



Introduction à R... en 1h45

Ecole de Bioinformatique AVIESAN/IFB – Novembre 2017

Hugo Varet – hugo.varet@pasteur.fr



Transcriptome & Epigenome Platform – Biomics Pole – Citech
Bioinformatics & Biostatistics Hub – C3BI & USR 3756 CNRS



CNRS UPMC
Station Biologique
Roscoff

R en quelques mots

Langage de programmation qui permet de :

1. manipuler des données : importer, transformer, exporter
2. faire des analyses statistiques plus ou moins complexes : description, exploration, modélisation...
3. créer des (jolies) figures

Disponible sur



Historique :

- 1993 : début du projet R
- 2000 : sortie de R 1.0.0



Avantages et inconvénients

Avantages :

- Souplesse d'utilisation pour réaliser des analyses statistiques
- R est libre et gratuit, même s'il existe maintenant des versions payantes de RStudio (shiny et/ou server)

Inconvénients :



Modes d'utilisation (liste non exhaustive)



Localement via le terminal



Localement via RStudio (utilisation classique)



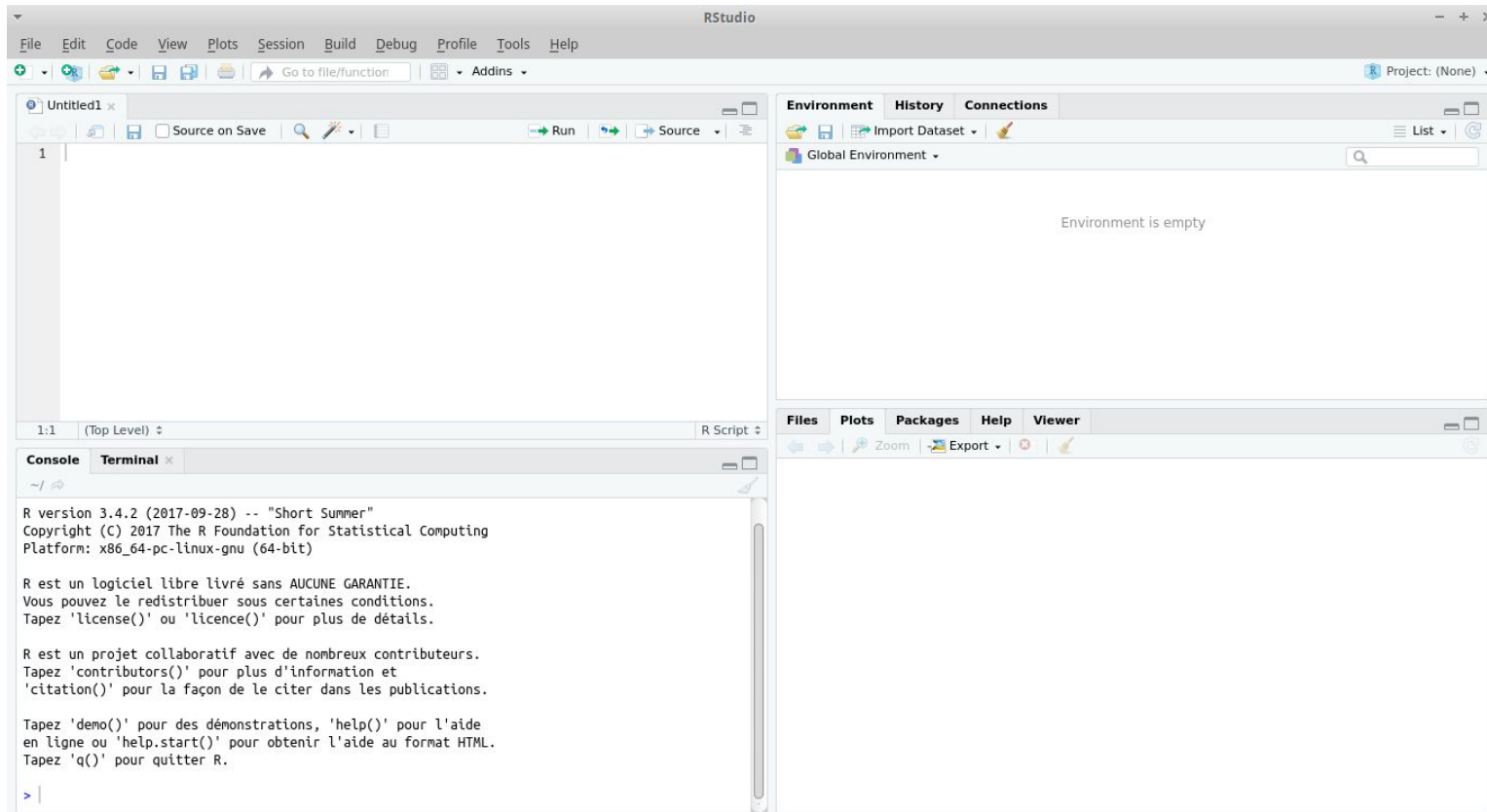
Sur un serveur via le terminal et une connexion ssh



