Studi Kasus 1

Adrian Nugraha Utama

2023-10-28

Pengenalan

Pada studi kasus ini, Anda akan membandingkan distribusi data rna-seq dan microarray. Kita juga akan mencoba melakukan fitting terhadap data rna-seq, dan melihat apakah count data memiliki distribusi negative binomial.

Pustaka

```
library(airway) # Dataset - cara akses: data(gse)
library(MASS) # Untuk fitting dengan function fitdistr
# Dan library-library lainnya (dplyr, ggplot2) yang mungkin dibutuhkan
```

Task 1: Cari dan load data microarray

Cari dan unduh satu data microarray yang tersedia secara public. Petunjuk: Pada dataset GEO, series microarray memiliki nomor GSExxxxx, dan sampel memiliki nomor GSMxxxxxx. Untuk mengamati distribusinya, Anda hanya perlu download untuk salah satu sampel saja, tidak perlu untuk keseluruhan eksperimennya.

Coba impor data yang telah diunduh tersebut ke dalam R!

Task 2: Bandingkan distribusi data rna-seq dan microarray secara kualitatif

Coba plot distribusi data rna-seq (dari library airway) dan microarray. Bandingkan kedua distribusi tersebut, dan berikan komentar secara kualitatif. Petunjuk: Anda mungkin harus melakukan log-scaling pada sumbu-x.

Task 3: Fitting data rna-seq

Coba fitting-kan distribusi negative binomial dan poisson terhadap data rna-seq. Petunjuk: Anda boleh menggunakan fungsi fitdistr yang terdapat pada modul MASS.

Tampilkan distribusi dan hasil fittingannya pada satu grafik. Petunjuk: Untuk membandingkan model dengan histogram, lebar dari histogram bin harus sama, dan Anda mungkin juga perlu melakukan sedikit normalisasi. Buatkan perbandingan pada nilai count yang rendah, dan juga nilai count yang tinggi.

Pertanyaan: Apakah hasil fitting sesuai dengan ekspektasi? Apabila tidak sesuai, coba jelaskan kenapa!

Optional Task: Grafik Varians - Rerata

Terkecuali untuk kasus-kasus spesial tertentu, hasil fitting tidak akan sesuai dengan histogram data. Untuk melihat bahwa data rna-seq lebih cenderung mengikuti distribusi negative binomial daripada distribusi Poisson, salah satu bentuk visualisasi yang boleh digunakan adalah hubungan varians dan rerata untuk tiap-tiap gen. Coba buatkan plot tersebut, dan tunjukkan bahwa model negative binomial lebih sesuai!