Table d'annotations génomique

Probabilités et statistique pour la biologie (STAT1)

Jacques van Helden

2017-09-14

But de ce tutoriel

Ce tutoriel vise à manipuler une table de données génomique (les annotations du génome de la levure) et à générer des graphiques pour représenter différents aspects liés à ces données.

Téléchargement des données

- Connectez-vous au site FTP d'EnsemblGenomes Fungi (http://fungi.ensembl.org/ (http://fungi.ensembl.org/)) et cliquez sur le lien **Download**.
- Sélectionnez le génome de *Saccharomyces cerevisiae* (boîte "Filter")
- Cliquez sur le lien **GTF**
- Lisez la documentation du format GTF sur les sites suivants:
  - Ensembl (http://www.ensembl.org/info/website/upload/gff.html (http://www.ensembl.org/info/website/upload/gff.html))
  - UCSC (https://genome.ucsc.edu/FAQ/FAQformat.html#format4 (https://genome.ucsc.edu/FAQ/FAQformat.html#format4))

Exercices

1. Ecrivez un script qui charge la table de données, en utilisant la fonction R `read.delim()`. Veillez à ignorer les lignes de commentaires (qui commencent par un caractère #).
2. Ajoutez une colonne intitulée "length" qui indique la longueur de chaque élément génomique annoté.
3. Comptez le nombre de lignes de la table correspondant à chaque type d'annotation (3ème colonne du GTF, "feature").
4. Sélectionnez les lignes correspondant à des gènes.
5. Comptez le nombre de gènes par chromosome.
6. Chargez la table des tailles de chromosomes chrom_sizes.tsv (../../data/Saccharomyces_cerevisiae/chrom_sizes.tsv), et calculez la densité de gènes pour chaque chromosome (nombre de gènes par Mb).
7. Dessinez la distribution de longueur des gènes.
8. Sur base de la taille du génome (12.156.679 bp) et des fréquences génomiques de codons définies au cours théorique, calculez la distribution attendue au hasard, et ajoutez-là au graphique.

Avant de terminer : conservez la trace de votre session

La traçabilité constitue un enjeu essentiel en sciences. La fonction ***R*** `sessionInfo()` fournit un résumé des conditions d'une session de travail: version de R, système opérateur, bibliothèques de fonctions utilisées.

```
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: macOS Sierra 10.12.6

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] knitr_1.15.1

loaded via a namespace (and not attached):
 [1] backports_1.0.5 magrittr_1.5    rprojroot_1.2   tools_3.3.2
 [5] htmltools_0.3.5 yaml_2.1.14     Rcpp_0.12.10    stringi_1.1.5
 [9] rmarkdown_1.5   stringr_1.2.0   digest_0.6.12   evaluate_0.10
```