

M2 Mycologie

Outils bioinformatiques

Utilisation des serveurs Galaxy publics

Techniques bioinformatiques

Algorithmes bioinformatiques

Utilisation des serveurs Galaxy publics

- 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

Authentication

Galaxy Europe

Tools
search tools

Get Data
Send Data
Collection Operations

GENERAL TEXT TOOLS
Text Manipulation
Convert Formats
Filter and Sort
Join, Subtract and Group

GENOMIC FILE MANIPULATION
Convert Formats
FASTA/FASTQ
Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore

COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Sequences / Alignments

GENOMICS ANALYSES
Annotation
Multiple Alignments
Assembly

Navigation icons: Home, Workflow, Visualize, Download outputs, Help, Authentication or Logout, Settings, Galaxy logo.

Using RSE

Peace to Ukraine!

The list of bioinformaticists that can host Ukrainian scientists can be found here. Galaxy Project has a number of positions at its EU and US sites. Contact us at ukraine@galaxyproject.org | galaxy@galaxyproject.org или используйте следующие номера для связи с нами на русском языке. Galaxy Project was supported by grants from European Union and American National Institutes of Health. The project is also supported by the Ministry of Science and Higher Education of the Russian Federation. The project is also supported by the Ministry of Science and Higher Education of the Russian Federation. The project is also supported by the Ministry of Science and Higher Education of the Russian Federation.

GCC23 is a wrap!

Thank you to all the authors, presenters, and sponsors.

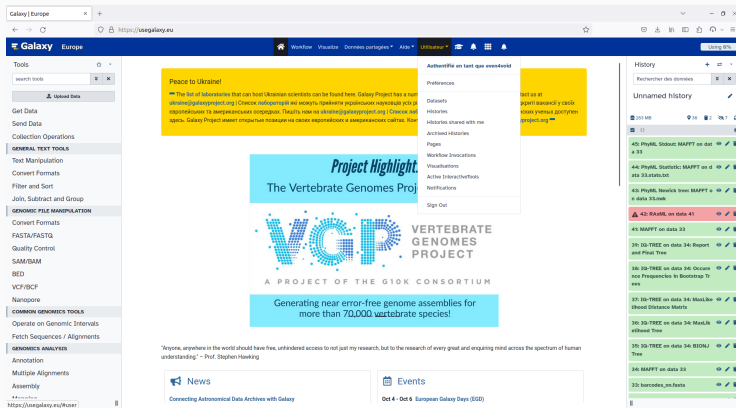
91 in-person participants	316 authors
49 virtual participants	48 talks
21 countries represented	49 posters
many koalas hugged	14 training workshops

See you next year in Brno! 🇨🇪

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." ~ Prof. Stephen Hawking

News Events

Gestion des données partagées



The screenshot shows the Galaxy Europe web interface. The top navigation bar includes the Galaxy logo, the text "Europe", and a "Workflow" button. Below the navigation bar, there is a sidebar on the left with various tool categories: "Tools" (with a search bar and "Upload Data" button), "GENERAL TEXT TOOLS" (including Get Data, Send Data, Collection Operations, Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group), "GENOMIC FILE MANIPULATION" (including Convert Formats, FASTA/FASTQ, Quality Control, SAM/BAM, BED, VCF/BCF, Nanopore), "COMMON GENOMICS TOOLS" (including Operate on Genomic Intervals, Fetch Sequences / Alignments), and "GENOMICS ANALYSIS" (including Annotation, Multiple Alignments, Assembly). The main content area features a "Project Highlight" for the "Vertebrate Genomes Project" (VGP), which is a project of the "10K Consortium". The highlight text reads: "Generating near error-free genome assemblies for more than 70,000 vertebrate species!". Below the highlight, there is a quote: "Anyone, anywhere in the world should have free, unbridled access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." - Prof. Stephen Hawking. The bottom of the interface shows a "News" section with the title "Connecting Astronomical Data Archives with Galaxy" and an "Events" section with the title "Oct 4 - Oct 6 European Galaxy Days (EGD)". On the right side, there is a "History" panel showing a list of recent workflows, including "PhyML, Subout: MAFFT on data 33", "PhyML, Statistics: MAFFT on data 33", "PhyML, Newick tree: MAFFT on data 33", "RAxML on data 41", "MAFFT on data 33", "3D-TREE on data 34: Report and Plot Tree", "3D-TREE on data 34: Occurrence Frequencies in Bootstrap Trees", "3D-TREE on data 34: Multi-Link Distance Matrix", "3D-TREE on data 34: Multi-Link Distance Tree", "3D-TREE on data 34: BIONJ Tree", "MAFFT on data 33", and "3D barcodes_m.fasta".

