Praktinė užduotis Nr. 2

Klasterizavimas

Jaroslav Rutkovskij, Danielė Stasiūnaitė, Karolis Augustauskas (JDK)

07/05/22

Pagrindinės užduotys

Prieš atliekant duomenų hierarchinį klasterizavimą buvo sukurti nauji mėginių pavadinimai, pakeičiant originalius pavadinimus, kurie buvo ilgi ir neinformatyvūs. Pavadinimai buvo pakeisti, vadovaujantis žemiau aprašytu principu:

1. Nurodomas donoro numeris. Pvz.: ***n37***.
2. Po apatinio brūkšnio nurodomas laikas, kada buvo imta biopsija (gali būti ***t1*** arba ***t2***, kur ***t1*** nurodo, kad biopsija buvo imta pirmaisiais metais, o ***t2*** nurodo, kad biopsija buvo paimta praėjus 10 metų po pirmosios biopsijos ėmimo).
3. Po apatinio brūkšnio nurodoma, iš kurios žarnos (proksimalinės ar distalinės) buvo imama biopsija. Žarnų tipai užrašomi sutrumpinus pilnus pavadinimus iki keturių raidžių:
   proximal ⇒ ***prox***,
   distal ⇒ ***dist***.
4. Po apatinio brūkšnio trijų raidžių kombinacija nurodoma, ar tiriamasis vartojo aspiriną:
   nnu ⇒ **aspirinas nevartojamas (non-user)**,
   lng ⇒ **aspirinas vartojamas (longterm user)**.
5. Po apatinio brūkšnio nurodomas moters amžius.

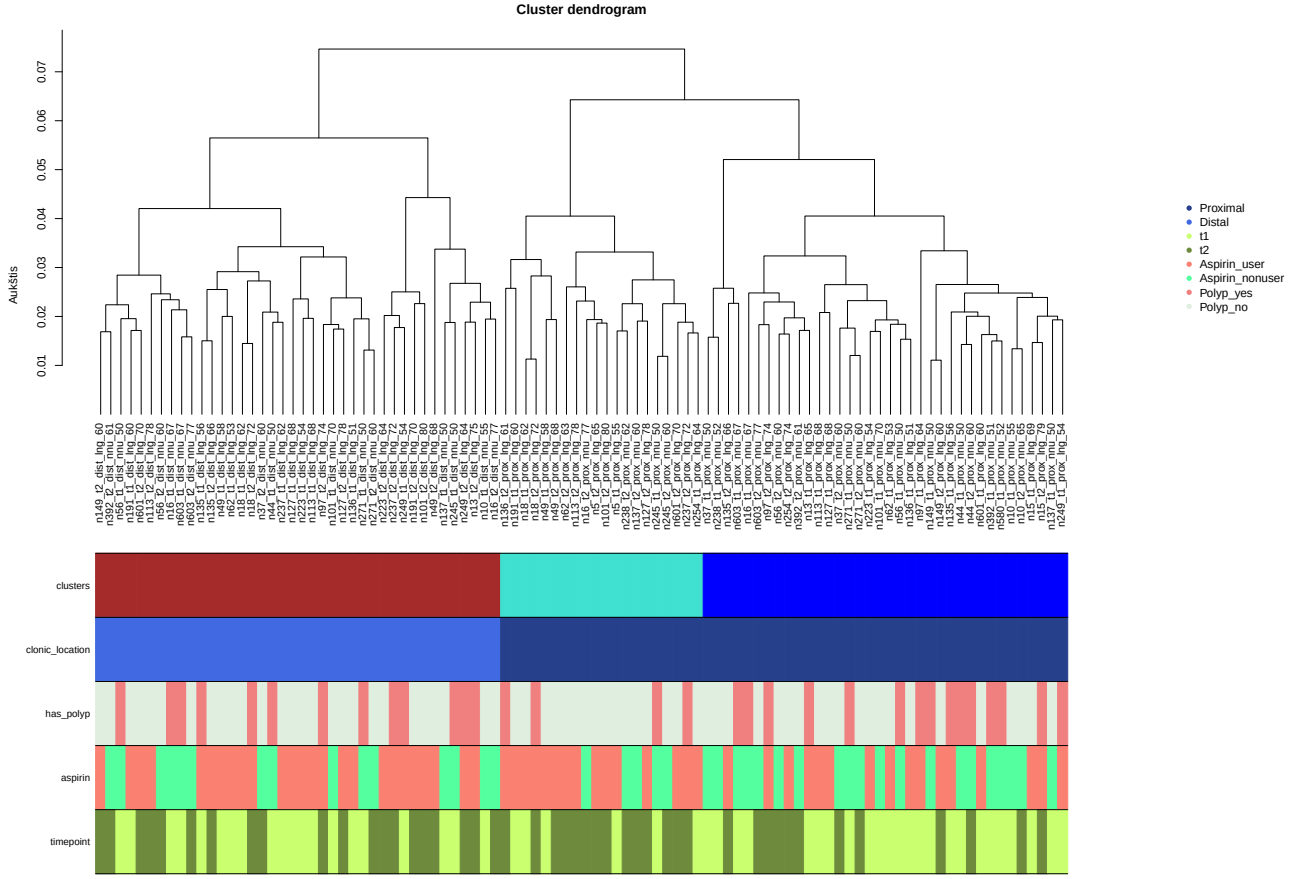
Pilnas naujai sukurto pavadinimo pavyzdys: ***n10__t1__prox__nnu__55***, kur:

- **n10** - donoro numeris;
- **t1** - biopsija imta pirmaisiais tyrimo metais;
- **prox** - mėginys gautas iš proksimalinės žarnos;
- **nnu** - tiriamasis nevartojo asprino;
- **55** - tiriamojo amžius.

Mėginių vardų pavyzdys realiems duomenims:

```
##            n5_t1_prox_lng_55 n5_t2_prox_lng_65
## cg14817997         0.7490543         0.8515831
## cg26928153         0.9521776         0.9509175
## cg16269199         0.8433516         0.8342750
```

1. Mėginių hierarchinis klasterizavimas



Atliekant šią užduotį išskyrėme 3 klasterius. Iš dendrogramos galime matyti 2 ryškius klasterius su distalinės žarnos ir proksimalinės žarnos mėginiais. Išskirtame viduriniame klasteryje matosi mėginiai, kurie turėjo polipų (polyp_yes), daugiau buvo ilgalaikių aspirino naudotojų (aspirin_user), taip pat daug mėginių paimtų vėlesniu laikotarpiu (t2 - paėmus mėginį po 10 metų). Daugiau smulkesnių susiskirstymų neišskyrėme.

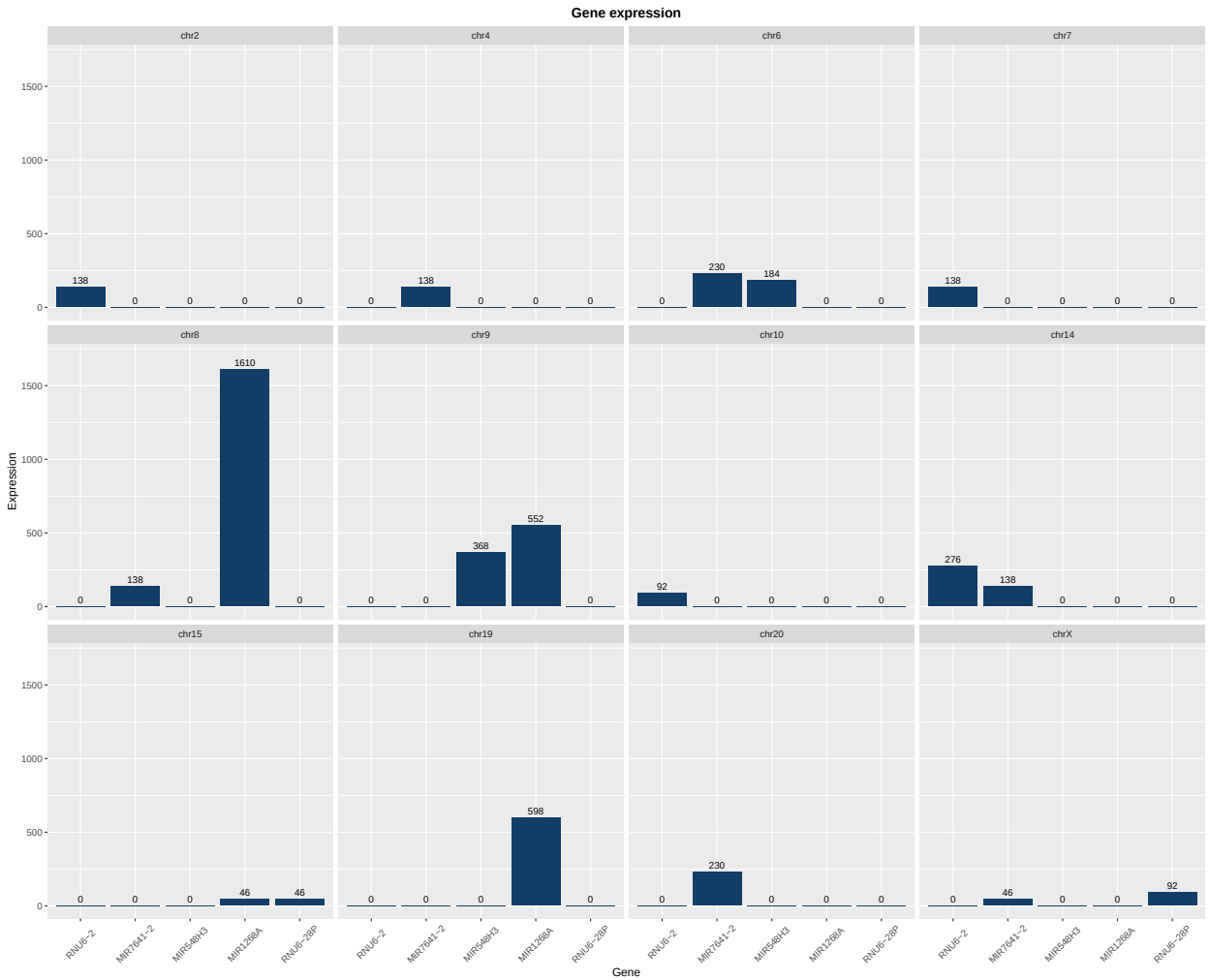
2. Mėginių atvaizdavimas “heatmap” pavidalu

