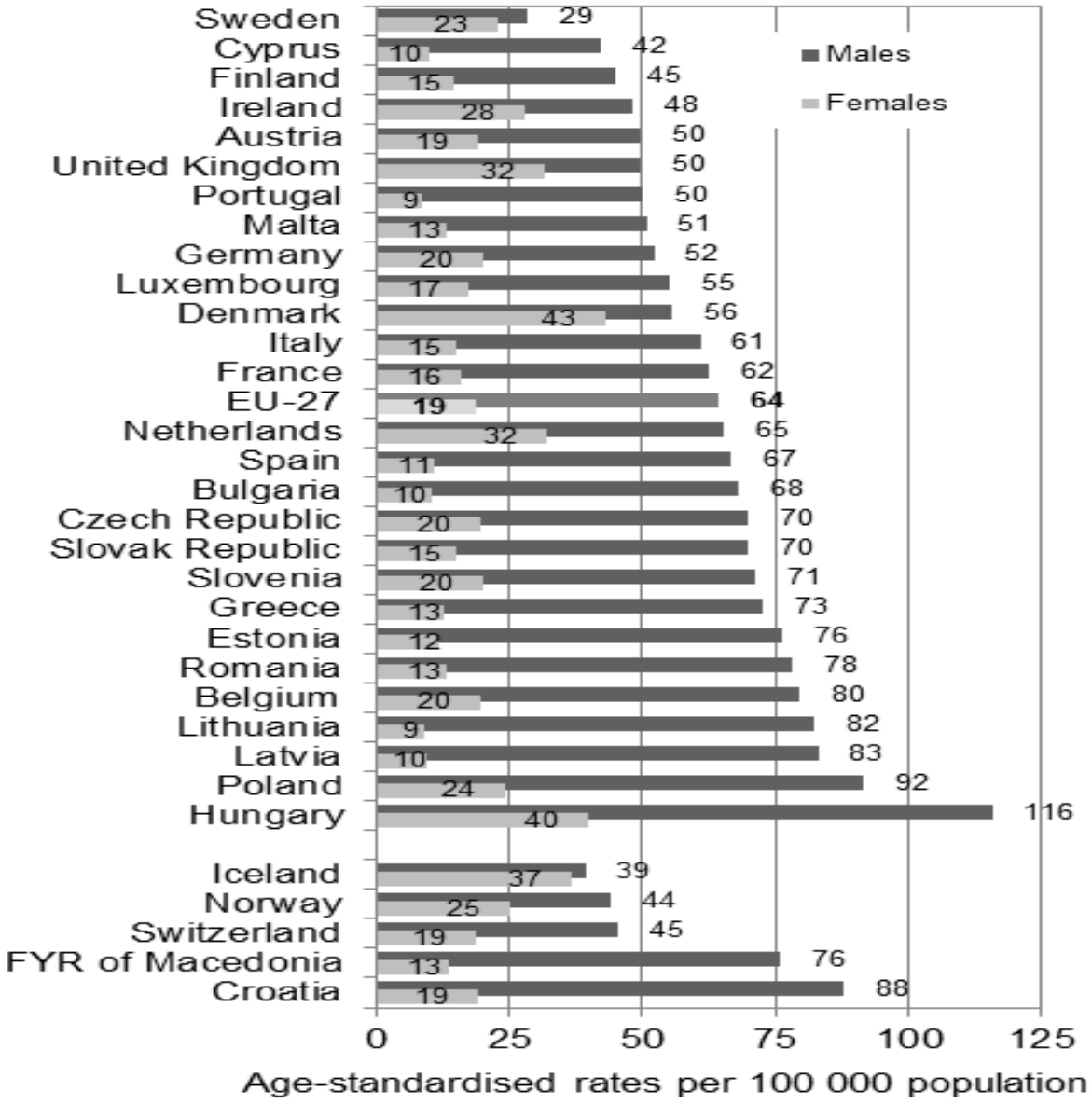
1.6. Plaučių vėžys

Kiekvienais metais Lietuvoje diagnozuojamas plaučių vėžys maždaug 1200 gyventojų. Jis gali užklupti netikėtai, nors jo pagrindinė priežastis yra rūkymas. Jeigu vėžys nustatomas ankstyvoje stadijoje jį įmanoma pagydyti ar bent jau sulėtinti ligos eigą.

Plaučių vėžys dažniausiai pasitaiko dviejų tipų: smulkialąstelinis (SLPV) ir nesmulkialąstelinis (NSLPV). Smulkialąstelinis plaučių vėžys dažniausiai prasideda lastelėse, kurios gamina hormonus arba nervinėse ląstelėse, esančiose plaučiuose. Šia vėžio forma dažniausiai serga rūkantys žmonės, labai retas atvejis, kad šia vėžio forma sirgtų niekada nerūkęs žmogus. SLPV plinta labai greitai [13]. NSLPV yra labiau paplitęs ir dažniausiai atsiranda epitelio ląstelėse, todėl forma apibūdinama pagal tai, kurioje vietoje vėžinės ląstelės pradėjo augti pirmiausia. Yra išskiriamos trys pagrindinės NSLPV vėžio rūšys, sudarytos pagal navikų vystymosi panašumą [14]:

- Plokščialąstelinis vėžys arba epidermoidinė karcinoma – dažniausiai yra aptinkama netoli plaučių centro, kuriame nors iš pagrindinių kvėpavimo takų (kairiajame arba dešiniajame bronche). Plinta ląstelėse, kurios yra išklojusios kvėpavimo takus;
- Adenokarcinoma – dažniausiai prasideda ląstelėse, kurios gamina gleives ir iškloja kvėpavimo takus;
- Stambialąstelinis vėžys arba stambialąstelinė karcinoma, dažniausiai plintanti kitose ląstelėse nei pirmosios dvi rūšys.

Šio vėžio gydymas yra susijęs su tuo ar jis yra išplitęs į kitus organus ar limfmazgius, kadangi plaučiai yra didelis organas ir navikai gali plisti ilgą laiką, o tokius simptomus kaip kosulys ar nuovargis daugelis žmonių sieja su kitokio pobūdžio ligomis. Būtent dėl šios priežasties pirmą ir antrą stadijas sunku nustatyti, daugumai žmonių jis diagnozuojamas 3 arba 4 stadijoje [14]. Nuo plaučių vėžio mirusių vyrų skaičius yra didesnis už moterų visoje Europoje. Tačiau plaučių vėžys yra ir viena pagrindinių vėžio rūšių pasiglemžiančių ir moterų gyvybes. 1 pav. vaizduojamas mirtingumas, vyrų ir moterų, nuo plaučių vėžio Europos šalyse 2010 metais [15].



Pav. 1: Vyrų ir moterų mirčių skaičius nuo plaučių vėžio, 2010 metais, Europos šalyse [15].

1.6.1. Nesmulkialąstelinis plaučių vėžys (NSLPV)

Bakalauro darbe analizuojami nesmulkialąstelinio plaučių vėžio duomenys, todėl šis plaučių vėžio tipas aptariamas plačiau. Nesmulkialąstelinio plaučių vėžio stadijos [16]:

- 1 stadijos - vėžys yra tik plaučiuose, neišplitęs į kitus organus ar limfmazgius,
- 2 stadijos – vėžys išplitęs plaučiuose ir netoli limfmazgių,
- 3 stadijos – vėžys išplitęs plaučiuose ir limfmazgiuose,
- 4 stadijos – vėžys išplitęs abiejuose plaučiuose, stipriai progresuoja. Taip pat gali būti išplitęs į kitas kūno dalis, pavyzdžiui kepenis.

1.6.2. NSLPV genai

Per pastarąjį dešimtmetį, buvo nustatyta, kad NSLPV stadijos tiksliau gali būt apibūdinamos molekuliniame lygmenyje, pasitelkiant pasikartojančias mutacijas, kurios atsiranda dėl kelių onkogenų, tokių kaip: AKT1, ALK, BRAF, EGFR, HER2, KRAS, MEK1, MET, NRAS, PIK3CA, RET ir ROS1 [17]. Mutacija sukelia mutantinių signalinių baltymų aktyvaciją, kuri skatina ir palaiko navikų augimą. Šios mutacijos dar vadinamos „driver mutation“. Jos įgyja išskirtinį pranašumą prieš ląsteles, iš kurių jos klonavosi, nes padidėja išgyvenimo ir reprodukcijos tikimybė toje pačioje ląstelių augimo aplinkoje. Ši mutacija taip pat linkusi sukelti pradinės ląstelės irimą. Jas galima rasti visose NSLPV stadijose, nepriklausomai nuo to ar žmogus rūkė ar ne. Tiems, kurie nerūkė, tačiau serga adenokarcinoma dažniausiai pasireiškia EGFR, HER2, ALK, RET ir ROS1 genų mutacijos [17].

2. Metodinis skyrius

Šiame skyriuje aprašomi apdorojami molekulinės biologijos eksperimentiniai duomenys, išankstinis šių duomenų apdorojimas ir reikšmingai skirtingos raiškos genų paieškai naudojami metodai. Visi toliau minimi failai pateikiami kartu su šiuo darbu.

2.1. Eksperimentiniai duomenys

Duomenys pasirinkti analizei: 40 vėžinių ląstelių mėginių genų raiškos profiliai gauti Affymetrix mikrogardelių pagalba. Duomenys „cel“ formatu buvo atrinkti ir parsiųsti iš http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE43580. Recipientams, iš kurių buvo paimti pavyzdžiai, buvo diagnozuotas plaučių vėžys. Taip pat pavyzdžiai buvo suskirstyti pagal keturis plaučių vėžio tipus kaip nurodoma 1 lentelėje.

Lentelė 1. Duomenų pasiskirstymas pagal plaučių vėžio tipus.