Sur un serveur via un navigateur web pour accéder à RStudio server



- Disponible depuis 2011
- Logiciel facilitant l'utilisation de R via 4 panneaux



- Utilisation avancée : génération de rapports PDF/HTML, création de packages, gestionnaire de versions



Fichiers à récupérer

1. Télécharger et enregistrer le script R :

<https://tinyurl.com/intro-r-roscoff>

→ contient du code très simple couvrant les bases de R.

2. Fichiers de données (plus tard) :

- **Comptages de lectures (“reads”) par gène**
 - `rnaseq_data.txt`
 - `rnaseq_data.csv`
- **Annotations génomiques en format [GTF](#) (description sur [ensembl.org](#))**
 - `Saccharomyces_cerevisiae.R64-1-1.90.gtf`

Ces fichiers sont disponibles sur le serveur de Roscoff dans le répertoire partagé :

`/projet/sbr/ggb/intro_R`



Connexion à RStudio server à Roscoff

<http://r.sb-roscoff.fr>

Sign in to RStudio

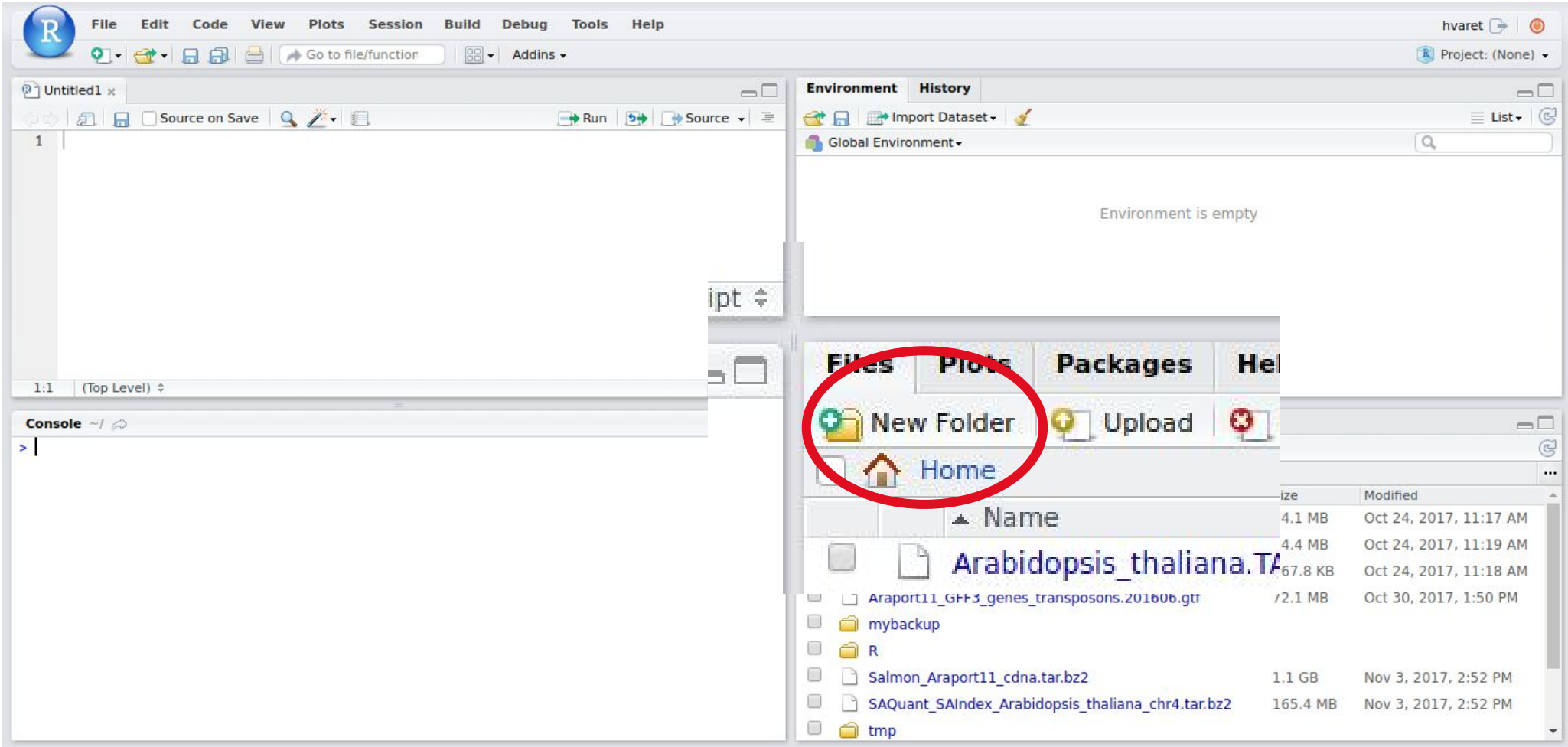
Username:

Password:

Stay signed in



Création d'un dossier "intro_R"



The screenshot displays the RStudio environment. The main editor window is titled 'Untitled1' and contains a single line of code: `ipt`. The 'Environment' pane on the right shows 'Global Environment' and 'Environment is empty'. The 'Files' pane at the bottom right shows a file explorer view with a red circle highlighting the 'New Folder' button. The file list includes:

Name	Size	Modified
Arabisopsis_thaliana.TA	67.8 KB	Oct 24, 2017, 11:18 AM
Araport11_GFF3_genes_transposons.201606.gtf	72.1 MB	Oct 30, 2017, 1:50 PM
mybackup		
R		
Salmon_Araport11_cdna.tar.bz2	1.1 GB	Nov 3, 2017, 2:52 PM
SAQuant_SAIIndex_Arabidopsis_thaliana_chr4.tar.bz2	165.4 MB	Nov 3, 2017, 2:52 PM
tmp		



Téléversement (“upload”) et ouverture du script R

The screenshot displays the RStudio environment. The main editor window is titled 'Untitled1' and contains a single line of code starting with '1'. The Environment pane on the right shows 'Global Environment' and 'Environment is empty'. The Packages pane at the bottom right is open, showing a list of installed and available packages. The 'Upload' button, represented by a yellow lightning bolt icon, is circled in red. Below the Packages pane, a file browser view shows a directory structure with a table of files.

Name	Size	Modified
Arabidopsis_thaliana.TAIR10	72.1 MB	Oct 30, 2017, 1:50 PM
R		
Salmon_Araport11_cdna.tar.bz2	1.1 GB	Nov 3, 2017, 2:52 PM
SAQuant_SAIIndex_Arabidopsis_thaliana_chr4.tar.bz2	165.4 MB	Nov 3, 2017, 2:52 PM
tmp		



Tutoriel

Exploration guidée des commandes R en parcourant le script R que nous avons téléchargé.



Exercice

1. **Copier le fichier** `Saccharomyces_cerevisiae.R64-1-1.90.gtf` **disponible dans** `/projet/sbr/ggb/intro_R/` **vers votre espace de travail.**
2. **Charger ce fichier dans R et afficher les 10 premières lignes. Remarque : le fichier contient les colonnes suivantes :**
 - `seqname,`
 - `source,`
 - `feature,`
 - `start,`
 - `end,`
 - `score,`
 - `strand,`
 - `frame,`
 - `attribute`
3. **Afin de gagner en lisibilité, supprimer la dernière colonne du tableau. Afficher à nouveau les 10 premières lignes. Combien y en a-t-il au total?**



Exercice (suite)

4. Combien le génome contient-il de gènes ? Créer une data.frame contenant uniquement les gènes.
5. Pour cette nouvelle data.frame, créer une variable donnant la longueur de chacun de ces gènes ($\text{end} - \text{start} + 1$)
6. Quelle est la longueur moyenne des gènes de cet organisme ? Tracer l'histogramme de ces longueurs.
7. A l'aide de boxplots, tracer la distribution de la longueur des gènes pour chacun des chromosomes.

