VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

Daugiamačių biomedicinos duomenų atvaizdavimo metodai

Multidimensional biomedical data plotting methods

Kursinis darbas

Atliko: 3 kursas, 1 grupė, Bioinformatika
Asta Kvedaraitė

Darbo vadovas: Karolis Koncevičius

Vilnius - 2020

Turinys

Įvadas

Biomedicina yra biologijos mokslų sritis, kuri nagrinėja gyvąsias sistemas. Tokie daugiamačiai tyrimų duomenys generuojami dideliais kiekiais, pvz. genomui tirti. Taip pat galima sudaryti daugiamačius duomenų rinkinius analizuojant įvairias ligas, jų progreso eigą, šalutinius poveikius. Tačiau analizuoti duomenis be vizualios informacijos ne visada pavyksta, o kartais tiesiog vizuali informacija lengviau interpretuojama. Ypač sunku analizuoti daugiamačius duomenis, todėl naudojamos diagramos, grafikai, schemos ir kiti atvaizdavimo metodai.

Darbo tikslas - išbandyti kelis daugiamačius duomenų atvaizdavimo metodus analizuojant realius epigenetinius DNR metilinimo duomenis, bei pateikti apžvelgtų metodų naudojimo epigenetiniams duomenims išvadas bei rekomendacijas.

Tikslui pasiekti išsikeliami šie uždaviniai:

1. Susirasti ir darbui paruošti realius DNR metilinimo duomenis.
2. Atliekant literatūros apžvalgą atsirinkti 6 daugiamačių duomenų atvaizdavimui tinkančius metodus.
3. Pasirinktus atvaizdavimo metodus pritaikyti pasiruoštiems DNR metilinimo duomenims.

Šiame darbe aptariami DNR metilinimo duomenų vizualizavimo sunkumai ir pateikiami duomenų atvaizdavimo pavyzdžiai pritaikyti realiems duomenims, taip pat atvaizdavimo rekomendacijos.

1. Epigenetika

Epigenetika - mokslas, tiriantis veiksnius, kurie reguliuoja genų raišką nekeisdami deoksiribonukleorūgšties, arba kitaip DNR, sekos. DNR modifikacijos, tokios kaip citozino arba adenino metilinimas, citozino hidroksilinimas, priklauso epigenetiniams veiksniams. Histonų modifikacijos - acetilinimas, metilinimas, fosforilinimas ir kita, taip pat trumposios ir ilgosios baltymo nekoduojančios reguliacinės RNR, priklauso epigenetiniams veiksniams, kurie atlieka pagrindinį vaidmenį ląstelės reguliatoriniuose procesuose. Jie keičia chromatino kompaktiškumą, kartu ir genų raišką (transkripciją), todėl siejami su žmogaus ligų atsiradimu. Vienas iš geriausiai ištirtų epigenetinių mechanizmų yra DNR metilinimas CpG sekose suformuojant 5-metilcitoziną [1].

1.1. DNR metilinimas

DNR metilinimo metu modifikuojami CG dinukleotidai, prie jų pridedama metilo grupė (CH_3) kovalentiniu būdu. DNR metilinimas yra katalizuojamas DNR metiltransferazės fermentų, kurie pakeičia penktą anglies atomą citozino žiede, CG nukreipti dinukleotidai iš 5’ į 3’ dar vadinami CpG vietomis [3].

Tokia modifikacija yra paplitusi žinduolių genome ir iš visų citozinų esančių CpG sekose, tik 60%-80% yra metilinti. Taip pat žinduolių genome randama daugiau nei viena citozino modifikacija. Norint išsiaiškinti citozino vaidmenį epigenetiniame reguliavime, reikia naudoti tam specialius metodus ir įrankius, kurie leidžia atskirti 5-metilcitozino modifikaciją. Tokioms modifikacijoms tinkami įvairūs metodai, kaip mikrogardelės ar sekoskaita [2].

1.2. Metilinimo mikrogardelės

DNR mikrogardeles sudaro DNR sekos (zondai) kurios yra užneštos ant kietos medžiagos nešiklio (pavyzdžiui stiklo) paviršiuje tiksliai nustatytose vietose esančių mikroskopinių mažų taškelių. Kai randama DNR sudėtyje sekų, atitinkančių mikrogardelės zondus, jie hibridizuojasi t.y. susiriša komplementarūs DNR regionai, tačiau niekur neprisijungę DNR fragmentai yra nuplaunami. Šis rišimasis yra išmatuojamas kompiuterio valdomo nuskaitymo metu ir yra vertinamas kaip teigiamas signalas [4].

DNR metilinimo mikrogardelės padeda nustatyti metilintus arba galimai metilintus regionus, CpG salas. Tam naudojamas Illumina 450k infinium metilinimo BeadChip, kuris apima 96% CpG salų, > 99% promotorių, metilintas vietas, kurios nėra CpG arba miRNR promotorius[3,5]. Tačiau DNR metilinimas nėra tikslus, kadangi jam turi įtakos tokie veiksniai kaip senėjimas, rūkymas, onkologinės ligos ir pan.

1.3. Epigenetiniai ligų tyrimai

Epigenetiškai sukelti genų ekspresijos pokyčiai yra susiję su įvairiomis vėžio formomis kaip žarnyno, plaučių, prostatos, gimdos, kepenų ir kitomis. DNR metilinimo lygiai, tokie kaip viršsvoris, rūkymas ir senatvė, veikia faktorius, kurie padidina vėžio susirgimo riziką. Epigenetika taip pat paveikia autoimunines ligas, tokias kaip reumatoidinis artritas, neurodegeneracines ligas kaip Alzheimerį, Parkinsono, Hantingtono ligas. Epigenetiniai pokyčiai buvo pastebėti ir turint intelektualumo negalias, kurios pasireiškia Fragilios X chromosomos, Retto, Angelmano sindromuose. Taip pat pasireiškia ligose kaip depresija, bipolinis sutrikimas ir šizofrenija [10].

Epigenetiniai modeliai gali būti naudojami kaip biomarkeriai ligų nustatymui, ankstyvų stadijų identifikavimui pacientams, kurie yra rizikos grupėje. Taip pat nustatyti patikimumą esantiems gydymo metodams ir stebėti jų efektyvumą.

2.Biomedicininių duomenų atvaizdavimo metodai

2.1. Dendrograma

Dendrograma, yra diagrama, kuri parodo hierarchinius ryšius tarp objektų, pavyzdžiui tarp koreliacinių metilinimo profilių. Individualūs metilinimo profiliai sudėliotos dendrogramos apačioje ir vadinamos vaikiniais elementais, o paskutiniai elementai - lapais. Vaikinius elementus sujungia tėviniai elementai(viršūnės). Metilinimo profilių grupės formuojamos sujungiant individualias metilinimo profilius arba egzistuojančias jų grupes į vieną tašką - viršūnę. Vertikali ašis parodo suskaičiuotą atstumą tarp metilinimo profilių arba jų grupių. Viršūnės aukštis nurodo atstumą tarp kairios ir dešiniosios vaikinių šakų grupių [7,11]. Atstumo skaičiavimui tarp dviejų grupių skaičiuojama šia formule :

$$D = 1 - C$$

Kur D reiškia atstumą, o C - koreliaciją tarp komponenčių grupių [8].
Jei komponentės koreliuoja, jos turės koreliacinę reikšmę arti vieneto, todėl D = 1 - C turės reikšmę arti nulio. Jei grupės koreliuoja, jos bus dendrogramos apačioje. Nekoreliuojančios grupės turės reikšmę 0, o neigiama koreliacija turi reikšmę lygi -1 (C = -1, o D = 2). Tačiau reikia atkreipti dėmesį į metrinę sistemą, kuri naudojama atstumų skaičiavimams ir klasterizavimui.

Dendrograma turi keletą variacijų - gali būti horizontali, vertikali, linijinė arba apvali. Apvalios dendrogramos versija išnaudoja grafinę erdvę efektyvesniam

atvaizdavimui. Jei reikia nurodyti ilgus mėginių pavadinimus, geriausia naudoti horizontalią dendrogramos versiją.

