

M2 Mycologie

Outils bioinformatiques



Outline

Utilisation des serveurs Galaxy publics

Techniques bioinformatiques

Algorithmes bioinformatiques

Références bibliographiques

1

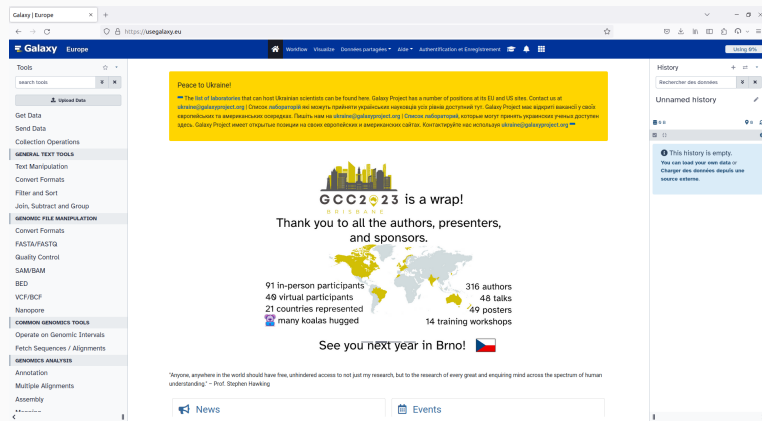
Utilisation des serveurs Galaxy publics

<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

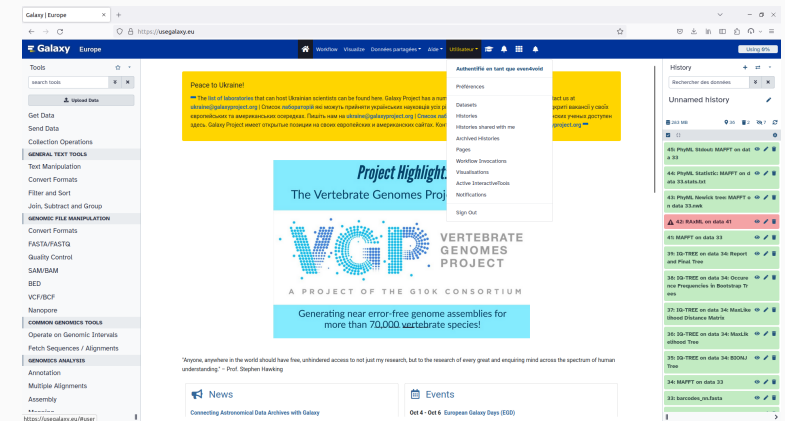
2

Authentification



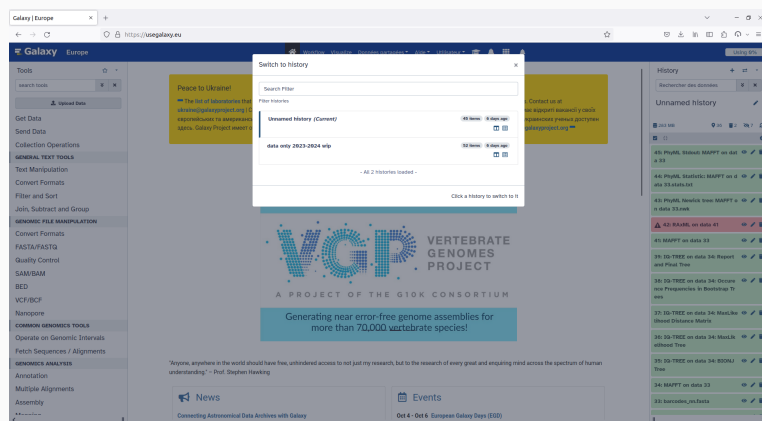
3

Gestion des données partagées



4

Gestion des historiques



5

Techniques bioinformatiques

Quelques ordres de grandeur

Podospira 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 (phymI)
- phylogénie codes barres : 12 à 15h (phymI)

6

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.)

7

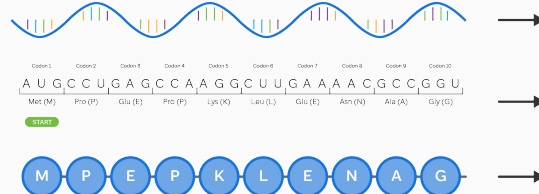
Recherche de motifs

Translation from mRNA to Protein

Mature mRNA

Nucleotides As Codons

Amino Acid Sequence



© Copyright 2022 St. Jude Children's Research Hospital, a not-for-profit, section 501(c)(3)