Plaučių vėžio tipas	*„cel“ failai*		
ACs I	GSM1065772.CEL GSM1065778.CEL GSM1065788.CEL GSM1065797.CEL	GSM1065774.CEL GSM1065783.CEL GSM1065789.CEL	GSM1065776.CEL GSM1065784.CEL GSM1065793.CEL
ACs II	GSM1065765.CEL GSM1065770.CEL GSM1065781.CEL GSM1065798.CEL	GSM1065767.CEL GSM1065777.CEL GSM1065785.CEL	GSM1065769.CEL GSM1065780.CEL GSM1065791.CEL
SCCs I	GSM1065739.CEL GSM1065756.CEL GSM1065768.CEL GSM1065786.CEL	GSM1065740.CEL GSM1065758.CEL GSM1065773.CEL	GSM1065743.CEL GSM1065766.CEL GSM1065782.CEL
SCCs II	GSM1065749.CEL GSM1065760.CEL GSM1065775.CEL GSM1065792.CEL	GSM1065750.CEL GSM1065764.CEL GSM1065779.CEL	GSM1065755.CEL GSM1065771.CEL GSM1065787.CEL

Rūkančiųjų ir nerūkančiųjų skaičius tyrime buvo subalansuotas. Taip pat buvo surinkta klinikinė informacija apie vėžio vystymosi istoriją, amžių, lytį, rūkymo istoriją.

2.2. Duomenų išankstinis apdorojimas

Bet kokie duomenų rinkiniai gauti iš mikrogardelių eksperimentų, tokie kaip genų raiškos profiliai, turi būti apdorojami prieš juos analizuojant ir interpretuojant. Išankstinis apdorojimas tai žingsnis, kurio metu atskiriamos prasmingos duomenų charakteristikos ir paruošiamas duomenų rinkinys, reikalingas atitinkamiems duomenų analizės metodams.

2.2.1. Foninis taisymas

Foninis taisymas yra skirtas reguliuoti nespecifiškai prisijungusius pavyzdžio RNR taikinius, kurių seka nepilnai atitinka zondo seką ant mikrogardelės. Nespecifinis prisijungimas gali būti įvertintas pagal fluorescencijos lygį šalia zondo [18]. Ant Affymetrix mikrogardelių, kurių zondai padengia visą paviršių, foninis lygmuo gali būti vertinamas pagal nesutampančius zondus (mismatch probes) [19]. Nesutampantys zondai yra identiški idealiai sutampantiems zondams, išskyrus vieną bazių porą, kuri yra viduryje zondo sekos. Taigi, nesutampančių zondų intensyvumo lygiai pateikia informaciją apie nespecifiškai prisijungusius RNR taikinius. Foniniam taisymui naudojamas tas pats metodas kaip ir normalizacijai „gcrma“.

2.2.2. Duomenų normalizavimas

Tipinis išankstinio duomenų apdorojimo pavyzdys yra duomenų normalizavimas arba normalizacija. Normalizacija yra išankstinis duomenų apdorojimas, kuris skirtas ištaisyti sisteminius skirtumus tarp genų arba mikrogardelių. Pavyzdžiui, dviejų spalvų DNR mikrogardelėje, pavyzdžio intensyvumai pažymėti žaliais dažais (Cy3) gali būti pastoviai didesnis, negu tuose pavyzdžiuose, kurie nudažyti raudonais dažais (Cy5). Dėl šios priežasties, atsižvelgus tik į raudonų ir žalių intensyvumų santykį būtų netiksliai atspindėtas RNR kiekis pavyzdyje. Šis disbalansas tarp dvispalvių mikrogardelių žinomas, kaip „dažų paklaida“ [20].

Affymetrix mikrogardelių pateikiami zondų intensyvumai duotoje gardelėje gali būti pastoviai didesni arba mažesni, negu kitokiose mikrogardelėse. Toks skirtumas yra žinomas kaip „gardelių paklaida“. Taigi, lyginant to paties zondo intensyvumus skirtingose gardelėse galima gauti rimtas klaidas, jeigu prieš lyginimą nėra atliekama normalizacija.

Tokių paklaidų yra ir daugiau, tačiau kai kurias paklaidas galime sumažinti normalizacijos pagalba. Išankstiniam mikrogardelių duomenų apdorojimui yra sukurta daugybė

programinės įrangos. Tarp jų yra ir jau minėtas Bioconductor projektas, turintis geriausius žinomus algoritmus išankstiniam duomenų apdorojimui.

Duomenų normalizavimui naudojamas „gcrma“ metodas. Šis metodas koreguoja mikrogardelių duomenų, kurie apima optinį triukšmą ir nespecifinį prisijungimą prie zondo, foninius intensyvumus. Pagrindinė funkcija – konvertuoti zondų intensyvumus į genų raiškos vertes naudojant tuos pačius normalizacijos metodus, kaip „rma“ (nuo angl. Robust Multiarray Average).

2.2.3. Mikrogardelės kokybės įvertinimas

RLE (nuo angl. Relative Log Expression) ir NUSE (nuo angl. Normalized Unscaled Standard Error) grafikai yra naudingi ir jautrūs norint įvertinti mikrogardelės kokybę [21]. Abu kilę iš zondų lygmens modelių (PLM).

RLE grafikas konstruojamas panaudojant logaritminius kiekvieno zondų rinkinio raiškos įverčius kiekvienai mikrogardelei („cel“ failui). Kiekvienam zondų rinkiniui ir kiekvienai mikrogardelei santykiai paskaičiuojami pagal zondų rinkinio raišką ir medianą, apskaičiuotą iš visų eksperimento mikrogardelių raiškų būtent šiam zondų rinkiniui. Kiekvienai mikrogardelei šios santykinės raiškos reikšmės pavaizduojamos, kaip stulpeliai. Kadangi tikimasi, kad daugumoje eksperimentų tik palyginti nedaug genų turės skirtingą raišką, stulpeliai turėtų būti panašių ribų ir jų centrai artimi 0.

NUSE pateikia normalizuotas standartines paklaidas iš PLM. Standartinė paklaida normalizuota kiekvienam zondų rinkiniui taip, kad standartinės paklaidos mediana kiekvienai mikrogardelei lygi 1. Kiekvienai mikrogardelei („cel“ failui) stulpeliais pavaizduojamos NUSE reikšmes. Mikrogardelės, kurių kokybė žemesnė, turės stulpelius, kurių centras bus aukščiau ir stulpelis bus labiau praplėstas negu tų, to paties eksperimento mikrogardelių, kurių kokybė gera. Dažniausiai, stulpeliai, kurių centras yra virš 1.1 parodo mikrogardeles, kurios turi rimtų kokybės problemų.

2.2.4. Genų pavadinimų generavimas

Mikrogardelių duomenų analizės metu dirbama su zondais, o ne su genais, todėl prieš pateikiant rezultatus būtina zondui suteikti jam priklausančio geno pavadinimą. Tai galima daryti rankiniu būdu internetinio įrankio „GeneAnnot“ pagalba [22]. Žinoma, tai nėra patogus būdas, kai reikia rasti daugiau genų pavadinimų atitikmenų.

Kitas būdas yra pasinaudoti Bioconductor biblioteka „hgu133plus2.db“. Kurios pagalba galima kiekvienam zondui iš sąrašo priskirti atitinkamą geno pavadinimą. Tai greitas ir efektyvus būdas genų pavadinimo generavimui. Būtent šią biblioteką naudoja sukurta programa.

2.3. Reikšmingai skirtingos raiškos genų paieška

Programos pagrindinis tikslas yra atrasti genus, kurių raiška tarp dviejų plaučių vėžio tipų reikšmingai skiriasi. Kadangi pasirinkti 4 plaučių vėžio tipai: ACs I, ACs II, SSCs1, SSCs2, programa atlieka 6 lyginimus, lygindama kiekvieną tipą, vieną su kitu ir ieško genų, kurių raiška reikšmingai skiriasi kiekvienu atveju.

2.3.1. Tiesinis modelis

Skirtingos raiškos genų paieškai naudojama Limma Bioconductor biblioteka, kuri pasižymi patogiai naudojamais tiesiniais modeliais. Šie modeliai leidžia analizuoti sudėtingus eksperimentus beveik taip pat lengvai, kaip paprastus. Šiems modeliams būtina aprašyti dvi matricas:

- Dizaino matrica, kuri reprezentuoja skirtingus tiriamuosius mėginius. Pasirinktų duomenų atveju tiriamieji mėginiai yra 4 tipų (ACs I, ACs II, SCCs I, SCCs II);
- Kontrasto matrica, kuri leidžia pagal dizaino matricą sudėlioti norimus kontrastus palyginimams. Pasirinktų duomenų atveju tai yra 6 lyginimai tarp visų 4 plaučių vėžio tipu vienų su kitais;

Programa pritaiko duotus duomenis pagal nurodytas matricas, taip užpildydama modelį.

2.3.2. Empirinis Bajeso metodas

Empirinis Bajeso metodas yra statistinis metodas, skirtas įvertinti nežinomus parametrus, stebint tarpusavyje susijusius duomenis. Jis pateikia kompromisą tarp modelio tikimo atskirai duomenims kiekvienu atveju, gaunant atvejui specifinius parametrų įverčius, ir tuo pat metu pritaikant modelį visiems atvejams su prielaida, kad įvertis yra tinkamas visiems atvejams [23]. Toks metodas idealiai tinka mikrogardelių tyrimams, kur duomenų kiekis yra didelis. Pasirinktų duomenų atveju, empirinis Bajesas atliekamas sukonstruotam tiesiniam

modeliui. Gaunamas genų sąrašas, kuriame kiekvienam genui yra priskirtos apskaičiuotos P-reikšmės, logaritminė pokyčio reikšmingumo slenksčio reikšmė ir kiti statistiniai parametrai. Aktuali yra P-reikšmė, pagal kurią randami grupėse skirtingos raiškos genai. Pasirenkami genai, kurių logaritminės pokyčio reikšmingumo slenksčio reikšmės modulis didesnis už 2, ir sugeneruojami tekstiniai „csv“ formato failai. Vaizduojant grafikus taip pat reikalinga logaritminė pokyčio reikšmingumo slenksčio reikšmė.