Kadangi **beta** matrica turi ***853368*** eilutes bei ***96*** stulpelius, pavaizduoti visas matricos reikšmes bei, atsižvelgus į spalvų intensyvumą ir atspalvį, daryti išvadas apie duomenų grupes - klasterius, naudojantis ***heatmap()*** funkcija yra negalima, iš eilutes aprašančios matricos buvo atrinktos tik tam tikros eilutės.

Eilučių atrinkimas buvo atliktas, remiantis žemiau pažingsniui aprašyta logika:

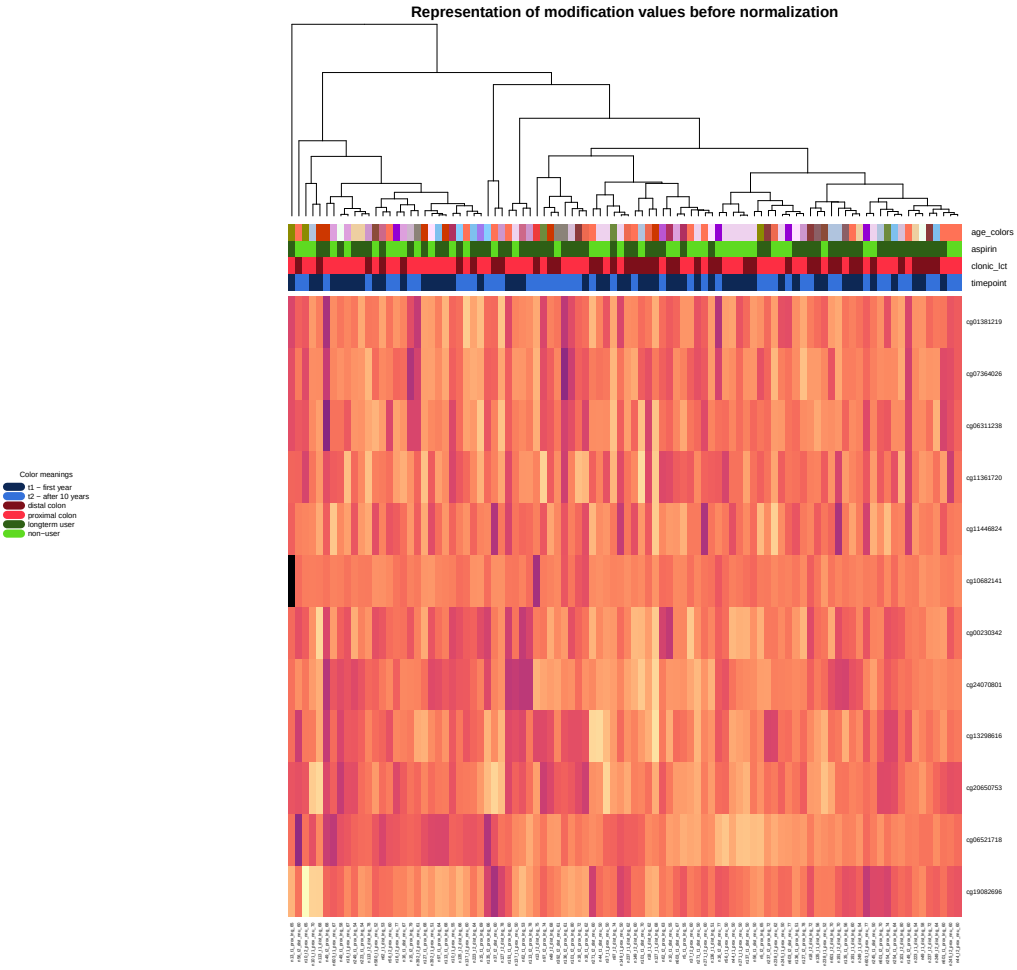
1. Iš pradžių iš eilutes aprašančios matricos buvo surasti genai (jų pavadinimai, remiantis eilutes aprašančios matricos ‘UCSC_RefGene_Name’ stulpeliu), kurie pasikartodavo tarp dviejų chromosomų. Gauti bendrų genų dažniai pavaizduoti lentele:
2. Turint dažnius buvo suformuotas objektas, suskaičiuojantis konkrečių genų dažnius chromosomose.
3. Iš matricos (ji buvo gauta iš data.frame objekto) buvo pašalintos eilutės (chromosomos), kuriose nebuvo nustatytas nei vienas genas (visos reikšmės eilutėje buvo lygios 0), ir buvo sukurtos stulpelinės diagramos, vizualizuojančios, kokiose chromosomose tam tikro geno buvo nustatyta daugiausiai.

