Biogenouest

ifb
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

EMBRC FRANCE
CENTRE NATIONAL DE RESSOURCES BIOLOGIQUES MARINES

OCEANOMICS

IDEALG
seaweed for the future

Abims 4

15/05/2014

Galaxy

Initiation

Le Corguillé – Monsoor

V1.01

UPMC SORBONNE UNIVERSITÉS

cnrs

INTRODUCTION / PROBLÉMATIQUE

Introduction

- Setup TP
  - http://galaxy.sb-roscoff.fr
  - Compte
    - login@sb-roscoff.fr
    - *********
- Support de cours :

http://application.sb-roscoff.fr/download/fr2424/abims/lecorguille/cours/galaxy-initiation.pdf

Introduction

Sélectionner votre niveau :

Level 1

“Je veux connaître l'expression de mes gènes”

Level 2

“Je veux faire un mapping et un comptage de mes reads sur mon génome de référence”

Level 3

“Je veux lancer les outils tophat2 et cufflinks

J'ai des fastq et mon génome au format fasta et gtf”

Level 4

"Je veux un espace projet avec 1To car je pense faire du ssh pour lancer tophat2 sur 8 procs via qsub

et soumettre le fichier bam à mon génome avec cufflinks

Sinon, pas besoin de vous :P"

Level 5

“J'ai un tas d'outils sympas !

Mais je suis le seul à pouvoir les lancer.

Des commentaires ?”

Pourquoi ?

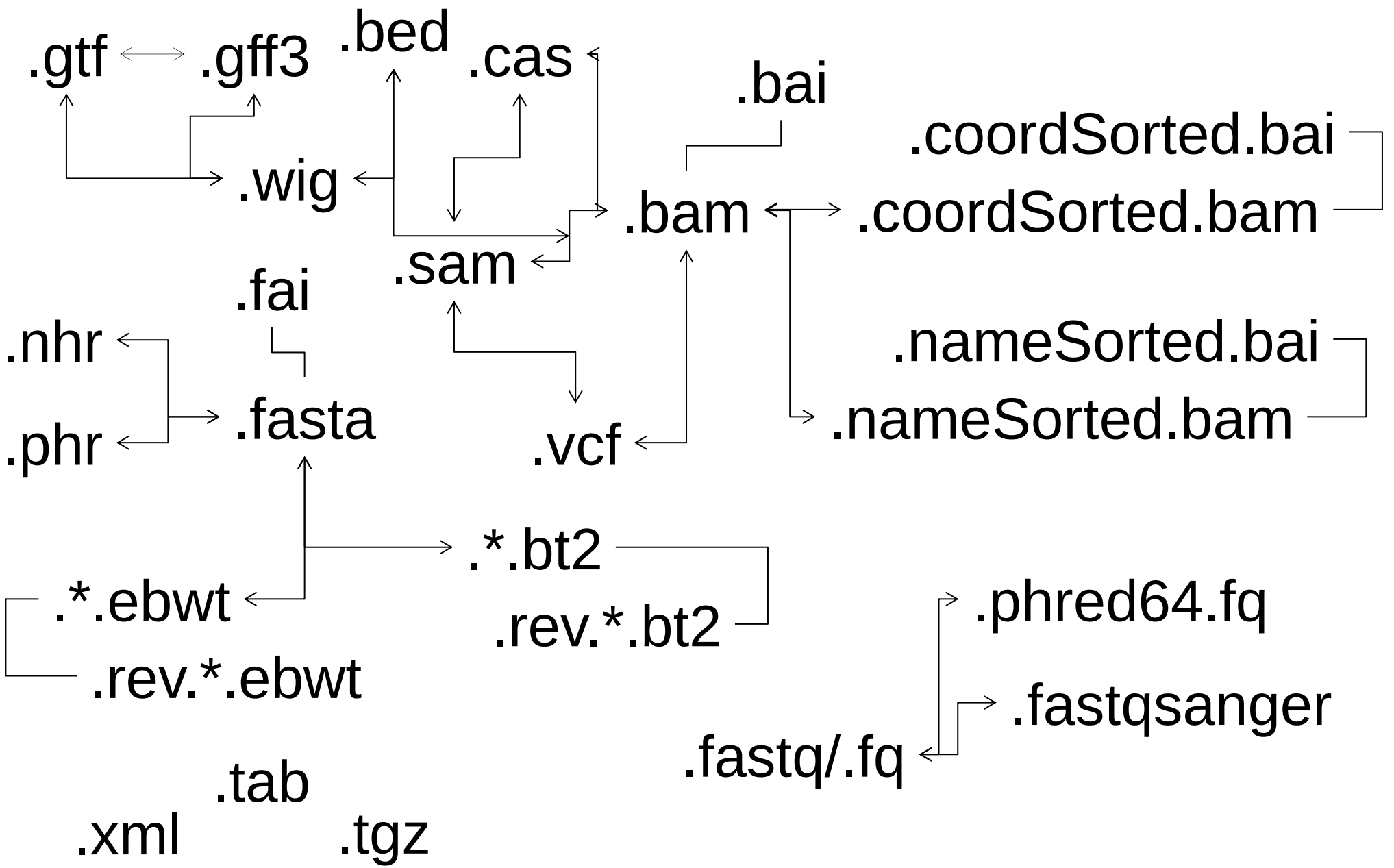
```
login@sbr4-1042:~$ ssh -Y login@bioinfo.sb-roscoff.fr
[...]
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
[login@n0 mapping]$ ls
accepted_hits.bam  junctions.bed    qsub.err  unmapped.bam
deletions.bed      logs             qsub.out
insertions.bed     prep_reads.info  tmp
[login@n0 mapping]$ cd ..
[login@n0 mapping]$ mkdir cufflinks
```

```
login@sbr4-1042:~$ ssh -Y login@bioinfo.sb-roscoff.fr
[...]
[login@n0 ~]$ cdprojet
[login@n0 login]$
[login@n0 mapping
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-r
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v1                                        .R1-2.fq
-GTF ../input/par
-num-threads 8

[login@n0 mapping                                       t.qsub
Your job 5338969
[login@n0 mapping
accepted_hits.bar                                       l.bam
deletions.bed       logs                qsub.out
insertions.bed      prep_reads.info     tmp
[login@n0 mapping]$ cd ..
[login@n0 mapping]$ mkdir cufflinks
```

Introduction

.gtf .gff3 .bed .cas .bai .coordSorted.bai .wig .bam .coordSorted.bam .sam .fai .nhr .nameSorted.bai .fasta .phr .nameSorted.bam .vcf .*.bt2 .*.ebwt .rev.*.bt2 .phred64.fq .rev.*.ebwt .fastqsanger .fastq/.fq .tab .xml .tgz

Introduction



Introduction

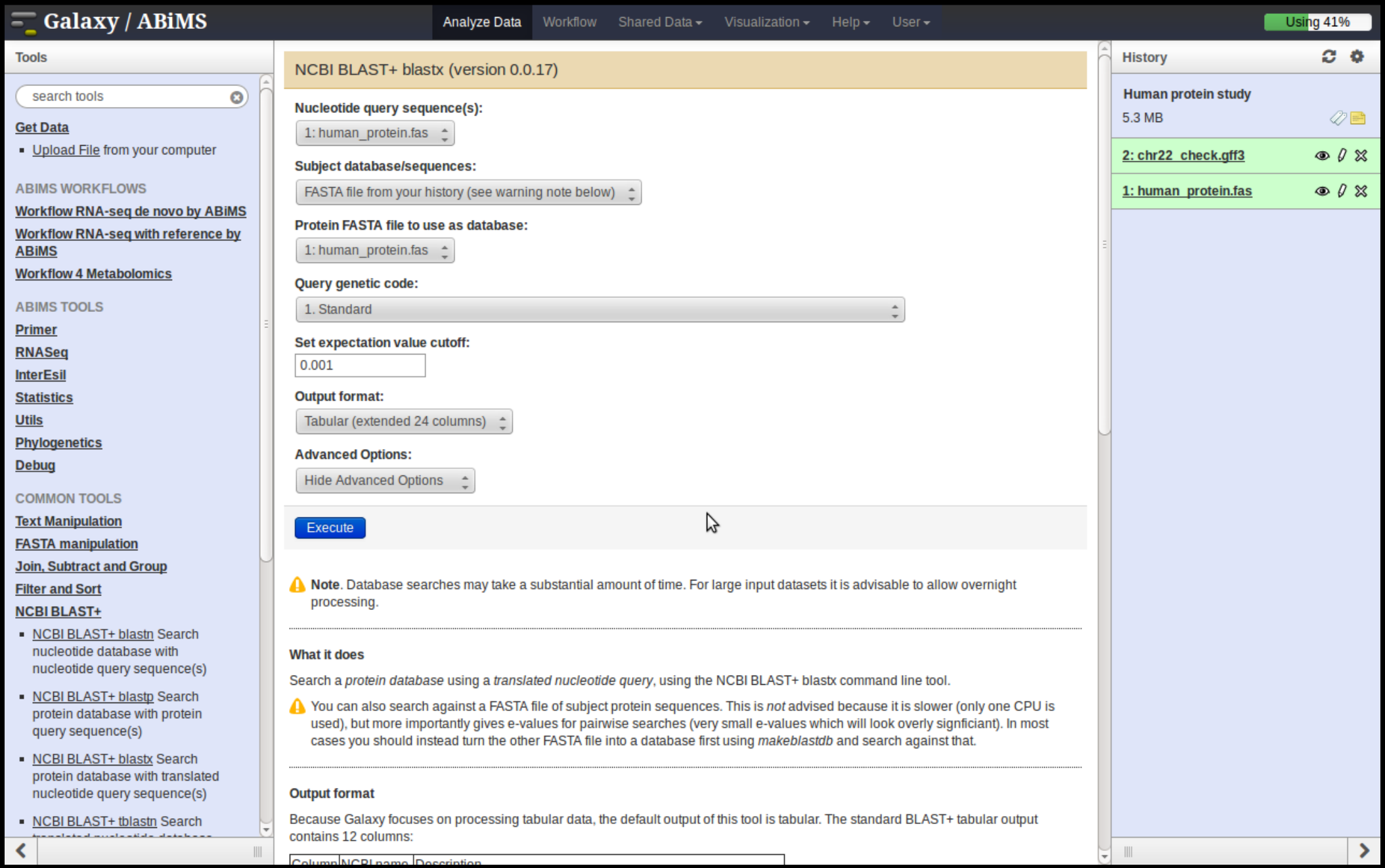
- Les outils avec interface graphique click bouton sous windows
  - + très ergonomique
  - – trop ergonomique → manque de souplesse
  - – faut pas rêver ! Vous avez déjà vu un thésard trouver le temps pour faire de beaux boutons verts ?

- Les outils en ligne sur Internet
  - + très ergonomique
  - – trop ergonomique → manque de souplesse
  - – une infime part des outils disponibles
  - – réparti un peu partout sur les différents sites universitaires
  - – souvent limité en terme de taille de soumission
  - – il ne faut pas être parano

Introduction

- Les outils en ligne de commande
  - + représente la quasi majorité des outils scientifiques
  - + très complet en terme de paramètre
  - + peuvent tourner sur des clusters de calcul
  - + les g33ks adorent car automatisable, workflowsable, ...
  - – nécessite un minimum de connaissance en linux
  - – manque cruellement d'ergonomie

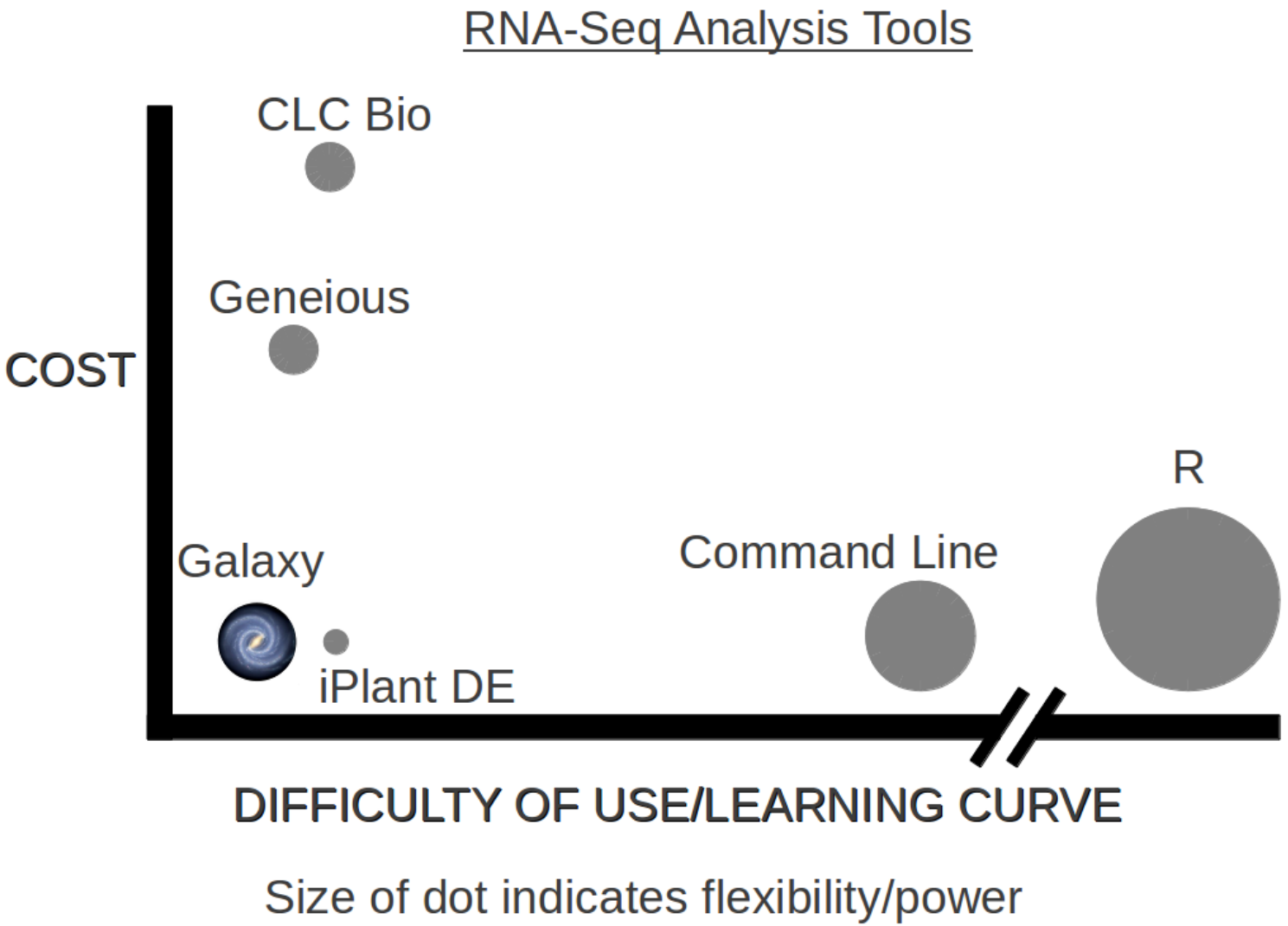
INTRODUCTION / GALAXY


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 41%
Tools
search tools
Get Data
Upload File from your computer
ABIMS WORKFLOWS
Workflow RNA-seq de novo by ABiMS
Workflow RNA-seq with reference by ABiMS
Workflow 4 Metabolomics
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Debug
COMMON TOOLS
Text Manipulation
FASTA manipulation
Join, Subtract and Group
Filter and Sort
NCBI BLAST+
NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
NCBI BLAST+ blastp Search protein database with protein query sequence(s)
NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
NCBI BLAST+ tblastn Search
NCBI BLAST+ blastx (version 0.0.17)
Nucleotide query sequence(s):
1: human_protein.fas
Subject database/sequences:
FASTA file from your history (see warning note below)
Protein FASTA file to use as database:
1: human_protein.fas
Query genetic code:
1. Standard
Set expectation value cutoff:
0.001
Output format:
Tabular (extended 24 columns)
Advanced Options:
Hide Advanced Options
Execute
Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.
What it does
Search a protein database using a translated nucleotide query, using the NCBI BLAST+ blastx command line tool.
You can also search against a FASTA file of subject protein sequences. This is not advised because it is slower (only one CPU is used), but more importantly gives e-values for pairwise searches (very small e-values which will look overly signficiant). In most cases you should instead turn the other FASTA file into a database first using makeblastdb and search against that.
Output format
Because Galaxy focuses on processing tabular data, the default output of this tool is tabular. The standard BLAST+ tabular output contains 12 columns:
Column
NCBI name
Description
History
Human protein study
5.3 MB
2: chr22_check.gff3
1: human_protein.fas

Introduction

- Galaxy c'est …
  - Plus besoin de lancer une ligne de commande dans un terminal
  - Plus besoin de connaître la programmation ou le scripting
  - Des jobs soumis de manière transparente sur un cluster de calcul
  - Un gestionnaire d'historique et de données sécurisées
  - Un système de partage de données ou de protocoles
  - Des boîtes à outils dans plusieurs domaines de la bioinformatique
    - NGS
    - Metabolomique
    - Statistique
    - Analyse d'image
    - Chimie
    - Etc ...
  - Une interface web

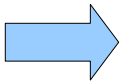
Introduction



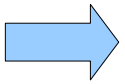
Introduction

- Pourquoi Galaxy ?
  - Accessibilité
    - Accès à des outils de bioinformatique sans connaissance en informatique
    - Ergonomie à géométrie variable
    - Modularité
  - Reproductibilité
    - Traçabilité des paramètres
  - Transparence
    - Partage des données et protocoles

Introduction

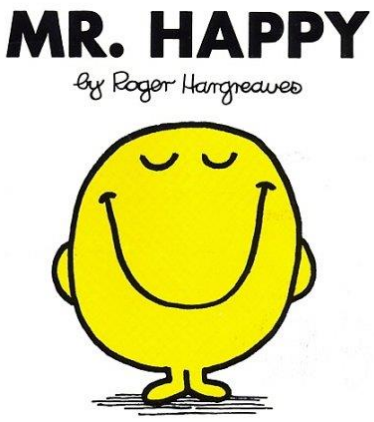


```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```



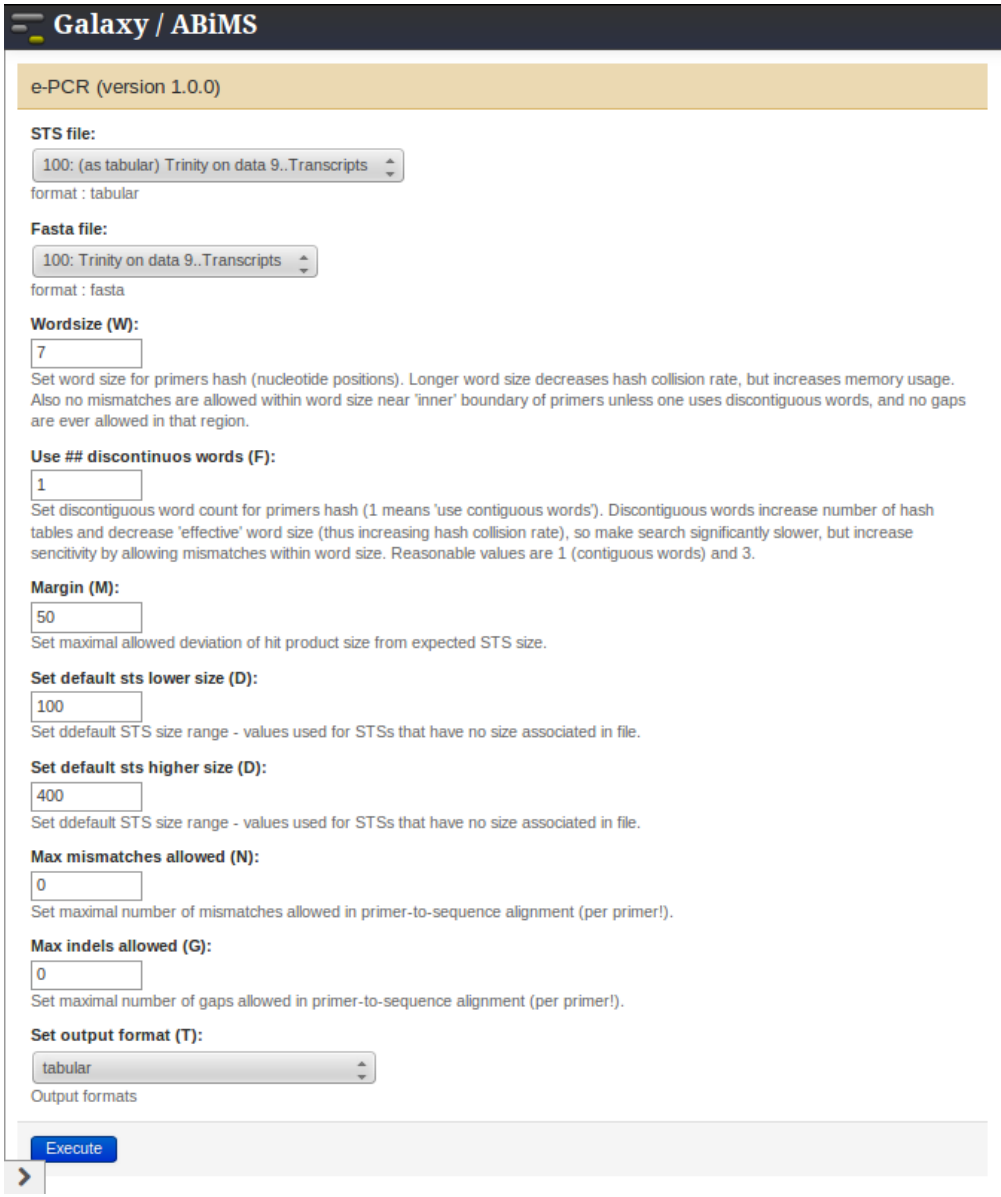
Introduction

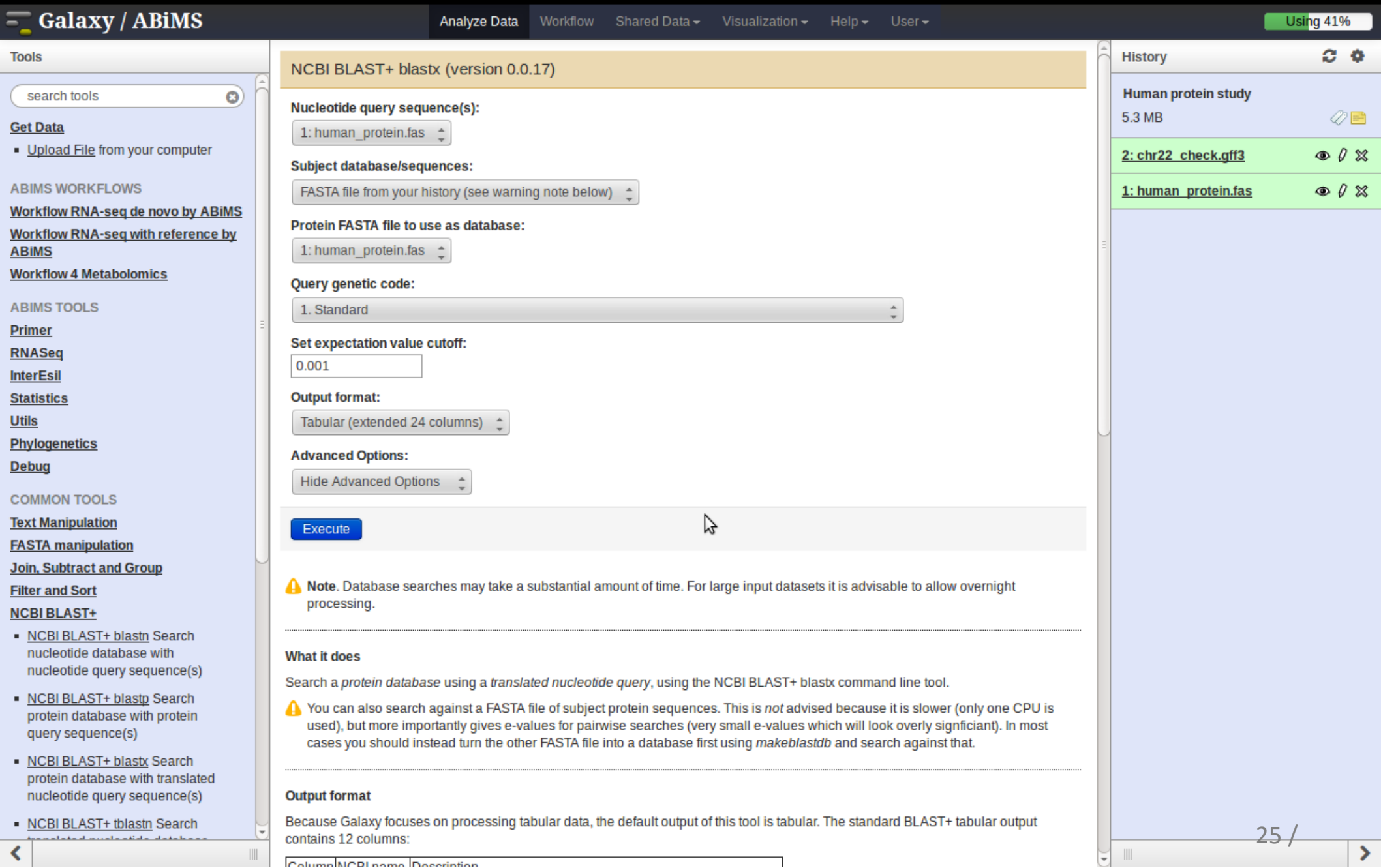
```
[login@n0 ~]$ cdprojet
[login@n0 login]$ cd 13-07-29-panda/tmp/mapping
[login@n0 mapping]$ cat tophat.qsub
#!/bin/bash
#$ -S /bin/bash
#$ -M login@sb-roscoff.fr
#$ -m bea
#$ -V
#$ -cwd
#$ -o qsub.out
#$ -e qsub.err

tophat2 panda_v121029 ../input/IllR1-1.fq ../input/IllR1-2.fq
-GTF ../input/panda_v121029.gtf --b2-sensitive -r 100
-num-threads 8

[login@n0 mapping]$ qsub -q long.q -pe thread 8 tophat.qsub
Your job 5338969 ("tophat.qsub") has been submitted
```



Introduction

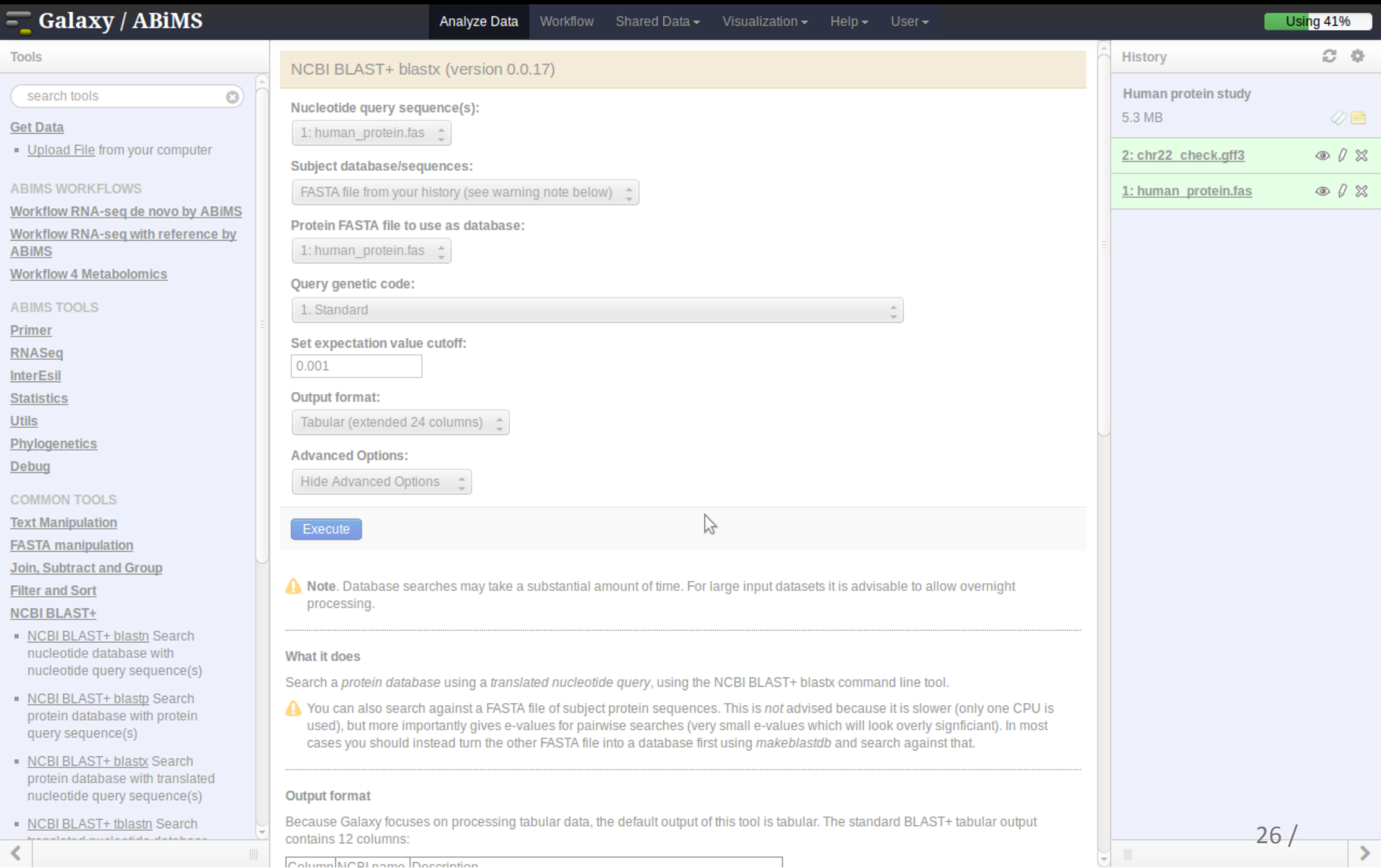
```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
          -m ##     Margin (default 50)
          -w ##     Wordsize  (default 7)
          -n ##     Max mismatches allowed (default 0)
          -g ##     Max indels allowed (default 0)
          -f ##     Use ## discontiguos words, slow if
                              ##>1
          -o ##     Set output file
          -t ##     Set output format:
                    1 - classic, range (pos1..pos2)
                    2 - classic, midpoint
                    3 - tabular
                    4 - tabular with alignment in
comments
                              (slow)
          -d##-##   Set default size range
                              (default 100-350)
          -p +-     Turn hits postprocess on/off
          -v ##     Verbosity flags
          -a a|f    Use presize alignmens (only if
                              gaps>0), slow
                     a - Allways or f - as Fallback
          -x +-     Use 5'-end lowercase masking of
                              primers (default -)
          -u +-     Uppercase all primers (default -)