2.3.3. Vulkano tipo grafikas

Vulkano tipo grafikas yra taškinis grafikas, kuris dažnai naudojamas mikrogardelių duomenų rinkinių analizei, svarbių genų peržiūrai [24]. Grafiką sudaro ant x ašies vaizduojama geno logaritminė pokyčio reikšmingumo slenksčio reikšmė ir ant y ašies neigiamas dešimtainis P-reikšmės logaritmas. Vaizduojami biologiškai ir satistiškai reikšmingi genai. Biologinė įtaka pokyčiui stebima ant x ašies, y ašis rodo statistinį pokyčio patikimumą. Mokslininkai iš tokių grafikų gali atrinkti svarbiausius atvejus tolimesniems tyrimams. Programa vulkano tipo grafike išskiria genus, kurių logaritminės pokyčio reikšmingumo slenksčio reikšmės modulis yra didesnis už 2 ir kurių P-reikšmė yra mažesnė už 0.05 ir genų, kurių logaritminės pokyčio reikšmingumo slenksčio reikšmės modulis yra didesnis už 2, skaičiaus santykį.

3. Rezultatai

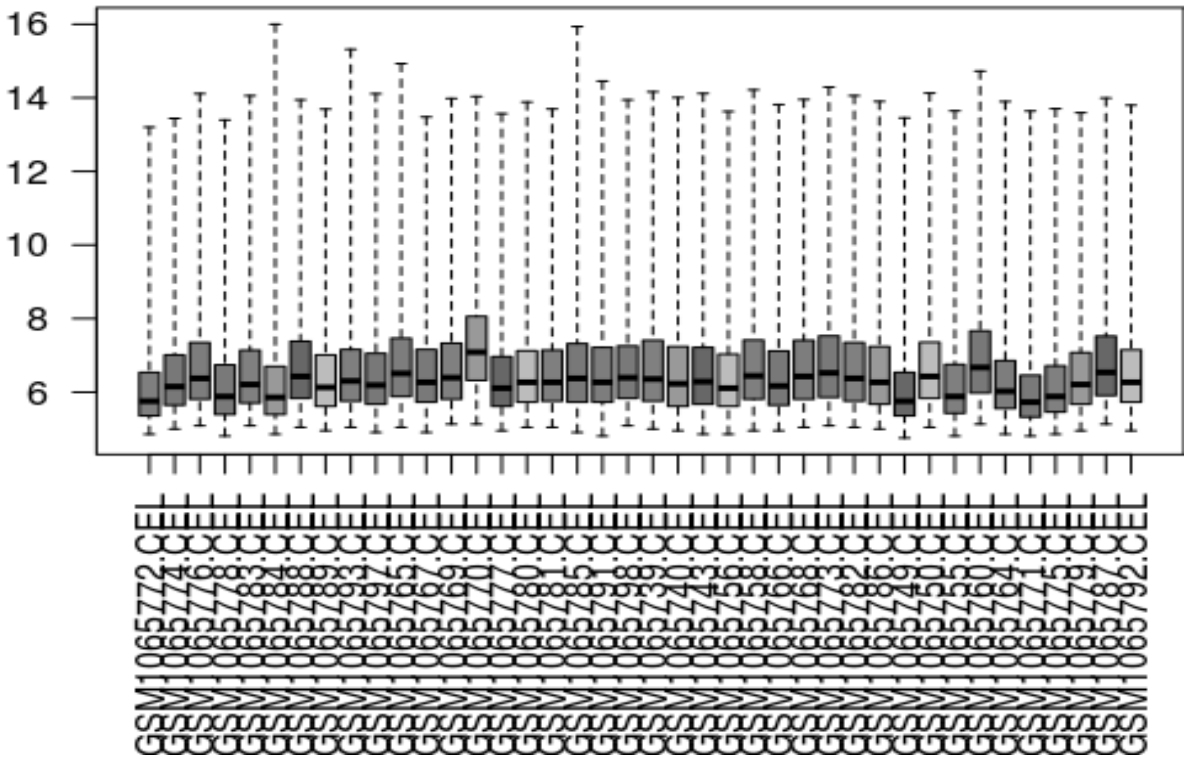
3.1. Programa

Visus toliau pateikiamus rezultatų paveikslus generuoja R ir Bioconductor pagalba sukurta programa. Programos kodas, bibliotekų parsiuntimo skriptas, naudojami duomenys, gaunami rezultatai ir trumpa instrukcija pateikiami kartu su šiuo darbu. Kadangi programa sukurta R pagalba, tai programai reikalinga pati R programa. Taip pat programa perrašyta „R markdawn“ formatu, tad turint „RStudio“ programą ir „knitr“ biblioteką jame, galima generuoti programos rezultatus „pdf“ formatu su trumpu aprašymų.

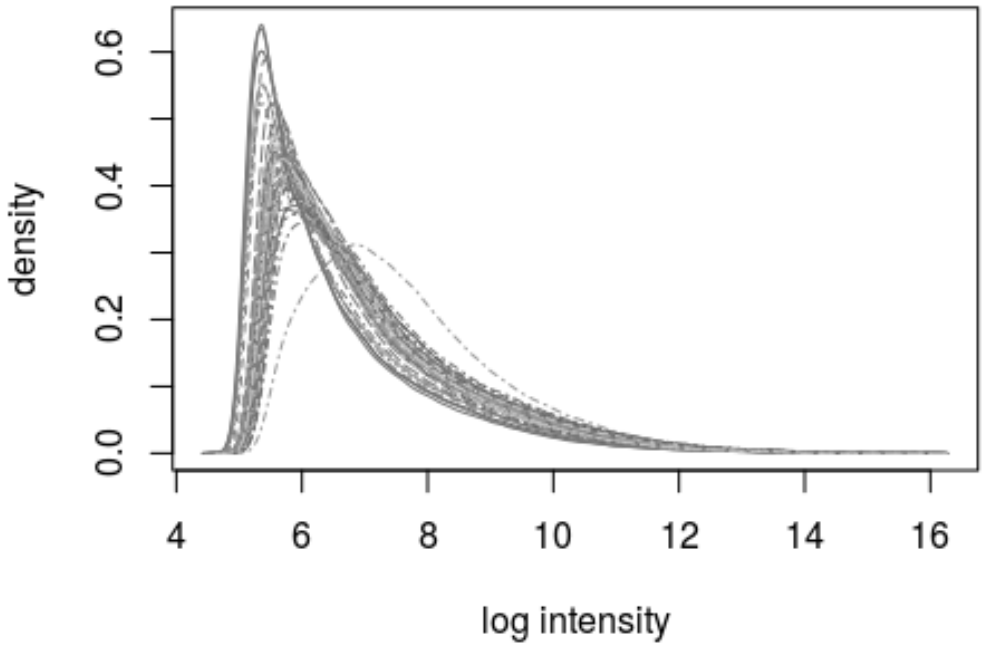
3.2. Duomenų apdorojimo rezultatai

3.2.1. Duomenų normalizavimo rezultatai

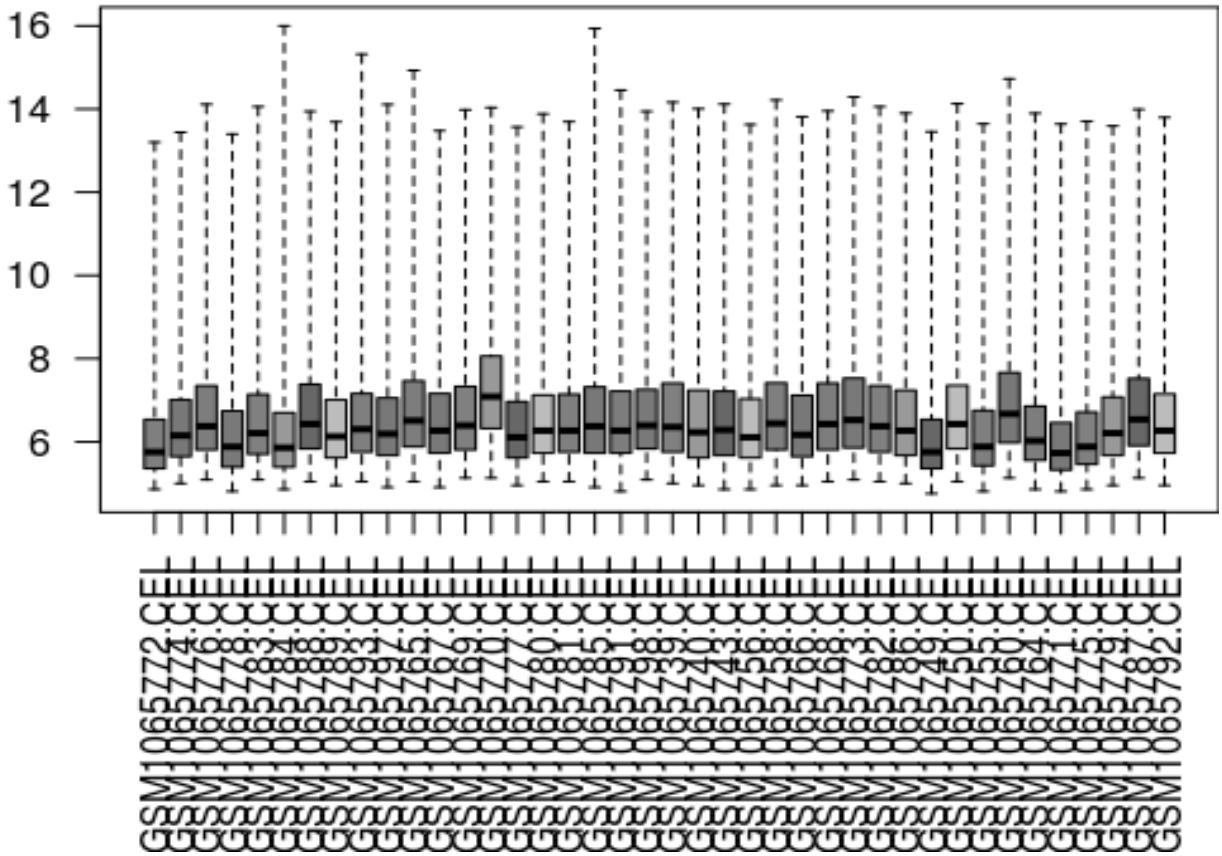
Normalizacijos rezultatai gerai matomi „boxplot“ ir logaritminio intensyvumo tankio grafikuose. Rezultatai prieš normalizaciją pateikiami 2 ir 3 pav., o jau normalizuoti duomenys pateikiami 4 ir 5 pav..