Modifikacijos įverčiams atvaizduoti ***heatmap*** pavidalu buvo pasirinkta chromosoma, turinti didžiausią tam tikro geno raišką.

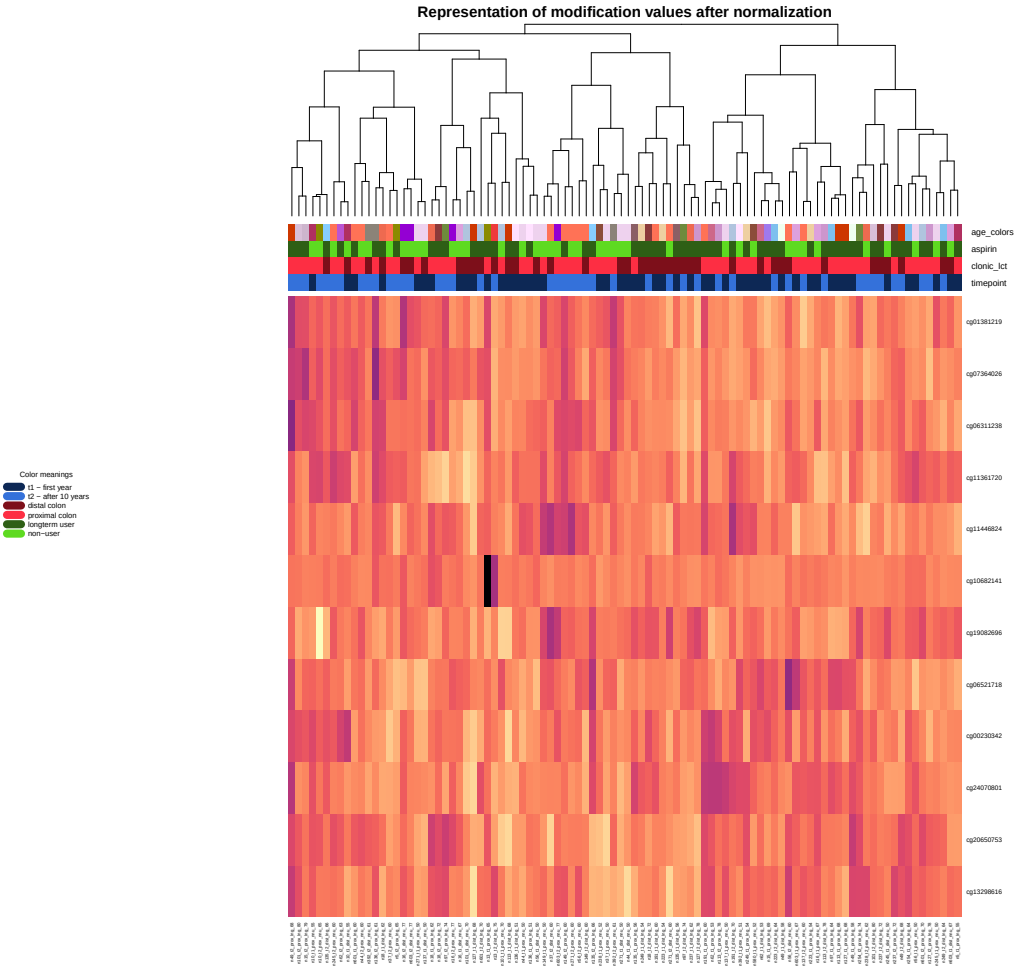


Remiantis aukščiau pavaizduotų stulpelinių diagramų duomenimis, ***heatmap*** buvo pasirinkta atvaizduoti **9 chromosomos** ir ***MIR1268A geno*** modifikacijos duomenis.

