M2 Mycologie

Outils bioinformatiques

Université Paris Cité

Outline

Utilisation des serveurs Galaxy publics

Techniques bioinformatiques

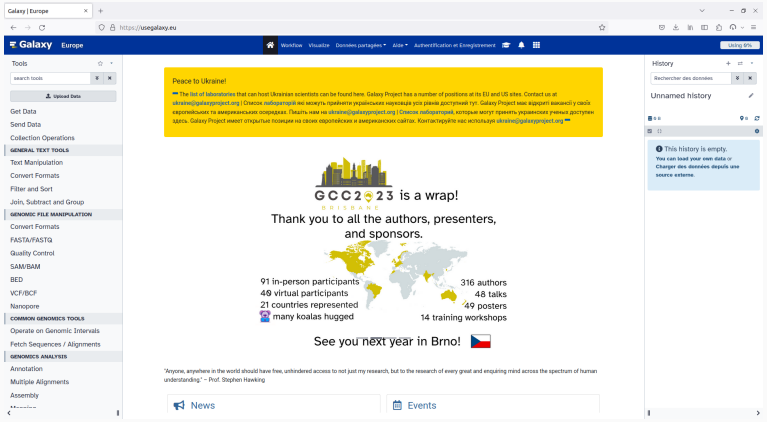
Algorithmes bioinformatiques

Utilisation des serveurs Galaxy publics

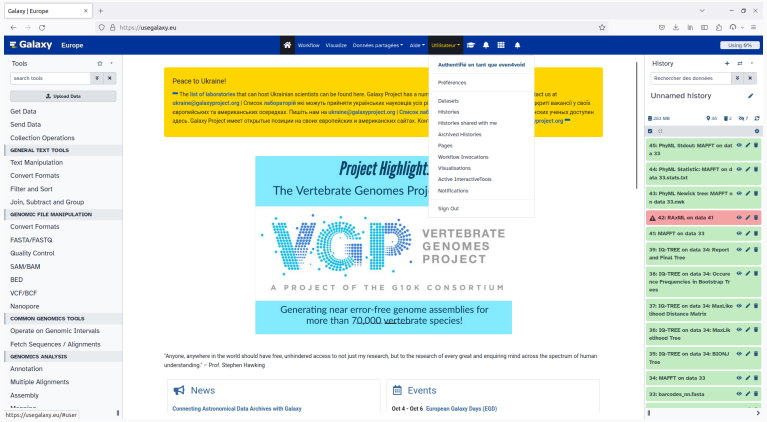
https://usegalaxy.eu

- créer un compte pour le transfert de fichiers et les notifications
- vérifier disponibilité des données partagées (”data only...”)
- temps de calcul variable (queue, batch job intercurrents, événements), sensibilité aux paramètres par défaut, disponibilité des utilitaires

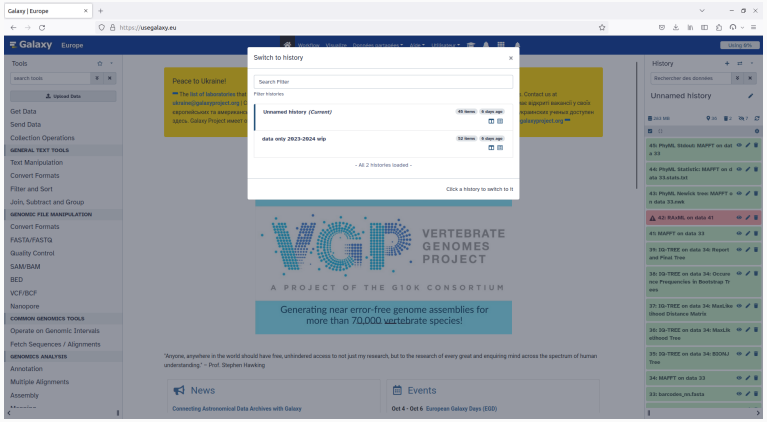
Authentification



Gestion des données partagées



Gestion des historiques



Techniques bioinformatiques

Quelques ordres de grandeur

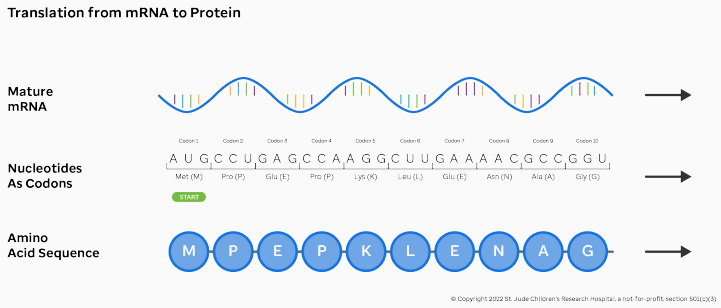
Podospora anserina

- génome 36 Mb (Fasta)
- données de séquençage 2 x 500-800 Mb (Fastq)
- 10k gènes annotés
- assemblage nouvelles souches : entre 4 et 8h (12 coeurs 3.5 GHz)
- phylogénie ITS seuls : 4h (phyml)
- phylogénie codes barres : 12 à 15h (phyml)

Implications informatiques

- serveur de calcul avec beaucoup de RAM (assemblage) et GPU (phylogénie)
- écriture de scripts shell et Python (ou R) pour les prétraitements et le développement de "workflow"
- serveur de stockage : 400 génomes ADN (+ 96 protéines) = 24 Go (en 2022)
- scripts de recherche/blast automatique (NCBI, JGI, etc.)