2.2. Paralelinės koordinatės

Paralelinių koordinačių grafikas palygina kokį nors bruožą, kurį turi stebimi individai, tam naudojama skaitinių reikšmių aibė. Vertikalūs stulpeliai nurodo kintamuosius ir dažniausiai turi atskirus mastelius, pagal duotas reikšmės. Kintamasis turi reikšmių rinkinį, kuris vaizduojama Y ašyje. Reikšmės yra sujungiamos pagal matricoje esančią eilutę kiekvienam mėginiui - stulpeliui [12].

Šio grafiko privalumas - stulpeliai gali žymėti skirtingus požymius, jei jie turi skaitines reikšmes, taip pat gali rodyti skirtingus diapazonus. Tačiau pateikiant per daug mėginių, jų grafikas bus neįskaitomas dėl per didelio tankio.

2.3. Medžio schema

Medžio schema (angl. *Treemap*) pavaizduoja hierarchinį duomenų rinkinį stačiakampiuose, kurie yra suglausti vienas prie kito, taip išnaudojant visą vietą. Kiekvieno stačiakampio dydis priklauso nuo pateiktų reikšmių kiekio. Naudojant spalvas ir interaktyvumą galima pateikti pagal grupes, subgrupes ir panašiai skaidant grupę į kelias subgrupes [13].

Medžio schema yra naudojama norint parodyti keletą požymių tame pačiame grafike:

- Parodyti grupių pasiskirstymą ir nustatyti kuris mėginys yra svarbiausias, kaip kiti mėginiai pasiskirstę.
- Parodo hierarchijos pasiskirstymą.

Grafikas gali būti naudojamas dideliems duomenų rinkiniams, taip pat be hierarchinio duomenų pasiskirstymo - tiesiog parodyti kelių mėginių reikšmes.

2.4. Veno diagrama

Veno diagrama taip pat vadinama pirmine diagrama, rinkinio diagrama arba logine diagrama. Kiekviena aibė turi būti baigtinė, kuri nubrėžiama apskritimu. Nubrėžti aibių apskritimai dažniausiai kertasi ir susikirtimo dydis priklauso nuo abiejų grupių. Veno diagramą geriausia naudoti 2 ar 3 grupėms, kadangi taikant diagramą daugiau nei trims grupėms, jį tiesiog sunkiau analizuoti [14].

Naudojant dvi grupes, Veno diagrama nubrėžia apskritimus pagal reikšmių dydžius.

2.5. Korelograma

Korelograma vadinama koreliacinė matrica, kuri leidžia analizuoti ryšius tarp kiekvienos poros skaitinės reikšmės. Ryšiai tarp porų yra atvaizduojami taškiniu grafiku, arba simboliu, kuris nurodo koreliaciją. Diagramos įstrižainė atvaizduoja kiekvienos reikšmės pasiskirstymą histogramos pavidalu arba tankio plotu [15].

Korelograma naudinga norint gauti tiriamąją analizę, kuri nurodo viso duomenų rinkinio reikšmių ryšius vienu bendru vaizdu. Kiekvienam taškiniam grafikui, esančiam korelogramos viduje, galima pridėti linijinę regresiją. Taip pat galima atvaizduoti grupes spalvomis, jei reikšmės turi kategoriją (pvz.: sergantys/nesergantys).

Tačiau reikia atsižvelgti, kad atvaizduojant ryšius tarp 10 tiriamųjų objektų ar daugiau, korelogramoje esantys taškiniai grafikai yra sunkiau suprantami.

2.6. Genominių savybių persidengimo atvaizdavimas

Genominių savybių persidengimams atvaizduoti naudojamas “kpPlotCoverage” funkcija, kuri labai panaši į tankio grafiką. Pagrindinis skirtumas yra atvaizdavimas - tankio grafike vaizduojama specifinė vieta, o persidengimo grafike vaizduojami kiekvienos bazių poros pasirinktos savybės persidengimai genome [16].

3.Duomenų rinkinio paruošimas

3.1. DNR metilinimo duomenų paieška

Tiriamieji duomenys yra integruoti DNR metilinimai ir genų raiškos skirtinguose smegenų regionuose tiriant Alzheimerio ligą. Eksperimento tipas buvo metilinimo analizavimas naudojant mikrogardeles. Duomenys sudaryti iš 420 852 DNR metilintų vietų. viso 269 mėginiai, iš kurių 49 sveiki asmenys ir 24 sergantys Alzheimerio liga. Iš kiekvieno asmens buvo paimti duomenys iš keturių smegenų regionų:

- dorsolateralinės priekinės žievės (DLPFC)
- entorinalinės žievės (ERC)
- hipokampo (HIPPO)
- smegenų žievės (CRB).

Duomenys atsisiųsti iš NCBI GEO, kurio ID: GSE125895, idat formatu [6].

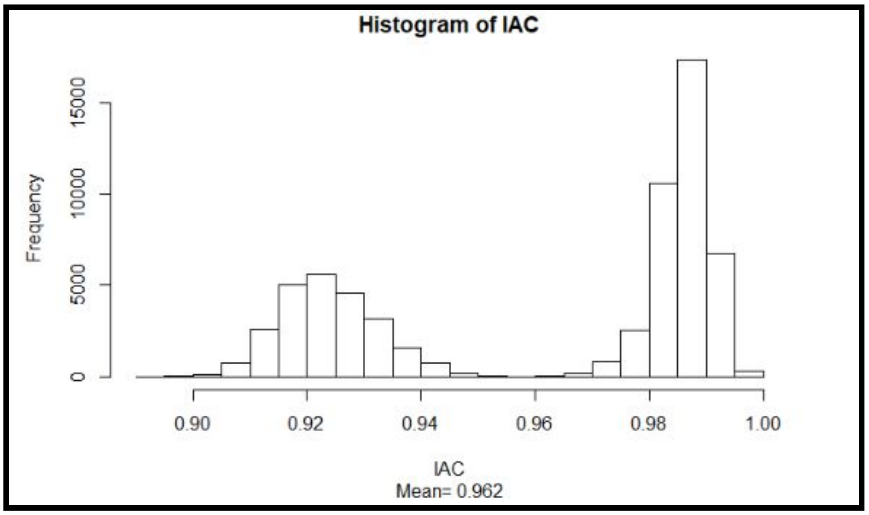
3.2. Duomenų normalizavimas

Atsiųsti duomenys turėjo būti pakeisti į matricą, duomenų tyrimui atlikti. Tam buvo naudota minfi [20] paketo funkcija “read.metharray.exp”, kuri nuskaito metilinimo duomenis idat failų formate ir kiekvienam matavimui nustato signalo patikimumo p-reikšmes, taip pat naudojant minfi paketo funkciją. Tačiau vien pasiversti į matricą neužtenka, duomenys turi būti normalizuoti - buvo naudojama “preprocessSWAN” funkcija. Prieš ir po normalizavimo buvo pasiriknta išmesti visas pozicijas, kurios turi daugiau nei 1% nepatikimų p-reikšmių. Taip pat tyrimui nereikalingos “CH” pozicijos ir tos pozicijos, kurios yra šalia DNR polimorfizmų (SNP). Čia panaudota minfi paketo “dropMethylationLoci” funkcija, kuriai reikia nurodyti prieš tai gautą SWAN duomenų rinkinį, “dropRS = TRUE”, kuris nurodo, kad bus išmesti polimorfizmai. Paskutinis argumentas “dropCH = TRUE”, kuris nurodo išmesti visus “CH”.

