VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
MATEMATINĖS INFORMATIKOS KATEDRA

Epigenetinių ligų profilių atvaizdavimas ir palyginimas

Visualization and Comparison of Epigenetic Disease Profiles

Bakalauro baigiamasis darbas

Atliko: 4 kurso Bioinformatikos studentas
Artūras Tarasenka
Darbo vadovas: Karolis Koncevičius

Vilnius – 2020

TURINYS

ĮVADAS

Atliekant plačios žmogaus ligų ir bruožų įvairovės tyrimus bei dirbant su dideliais daugiamačių biomedicinos duomenų kiekiais, efektyvus sprendimas yra tirti kelias, nors iš pirmo žvilgsnio niekuo nesusijusias, ligas ir lyginti jas ieškant panašumų pagal skirtingus požymius, siekiant nustatyti dėsningumą pagal kurį galima būtų atlikti išvadas apie visas lyginamas ligas. Epigenetiniams ligos profilių tyrimams yra tinkami daugelis daugiamačių biomedicinos duomenų atvaizdavimui skirtų metodų, nuo paparstų iki sudėtingesnių, tačiau ne visi iš jų yra vienodai efektyvus.

Šio darbo tikslas - sukurti R programavimo kalbos funkcijas pritaikančias įvairius daugiamačių biomedicinos duomenų atvaizdavimo metodus epigenetiniams ligų tyrimams ir atvaizduoti bei palyginti DNR metilinimo pokyčių profilius įvairiose žmogaus ligose.

Šiam tikslui pasiekti darbe iškelti šie uždaviniai:

1. Parsisiųsti bei tyrimui paruošti ,,EWAS Atlas” DNR metilinimo rezultatų žmogaus ligų tyrimuose duomenų bazę.
2. Naudojant R programavimo kalbą sukurti įrankį galintį atvaizduoti bei palyginti skirtingų ligų DNR metilinimo pokyčių profilius ir paruošti palyginimo ataskaitą.
3. Sugalvoti atstumo matą, galintį įvertinti panašumą tarp dviejų skirtingų DNR metilinimo tyrimo rezultatų.
4. Sugrupuoti ligų tyrimų rezultatus esančius "EWAS Atlas" duomenų bazėje.
5. Apibendrinti bei interpretuoti tyrimo rezultatus, pateikti išvadas.

1. EPIGENETIKA

Šiame skyriuje aprašomas epigenetikos mokslas bei jam būdingi ligų tyrimai ir šiame darbe naudota epigenetinių duomenų bazė „EWAS Atlas“.

1.1 Epigenetikos mokslas

Beveik šimtmetį nuo termino „epigenetika“ pirmojo pasirodymo spausdintame lape tyrėjai bei medikai yra pasinėrę į genų tyrinėjimus, bandydami išnarplioti užuominas, kad geno funkciją gali pakeisti ne tik pokyčiai sekoje. Šiandien daugybė įvairių ligų, elgesio ir kitų sveikatos rodiklių turi tam tikro lygio įrodymų, siejančių juos su epigenetiniais mechanizmais, įskaitant beveik visų tipų vėžį, kognityvinę disfunkciją ir kvėpavimo, širdies ir kraujagyslių, reprodukcines, autoimunines ir neuro elgesio ligas. Tarp žinomų ar įtariamų veiksnių įtakojančių epigenetinius procesus yra sunkieji metalai, pesticidai, dyzelino išmetamosios dujos, tabako dūmai, policikliniai aromatiniai angliavandeniliai, hormonai, radioaktyvumas, virusai, bakterijos ir pagrindinės maistinės medžiagos.

Pastaruoju metu keli novatoriški tyrimai naujai atkreipė dėmesį į epigenetiką. Susidomėjimas pradėjo didėti, kai tapo akivaizdu, kad epigenetikos ir epigenomikos - epigenetinių pokyčių pasiskirstymo genomo mastu - supratimas bus būtinas darbuose, susijusiuose su daugeliu kitų temų, reikalaujančių išsamaus supratimo apie visus genetikos aspektus, tokius kaip kamieninės ląstelės, klonavimas, senėjimas, sintetinė biologija, rūšių išsaugojimas, evoliucija ir žemės ūkis.

Žodis „epigenetinis“ pažodžiui reiškia „papildomai prie genetinės sekos pokyčių“. Terminas išsivystė taip, kad apima bet kurį procesą, keičiantį genų aktyvumą nekeičiant DNR sekos ir sukeliantį modifikacijas, kurios gali būti perduodamos į dukterines ląsteles (nors eksperimentai rodo, kad kai kurie epigenetiniai pokyčiai gali būti grįžtami). Greičiausiai ir toliau bus diskutuojama, ką tiksliai reiškia ši sąvoka ir ką ji apima.

Nustatyta daugybė epigenetinių procesų tipų - tarp jų metilinimas, acetilinimas, fosforilinimas. Epigenetiniai procesai yra natūralūs ir būtini daugeliui organizmo funkcijų, tačiau vykstant netinkamai, jie gali turėti didelį neigiamą poveikį sveikatai ir fenotipui.

Ko gero, geriausiai žinomas epigenetinis procesas, iš dalies todėl, kad buvo lengviausias tirti naudojant esamas technologijas, yra DNR metilinimas. Tai yra metilo grupės (CH3) pridėjimas arba šalinimas, vykstantis daugiausia kai DNR sekoje iš eilės susideda kelios citozino bazės. 1983 m. pirmą kartą patvirtinta, kad DNR metilinimas vyksta žmogaus vėžyje ir nuo to laiko jis buvo

tiriamas daugelyje kitų ligų ir sveikatos būklių.

Kitas reikšmingas epigenetinis procesas yra chromatino modifikacija. Chromatinas yra baltymų (histonų) ir DNR kompleksas, tvirtai surištas, kad tilptų į branduolį. Kompleksą gali modifikuoti tokios medžiagos kaip acetilgrupės (procesas vadinamas acetilinimu), fermentai ir kai kurios RNR formos, tokios kaip mikroRNR ir mažos RNR. Ši modifikacija keičia chromatino struktūrą ir daro įtaką genų ekspresijai. Paprastai sandariai sulankstytas chromatinas yra linkęs būti neaktyviu, tuo tarpu atviresnis chromatinas veiklus.

Vienas iš tokių procesų padarinių yra imprintingas. Genetikoje jis apibūdina būklę, kai vienas iš dviejų tipiškų genų porų alelių yra nutildytas epigenetinių procesų, tokių kaip metilinimas arba acetilinimas. Tai tampa problema, jei išreikštas alelis yra pažeistas arba jame yra variantas, padidinantis organizmo jautrumą mikrobams, toksiškoms medžiagoms ar kitoms kenksmingoms medžiagoms. Pirmą kartą imprintingas buvo nustatytas 1910 m. kukurūzuose, o žinduoliams pirmą kartą patvirtintas 1991 m. [Wei06]

1.2 Epigenetiniai ligų tyrimai

Pirmoji žmonių liga susieta su epigenetika buvo vėžys. 1983 m. tyrėjai nustatė, kad pacientams, sergantiems gaubtinės ir tiesiosios žarnos vėžiu, pažeistame audinyje globalus DNR metilinimas buvo mažesnis nei normaliame tų pačių pacientų audinyje [FVo83]. Kadangi metilinti genai paprastai yra išjungiami, DNR metilinimo praradimas gali sukelti neįprastai didelį genų aktyvavimą, pakeisdamas chromatino išdėstymą. Kita vertus, per didelis metilinimas gali panaikinti apsauginių navikų slopintuvų genų darbą.

Trapios X chromosomos sindromas yra dažniausiai paveldima, ypač vyrams, psichinė negalia. Ši liga gali paveikti abi lytis, tačiau kadangi vyrai turi tik vieną X chromosomą, liga juos paveiks stipriau: juo serga maždaug 1 iš 4000 vyrų ir 1 iš 8 000 moterų. Šį sindromą turintys žmonės turi sunkią intelekto negalią, uždelstą žodinį vystymąsi ir „į autizmą panašų“ elgesį. Trapios X chromosomos sindromas nėra vienintelis su protiniu atsilikimu susijęs sutrikimas siejamas su epigenetiniais pokyčiais. Kitos tokios ligos yra „Rubenstein-Taybi”, „Coffin-Lowry”, „Prader-Willi”, „Angelman”, „Beckwith-Wiedemann”, „ATR-X” ir „Rett” sindromai.

Kadangi tiek daug ligų, pavyzdžiui, vėžys, sukelia epigenetinius pokyčius, atrodo pagrįsta pabandyti neutralizuoti šias modifikacijas epigenetiniu gydymu. Šie pokyčiai atrodo idealus

taikinys, nes jie, priešingai nei DNR sekos mutacijos, iš prigimties yra grįžtami. Populiariausiu iš šių gydymo būdų siekiama pakeisti DNR metilinimą arba histono acetiliaciją.

Epigenetinę terapiją naudoti reikia atsargiai, nes epigenetiniai procesai ir pokyčiai yra labai platūs. Kad epigenetinis gydymas būtų sėkmingas, jis turi būti selektyvus nenormalioms ląstelėms, priešingu atveju, suaktyvinus genų transkripciją normaliose ląstelėse, jos gali virsti vėžinėmis, todėl gydymas gali sukelti tuos pačius sutrikimus, kuriuos jie bando neutralizuoti. Nepaisant šio galimo trūkumo, tyrėjai randa būdų, kaip nukreipti gydymą į nenormalias ląsteles kuo mažiau pažeidžiant normalias ląsteles, o epigenetinė terapija pradeda atrodyti vis perspektyvesnė. [Sim08]

Vykdant epigenetinius ligų tyrimus reikia lyginti DNR metilinimą tarp sveikų ir sergančių žmonių grupių. Tam atlikti reikalingi metodai, gebantys nustatyti DNR metilinimo pokyčius genomo mąstu.

1.3 DNR metilinimo mikrogardelės

Ligų atveju DNR metilinimo skirtumams nustatyti dažniausiai naudojamos metilinimo mikrogardelės.

Pagrindinis mikrogardelių principas yra hibridizacija tarp dviejų DNR grandžių, komplementarių nukleorūgščių sekų savybė specifiškai poruotis viena su kita, sudarant vandenilinius ryšius tarp komplementarių nukleotidų bazių porų. Didelis komplementarių bazių porų skaičius nukleotidų sekoje reiškia tvirtesnį nekovalentinį ryšį tarp dviejų grandinių. Nuplovus nespecifines sekas, tik stipriai suporuotos grandinės lieka hibridizuotos. Fluorescenciškai pažymėtos tikslinės sekos, kurios jungiasi su zondo seka, sukuria signalą, kuris priklauso nuo hibridizacijos sąlygų (pavyzdžiui, temperatūros) ir plovimo po hibridizacijos. Bendras signalo stiprumas, skleidžiamas taško (požymio), priklauso nuo to, kiek tikslinio mėginio jungiasi su toje vietoje esančiais zondais. Mikrogardelėse naudojama santykinė kiekybinė analizė, kurios metu požymio intensyvumas lyginamas su to paties požymio intensyvumu esant kitoms sąlygoms, o požymio tapatumas žinomas pagal jo padėtį.

DNR metilinimo mikrogardelės - DNR mikrogardelių tipas, pritaikytas metilinimo skirtumų tyrimams. Naudojant tokio tipo mikrogardeles DNR gauta iš ląstelės apdorojama bisulfitine konversija. Bisulfito konversija metilintus citozinus palieka kaip citozinus, o nemetilintus paverčia i uracilą. Tokiu būdu mikrogardelėje lyginant intensyvumą tarp dviejų zondų - vieno prie kurio

hibridizuojasi nekonvertuotos sekos (su citozinais) ir kito, prie kurio jungiasi to pačio fragmento konvertuotos sekos (su uracilu) gaunamas metilinimo procentas.

Tyrėjai, savo tyrimuose naudojantys DNR metilinimo mikrogardeles, įprastai rezultatais pasidalina paskelbdami citozinų, kuriuose metilinimo lygis reikšmingai skyrėsi tarp sveikų ir tiriamą sutrikimą turinčių individų, pozicijomis, jų efektų dydžiais bei p-vertemis.