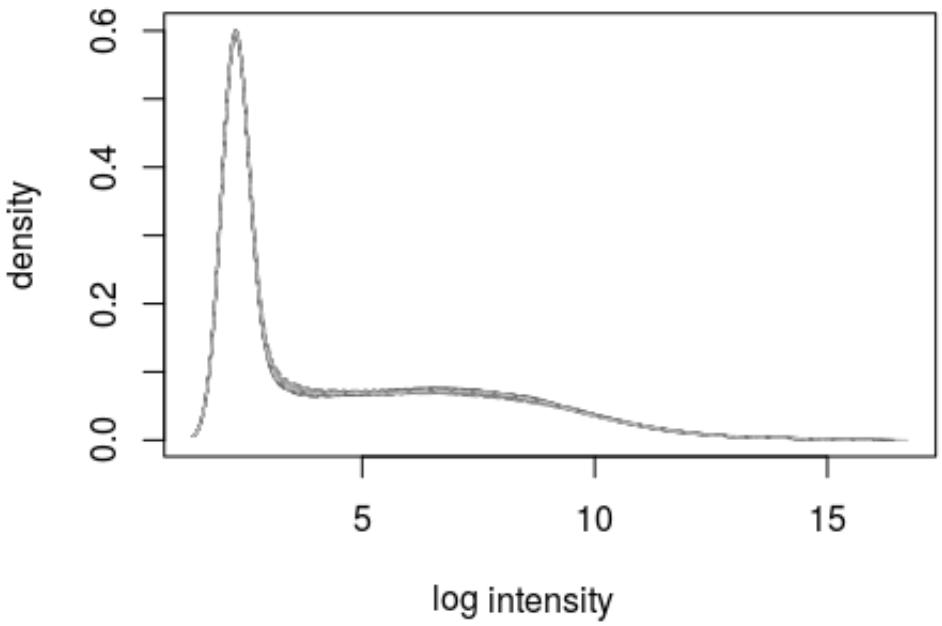
Pav. 2: Duomenys prieš normalizaciją pavaizduoti „boxplot“ tipo grafike.



Pav. 3: Nenormalizuotų duomenų logaritminio intensyvumo tankio grafikas.



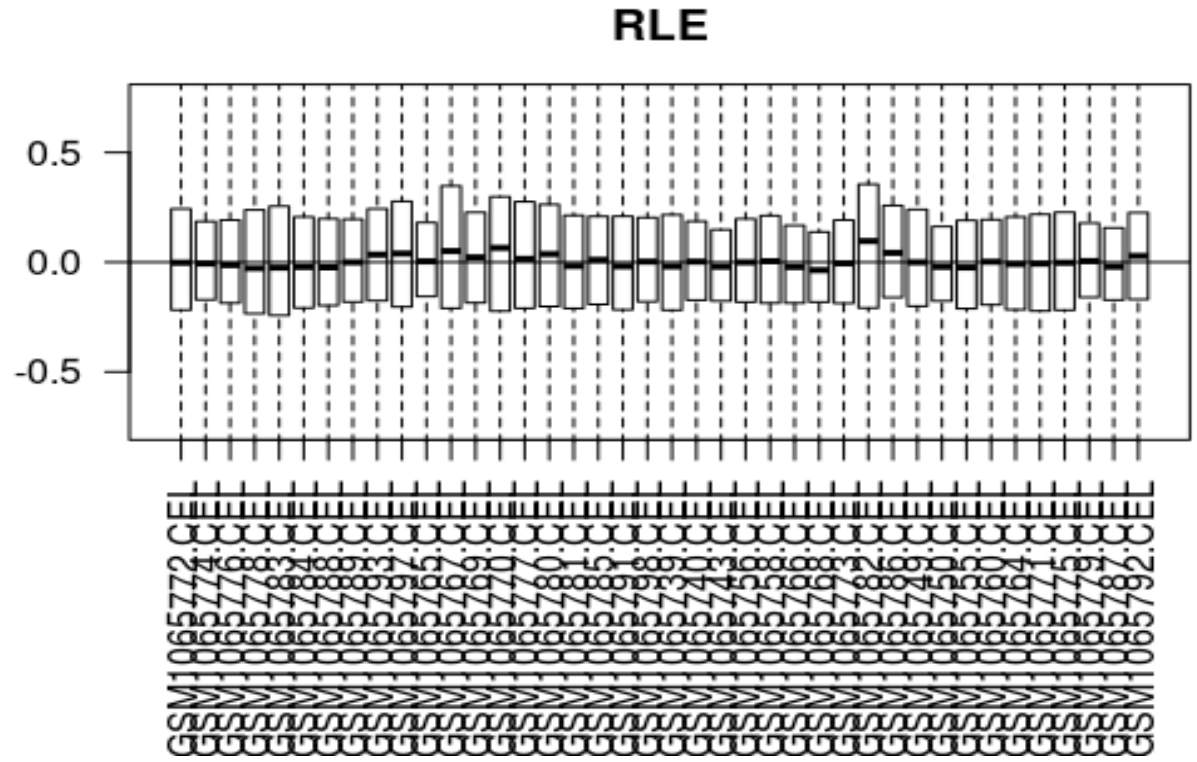
Pav. 4: Duomenys po normalizaciją pavaizduoti „boxplot“ tipo grafike.



Pav. 5: Normalizuotų duomenų logaritminio intensyvumo tankio grafikas.

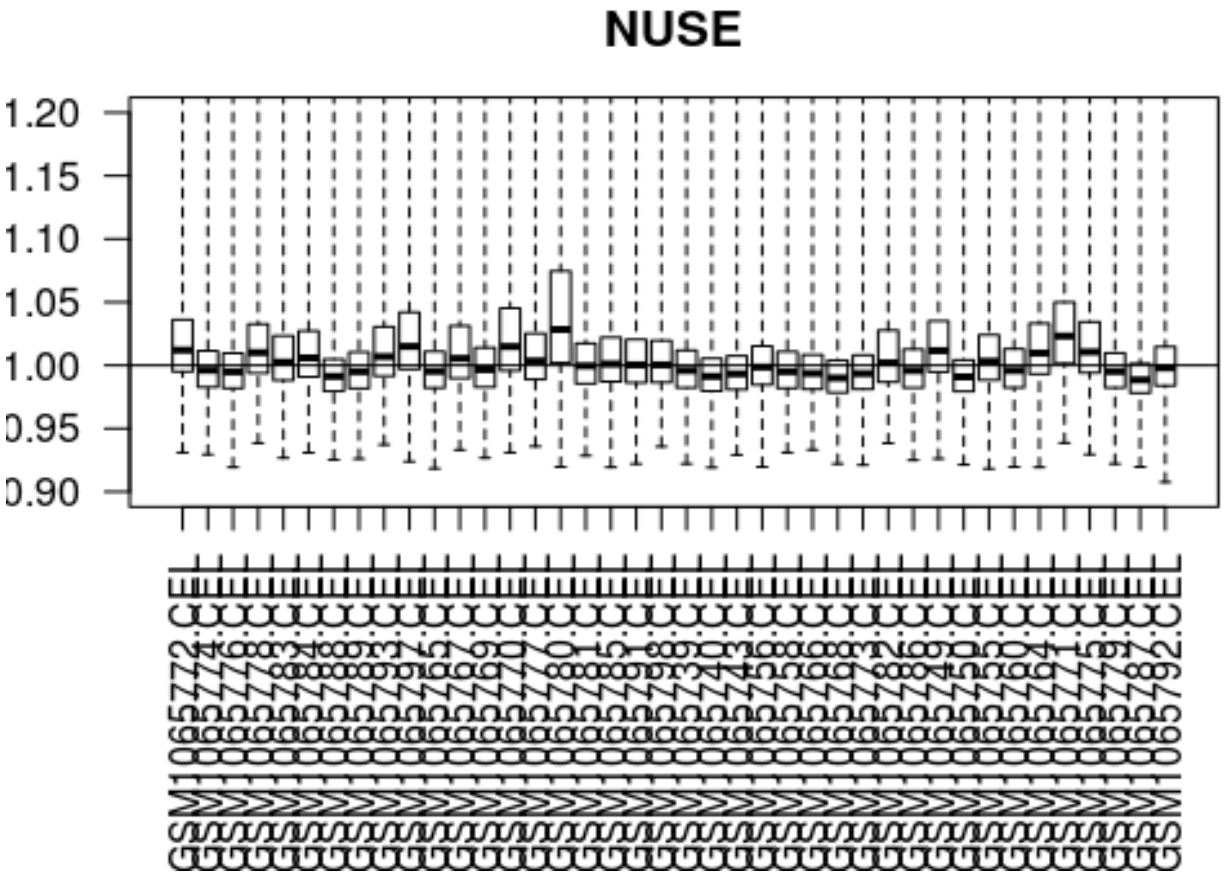
3.2.2. Mikrogardelių kokybės įvertinimo rezultatai

Pateikiami RLE ir NUSE grafikai, iš kurių yra sprendžiama apie pasirinktų mikrogardelių kokybę (6 ir 7 pav.).