Recherche de motifs



- blast (shell ou en ligne au NCBI)
- scripts (Python, Perl, R, Bash, etc.)

Alignement de séquence

```
RLA0_METVA  --MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDK
RLA0_METJA  ---METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDK
RLA0_PYRAB  --------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRL
RLA0_PYRHO  --------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRL
RLA0_PYRFU  --------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRL
RLA0_PYRKO  --------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDK
RLA0_HALMA  MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRD
RLA0_HALVO  MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRE
RLA0_HALSA  MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRG
RLA0_THEAC  --------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRIRQIQDIRGK
RLA0_THEVO  --------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRIRQMQDIRAK
RLA0_PICTO  --------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNS
```

- clustal
- mafft (*)
- muscle[1]
- visualisateurs : jalview, seaview

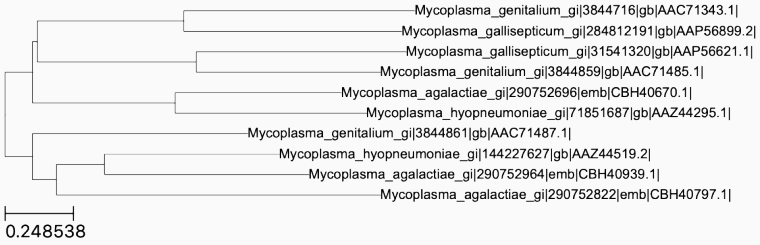
[1] https://bioinformaticsreview.com/20151018/multiple-sequence-alignment/

Mapping et assemblage (de novo)

- hisat2, tophat, bowtie2
- bwa[2]
- unicycler (spades) (*)
- abyss[3]

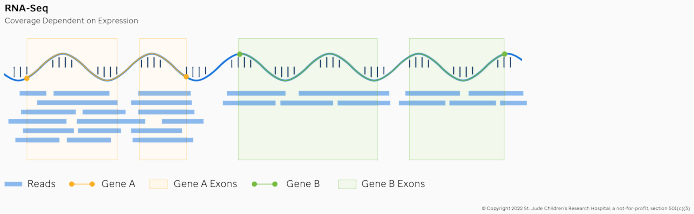
[2] Benchmarking short sequence mapping tools
[3] A biologist's guide to de novo genome assembly using next-generation sequence data

Phylogénie moléculaire



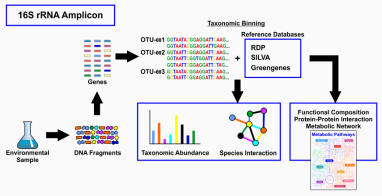
- fasttree
- IQ-TREE (*)
- RaXML
- MEGA
- NGPhylogeny
- visualisateurs : seaview (phylip), figtree, itol (payant)

RNA-Seq



- TopHat2 + HTSeq (ou assimilé)
- kallisto + DESeq2 (R) (*)
- Blast2Go (payant, version académique limitée)

Métagénomique

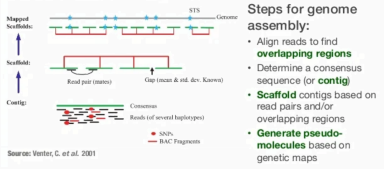


- *species*[a] vs. gene-centric
- FROGS (workflow Galaxy, base de données ITS)
- Kraken (bases de données pré-existantes) (*)

[a] Chapter 12: Human Microbiome Analysis, PLoS Computational Biology 8(12):e1002808

Algorithmes bioinformatiques

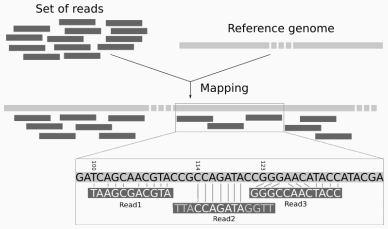
Assemblage de génome *de novo*



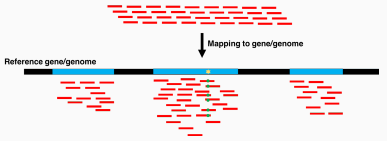
- Données : short et/ou long reads (FASTQ)
- The present and future of de novo whole-genome assembly

Alignement sur un génome de référence (mapping)

- Données : short reads (FASTQ), génome de référence (FASTA)
- Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis



Détection de mutation (variant calling)