1.4 EWAS Atlas duomenų bazė

EWAS (Epigenome-Wide Association Study) - pastaruoju metu tapo veiksminga strategija tiriant sudėtingų bruožų ir ligų epigenetinius pagrindus. Naujausi sekvenavimo bei DNR metilinimo mikrogardelių technologijų pasiekimai leido pradėti plataus masto su žmonių ligomis susijusios epigenetinės variacijos tyrimus. Daugėjant tyrimų atsirado poreikis kaupti bei palyginti skirtingų tyrėjų rezultatus ir tokiu tikslu buvo sukurtas „EWAS Atlas".

„EWAS Atlas" yra bendro formato ir laisvai prieinama DNR metilinimo rezultatų duomenų bazė, kurioje kaupiami ligų DNR metilinimo tyrimų rezultatai iš viso pasaulio. Tai vienas pagrindinių Kinijos nacionalinio genomikos duomenų centro šaltinių, skirtas pateikti išsamų aukštos kokybės EWAS asociacijų rinkinį, kuris padėtų sistemingai tirti epigenetinius molekulinius mechanizmus, susijusius su skirtingais biologiniais požymiais. Ši duomenų bazė yra joje kaupiami rezultatai surinkti iš pasaulyje vykdomų EWAS tyrimų. Kadangi duomenų bazė teikia atvirą prieigą prie visų kuruojamų duomenų, tokiu būdu ji yra naudingas šaltinis pasaulinei tyrimų bendruomenei, nuo 2009 metų sukaupęs rezultatus iš 718 skirtingų tyrimų.

Norėdami pateikti aukštos kokybės informaciją, kuruojamą iš EWAS leidinių, EWAS sukūrė standartizuotą kuravimo procesą, apimantį tris pagrindinius etapus - literatūros paiešką, tyrimų ir asociacijų kuravimą. Pirmiausia atliekama literatūros paieška PubMed naudojant iš anksto apibrėžtus raktinius žodžius. Leidiniai įtraukiami į „EWAS Atlas" tik tuo atveju kai juose randamas būtinas tirtų bruožų ir svarbių EWAS asociacijų aprašymas. Antrasis žingsnis yra tyrimo kuravimas - rankinis išsamios tyrimų informacijos, gautos iš publikacijų, kuravimas, įskaitant bruožo pavadinimą (-us), trumpas tiriamų ir kontrolinių grupių aprašymas, klinikinės ir patologinės tyrimo populiacijų savybės. Siekiant suvienodinti biologinių bruožų vaizdavimą, subjektai priskiriami standartizuotiems terminams Eksperimentinėje Faktorių Ontologijoje, apjungiančioje kelių biologinių ontologijų dalis, tokias kaip Ligų Ontologija ir Genų Ontologija . Galiausiai atliekamas asociacijų kuravimas siekiant rankiniu būdu rinkti informaciją apie tinkamas asociacijas (citozinų

pozicijas kurių metilinimo skirtumų p-vertė tyrime buvo mažesnė nei 0,0001 arba pakoreguota dėl daugkartinio testavimo p-vertė mažesnė nei 0,05), įskaitant koreliacijas tarp DNR metilinimo lygių ir eksperimentinių kintamųjų, taip pat jų rangus konkrečiuose tyrimuose. Be to, atsižvelgiant į įvairias anotavimo sistemas, priimtas skirtingose platformose, visi Illumina kompanijos DNR metilinimo mikrogardelių (27K, 450K ir 850K) zondai yra anotuojami remiantis GENCODE 28 leidimu (GRCh37), kad būtų išlaikytas jų nuoseklumas. Duomenų kuravimą ,,EWAS Atlas" gali pasiekti keli kuratoriai, naudodamiesi patogiomis interneto sąsajomis, leidžiančiomis bendradarbiauti kuruojant ir didinant kuravimo proceso efektyvumą. [LZL+18]

2. DAUGIAMAČIŲ BIOMEDICINOS DUOMENŲ ATVAIZDAVIMO METODAI

Epigenetiniams ligų profiliams palyginti yra tinkami daugelis daugiamačiams duomenims atvaizduot skirtų metodų ir juos įgyvendinančių R programavimo kalbos funkcijų. Pradedant paprastomis bazinėmis funkcijomis ir baigiant daug sudėtingesnę duomenų analizę atliekančiomis. Tačiau ne visos iš jų yra vienodai efektyvios šioje srityje.

2.1 Baziniai atvaizdavimo metodai

Ligų profiliams atvaizduoti galima naudoti ir pačius paprasčiausius metodus. Vienas iš populiariausių duomenų atvaizdavimo būdų yra ,,scatterplot", galintį vizualiai parodyti ryšį tarp dviejų kintamųjų (kuo labiau koncentruoti taškai linijoje, tuo stipresnis santykis tarp jų), ar tarp kintamųjų yra teigiamas arba neigiamas ryšys (ar nuolydis yra teigiamas arba neigiamas), ar duomenų modelis yra tiesinis (tiesus) ar netiesinis (išlenktas) ir ar duomenų rinkiniuose egzistuoja neįprastos ypatybės, pvz., išskirtys, klasteriai ir spragos. [EMS14]

Chromosomų vietų, kuriose kaupiasi rasti skirtumai, vaizdavimui tinka naudoti ,,Manhattan plot" metodas. Tai ,,scatterplot" tipas, plačiai naudojamas genominių asociacijų tyrimuose. X ašyje pavaizduojama chromosomos pozicija, o y ašyje: -log10(p-vertė), parodanti kiekvienos pozicijos metilinimo skirtumų patikimumo lygį.

Norint atvaizduoti su liga susijusių lokusų pasiskirstymą po, pvz., CpG salų regionus galima naudoti Barplot metodą, atvaizdavimui naudojantį stulpelines diagramas. Kiekvienam regionui sudaromas atskiras stulpelis, skirtingi stulpeliai išdėstomi x ašyje, y ašyje rodomas citozinų, kurie

rodė skirtumus tarp lyginamų grupių tyrime, procentas patenkantis į konkretų genominį elementą. Tačiau palyginti šimtus ligų profilių tarpusavyje naudojantis šiais metodais ganėtinai sunku.

2.2 Genų ontologijos analizė

Pirmasis iš sudėtingesnių metodų yra genų ontologijos terminų, susijusių su tyrimo rezultatais, palyginimas.

Genų ontologija yra bioinformatikos iniciatyva, kuria siekiama suvienodinti visų rūšių genų ir genų produktų anotacijas. Šis projektas siekia išlaikyti ir plėtoti savo kontroliuojamą genų ir genų produktų savybių žodyną, komentuoti genus ir genų produktus, įsisavinti ir skleisti anotacijos duomenis ir aprūpinti tyrėjus įrankiais, kuriais galima lengvai prieiti prie visų projekto teikiamų duomenų aspektų ir sudaryti sąlygas funkciniam eksperimentinių duomenų aiškinimui naudojant GO.

Atliekant genų ontologijos analizę randami kiekvienai ligai būdingi genų ontologijos terminai ir pagal juos lyginamos tiriamos ligos, ieškant bendrų bei skirtingų terminų. Tokiu būdu ligos, paveikiančios panašią funkciją atliekančių genų raišką, turėtų turėti panašius genų ontologijos terminus.

2.3 Klasterizavimas

Hierarchinis klasterizavimas, dar vadinamas hierarchinių klasterių analize, yra algoritmas, grupuojantis panašius objektus į grupes, vadinamas klasteriais. Rezultatas yra klasterių rinkinys, kuriame kiekvienas klasteris skiriasi tarpusavyje, o kiekvienos iš jų objektai yra iš esmės panašūs vienas į kitą.

Šis metodas gali būti atliekamas naudojant atstumų matricą. Jis pradedamas traktuojant kiekvieną stebėjimą kaip atskirą klasterį. Tuomet pakartotinai vykdomi du veiksmai: identifikuojamos ir sujungiamos dvi arčiausiai viena kitos esančios grupės. Šis procesas tęsiasi tol, kol visos grupės paskutiniame žingsnyje yra apjungtos į vieną klaterį. Galutinis rezultatas yra dendrograma, kuri rodo hierarchinį ryšį tarp grupių bei atstumus tarp jų. Tokiu būdu grupuojant ligas pagal skirtingus požymius išryškėja panašumai bei skirtumai tarp ligų pagal tiriamus požymius. [Boc18]

2.4 t-SNE

Vienas iš sudėtingesnių duomenų atvaizdavimo ir palyginimo metodų yra t-Distributed Stochastic Neighbour Embedding (t-SNE). Tai netiesinė požymių mažinimo technika, kuri ypač gerai tinka didelės apimties daugiamačių duomenų rinkinių vizualizavimui. Ji plačiai taikoma vaizdo apdorojimo, neurolingvistiniame programavime, genomo duomenų ir kalbos apdorojimo srityse. Šios technikos veikimui, kaip ir hierarchinio klasterizavimo atveju, užtenka atstumų tarp visų tiriamų objektų matricos.

t-SNE algoritmas skaičiuoja panašumo tikimybę pilnoje objektų požymių erdvėje ir bando kuo tiksliau ją perteikti naudojant mažesnės dimensijos erdvę. Taškų panašumas apskaičiuojamas kaip sąlyginė tikimybė, kad taškas A pasirinktų savo kaimynu tašką B, jei kaimynai būtų renkami proporcingai jų tikimybės tankiui pagal Gauso (normalų) pasiskirstymą, centruotą ties A. Toliau algoritmas mažina skirtumą tarp šių sąlyginių tikimybių aukštesnėje ir žemesnėje erdvėse, kad duomenų taškai būtų kuo tiksliau vaizduojami mažesnių matmenų erdvėje. Tam, kad išmatuotų sąlyginių tikimybių skirtumo sumažinimą, t-SNE sumažina visų duomenų taškų Kullback-Leibler nuokrypio (matas to, kaip vienos tikimybės pasiskirstymas skiriasi nuo tikėtinos tikimybės pasiskirstymo) sumą naudodamas gradiento nusileidimo metodą.

Paprastai tariant, t-SNE metodas mažina skirtumą tarp dviejų pasiskirstymų: to, kuris nurodo įvesties objektų porinį panašumą ir paskirstymo, kuris matuoja atitinkamų taškų porinį panašumą sumažintoje požymių erdvėje. Tokiu būdu t-SNE daugiamačius duomenis paskirsto žemesnėje erdvėje ir ieško duomenų modelių, nustatydamas stebimas grupes, pagrįstas duomenų taškų panašumu pagal daugelį bruožų. Tačiau po šio proceso įvestis nebebus atpažįstama, todėl negalima daryti išvadų, pagrįstų tik t-SNE išvestimi. Taigi tai daugiausia duomenų paieškos ir vizualizavimo technika. [MHi08]

Pritaikius šį metodą „EWAS Atlas" duomenims naudojant pasirinktus atstumo tarp ligų matus turėtų išsiskirti skirtingos ligų grupės pagal jų panašumus ir skirtumus.

2.5 Atvaizdavimo įrankiai

Šiame darbe aprašomų atvaizdavimo metodų kūrime daugiausiai naudojama nemokama R programavimo kalba ir aplinka, skirta statistiniams skaičiavimams ir atvaizdavimui. Tai yra GNU („GNU's Not Unix") projektas, panašus į S kalbą ir aplinką, kurį sukūrė „Bell Laboratories" (anksčiau „AT&T", dabar „Lucent Technologies") darbuotojai John Chambers ir jo kolegos. R gali

būti vertinama kaip kitoks S įgyvendinimas. Yra keletas svarbių skirtumų, tačiau daug S parašyto kodo veikia be didelių pakeitimų ir R aplinkoje.

R teikia daugybę statistinių (tiesinis ir netiesinis modeliavimas, klasikiniai statistiniai testai, klasifikavimas, grupavimas ir t.t.) ir grafinių metodų. Grafiniam atvaizdavimui R kalboje dažnai naudojama „ggplot2" biblioteka [WCH+16] bei kitokie įrankiai, iš kurių daugelis prieinami „Bioconductor" sistemoje, pavyzdžiui šiame darbe atvaizdavimui kartu su „ggplot2" naudota „karyoploteR" biblioteka [GSe17].

3. EPIGENETINIŲ LIGOS PROFILIŲ PALYGINIMAS

3.1. Atstumas tarp dviejų ligos profilių

Norint aukščiau aprašytiems daugiamačiams duomenims pritaikytus metodus (hierarchinį klasterizavimą, t-SNE) pritaikyti ligoms, reikia gebėti palyginti dviejų ligų profilius tarpusavyje. Paprasčiausias būdas tai padaryti - ligų panašumą apibrėžti kaip procentinę dalį citozinų (CpG pozicijų), kurie rodo skirtumus tarp sveikų ir sergančių grupių abejose lyginamose ligose.