[...]
```




Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 41%
Tools
search tools
Get Data
Upload File from your computer
ABIMS WORKFLOWS
Workflow RNA-seq de novo by ABiMS
Workflow RNA-seq with reference by ABiMS
Workflow 4 Metabolomics
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Debug
COMMON TOOLS
Text Manipulation
FASTA manipulation
Join, Subtract and Group
Filter and Sort
NCBI BLAST+
NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
NCBI BLAST+ blastp Search protein database with protein query sequence(s)
NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
NCBI BLAST+ tblastn Search
NCBI BLAST+ blastx (version 0.0.17)
Nucleotide query sequence(s):
1: human_protein.fas
Subject database/sequences:
FASTA file from your history (see warning note below)
Protein FASTA file to use as database:
1: human_protein.fas
Query genetic code:
1. Standard
Set expectation value cutoff:
0.001
Output format:
Tabular (extended 24 columns)
Advanced Options:
Hide Advanced Options
Execute
Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.
What it does
Search a protein database using a translated nucleotide query, using the NCBI BLAST+ blastx command line tool.
You can also search against a FASTA file of subject protein sequences. This is not advised because it is slower (only one CPU is used), but more importantly gives e-values for pairwise searches (very small e-values which will look overly signficiant). In most cases you should instead turn the other FASTA file into a database first using makeblastdb and search against that.
Output format
Because Galaxy focuses on processing tabular data, the default output of this tool is tabular. The standard BLAST+ tabular output contains 12 columns:
History
Human protein study
5.3 MB
2: chr22_check.gff3
1: human_protein.fas

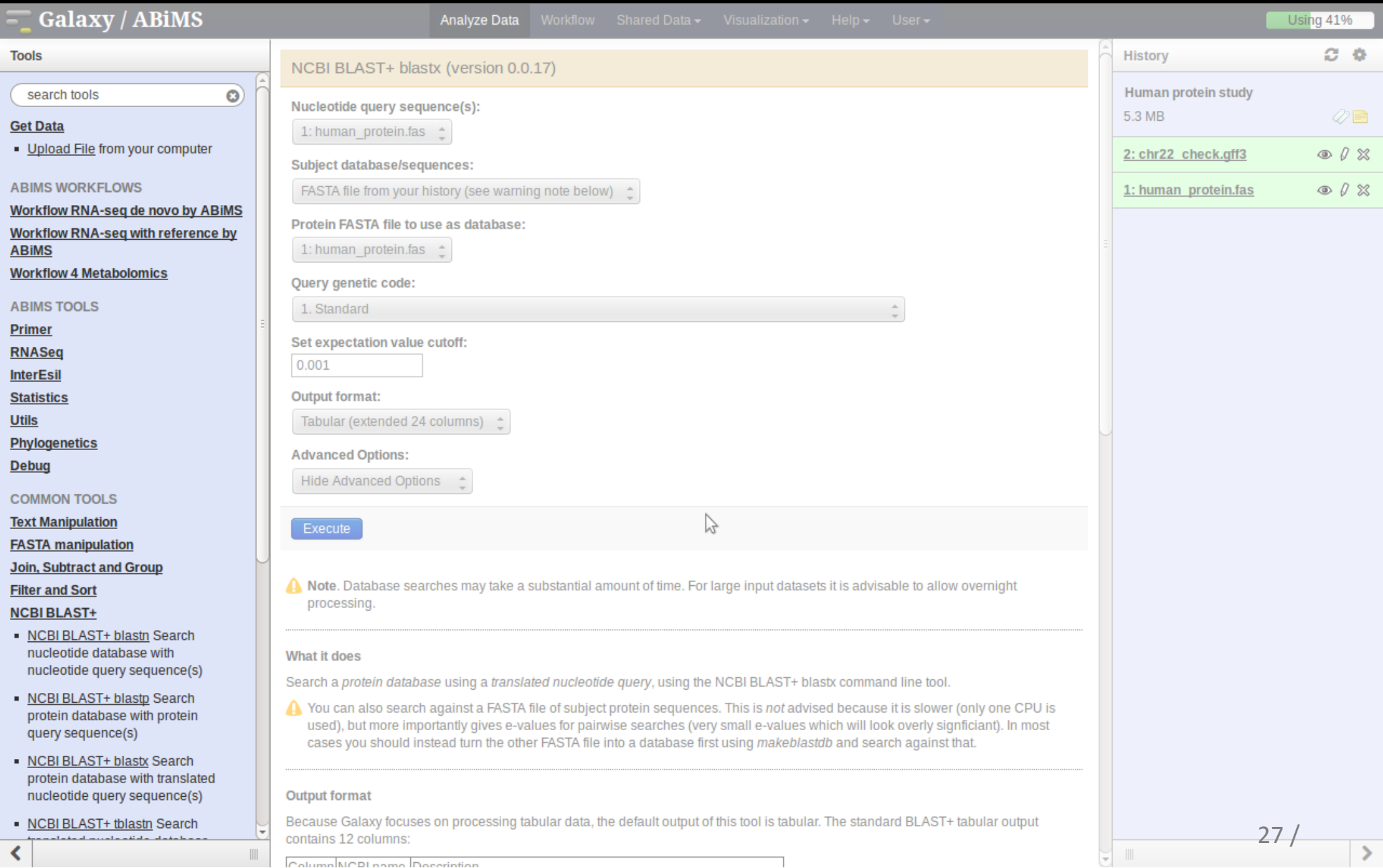
L'interface de Galaxy

le menu



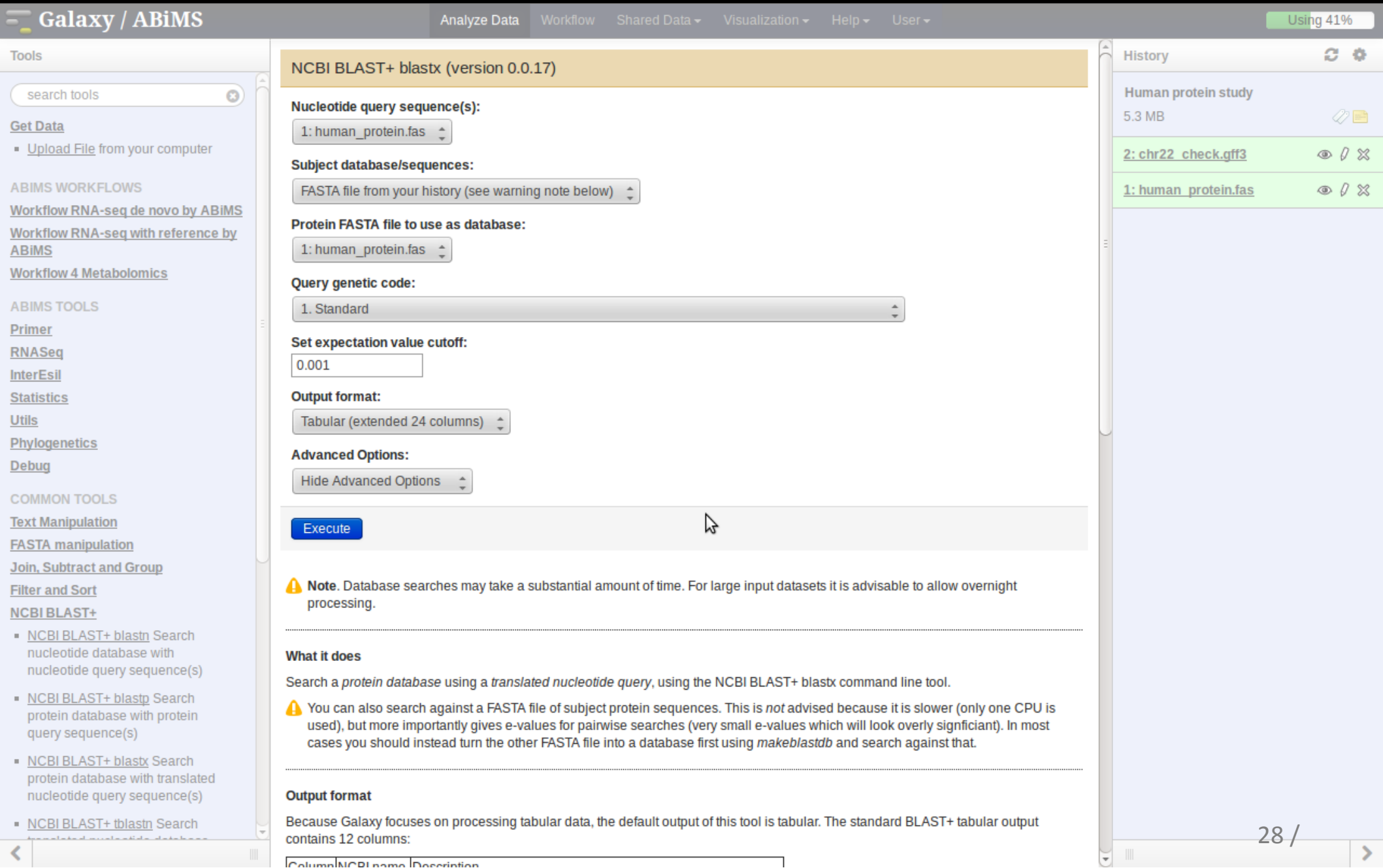
L'interface de Galaxy

la liste des outils



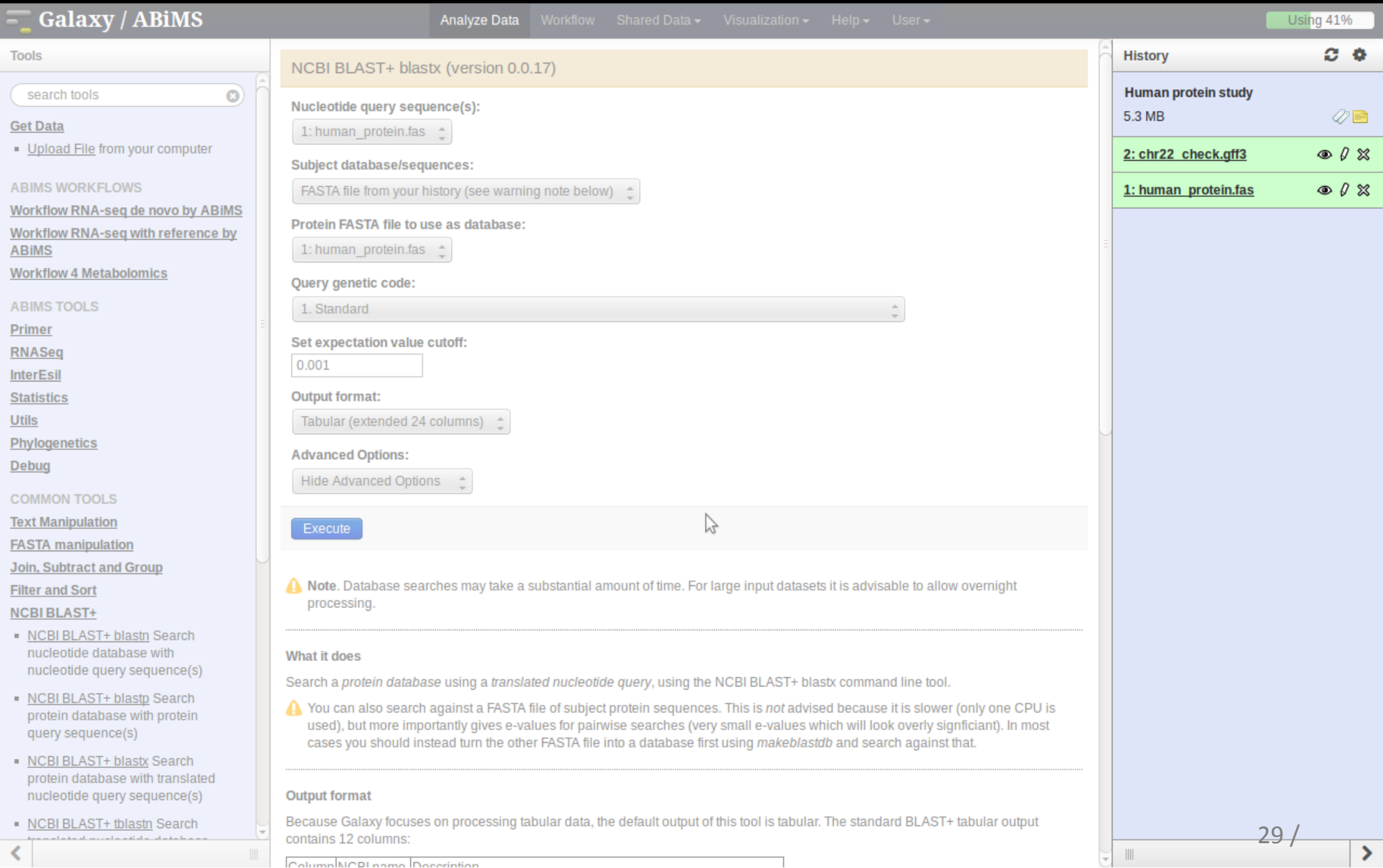
L'interface de Galaxy

formulaire / visualisation / information diverse



L'interface de Galaxy

l'historique

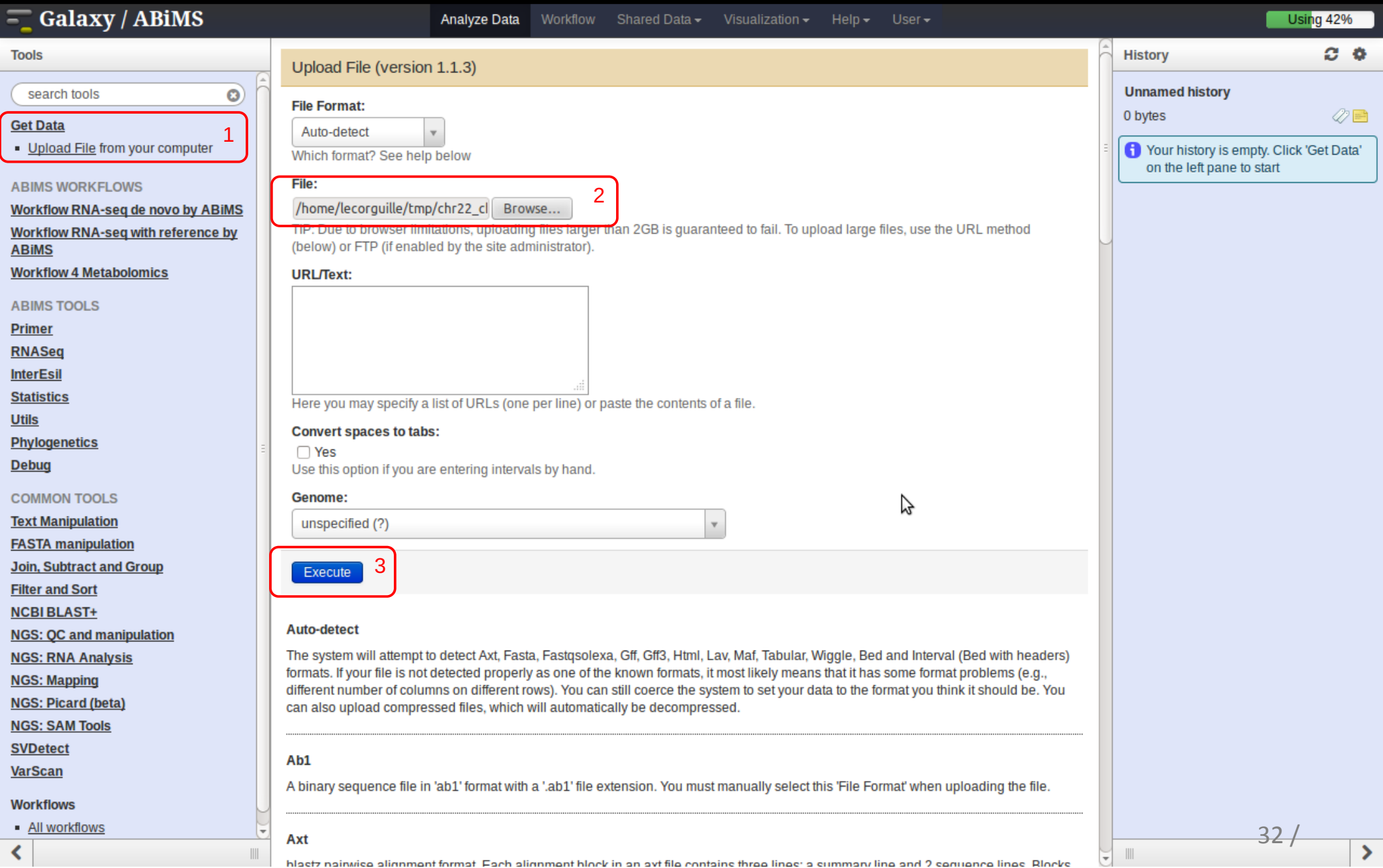


IMPORT DES DONNÉES

IMPORT DES DONNÉES

< 2 GO

Import d'un fichier de moins de 2 Go stocké sur votre disque local

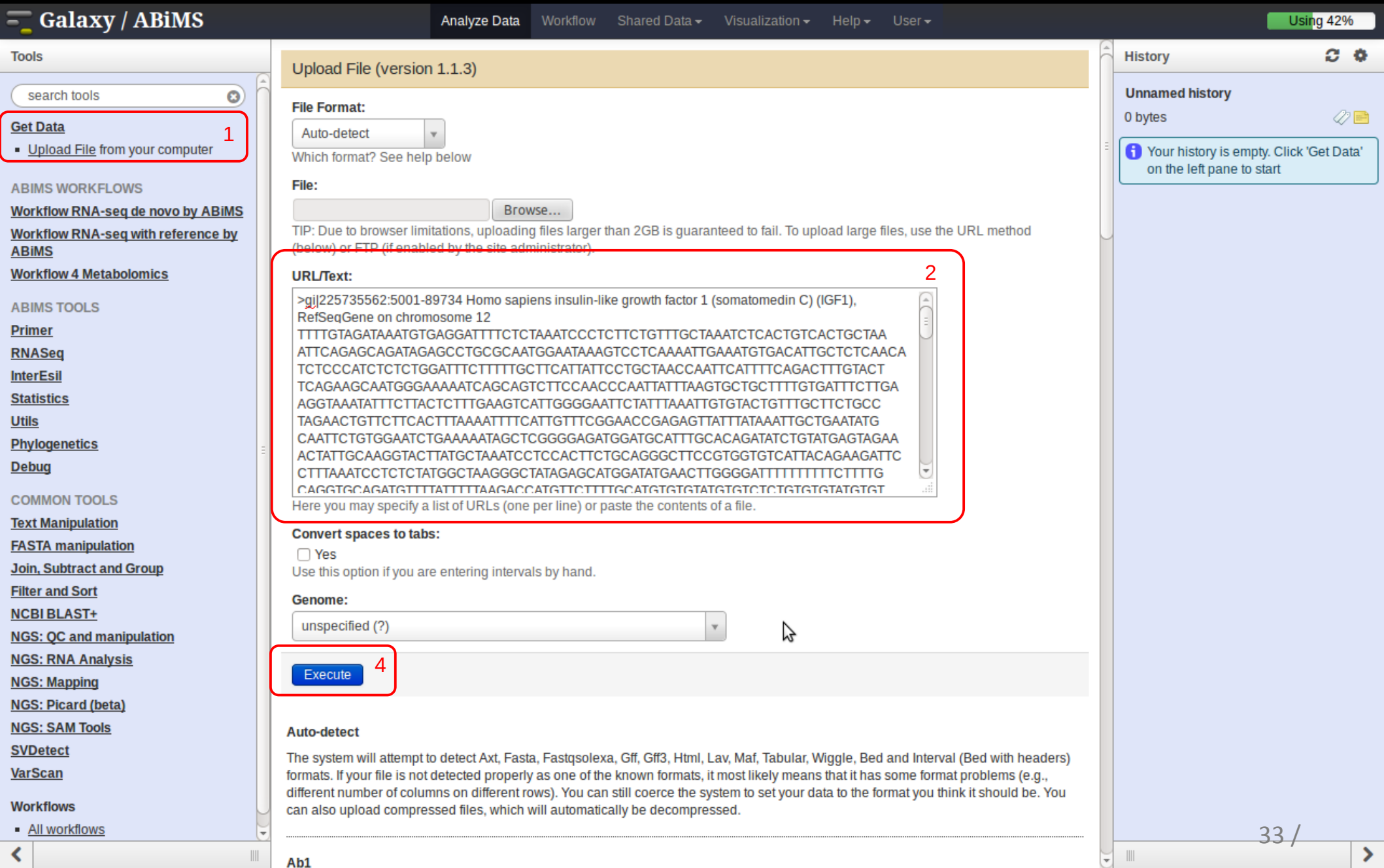


Import par copier/coller dans la zone de texte

Il est aussi possible d'y recopier une url ftp d'un fichier en ligne zippé (zip, tgz, gz)

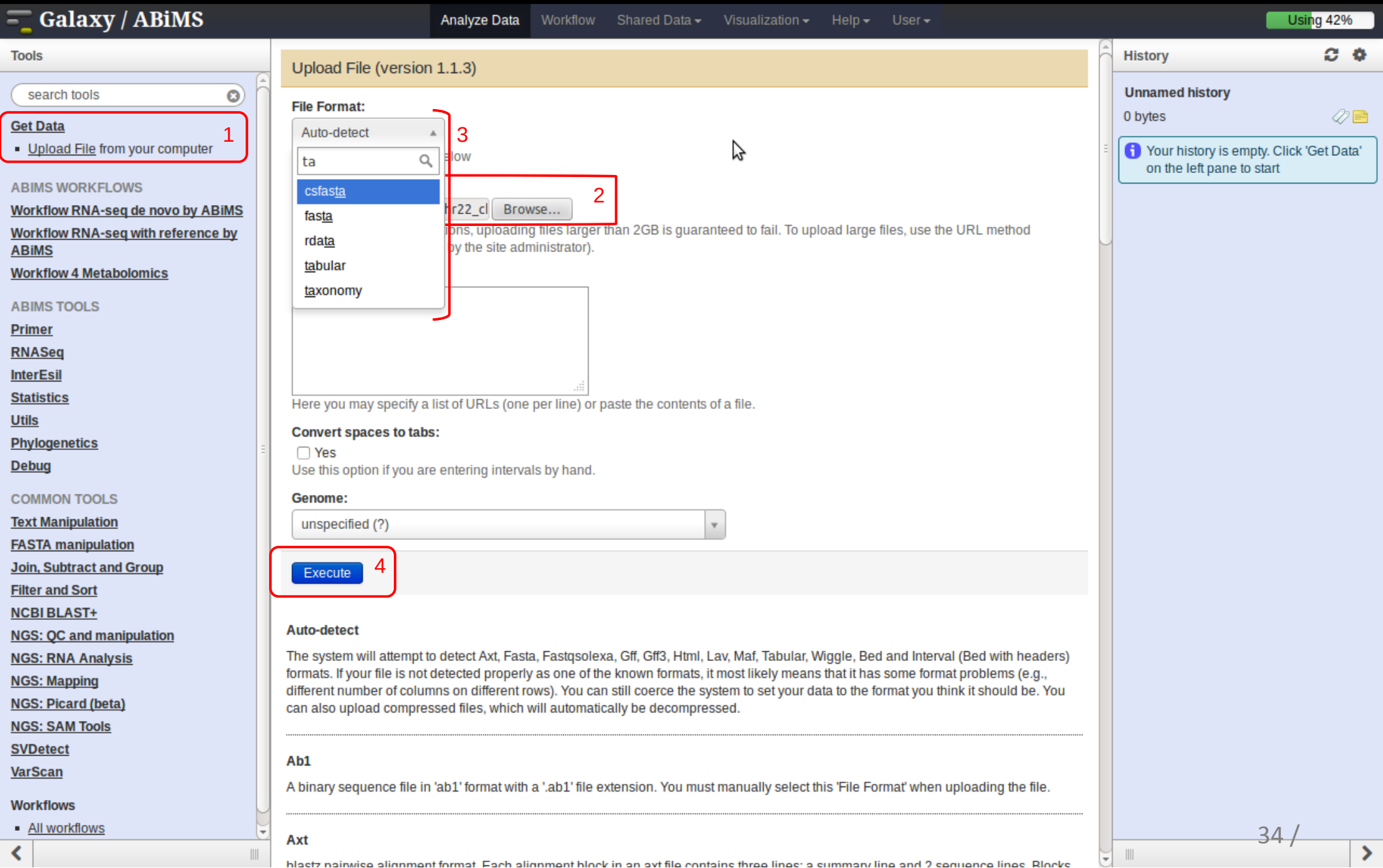
ex : ftp://ftp.ncbi.nih.gov/blast/db/FASTA/mito.aa.gz

Mais attention à la taille !



Si la détection automatique du format échoue ou pour forcer un format :

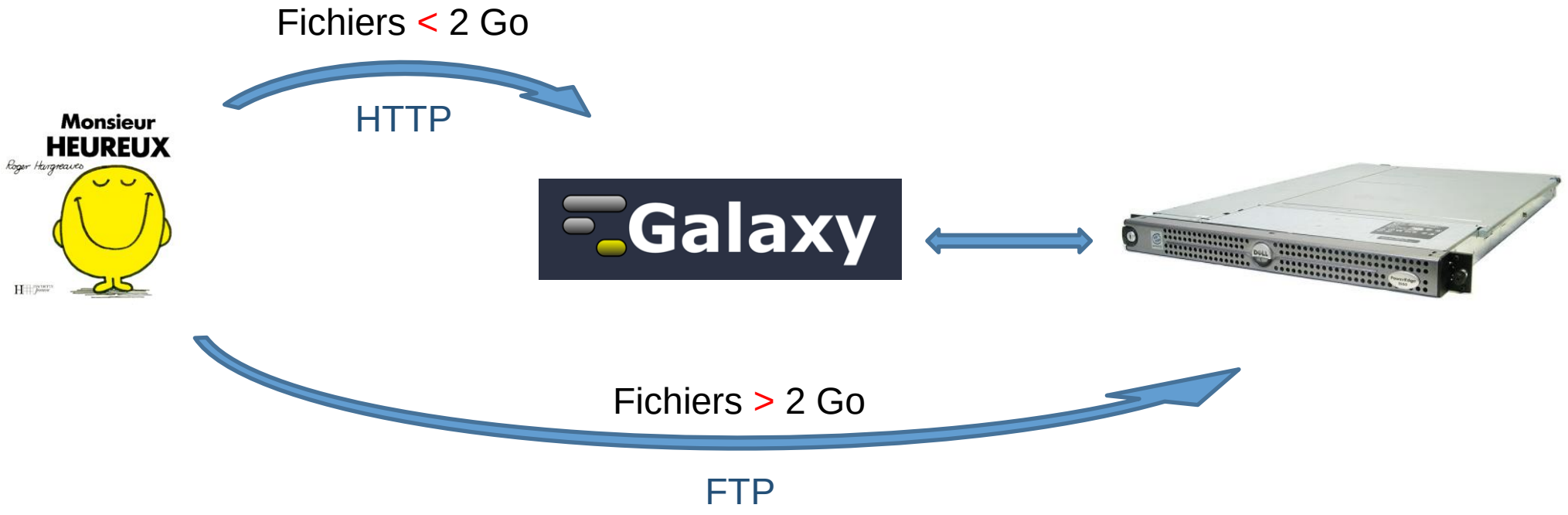
- tabular → gff
- fastq → fastqsanger
- xml → blastxml



IMPORT DES DONNÉES
> 2 GO - PART I – CÔTÉ TRANSFERT

Import FTP

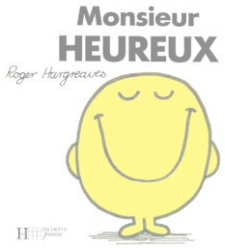
- Création d'une librairie
  - L'import des fichiers supérieurs à 2 Go doivent être déposés sur le serveur via le protocole FTP (File Transfert Protocol)



Import FTP

- Création d'une librairie