Pav. 6: Analizuojamų duomenų RLE grafikas.

Iš RLE grafiko matome, kad „GSM1065782.CEL“ stulpelio centras ganėtinai nutolęs nuo 0, tačiau šis nuokrypis nėra toks reikšmingas, kad įtakotų rezultatus, todėl paliekame tą patį „cel“ failą.



Pav. 7: Analizuojamų duomenų NUSE grafikas.

Iš grafiko galima matyti, kad nėra mikrogardelių, kurių stulpelio centras viršytų 1.1 ribą. Taigi galima teigti, kad pasirinktos mikrogardelės yra geros kokybės ir tolimesni rezultatai, nebus iškreipti dėl kokybės problemų.

3.3. Reikšmingai skirtingos raiškos genai pateikti tekstiniu formatu

Programa sugeneruoja reikšmingai skirtingos raiškos genų paieškos rezultatus ir juos pateikia 6 tekstinėmis ir 6 grafinėmis bylomis. Dviejų skirtingų plaučių vėžio tipų genų raiškos skirtumų analizės rezultatas yra dvi bylos.

Tekstinės bylos išsaugomos „csv“ formatu. Jose išrenkami genai, kurių logaritminės pokyčio reikšmingumo slenksčio reikšmės modulis didesnis už 2. Bylos pavyzdys pateikiamas 8 pav.

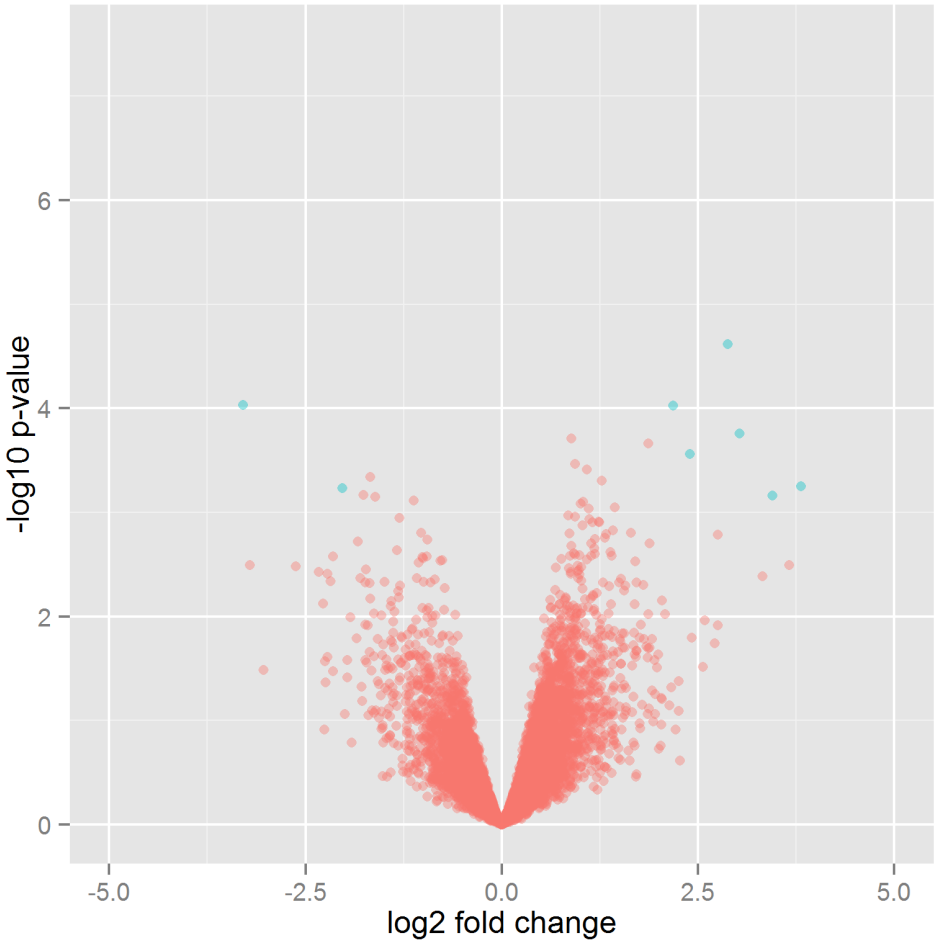
	A	B	C	D	E	F	G	H
1	probe.id	logFC	AveExpr	t	P.Value	adj.P.Val	B	gene.symbols
2	233772_at	2.87956250199663	4.14549683007006	4.7715174248204	2.40557270428784e-05	0.238753090900568	1.65013267439631	POU6F2-AS2
3	228462_at	-3.29143774728401	7.35213797306384	-4.34274556049547	9.22708826867022e-05	0.312936739322245	0.689572680338478	IRX2
4	222925_at	2.18036721165253	3.53184782688095	4.33472696056914	9.45904501729707e-05	0.312936739322245	0.671744678509327	DCDC2
5	220196_at	3.02967263270146	3.52456125889929	4.13834005303945	0.000173018196240964	0.359134322183899	0.23732170195157	MUC16
6	231849_at	2.397891828821	5.81223081037307	3.98720356608366	0.000273711262458122	0.388083468556694	-0.0935984612358238	KRT80
7	1559258_a_at	3.81212341622854	5.04321025499771	3.74697648525638	0.00056049332555541	0.438595391920763	-0.611899182987393	CXorf61
8	203440_at	-2.03219011496932	4.8294198255878	-3.73484865879168	0.000580892807294652	0.438595391920763	-0.637778119435682	CDH2
9	206884_s_at	3.44779141520984	6.79399026126914	3.67688881420692	0.000688694727225568	0.438595391920763	-0.761035032340025	SCEL
10	223631_s_at	2.75287895728847	8.6211151946868	3.37710355512515	0.00163250183889358	0.462930878600537	-1.38617466867452	C19orf33
11	229002_at	-2.149788124793	5.37692678984727	-3.20572132340467	0.00263558640602756	0.485199135355979	-1.73281704741856	FAM69B
12	204351_at	3.65818379721726	9.34325681449894	3.13739991169843	0.00317993331641945	0.494716312125519	-1.86852599384248	S100P
13	214023_x_at	-3.2060458454769	6.39672704679781	-3.13623121683679	0.00319011022428547	0.494716312125519	-1.87083442783914	TUBB2B
14	218625_at	-2.6264318291918	5.73585090255441	-3.12319549368069	0.00330573050978733	0.497111747115746	-1.8965527637219	NRN1
15	230493_at	-2.32777133179933	7.83594400418726	-3.07737295535748	0.00374434015039248	0.506088120525464	-1.98651192135248	SHISA2
16	241436_at	-2.21898269238495	6.4878516335875	-3.06408386856721	0.00388146611930126	0.512679519637744	-2.01246962640596	SCNN1G
17	223861_at	3.32014991928234	6.07930170349058	3.04345850345879	0.00410369702972465	0.512679519637744	-2.05263824452169	HORMAD1
18	229084_at	-2.17694642762121	4.4977327607521	-3.00259848559657	0.00457975790812842	0.512679519637744	-2.13178003750337	CNTN4
19	203726_s_at	2.04378517644524	7.23316377674301	2.84437529623471	0.00695665998441808	0.615934558970307	-2.43254548675102	LAMA3
20	226612_at	-2.27351605967364	5.25605588675778	-2.81365946875742	0.00753479101242168	0.635831788177493	-2.489835083188	UBE2QL1
21	212094_at	2.08311855973128	6.50479193006167	2.7246144201011	0.00947290128662073	0.635831788177493	-2.65379181620987	PEG10
22	219508_at	2.58956392125384	4.99570795346243	2.67117320116111	0.0108478671447989	0.656494398854446	-2.75062591296665	GCNT3
23	209277_at	2.75656134847765	5.89758754503314	2.62503493009675	0.0121805466791149	0.68261703902429	-2.83325050391467	TFPI2
24	206125_s_at	2.42458378913988	4.55522132182725	2.51602677807035	0.0159474179870381	0.68261703902429	-3.02473736790248	KLK8
25	210397_at	2.714317277965	4.97404226469861	2.46659375344573	0.0179833660885152	0.700823978371006	-3.10978856115628	DEFB1

Pav. 8: "Csv" bylos pavyzdys ACs I ir ACs II tipų genų raiškos lyginimo rezultatas.