Žemiau atvaizduojami modifikacijos įverčiai ***heatmap*** grafiko pavidalu prieš atliekant normalizavimą:

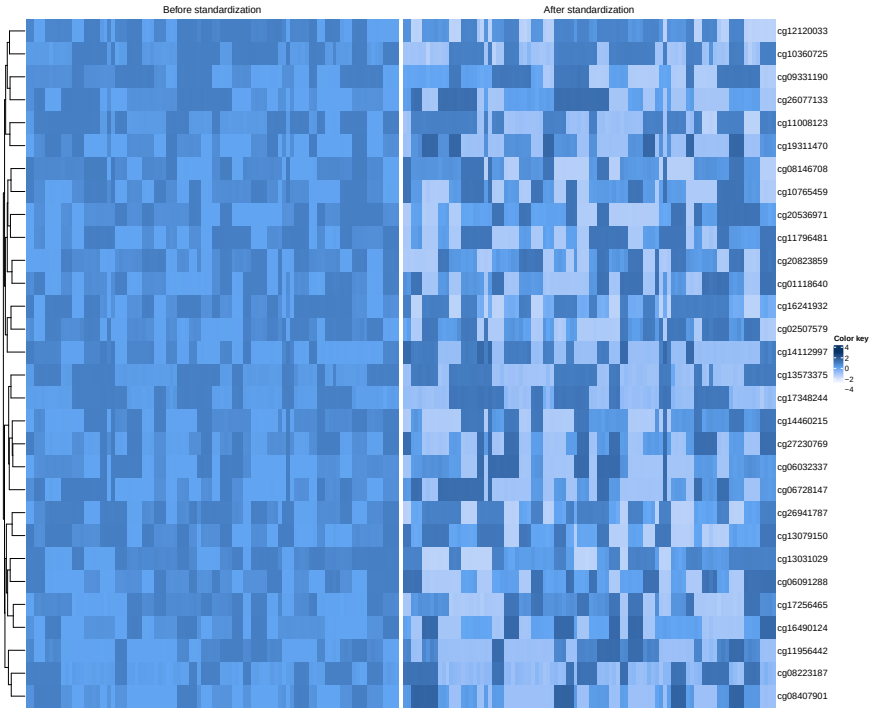


Žemiau atvaizduojami modifikacijos įverčiai ***heatmap*** grafiko pavidalu atlikus normalizavimą:



Apibendrinus gautus rezultatus galima padaryti išvadą, kad net ir sumažinus modifikacijos pozicijų skaičių, nagrinėjant tik devintos chromosomos ***MIR1268A*** geną, negalima įžvelgti aiškių grupių, kurios gali būti analizuojamos toliau, ieškant bendrumų tarp grupės objektų.

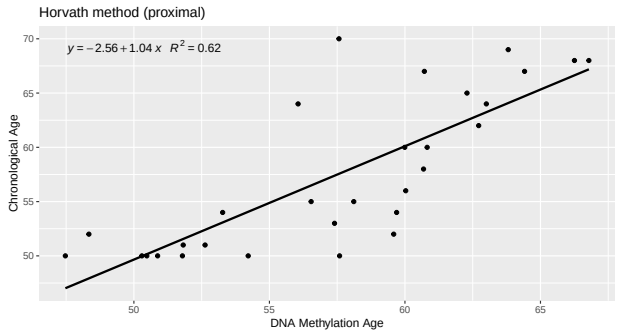
Siekiant gauti ***heatmap*** grafiką, kuriame galima matyti aiškiai susidariusias grupes (klasterius), buvo nuspręsta iš modifikacijos įverčių matricos atrinkti 30 variabiliausių pozicijų ir sukurti **heatmap** grafiką būtent šioms pozicijoms, pavaizduojant ***heatmap*** grafikus prieš atliekant normalizavimą bei atlikus jį.

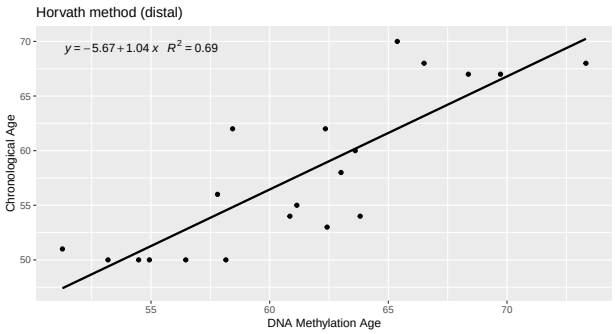


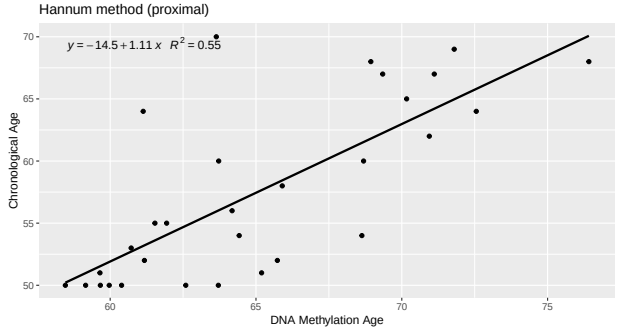
Aukščiau pavaizduotuose ***heatmap*** grafikuose aiškiai matomas spalvų pasikeitimas, atlikus modifikacijos pozicijų normalizavimą taip, jog eilučių vidurkis būtų lygus 0, o standartinis nuokrypis būtų lygus 1.