  - L'import des fichiers supérieurs à 2 Go doivent être déposés sur le serveur via le protocole FTP (File Transfert Protocol)

http://abims.sb-roscoff.fr/galaxyproject



HTTP

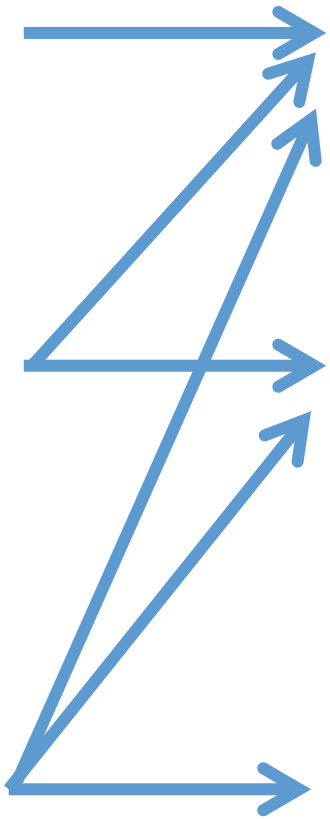
Galaxy

Fichiers > 2 Go

FTP

Import FTP

- Les programmes de FTP



Cyberduck

Libre FTP, SFTP, WebDAV & cloud storage browser for Mac & Windows.

WinSCP

Import FTP

- Les programmes de FTP

Avoid:
Malwares inside

Cyberduck

Libre FTP, SFTP, WebDAV & cloud storage browser for Mac & Windows.

WinSCP

Import FTP

- Les programmes de FTP

FileZilla

The FREE FTP solution

Cyberduck

Libre FTP, SFTP, WebDAV & cloud storage browser for Mac & Windows.

WinSCP

Import FTP

- Les paramètres
  - Server / Host → bioinfo.sb-roscoff.fr
  - Protocol → SFTP ou SCP
  - Username / Login → Abims
  - PATH →

    /projet/externe/<group>/<login>/galaxylibrary

Import FTP

- Les

  –

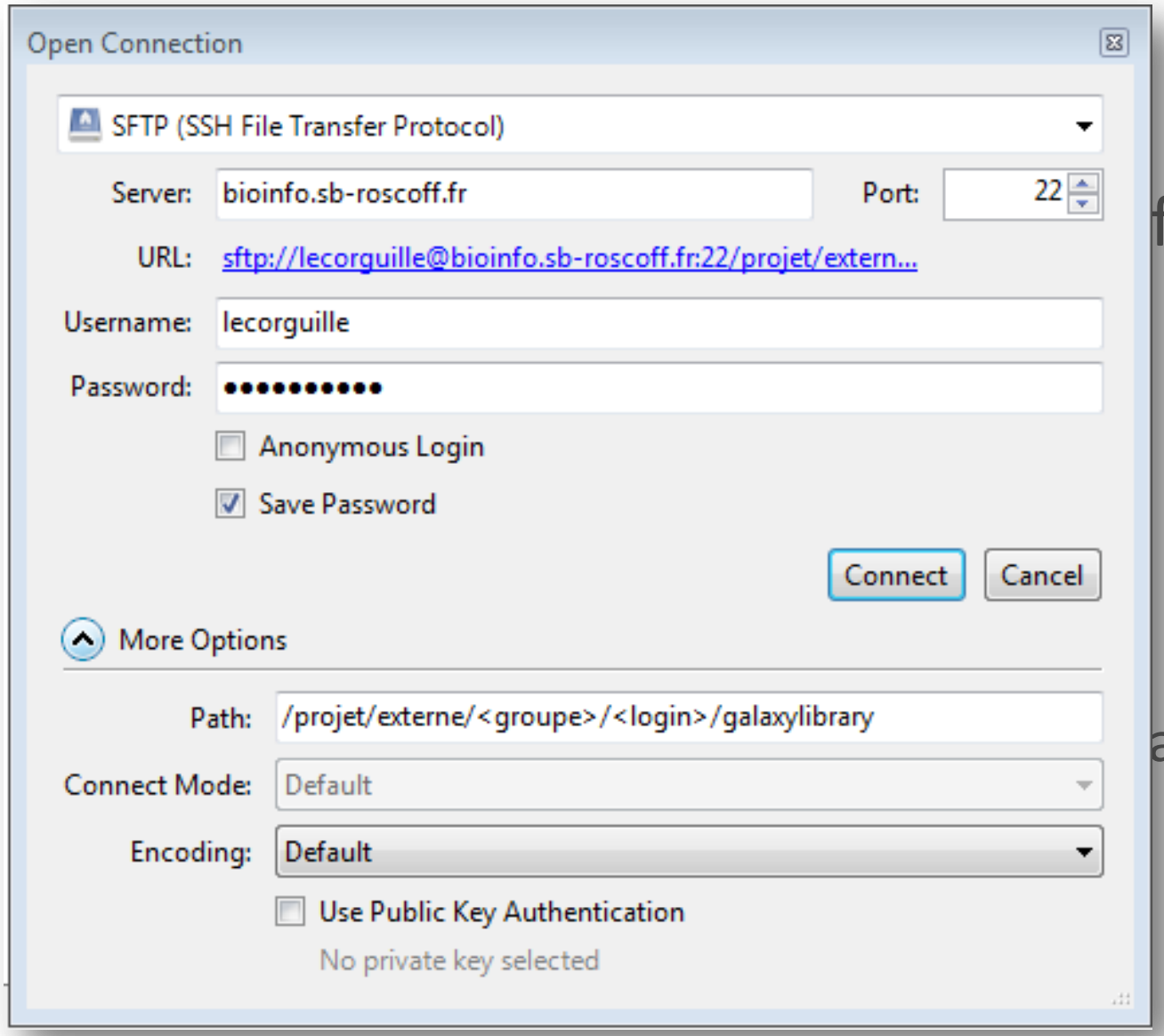
  –

  –

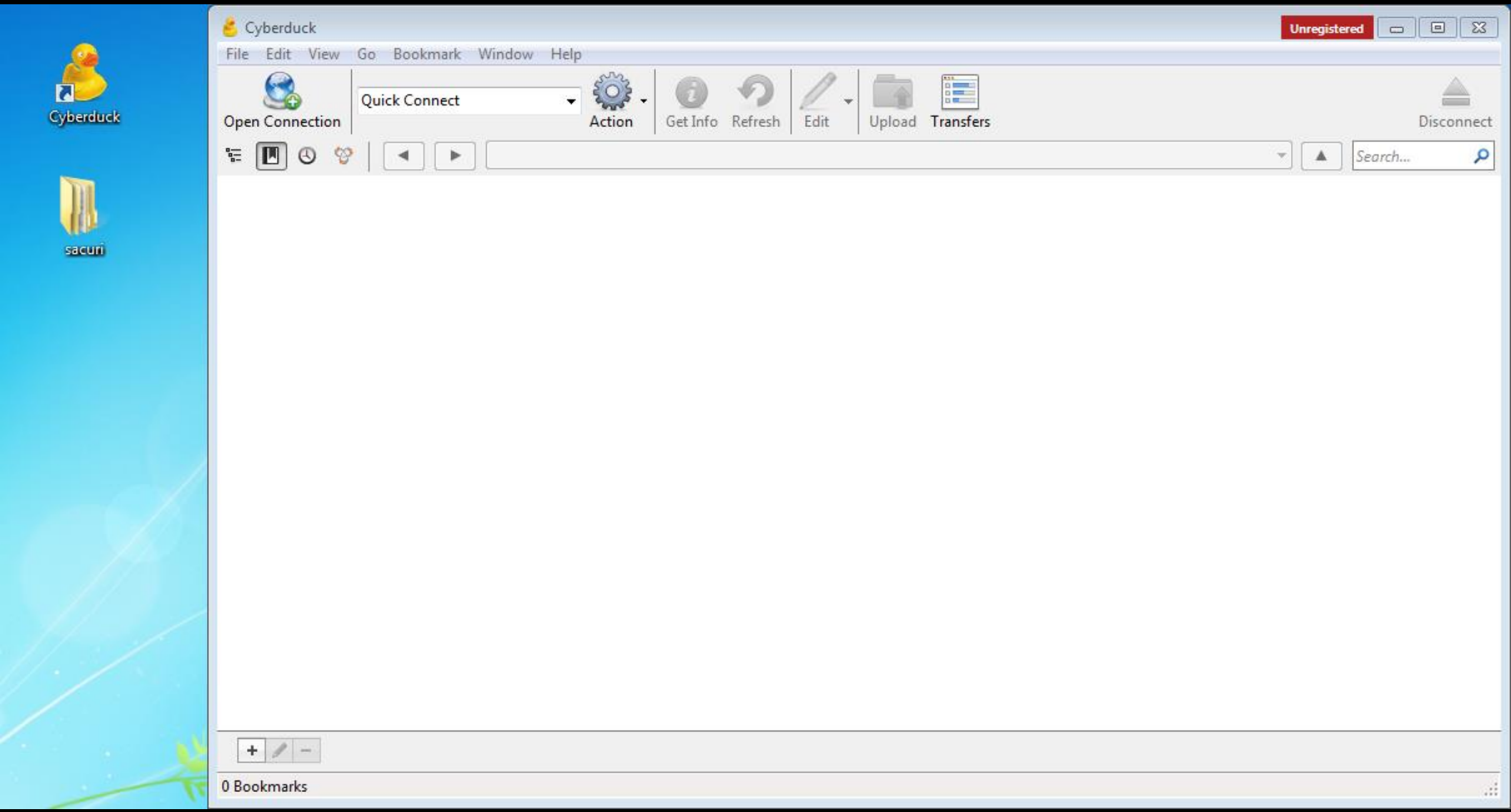
  –

ff.fr

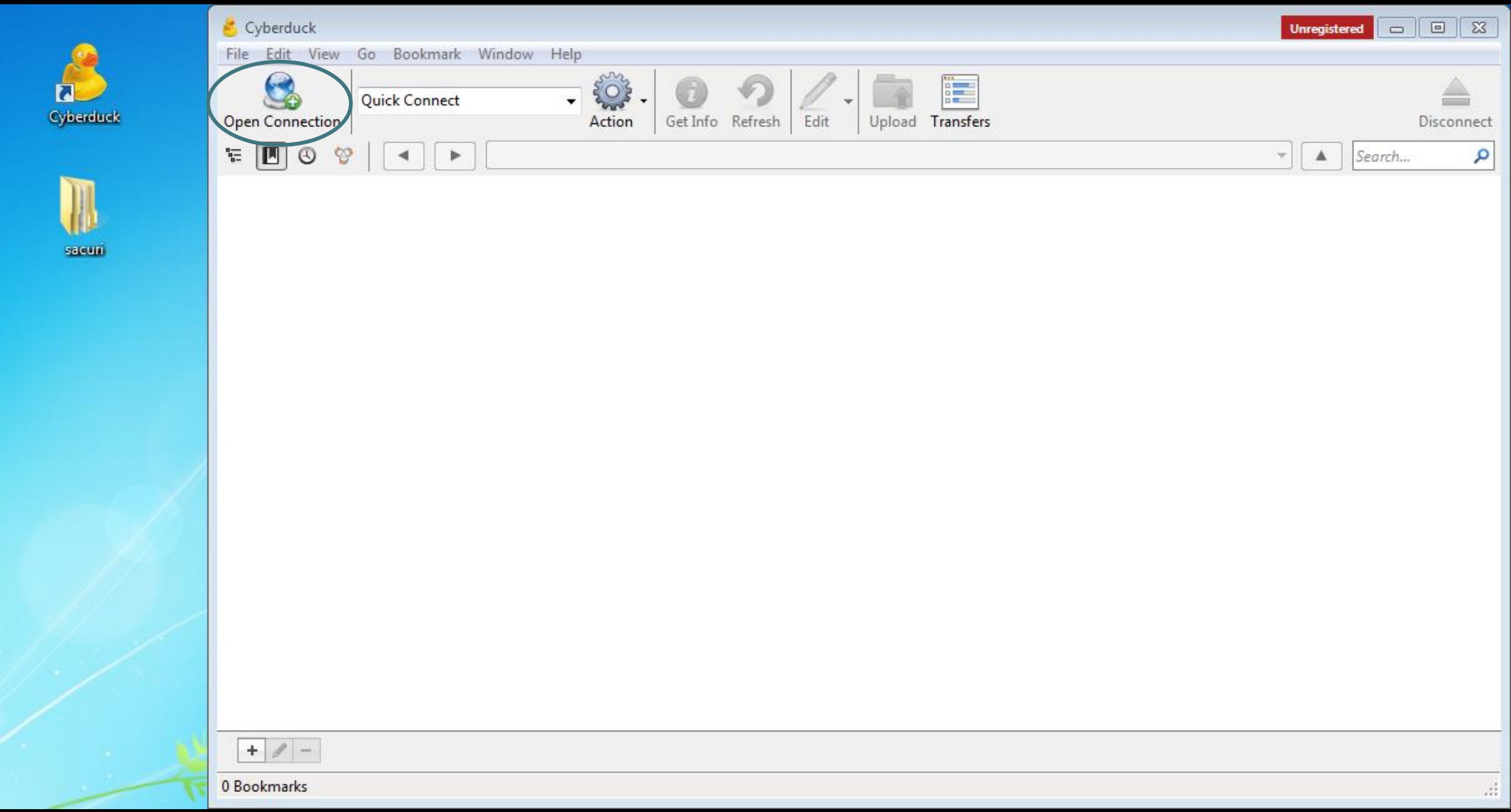
ary



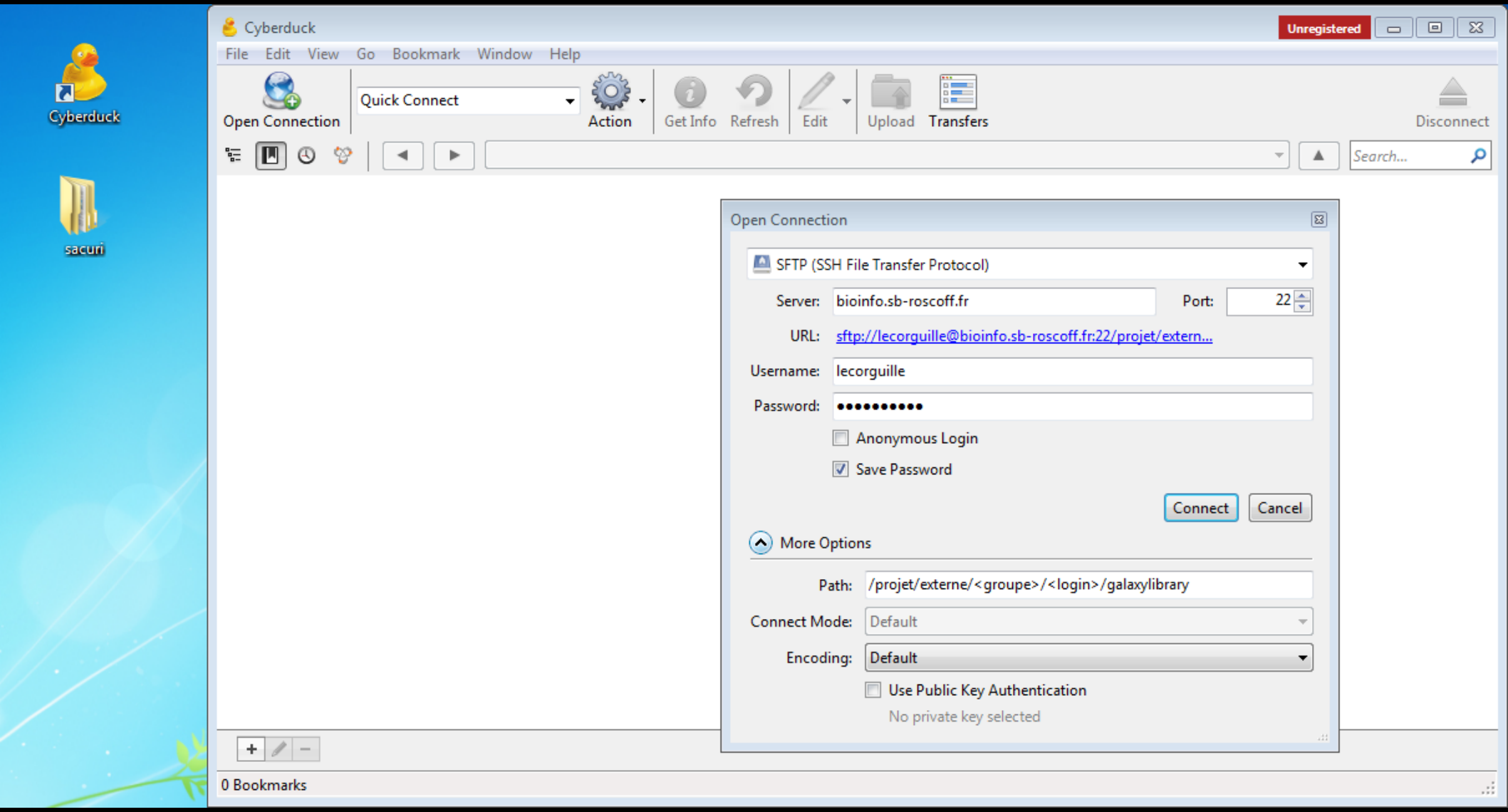
Import d'un fichier > 2 Go



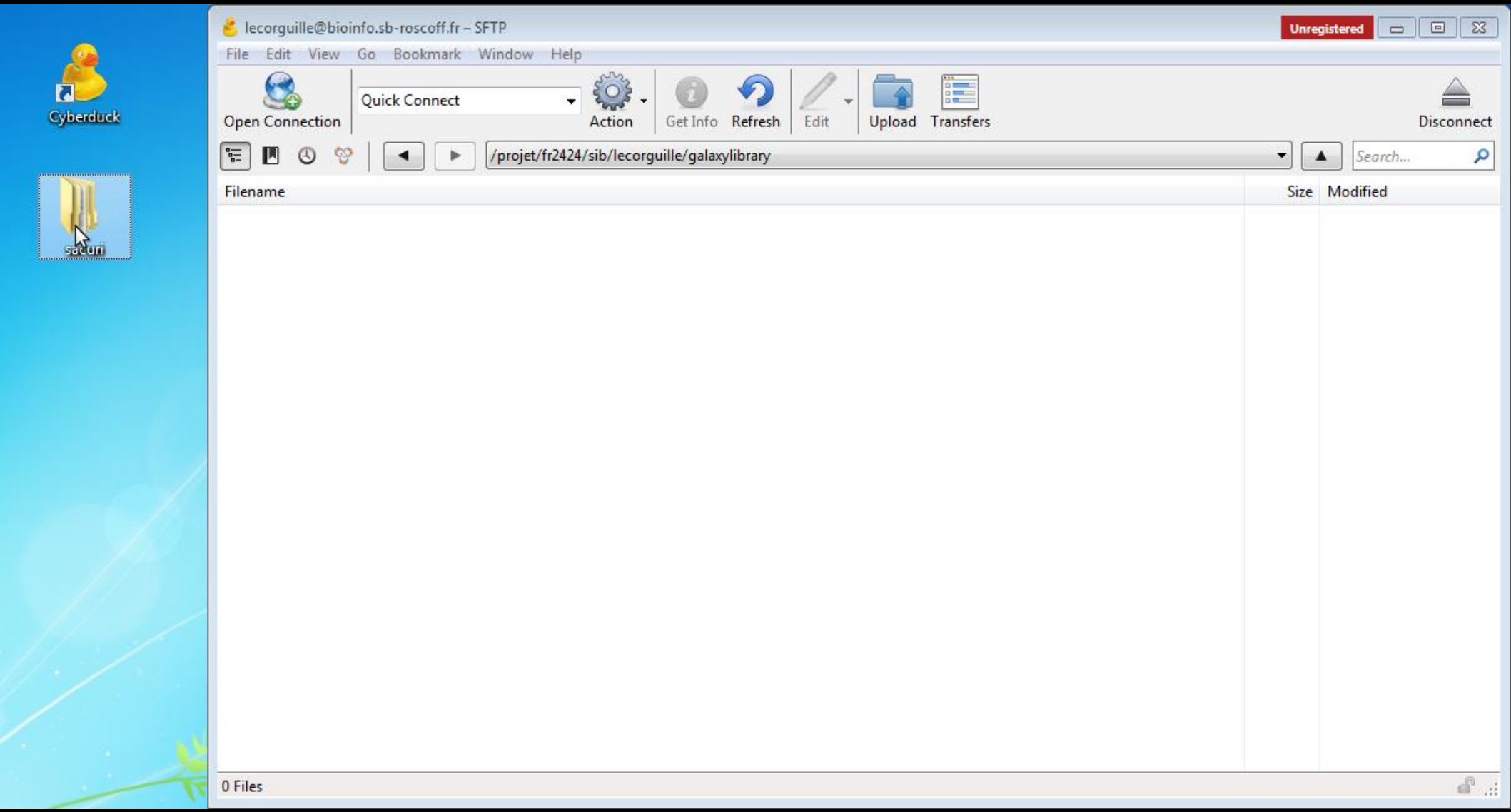
Import d'un fichier > 2 Go



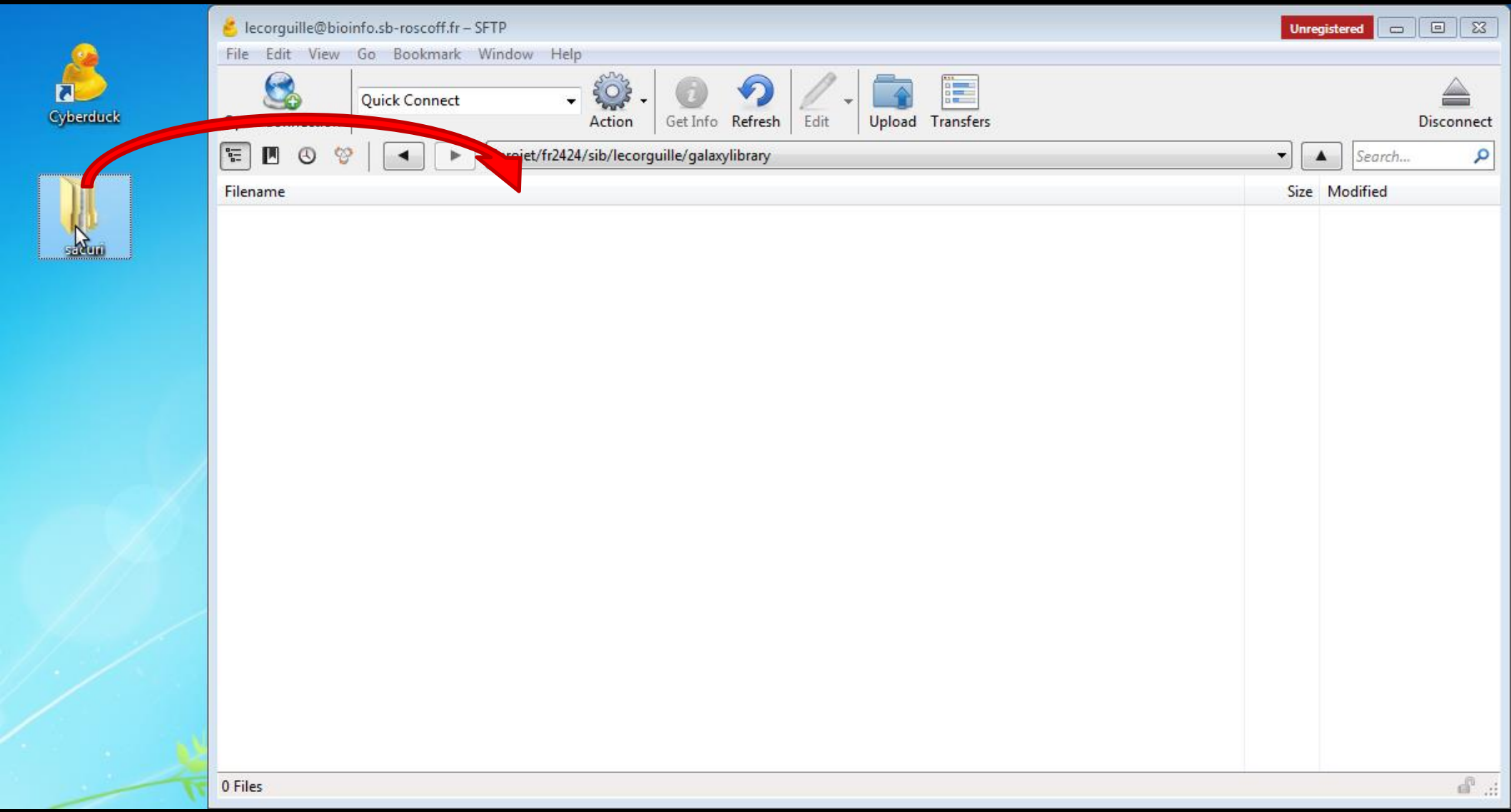
Import d'un fichier > 2 Go



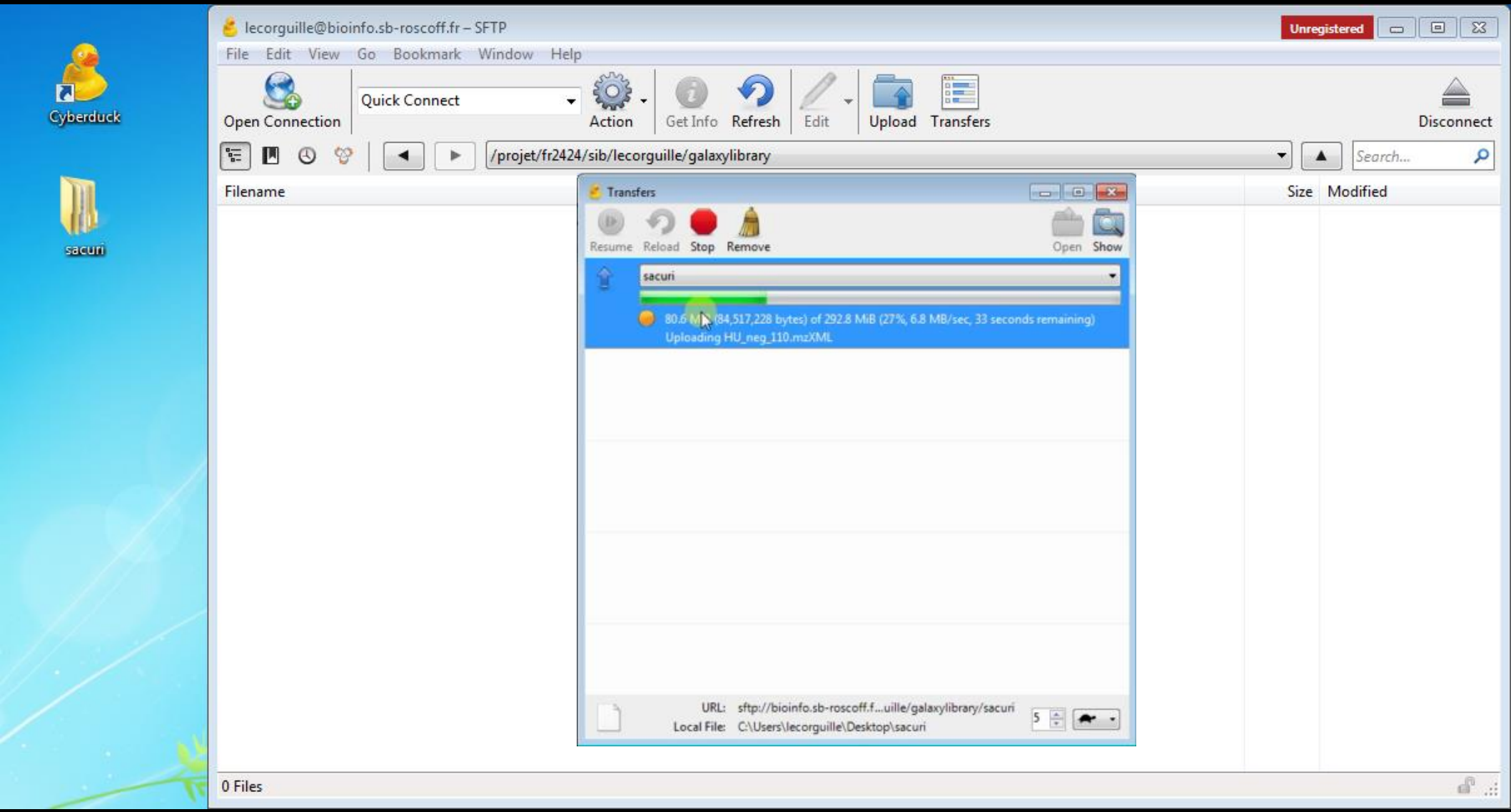
Import d'un fichier > 2 Go



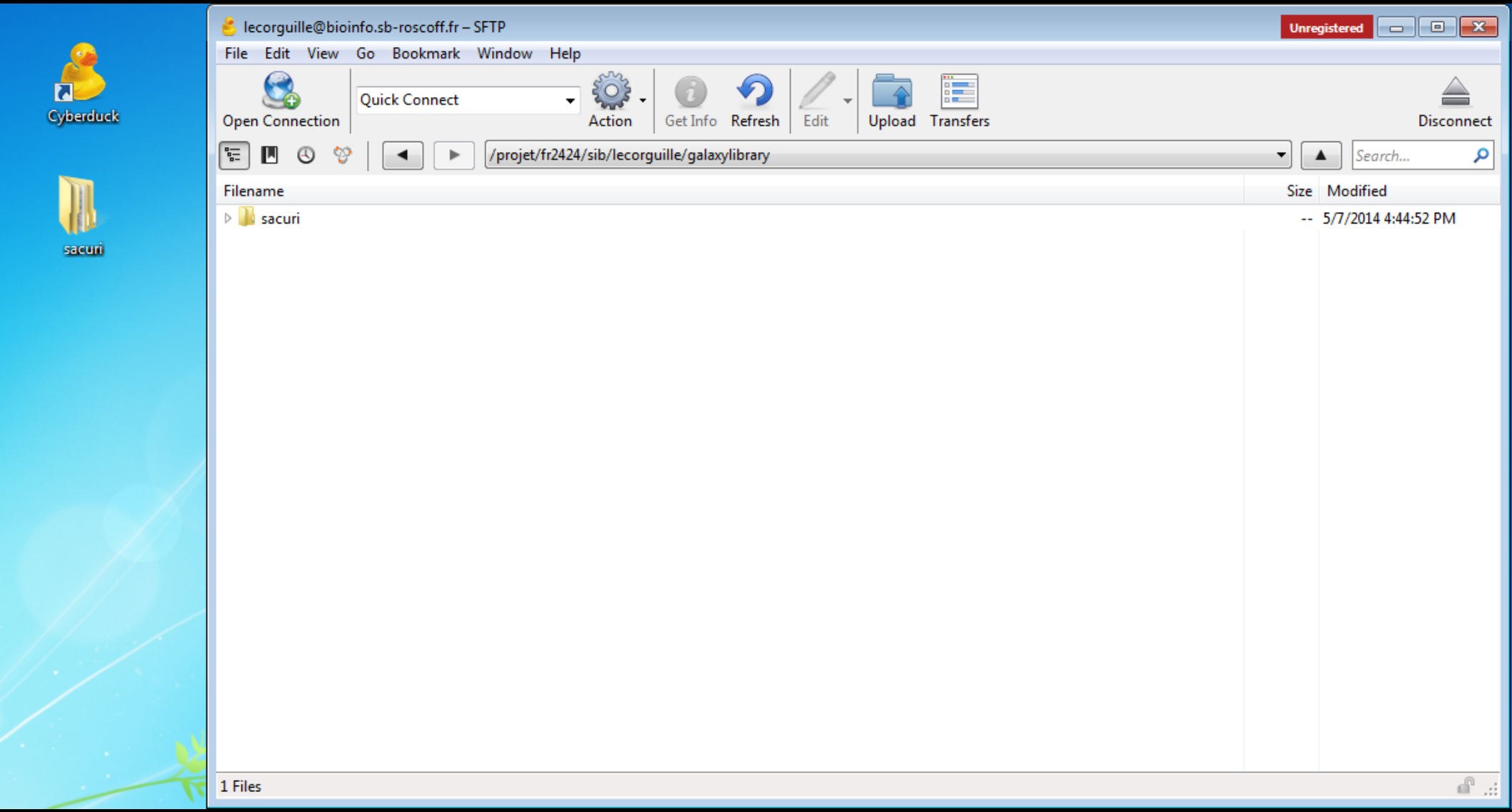
Import d'un fichier > 2 Go



Import d'un fichier > 2 Go

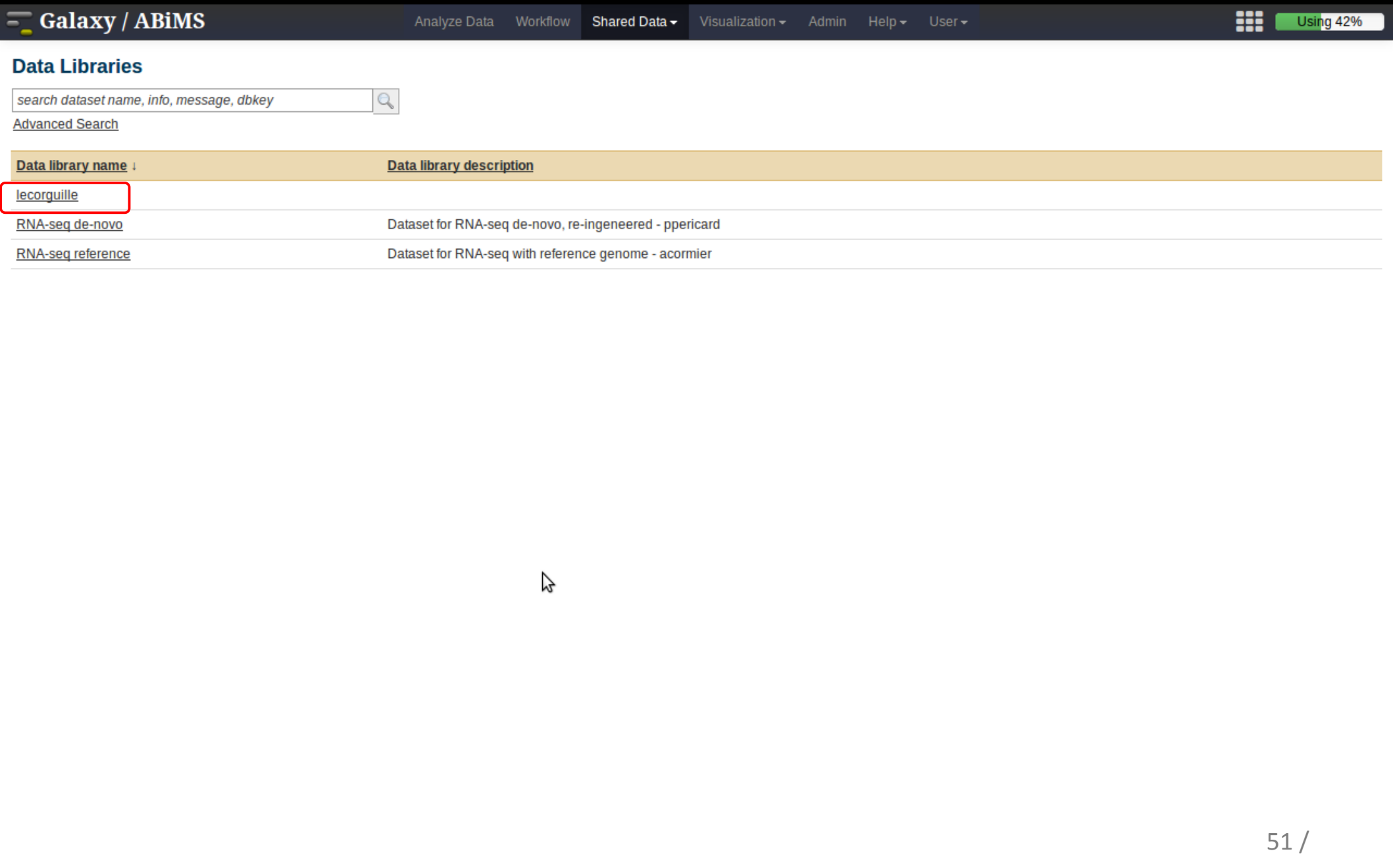


Import d'un fichier > 2 Go

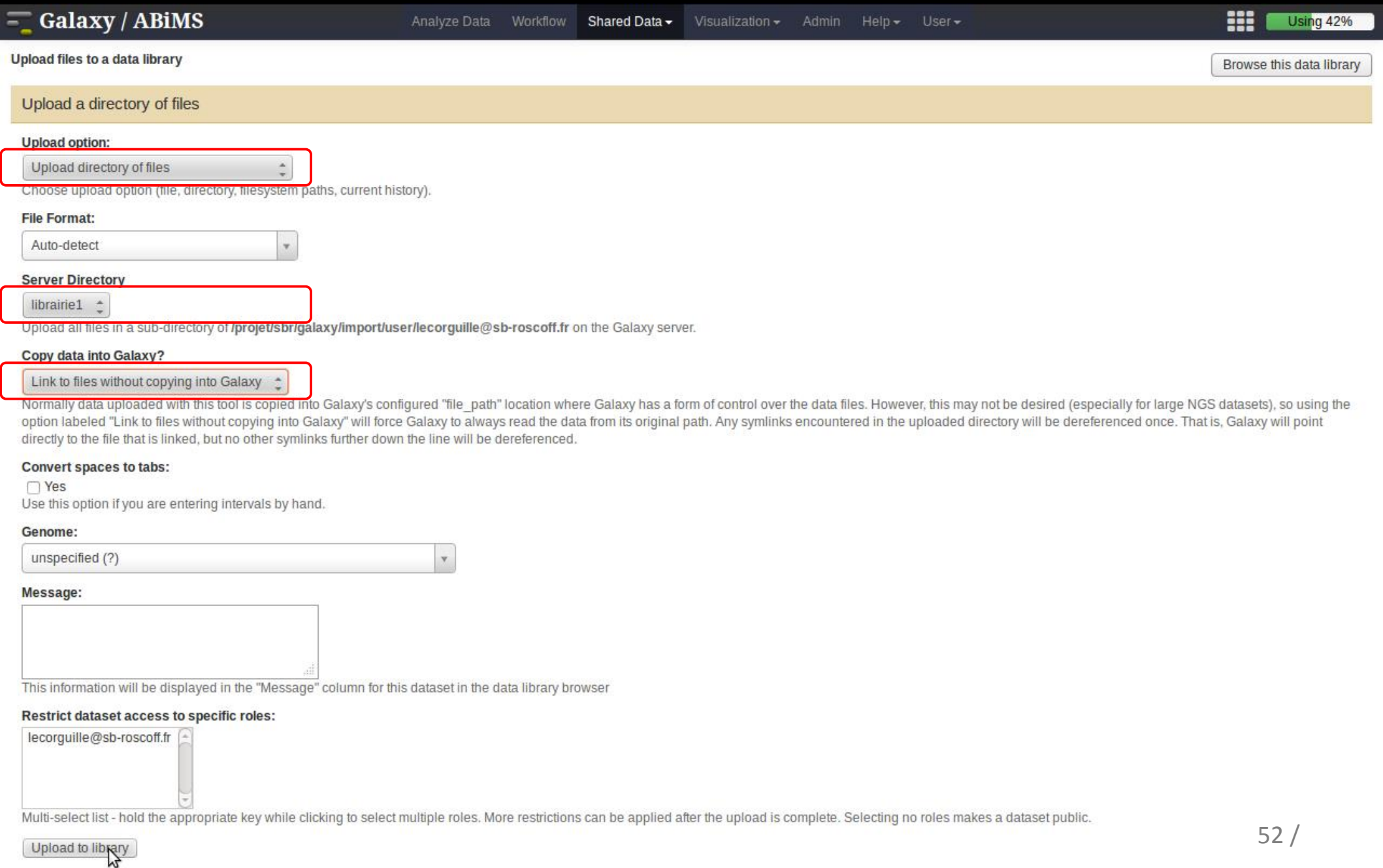


IMPORT DES DONNÉES
> 2 GO - PART II – CÔTÉ GALAXY

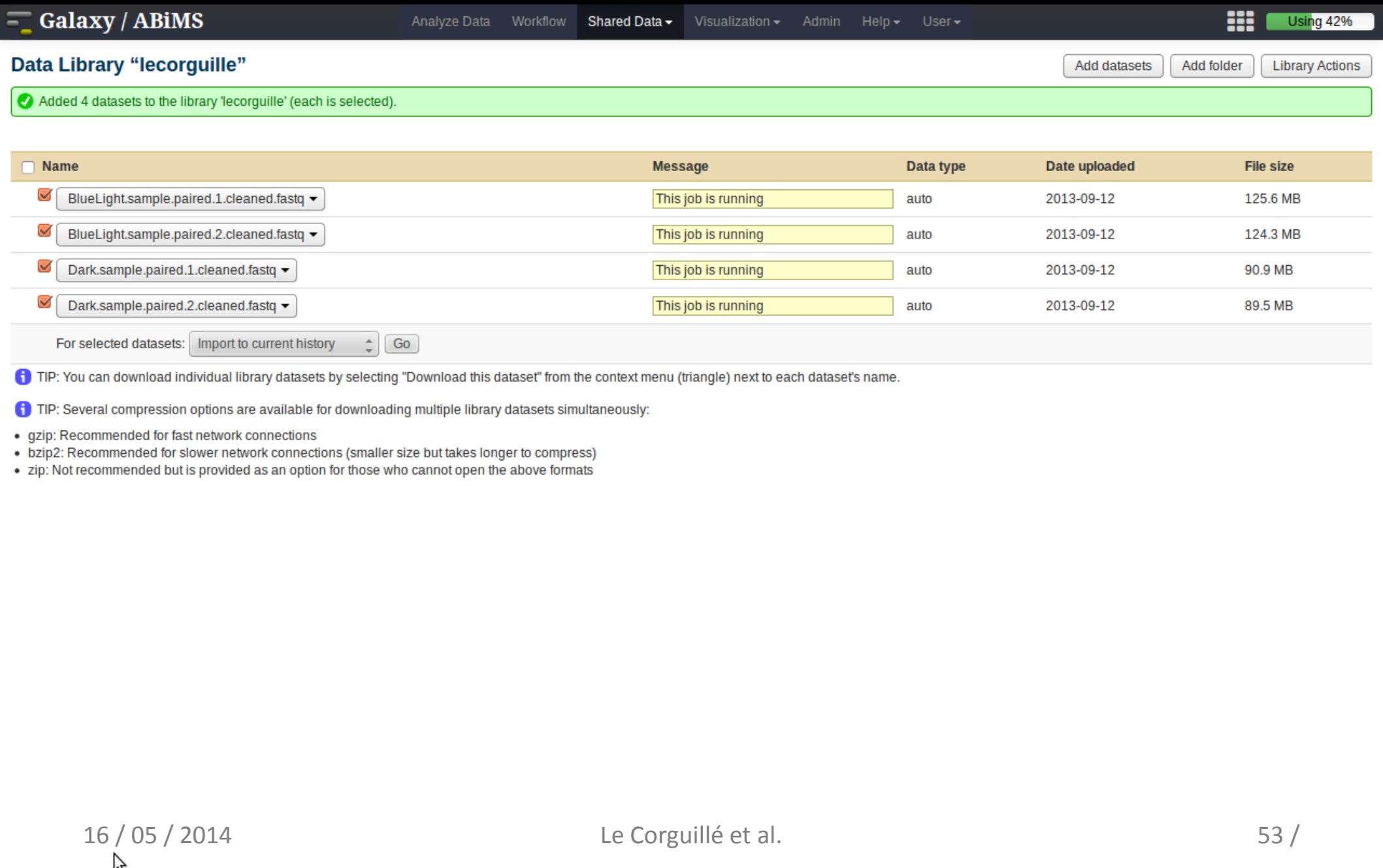
Import d'un fichier de plus de 2 Go



Import d'un fichier de plus de 2 Go



Import d'un fichier de plus de 2 Go

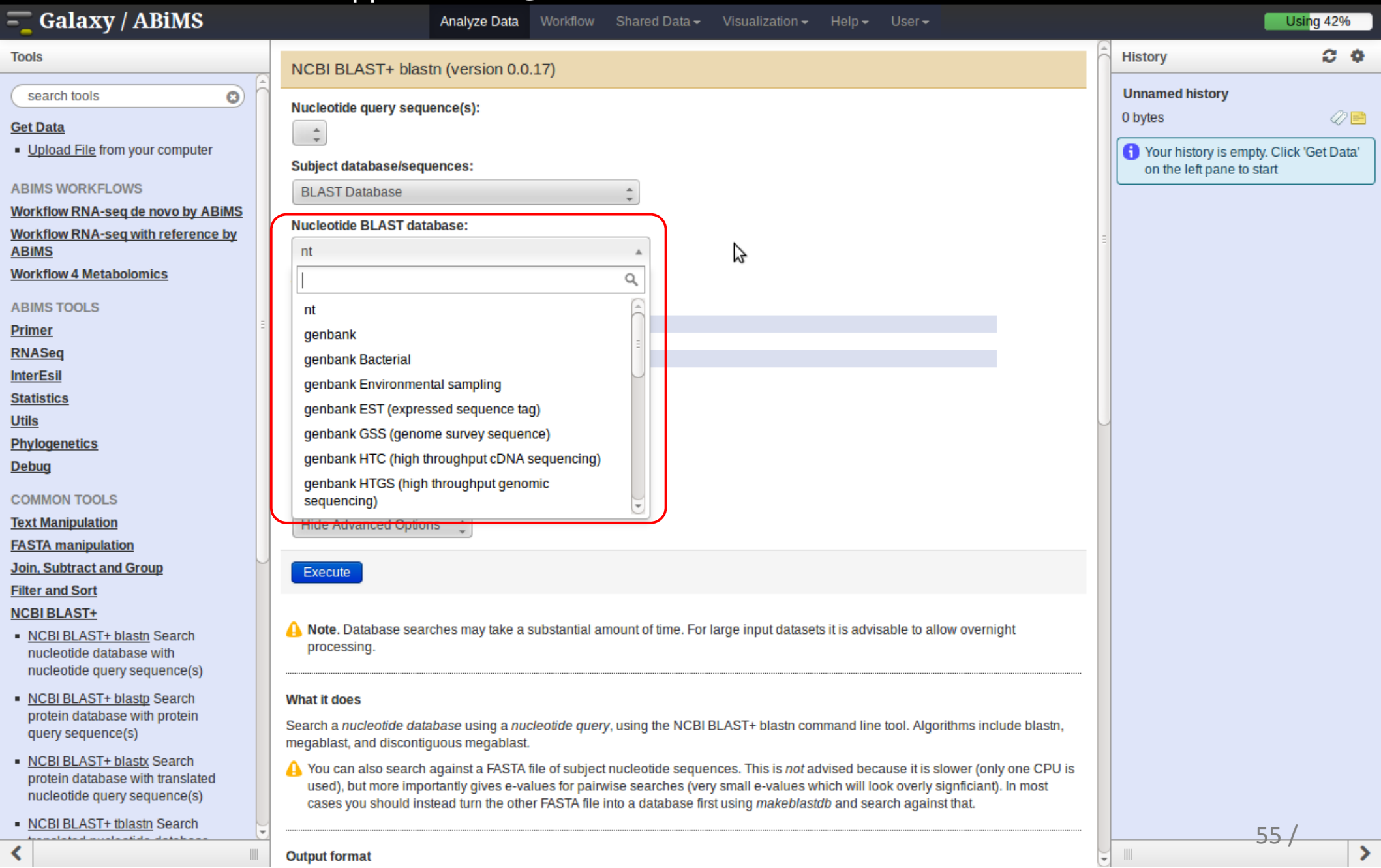


~~IMPORT DES DONNÉES~~

Pour les ressources conséquentes et publiques

ex : swissprot au format blast
chromosome 11 du génome humain au format bowtie2

Faire une demande à support.abims@sb-roscoff.fr



Résumé

IMPORT DES DONNÉES

Import

- Résumé
  - Si ressource locale :
    - Si < 2 Go :
      - Import depuis Get Data
      - Copier/Coller dans la zone de texte
    - Si > 2 Go :
      - Import via le protocol FTP
  - Si ressource distante :
    - Si < 2 Go :
      - Copier/Coller de l'adresse ftp://
  - Si ressource type banque publique (brute ou reformatée)
    - Demande à support.abims@sb-roscoff.fr pour mise à disposition

Import

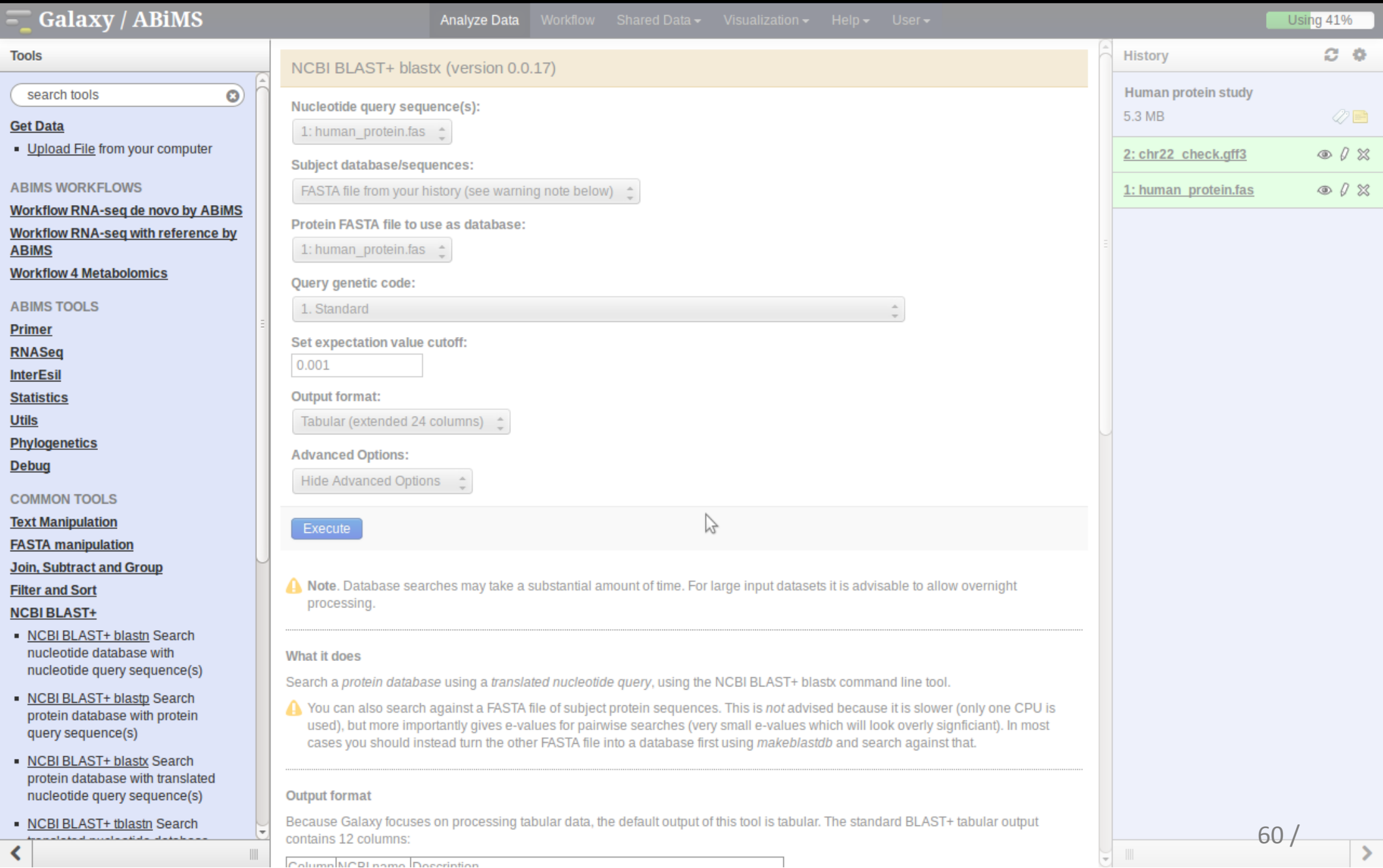
- Exercice
  - Récupérer sous Galaxy un fichier de séquence

http://application.sb-roscoff.fr/download/fr2424/abims/lecorguille/cours/JohnDoe.fasta

OUTILS

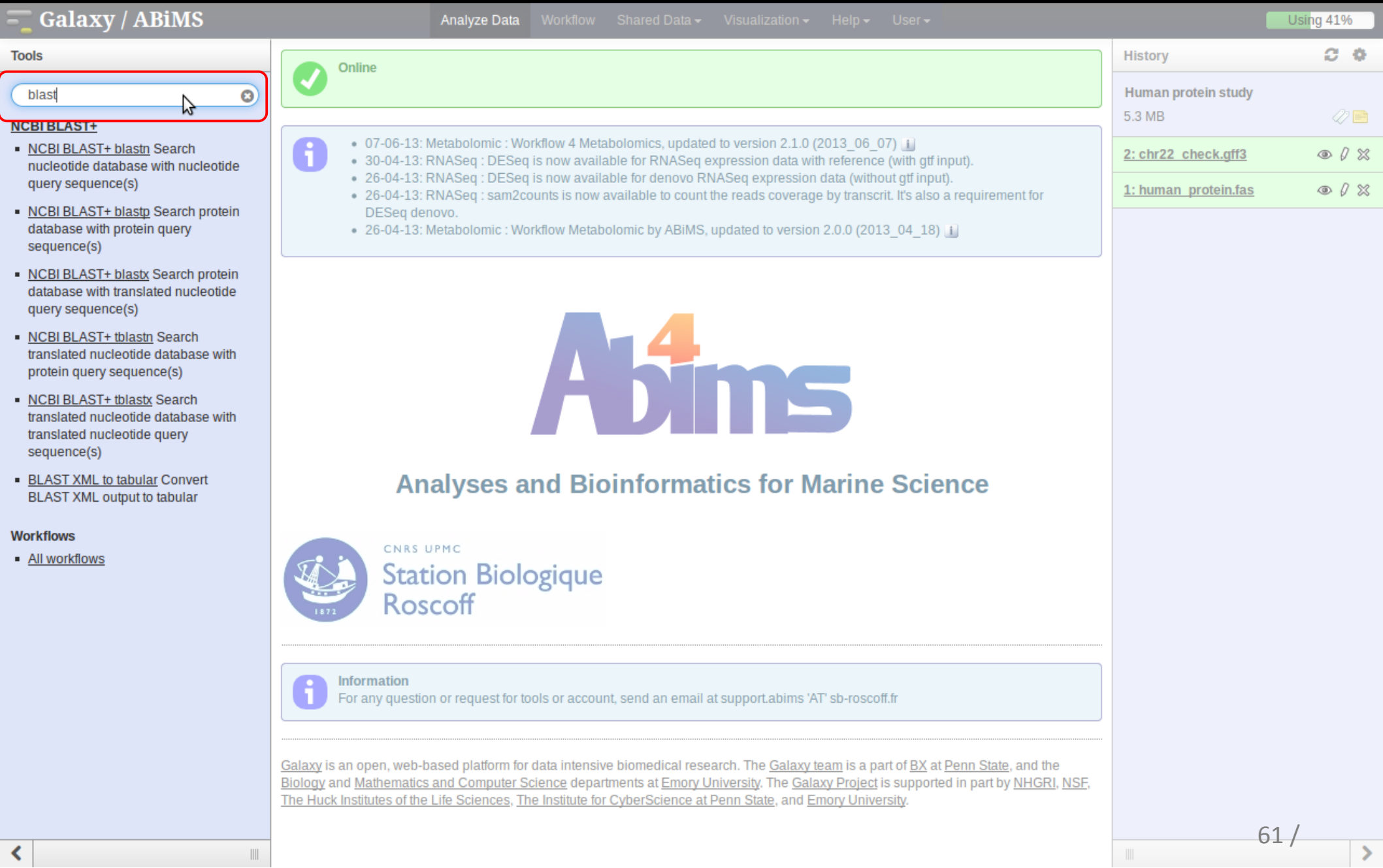
L'interface de Galaxy

la liste des outils



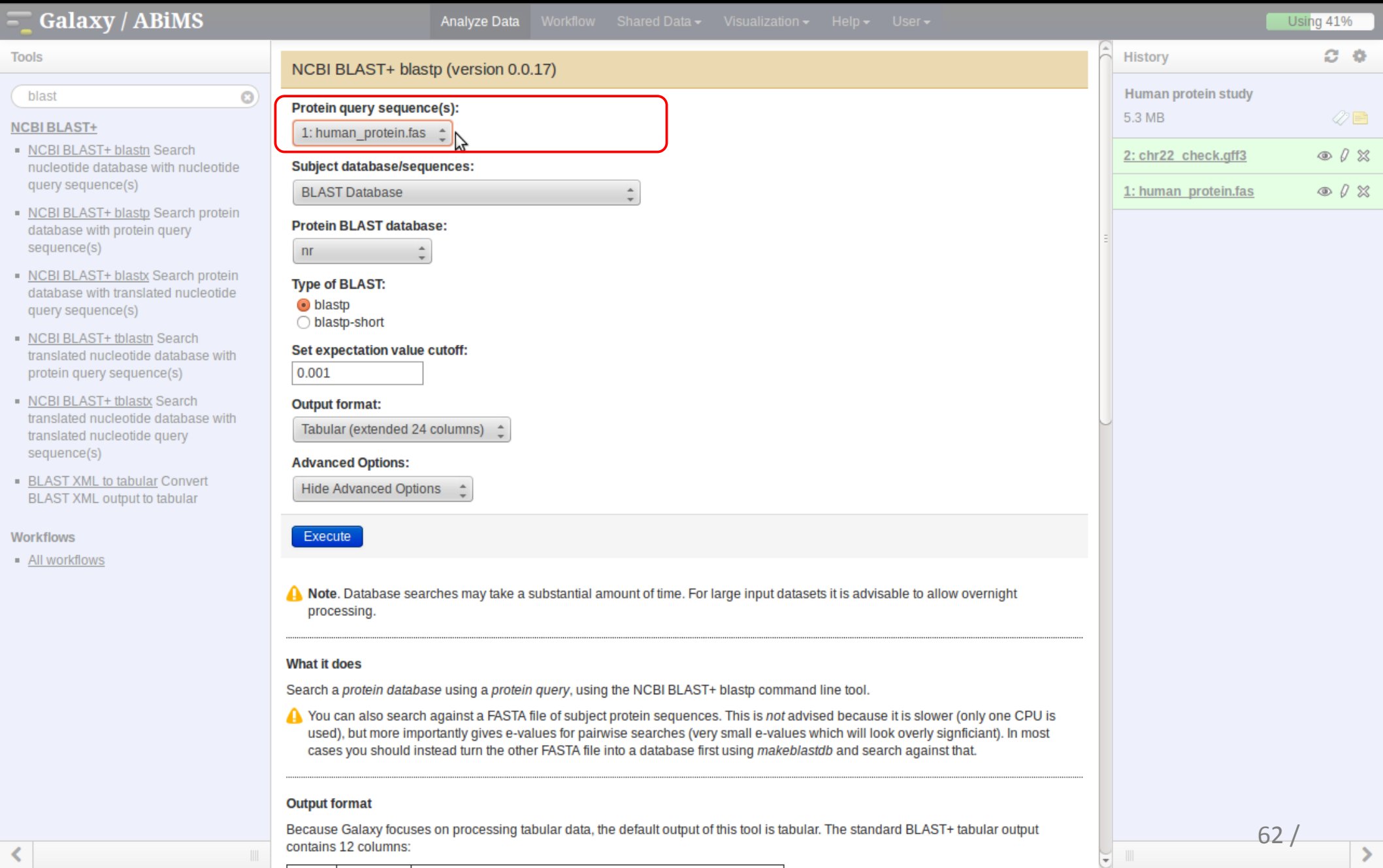
L'interface de Galaxy

la liste des outils



Exemple de formulaire :

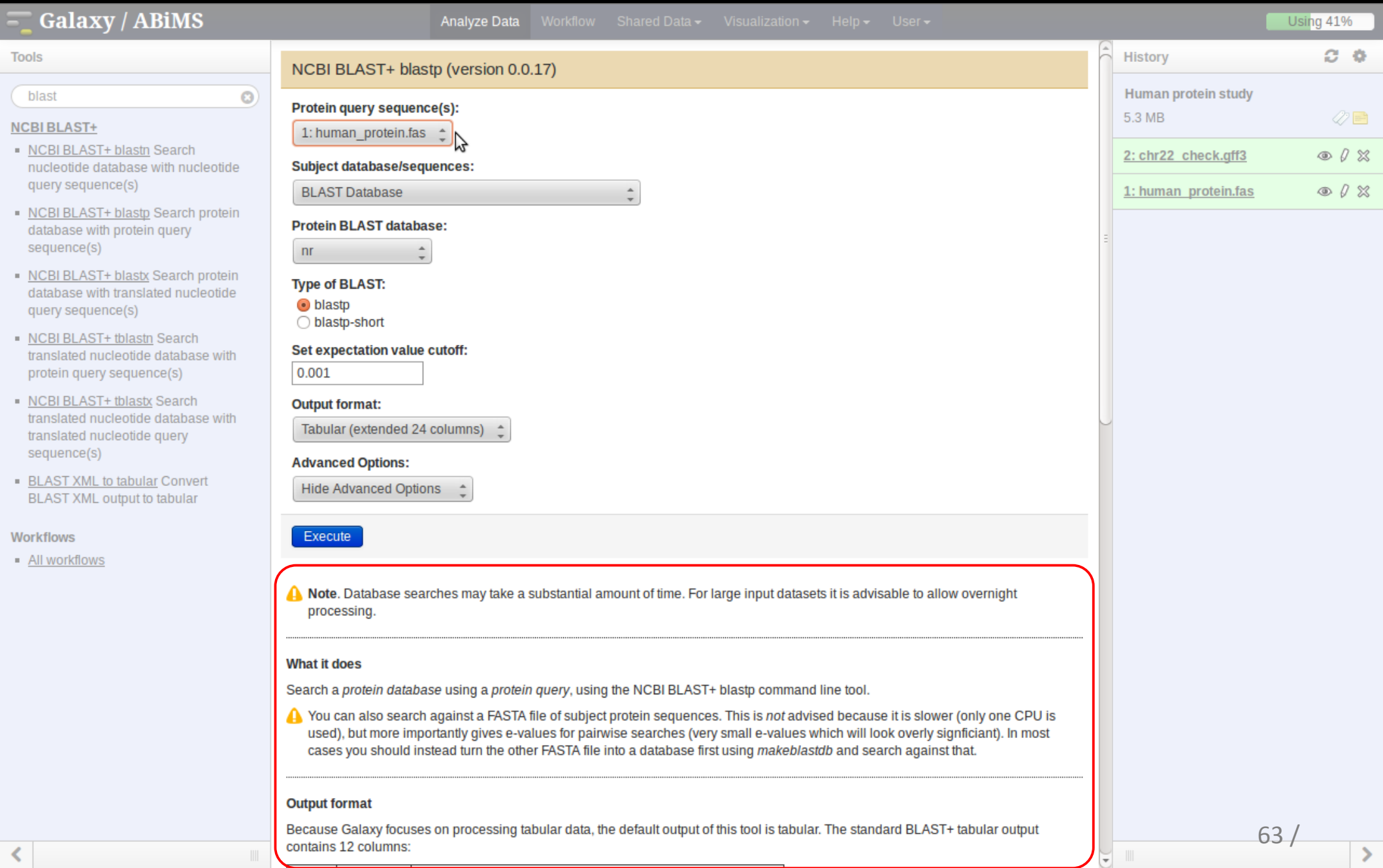
- Sélection du fichier d'entrée – filtre sur le type de fichier (ici : fichier fasta)



Exemple de formulaire :

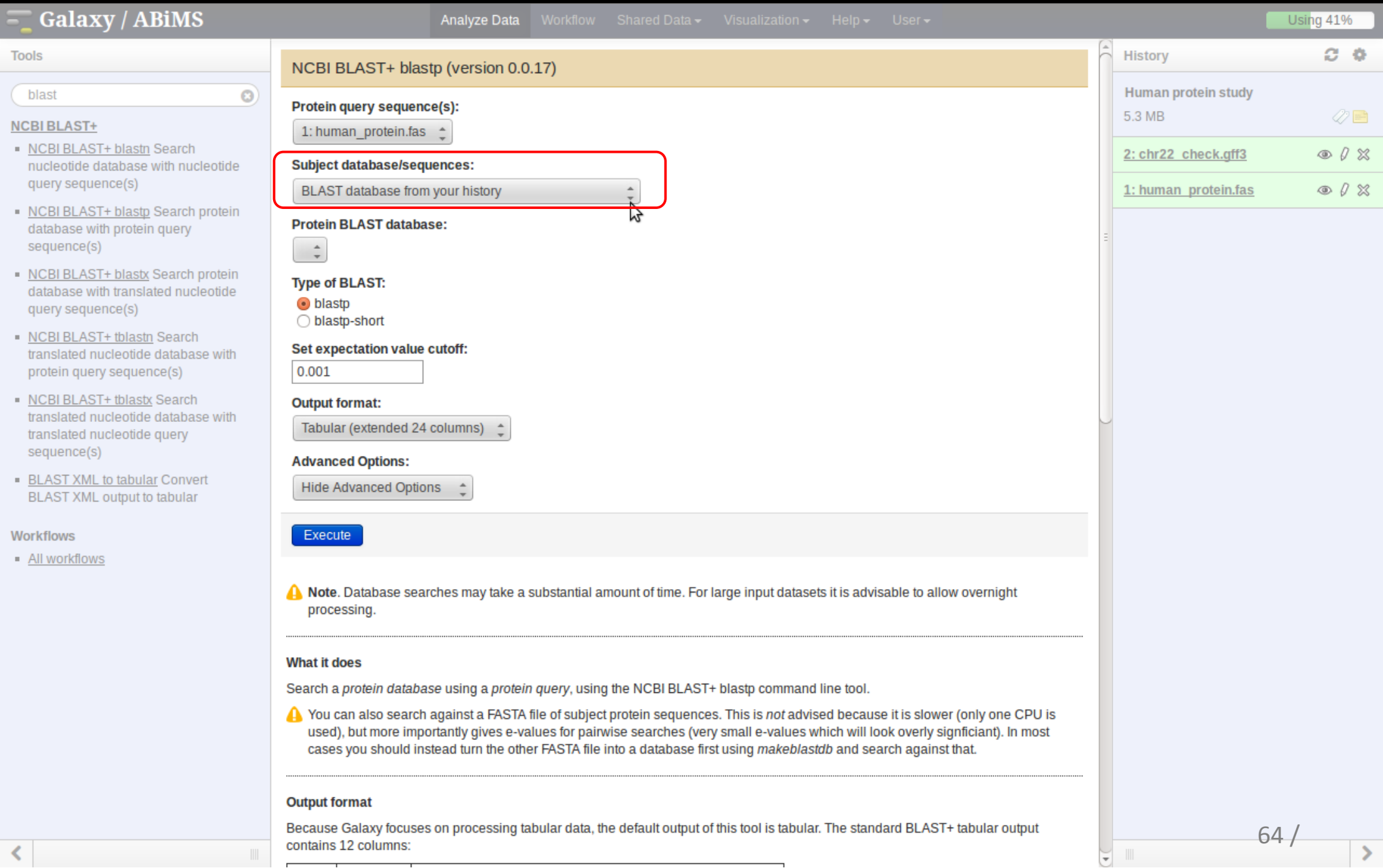
- Aide sur l'outil

Dans la limite des stocks disponibles



Exemple de formulaire :

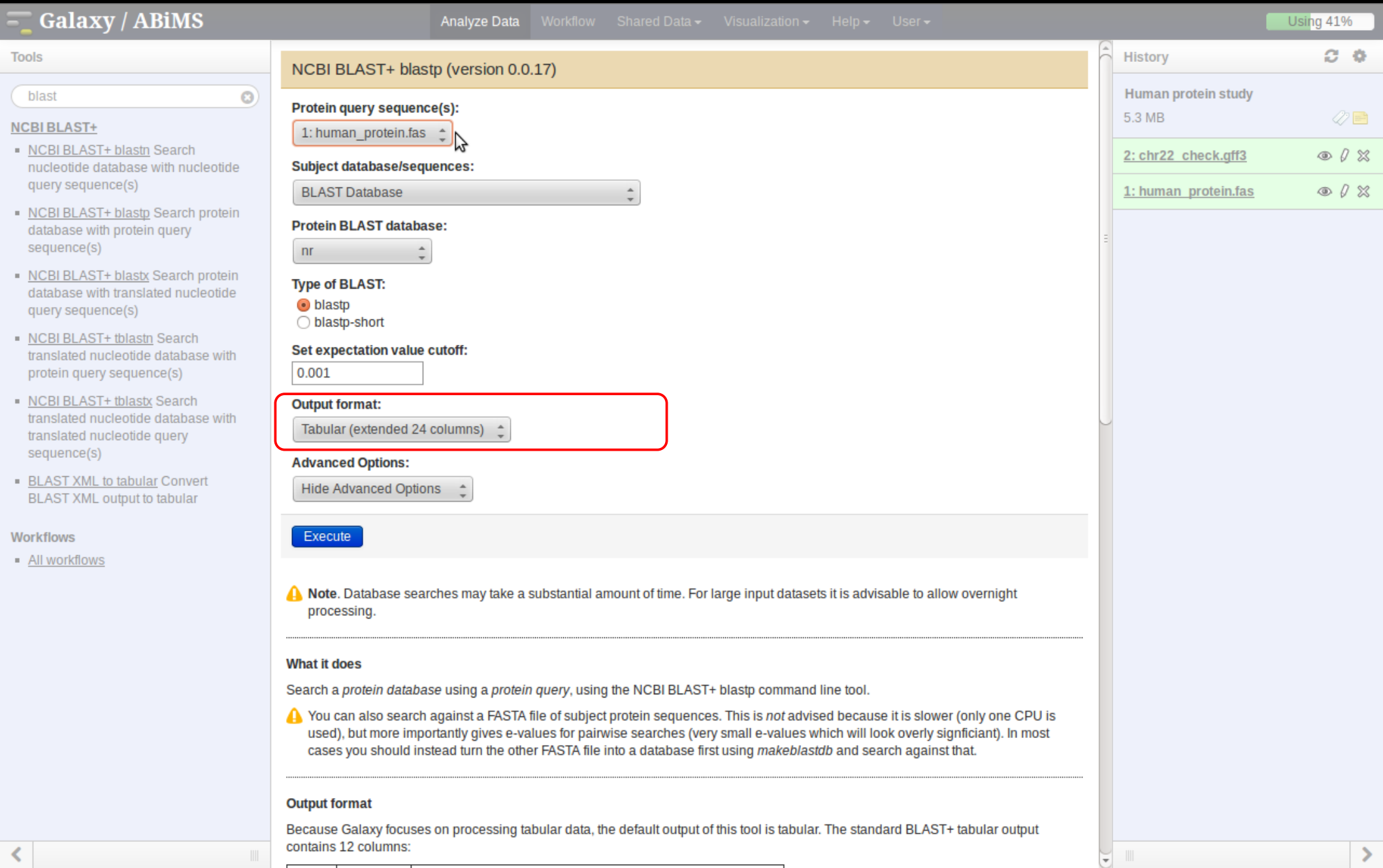