3.3. Išskirčių pašalinimas

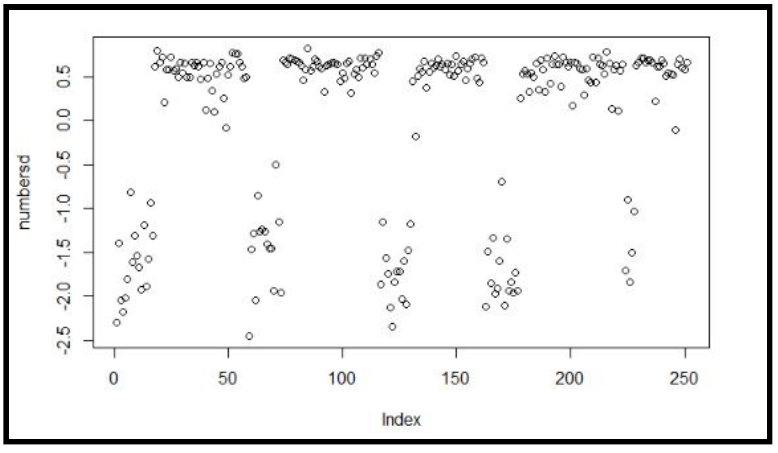
Po duomenų normalizavimo, gauti duomenys nėra tinkami tyrimui, reikia išmesti išskirtis. Šiam procesui naudojamas išskirčių šalinimo metodas, dar vadinamas IAC [17]. Normalizuotą matricą turi būti perduota kaip parametras “cor” funkcijai. Tuomet reikalinga nubraižyti grafiką apžiūrai, patogiausia su “hist” funkcija. Paveikslėlyje pavaizduota histograma (pav. 1) parodo, kad duomenų rinkinys turi reikšmes intervale [0.9 ,1.00]. Reikšmių dažnio intervalas yra nuo 0 iki 15 000. Taip pat nurodytas vidurkis, kuris yra 0.962. Įsitikinimui, darytas kitas atvaizdavimo būdas (pav. 2), kurio metu surandami stulpelių reikšmių vidurkiai, randamas standartinis

nuokrypis naudojant funkciją “sd”. Sekančiame žingsnyje rasti standartinio nuokrypio duomenys, naudojami atvaizdavime. Pagrindinis duomenų rinkinys gautas atrinkus visus stulpelius, kurių standartinis nuokrypis didesnis nei riba. Šį žingsnį reikia kartoti keletą kartų. Šio rinkinio atveju užtenka dviejų iteracijų - atvaizdavus atrinktus duomenis matome, kad jie išskirčių neturi, kadangi abiejuose atvaizduoti duomenys nepasikeičia.



1 pav. Histograma, po išskirčių šalinimo

Metilinimo reikšmių koreliacijos dažnių pasiskirstymas ir vidurkis. Histogramoje duomenys pasiskirstę dviejose vietose, tačiau tai nėra išskirstys.



2 pav. Standartinių nuokrypių pasiskirstymas

Duomenų pasiskirstymas be išskirčių. Duomenys nesiekia ribos, todėl grafike jos nerodoma.

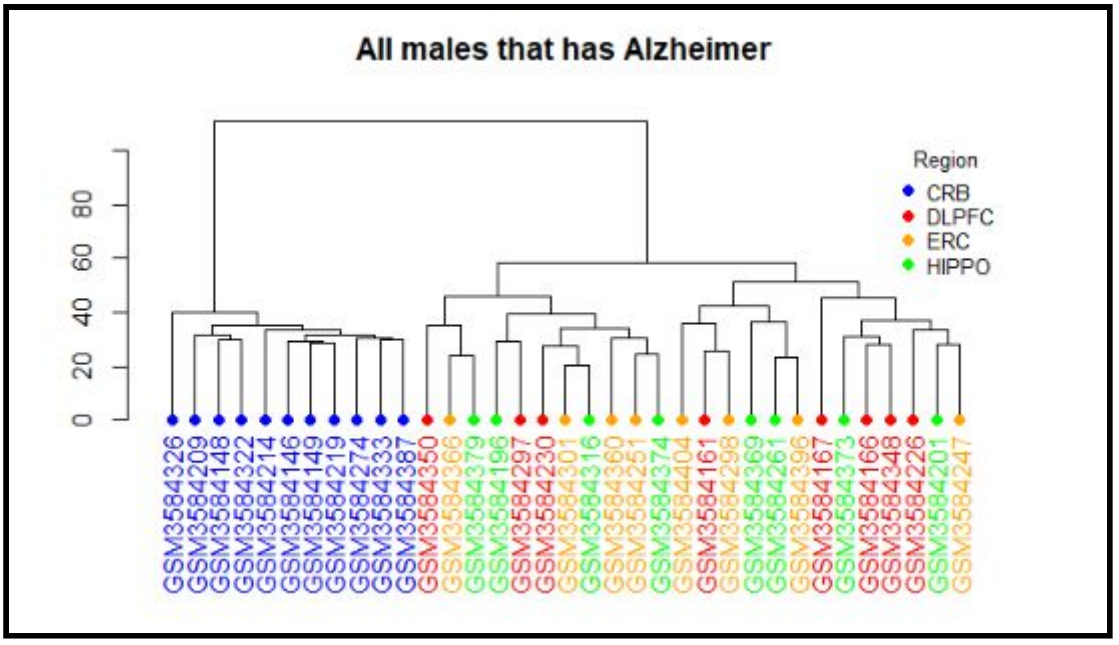
Po normalizavimo ir išskirčių pašalinimo duomenų rinkinio matrica turi 251 stulpelį ir 468 507 eilutes.

4. Duomenų rinkiniui pritaikyti atvaizdavimo metodai

4.1. Dendrograma

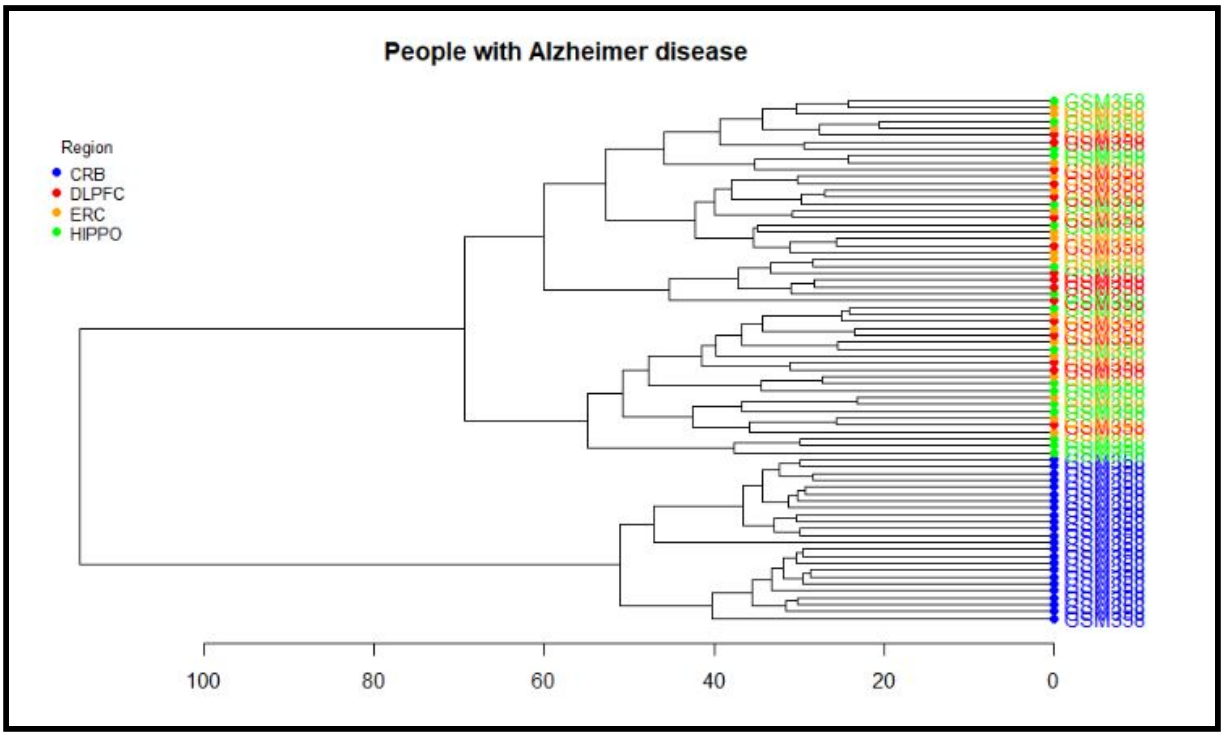
Dendrograma naudojama vizualiam hierarchiniam klasterizavimui - išskirti objektus į grupes. Dendrogramos analizavimui (pav. 3) buvo atrinkta tik dalis mėginių - sergantys Alzheimerio liga ir yra vyriškos lyties. Sekančios dendrogramos (pav. 4) analizavimui naudoti visi sergančių asmenų mėginiai. Turint mėginius su ilgais pavadinimais, rekomenduojama dendrogramą atvaizduoti horizontaliai (pav. 4), kai atstumai yra X ašyje, o mėginių pavadinimai Y ašyje.