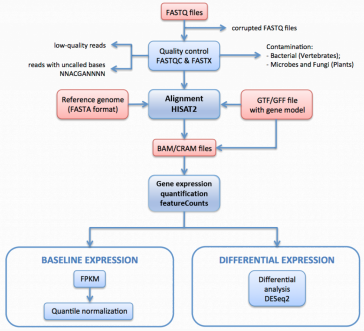


- Données : reads (FASTQ), génome de référence (FASTA)
- Fichier VCF comprenant les positions identifiées et les nucléotides associés (% et probabilité)
- Haute sensibilité aux paramètres de filtrage (cf. tutoriel Galaxy dans le cas des champignons)

RNA-Seq : mapping & quantification

- Données : reads (FASTQ), génome de référence (FASTA)
- RPKM (reads per kilobase of exon model per million reads), FPKM (fragments per kilobase of exon model per million reads mapped) : prise en compte de la longeuur des gènes et de la taille de la bibliothèque
- Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis

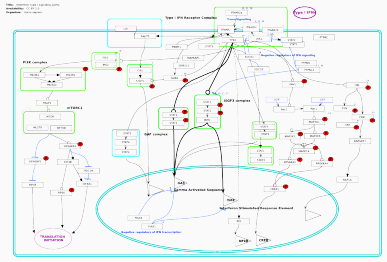
RNA-Seq : analyse différentielle



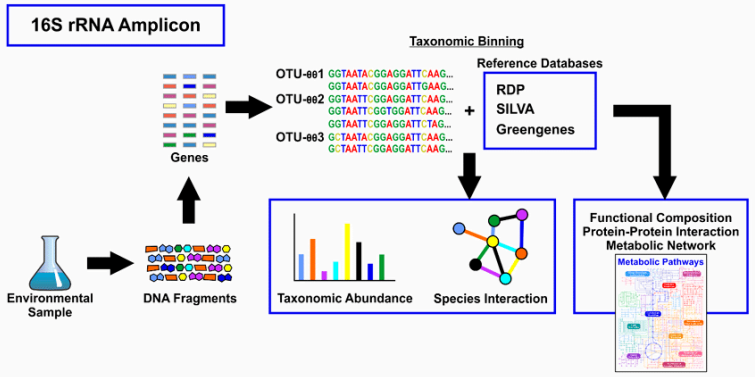
- Données : RPKM ou FPKM
- Approche fréquentiste ou bayésienne pour décider si les données de comptage moyennées sur les réplicats techniques et normalisées pour chaque réplicat biologique sont dûes au hasard ou non (gène sur- ou sous-exprimé par analyse de contraste sur condition de référence).

RNA-Seq : analyse d'enrichissement

- Données : tableau de quantification, annotation (go-terms, interpro)
- Approche par classification (3 classes/ontologies pour les go-terms : cellular component, biological process or molecular function) et "pathway"/"network" analysis (processus biologiques ou fonction moléculaire, et évenements régulatoires)

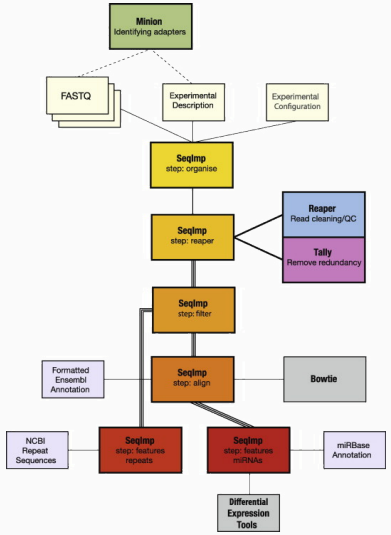


Métagénomique : Principe général



Métagénomique : Kraken

- Utilisation de base de données pré-définies, que l'on peut augmenter avec des souches de référence
- Mise en oeuvre rapide et rapport importable dans les suites d'analyses statistiques



Références i

[1] Mostafa M. Abbas, Qutaibah M. Malluhi, and Ponnuraman Balkrishnan. "Assessment of de novo assemblers for draft genomes: a case study with fungal genomes". In: *BMC Genomics* 15.S10 (2014), pp. 1–12.

[2] Scot A. Kelchner and Michael A. Thomas. "Model Use in Phylogenetics: Nine Key Questions". In: *TRENDS in Ecology and Evolution* 22.2 (2006), pp. 87–94.

[3] Bo Li et al. "RNA-Seq gene expression estimation with read mapping uncertainty". In: *Bioinformatics* 26.4 (2010), pp. 493–500.

[4] Ernesto Picardi. *RNA Bioinformatics*. Humana, 2021.

Références ii

[5] Ziheng Yang and Bruce Rannala. "Molecular phylogenetics: principles and practice". In: *Nature Reviews Genetics* 13 (2012), pp. 303–314.

Source principale des illustrations : `https://learngenomics.dev/`, `https://is.gd/xRcxsR`