- Ici : possibilité de choisir une banque personnelle dans l'historique … mais j'en ai pas !



Exemple de formulaire :

- Possibilité de choisir le format du fichier de sortie

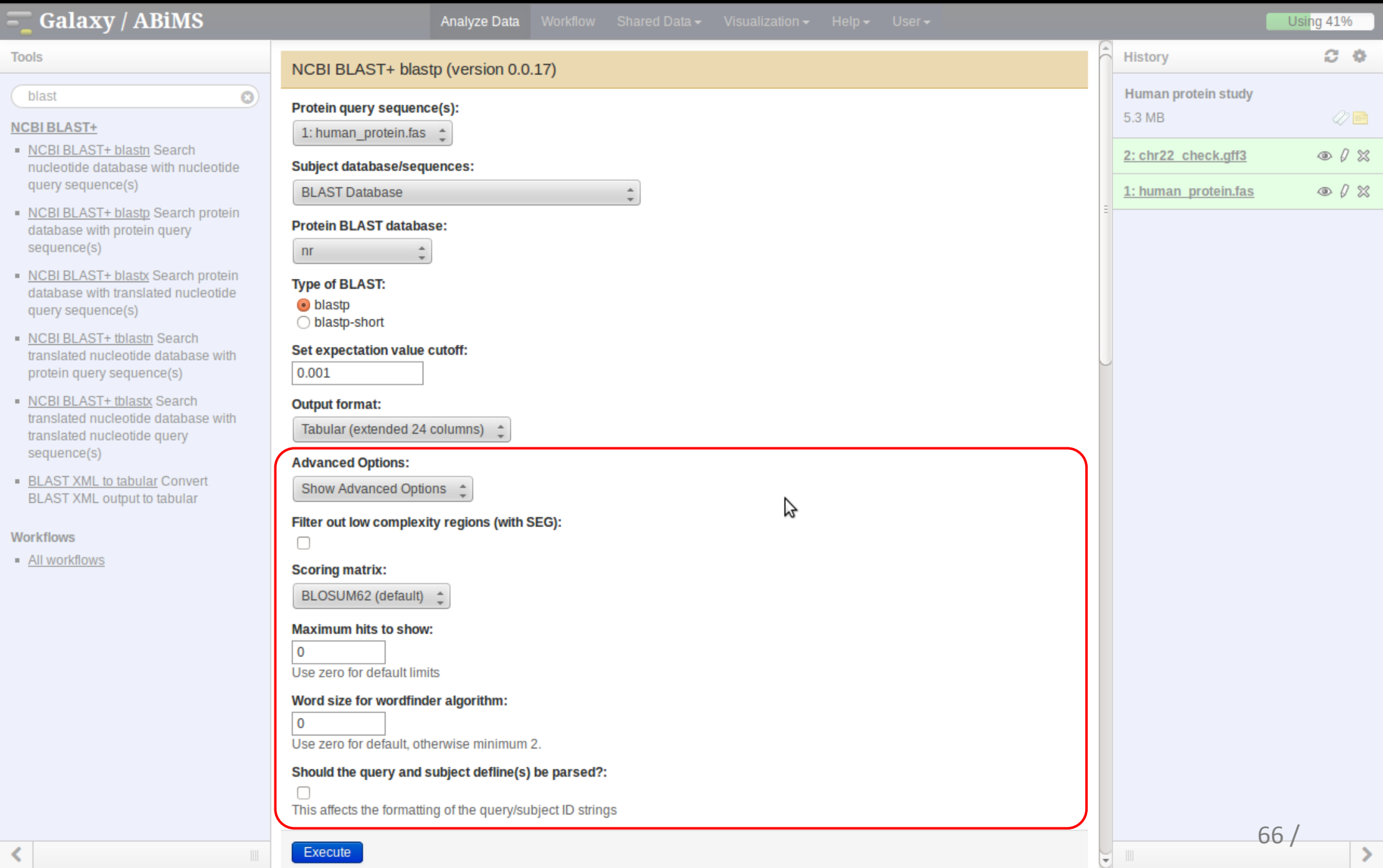
Dans la limite des stocks disponibles



Exemple de formulaire :

- Possibilité d'accéder à des options avancées = accessibilité / ergonomie

Dans la limite des stocks disponibles



Outils

- Les status

Job en attente de soumission

= le job est dans la « queue » de l'ordonnanceur

La durée de ce statut dépend du nombre de jobs actuellement en attente et du nombre de cpu demandé

Outils

- Les status

Job en cours d'exécution

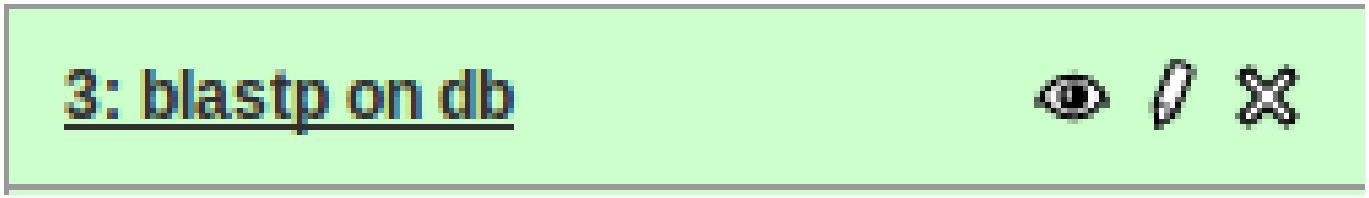
= le job tourne actuellement sur le cluster de calcul

La durée de ce status dépend complètement de la nature du job et de la puissance de calcul allouée

Certains programmes vont pouvoir tourner sur plusieurs processeurs et disposer de 4, 8 ou 16 Go de RAM. Et d'autres sont mono-CPU.

Outils

- Les status

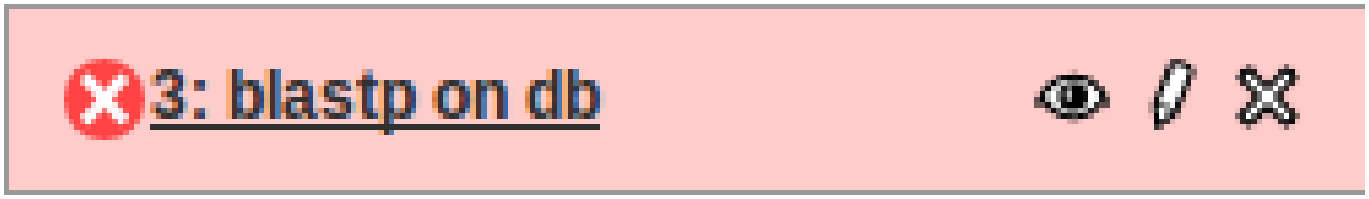


Job terminé

Son statut est OK mais des warnings ou des erreurs peuvent se cacher derrière. Ah hum !

Outils

- Les status



Job terminé mais en erreur

= le programme renvoie une erreur

Un programme si il est bien développé renvoie un code d'erreur :

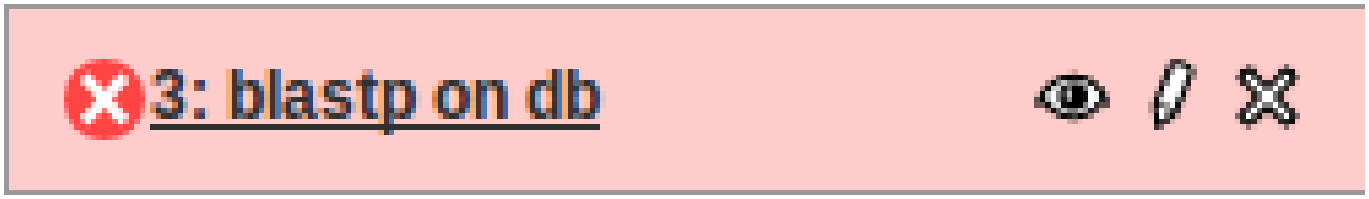
- = 0 si tout s'est bien terminé
- > 0 si il a rencontré une erreur

Le programme explique souvent d'où vient l'erreur

et parfois … pas.

Outils

- Les status

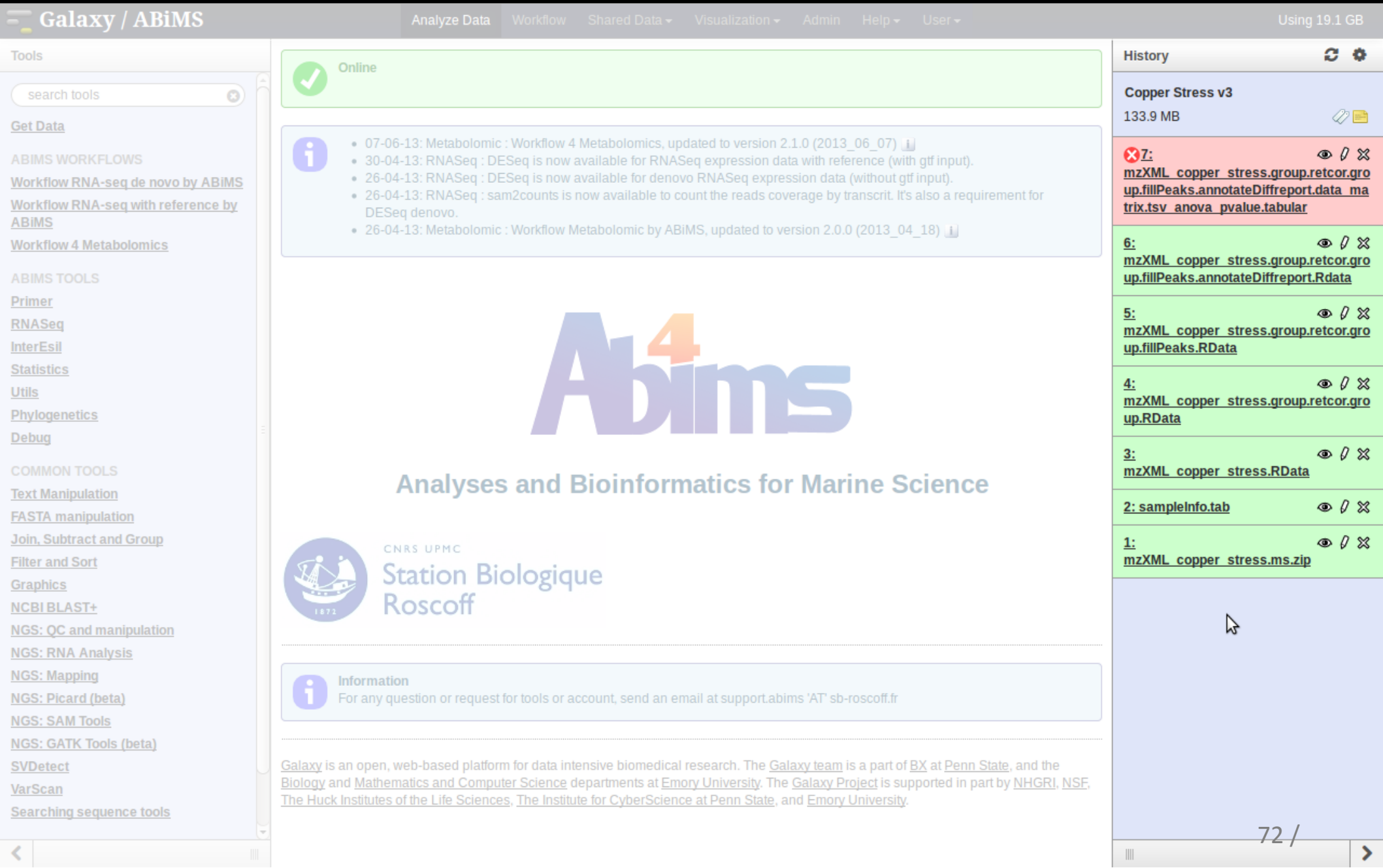


Job terminé mais en erreur

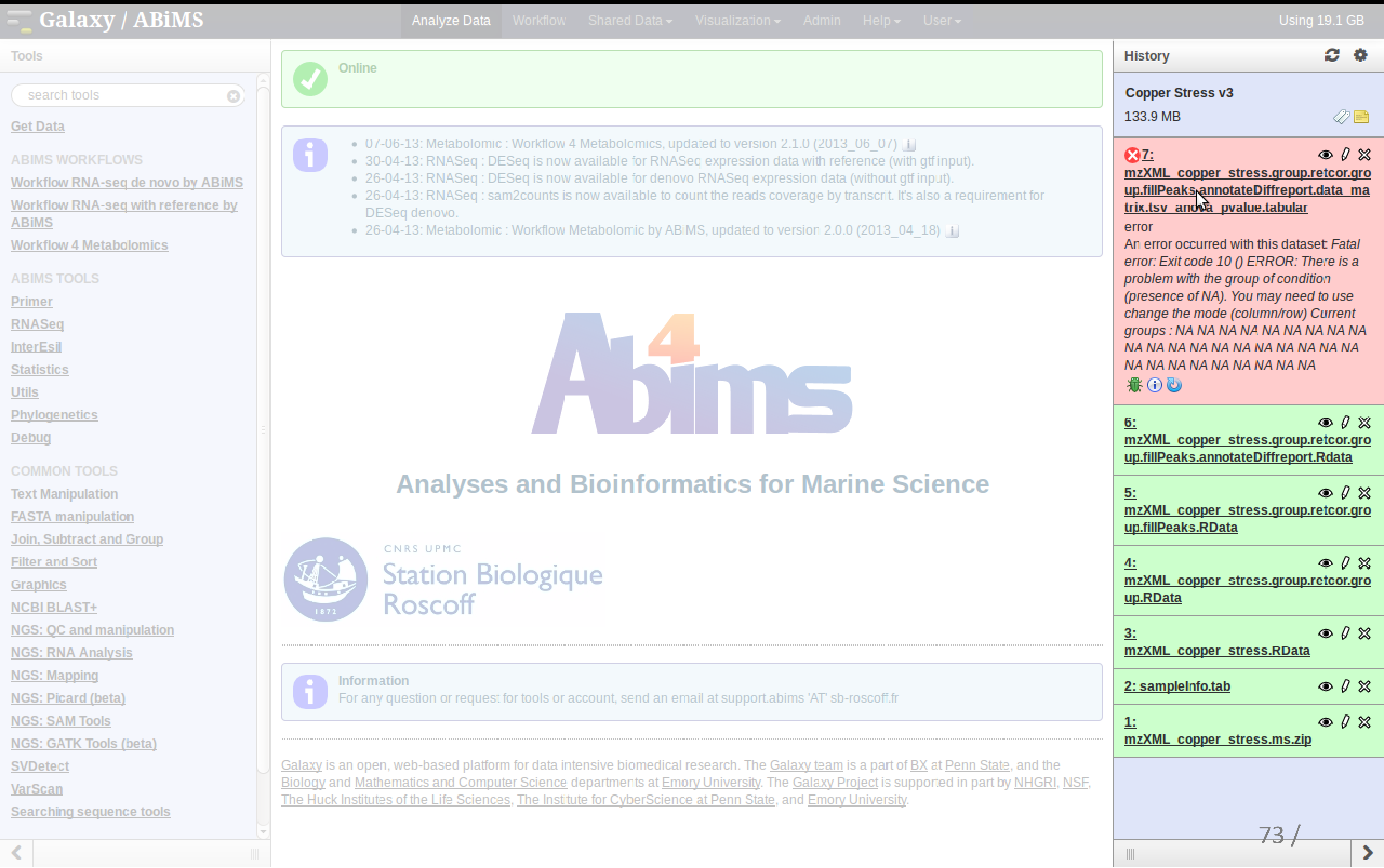
Les sources d'erreur :

- ~~L'utilisateur~~ :P
- Mauvaise utilisation : fichier d'entrée ou format ou option
- Mauvais portage du programme sous Galaxy … désolé :/
- Plantage non anticipé du programme

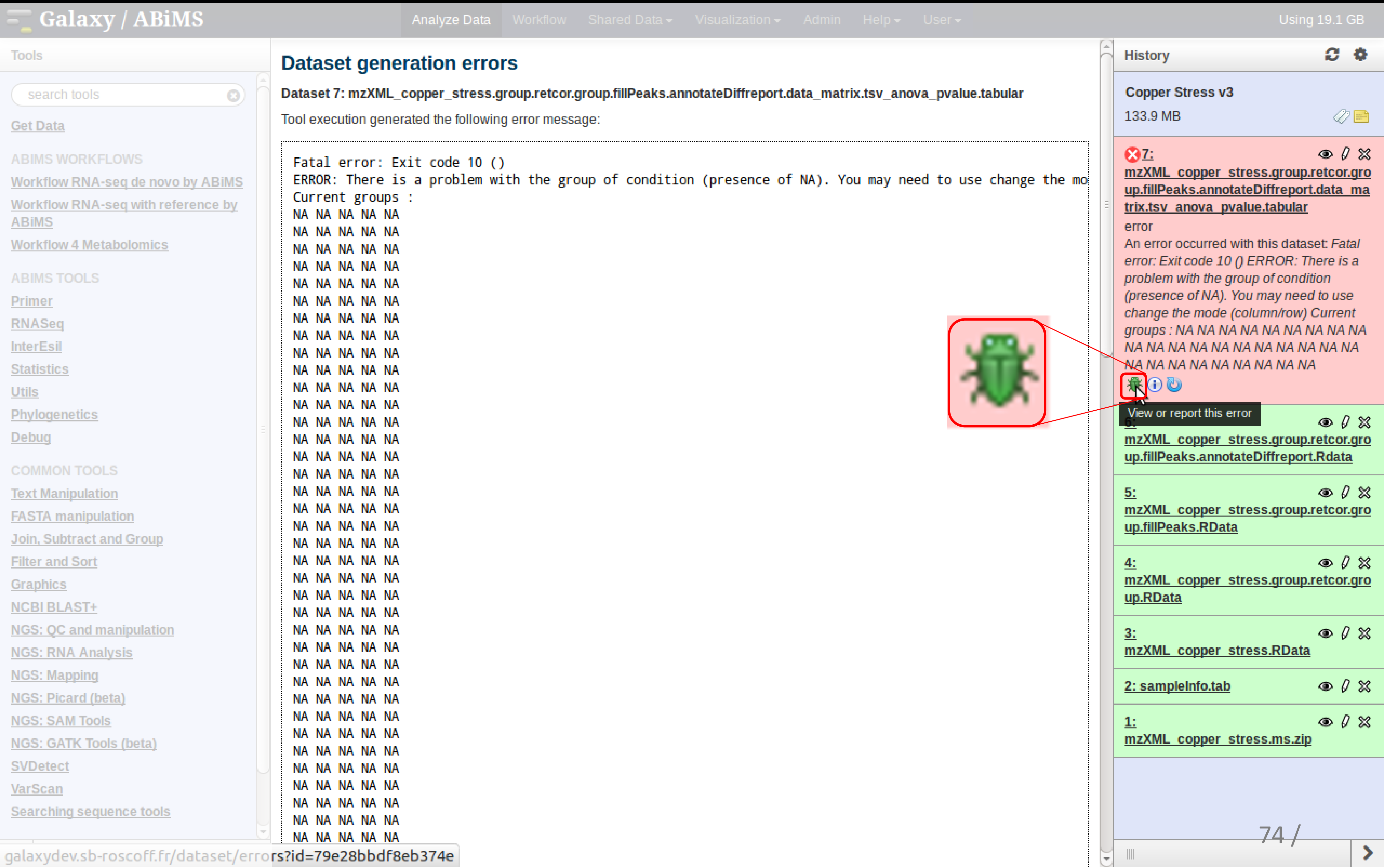
Exemple d'erreur



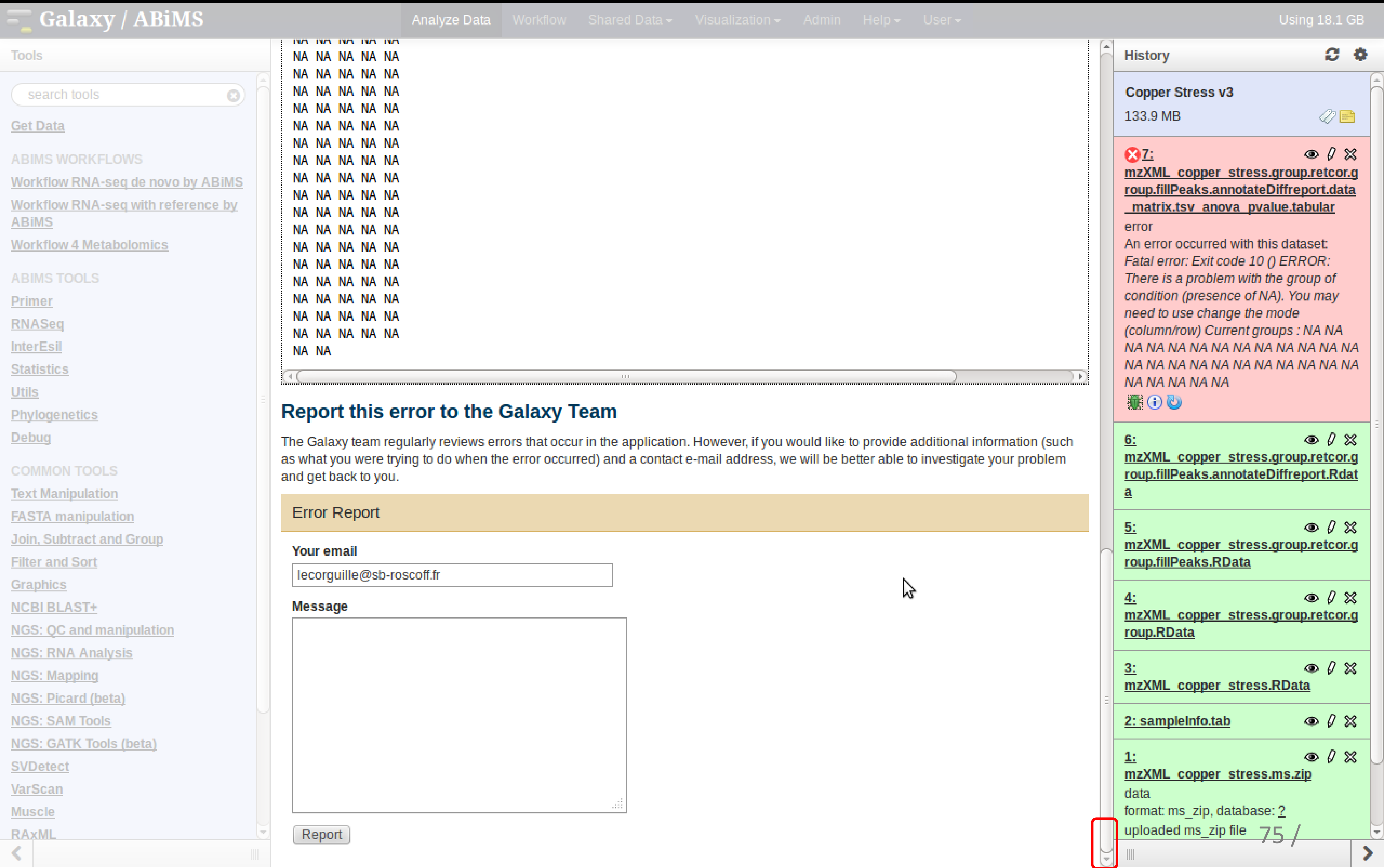
Exemple d'erreur



Exemple d'erreur



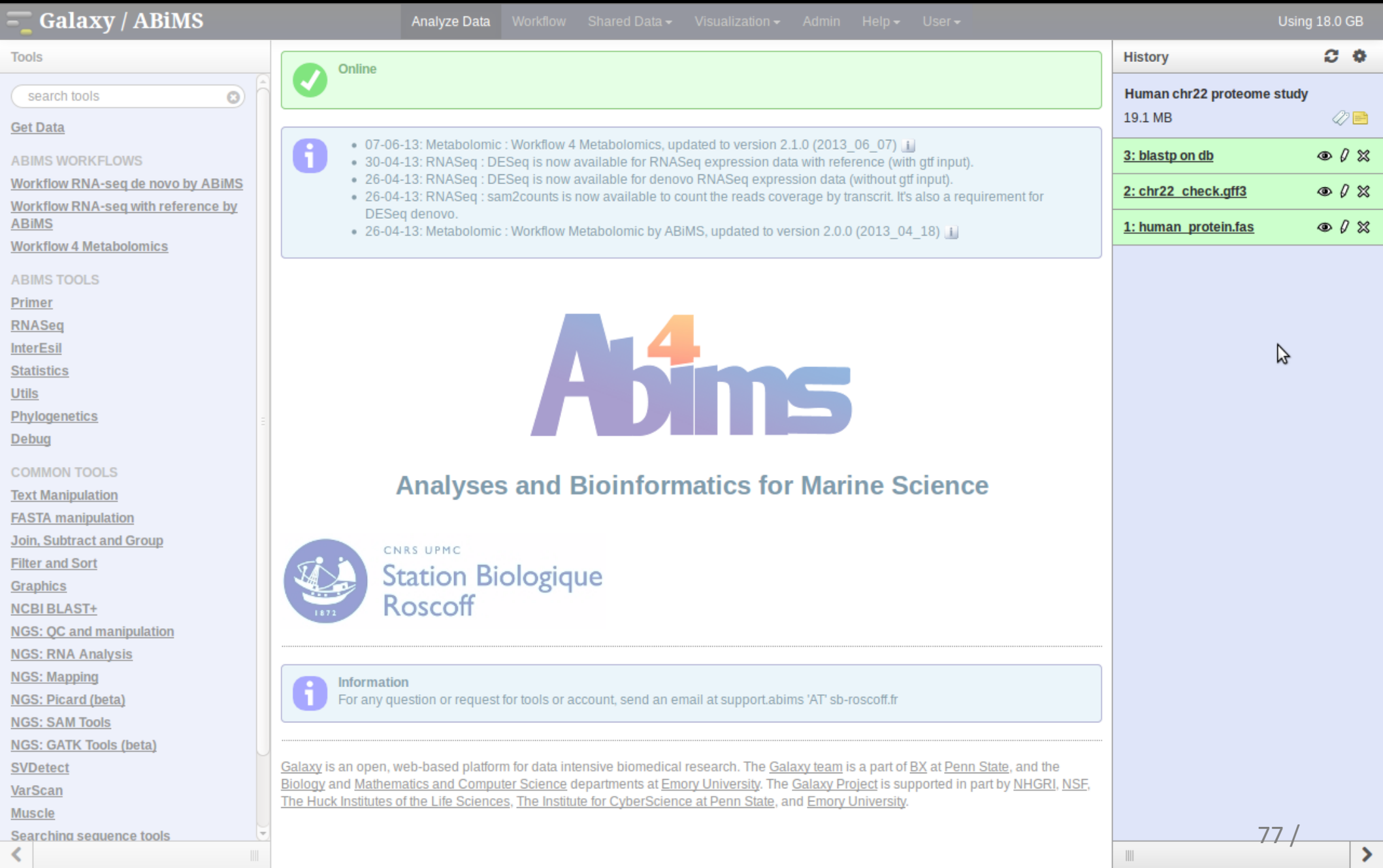
Exemple d'erreur



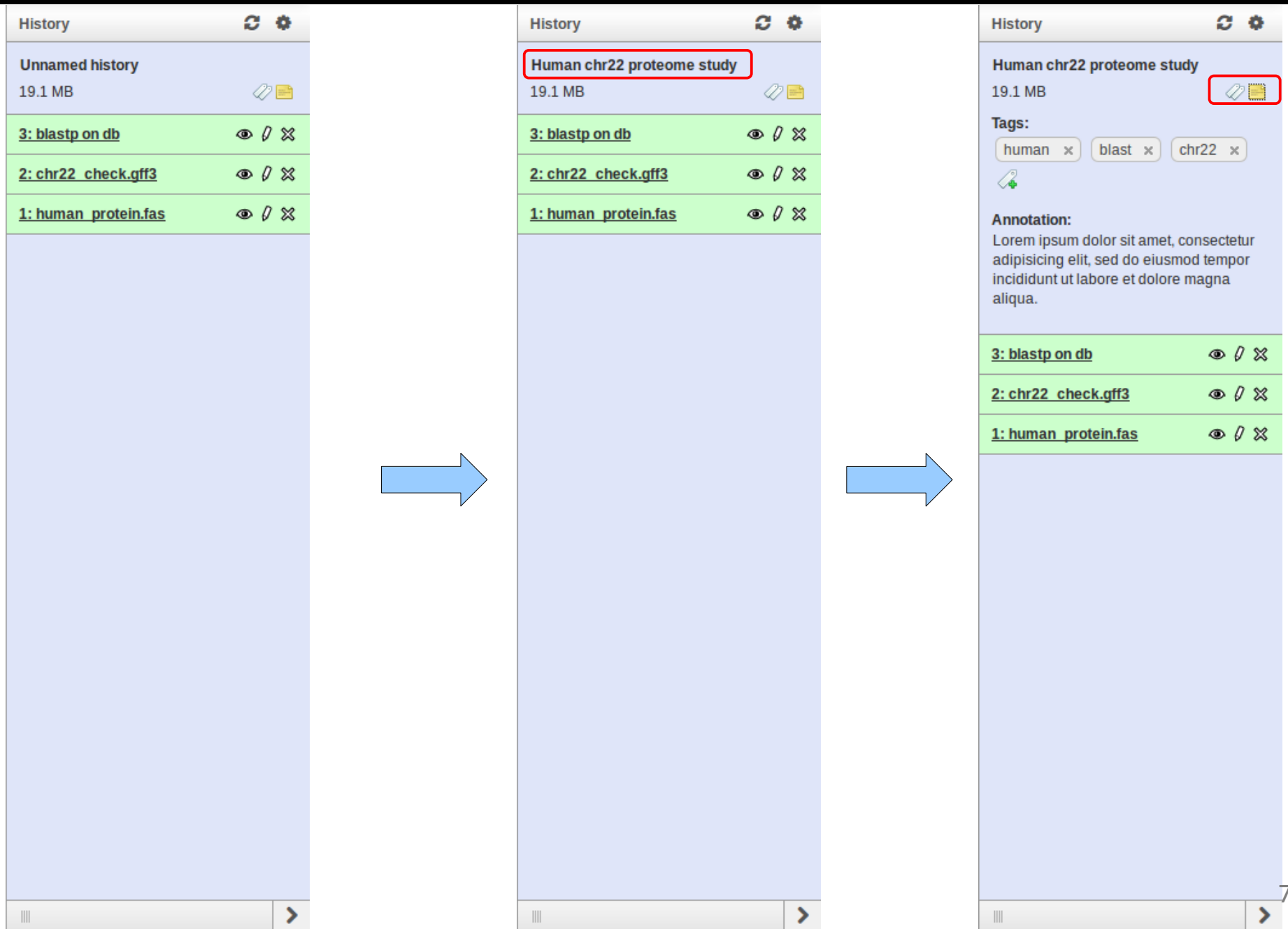
HISTORIQUE

L'interface de Galaxy : l'historique

Contient aussi bien les entrées que les sorties



Renommer, tagger et annoter



Information sur l'objet :

- Code erreur
- Sorties standard / error

Tool: NCBI BLAST+ blastp

Name:	blastp on db
Created:	Aug 09, 2013
Filesize:	13.9 MB
Dbkey:	?
Format:	tabular
Galaxy Tool Version:	0.0.17
Tool Version:	blastp: 2.2.27+ Package: blast 2.2.27, build Sep 11 2012 10:14:26
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	8a4a8f9f3df4a393
Full Path:	/w/galaxy/dev/galaxy-dist/database/files/001/dataset_1534.dat

Input Parameter	Value	Note for rerun
Protein query sequence(s)	1: human_protein.fas	
Subject database/sequences	db	
Protein BLAST database	uniprot_swissprot	
histdb		
subject		
Type of BLAST	blastp	
Set expectation value cutoff	0.0001	
Output format	Tabular (extended 24 columns)	
Advanced Options	basic	

Inheritance Chain

blastp on db

↑

'blastp on db' in Blastp

History

Human chr22 proteome study

19.1 MB

3: blastp on db

21,556 lines

format: tabular, database: ?