Gautam duomenų rinkiniui pritaikytas hierarchinis klasterizavimas, o atstumas gautas euklidiniu metodu naudojant “dist” funkciją. Taip pat naudota “dendroapply” funkcija, kuri leidžia pritaikyti sukurtą funkciją kiekvienam dendrogramos išsišakojimui. Lapai - paskutiniai elementai nuspalvinti atitinkamom spalvom pagal audinio regioną, kurie nurodyti legendoje. Trečiame paveikslėlyje CRB arba smegenų žievė pavaizduota kaip atskira grupė, o kairesnioji grupė turi maišytas vidines grupes. Ketvirtame paveikslėlyje vyrų ir moterų CRB regionas taip pat pavaizduotas kaip atskira grupė.



3 pav. Vyrų, kurie serga Alzheimeriu, smegenų regionų pasiskirstymas

Mažas duomenų kiekis aiškiau vaizduojamas grafike, tiriant tik smegenų regionų metilinimo reikšmes vyrams. Dendrogramoje vaizduojamos dvi pagrindinės grupės, tai CRB regiono, kurio atstumai panašūs tik tame regione. Kiti regionai nebeatsiskiria vienas nuo kito.



4 pav. Sergančių asmenų smegenų regionų pasiskirstymas

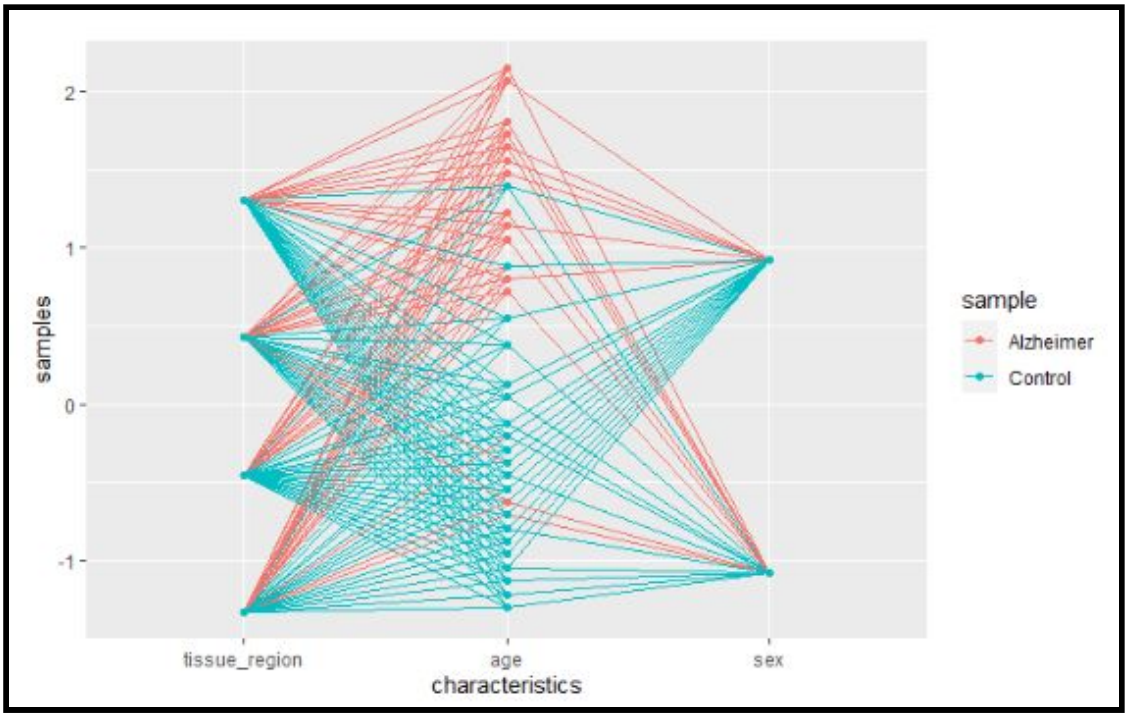
Vyrų ir moterų, sergančių Alzheimeriu, pavaizduoti regionai, taip pat turi dvi pagrindines grupes, CRB regiono ir kitų, kurie nėra grupuojami pagal konkretų regioną detaliau.

4.2. Paralelinės koordinatės

Paralelinės koordinatės grafikas naudojamas charakteristikų ir jų ryšių palyginimui. Galima palyginti skirtingus objektus, kurie turi tą pačią savybę. Penkto paveikslėlio paralelinės koordinatės sujungtos taškais, nurodo stulpeliuose esančias mėginių charakteristikas - audinio regioną, amžių ir lytį. Kiekviena eilutė yra sergantis arba sveikas mėginys, pagal kurį atitinkamai nuspalvinama jungiamoji linija. Čia nenaudojami daugiamačiai duomenys, tik “samplekey” rinkinys bendram vaizdui susidaryti (pav. 5), todėl vizualiai galima matyti, jog daugiausia Alzheimeriu sergantys asmenys yra vyresnio amžiaus tiriamieji asmenys. Tačiau kitame grafike (pav. 6) panaudoti daugiamačiai duomenys.

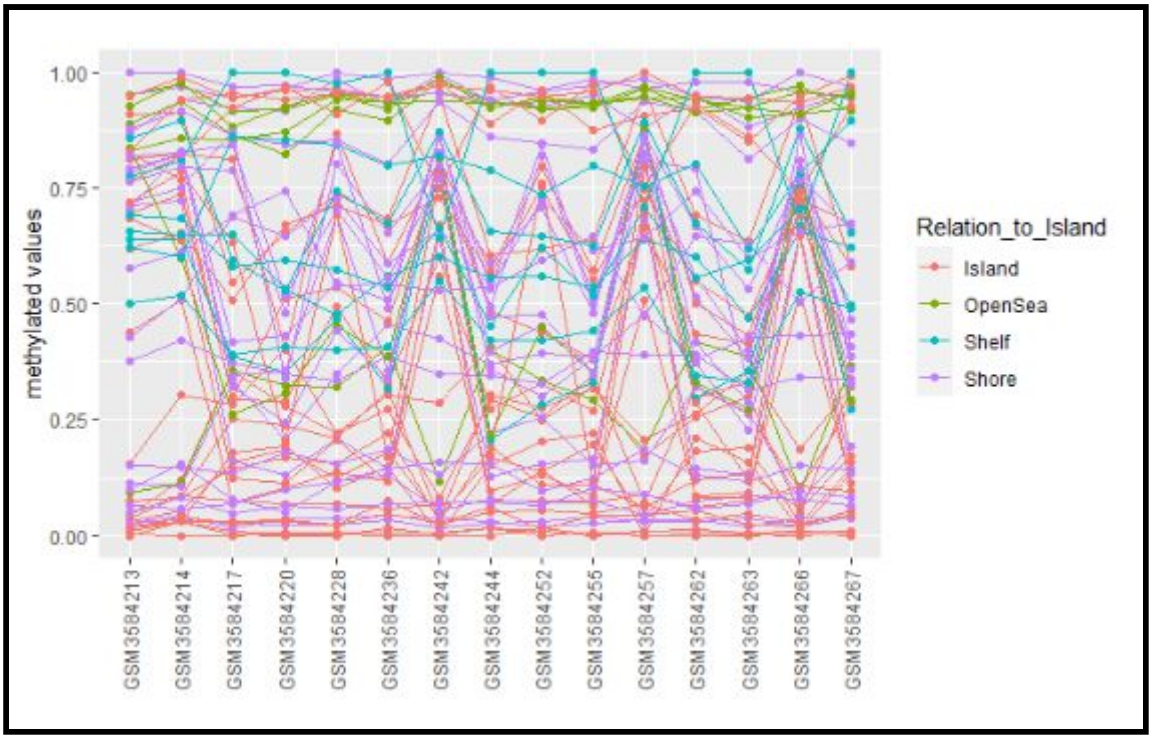
Vietoje charakteristikos X ašyje, galima naudoti mėginius, o Y ašyje eilutes - pozicijas. Kadangi paralelinių koordinačių grafiką patogiau stebėti kai yra nedaug duomenų, tai atrinkti tik tie mėginiai, kurie serga Alzheimerio liga, yra Afro-Amerikiečių kilmės, taip pat antros mėginio lėkštelės - tokių gauta 15 mėginių. Pozicijų taip pat žymiai per daug, todėl naudotos Y chromosomos pozicijos. Gauti duomenys pritaikyti antrojo paveikslėlio grafike. Jungiamosios linijos nuspalvintos atitinkamai pagal CpG salas kiekvienai pozicijai naudotos keturi regionai - sala, šiaurinis ir pietinis rifai kartu, jūra, pietinis ir šiaurinis krantai taip pat kartu (pav. 6). Tačiau linijų gan nemažai, todėl tinkamas bendram vaizdui susidaryti. Matosi, kad

salų regionus turinčios pozicijos yra apačioje - mažos reikšmės. Rifo ir kranto regionai pasiskirstę per vidurį grafiko, o jūra daugiau viršuje. Turint mažiau pozicijų, lengviau analizuoti kiekvienos salos regiono linijas (pav. 7).