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

8

Alignement de séquence

```
RLA0_METVA --MIDAKSEHKIAPWKIEEYNALKELIKSANVIALIDHMEVPAVLOEIRDK
RLA0_METJA ---METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAQLQEIIRDK
RLA0_PYRAB -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRRL
RLA0_PYRHO -----MAHVAEWKKKEVEELAKLIKSYPIVIALVDVSSMPAYPLSQMRRRL
RLA0_PYRFU -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVSSMPAYPLSQMRRRL
RLA0_PYRKO -----MAHVAEWKKKEVEELANLIKSYPIVIALVDVAGVPAYPLSKMRDK
RLA0_HALMA MSAESERKTETIPWKQEEVDAIVMIESYESVGVVNIAGIPSRLODMRRD
RLA0_HALVO MSAEVRQTEVIPWKREEVDLVDFIESYESVGVVGVAGIPSRLODMRRD
RLA0_HALSA MSAEEQRTTEEVPEWKRQEVAVELVDLLETYSVGVVNVTCIPSKLODMRRD
RLA0_THEAC -----MKEVSQKKELVNEITRIKASRSVAIVDTAGIRTRQIODIRGK
RLA0_THEVO -----MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTROMDIRAK
RLA0_PICTO -----MTEDAQWKIDFVKNLENEINSRKVAIVSISIKGLRNNEFOXIRNS
```

- clustal
- mafft (*)
- muscle¹
- visualisateurs : jalview, seaview

¹<https://bioinformaticsreview.com/20151018/multiple-sequence-alignment/>

9

Mapping et assemblage (de novo)

Whole Genome Sequencing

30-60x Coverage



© Copyright 2020 St. Jude Children's Research Hospital. All rights reserved. S002225

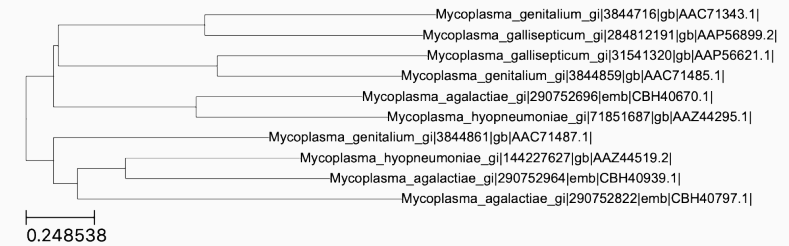
- hisat2, tophat, bowtie2
- bwa²
- unicycler (spades) (*)
- abyss³

²Benchmarking short sequence mapping tools

³A biologist's guide to de novo genome assembly using next-generation sequence data

10

Phylogénie moléculaire



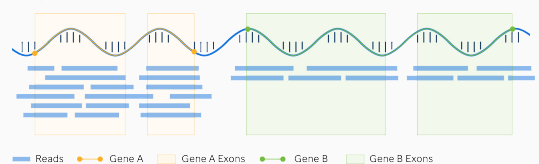
- fasttree
- IQ-TREE (*)
- RaXML
- MEGA
- NGPhylogeny
- visualisateurs : seaview (phylip), figtree, itol (payant)

11

RNA-Seq

RNA-Seq

Coverage Dependent on Expression

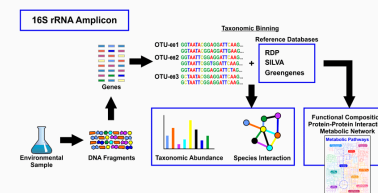


© Copyright 2020 St. Jude Children's Research Hospital. All rights reserved. S002225

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

12

Métagénomique



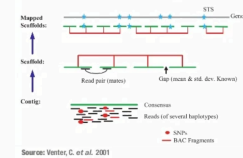
- *species*^a vs. gene-centric
- FROGS (workflow Galaxy, base de données ITS)
- Kraken (bases de données pré-existantes) (*)

^aChapter 12: Human Microbiome Analysis, PLoS Computational Biology 8(12):e1002808

13

Algorithmes bioinformatiques

Assemblage de génome *de novo*



Steps for genome assembly:

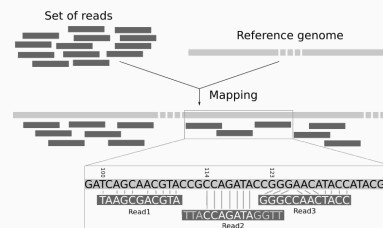
- Align reads to find **overlapping regions**
- Determine a consensus sequence (or **contig**)
- Scaffold** contigs based on read pairs and/or overlapping regions
- Generate **pseudo-molecules** based on genetic maps

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

14

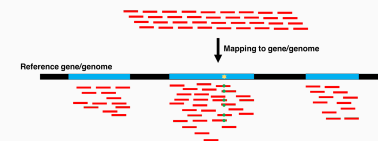
Alignement sur un génome de référence (mapping)