```
1          2          3          4          5          6
sp|P31946|1433B_HUMAN   sp|A4K2U9|1433B_PONAB   100.00  246
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT
sp|P31946|1433B_HUMAN   sp|Q4R572|1433B_MACFA   100.00  246
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT
```

2: chr22_check.gff3

1: human_protein.fas

Visualisation de l'objet

- Tableau
- Image
- HTML

Galaxy / ABiMS | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 18.0 GB

sp|P31946|1433B_HUMAN	sp|A4K2U9|1433B_PONAB	100.00	246	0	0	1	246	1	246	0.0	508	sp|A4K2U9|1433B_PO
sp|P31946|1433B_HUMAN	sp|Q4R572|1433B_MACFA	100.00	246	0	0	1	246	1	246	0.0	508	sp|Q4R572|1433B_MA
sp|P31946|1433B_HUMAN	sp|P31946|1433B_HUMAN	100.00	246	0	0	1	246	1	246	0.0	508	sp|P31946|1433B_HU
sp|P31946|1433B_HUMAN	sp|P68250|1433B_BOVIN	99.59	246	1	0	1	246	1	246	0.0	506	sp|P68250|1433B_BO
sp|P31946|1433B_HUMAN	sp|Q9CQV8|1433B_MOUSE	98.78	246	3	0	1	246	1	246	3e-179	499	sp|Q9CQV8|1433B_M
sp|P31946|1433B_HUMAN	sp|P35213|1433B_RAT	98.37	246	4	0	1	246	1	246	2e-178	497	sp|P35213|1433B_RA
sp|P31946|1433B_HUMAN	sp|Q5ZLQ6|1433B_CHICK	97.95	244	5	0	3	246	1	244	3e-177	494	sp|Q5ZLQ6|1433B_CH
sp|P31946|1433B_HUMAN	sp|Q6UFZ9|143B1_ONCMY	91.80	244	20	0	3	246	1	244	1e-164	462	sp|Q6UFZ9|143B1_ON
sp|P31946|1433B_HUMAN	sp|Q5PRD0|143BA_DANRE	91.39	244	21	0	3	246	1	244	3e-164	461	sp|Q5PRD0|143BA_DA
sp|P31946|1433B_HUMAN	sp|Q5XHK2|143BA_XENLA	88.11	244	29	0	3	246	1	244	2e-159	449	sp|Q5XHK2|143BA_XE
sp|P31946|1433B_HUMAN	sp|Q8AVQ3|143BB_XENLA	87.70	244	30	0	3	246	1	244	6e-159	447	sp|Q8AVQ3|143BB_XE
sp|P31946|1433B_HUMAN	sp|Q5XGC8|1433B_XENTR	88.52	244	28	0	3	246	1	244	6e-159	447	sp|Q5XGC8|1433B_XE
sp|P31946|1433B_HUMAN	sp|P63102|1433Z_RAT	88.02	242	29	0	3	244	1	242	2e-157	444	sp|P63102|1433Z_RAT
sp|P31946|1433B_HUMAN	sp|P63101|1433Z_MOUSE	88.02	242	29	0	3	244	1	242	2e-157	444	sp|P63101|1433Z_MO
sp|P31946|1433B_HUMAN	sp|P29361|1433Z_SHEEP	87.60	242	30	0	3	244	1	242	1e-156	442	sp|P29361|1433Z_SH
sp|P31946|1433B_HUMAN	sp|Q5R651|1433Z_PONAB	87.60	242	30	0	3	244	1	242	3e-156	441	sp|Q5R651|1433Z_PO
sp|P31946|1433B_HUMAN	sp|P63104|1433Z_HUMAN	87.60	242	30	0	3	244	1	242	3e-156	441	sp|P63104|1433Z_HU
sp|P31946|1433B_HUMAN	sp|P63103|1433Z_BOVIN	87.60	242	30	0	3	244	1	242	3e-156	441	sp|P63103|1433Z_BO
sp|P31946|1433B_HUMAN	sp|Q5ZKC9|1433Z_CHICK	87.19	242	31	0	3	244	1	242	7e-156	440	sp|Q5ZKC9|1433Z_CH
sp|P31946|1433B_HUMAN	sp|Q7T356|143BB_DANRE	88.52	244	26	1	3	246	1	242	2e-155	439	sp|Q7T356|143BB_DA
sp|P31946|1433B_HUMAN	sp|P29309|1433_XENLA	87.23	235	30	0	12	246	1	235	6e-152	429	sp|P29309|1433_XEN
sp|P31946|1433B_HUMAN	sp|Q6UFZ8|143B2_ONCMY	88.41	233	27	0	3	235	1	233	3e-150	426	sp|Q6UFZ8|143B2_ON
sp|P31946|1433B_HUMAN	sp|Q6P4Z5|1433Z_XENTR	84.71	242	37	0	3	244	1	242	4e-150	426	sp|Q6P4Z5|1433Z_XE
sp|P31946|1433B_HUMAN	sp|Q91896|1433Z_XENLA	83.47	242	40	0	3	244	1	242	3e-148	421	sp|Q91896|1433Z_XE
sp|P31946|1433B_HUMAN	sp|Q5ZMD1|1433T_CHICK	82.23	242	43	0	3	244	1	242	1e-147	419	sp|Q5ZMD1|1433T_CH
sp|P31946|1433B_HUMAN	sp|Q5RFJ2|1433T_PONAB	81.82	242	44	0	3	244	1	242	2e-146	416	sp|Q5RFJ2|1433T_PO
sp|P31946|1433B_HUMAN	sp|P27348|1433T_HUMAN	81.82	242	44	0	3	244	1	242	2e-146	416	sp|P27348|1433T_HU
sp|P31946|1433B_HUMAN	sp|Q3SZI4|1433T_BOVIN	81.82	242	44	0	3	244	1	242	2e-146	416	sp|Q3SZI4|1433T_BO
sp|P31946|1433B_HUMAN	sp|Q52M98|1433T_XENLA	81.82	242	44	0	3	244	1	242	2e-146	416	sp|Q52M98|1433T_XE
sp|P31946|1433B_HUMAN	sp|P68255|1433T_RAT	81.82	242	44	0	3	244	1	242	2e-146	416	sp|P68255|1433T_RA
sp|P31946|1433B_HUMAN	sp|Q6Q6X0|1433T_RABIT	81.82	242	44	0	3	244	1	242	2e-146	416	sp|Q6Q6X0|1433T_RA
sp|P31946|1433B_HUMAN	sp|P68254|1433T_MOUSE	81.82	242	44	0	3	244	1	242	2e-146	416	sp|P68254|1433T_MO
sp|P31946|1433B_HUMAN	sp|Q2F637|1433Z_BOMMO	79.92	244	49	0	1	244	1	244	3e-143	408	sp|Q2F637|1433Z_BO
sp|P31946|1433B_HUMAN	sp|P29310|1433Z_DROME	79.42	243	50	0	2	244	3	245	3e-142	405	sp|P29310|1433Z_DR
sp|P31946|1433B_HUMAN	sp|Q1HR36|1433Z_AEDAE	78.60	243	52	0	2	244	3	245	2e-140	401	sp|Q1HR36|1433Z_AE
sp|P31946|1433B_HUMAN	sp|Q20655|14332_CAEEL	78.78	245	52	0	1	245	1	245	7e-140	399	sp|Q20655|14332_CA
sp|P31946|1433B_HUMAN	sp|P41932|14331_CAEEL	78.10	242	48	2	7	244	7	247	1e-131	378	sp|P41932|14331_CA

History

Human chr22 proteome study

19.1 MB

3: blastp on db

21,556 lines

format: tabular, database: ?

View data

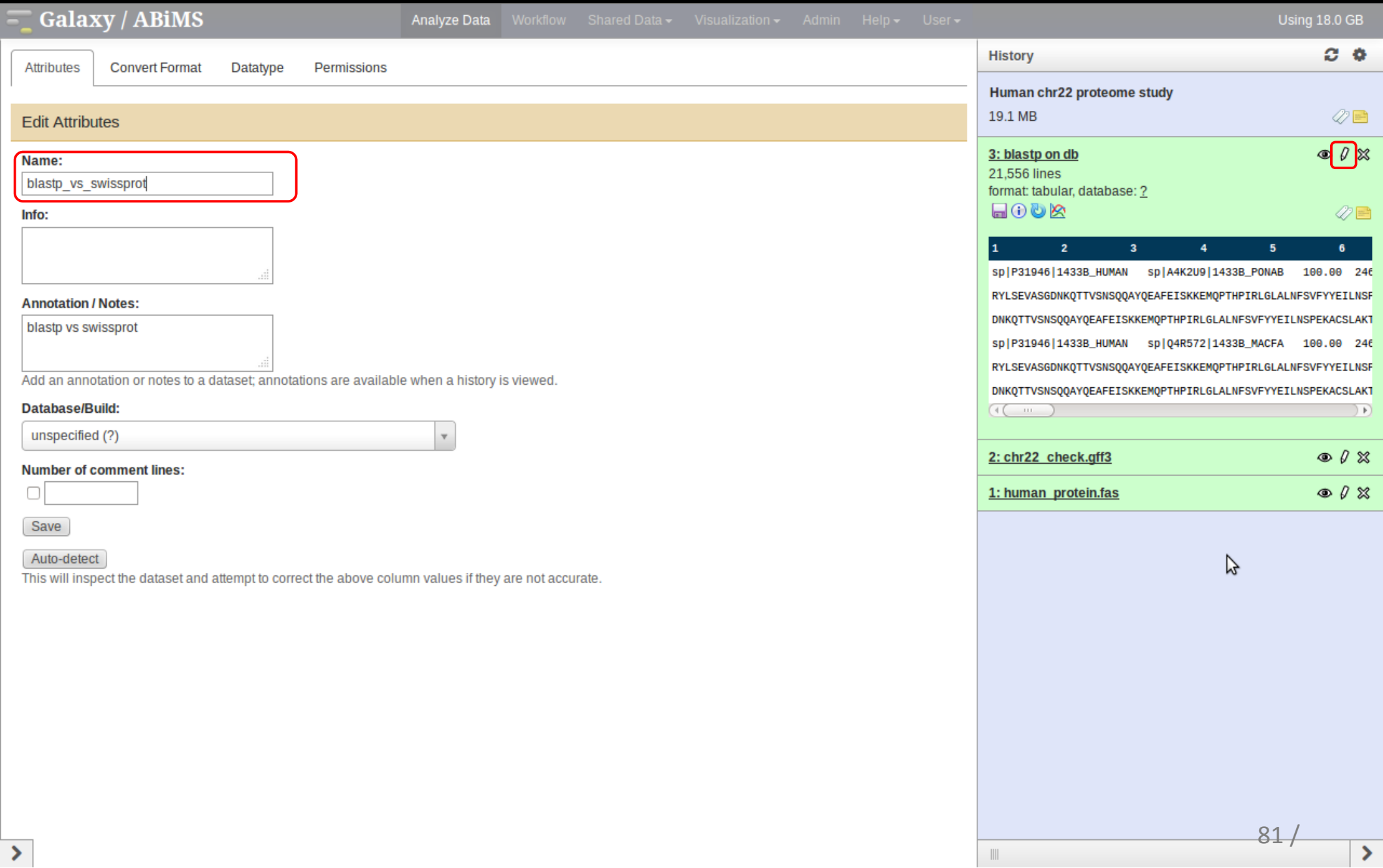
1	2	3	4	5	6
sp|P31946|1433B_HUMAN	sp|A4K2U9|1433B_PONAB	100.00	246		
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP					
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT					
sp|P31946|1433B_HUMAN	sp|Q4R572|1433B_MACFA	100.00	246		
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP					
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT					

2: chr22_check.gff3

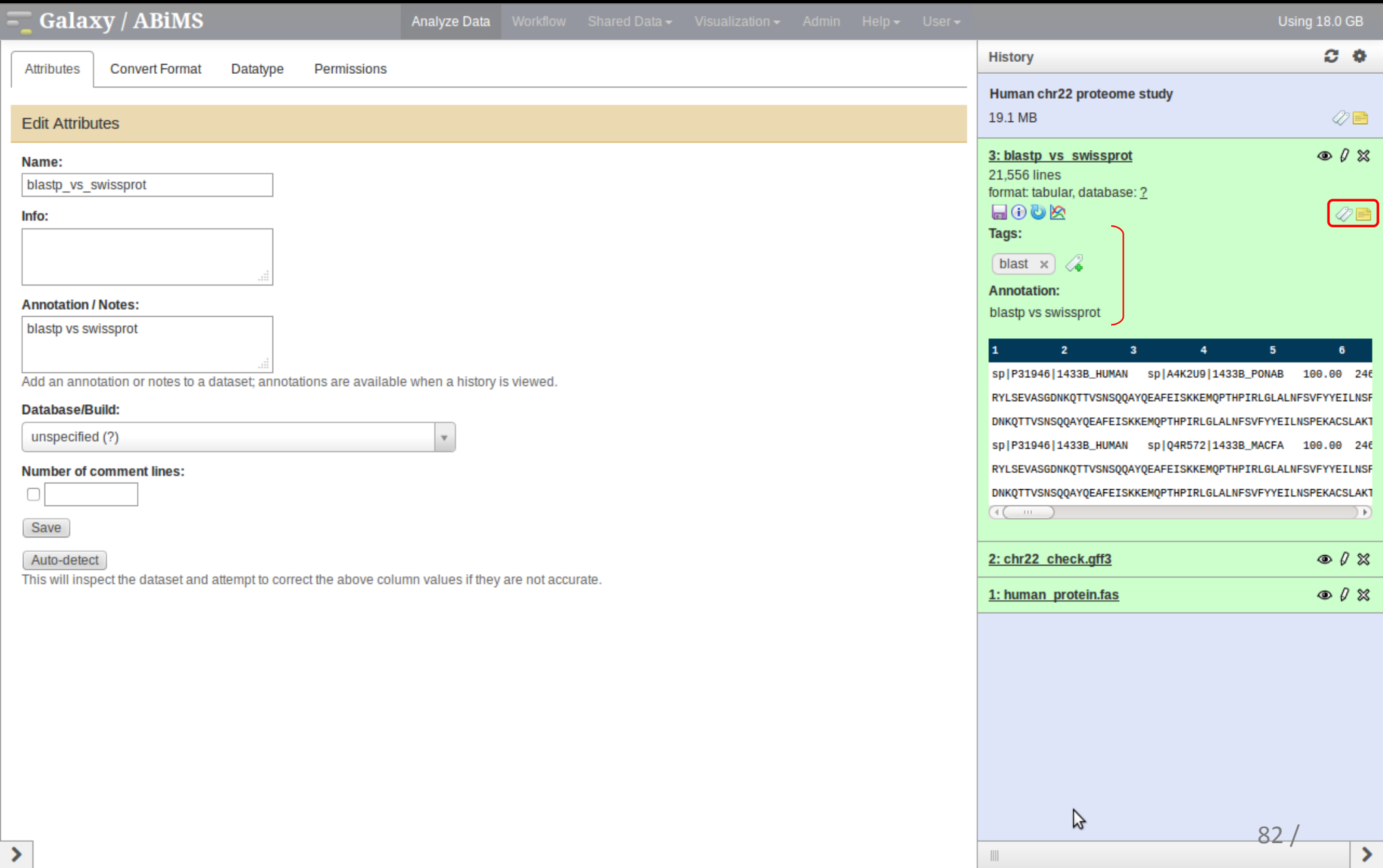
1: human_protein.fas

Modification des attributs

- Renommer
- Annoter…

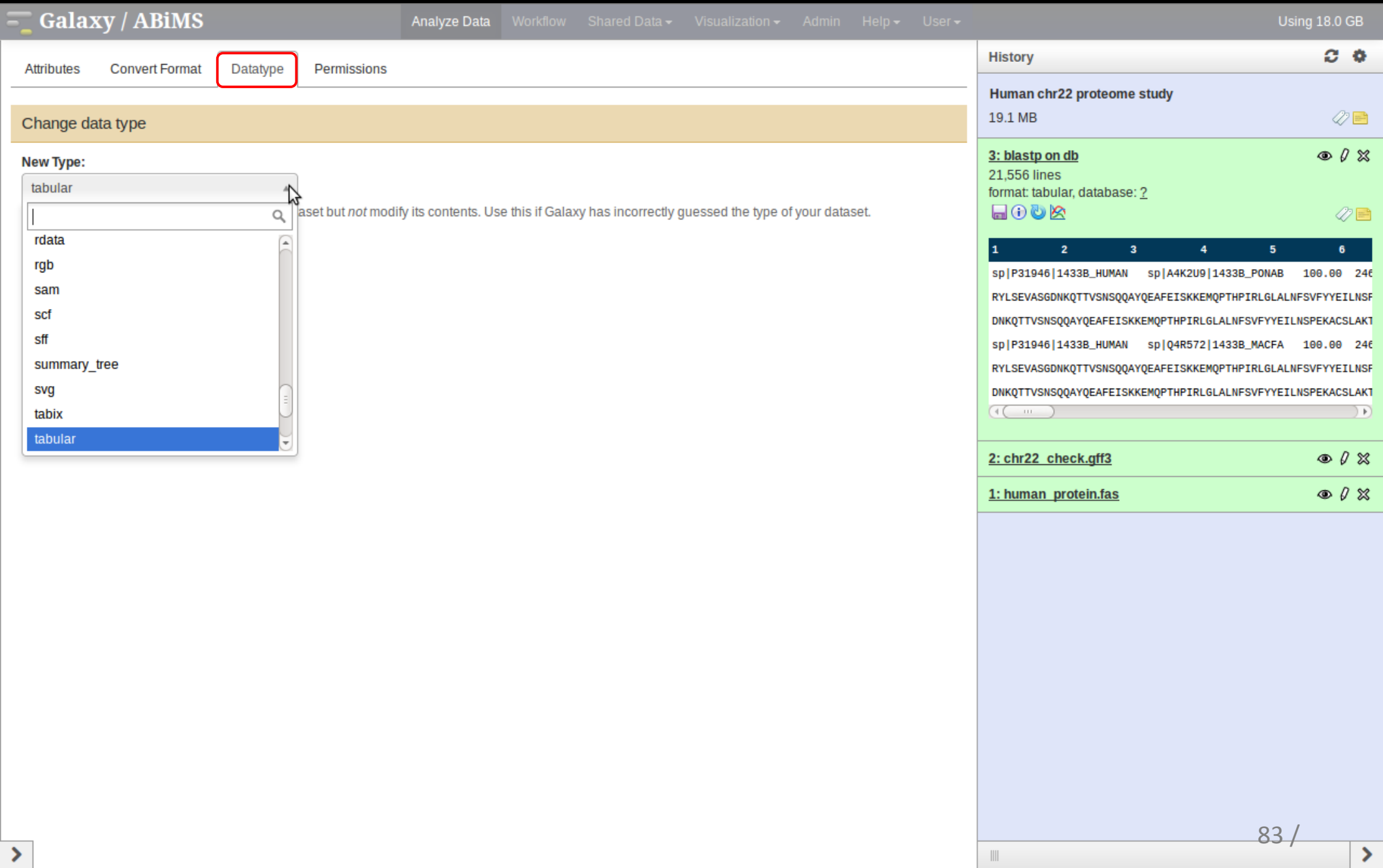


Ajout de tags et d'annotations



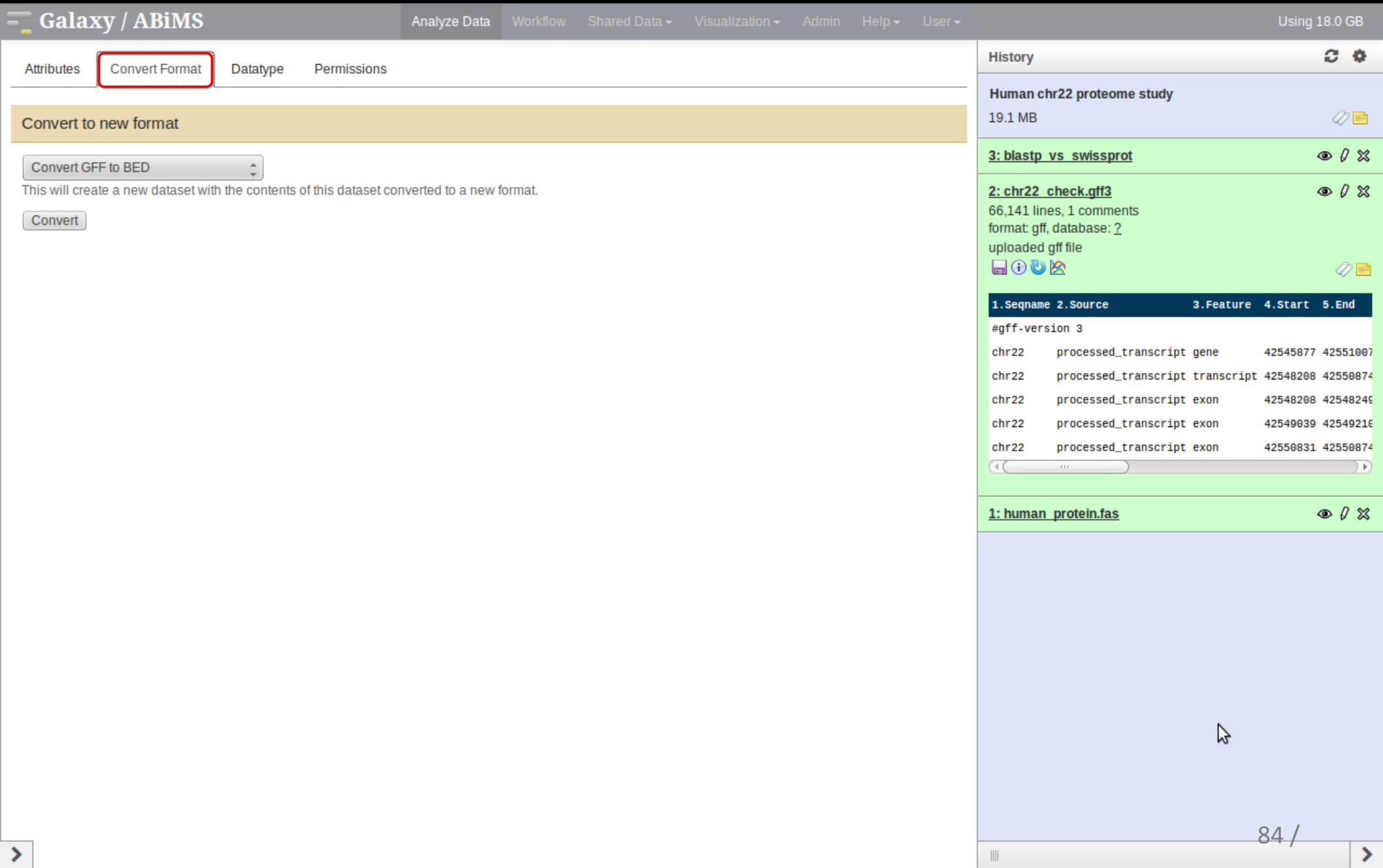
Modification du type de fichier

- Tabular → gtf
- fastq → fastqsanger (phred33)



Conversion de type

Dans la limite des stocks disponibles

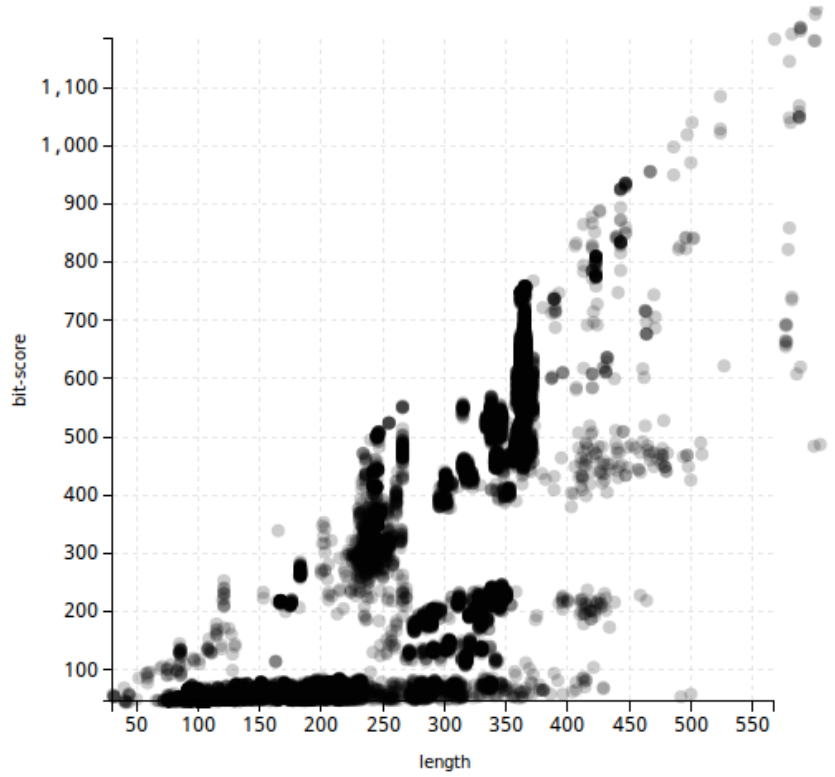


Visualisation

- Tabular → graphique de type scatterplot
- gff/bam → visualisation dans un système de track

Scatterplot of 'blastp_vs_swissprot'

Data Controls
Chart Controls
Statistics
Chart



History

Human chr22 proteome study
19.1 MB

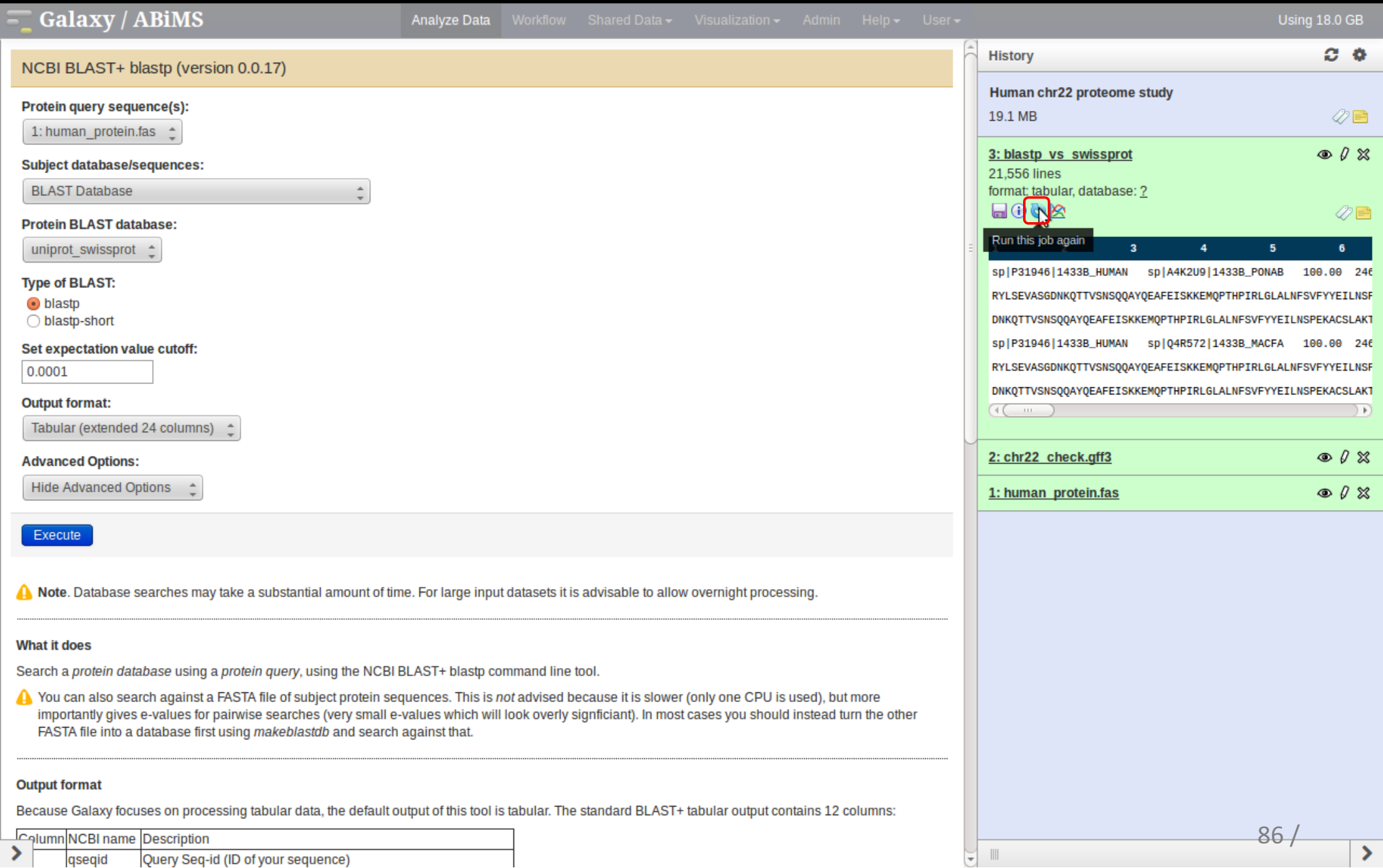
3: blastp_vs_swissprot
21,556 lines
format: tabular, database: ?

1	2	3	4	5	6
sp|P31946|1433B_HUMAN	sp|A4K2U9|1433B_PONAB	100.00	246		
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP					