Tačiau toks būdas ganėtinai naivus, nes ligos gali neturėti nei vieno bendro ligos paveikto citozino, tačiau vis tiek rodyti skirtumus funkciškai panašiose genomo vietose. Todėl matuojant atstumą tarp dvieju EWAS rezultatų reikėtų atsižvelgti ne tik į konkretaus citozino poziciją, bet ir į jo funkcinę svarbą, galbūt į atstumo matą įtraukiant ir genominius atstumus arba lyginamoms ligoms bendras genų ontologijos anotacijas.

3.2. Ligų grupavimas

Turint atstumo matą tarp dvieju ligų, galima grupuoti ligas į klasterius, pagal jų epigenetinių profilių panašumą. Toks grupavimas svarbus keletu aspektų. Jis mums padėtų įvertinti EWAS tyrimų pakartojamumą - pasakyti, kiek skirtingi tyrimai, tiriantys tą pačią ligą, yra panašūs vienas į kitą. Taip pat, matydami kurios ligos yra panašios viena į kitą, galėtume geriau suprasti ligų veikimo principus bei jų kilmę.

4. PRAKTINĖ TIRIAMOJI DALIS

Šiame skyriuje aprašoma praktinė darbo dalis. Jos metu pirmiausia buvo paruošiami „EWAS Atlas" duomenų bazės duomenys apie tris pasirinktas ligų grupes - kepenų ląstelių karcinomos,

antrojo tipo cukrinio diabeto ir sisteminės raudonosios vilkligės. Kiekvieną grupę sudaro trijų atskirų tyrimų, nagrinėjusių metilinimo pokyčius pasirinktose ligose, rezultatai. Toliau buvo sukurtos R funkcijos, įgyvendinančios skirtingus ligų palyginimo metodus taip, kad jų rezultatus būtų galima lyginti vieną su kitu ir spręsti ar skirtingų tyrimų, tyrusių tą pačia ligą, rezultatai yra panašesni lyginant juos su kitokio tipo ligų rezultatais. Jų palyginimui sukurta atskira funkcija, iš visų funkcijų rezultatų sudaranti ataskaitą ir kartu su kitomis patalpinta R pakete. Toliau pagal šių atvaizdavimo metodų rezultatus pasirinktas atstumų tarp ligų matas ir nebe kelioms grupėms, o visoms ,,EWAS Atlas" duomenų bazėje esančioms ligoms atliktas palyginimas naudojant atstumų matricą sudarytą naudojant pasirinktą atstumų matą.

4.1. „EWAS Atlas" duomenų bazės paruošimas

Darbo metu buvo naudojami trys pradiniai ,,EWAS Atlas" duomenų rinkiniai - CpG sričių anotacijos, ligų ir tyrimų duomenys. Iš nefiltruotų pradinių duomenų buvo paliktos tik tos ligos, kurias atitinkantys tyrimai naudojo Illumina 450k tipo DNR metilinimo mikrogardeles bei kuriuose tirtų mėginių skaičius buvo didesnis nei dvidešimt. Tuomet iš likusių ligų buvo paliktos tik tos, kurių rezultatuose buvo rasta bent penkiasdešimties skirtingų citozinų rodančių patikimus metilinimo skirtumus tarp sveikų ir sergančių grupių. Atlikus filtravimą išliko 304 atskirų tyrimų duomenys.

Šiame darbe skirtingų funkcijų ataskaitos kūrimui naudojami devynių ,,EWAS Atlas" žinių bazėje esančių tyrimų duomenys, suskirstyti į tris grupes - kepenų ląstelių karcinomos, antro tipo cukrinio diabeto ir sisteminės raudonosios vilkligės. Toliau atliekant ligų palyginimą tiriamos visos ligos.

4.1.1 Kepenų ląstelių karcinoma

Kepenų ląstelių karcinoma yra labiausiai paplitęs pirminio kepenų vėžio tipas. Ši liga dažniausiai pasireiškia žmonėms, sergantiems lėtinėmis kepenų ligomis, tokiomis kaip cirozė, kurią sukelia hepatito B arba hepatito C infekcija. Ji labiau būdinga žmonėms, kurie geria daug alkoholio ir kurių kepenyse kaupiasi riebalai.

Kepenų ląstelių karcinomos, labiausiai paplitusios kepenų vėžio rūšies, rizika yra didesnė žmonėms, sergantiems ilgalaikėmis kepenų ligomis. Tai taip pat didesnė, jei kepenis skauda dėl infekcijos su hepatitu B ar hepatitu C.

4.1.2 Antrojo tipo cukrinis diabetas

Per keletą kartų dramatiškai padaugėjo antro tipo diabeto ir nutukimo atvejų, jie šiuo metu siekia epidemijos lygį. Be genetinių rizikos veiksnių, tiriami epigenetiniai mechanizmai, kuriuos sukelia besikeičianti aplinka, atsižvelgiant į jų vaidmenį šių sudėtingų ligų patogenezėje. EWAS atskleidė reikšmingą diabeto, nutukimo ir kūno masės indekso ryšį su DNR metilinimu. Tačiau populiacijos iš Vidurinių Rytų, kur antrojo tipo diabeto ir nutukimo rodikliai yra aukščiausi visame pasaulyje, iki šiol nebuvo ištirti. [MAZ+16]

4.1.3 Sisteminė raudonoji vilkligė

Kad organizmas išliktų sveikas, imuninė sistema kovoja su pavojingomis infekcijomis ir bakterijomis. Autoimuninė liga atsiranda, kai imuninė sistema puola kūną, nes supainioja jį su svetimkūniu. Yra daug autoimuninių ligų, įskaitant sisteminę raudonąją vilkligę (SRV).

SRV yra lėtinė liga, kuriai būdingi sunkesnių simptomų etapai, pakaitomis su lengvesniais simptomais. Daugelis žmonių, sergančių SRV, gydomi gali gyventi normalų gyvenimą.

Remiantis Amerikos Lupus fondo duomenimis, mažiausiai 1,5 mln. Amerikiečių gyvena su diagnozuota vilklige. Fondas mano, kad realiai šia liga sergančių žmonių yra daug daugiau ir kad daugelis atvejų yra nediagnozuoti. [Her16]

4.2. Ligos profilių atvaizdavimo ataskaita

Ligos profilių atvaizdavimui buvo sukurta ataskaita, padedanti palyginti įvairius profilių aspektus: pasiskirstymą po genominius elementus, paveiktų pozicijų pasiskirstymą chromosomose, jų atstumus iki skirtingų genominių elementų ir genų ontologijos terminų skirtumus. Šios ataskaitos generavimui kurta R biblioteka, turinti funkciją, gebančią visų ligų palyginimo funkcijų rezultatus pateikti kartu.

4.2.1 Duomenų apie ligas lentelė

Pirmajame ataskaitos puslapyje esanti lentelė yra skirta supažindinti su tiriamomis ligomis ir bruožais bei su duomenis teikiančiais šaltiniais. Lentelėje matomi duomenis teikiančių tyrimų identifikatoriai bei tyrimų straipsnių identifikatoriai PubMed archyve. Taip pat atvaizduojami ligos paveiktų CpG sričių skaičius, ligų ir audinių pavadinimai. (1 pav.)

study_id	probes	trait	tissue	PMID
ES00435	4617	hepatocellular carcinoma (HCC)	liver	23208076
ES00714	109	hepatocellular carcinoma (HCC)	liver, whole blood	29848370
ES00975	6510	hepatocellular carcinoma (HCC)	tumor tissue, liver	29988590
ES00298	951	type 2 diabetes (T2D)	whole blood	26823690
ES00290	1649	type 2 diabetes (T2D)	pancreatic islet	24603685
ES00295	251	type 2 diabetes (T2D)	liver	26418287
ES00546	7625	systemic lupus erythematosus (SLE)	whole blood	29437559
ES00911	58	systemic lupus erythematosus (SLE)	peripheral blood mononuclear cell	30301579
ES01237	7585	systemic lupus erythematosus (SLE)	whole blood	31428085

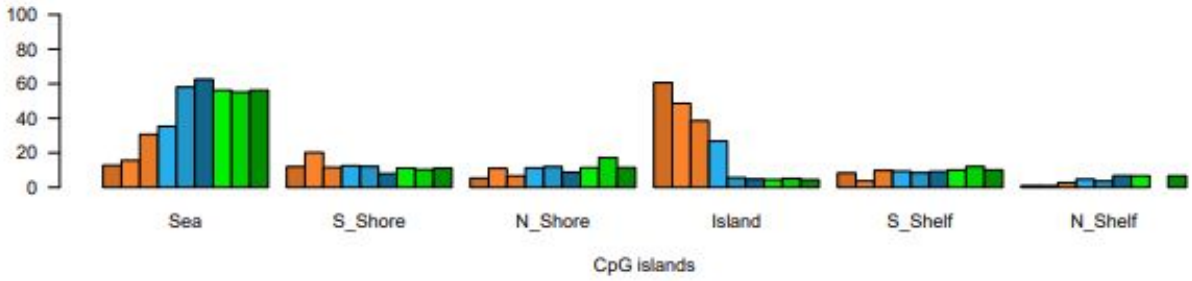
1 pav. Duomenų apie ligas lentelė

4.2.2 Pasiskirstymas po genominius elementus

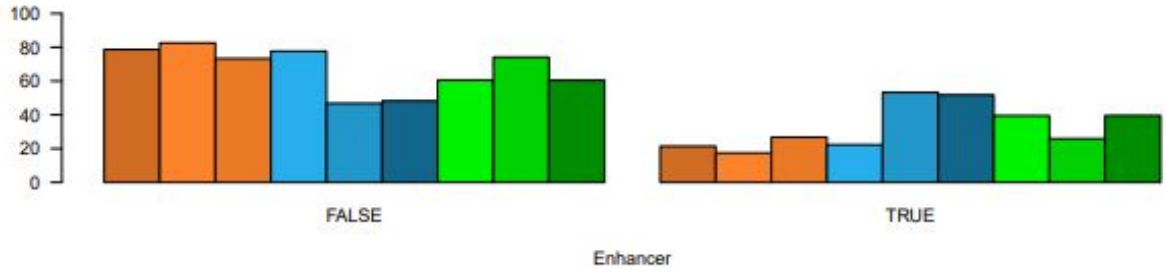
Pirmosios R funkcijos rezultato grafikas - stulpelinė diagrama, kurioje kiekvienai ligai skiriamas atskiras stulpelis, atvaizduojantis su ta liga susijusių metilinimo pokyčius turinčių CpG pozicijų, esančių tiriamuose genominiuose elementuose, dalį procentais.

Funkcijos įvestis yra norimų palyginti ligų CpG pozicijų identifikatorių sąrašas ir lentelė iš dviejų stulpelių - visų DNR metilinimo gardele tirtų CpG pozicijų identifikatorių bei kiekvienai tokiai pozicijai priskiriamo genominio elemento. Taip pat funkcijai paduodamas kiekvienai tiriamai ligai skiriamų spalvų sąrašas.

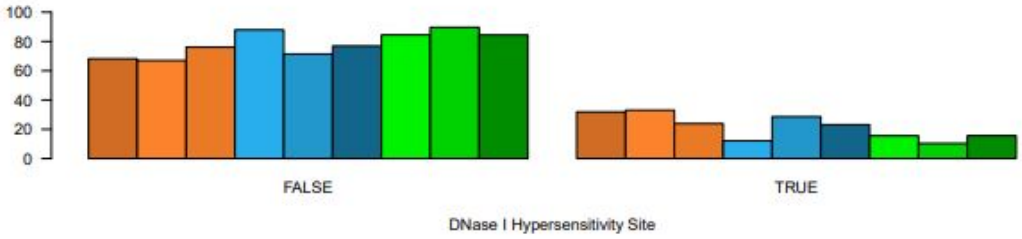
Pritaikius šią funkciją kepenų ląstelių karcinomos, antro tipo cukrinio diabeto ir sisteminės raudonosios vilkligės duomenims, gautas grafikas (2 pav.), kuriame matoma jog ligos labiausiai išsiskiria pagal pasiskirstymą CpG saloje „Island” bei už salų ribų („Sea”) (2 pav.), skirtingai metilinamų regionų tipus (5 pav.) ir skirtingas genų dalis (6 pav.). Skirtumai tarp sveikų ir sergančių vilklige pagrinde buvo randami už CpG salų (Sea), o metilinimo skirtumai tarp sveikų ir sergančių kepenų karcinoma daugiausia kaupiasi CpG salų viduje. Taip pat matoma, jog vilkligė, tirta skirtinguose tyrimuose, rodo panašius rezultatus. Didžiausias panašumas tarp ligų matomas DNazės I padidinto jautrumo srityse (4 pav.)



