Praktinė užduotis Nr. 2

Klasterizavimas

Jaroslav Rutkovskij, Danielė Stasiūnaitė, Karolis Augustauskas (JDK)

4/24/2022

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 procedūros).

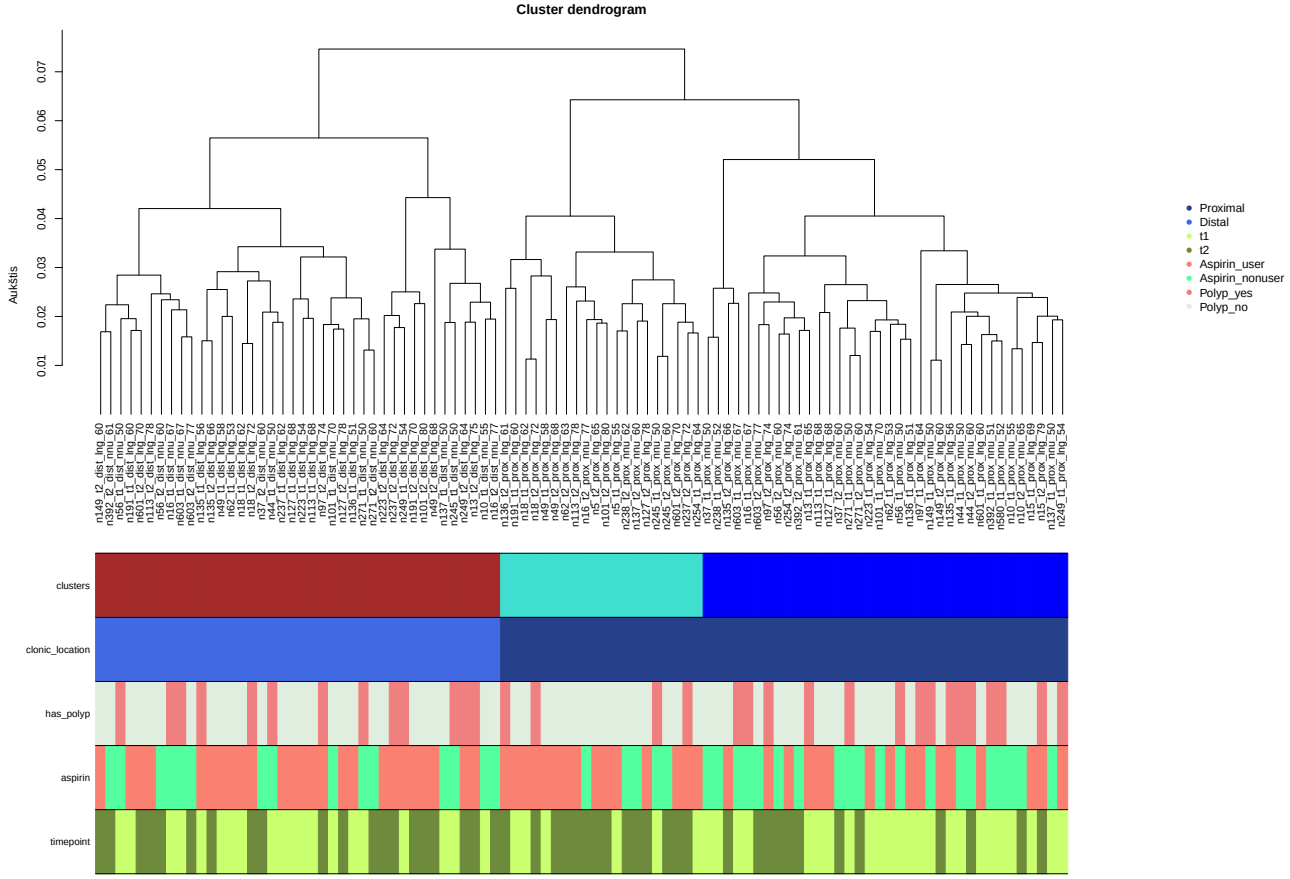
3. Po apatinio brūkšnio nurodoma, iš kurios žarnos (proksimalinės ar distalinės) buvo imama biopsija. Žarnų tipai užrašomi sutrumpinus pilnus pavadinimus iki keturių raidžių:
proximal ⇒ ***prox***, **distal** ⇒ ***dist***.

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.

1. Mėginių hierarchinis klasterizavimas



Išskyrėme 3 klasterius. Iš dendrogramos galime matyti du ryškius klasterius su distalinės žarnos ir proksimalinės žarnos mėginiais. Išskirtame viduriniame klasteryje matosi mėginiai, kurie turėjo polypų(polyp_yes), daugiau buvo ilgalaikių aspirino naudotojų(aspirin_user), taip pat daug mėginių paimtų vėlesniu laikotarpiu(t2). Daugiau smulkesnių susiskirstymų neišskyrėme.

Dabar išbandysime “t-SNE” klasterizavimo metodą

Kaip ir preitoje diagramoje matome du ryškius klasterius, vieną su distalinės žarnos mėginiai(apskritimai), kitą su proksimalinės žarnos mėginiais(trikampiai).

2. Mėginių atvaizdavimas “heatmap” pavidalu

Sukūrus aukščiau pavaizduotas stulpelines diagramas buvo nustatyta, kokie genai buvo dažniausi bei kokioje chromosomoje jų buvo daugiausiai.

Nuskaitytiems duomenims pritaikyta standartinė ***heatmap()*** funkcija, iliustruojanti **beta** matricos reikšmes ir pagal spalvų intensyvumą bei atspalvį, leidžianti daryti išvadas apie duomenų grupes - klasterius.
Kadangi **beta** matrica turi ***853368*** eilutes bei ***96*** stulpelius, pavaizduoti visas matricos reikšmes, naudojantis ***heatmap()*** funkcija yra negalima.

3. Tikro ir nuspėto amžiaus bei senėjimo “pagreitėjimo” nustatymas