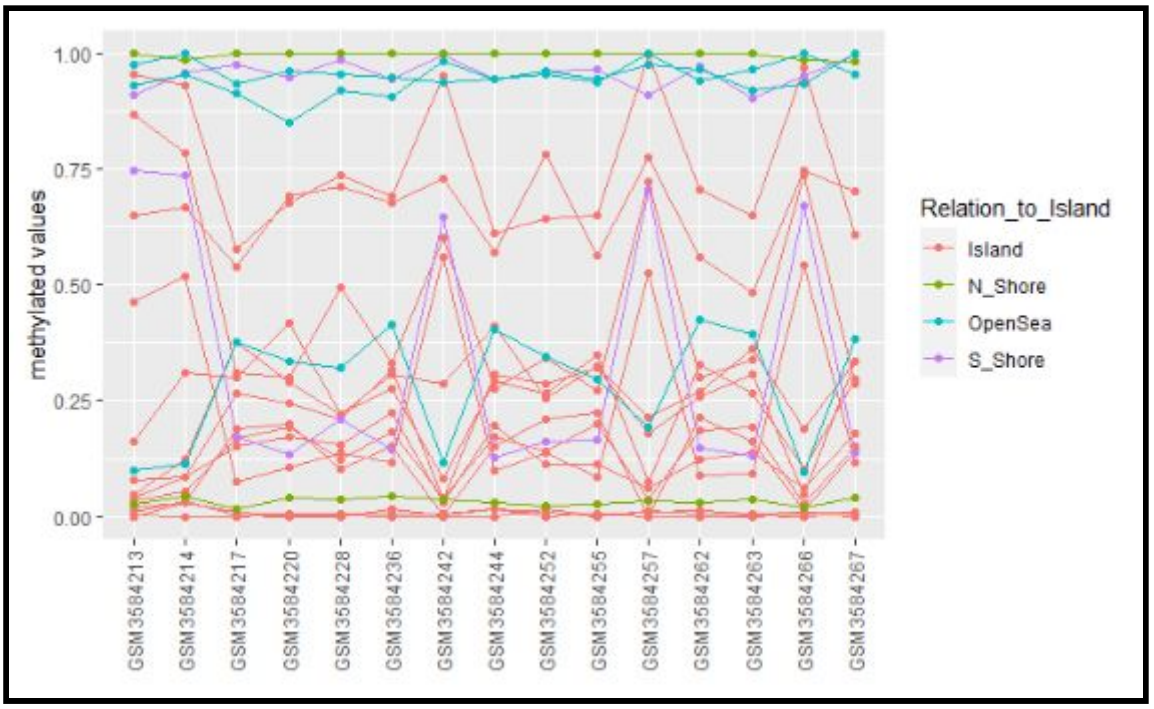
5 pav. Mėginių charakteristikų sujungimas

Kiekvienas regionas turi sergančių mėginių, bet daugiausia jų vyresnio amžiaus pacientai, tiek vyrai tiek moterys.



6 pav. Pozicijų metilintų reikšmių atvaizdavimas, pagal CpG salų regionus

CpG salos vaizduojamos atitinkamomis spalvomis, kiekvienai Y chromosomos pozicijai, kurios turi mėginių metilintas reikšmes nuo 0 iki 1 ir yra sujungtos linijomis. CpG salų regionai daugiausia apačioje, rifas ir krantai per vidurį o jūra viršuje, reikšmių atžvilgiu.



7 pav. Pozicijų metilintų reikšmių sujungimas

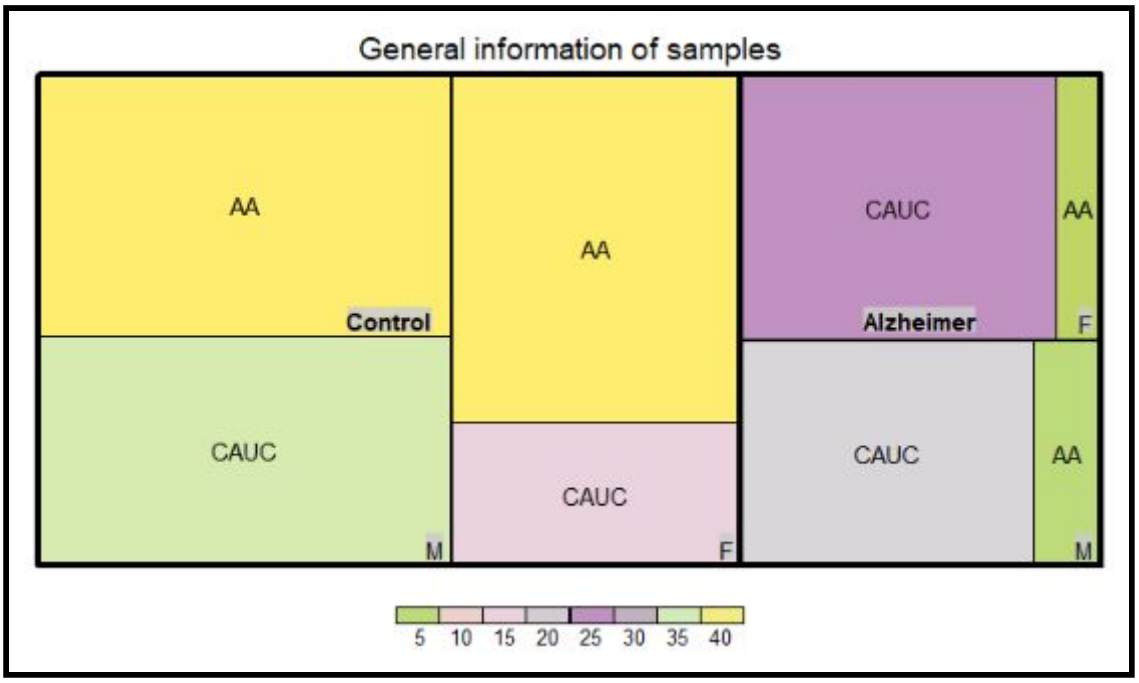
Naudojant mažiau pozicijų, t.y. pirmas 20, aiškiau matosi linijų sujungimai ir regionų panašumai mėginiuose, pasiskirstymas. Nemažai salų regionų (angl. *Island*) pasiskirsčiusių visame diapazone, nei jūros regiono, kurio šiuo atveju matosi mažiau.

4.3. Medžio schema

Medžio schema naudojama hierarchinių duomenų vizualizavimui medžio tipo struktūra. Duomenys organizuojami į šakas ir jų vaikines šakas, kurie vaizduojami stačiakampiuose.

Šiam grafikui reikia vienos reikšmių eilutės, todėl naudojamas mėginio amžius ir skirstomas į grupes serga/neserga, lyties subgrupę ir grupuojama pagal asmens kilmę. Medžio struktūra sudaryta panašiai kaip dendrogramoje iš viršūnių ir lapų. Šio duomenų rinkinio atveju bendra viršūnė sudaryta iš dviejų vaikinių elementų, t.y. elemento grupė yra sergančių asmenų mėginiai, o kita nesergančių asmenų mėginiai. Abi grupės skirstomos į dar daugiau elementų. Gauta sergančių asmenų grupė suskirstyta į pagal lytį ir vėliau kilmę. Vyriškos lyties grupė taip pat suskaidoma atitinkamai.

Šiame grafike naudojami duomenys yra tik iš “samplekey”, tačiau grafikas pateikia bendrą informaciją apie jame esančius duomenis, kokie yra skirtumai tarp mėginių grupių.