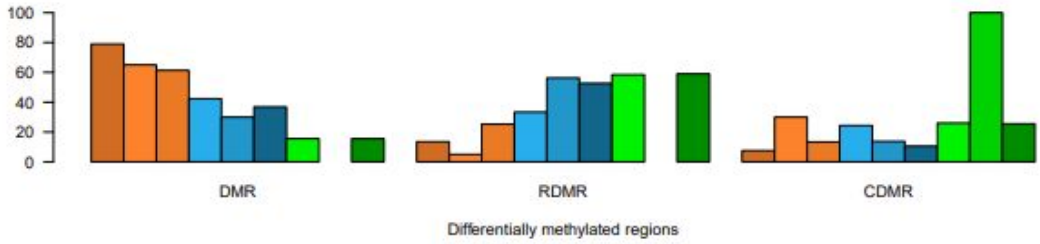
2 pav. Ligų paveiktų citozinų pasiskirstymas CpG salose



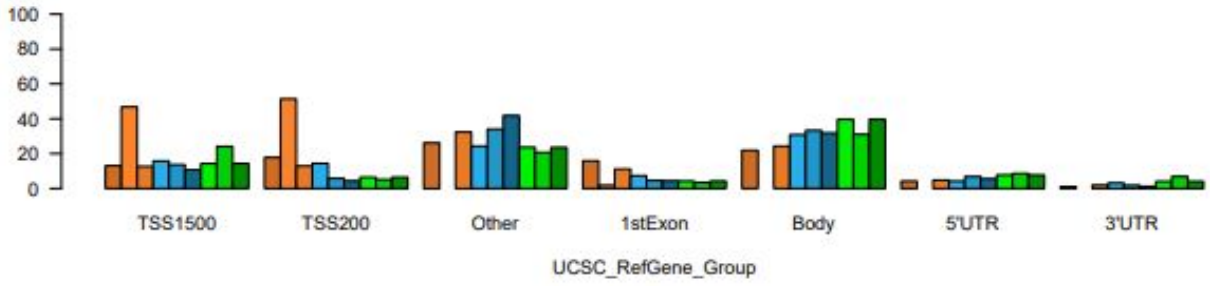
3 pav. Ligų paveiktų citozinų pasiskirstymas enhanceriuose



4 pav. Ligų paveiktų citozinų pasiskirstymas DNazės I padidinto jautrumo srityse



5 pav. Ligų paveiktų citozinų pasiskirstymas skirtingai metilinamuose regionuose

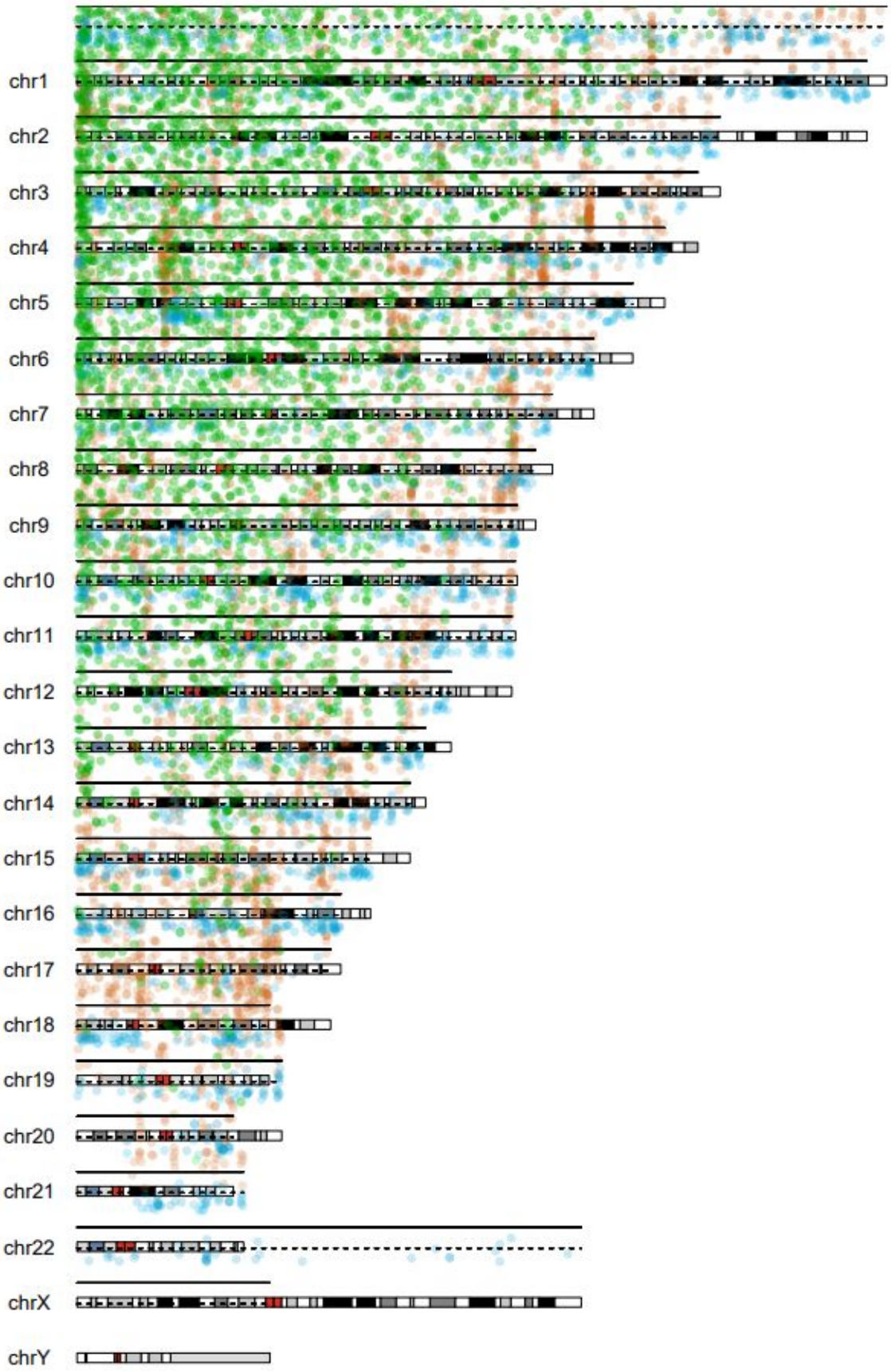


6 pav. Ligų paveiktų citozinų pasiskirstymas skirtingose genų pozicijose

4.2.3 Paveiktų pozicijų pasiskirstymas chromosomose

Sekanti ligų palyginimo funkcija analizuoja skirtumų kaupimasi chromosomų pozicijose. Rezultato grafike atskirai atvaizduodama kiekviena chromosoma, o taškai ant jų, rodo kiekvieno citozino, rodančio metilinimo pokyčius ligos atveju, patikimumo įvertį: p-reikšmę. Skaičiavimams funkcija naudoja ligų CpG pozicijų identifikatorius bei jiems priskiriamus spalvų sąrašus ir duomenis iš „EWAS Atlas".

Funkcijos grafike (7 pav.) matoma, jog didžiausią kiekį metilinimo pakitimų turi sisteminė raudonoji vilkligė. Jos paveiktose pozicijose kartu kaupiasi dauguma kepenų ląstelių karcinomos paveiktų citozinų. Tuo tarpu antrojo tipo cukrinio diabeto poveikis matomas kitose pozicijose nei kitų dviejų ligų, taigi ji atsiskiria pagal paveiktas chromosomų pozicijas. Grafike antroje, ketvirtoje ir penktoje chromosomose matomi ryškūs rudos spalvos stulpeliai, kuriuose kaupiasi kepenų ląstelių karcinomos paveikti citozinai.



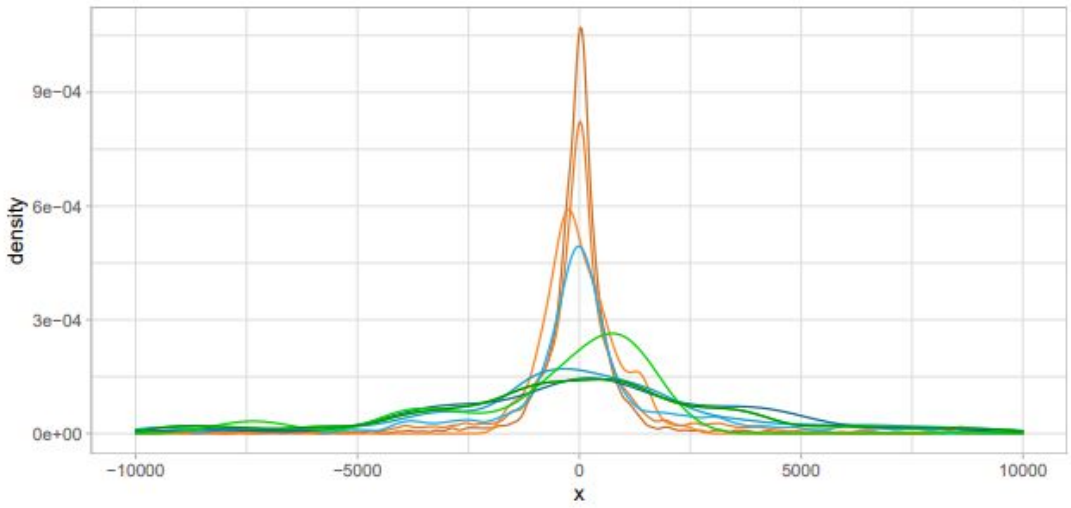
7 pav. Paveiktų pozicijų pasiskirstymas chromosomose

4.2.4 Atstumai iki skirtingų genominių elementų

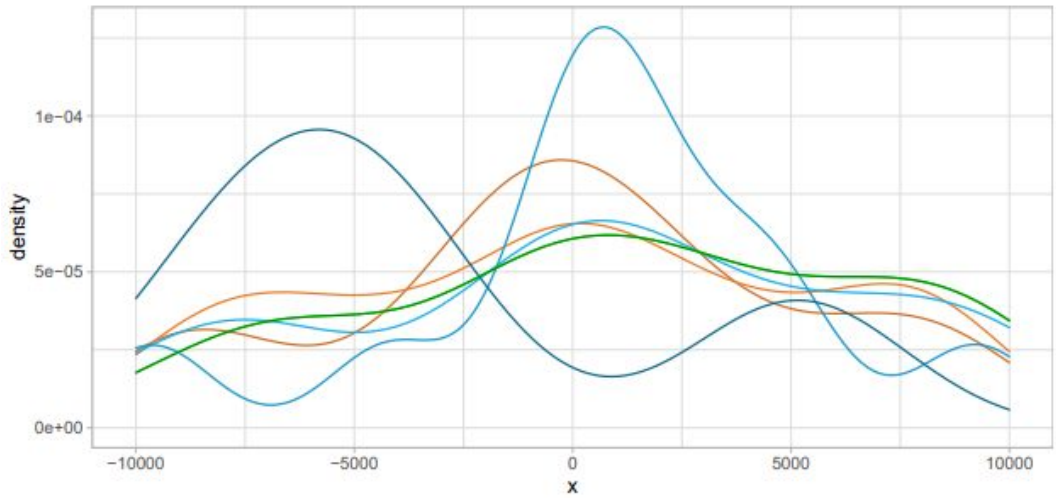
Sekančios funkcijos tikslas - atvaizduoti suvidurkintą konkretaus genominio elemento vaizdą nurodant kuriose pozicijose aplink tą elementą išsidėlioja ligos paveikti citozinai

Tai įgyvendinta sukūrus funkciją, kuri, pateikus lentelę, sudarytą iš visų mus dominančių genominių elementų centrų pozicijų, paskaišiuoja atstumus nuo kiekvieno ligos paveikto citozino iki jam artimiausio genominio elemento centro ir sukuria šių vidurkių tankio funkcijų grafiką. Genomiai elementai, tirti šio darbo metu, buvo CpG salos, enhanceriai bei genų transkripcijos pradžios bei pabaigos pozicijos.