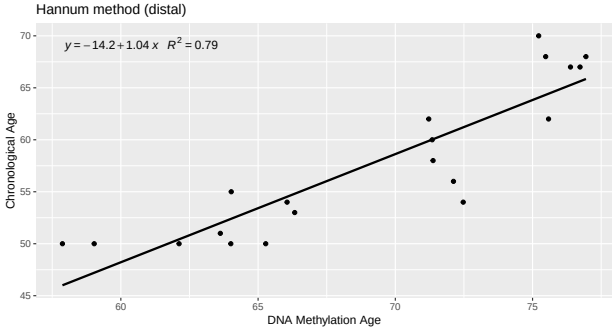
Remiantis "Color key" skale galima pastebėti, jog modifikacijos pozicijos, kurios prieš atliekant standartizavimą buvo vizualizuojamos šviesesne spalva, atlikus standartizavimą buvo pavaizduotos dar šviesesne spalva (reikšmės buvo sumažintos). Tuo tarpu modifikacijos pozicijos, kurios prieš normalizavimą buvo vaizduojamos tamsesne spalva, po normalizavimo įgijo dar didesnę reikšmę ir buvo atvaizduotos dar tamsesne spalva.

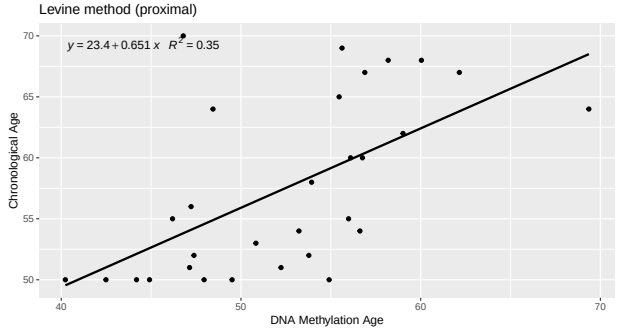
3. Tikro ir nuspėto amžiaus bei senėjimo "pagreitėjimo" nustatymas