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT					
sp|P31946|1433B_HUMAN	sp|Q4R572|1433B_MACFA	100.00	246		
RYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSP					
DNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYYEILNSPEKACSLAKT					

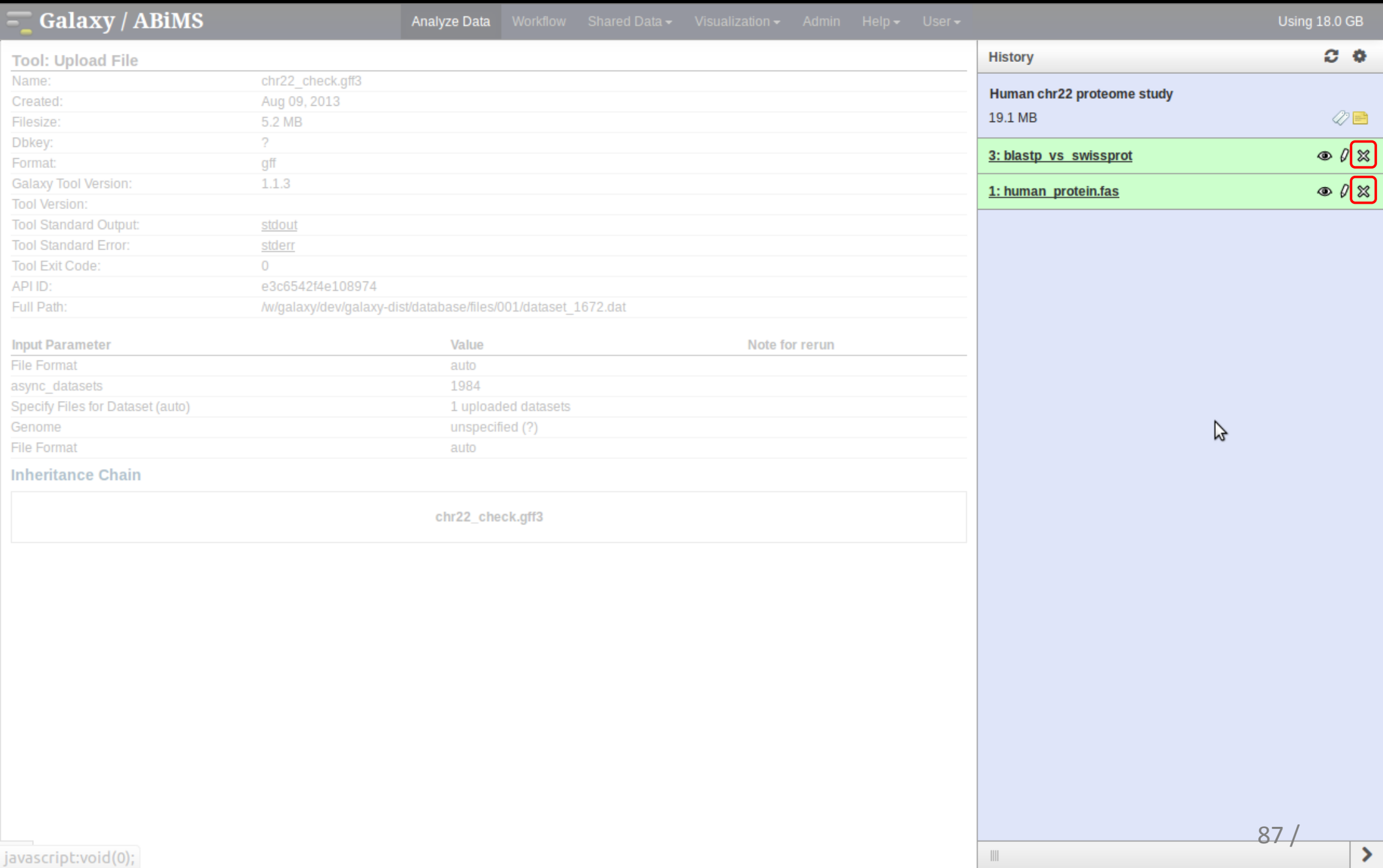
2: chr22_check.gff3

1: human_protein.fas

Re-ouverture du formulaire à l'origine de l'objet avec ces options pré-remplies

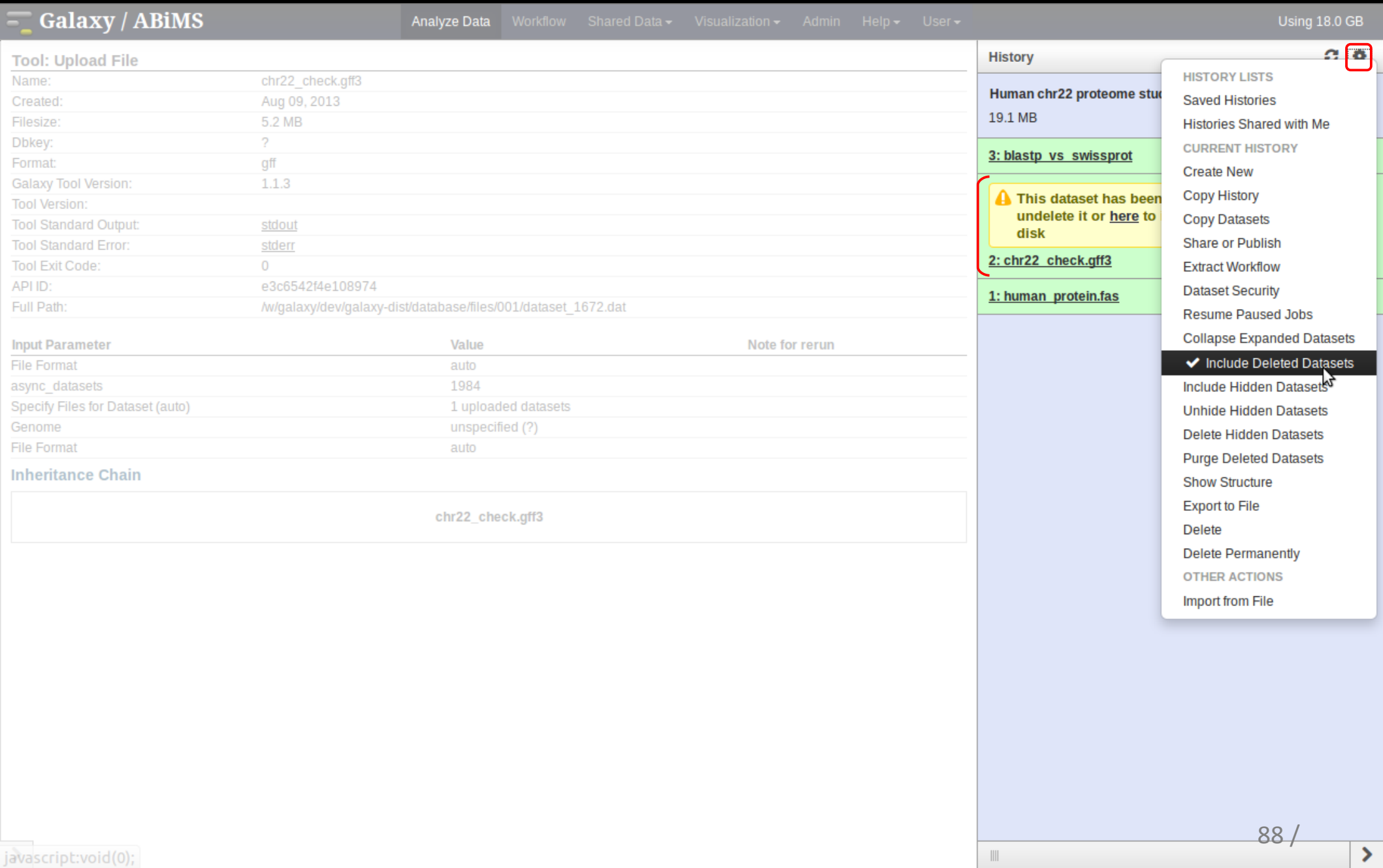


Suppression d'un objet



Suppression d'un objet

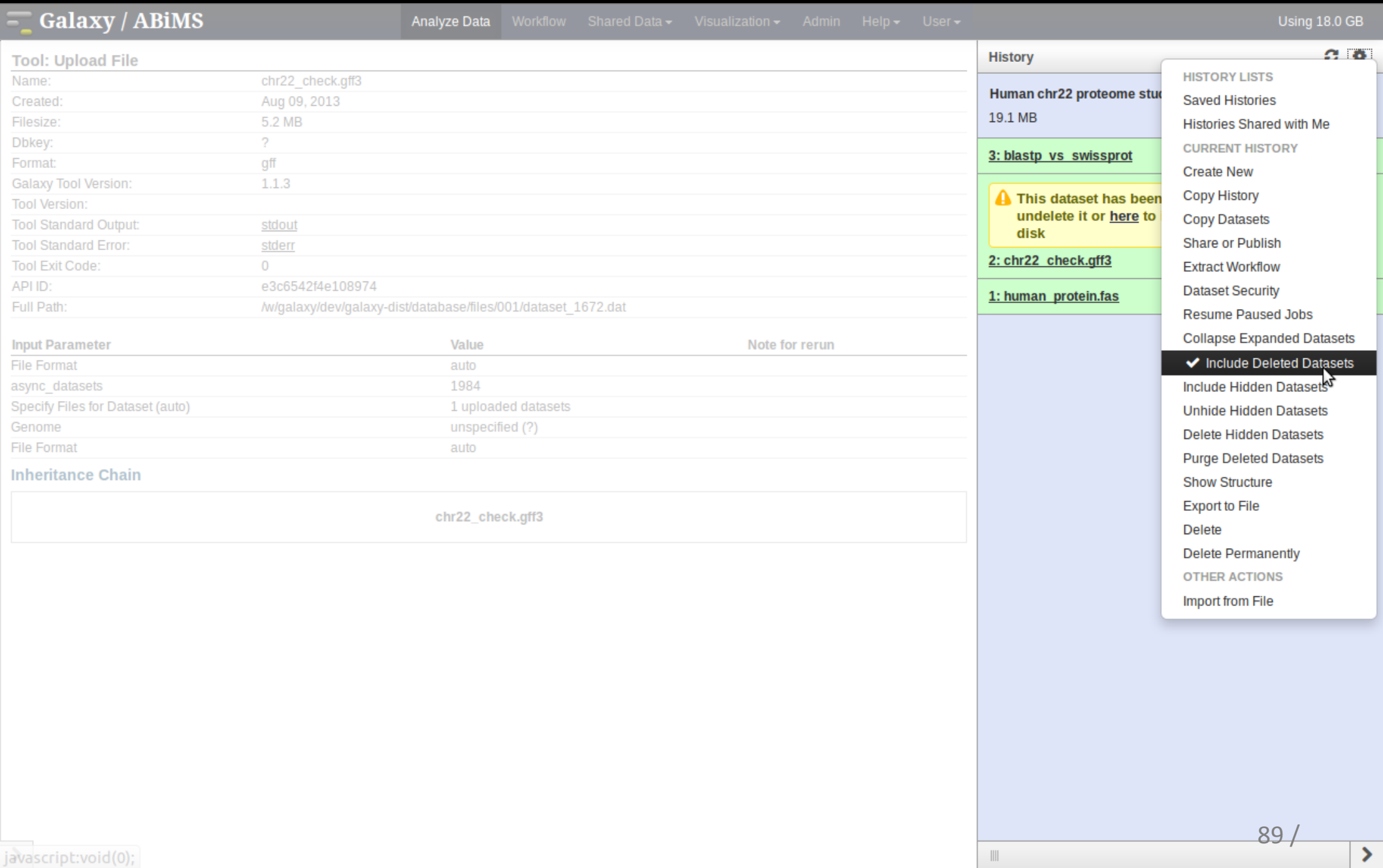
Mais attention, l'objet est dans une sorte de corbeille



Suppression d'un objet

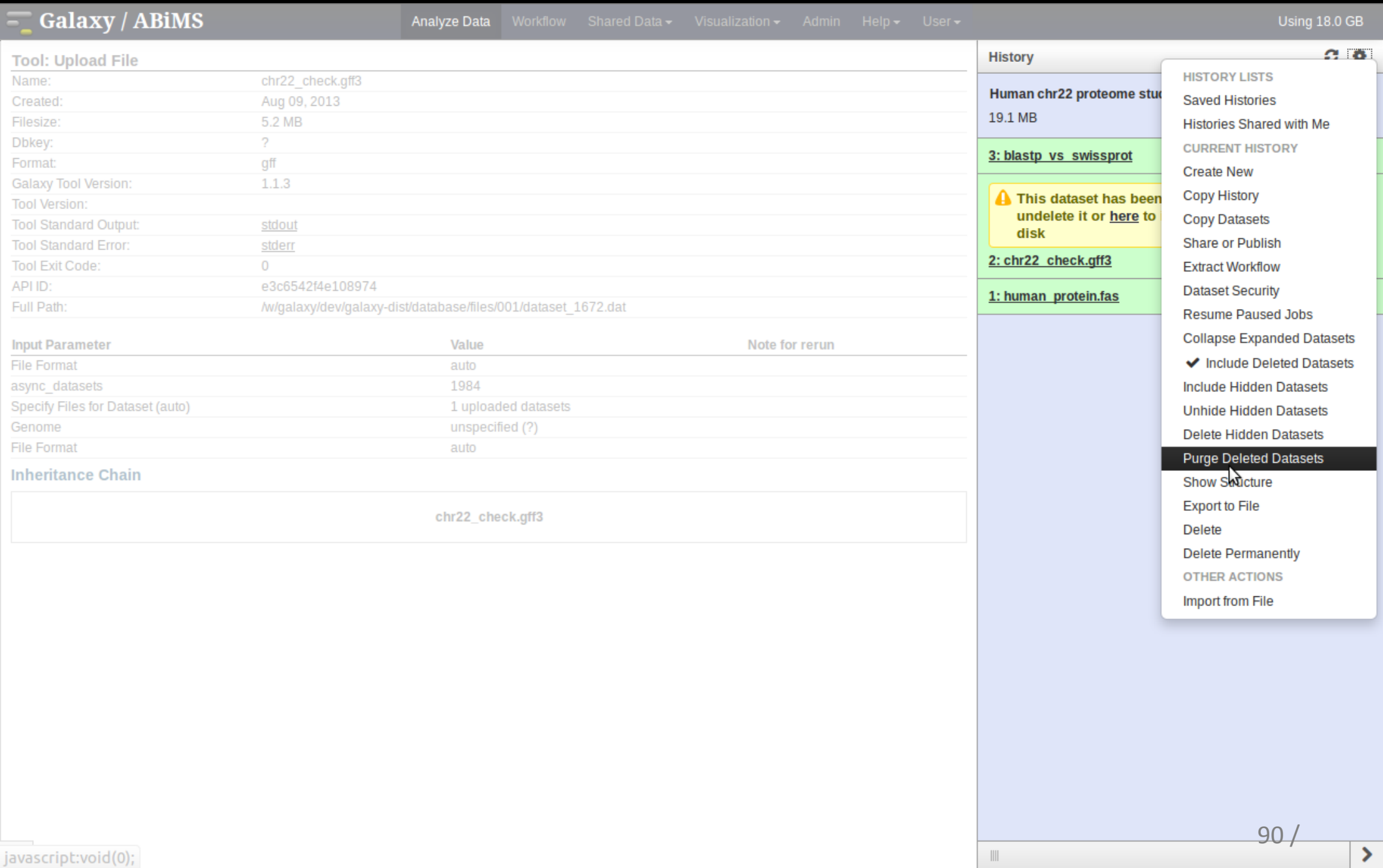
Mais attention, l'objet est dans une sorte de corbeille

« Du coup, l'espace disque n'est pas libéré »



Purge des objets deletés

Mais attention, là, c'est pour de bon !

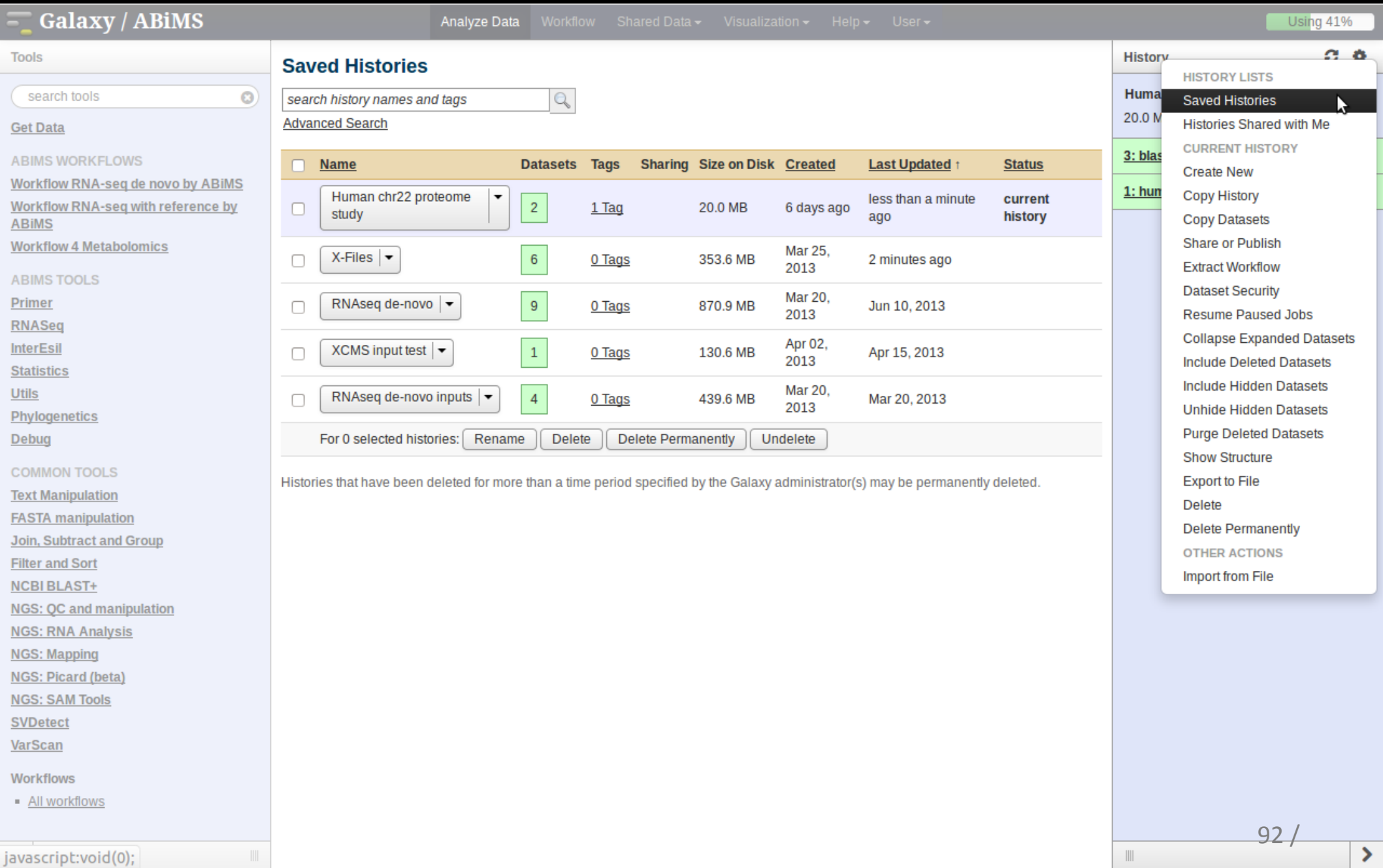


Historique

- Cycle de vie des données
  - http://abims.sb-roscoff.fr/galaxyproject
  - Gestion par l'utilisateur de son quota
  - Suppression automatique en cours de réflexion
    - 6 mois ou 1 an

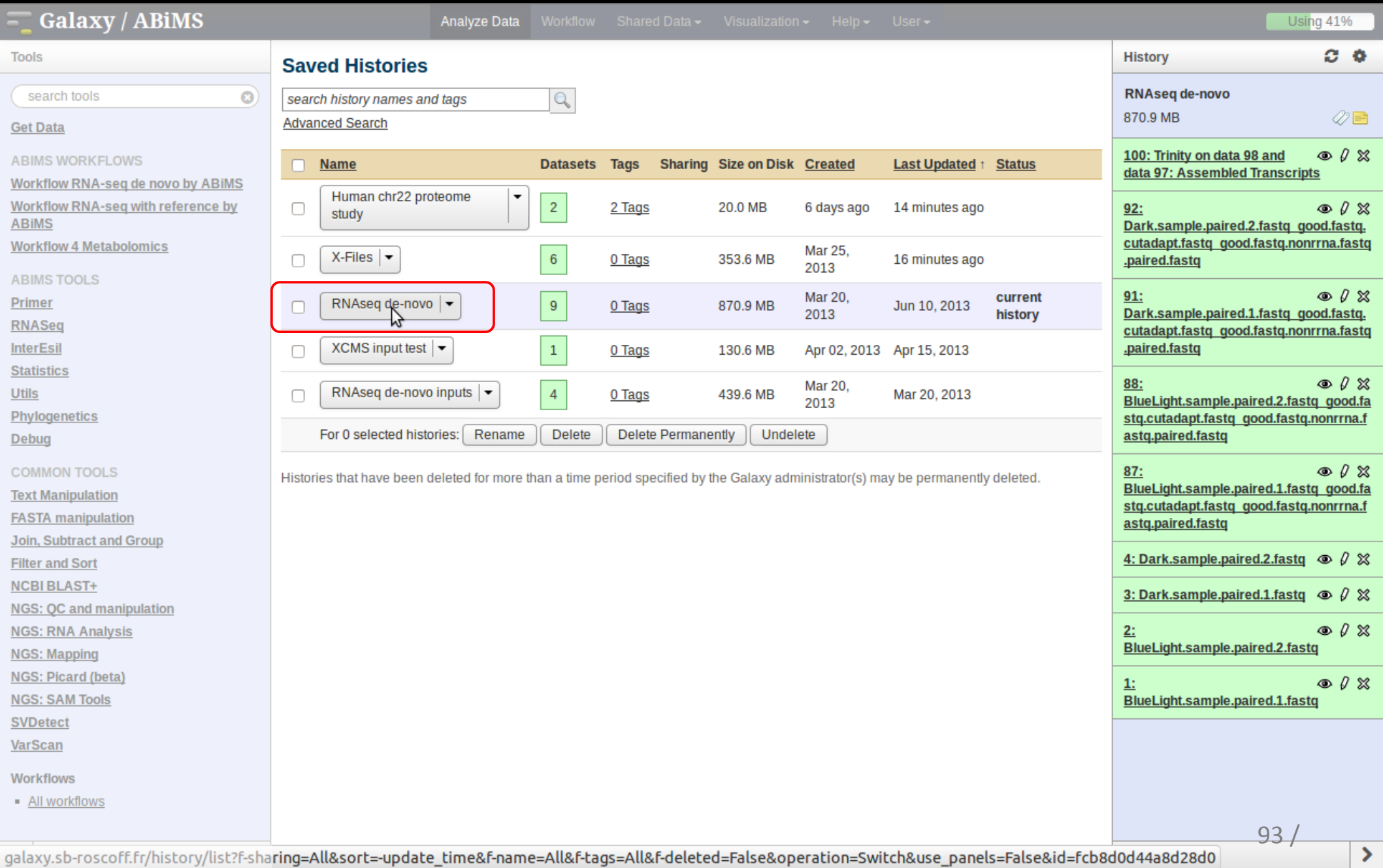
Saved history :

- Renommer
- Supprimer [définitivement]



Saved history :

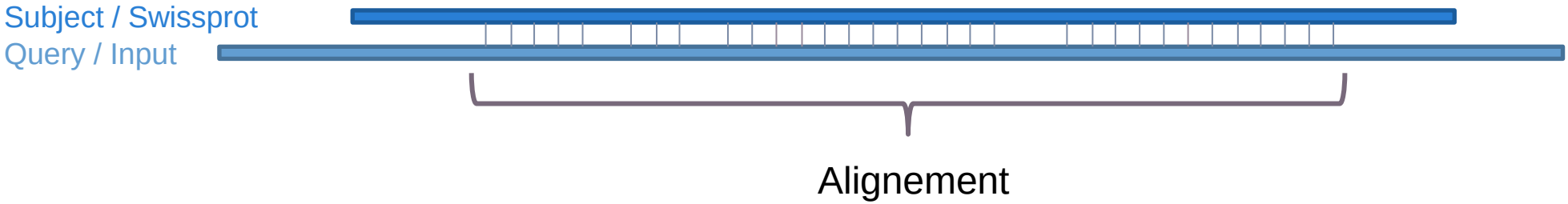
- Renommer
- Supprimer [définitivement]



OUTILS / EXERCICE 1

Outils

- Exercice 3e kyu :
  - Récupérer les identifiants des hits obtenus à partir de vos séquences « blastées » sur swissprot :
    - Identité > 75 %
    - Couverture de l'alignement > 75 %



Couverture = Longueur alignement / Longueur query

Outils

- Exercice 4e kyu :
  - Lancer un blast de vos séquences sur swissprot
    - Regarder quel blast convient à vos données d'entrée et à votre banque
  - Calculer le pourcentage de couverture de l'alignement par rapport à la longueur effective de la sequence query (pensez à regarder la signification des colonnes dans l'aide de blast)
  - Filtrer
    - Identité > 75 %
    - Couverture de l'alignement > 75 %

Outils

- Exercice 5e kyu :
  - Lancer un **blastx** de vos séquences sur swissprot
  - Calculer les longueurs de vos séquences
  - Fusionner le tableau de résultats du blast avec le résultat des longueurs de séquence (données tabulées)
  - Calculer le pourcentage de couverture de l'alignement par rapport à la longueur effective de la sequence query (pensez à regarder la signification des colonnes dans l'aide du blast)
  - Filtrer
    - Identité > 75 %
    - Couverture de l'alignement > 75 %

Outils

- Exercice 6e kyu :
  - Lancer un **blastx** de vos séquences sur swissprot
  - Calculer les longueurs de vos séquences (**Compute sequence length**)
  - Fusionner le tableau de résultats du blast avec le résultat des longueurs de séquence (données tabulées)
    - Si la fusion ne fonctionne pas, c'est qu'il y a un piège avec les identifiants.
  - Calculer le pourcentage de couverture de l'alignement par rapport à la longueur effective de la sequence query (pensez à regarder la signification des colonnes dans l'aide du blast)
  - Filtrer
    - Identité > 75 %
    - Couverture de l'alignement > 75 %

Outils

- Exercice 7e kyu :
  - Lancer un **blastx** de vos séquences sur swissprot
  - Calculer les longueurs de vos séquences (**Compute sequence length**)
  - Fusionner le tableau de résultats du blast avec le résultat des longueurs de séquence (données tabulées)
    - Si la fusion ne fonctionne pas, c'est qu'il y a un piège avec les identifiants.
    - Une des pistes est de travailler le tableau des longueurs de séquences pour éliminer "path ...."
  - Calculer le pourcentage de couverture de l'alignement par rapport à la longueur effective de la sequence query (pensez à regarder la signification des colonnes dans l'aide du blast)
  - Filtrer
    - Identité > 75 %
    - Couverture de l'alignement > 75 %

Outils

- Exercice 8e kyu :
  - Lancer un **blastx** de vos séquences sur swissprot
  - Calculer les longueurs de vos séquences (**Compute sequence length**)
  - Fusionner (**join**) le tableau de résultats du blast avec le résultat des longueurs de séquence (données tabulées)
    - Si la fusion ne fonctionne pas, c'est qu'il y a un piège avec les identifiants.
    - Une des pistes est de travailler le tableau des longueurs de séquences pour éliminer "path ...." en utilisant la commande **cut** et en jouant sur le délimitateur de champs. Vous pouvez ainsi obtenir 2 datasets que vous *collez* l'un à côté de l'autre
  - Calculer (**compute**) le pourcentage de couverture de l'alignement par rapport Ã la longueur effective de la sequence query (pensez à regarder la signification des colonnes dans l'aide du blast)
  - Filtrer (**filter**)
    - Identité > 75 %
    - Couverture de l'alignement > 75 %

Outils

- Exercice 9e kyu :
  - Lancer un **blastx** de vos séquences sur swissprot
  - Calculer les longueurs de vos séquences (**Compute sequence length**)
  - Fusionner (**join**) le tableau de résultats du blast avec le résultat des longueurs de séquence (données tabulées)
    - Si la fusion ne fonctionne pas, c'est qu'il y a un piège avec les identifiants.
    - Une des pistes est de travailler le tableau des longueurs de séquences pour éliminer "path ...."
      - Un **cut** pour récupérer la premier colonne en utilisant l'espace
      - Un **cut** pour récupérer la seconde colonne en utilisant la tabulation
      - Un **paste** pour coller les 2 tableaux
  - Calculer (**compute**) le pourcentage de couverture de l'alignement par rapport Ã la longueur effective de la sequence query : ((c8-c7+1)/c14)*100
  - Filtrer (**filter**)
    - (c3 >= 75 and abs(c15) >= 75)
    - Identité > 75 % et Couverture de l'alignement > 75 %

OUTILS / EXERCICE 2