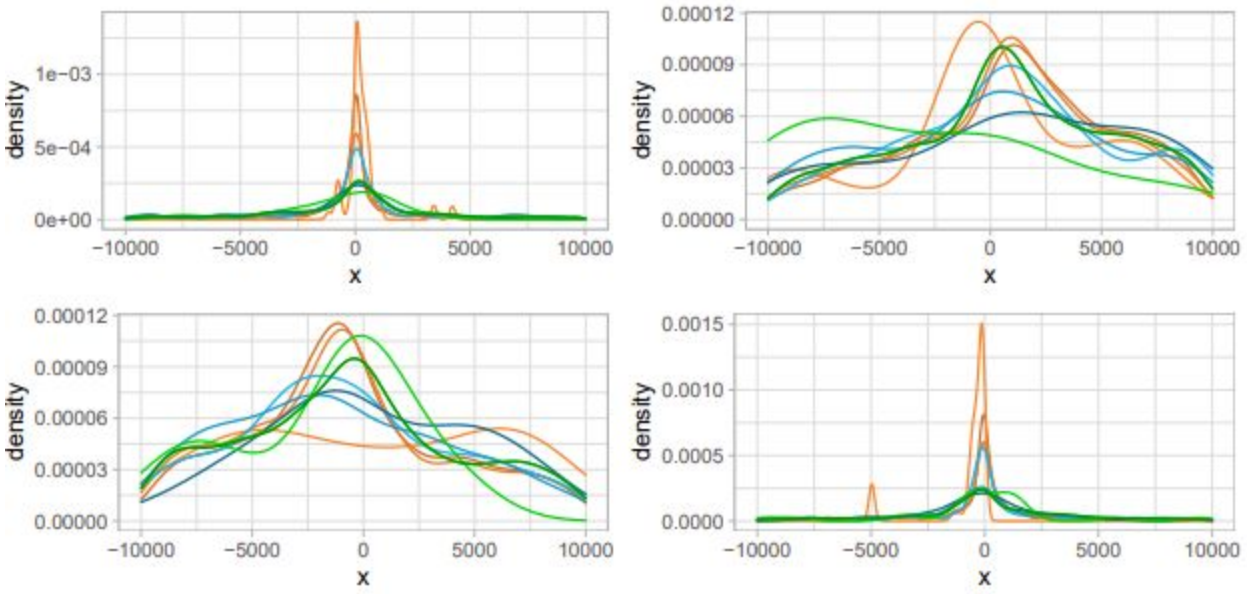
Funkcijos grafikuose matomi tiriamų ligų atstumai iki skirtingų genominių elementų - CpG salų (8 pav.), enhancerių (9 pav.) bei skirtingų genų pozicijų (10 pav.). Atstumo iki genų pozicijų grafikas yra padalintas į keturis smulkesnius, kuriuose atvaizduojami ligų atstumai iki genų pradžios ir pabaigos abiejose chromosomų grandinėse. Rezultate matoma, kad visos ligos pagal skirtingus atstumus išsiskiria. Ruda spalva vaizduojama karcinoma turi daugiausiai pokyčių genų transkripcijos pradžioje, o žalia spalva pažymėta vilkligė turi mažiausiai. Mėlyna spalva pavaizduotas diabetas užima tarpinę poziciją. Pagal atstumą iki ehancerių ryškiai išsiskiria skirtingų tyrimų tirtas diabetas. Taigi tai yra patikimas matas skirtumams tarp ligų išryškinti.



8 pav. Ligų atstumai iki CpG salų



9 pav. Ligų atstumai iki enhancerių



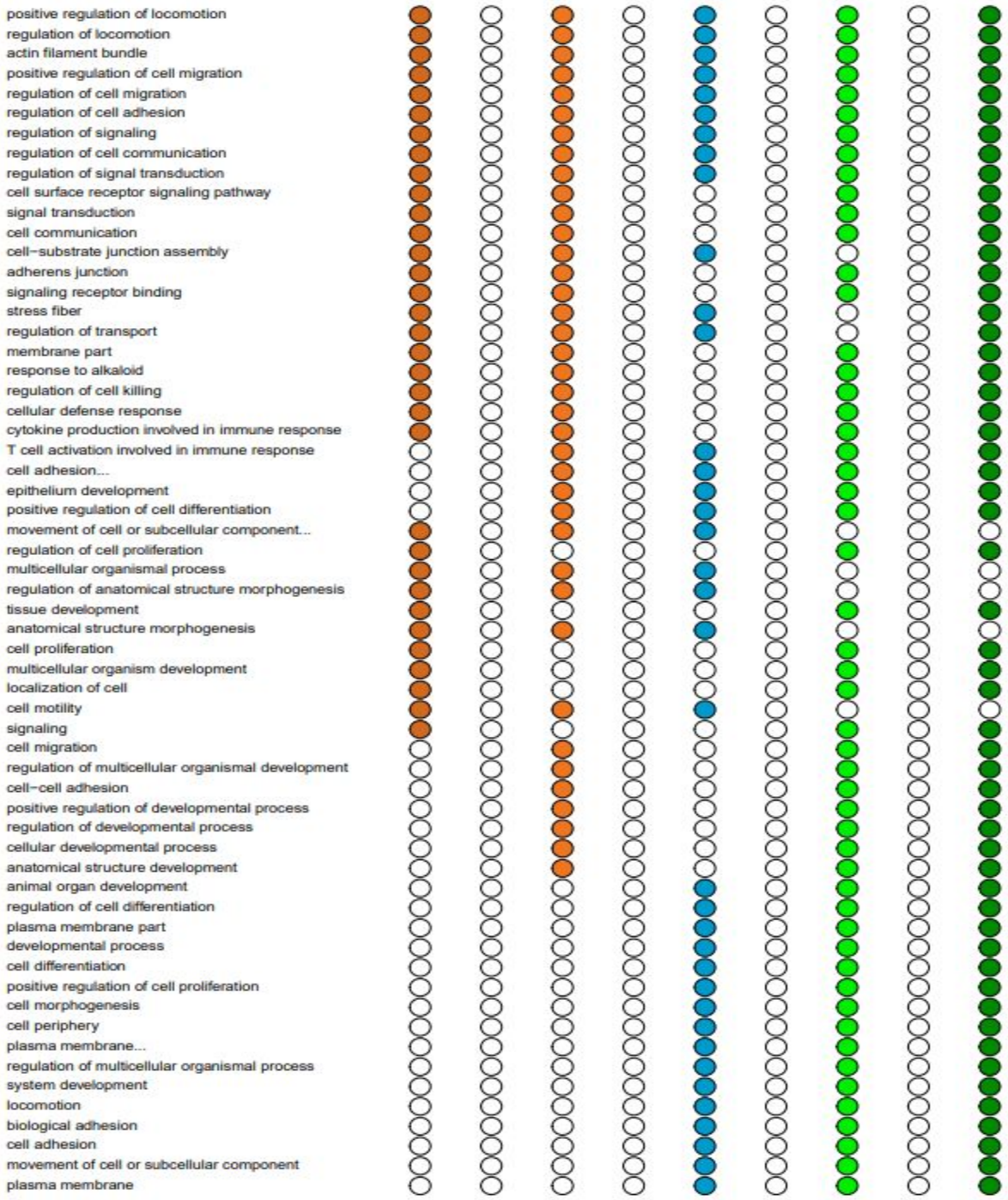
10 pav. Ligų atstumai iki geno pradžios ir pabaigos abiejose chromosomų grandinėse

4.2.5 Genų ontologijos terminų skirtumai

Paskutinė šio darbo funkcija buvo skirta genų ontologijos terminų palyginimui. Funkcija tai įgyvendina kiekvienai tiriamai ligai iš genų ontologijos resurso (Gene Ontology Resource) išrinkdama jos paveiktoms citozinų pozicijoms būdingus genų anotacinius terminus ir atvaizduoja juos grafiškai.

Funkcijos rezultato grafike (11 pav.) matomi visi genų ontologijos terminai, kurie yra būdingi bent vienai tiriamai ligai ir jiems iš dešinės atvaizduojami taškai, žymintys kurioms būtent ligoms buvo priskirti konkretūs terminai. Atvaizduojami tik šešiasdešimt dažniausiai tarp ligų

pasikartojančių terminų juos rūšiuojant pagal skaičių ligų, kurioms tie terminai yra būdingi. Daugiausia terminų pasižymi sisteminė raudonoji vilkligė, todėl didelė terminų dalis sutampa su kitomis ligomis, kurios jų turi mažiau, taigi skirtingos ligos pagal savo terminus ryškiai neišsiskyrė, todėl taikant šį metodą ryškių dėsningumų tarp ligų nerasta. Daugiausiai ligų priklausantys terminai - teigiamas judėjimo reguliavimas, judėjimo reguliavimas, rūgšties gijų pluoštas, teigiamas ląstelių migracijos reguliavimas, ląstelių migracijos reguliavimas, ląstelių adhezijos reguliavimas, bendravimo tarp ląstelių, signalų ir jų perdavimo reguliavimas.



11 pav. Genų ontologijos terminų skirtumai

4.3 Atstumo tarp ligos profilių kūrimas

Remiantis gautais R funkcijų rezultatais matyti, jog ligos daugiausia viena nuo kitos skiriasi pagal citozinų pasiskirstymą CpG salose, skirtingai metilinamuose regionuose ir skirtingose genų pozicijose bei pagal atstumus iki skirtingų genominių elementų. Darbo metu buvo praktiškai išbandyta daugiau nei viena ataskaita naudojant skirtingas ligas ir tarp jų buvo nemažai tokių kurios

rodė skirtumus genų ontologijos terminuose. Todėl, apjungiant visus ataskaitų rezultatus, atstumo tarp ligų matais buvo pasirinkti genų ontologijos terminai ir CpG salos kartu su ligų paveiktų citozinų identifikatoriais.

Skaičiuojant bendrą visų ligų atstumų matricą, kiekvienam požymiui atskirai suskaičiuotos atstumų matricos taikant koreliaciją tarp šio požymio įverčių kiekvienai ligai bei atliekant palyginimą skirtingų ligų paveiktiems citozinams. Atskiros matricos sudėtos, kiekvienai taikant vienos trečiosios svorį:

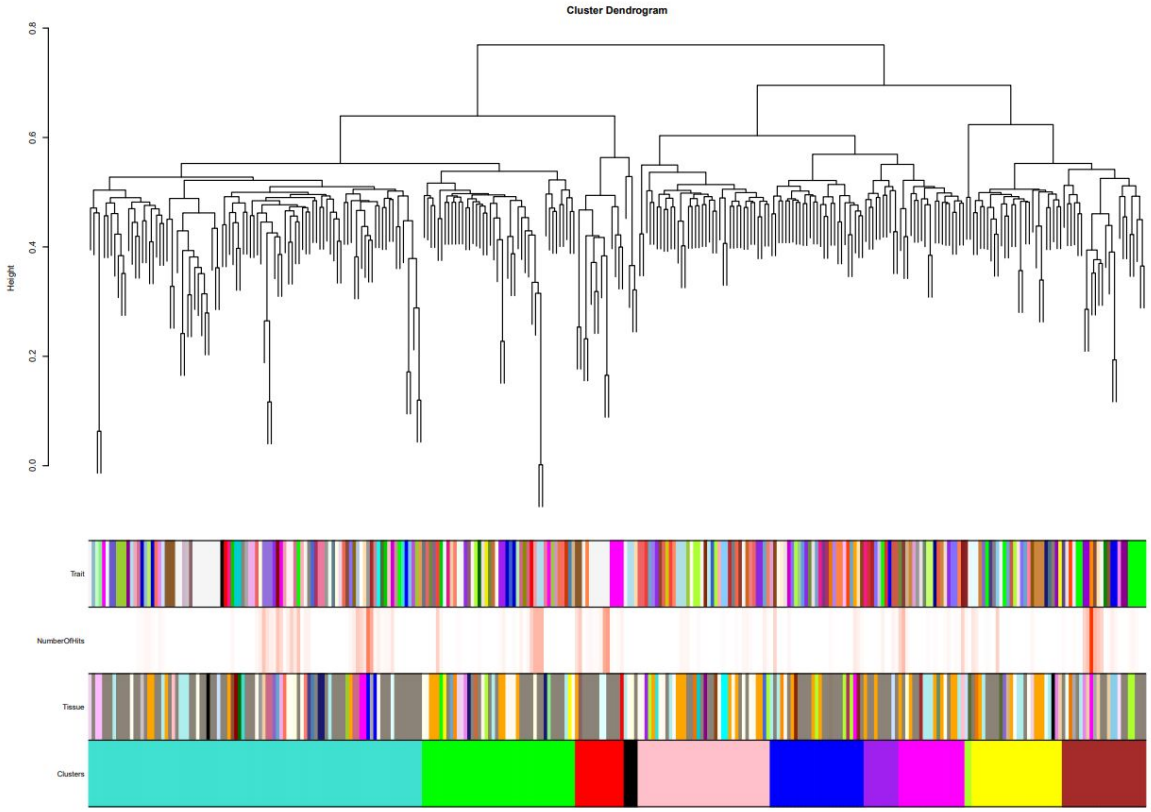
$$d_{ij} = 1 - (1/3 \frac{|CGSi \cap CGSj|}{|CGSi \cup CGSj|} + 1/3 \frac{cor(GOPi, GOPj) + 1}{2} + 1/3 \frac{cor(ISLANDSi, ISLANDSj)+1}{2})$$

Formulėje CGS žymimi ligų paveikti citozinai, GOP genų ontologijos terminų p-reikšmės, o ISLANDS - pasiskirstymo įverčių visose CpG salose suma. Taip pat cor - tai Pearsono koreliacija, o d_{ij} - atstumo matas tarp dviejų ligų rezultatų: i ir j.