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



15

Détection de mutation (variant calling)

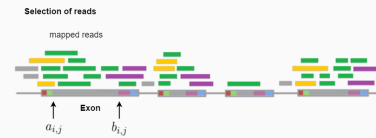


- Données : reads (FASTQ), génome de référence (FASTA)
- Fichier VCF comprenant les positions identifiées et les nucléotides associés (% et probabilité)
- Haute sensibilité aux paramètres de filtrage (cf. tutoriel Galaxy dans le cas des champignons)

16

RNA-Seq : mapping & quantification

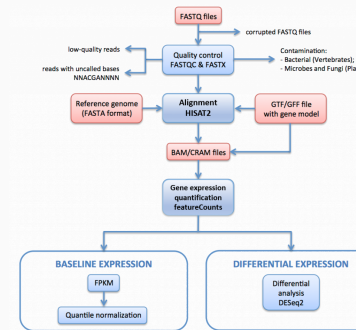
- Données : reads (FASTQ), génome de référence (FASTA)
- RPKM (reads per kilobase of exon model per million reads), FPKM (fragments per kilobase of exon model per million reads mapped) : prise en compte de la longueur des gènes et de la taille de la bibliothèque
- Systematic comparison and assessment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



17

RNA-Seq : analyse différentielle

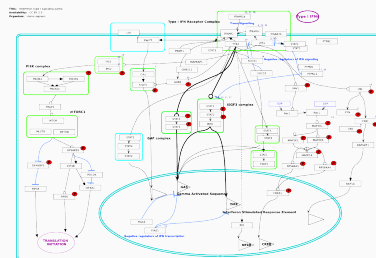
- Données : RPKM ou FPKM
- Approche fréquentiste ou bayésienne pour décider si les données de comptage moyennées sur les réplicats techniques et normalisées pour chaque réplicat biologique sont dûes au hasard ou non (gène sur- ou sous-exprimé par analyse de contraste sur condition de référence).



18

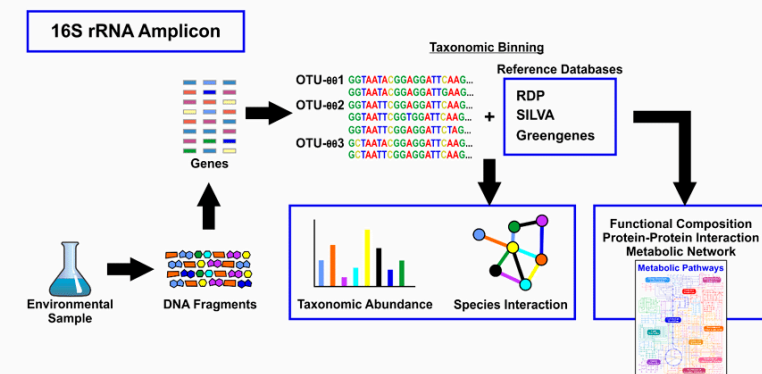
RNA-Seq : analyse d'enrichissement

- 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 événements régulatoires)



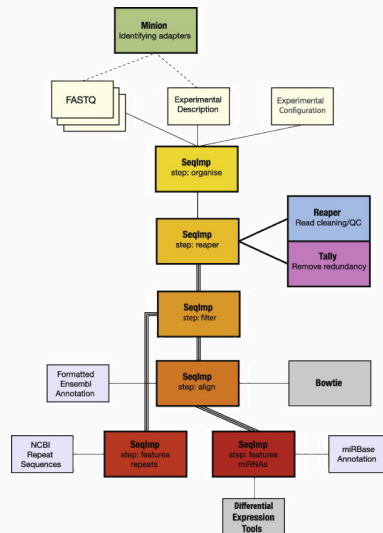
19

Métagénomique : Principe général



20

- 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



21

Références bibliographiques

- Ernesto Picardi (2021) *RNA Bioinformatics*, Springer
- Ziheng Yang & Bruce Rannala (2012) 'Molecular phylogenetics: principles and practice', *Nature Reviews Genetics* 13, 303–314
- Scot A. Kelchner & Michael A. Thomas (2006) 'Model Use in Phylogenetics: Nine Key Questions', *TRENDS in Ecology and Evolution* 22, 87–94
- Mostafa M. Abbas, Qutaibah M. Malluhi & Ponnuraman Balkrishnan (2014) 'Assessment of de novo assemblers for draft genomes: a case study with fungal genomes', *BMC Genomics* 15, 1–12
- Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson & Colin N. Dewey (2010) 'RNA-Seq gene expression estimation with read mapping uncertainty', *Bioinformatics* 26, 493–500

Source principale des illustrations : <https://learn.genomics.dev/>,
<https://is.gd/xRcxSR>

21