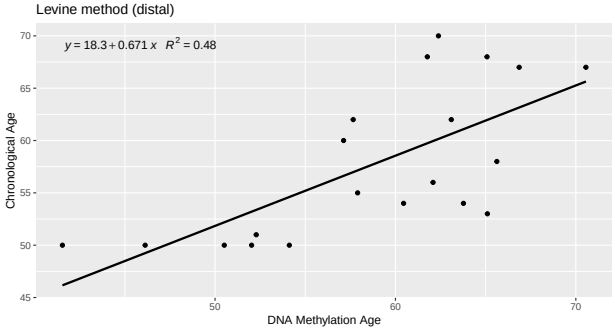


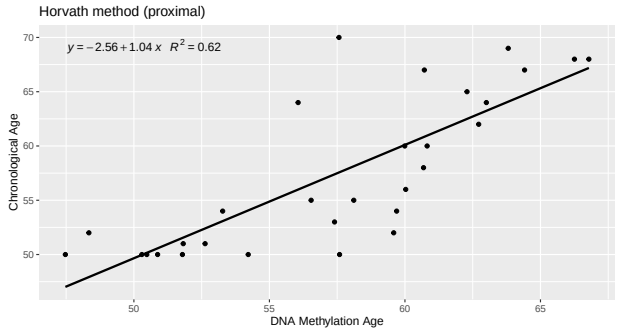


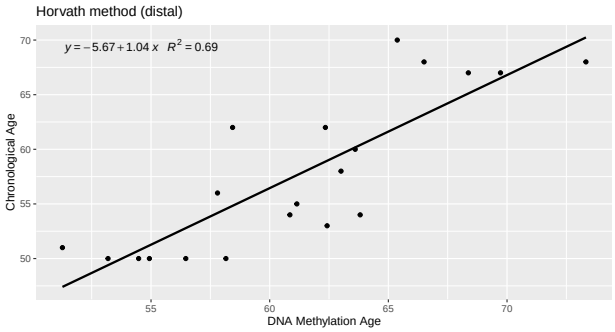


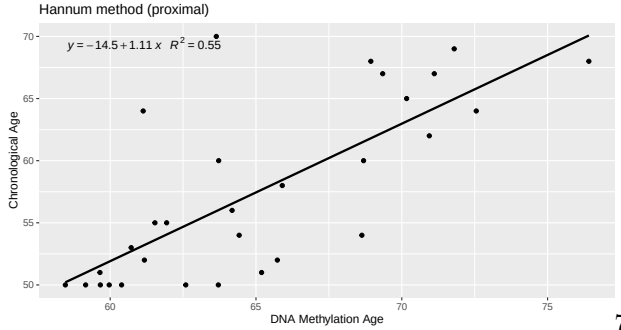


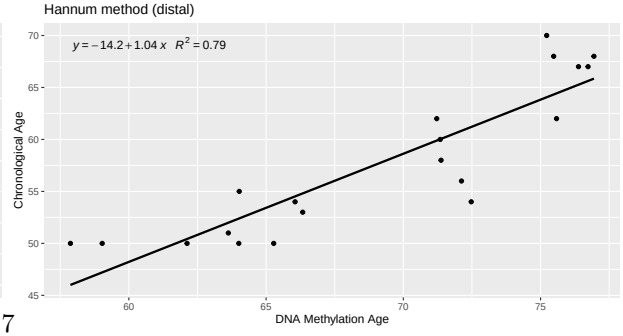


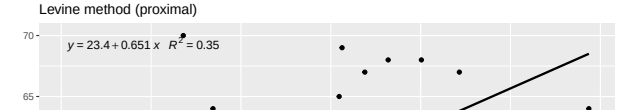


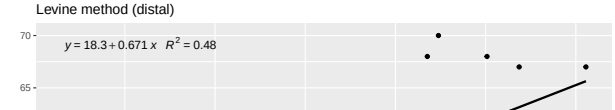






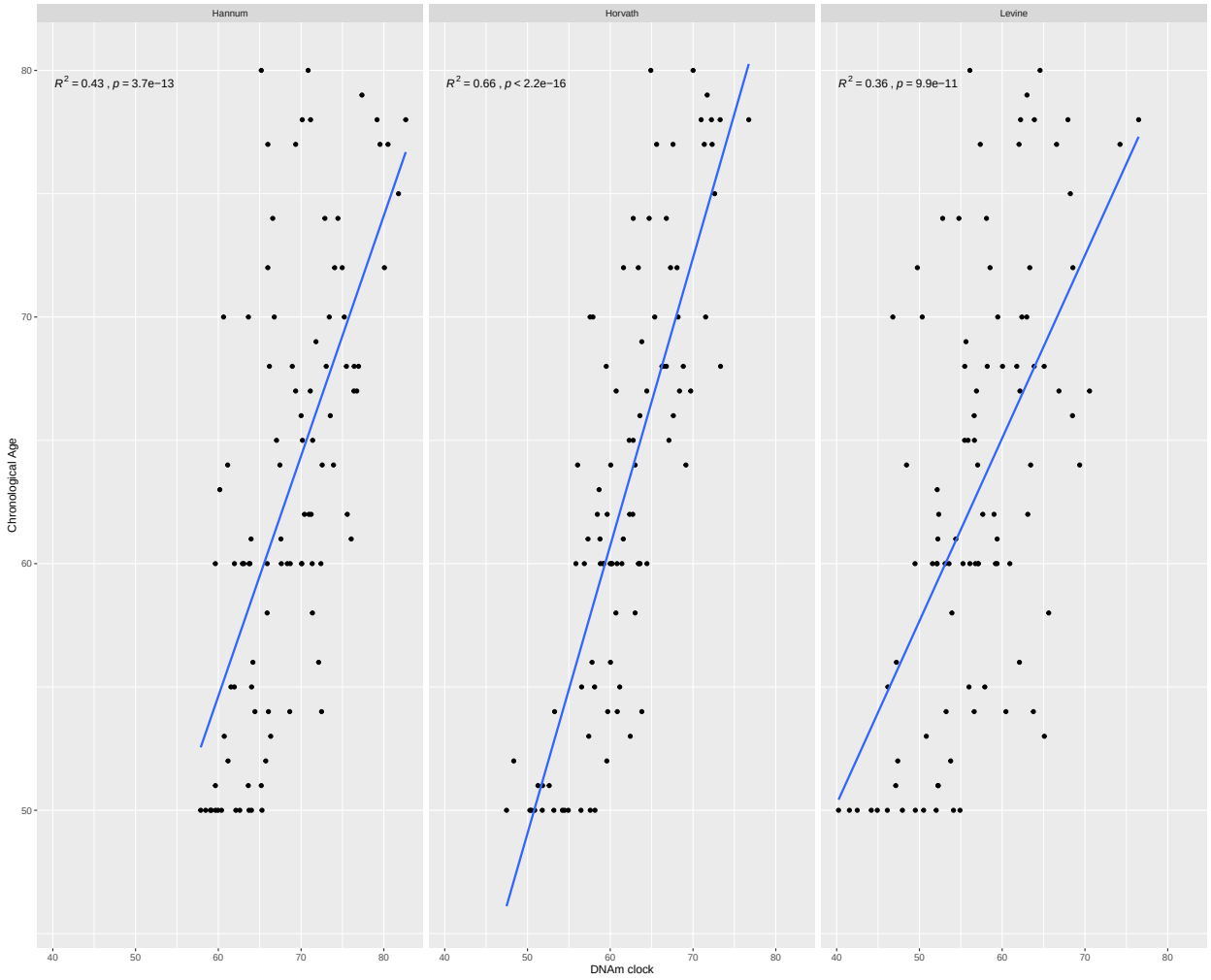




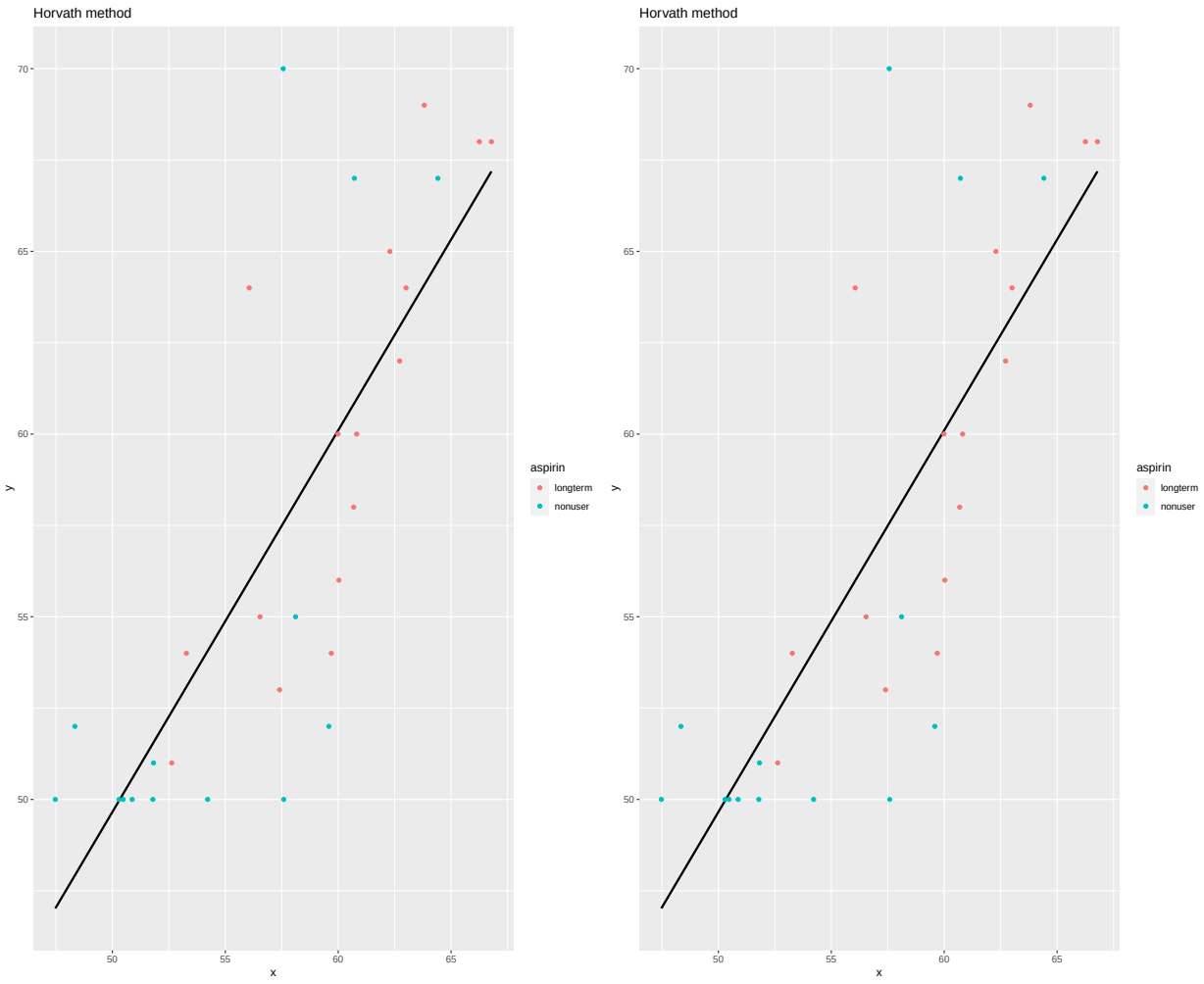


Parodoma, kuris iš trijų naudojamų laikrodžių yra tiksliausias ir kuris yra mažiausiai tikslus. Naudododami *DNAmAge* funkciją pamatome, kad tiksliausias yra ***Horvarth laikrodis***, kurio nuklydimo vidurkis yra ***Horvarth***, o mažiausiai tikslus yra ***Levine laikrodis***, kurio vidurkis yra lygus ***Levine***.

Žemiau yra vaizduojama koreliacija tarp visų trijų laikrodžių.



Žemiau esančiuose grafikuose vaizduojama, kaip nuspėti amžiai skiriasi tarp aspiriną vartojančių ir jo nevartojančių.



Papildoma užduotis

t-SNE klasterizavimo metodo pritaikymas

Kaip ir pirmosios užduoties atveju diagramoje matome du ryškius klasterius: vieną su distalinės žarnos mėginiais (apskritimai), kitą su proksimalinės žarnos mėginiais (trikampiai).