4.4. Ligų rezultatų grupavimas ir palyginimas

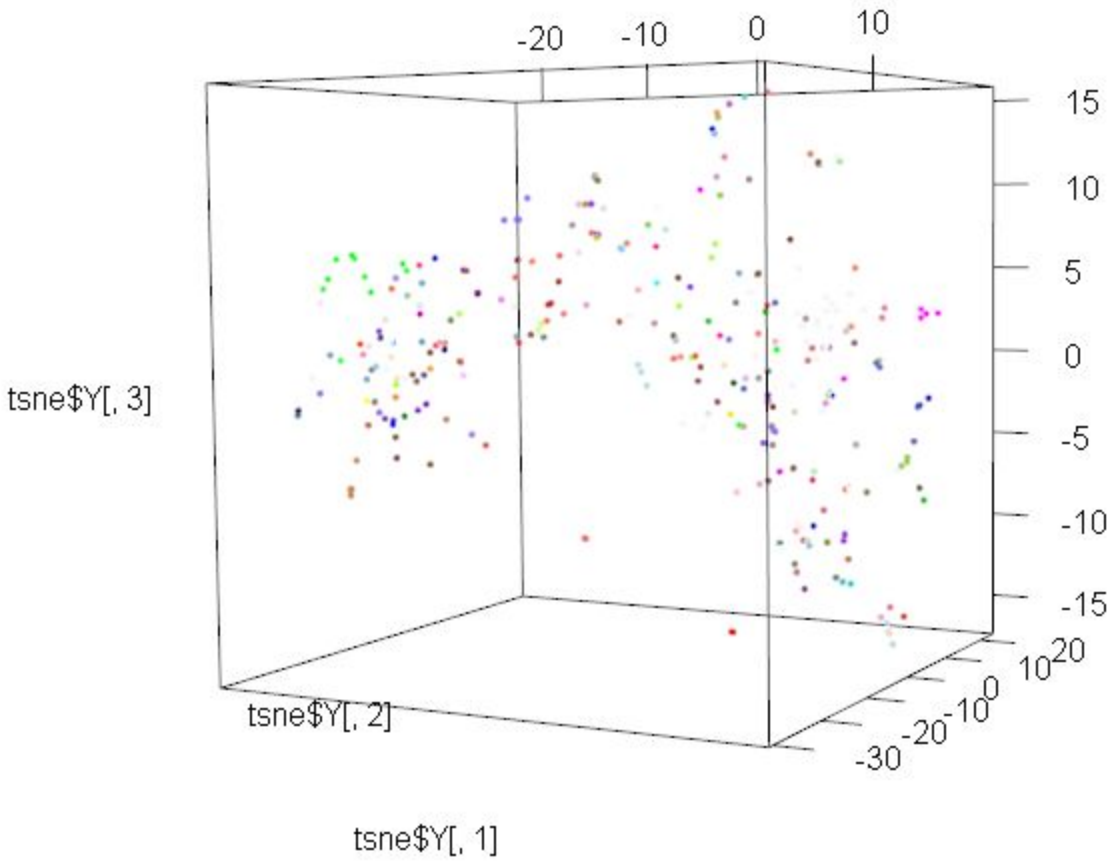
Sukurtas atstumo matas pritaikytas „EWAS Atlas" duomenims naudojant hierarchinio klasterizavimo ir t-SNE metodus. Hierarchinio klasterizavimo metodu sugeneruotoje dendrogramoje (12 pav.) matoma jog visos ligos išsiskiria į dvi pagrindines grupes bei panašios ligos į dešimt smulkesnių grupių.

Grafike ryškiausiai matomos baltos spalvos rūkymo ir žalios spalvos senėjimo grupės, kuriose skirtingų tyrimų tirti požymiai klasterizuojasi kartu. Taip pat grupuojant pagal tirtą audinį matomos viso kraujo grupės. Pagrindines klasterių grupes sudaro skirtingo tipo ligos - autoimuninės, neurologinės, vėžiai.



12 pav. Hierarchinio klasterizavimo dendrograma

„EWAS Atlas" duomenims pritaikius t-SNE metodą gautas trimatis visų ligų grupavimo grafikas (13 pav.), kuriame vienodos ligos tirtos skirtingų tyrimų nuspalvintos viena spalva. Jame ryškiausiai matomos žalios bei baltos spalvos ligų grupės, kuriose grupuojasi skirtingų tyrimų tirti atitinkamai rūkymas ir senėjimas.



13 pav. Trimatis t-SNE metodo grafikas

4.5. R paketas „methprofiler”

Darbo metu sukurtas R paketas, galintis atlikti visus aukščiau aprašytus analizės ir atvaizdavimo žingsnius ir juos atlikęs pateikti visų žingsnių rezultatus kartu vienoje ataskaitoje, kas padeda daryti tyrimo išvadas atsižvelgiant į visų palyginimo žingsnių grafikus.

„EWAS Atlas” plečiantis toliau, šis paketas gali pasitarnauti įtraukiant naujus rezultatus, naudojant rezultatus gautus iš kitų mikrogardelių platformų.

SANTRAUKA

Šio darbo metu įvairūs daugiamačių biomedicinos duomenų atvaizdavimo metodai buvo pritaikyti epigenetinių ligos profilių atvaizdavimui ir palyginimui. Sukurtas R paketas, galintis atlikti citozinų pasiskirstymo po genominius elementus, paveiktų pozicijų pasiskirstymo chromosomose, jų atstumų iki skirtingų genominių elementų ir genų ontologijos terminų palyginimą. Visi šie rezultatai pateikti vienoje ataskaitoje, kartu su lentele aprašančia tyrimų apie ligas duomenis pirmajame jos puslapyje.

Įgyvendinus šiuos metodus buvo atsižvelgta į išryškėjusius panašumus bei skirtumus tarp ligų ir dalis šių požymių buvo panaudoti atstumo mato tarp dviejų ligos profilių sukūrimui. Naudojant sukurtą atstumo matą buvo gauta ligų tyrimų atstumų matrica aprašanti atstumus tarp visų „EWAS Atlas" duomenų bazėje esančių tyrimų porų. Šiai matricai pritaikyti hierarchinio klasterizavimo bei t-SNE metodai ir aliktas visų „EWAS Atlas" duomenų palyginimas, kuris parodė, kad panašios ligos grupuojasi į bendrus klasterius.

SUMMARY

In the course of this work, various methods for visualizing high-dimensional biomedical data were applied to the visualization and comparison of epigenetic disease profiles. An R package was developed, capable of comparing the distribution of cytosines by genomic elements, the distribution of affected positions on chromosomes, their distances to different genomic elements, and gene ontology terms. All of these results are presented in a single report, along with a table describing the disease research data on its first page.

After the implementation of these methods, the similarities and differences between the diseases were taken into account and some of these features were chosen as measures of distance to form a matrix of distances between diseases. Hierarchical clustering and t-SNE methods were applied to this matrix and comparison of all „EWAS Atlas" data was made, which showed that similar diseases clustered into common groups.

REZULTATAI IR IŠVADOS

Darbo metu buvo apžvelgti skirtingi paprasti ir sudėtingi daugiamačių biomedicinos duomenų atvaizdavimo metodai. Pasirinkus kelis iš jų, tinkamiausius epigenetinių ligos profilių atvaizdavimui ir palyginimui buvo sukurtas R paketas su funkcijomis įgyvendinančiomis šiuos metodus.

Sukurtos funkcijos geba atlikti ligų paveiktų CpG pozicijų pasiskirstymo genominius elementuose, chromosomų pozicijose, atstumų iki skirtingų genominių elementų ir genų ontologijos terminų analizę. Visų funkcijų rezultatų apibendrinimui bei jų išvadų darymui palengvinti sukurta papildoma funkcija, visus rezultatus pateikianti bendroje ataskaitoje. R paketo ataskaitos funkcijos panaudotos kepenų ląstelių karcinomos, antrojo tipo cukrinį diabeto bei sisteminės raudonosios vilkligės palyginimui. Rezultatai parodė, kad šios ligos labiausiai išsiskiria pagal pasiskirstymą CpG salose „Island" ir už visų salų ribų („Sea"), skirtingai metilinamuose regionuose bei pagal ligų paveiktų citozinų atstumus iki skirtingų genominių elementų. Visos ligos yra panašiausios jas lyginant pagal genų ontologijos terminus, kur maža dalis terminų nepriklauso visoms tiriamoms ligoms.

Kadangi tirtos ligos labiausiai skyrėsi pagal pasiskirstymą CpG salose ir lyginant pagal genų ontologijos terminus, šie požymiai kartu su pačių paveiktų citozinų identifikatoriais buvo pasirinkti kaip atstumo matai atstumų tarp ligų matricos sudarymui. Sukurta matrica panaudota visų „EWAS Atlas" duomenų palyginimui hierarchinio klasterizavimo ir t-SNE metodais. Palyginimo rezulatai parodė, kad visos panašiausios ligos grupuojasi į bendras grupes, skirtingų tyrimų tirti senėjimas ir rūkymas klasterizavosi į atitinkamas grupes hierarchinio klasterizavimo dendrogramoje bei t-SNE metodo grafike.

SĄVOKŲ APIBRĖŽIMAI

- DNR metilinimas – DNR azotinių bazių adenino ir citozino metilinimas, susidarant N6-metiladeninui(m6A), 5-metilcitozinui(m5C) ir N4-metilcitozinui (m4C).
- Epigenetika - biologijoje epigenetika yra paveldimų fenotipo pokyčių, nesusijusių su DNR sekos pokyčiais, tyrimas.
- Genų ontologija - ontologija yra formalus žinių visumos atvaizdavimas tam tikroje srityje. Ontologijas paprastai sudaro klasių rinkinys (arba terminai ar sąvokos) su santykiais tarp jų. Genų ontologija apibūdina biologijos žinias trimis aspektais: molekulių funkcijos, ląstelių struktūros ir biologinių procesų.
- EWAS (Epigenome-Wide Association Study) - kiekybiškai įvertinamų epigenetinių požymių, tokių kaip DNR metilinimas, rinkinių tyrimas siekiant nustatyti ryšius tarp epigenetinės variacijos ir konkretaus atpažįstamo fenotipo.

ŠALTINIAI

[Boc18] T. Bock. *What is Hierarchical Clustering?*, 2018. Prieiga per internetą: https://www.displayr.com/what-is-hierarchical-clustering/.

[EMS14] S. Evergreen, A. Mountain, M. Salm. *Scatterplot*, 2014. Prieiga per internetą: https://www.betterevaluation.org/en/evaluation-options/scatterplot.

[FVo83] A.P Feinberg, B. Vogelstein. *Hypomethylation distinguishes genes of some human cancers from their normal counterparts*. 1983. Prieiga per internetą https://www.nature.com/scitable/content/Hypomethylation-distinguishes-genes-of-some-human-cancers-11329/.