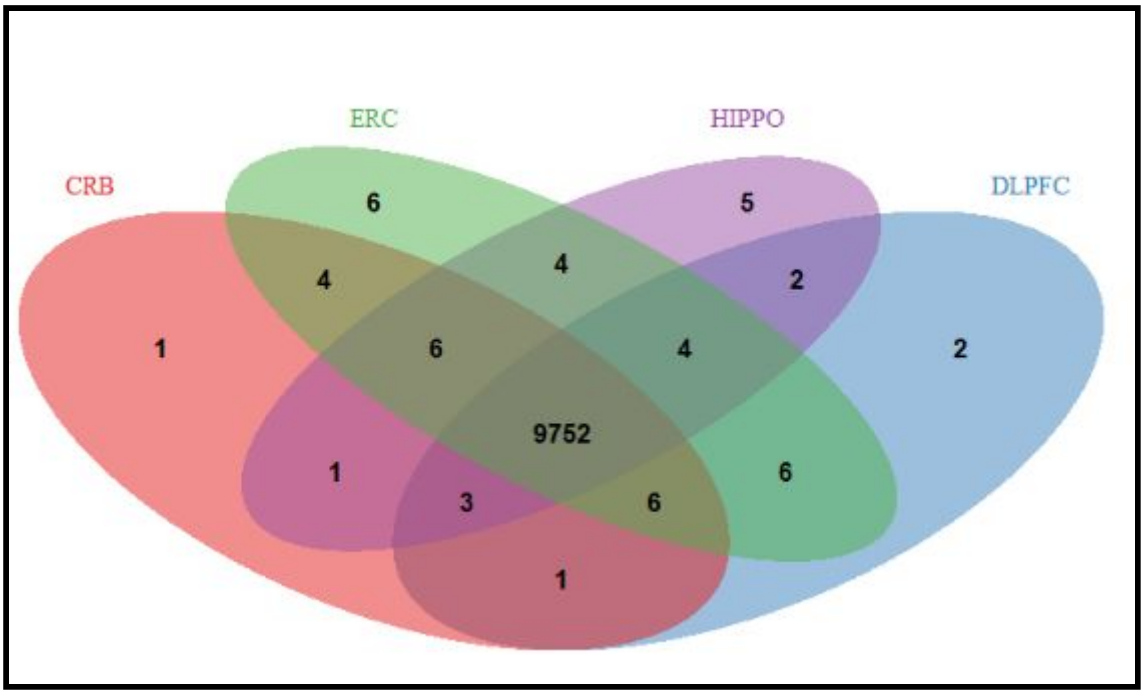


8 pav. Sergančių ir nesergančių mėginių grupavimas pagal jų kiekį

Schema parodo grupes ir jų skaidymą. Matomi duomenų rinkinyje esantys skirtumai - sergančių asmenų daugiau. Sveikų mėginių grupėje, tiriamų moterų kilmė daugiausia Afro-Amerikiečių.Tačiau sergančiųjų grupėje daugiausia Kaukazo tautybės asmenys ir tik mažuma Afro-Amerikiečių. Legenda, pagal spalvas, nurodo mėginių kiekį tam tikroje grupėje.

4.4. Veno diagrama

Veno diagrama parodo visus galimus ryšius tarp keletos dar daugiau duomenų rinkinių. Šiai diagramai naudoti skirtingų smegenų regionai - keturios grupės. Kiekviena grupė turi mėginius pagal smegenų regioną, o tie mėginiai turi pozicijas su metilintomis reikšmėmis. Norint rasti panašumus, galima naudoti eilučių suapvalintus vidurkiai, kiekvienai grupei. Suapvalinus vidurkius tik vienu skaičiumi po kablelio, visi regionai turėjo tokius pat rezultatus, todėl skaičius buvo padidintas iki 4 ženklų po kablelio. Tačiau visi regionai turi didžiąją dalį panašumų - 9752 reikšmes turi vienodas reikšmes ir tik keletą skirtingų. Kiekvienas audinio regiono duomenų rinkinys sudarytas iš metilintų reikšmių masyvo. Lyginant tik dvi grupes, pavaizduoto apskritimo dydžiai priklauso nuo jų nepanašių reikšmių ar savybių skaičiaus. Veno diagramoje sutampančias vietas galima vaizduoti procentais vietoje sutampančių savybių skaičiaus.

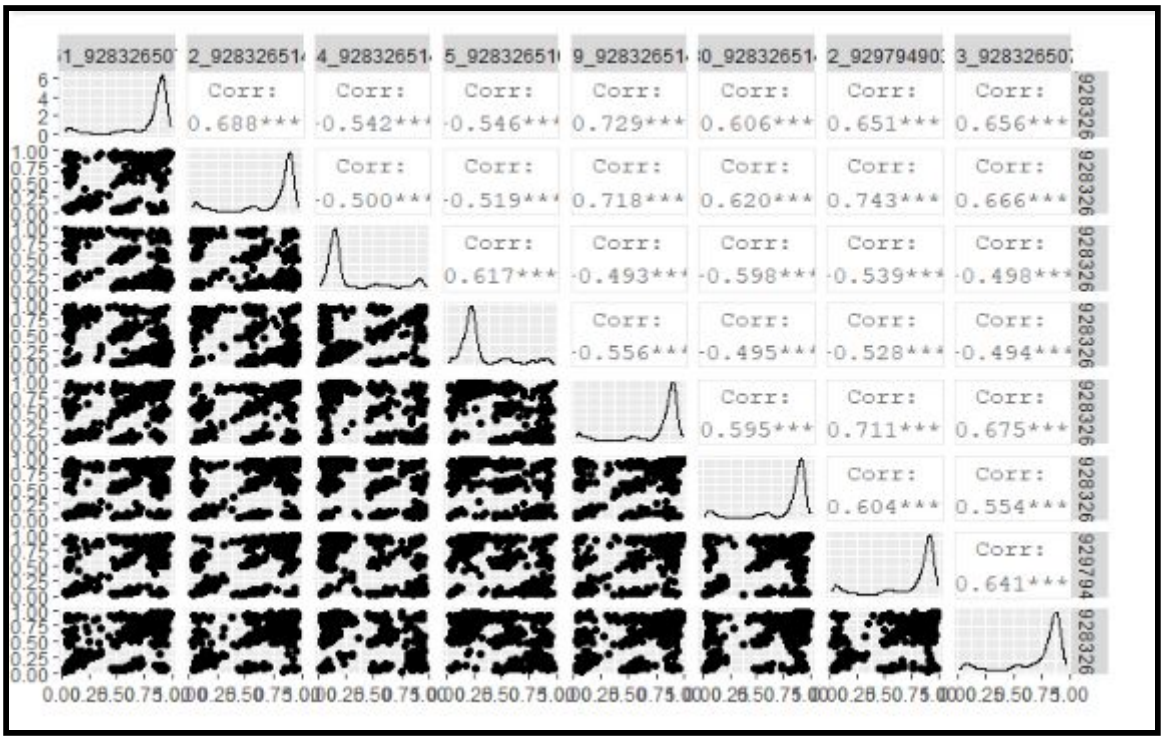


9 pav. Kiekvienam smegenų regionui išvestų metilinimo pozicijų vidurkių sutapimai

Tiriamos 4 smegenų grupės, kurios turi metilintas reikšmes, rastas pagal mėginį, kurie priskirti toms grupėms. Panašumų ir skirtumų vizualizavimui naudoti metilintų reikšmių vidurkiai kiekvienai eilutei, pagal grupę. Vidurkis buvo rastas 4 ženklų po kablelio tikslumu. Diagrama parodo, kad dauguma reikšmių yra visose smegenų regionų grupėse ir tik mažuma turi nesutampančius reikšmių vidurkius.