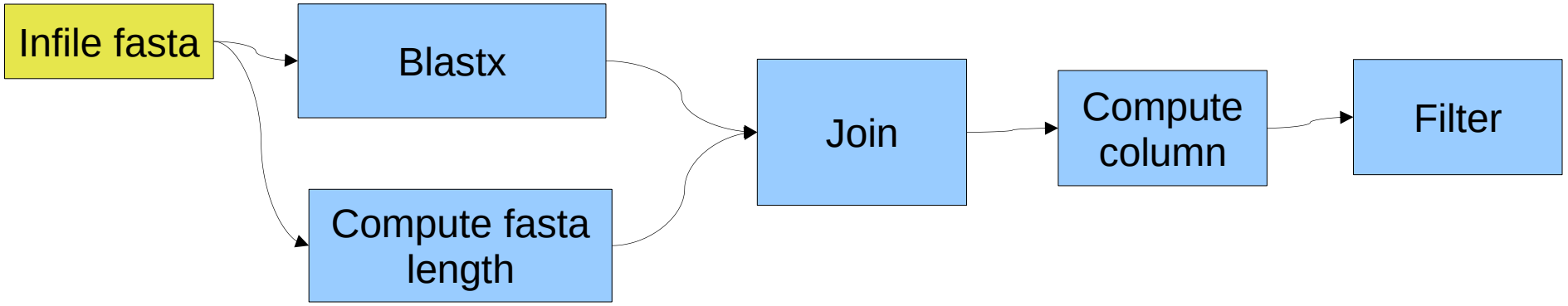
WORKFLOW

Workflow

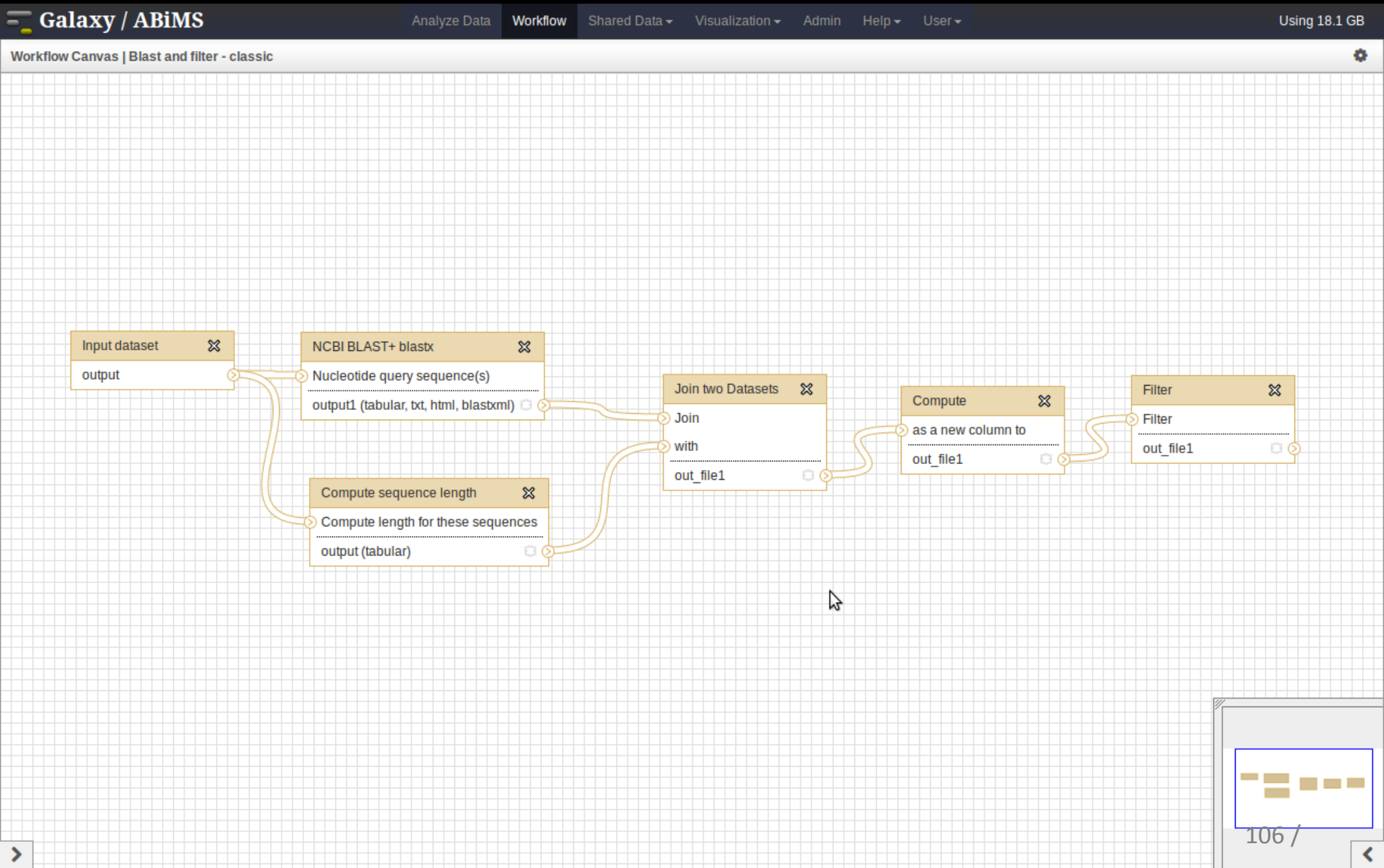
- Un workflow est un enchaînement d'outils et de paramètres
- Peut correspondre au protocole de l'expérience
- Un workflow est construit pour être rejoué (de manière plus ou moins stricte)

Workflow

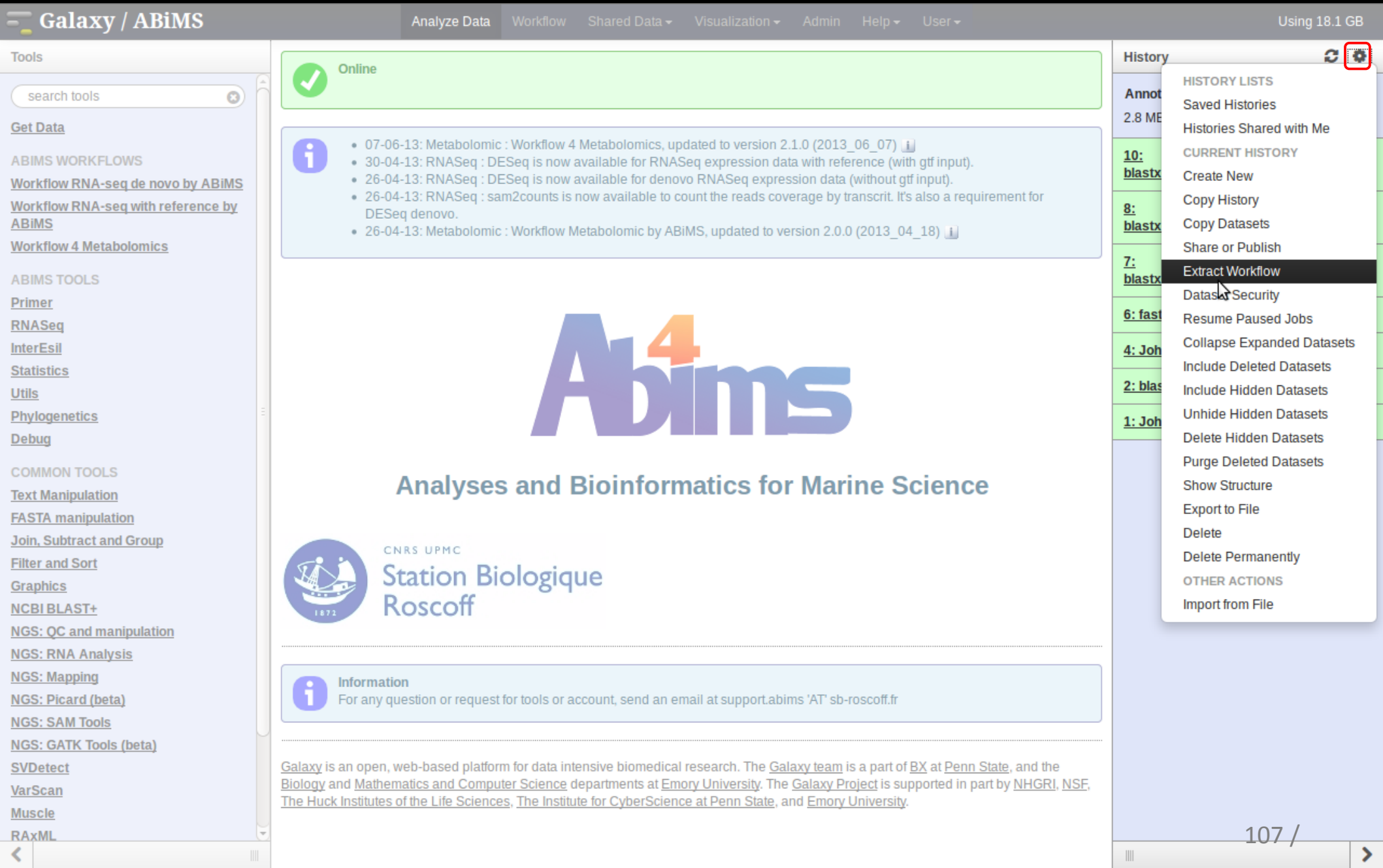
- Notre workflow



Notre workflow sauce Galaxy

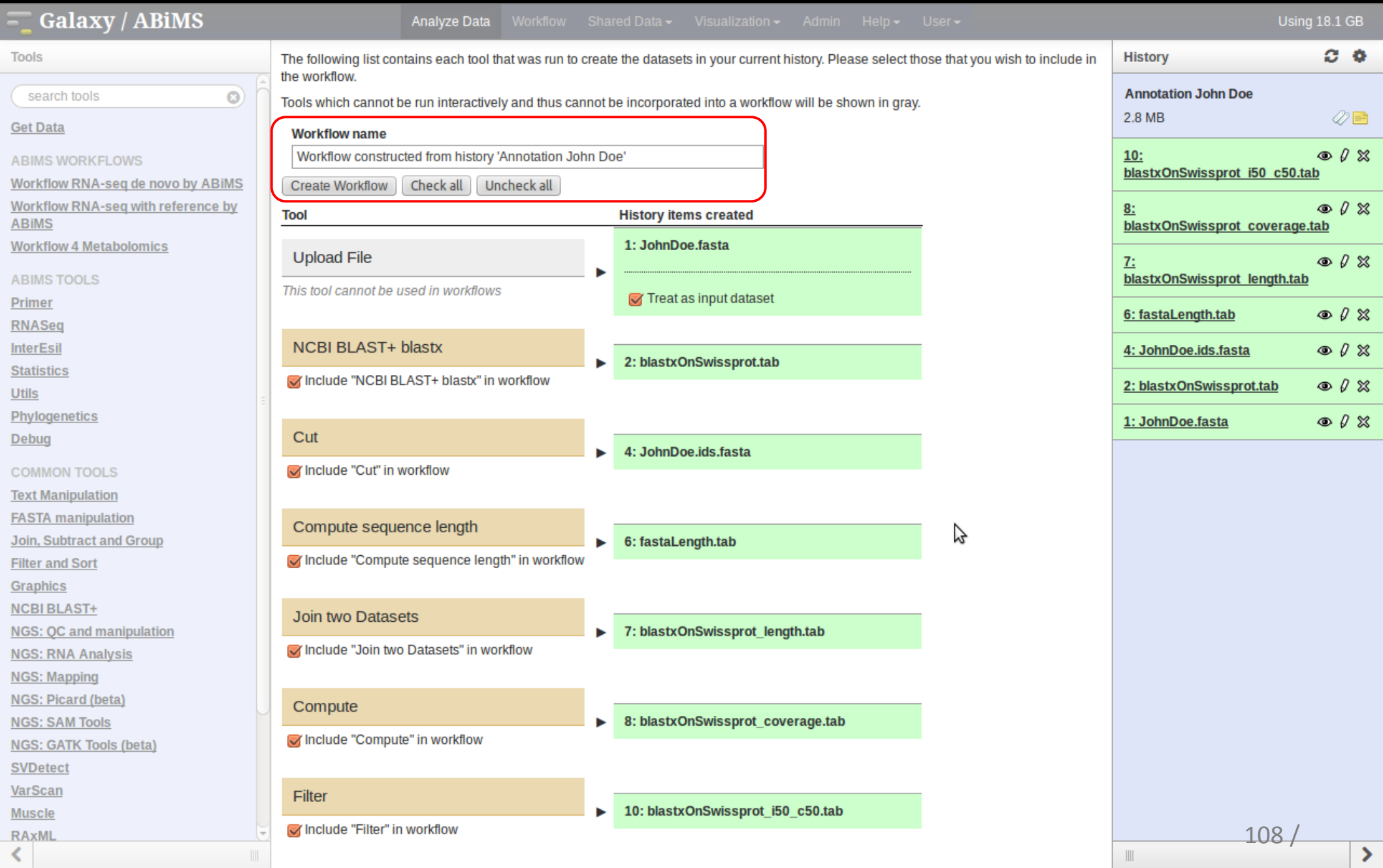


Création d'un workflow à partir d'un historique



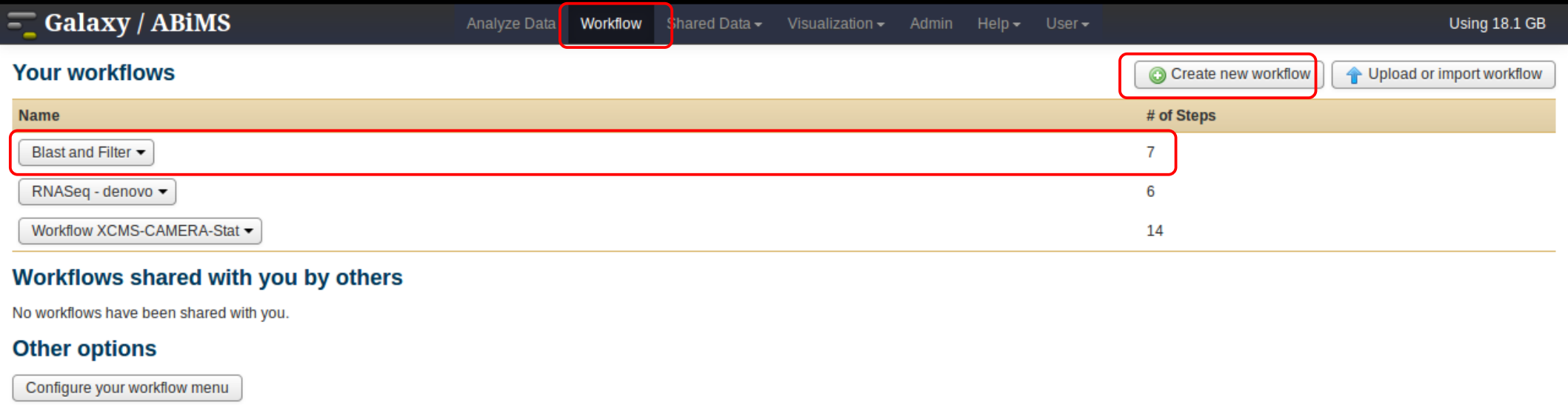
Création d'un workflow à partir d'un historique

- Changement du nom
- Possibilité de décocher des étapes

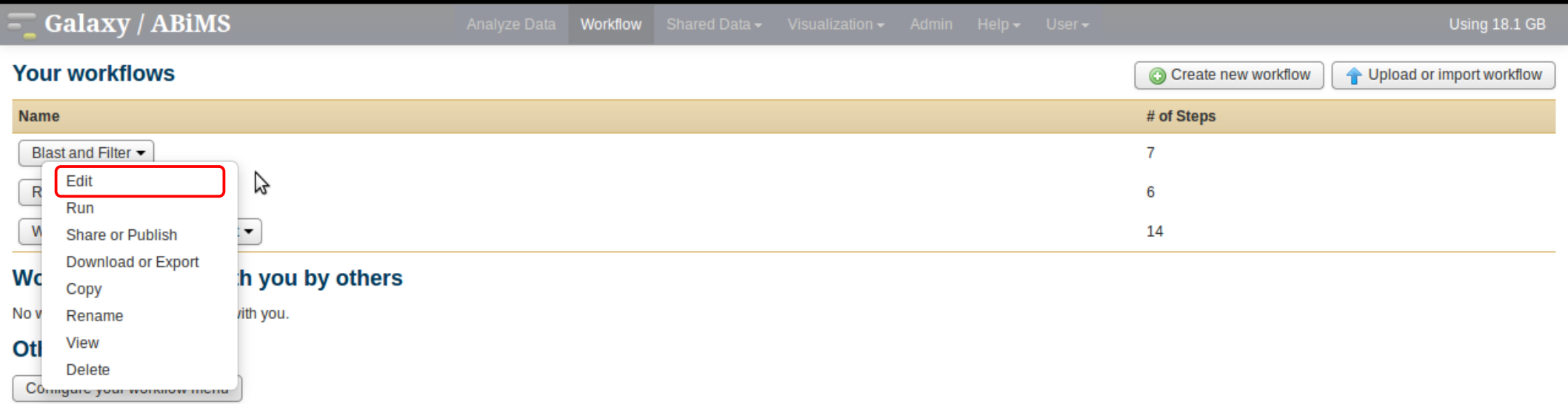


Liste des workflows

- Possibilité de créer un workflow de zéro



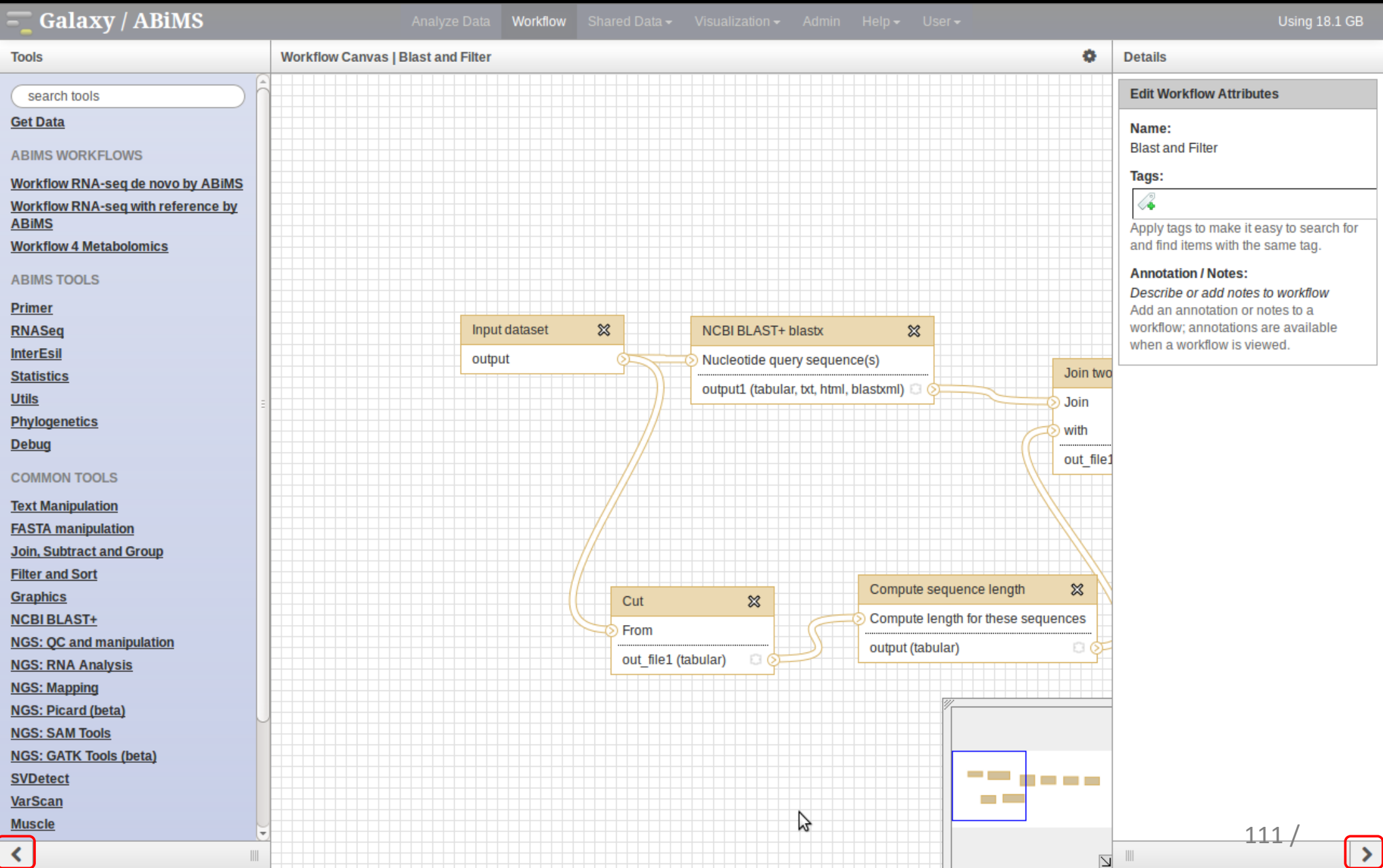
Edition d'un workflow



Edition d'un workflow :

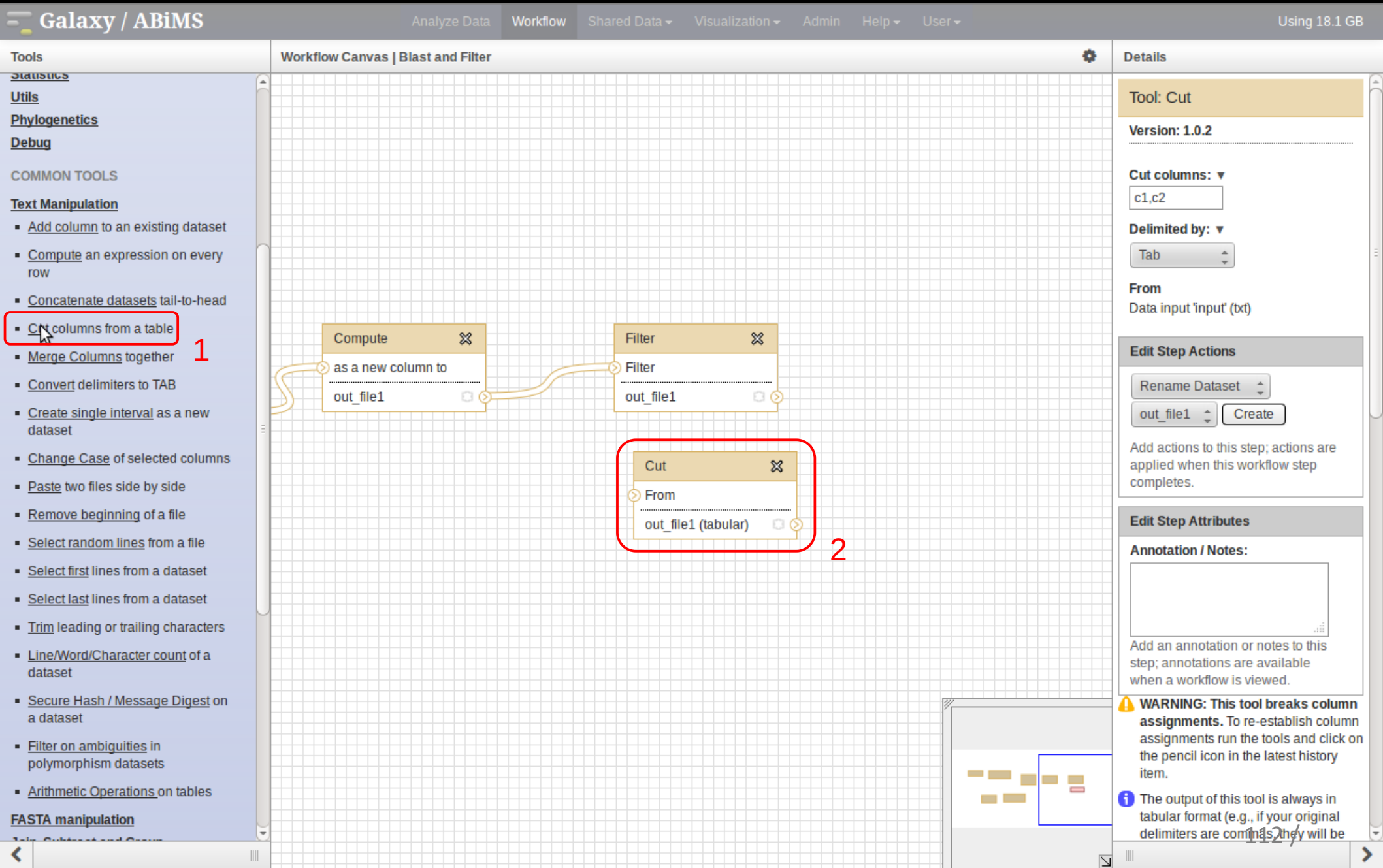
Le canevas

- Glisser-Déposé



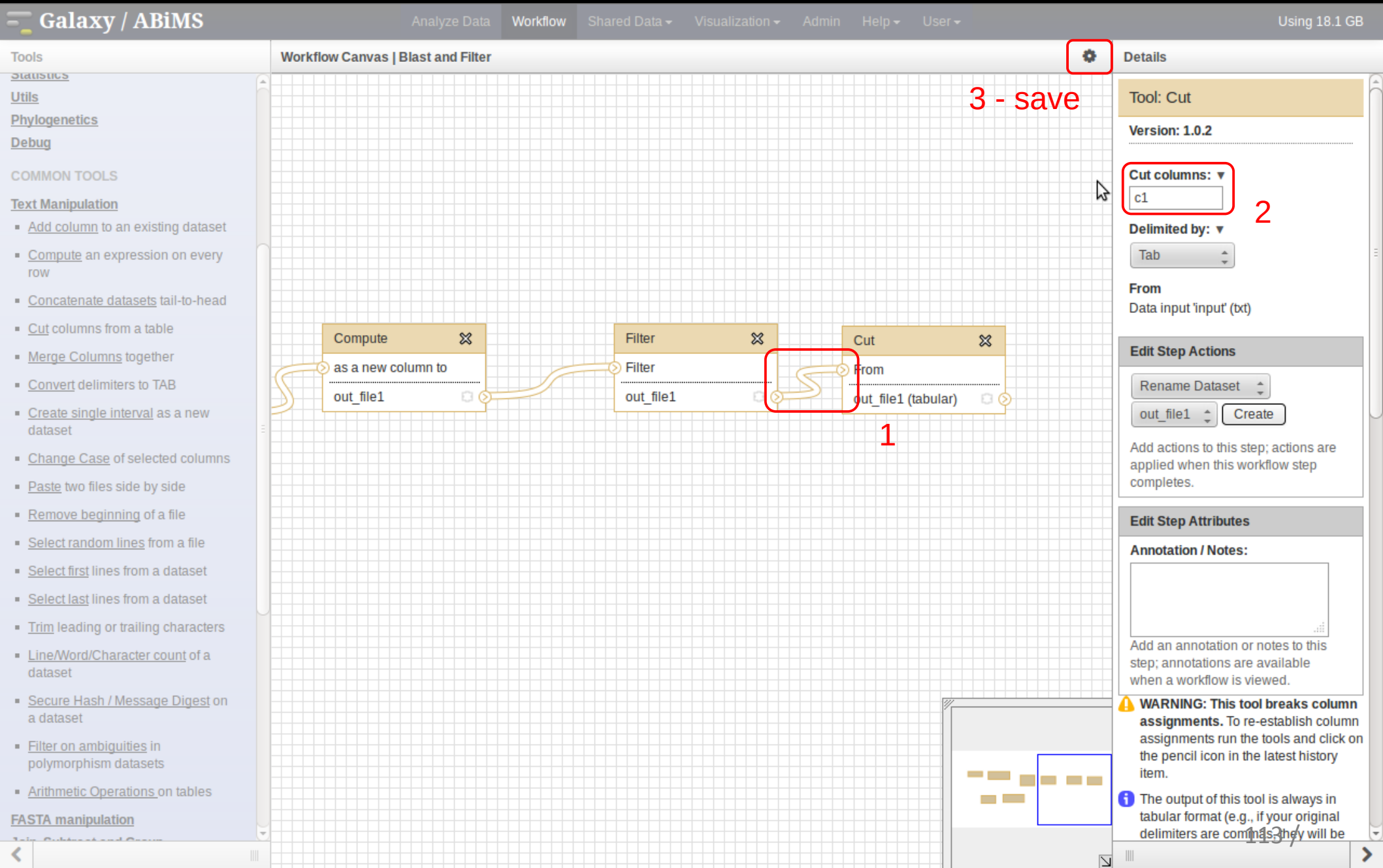
Edition d'un workflow :

- Ajout d'un nouvel outil dans le workflow



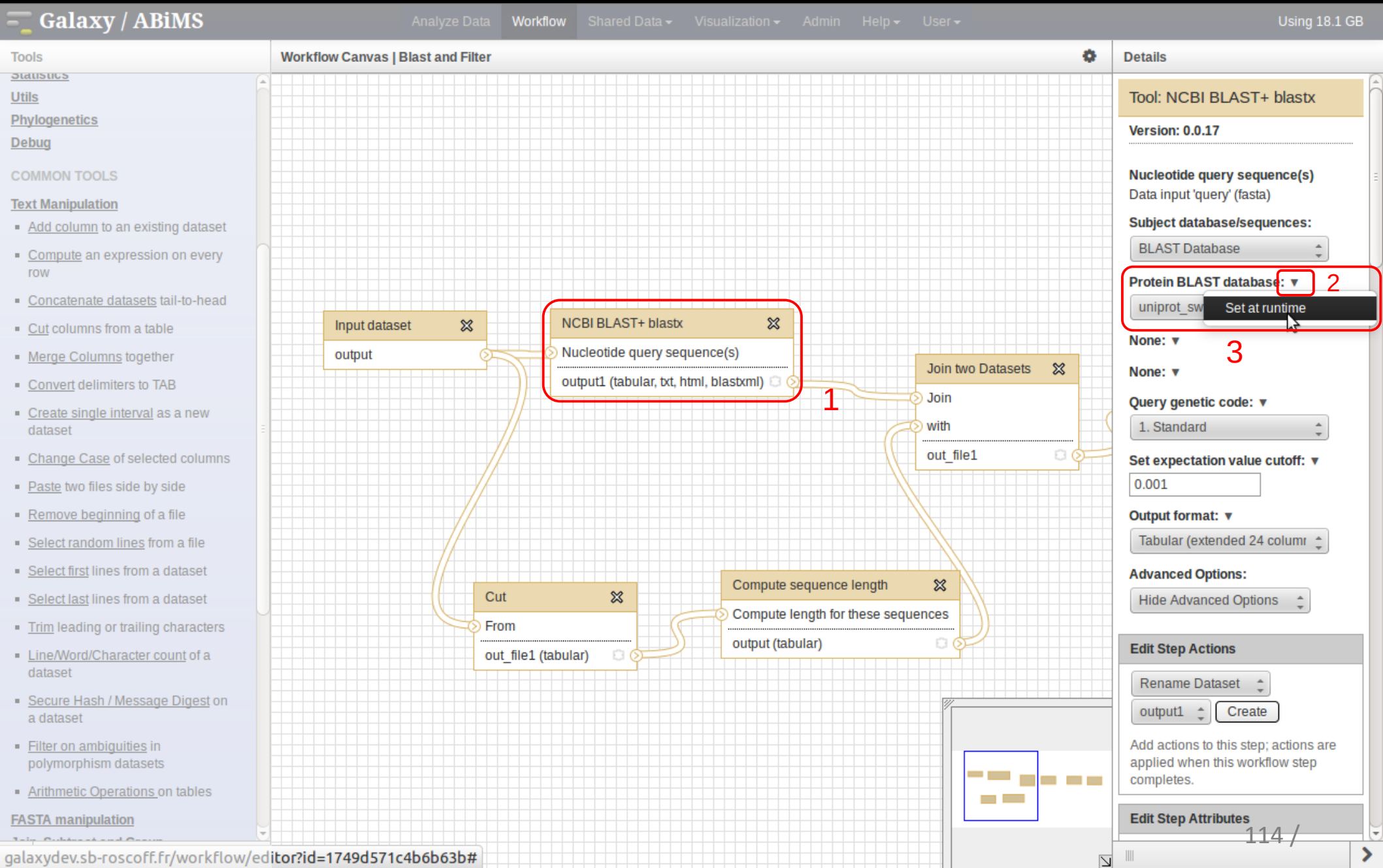
Edition d'un workflow :

- Ajout d'un nouvel outil dans le workflow

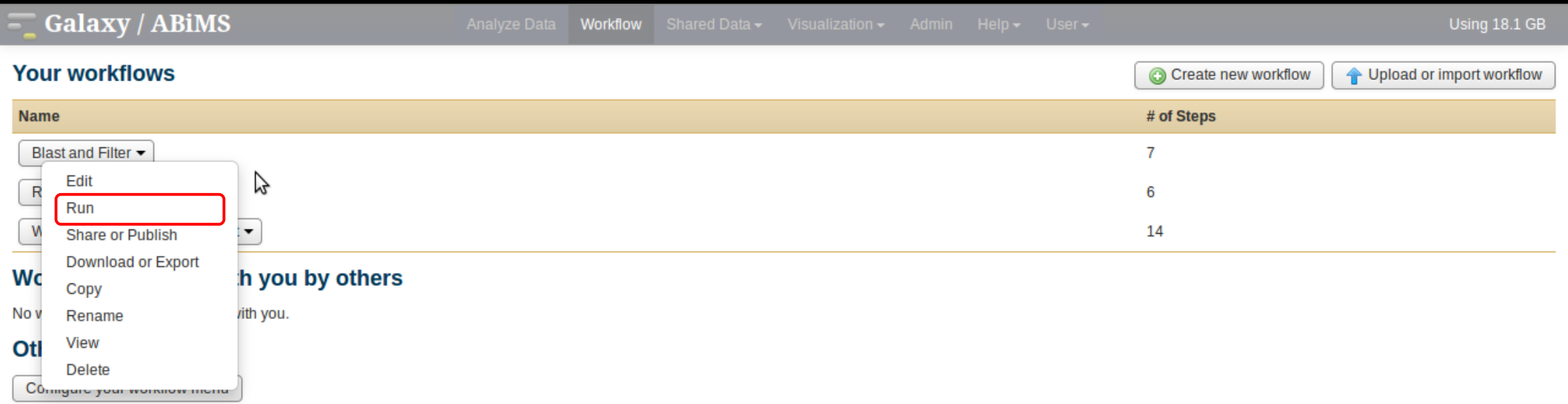


Edition d'un workflow :

- Rendre paramétrable un paramètre au moment de l'exécution

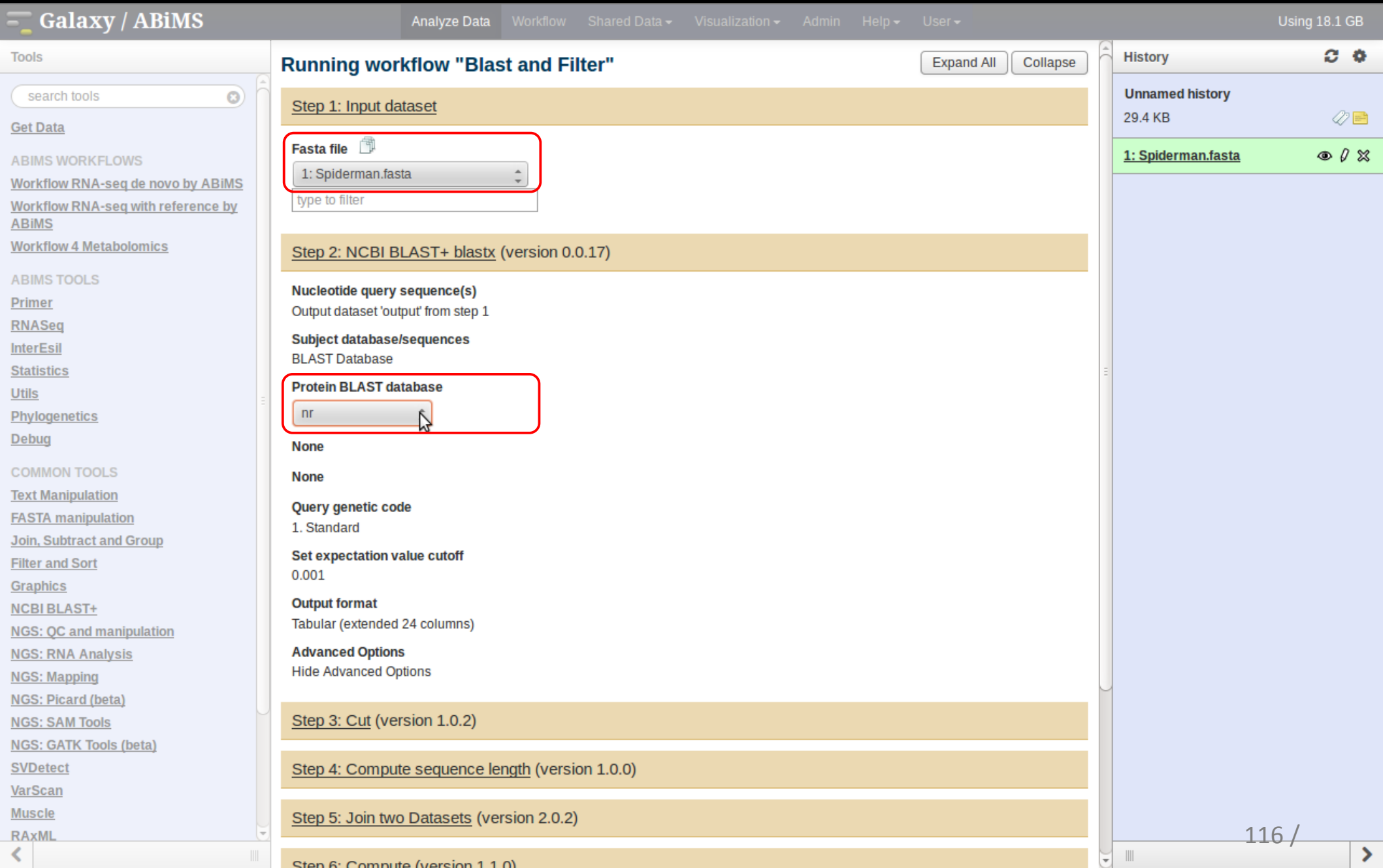


Lancer un workflow



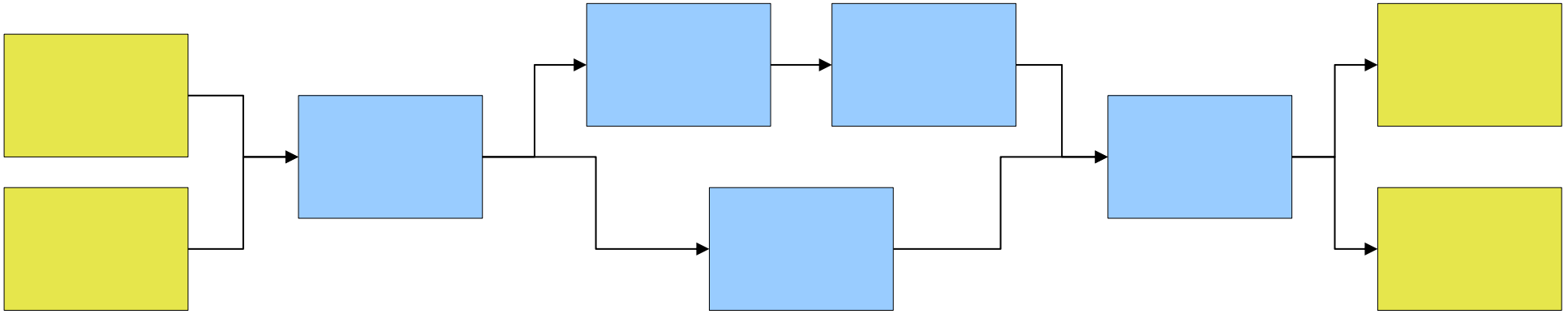
Lancer un workflow

- Paramètres par défaut
- Paramètres à renseigner

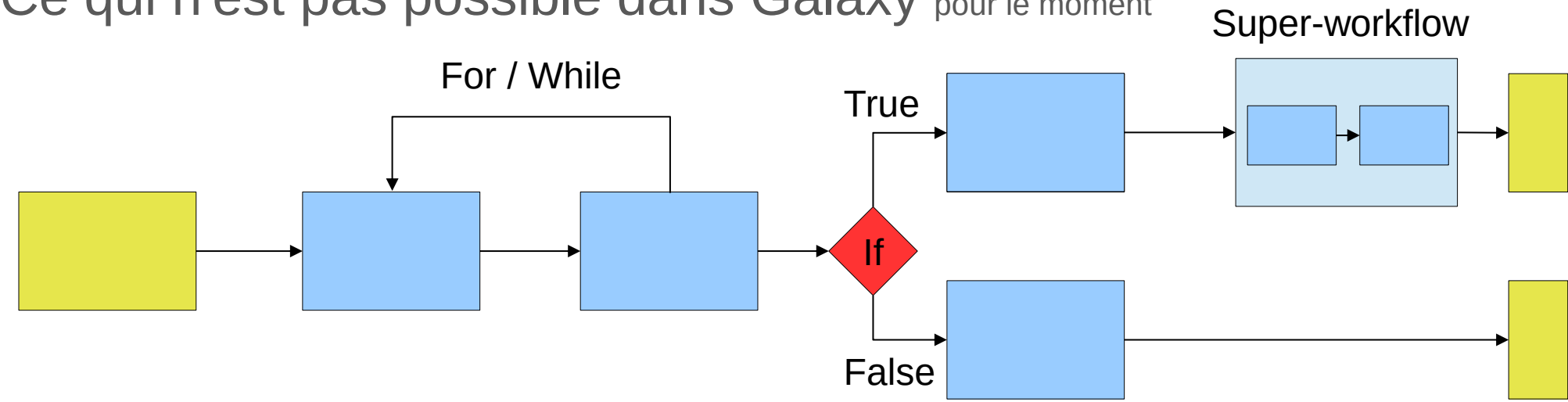


Workflow

- Ce qui est possible dans Galaxy



- Ce qui n'est pas possible dans Galaxy pour le moment



PARTAGE

Partage

« biologiste » ↔ « biologiste »

- Partage de datasets
  - Historique en entier
    - Avec ou sans workflow associé (Extract workflow)
  - Sous ensembles

Partage

bioanalyste → « biologiste »

- Partage de workflows
  - Paramètres pré-configués
  - Paramètres à configurer (Set at runtime)
  - => Suivant le niveau de l'utilisateur
- Partage de pages

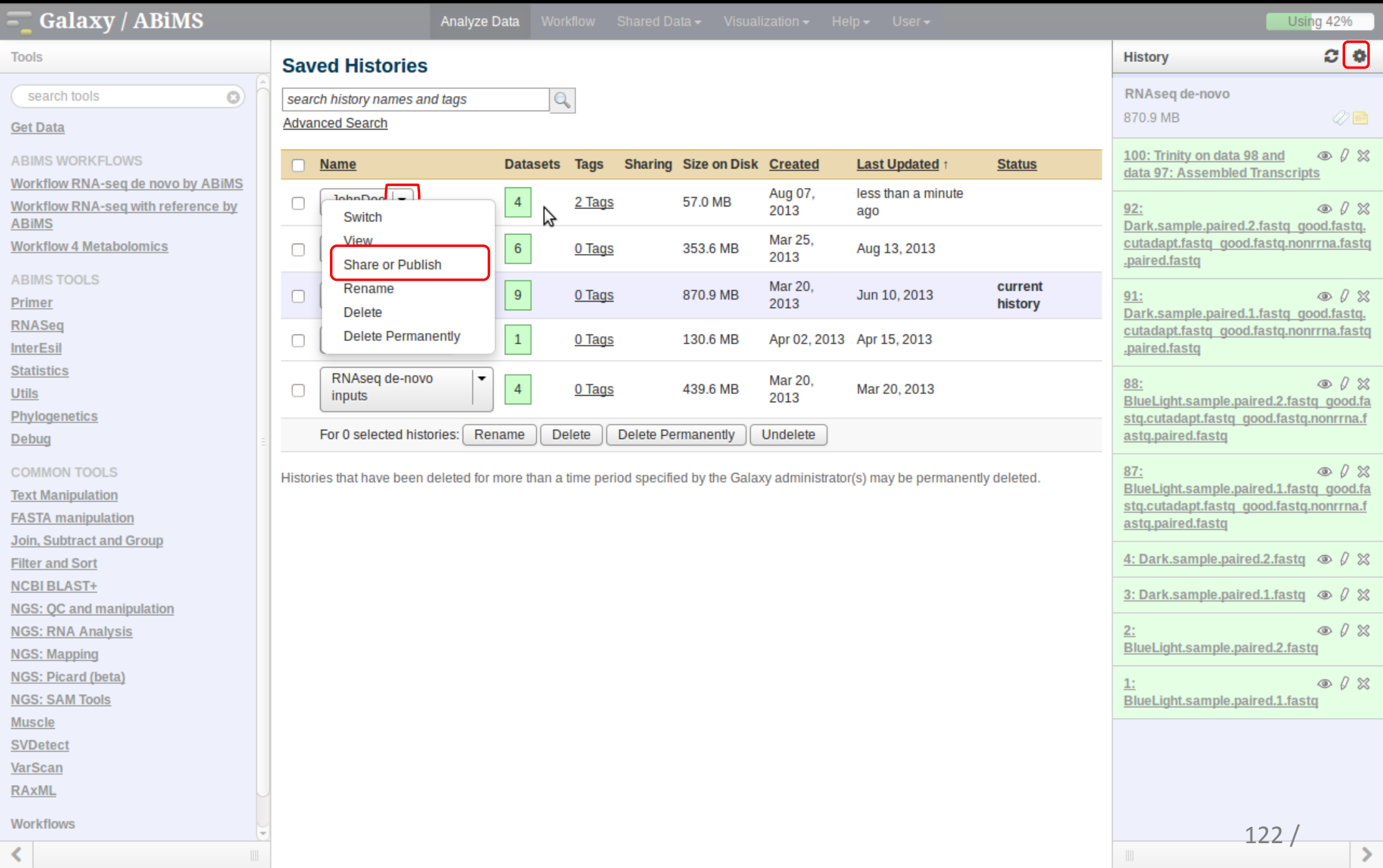
  (voir plus loin)

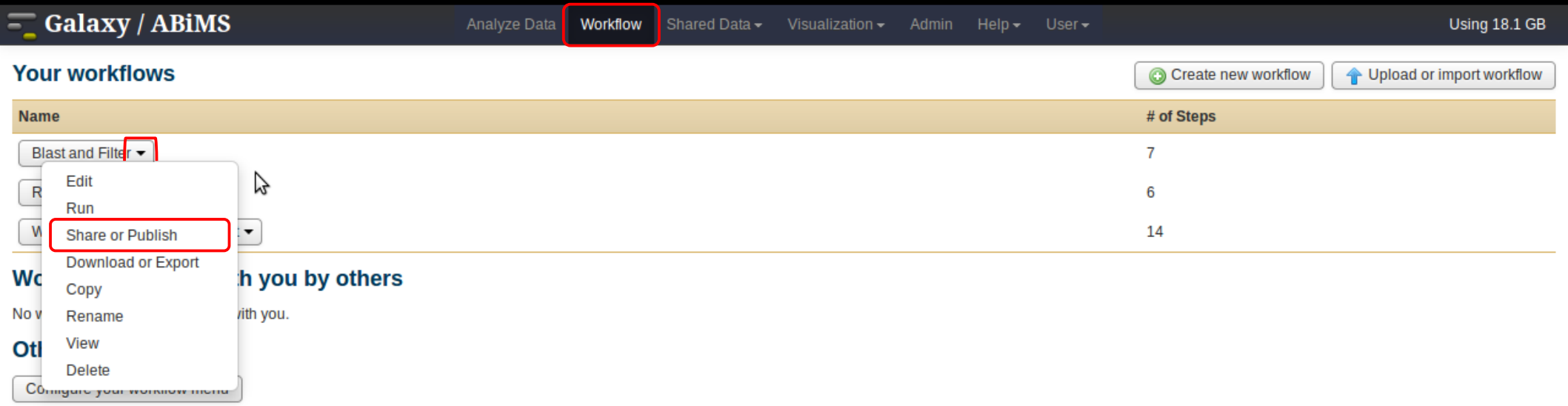
Partage

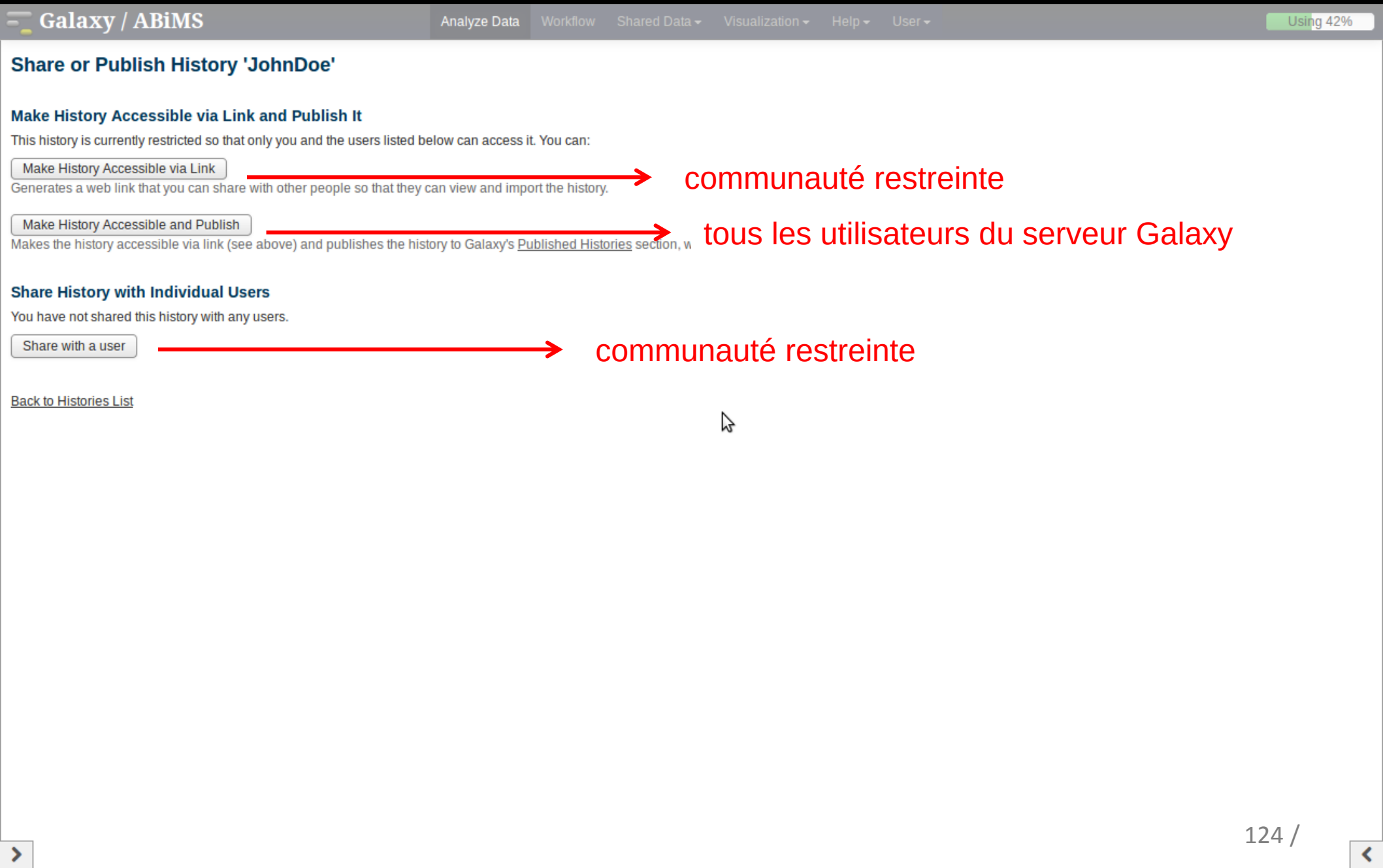
bioinformaticien → bioinformaticien

- Partage de descriptions d'outils et/ou de scripts

Toolshed Galaxy

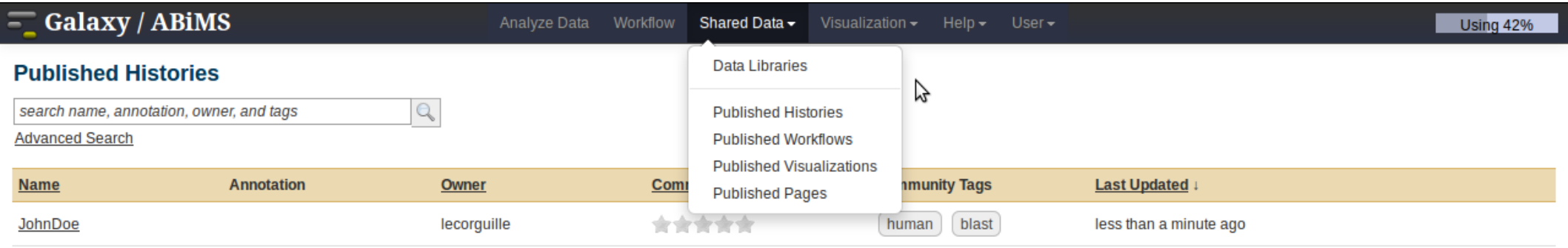

Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 42%
Tools
search tools
Get Data
ABIMS WORKFLOWS
Workflow RNA-seq de novo by ABiMS
Workflow RNA-seq with reference by ABiMS
Workflow 4 Metabolomics
ABIMS TOOLS
Primer
RNASeq
InterEsil
Statistics
Utils
Phylogenetics
Debug
COMMON TOOLS
Text Manipulation
FASTA manipulation
Join, Subtract and Group
Filter and Sort
NCBI BLAST+
NGS: QC and manipulation
NGS: RNA Analysis
NGS: Mapping
NGS: Picard (beta)
NGS: SAM Tools
Muscle
SVDetect
VarScan
RAxML
Workflows
Saved Histories
search history names and tags
Advanced Search
Name
Datasets
Tags
Sharing
Size on Disk
Created
Last Updated
Status
Switch
View
Share or Publish
Rename
Delete
Delete Permanently
4
2 Tags
57.0 MB
Aug 07, 2013
less than a minute ago
6
0 Tags
353.6 MB
Mar 25, 2013
Aug 13, 2013
9
0 Tags
870.9 MB
Mar 20, 2013
Jun 10, 2013
current history
1
0 Tags
130.6 MB
Apr 02, 2013
Apr 15, 2013
RNAseq de-novo inputs
4
0 Tags
439.6 MB
Mar 20, 2013
Mar 20, 2013
For 0 selected histories:
Rename
Delete
Delete Permanently
Undelete
Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.
History
RNAseq de-novo
870.9 MB
100: Trinity on data 98 and data 97: Assembled Transcripts
92: Dark.sample.paired.2.fastq_good.fastq.cutadapt.fastq_good.fastq.nonrrna.fastq.paired.fastq
91: Dark.sample.paired.1.fastq_good.fastq.cutadapt.fastq_good.fastq.nonrrna.fastq.paired.fastq
88: BlueLight.sample.paired.2.fastq_good.fastq.cutadapt.fastq_good.fastq.nonrrna.fastq.paired.fastq
87: BlueLight.sample.paired.1.fastq_good.fastq.cutadapt.fastq_good.fastq.nonrrna.fastq.paired.fastq
4: Dark.sample.paired.2.fastq
3: Dark.sample.paired.1.fastq
2: BlueLight.sample.paired.2.fastq
1: BlueLight.sample.paired.1.fastq


Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 18.1 GB
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
Blast and Filter
7
6
14
Edit
Run
Share or Publish
Download or Export
Copy
Rename
View
Delete

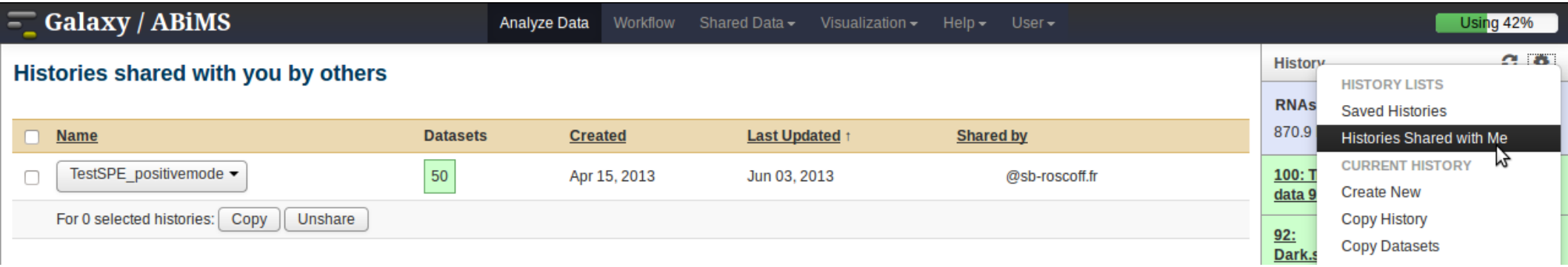

Galaxy / ABiMS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 42%
Share or Publish History 'JohnDoe'
Make History Accessible via Link and Publish It
This history is currently restricted so that only you and the users listed below can access it. You can:
Make History Accessible via Link
Generates a web link that you can share with other people so that they can view and import the history.
communauté restreinte
Make History Accessible and Publish
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, w
tous les utilisateurs du serveur Galaxy
Share History with Individual Users
You have not shared this history with any users.
Share with a user
communauté restreinte
Back to Histories List

Partage

- Import d'un published *



- Import d'un shared history



Partage

Level 5

- Partage d'outils et de descriptions via le toolshed

Level 4

- Lancement des outils de manière autonome.
- Utilisation des options avancées.
- Utilisation de l'API Galaxy
- Propose des workflows aux collègues de niveau --

Level 3

- Lancement des outils de manière autonome.

- Utilisation d'un workflow plus ou moins préconfiguré

Level 2

- Utilisation d'un workflow préconfiguré

Level 1

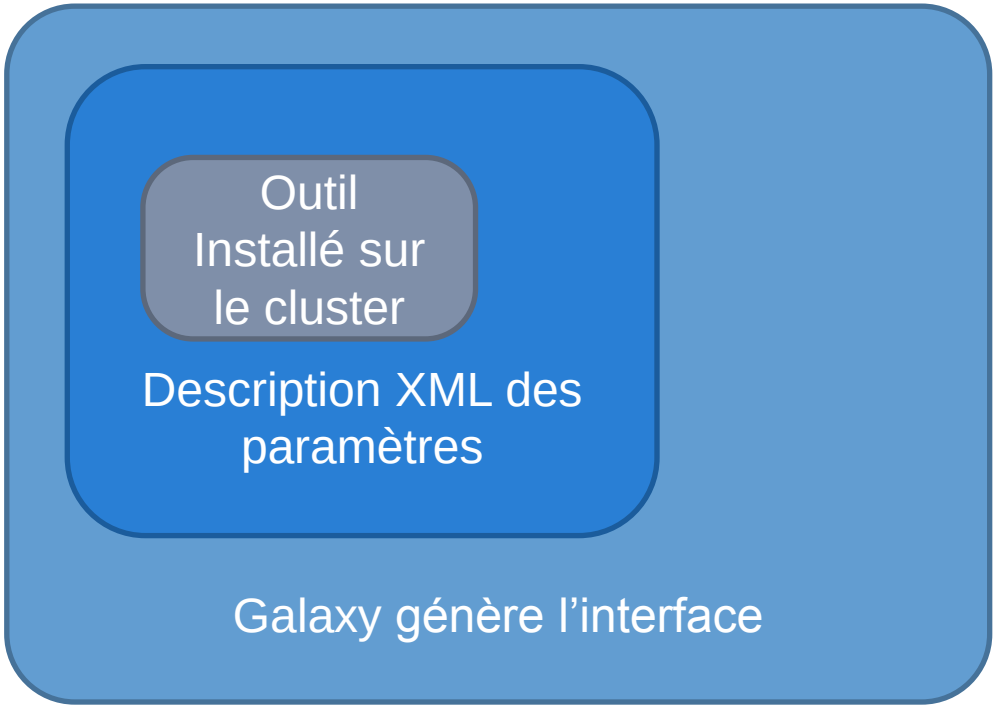
- Partage des données avec un collègue de niveau ++

BONUS

BONUS - MECANIQUE

Bonus

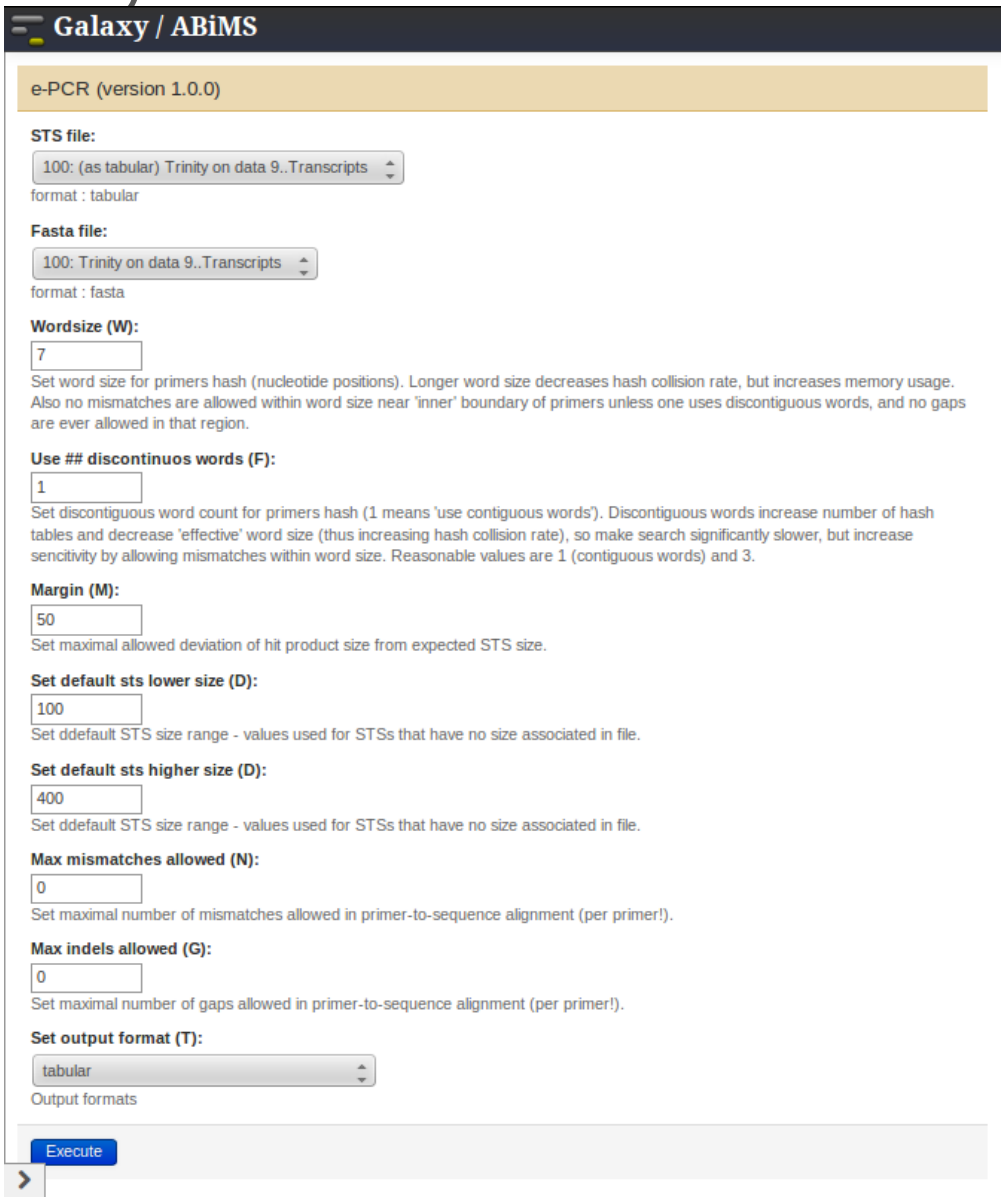
- Comment un outil arrive dans Galaxy ?



Bonus

- Comment un outil arrive dans Galaxy ?