Gestion des historiques

The screenshot shows the Galaxy web interface with a 'Switch to history' dialog box open. The dialog box has a search filter and a list of history items. The first item is 'Unnamed history (Current)' with a size of 45 items and a date of 6 days ago. The second item is 'data only 2023-2024 w/p' with a size of 52 items and a date of 6 days ago. Below the list, it says '- All 2 histories loaded -'. At the bottom of the dialog box, it says 'Click a history to switch to it'.

The background interface shows the Galaxy logo, a search bar, and a list of tools on the left. The main content area displays a banner for the Vertebrate Genomes Project (VGP) with the text 'Generating near error-free genome assemblies for more than 70,000 vertebrate species!'. Below the banner, there is a quote: 'Anyone, anywhere in the world should have free, unbridled access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding.' - Prof. Stephen Hawking. At the bottom, there are sections for 'News' and 'Events'.

Techniques bioinformatiques

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)

- 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.)

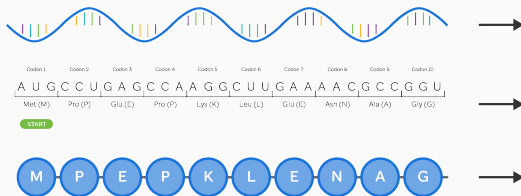
Recherche de motifs

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.)

Alignement de séquence

```
RLA0_METVA  --MIDAKSEHKTIAPWKIEEVNALKELLKSANVIALIDMMEVPAVLOEIRDK
RLA0_METJA  ---METKVKAHVADPKIEEVKTLKGLIKSKPVVAIVDMMDVPAPLOEIRDK
RLA0_PYRAB  -----MAHVAEWKKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRL
RLA0_PYRHO  -----MAHVAEWKKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRL
RLA0_PYRFU  -----MAHVAEWKKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRL
RLA0_PYRKO  -----MAHVAEWKKKKEVEELANLIKSYPVIALVDVAGVPAYPLSKMRDK
RLA0_HALMA  MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRLODMRRD
RLA0_HALVO  MSESEVRQTEVIPQWKREEVDLVDFIESYESVGVVGVAGIPSRLODSMRRE
RLA0_HALSA  MSAEEQRTTEEVPEWKQEEVDELVDLLETYDSVGVVNVGTGIPSKOLODMRRG
RLA0_THEAC  -----MKEVSQKKELVNEITRIKASRSVAIVDTAGIRTRQIQDIRGK
RLA0_THEVO  -----MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTROMODIRAK
RLA0_PICTO  -----MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNS
```

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

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

Mapping et assemblage (de novo)

Whole Genome Sequencing

30-60x Coverage



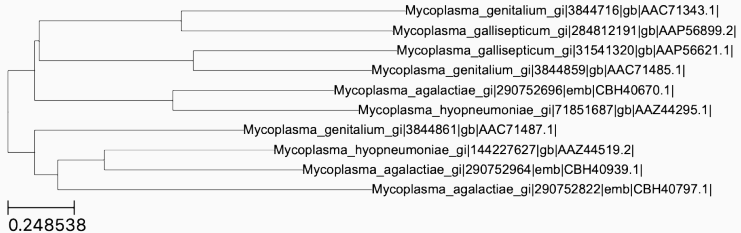
© Copyright 2022 by Jude Children's Research Hospital, a not-for-profit, section 501(c)(3)

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

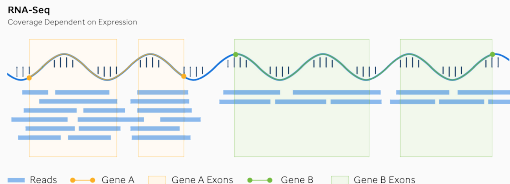
²Benchmarking short sequence mapping tools

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

Phylogénie moléculaire

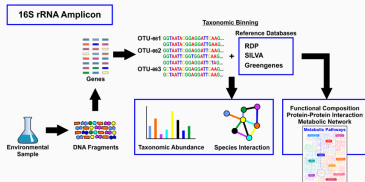


- fasttree
- IQ-TREE (*)
- RaXML
- MEGA
- NGPhylogeny
- visualisateurs : seaview (phylip), figtree, itol (payant)



© Copyright 2002-05, Jude Children's Research Hospital, a not-for-profit, section 501(c)(3)