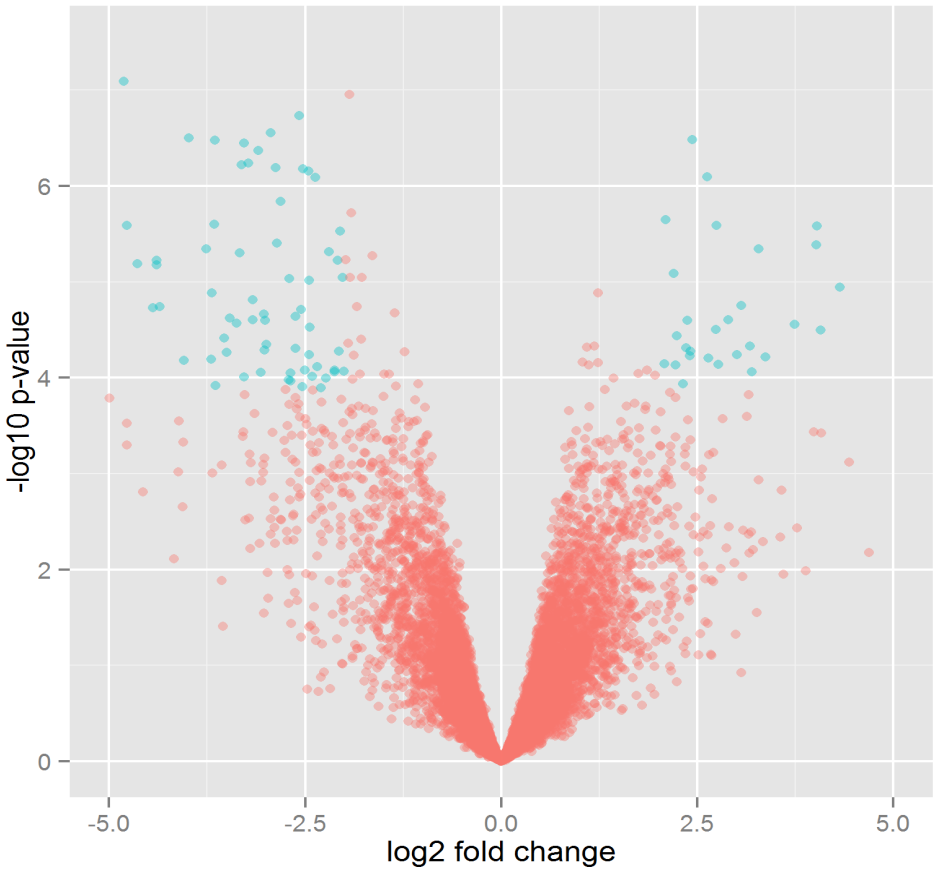
Matoma, kas yra saugoma, dvejų plaučių vėžio tipų genų raiškos lyginimo, tekstiniame rezultatų faile. Tai yra įvairūs statistiniai rodikliai. Genai byloje surikiuoti P-reikšmės didėjimo tvarka. Tai reiškia, kad reikšmingiausi raiškos skirtumai, tarp atitinkamų vėžio tipų, yra tų genų, kurie yra aukščiau šiame faile. 8 pav. matome, kad aukščiausiai yra genas „POU6F-AS2". Tai reiškia, kad tarp plaučių vėžio tipų ACs I ir ACs II šio geno raiška, programos pagalba, nustatyta kaip reikšmingiausiai skirtinga. Žinoma, šis genas gali būti ir niekuo neypatingas, galbūt net nesusijęs su plaučių vėžiu, bet tuo įsitikinti galima tik atlikus papildomus tyrimus tarp šių dviejų plaučių vėžio tipų stebint būtent šio geno raišką.

3.4. Reikšmingai skirtingos raiškos genai pateikti grafiniu formatu

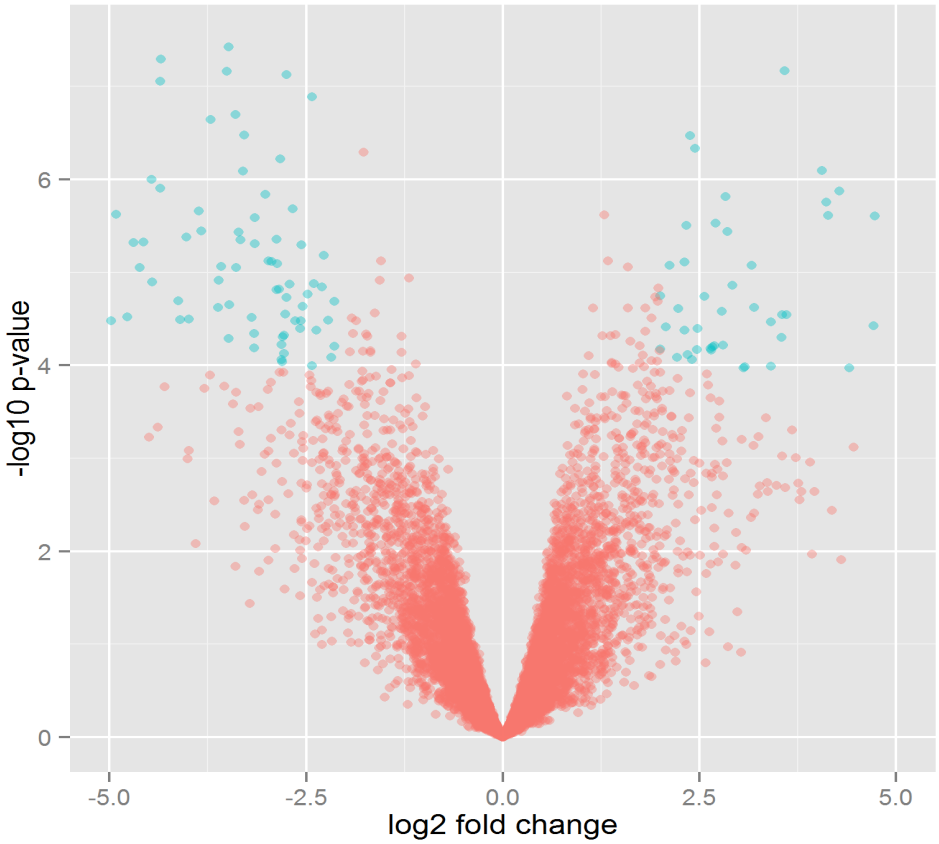
Kitos rezultatų bylos yra grafinės ir pateikiamos „png" formatu. Jose vaizduojami vulkano tipo grafikai, kuriuose puikiai matomi reikšmingai skirtingos raiškos genai. Mėlyni taškai rodo būtent tokius genus. Žemiau pateikiami visų genų raiškos lyginimų, tarp plaučių vėžio tipų, vulkano tipo grafikai, kuriuose melsvai pažymėti reikšmingai skirtingos raiškos genai (9 – 14 pav.).



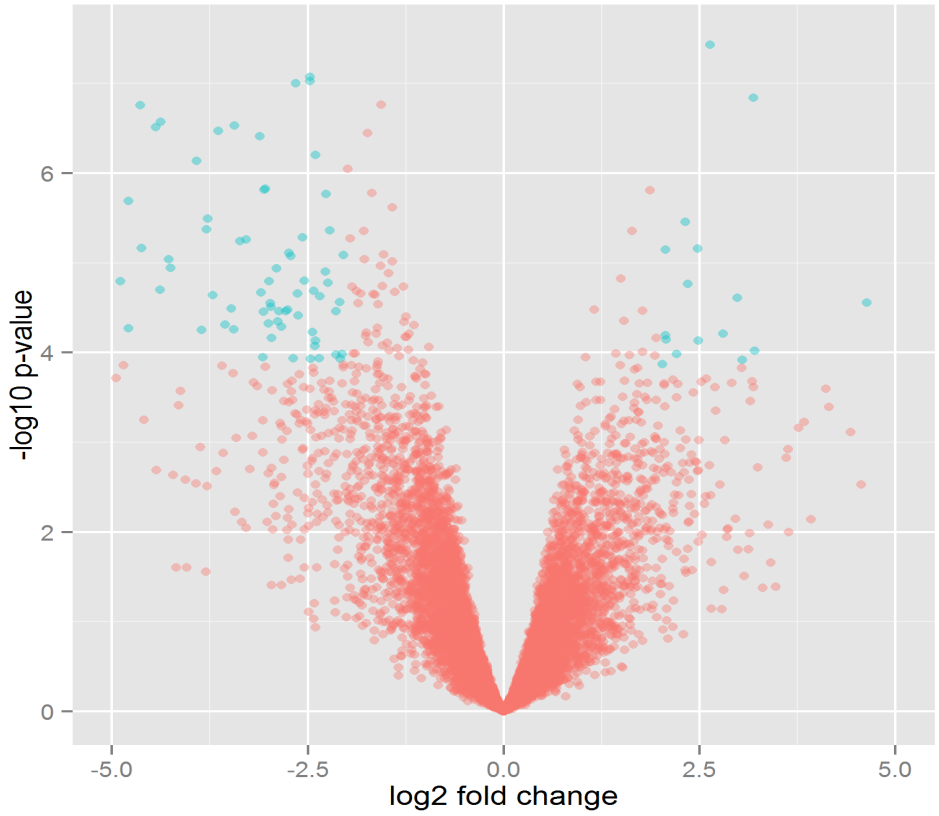
Pav. 9: Reikšmingai skirtingos raiškos genų, tarp ACs I ir ACs II tipų, vulkano grafikas.



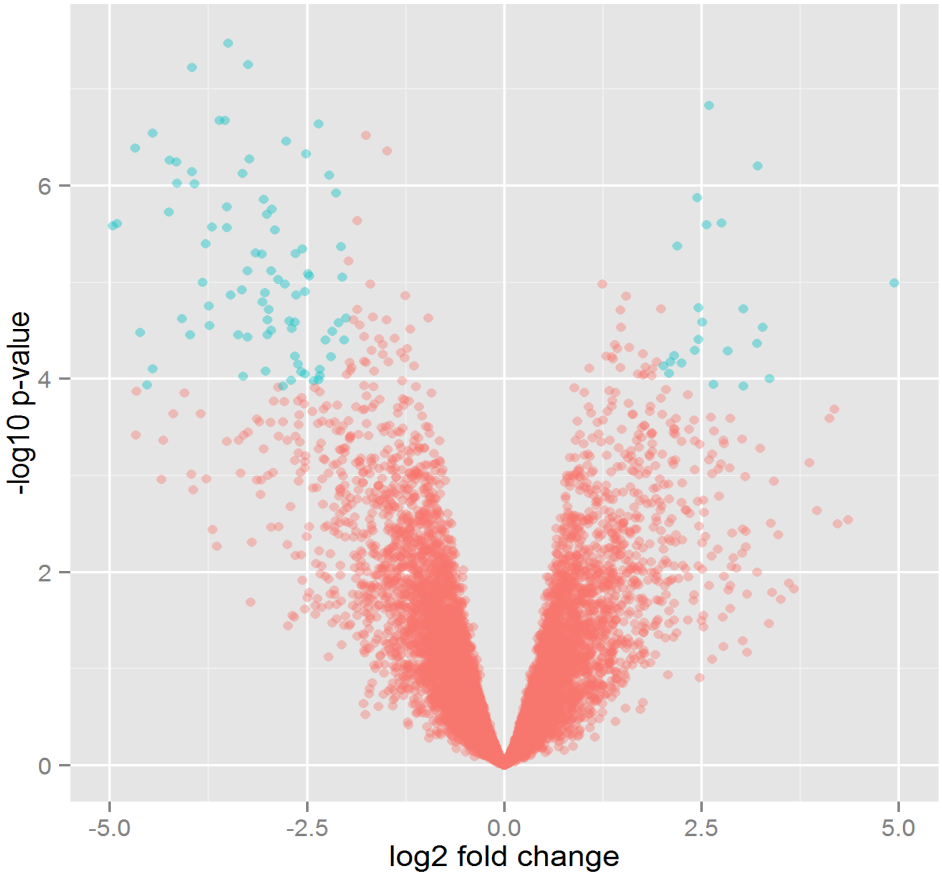
Pav. 10: Reikšmingai skirtingos raiškos genų, tarp ACs I ir SCCs I tipų, vulkano grafikas.



