Cours LSTAT2340 – Traitement statistique de données Omiques

Projet 4: Analyse de données omiques par tests multiples

Objectif

On dispose de données d'expression de gènes pour 3051 et 38 tumeurs émanant de 38 patients atteints par deux types de Leucémie: acute lymphoblastic leukemia (ALL) et acute myeloid leukemia (AML). On se demande quels sont les gènes qui permettent de différencier ces deux groupes.

But de votre travail : Rechercher dans les 3051 gènes ceux qui ont une expression significativement différente entre les deux groupes.

Les données sont décrites dans l'article Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, Vol. 286:531-537. http://www-genome.wi.mit.edu/MPR/

Votre contribution

Votre code R Markdown ainsi que le fichier html généré.

Consignes plus précises

Les données sont disponibles dans des fichiers .txt. Les trois jeux de données sont:

- golub = la matrice des expressions
- golub_cl = les classes des patients 0=ALL et 1=AML
- golub_gnames = noms des gènes (prendre la colonne 3)

Etapes de votre travail :

- Faites une analyse descriptive des données afin de choisir quel test statistique vous allez utiliser pour comparer les 2 groupes. Justifiez votre choix
- Calculer toutes les p-valeurs (non corrigées) et en faire un histogramme. Combien de p-valeurs sont plus petites que 0.05 ?
- Pour 3 procédures de correction pour multiplicité des tests (par exemple, Bonferroni, contrôle du FWER par la méthode de Sidak et calcul des q-valeurs), donnez :
    - Le nombre de tests considérés comme significatifs avec cette méthode
    - Les noms des 10 gènes qui ont les p-valeurs les plus petites et afficher ces noms dans l'ordre allant de la plus petite à la plus grande.

Délais

Votre rapport Rmarkdown est à rendre en semaine 5.