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

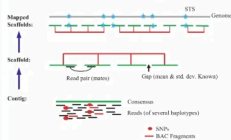


- *species*^a vs. gene-centric
- FROGS (workflow Galaxy, base de données ITS)
- Kraken (bases de données pré-existantes) (*)

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

Algorithmes bioinformatiques

Assemblage de génome *de novo*



Source: Venter, C. et al. 2001

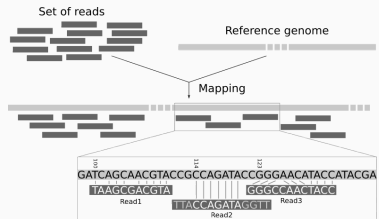
Steps for genome assembly:

- 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

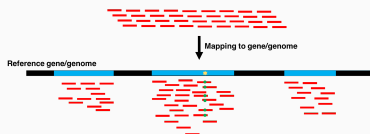
- Données : short et/ou long reads (FASTQ)
- The present and future of *de novo* whole-genome assembly

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



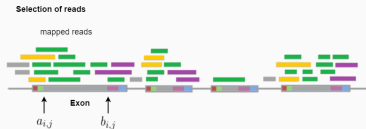
Détection de mutation (variant calling)



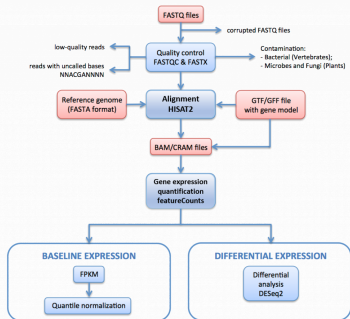
- Données : reads (FASTQ),
génom de référence (FASTA)
- Fichier VCF comprenant les
positions identifiées et les
nucléotides associés (% et
probabilité)
- Haute sensibilité aux
paramètres de filtrage (cf.
tutoriel Galaxy dans le cas
des champignons)

RNA-Seq : mapping & quantification

- Données : reads (FASTQ),
génomme 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



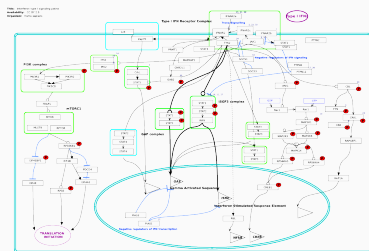
RNA-Seq : analyse différentielle



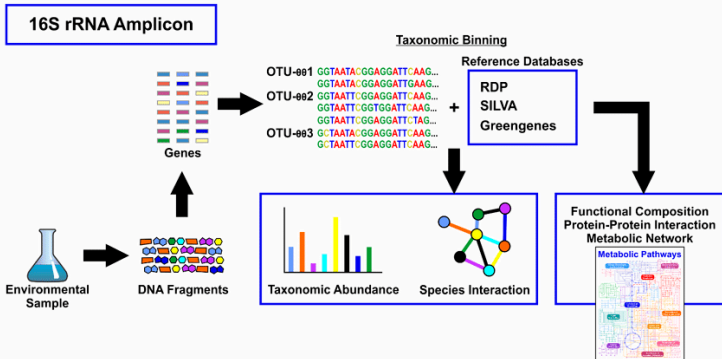
- Données : RPKM ou FPKM
- Approche fréquentiste ou bayésienne pour décider si les données de comptage moyennées sur les réplicats techniques et normalisées pour chaque réplicat biologique sont dûes au hasard ou non (gène sur- ou sous-exprimé par analyse de contraste sur condition de référence).

RNA-Seq : analyse d'enrichissement

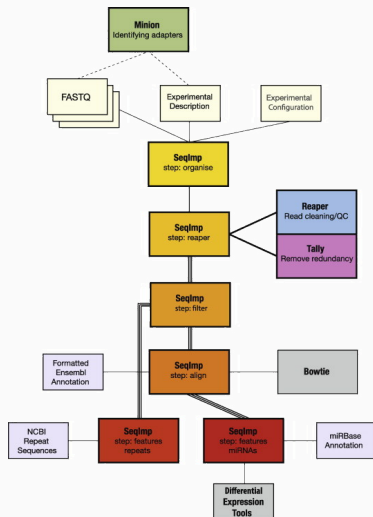
- Données : tableau de quantification, annotation (go-terms, interpro)
- Approche par classification (3 classes/ontologies pour les go-terms : cellular component, biological process or molecular function) et "pathway"/"network" analysis (processus biologiques ou fonction moléculaire, et événements régulatoires)



Métagénomique : Principe général



- 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



- [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.

- [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>