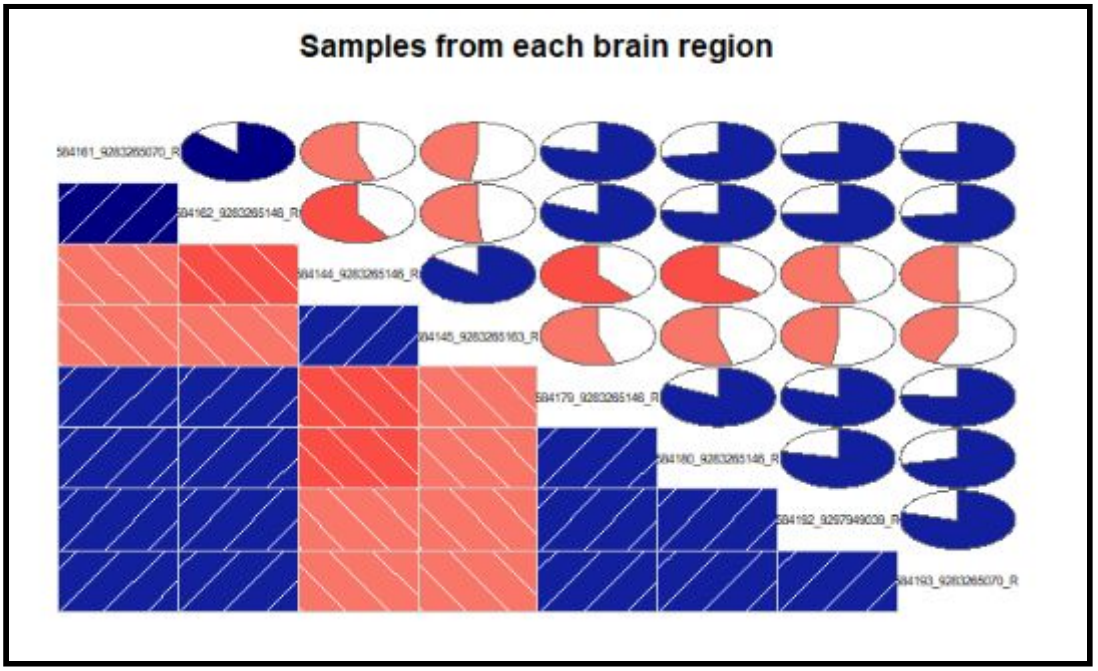
4.5. Korelograma

Korelograma leidžia analizuoti ryšius tarp porose esančių skaitinių reikšmių. Korelogramą galima atvaizduoti įvairiais būdais - ggpairs, corrgram, corrplot ir pan. Funkcija ggpairs reikalauja paketo “GGally”. Bandymas pavaizduotas paveiksle 10. Čia naudota 8 kiekvieno regiono pirmi du mėginiai ir 1000 pačių variabiliausių pozicijų. Korelogramoje pavaizduoda dalis koreliacijų taškiniu būdu, o viršutinėje dalyje parašyta koreliacija. Kitai funkcijai naudota duomenų matrica 10 000 pačių variabiliausių pozicijų ir 8 stulpeliai - iš kiekvieno smegenų regiono po 2 pirmus mėginius. Šie duomenys naudoti funkcijai - “corrgram”. Vienuoliktame paveiksle, puse grafo yra taškų koreliacija o kita pusė - pyrago diagrama, įstrižainėse pažymėti mėginiai. Raudona spalva žymi neigiamas koreliacijos reikšmes, o mėlynos - teigiamas. Jei spalva yra tamsesnė, vadinasi koreliacijų atstumas yra didesnis ir atitinkamai šviesesnės spalvos žymi mažesnius atstumus.



10 pav. Koreliacijos, naudojant ggpairs, korelograma

Naudota mažiau pozicijų (1000 pačių variabiliausių pozicijų), tačiau tarp jų koreliacija matosi, bet neaiškiai. Taškinės grafo dalys panašios viena į kitas. Įstrižainės parodo, kad 3 ir 4 mėginiai skiriasi nuo kitų.



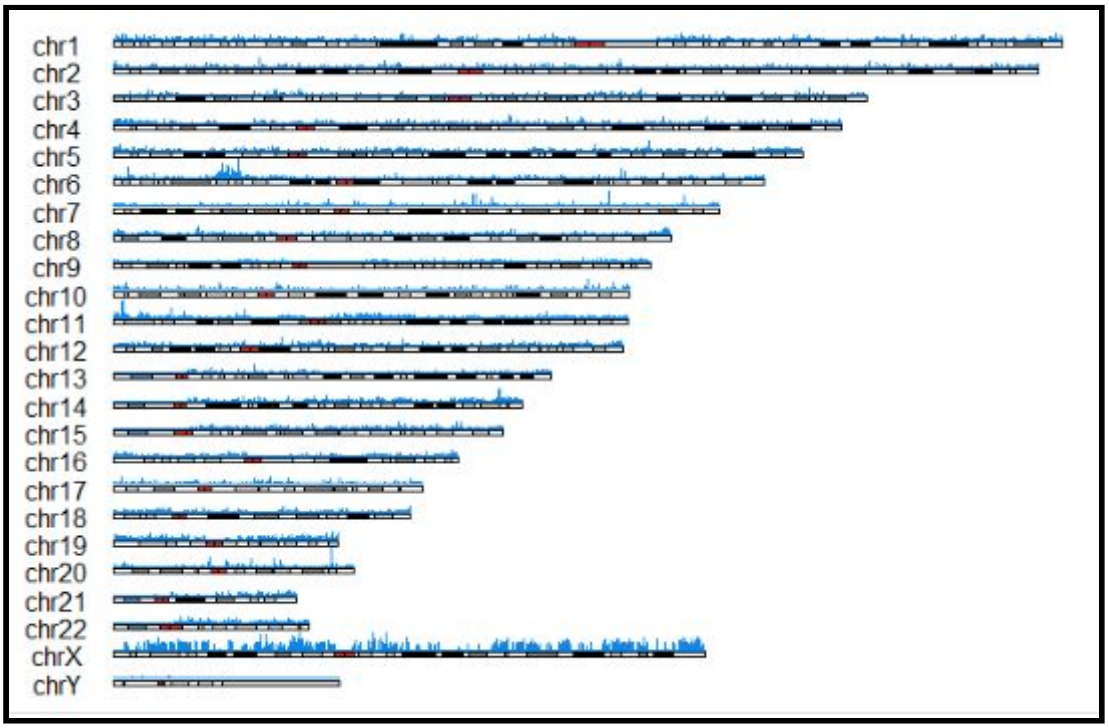
11 pav. Mėginių, iš kiekvieno smegenų regiono, koreliacija

Šiame paveiksle pavaizduotos metilintų reikšmių koreliacijos kiekvienam mėginiui (kurie yra įstrižainėse). Pyrago formos ir taškiniai grafikai parodo, kad koreliacija yra tarp kiekvieno iš mėginių. Mėlyna spalva žymi teigiamas reikšmes, o raudona neigiamas. Kuo

spalva ryškesnė, tuo atstumas tarp koreliacijų didesnis. 3 ir 4 mėginiai, kurie priklauso CRB smegenų regionui, mažiausiai koreliuoja su likusiais.

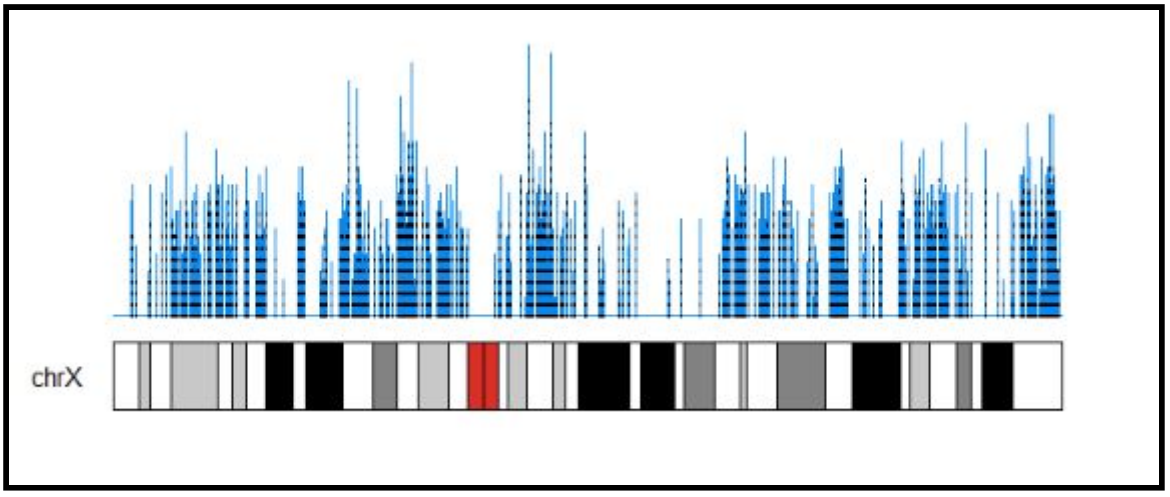
4.6. Genominių savybių persidengimo atvaizdavimas