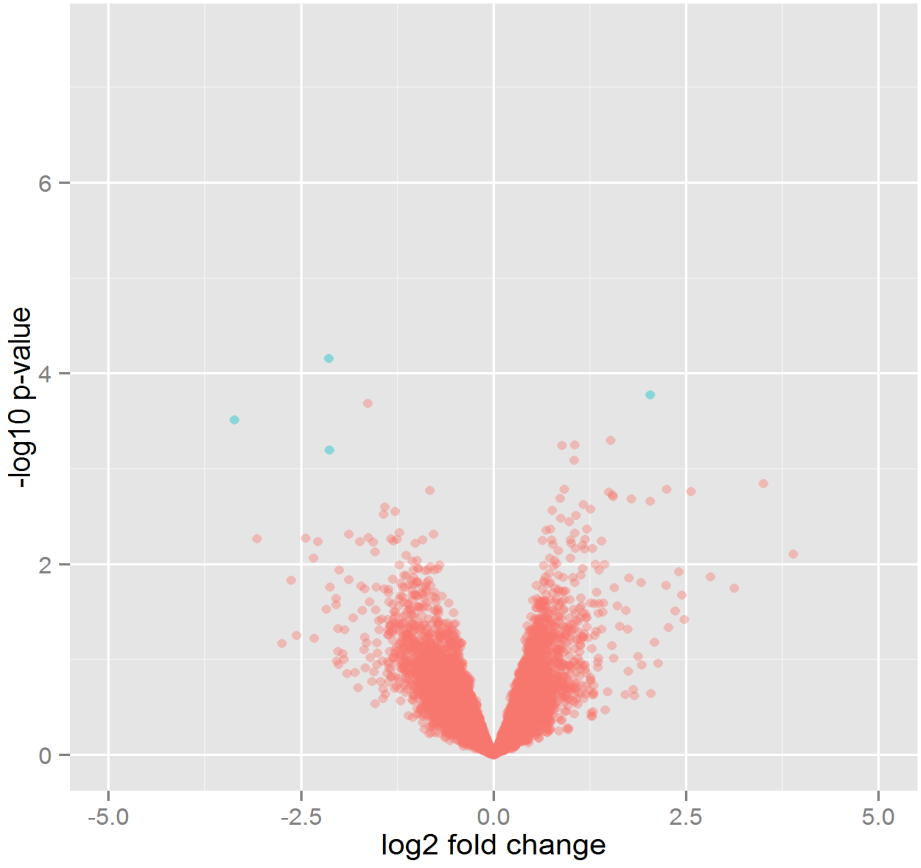
Pav. 11: Reikšmingai skirtingos raiškos genų, tarp ACs I ir SCCs II tipų, vulkano grafikas.



Pav. 12: Reikšmingai skirtingos raiškos genų, tarp ACs II ir SCCs I tipų, vulkano grafikas.



Pav. 13: Reikšmingai skirtingos raiškos genų, tarp ACs II ir SCCs I tipų, vulkano grafikas.



Pav. 14: Reikšmingai skirtingos raiškos genų, tarp SCCs I ir SCCs II tipų, vulkano grafikas.

Iš vulkano tipo grafikų taip pat galime pastebėti, kad daug daugiau reikšmingai skirtingos raiškos genų yra tarp AC ir SCC tipų nei tarp AC ir AC arba SCC ir SCC tipų. Tai rodo, kad reikšmingai skirtingos raiškos genų lyginant, to paties plaučių vėžio tipo, bet skirtingų vėžio stadijų pavyzdžius, yra randama gerokai mažiau, negu lyginant pavyzdžius tarp skirtingų plaučių vėžio tipų.

3.5. Unikalūs reikšmingai skirtingos raiškos genai

Šiame darbe analizuojami 4 plaučių vėžio tipai, ir pateikiami genai, kurių raiška reikšmingai skiriasi tarp tų tipų. Tačiau tolimesniems tyrimams, pavyzdžiui, norint klasifikuoti plaučių vėžio atvejus, naudinga turėti tik unikalius reikšmingai skirtingos raiškos genus kiekvienam lyginimui tarp tipų (2 lentelė).

Lentelė 2. Unikalūs reikšmingai skirtingos raiškos genai, kiekvienam plaučių vėžio tipų lyginimui.

Lyginimas	*Unikalūs reikšmingai skirtingos raiškos genai*					
ACs I su ACs II	UBE2QL1	SFRP1	KLK8	DEFB1	CDH2	
ACs I su SCCs I	TSPAN7	RBP4	OVOL1	IDO1	CSRP2	BTC
	TIAM1	RASD1	NR3C2	HPGD	CRYAB	BCAS1
	TFPI	RAB3B	NDUFA4L2	HMGB3	CYP4B1	ARSD
	TCN1	PTPN22	MICALL1	HEY1	CHST2	AQP4
	ST315	PRKX	LPIN2	GSTM3	CHL1	APOC1
	SEMA6A	PLCH1	LONRF2	GPRC5B	CEL	ACOX2
	SDK2	PDK4	KIF13B	GOLM1	CCDC68	
	SCGB1A1	PAX6	INPP4B	DNAJC15	C8orf4	
ACs I su SCCs II	ZNF818P	RASSF10	LRRC15	HNF4G	FA2H	CELSR2
	USP31	RASGEF1A	LMO7	GPR161	ERC2	CDH13
	TWIST1	PRRX1	LLGL2	GLI3	ERBB3	CACNA2D1
	TRPM8	PRKAA2	LIPH	GDPD1	ELF3	ATP6V0A4
	THSD7A	PLA216	LGR5	GCNT2	EFNA5	ATP11A
	STEAP4	OXR1	KRT8	GAS1	DUSP7	ACOT4
	SPOCK1	NOV	KRT18	FUT2	DLL1	
	SYNE4	MPPED2	KCNN4	FOXC2	CST6	
	SFRP2	MYO5C	IL8	FMO1	COL3A1	
	SCGB3A1	MISP	IGF2BP3	FAM57A	COL27A1	
	RBP1	MFI2	HOXC10	FAM43A	COBL	

ACs II su SCCs I	TINCR	RNASE4	PCGF3	HLA-DRB4	CTSV	ATP2A3
	SNTB1	RHOV	NINJ2	HIST3H2A	CBR1	AIM1L
	SLC351	RAPGEFL1	LPAR5	GRHL1	CATSPERB	ADIRF
	SYT1	POF1B	LOXL4	GPR133	C1M3	ACSM3
	SCNN1B	PKP2	IRX5	FAM149A	BDNF	
	SBK1	PHLDB2	HSD17B1	ELOVL6	ATP8A1	
ACs II su SCCs II	ZNF320	RNF183	METTL8	HAS2	DDC	C11orf82
	TMEM170B	RND3	MAD2L1	GABRB2	CTTN	BPIFA1
	TLE2	RHOU	LRP2	FOXQ1	CDK6	BCHE
	TCTEX1D2	RGS20	LAMB3	FOXF2	CDH3	ARL4D
	TBL1XR1	RGN	KLK7	FOSL1	CAV2	APOBEC3A
	SPAG4	RAP2B	KLK12	FDCSP	CAV1	AHNAK2
	SHANK2	PRNP	KLF5	ELF5	C4orf19	AFAP1L2
	SEPT10	PPP1R3C	KYNU	EIF5A2	C4BPB	
	SELE	POPDC3	YPEL1	ECHDC3	C3orf67	
	SCIN	PLEKHS1	ITGB8	DPP4	C16orf74	
SCCs I su SCCs II	PCOLCE2	HLA-DQA1	GLDC	GABRP	DNAJC6	

3 lentelėje pateikiami keli unikalūs reikšmingai skirtingos raiškos genai, kiekvienam genų raiškos lyginimui tarp plaučių vėžio tipų. Taip pat pateikiamos „PubMed“ duomenų bazėje rastos atitinkamo geno sąsajos su plaučių vėžiu.

Lentelė 3. Keleto unikalių reikšmingai skirtingos raiškos genų sąsajos su plaučių vėžiu.