[GSe17] B. Gel, E. Serra. *karyoploteR : an R / Bioconductor package to plot customizable genomes displaying arbitrary data,* 2017. Prieiga per internetą: https://academic.oup.com/bioinformatics/article/33/19/3088/3857734.

[Her16] J. Herndon. Systemic Lupus Erythematosus (SLE), 2016. Prieiga per internetą: https://www.healthline.com/health/systemic-lupus-erythematosus.

[LZL+18] M. Li, D. Zou, Z. Li, R. Gao, J. Sang, Y. Zhang, R. Li, L. Xia, T. Zhang, G. Niu, Y. Bao, Z. Zhang. *EWAS Atlas: a curated knowledgebase of epigenome-wide association studies,* 2018. Prieiga per internetą: https://academic.oup.com/nar/article/47/D1/D983/5144953.

[MAZ+16] M W. A. Al Muftah, M. Al-Shafai, S. B. Zaghlool, A. Visconti, P. Tsai, P. Kumar, T. Spector, J. Bell, M. Falchi, K. Suhre. *Epigenetic Associations of Type 2 Diabetes and BMI in an Arab Population,* 2016. Prieiga per internetą: https://pubmed.ncbi.nlm.nih.gov/26823690/.

[MHi08] L. van der Maaten, G. Hinton. *Visualizing Data using t-SNE,* 2008. Prieiga per internetą:

http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[Sim08] D. Simmons. *Epigenetic influence and disease*, 2008. Prieiga per internetą: https://www.nature.com/scitable/topicpage/epigenetic-influences-and-disease-895/.

[WCH+16] H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, D. Dunnington. *ggplot2: Elegant Graphics for Data Analysis,* 2016. Prieiga per internetą: https://ggplot2.tidyverse.org/.

[Wei06] B. Weinhold. *Epigenetics: The Science of Change*, 2006. Prieiga per internetą: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1392256/.

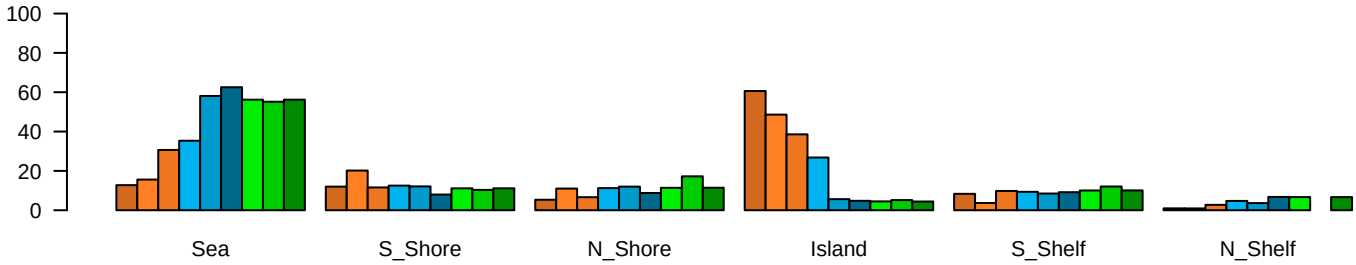
PRIEDAI

Prieiga internete prie šiame darbe sukurto R paketo:

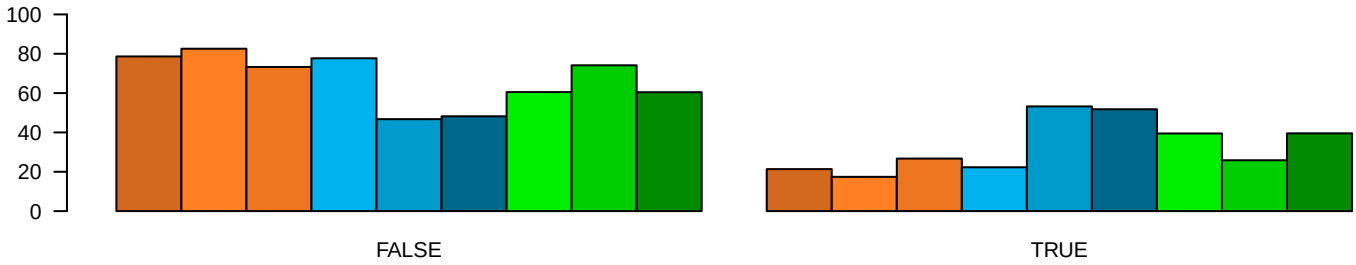
https://git.mif.vu.lt/arta3682/bakalaurinis

Toliau pateikiamas darbo metu sukurto R paketo funkcijų rezultatų ataskaitos pavyzdys.

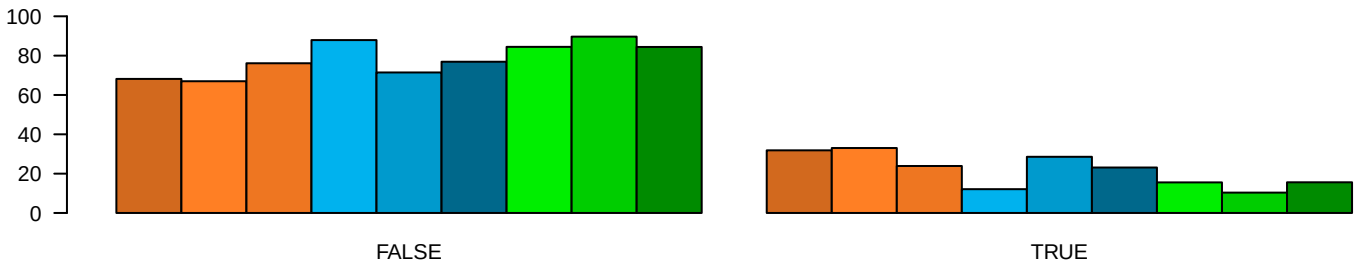
study_id	probes	trait	tissue	PMID
ES00435	4617	hepatocellular carcinoma (HCC)	liver	23208076
ES00714	109	hepatocellular carcinoma (HCC)	liver, whole blood	29848370
ES00975	6510	hepatocellular carcinoma (HCC)	tumor tissue, liver	29988590
ES00298	951	type 2 diabetes (T2D)	whole blood	26823690
ES00290	1649	type 2 diabetes (T2D)	pancreatic islet	24603685
ES00295	251	type 2 diabetes (T2D)	liver	26418287
ES00546	7625	systemic lupus erythematosus (SLE)	whole blood	29437559
ES00911	58	systemic lupus erythematosus (SLE)	peripheral blood mononuclear cell	30301579
ES01237	7585	systemic lupus erythematosus (SLE)	whole blood	31428085