Genominių savybių grafikas leidžia atvaizduoti genome arba pasirinktoje chromosomoje persidengusiais vietas. Genominės savybės tokios kaip snps, mutacijos, genai ar kitos savybės, kurios turi vietas genome, gali turėti persidengimus - toje pačioje genomo vietoje keletas pozicijų ar pan. Tokios turimos pasikartojančios pozicijos vietos chromosomoje, randamos pagal pozicijos pradžią ir pabaigą. Genome esančius bazių pasikartojimus galima suskaičiuoti ir pavaizduoti jų tankį. Šiam atvaizdavimui reikalingas “karyoploteR” paketas, ir naudojama funkcija, kuriai reikalingas genominis duomenų rinkinys “GRanges”. Šis rinkinys yra sudarytas iš chromosomos, pozicijos pradžios, galo, teigiamos arba neigiamos grandinės, todėl reikia turimą duomenų rinkinį pasiversti į tinkamą naudojimui [18]. Tokiu grafiku galima vizualiai pamatyti vietas, kur persidengia pozicijos lyčių chromosomose, jos pavaizduoti paveiksle 13. Naudotas duomenų rinkinys sudarytas iš atrinktų pozicijų. Joms rasti buvo atliktas nepriklausomų imčių testas pagal lytis, jei p-reikšmė mažiau už 0.05, tuomet pozicija tinkama. Tačiau galima pasirinkti ir specifinę chromosomą pavyzdžiui chromosomą X, kuri matosi detaliau paveiksle 14.



13 pav. Pozicijų persidengimai chromosomose

Genai turi genome esančias pozicijas, iš kurių atrinktos ir naudotos tik tos reiškmės, kurios po neprikausomų imčių testo (tarp lyčių) turi mažesnes p-reikšmes nei 0.05 (alfa). Todėl labiausiai skiriasi X chromosoma - turi daugiausia persidengimų.



14 pav. X chromosomoje esantys pozicijų persidengimai

Galima pavaizduoti pasirinktą chromosomą, kuri nurodo kiek šioje vietoje yra persidengiančių pozicijų kiekvienoje genomo bazėje.

Išvados

Tinkamai pasirinkus duomenis, jų analizavimui galima vizualiai pamatyti išskirtis dendrograma, ypač kai nuspalvinti lapai atitinkamomis spalvomis pagal grupę. Turint ilgus pavadinimus rekomenduojama naudoti horizontaliai. Taip pat duomenų panašumus ar skirtumus galima pamatyti paralelinių koordinačių grafiku, kuris leidžia pavaizduoti skirtingas charakteristikas įvairiems mėginiams linijomis, tačiau nepatartina analizuoti daug charakteristikų - grafikas tampa neįskaitomas.

Norint suskaičiuoti skirtumus ir panašumus tarp grupių tinka naudoti Veno diagramą, kuri gali naudoti daug grupių, tačiau kaip ir daugumai rekomenduojama naudoti tik keletą grupių analizavimui. Spalvų naudojimas grafike palengvina suvokimą.

Medžio diagrama tinka dideliems duomenų rinkiniams ir puikiai parodo kaip duomenys pasiskirsto pagal grupes, kurias galima skaidyti dar smulkiau - grupės grupė ir t.t. Grafikas ypač tinkamas mėginių požymių pasiskirstymui eksperimente įvertinti.

Kolerolgramoje naudojant daugiau nei 10 tiriamųjų tampa neįskaitomas. Grafikas suteikia galimybę koreliaciją pavaizduoti ne vien taškiniu būdu bet ir dviem skirtingais iš karto, kur vienas aukščiau o kitas žemiau. Taip pat spalvos vizualiai pateikia teigiamas ir neigiamas reikšmes. Toks grafikas mažiau naudojamas, tiriant koreliacijas geriau naudoti dendrogramą.

Genominių savybių persidengimo grafikas atvaizduoja visą genomą, nors galima pasirinkti ir gauti tik vienai chromosomai atvaizdavimą. Toks grafikas naudingas, norint pamatyti genome esančius pozicijų persidengimus, tarp tiriamų grupių, kurios rodo metilinimo skirtumus.

Šaltiniai

1. Visuotinė Lietuvių enciklopedija internetinė svetainė [interaktyvus] 2020. [žiūrėta 2020-04-04] Prieiga per: https://www.vle.lt/Straipsnis/epigenetika-113259
2. Lietuvos Mokslo taryba internetinė svetainė [interaktyvus] [žiūrėta 2020-06-15] Prieiga per: https://informacija.lmt.lt/EKSPERTAI/Anotacija_lt.php?Pr_kodas=60&Baze=VID&Proj_pavad=Molekulin%EBs%20%E1rankin%EBs%20epigenomikai%20ir%20RNomikai
3. Labome internetinė svetainė [interaktyvus] 2020. [žiūrėta 2020-04-04] Prieiga per: https://www.labome.com/method/DNA-Methylation.html
4. Dimura internetinė svetainė [interaktyvus] 2020. [žiūrėta 2020-04-06] Prieiga per: http://www.dimuna.com/lt/produktai/euroarray
5. Epigenie internetinė svetainė [interaktyvus] 2020. [žiūrėta 2020-04-06] Prieiga per: https://epigenie.com/epigenetics-research-methods-and-technology/methylcytosine-5mc-analysis/dna-methylation-arrays/
6. NCBI GEO internetinė svetainė [interaktyvus] [žiūrėta 2020-02-21] Prieiga per: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125895
7. R graph gallery Dendrogram internetinė svetainė [interaktyvus] 2018. [žiūrėta 2020-05-10] Prieiga per: https://www.r-graph-gallery.com/dendrogram.html
8. Nonlinear dynamics internetinė svetainė [interaktyvus] [žiūrėta 2020-05-14] Prieiga per: http://www.nonlinear.com/support/progenesis/comet/faq/v2.0/dendrogram.aspx
9. BMC Part of Spring Nature internetinė svetainė [interaktyvus] 2020. [žiūrėta 2020-05-15] Prieiga per: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-019-0646-6
10. Epigenesys internetinė svetainė [interaktyvus] 2011-2017. [žiūrėta 2020-05-17] Prieiga per: https://www.epigenesys.eu/sv/about-us/why-do-we-need-epigenetics-research
11. Data to Viz Dendrograma internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://www.data-to-viz.com/graph/dendrogram.html
12. Data to Viz Parallel coordinates plot internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://www.data-to-viz.com/graph/parallel.html
13. Data to Viz Treemap internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://www.data-to-viz.com/graph/treemap.html

14. Data to Viz Venn Diagram internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://www.data-to-viz.com/graph/venn.html
15. Data to Viz Correlogram internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://www.data-to-viz.com/graph/correlogram.html
16. KaryoploteR internetinė svetainė [interaktyvus] [žiūrėta 2020-05-25] Prieiga per: https://bernatgel.github.io/karyoploter_tutorial//Tutorial/PlotCoverage/PlotCoverage.html
17. Identification and Removal of Outlier Samples internetinis straipsnis [interaktyvus] [žiūrėta 2020-05-28] Prieiga per: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/HumanBrainTranscriptome/Identification%20and%20Removal%20of%20Outlier%20Samples.pdf
18. R documentation GRanges internetinė svetainė [interaktyvus] [žiūrėta 2020-06-10] Prieiga per: https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/GenomicRanges/html/makeGRangesFromDataFrame.html
19. Epigenetiniai tyrimai internetinis straipsnis [interaktyvus] [žiūrėta 2020-05-11] Prieiga per: https://www.upc.smm.lt/tobulinimas/renginiai/medziaga/biotechno/Epigenetika._genomo_abeceles_diakritiniai_zenklai._Prof._Saulius_Klimasauskas.pdf
20. The minfi User's Guide internetinė svetainė [interaktyvus] [žiūrėta 2020-05-11] Prieiga per: https://bioconductor.org/packages/release/bioc/vignettes/minfi/inst/doc/minfi.html