Lyginimas	*Genas*	*Geno sąsaja su plaučių vėžiu*
ACs I su ACs II	SFRP1	Gali būti panaudojamas nesmulkialąstelinio plaučių vėžio naviko augimo slopinimui.
	KLK8	Jeigu geno raiška labai padidinama stipriai invazinėse plaučių vėžio ląstelėse, tai tų ląstelių invaziškumas slopinamas.
	CDH2	Susijęs su nesmulkialąstelinio plaučių vėžio metastazavimu į smegenis.
ACs I su SCCs I	TIAM1	Šio geno raiška reikšmingai didesnė plaučių vėžio pavyzdžiuose lyginant su šio geno raiška normaliose plaučių epitelinėse ląstelėse.
	TFPI	Šis genas svarbus plaučių vėžio metastazavimui ir invaziškumui.
	AQP4	Dalyvauja plaučių vėžio invazijos mechanizmuose.
ACs I su SCCs II	TRPM8	Kartu su kitu genu TRPA1 prisideda prie invazinio fenotipo plaučių vėžyje.
	LGR5	Gali būti kaip žymuo nesmulkialąstelinio plaučių vėžio ląstelėms aptikti.
	IL8	Stimuliuoja nesmulkialąstelinių plaučių vėžio ląstelių proliferaciją.
ACs II su SCCs I	RHOV	Dauguma plaučių vėžio atveju baltymo kiekis yra padidėjęs.
	HSD17B1	Nesmulkialąstelinio plaučių vėžio ląstelėse stipriai padidėjęs baltymo kiekis.
	CBR1	Šio geno raiška yra veikiama cigaretėse esančių medžiagų.
ACs II su SCCs II	RND3	Reguliuoja plaučių vėžio ląstelių proliferaciją.
	LAMB3	Šio ir ITGB1 geno išjungimas (knockdown) slopintų vėžinių ląstelių invaziją ir metastazavimą.
	FOXQ1	Gali būti naudojamas kaip nesmulkialąstelinio plaučių vėžio prognostinis veiksnys.
SCCs I su SCCs II	HLA-DQA1	Susijęs su plokščialąstelinio plaučių vėžio susirgimo rizika.
	GLDC	Šio geno raiška reikšmingai susijusi su nesmulkialąsteline plaučių vėžio forma sergančių individu išgyvenimo rodikliu.

Išvados

1. R ir Bioconductor yra tinkami įrankiai visuminių tyrimų duomenų išankstiniam apdorojimui.
2. Bioconductor yra tinkamas įrankis ieškant reikšmingai skirtingos raiškos genų tarp skirtingų plaučių vėžio tipų
3. Vulkano grafikai tinka atvaizduoti reikšmingai skirtingos raiškos genus.
4. Lyginant tos pačios plaučių vėžio rūšies tačiau skirtingų vėžio stadijų pavyzdžius yra randama gerokai mažiau reikšmingai skirtingos raiškos genų, negu lyginant pavyzdžius tarp skirtingų plaučių vėžio rūšių.
5. Lyginant ACs I ir ACs II plaučių vėžio tipus randama unikalių reikšmingai skirtingų genų, kurių nerandama lyginant ACs I su SCCs I arba su SCCs II (kitais atvejais taip pat).
6. Unikalūs reikšmingai skirtingos raiškos genai yra siejami su vėžinių ląstelių augimo reguliavimu, invazyvumu ir metastazavimu.
7. Analizuojant ACs I ir SCCs II plaučių vėžio tipus gautas, vienas iš reikšmingai skirtingos raiškos genų, LGR5 genas gali būti naudojamas kaip nesmulkialąstelinio plaučių vėžio žymuo.
8. Analizuojant ACs II ir SCCs II plaučių vėžio tipus gautas, vienas iš reikšmingai skirtingos raiškos genų, FOXQ1 genas gali būti naudojamas kaip nesmulkialąstelinio plaučių vėžio prognostinis veiksnys.
9. Analizuojant ACs I ir SCCs I plaučių vėžio tipus gautas, vienas iš reikšmingai skirtingos raiškos genų, TIAM1 genas, kurio raiška plaučių vėžio pavyzdžiuose reikšmingai didesnė, lyginant su jo raiška normaliose plaučių epitelinėse ląstelėse.
10. Unikalius skirtingos raiškos genus tarp plaučių vėžio tipų, galima panaudoti molekulinių plaučių vėžio atvejų klasifikatorių kūrimui.
11. Sukurta programa gali būti pritaikoma kitokios rūšies vėžio reikšmingai skirtingos raiškos genų paieškai redaguojant dizaino ir kontrasto matricas.

Santrauka

Plaučių vėžys pasaulyje yra antra dažniausiai tiek moterims tiek vyrams diagnozuojama vėžio forma. Dauguma plaučių vėžio simptomų išryškėja ligai pasiekus vėlyvas stadijas. Siekiant, kad plaučių vėžiu sergančių pacientų gydymas būtų kuo veiksmingesnis būtina sukurti metodus greitai progresuojančių plaučių vėžio atvejų identifikavimui (plaučių vėžio atvejų klasifikavimo metodai) dar iki progresavimo. Vienas iš tokių metodų kūrimo kelių – visuminių tyrimų duomenų analizė. Pagrindinis darbo tikslas yra R ir Bioconductor pagalba sukurti programinę įrangą plaučių vėžio visuminių tyrimų duomenų analizei.

Darbe pristatoma R programavimo aplinka ir Bioconductor taip pat atrinktos jo bibliotekos, kurios naudojamos visuminėms analizėms. Trumpai aprašomi plaučių vėžio tipai. Bioconductor bibliotekų pagalba sukurta programa, kuri atlieka reikšmingai skirtingos raiškos genų, tarp dviejų plaučių vėžio tipų, paieška. Iš gautų rezultatų galima teigti, kad R yra tinkamas įrankis visuminių tyrimų duomenų analizėms atlikti. Sukurta programa tekstiniu ir grafiniu formatu pateikia reikšmingai skirtingos raiškos genus tarp skirtingų plaučių vėžio tipų. Rasti unikalūs skirtingos raiškos genai tarp plaučių vėžio tipų, gali būti panaudoti molekulinių plaučių vėžio atvejų klasifikatorių kūrimui. Taip pat programa gali būti lengvai pritaikoma kitos rūšies vėžio reikšmingai skirtingos raiškos genų paieškai. Programa perrašyta ir R markdawn formatu, kad būtų galima lengvai konvertuoti tyrimo rezultatus į „pdf“ formato failą.

Mantas Bernotas

Experimental Data Processing in Molecular Biology Using R

Summary

Lung cancer is the second most common diagnosed cancer in the world for both women and men. Most of lung cancer symptoms appear when disease reaches an advanced stage. In order for lung cancer patient's treatment be as effective as possible it is necessary to develop methods of rapidly progressive lung cancer identification before progression. The main goal of this work is to create software for lung cancer high-throughput data analysis by using R and Bioconductor. In this paper, R programming environment and Bioconductor with libraries used for high-throughput analysis are briefly presented. Types of lung cancer are also briefly described. Bioconductor libraries are used for searching differential expressed genes between two types of lung cancer. The results suggest that R is suitable tool for high-throughput data analysis. The program provides differential expressed genes among different types of lung cancer in a textual and in a graphical format. Discovered unique differential expressed genes between lung cancer types may be used for molecular classificator of different lung cancer cases development. The program can be easily adapted to search differential expressed genes in other types of cancer. The program is rewritten with R markdawn, for making it easier to convert research results into pdf file format.

Literatūra

1. Fox, John and Andersen, Robert (January 2005). "Using the R Statistical Computing Environment to Teach Social Statistics Courses" (PDF). Department of Sociology, McMaster University. Retrieved 2006-08-03.
2. Tippmann, Sylvia (29 December 2014). "Programming tools: Adventures with R". Nature (517): 109–110.
3. Gentleman, R. (2008). R Programming for Bioinformatics. Chapman & Hall/CRC.
4. http://www.bioconductor.org
5. C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science USA, 98:31 36, 2001.
6. Affymetrix. Affymetrix Microarray Suite User Guide. Affymetrix, Santa Clara, CA, version 4 edition, 1999.
7. Affymetrix. Affymetrix Microarray Suite User Guide. Affymetrix, Santa Clara, CA, version 5 edition, 2001.
8. Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003.
9. Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research Vol 43 (accepted 6 January 2015).
10. Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology Volume 3, Issue 1, Article 3.
11. Kirk J Mantione, Richard M. Kream, Hana Kuzelova, Radek Ptacek, Jiri Raboch, Joshua M. Samuel, and George B. Stefano. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. Med Sci Monit Basic Res. 2014; 20: 138–141.
12. http://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html
13. Ramaswamy Govindan, Nathan Page, Daniel Morgensztern, William Read, Ryan Tierney, Anna Vlahiotis, Edward L. Spitznagel and Jay Piccirillo. Changing Epidemiology of Small-Cell Lung Cancer in the United States Over the Last 30 Years: Analysis of the Surveillance, Epidemiologic, and End Results Database. Accepted May 16, 2006.
14. Mia A. Levy, Christine M. Lovly, and William Pao. Translating genomic information into clinical medicine: Lung cancer as a paradigm. 2012.

15. http://www.oecd-ilibrary.org/sites/9789264183896-en/01/05/index.html?itemId=/content/chapter/9789264183896-8-en.
16. Chien-Chou Pan, Pei-Tseng Kung, Yueh-Hsin Wang, Yu-Chia Chang, Shih-Ting Wang, Wen-Chen Tsai. Effects of Multidisciplinary Team Care on the Survival of Patients with Different Stages of Non-Small Cell Lung Cancer: A National Cohort Study. Published: May 12, 2015.
17. American Cancer Society. Cancer Facts & Figures 2015. Atlanta: American Cancer Society; 2015.
18. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002;11:108–36.
19. Affymetrix. Statistical algorithms description document. Affymetrix, Inc.; 2002.
20. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2:418–27. [PubMed:11389458]
21. CL Wilson, SD Pepper, Y Hey, and CJ Miller. Amplification protocols introduce systematic but reproducible errors into gene expression studies. Biotechniques, 36:498–506, 2004.
22. http://genecards.weizmann.ac.il/geneannot/index.shtml.
23. Belinda Phipson, Stanley Lee, Ian J. Majewski, Warren S. Alexander, Gordon K. Smyth. Empirical Bayes in the presence of exceptional cases, with application to microarray data. 2013.
24. Storey, J.D. and R. Tibshirani, Statistical significance for genomewide studies. Proc Natl Acad Sci U S A, 2003. 100 (16): p. 9440-5.