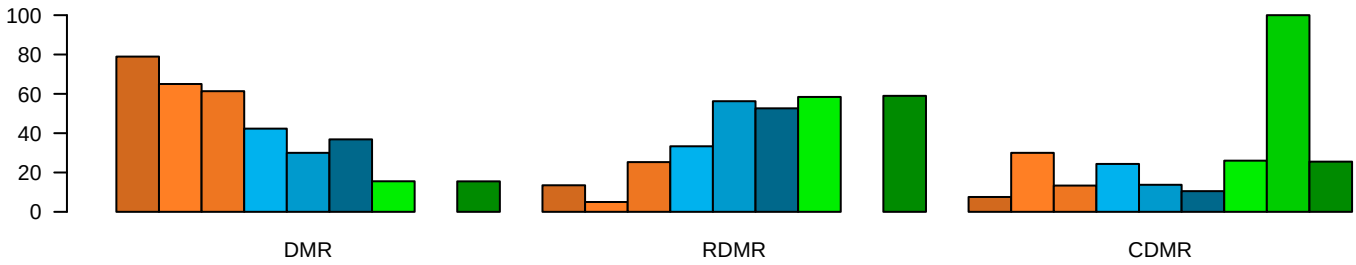
CpG islands



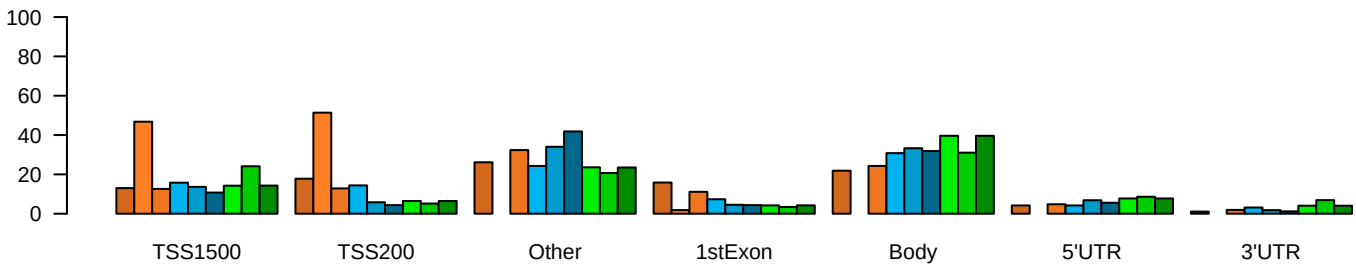
Enhancer



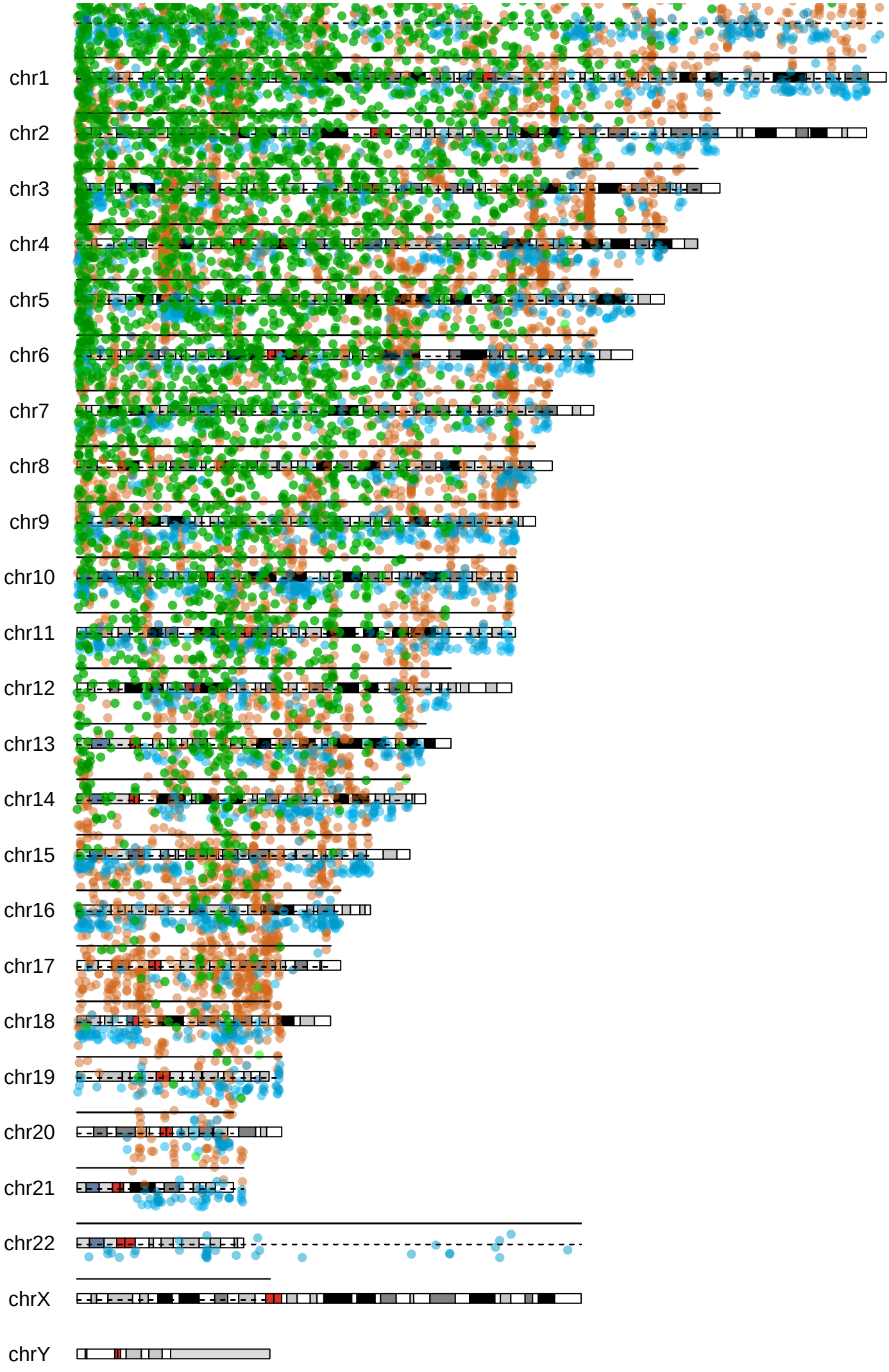
DNase I Hypersensitivity Site

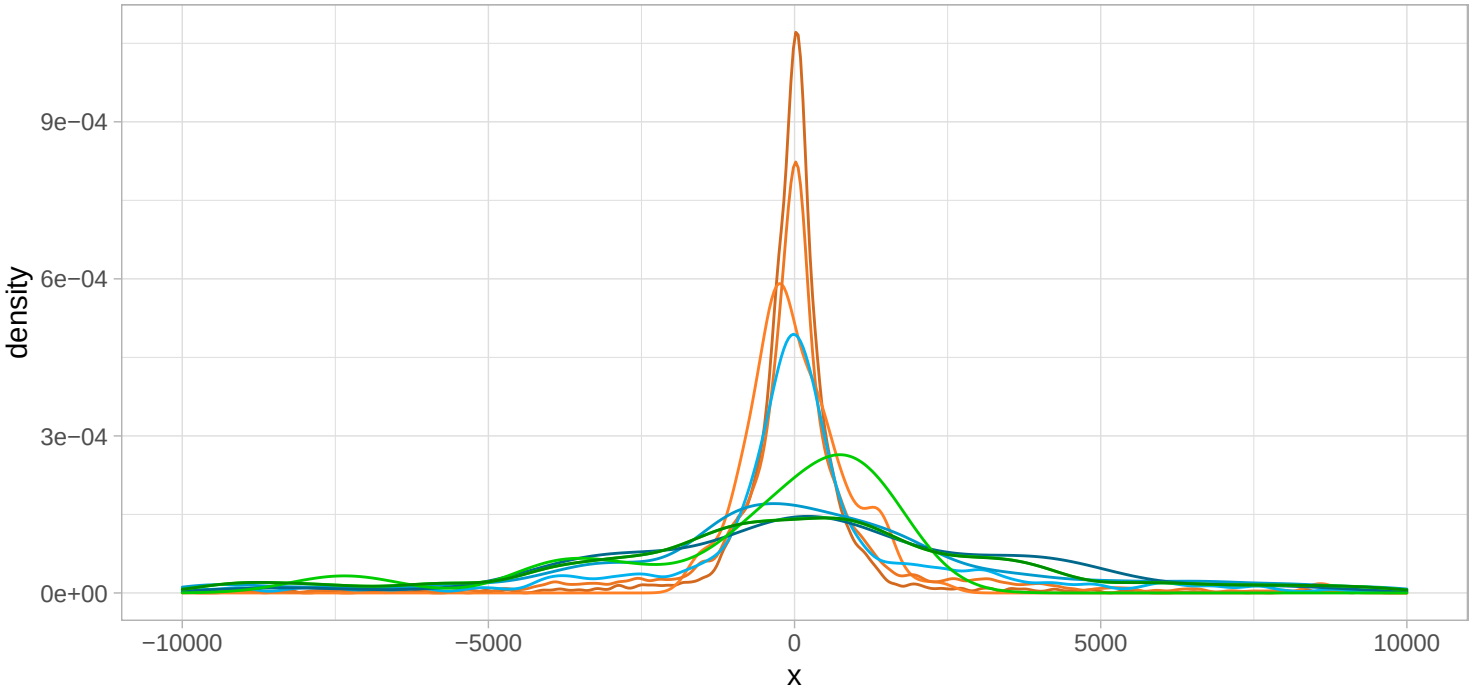


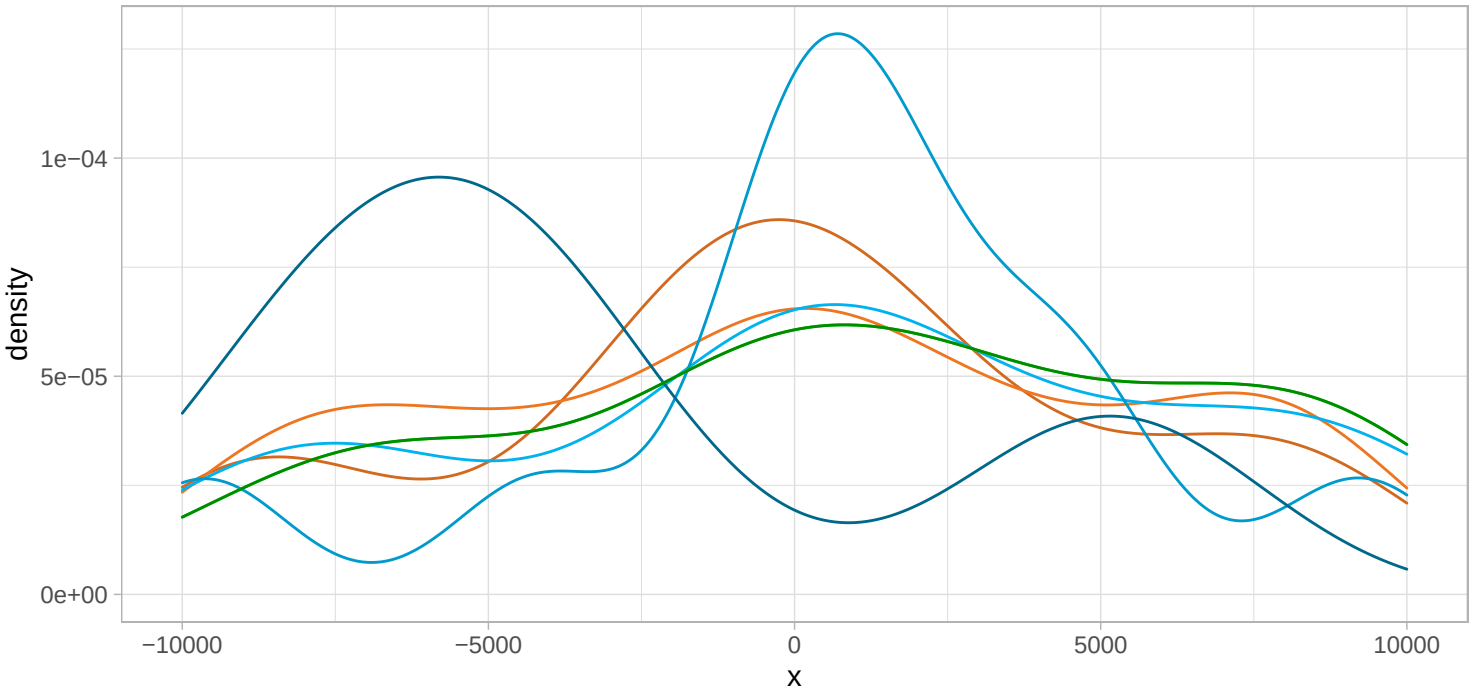
Differentially methylated regions

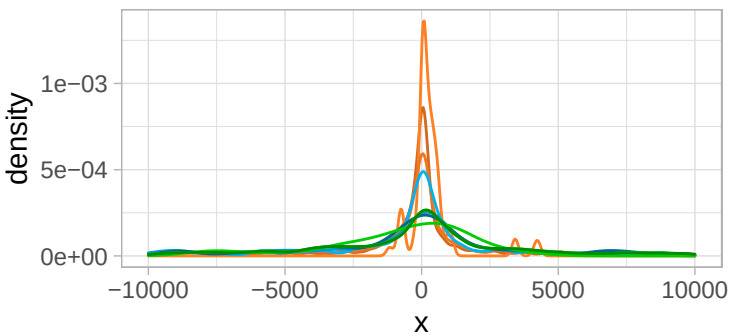


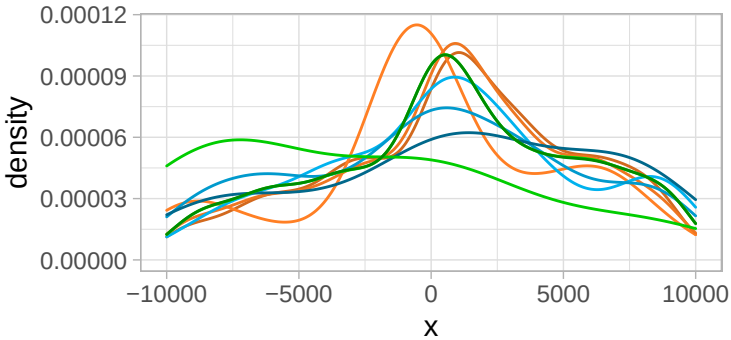
UCSC_RefGene_Group

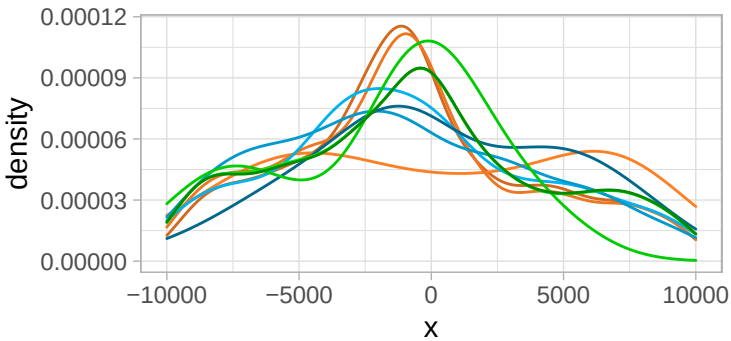

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY

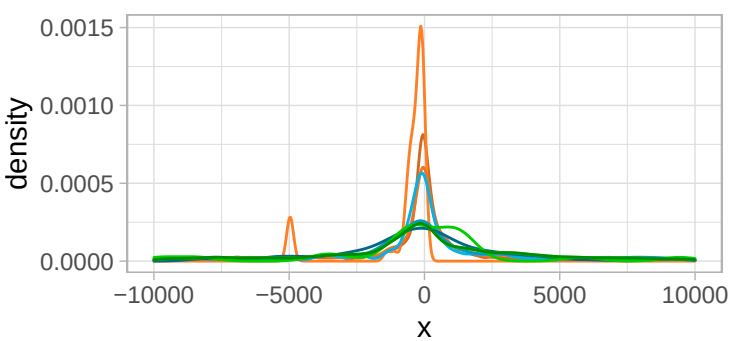

density
9e−04
6e−04
3e−04
0e+00
−10000
−5000
0
5000
10000
x


density
1e−04
5e−05
0e+00
−10000
−5000
0
5000
10000
x


density
1e−03
5e−04
0e+00
−10000
−5000
0
5000
10000
x


density
0.00012
0.00009
0.00006
0.00003
0.00000
−10000
−5000
0
5000
10000
x


density
0.00012
0.00009
0.00006
0.00003
0.00000
−10000
−5000
0
5000
10000
x


density
0.0015
0.0010
0.0005
0.0000
−10000
−5000
0
5000
10000
x

GO terms


positive regulation of locomotion
regulation of locomotion
actin filament bundle
positive regulation of cell migration
regulation of cell migration
regulation of cell adhesion
regulation of signaling
regulation of cell communication
regulation of signal transduction
cell surface receptor signaling pathway
signal transduction
cell communication
cell−substrate junction assembly
adherens junction
signaling receptor binding
stress fiber
regulation of transport
membrane part
response to alkaloid
regulation of cell killing
cellular defense response
cytokine production involved in immune response
T cell activation involved in immune response
cell adhesion...
epithelium development
positive regulation of cell differentiation
movement of cell or subcellular component...
regulation of cell proliferation
multicellular organismal process
regulation of anatomical structure morphogenesis
tissue development
anatomical structure morphogenesis
cell proliferation
multicellular organism development
localization of cell
cell motility
signaling
cell migration
regulation of multicellular organismal development
cell−cell adhesion
positive regulation of developmental process
regulation of developmental process
cellular developmental process
anatomical structure development
animal organ development
regulation of cell differentiation
plasma membrane part
developmental process
cell differentiation
positive regulation of cell proliferation
cell morphogenesis
cell periphery
plasma membrane...
regulation of multicellular organismal process
system development
locomotion
biological adhesion
cell adhesion
movement of cell or subcellular component
plasma membrane