```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
        -m ##    Margin (default 50)
        -w ##    Wordsize  (default 7)
        -n ##    Max mismatches allowed (default
        -g ##    Max indels allowed (default 0)
        -f ##    Use ## discontiguos words, slow
                        ##>1
        -o ##    Set output file
        -t ##    Set output format:
                 1 - classic, range (pos1..pos2)
                 2 - classic, midpoint
                 3 - tabular
                 4 - tabular with alignment in co
                        (slow)
        -d##-##  Set default size range
                        (default 100-350)
        -p +-    Turn hits postprocess on/off
        -v ##    Verbosity flags
        -a a|f   Use presize alignmens (only if
                        gaps>0), slow
                  a - Allways or f - as Fallback
        -x +-    Use 5'-end lowercase masking of
                        primers (default -)
        -u +-    Uppercase all primers (default -

[...]
```



Bonus

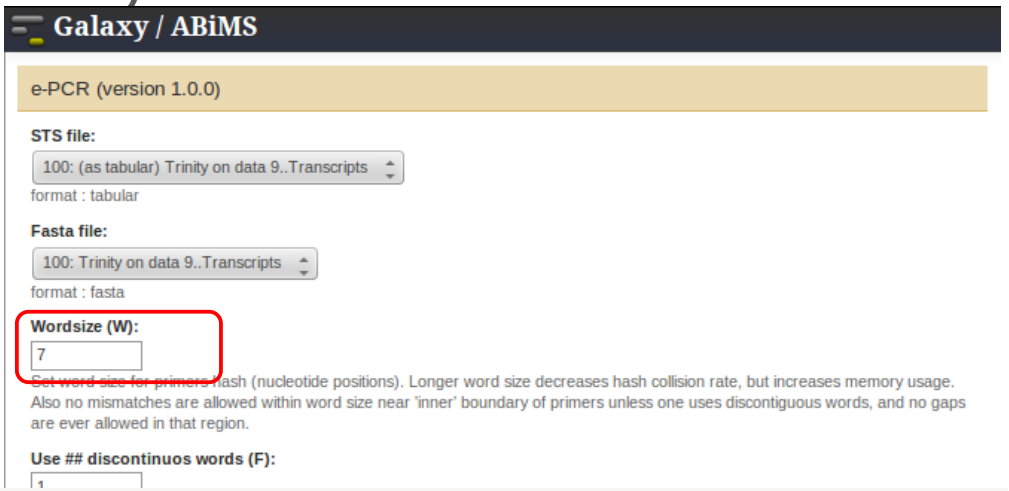
- Comment un outil arrive dans Galaxy ?

```
<tool id="abims_epcr" name="e-PCR">
    <!-- author : lecorguille@sb-roscoff.fr -->
    <!-- date : 11-05-12 -->
    <description>e-PCR parses stsfile in unists format, then reads nucleotide sequence data in FASTA format from files listed in commandline if any, or from stdin otherwise. For input sequences e-PCR finds matches and prints output in one of three formats.</description>
    <command>e-PCR -w $wordsize -f $wordcnt -m $margin -d$sts_size_lo-$sts_size_hi -n $max_mismatch -g $max_gap -t $output_format $infile_stsfile $infile_fasta > $output</command>

    <inputs>
        <param name="infile_stsfile" type="data" label="STS file" format="tabular" help="format : tabular" />
        <param name="infile_fasta" type="data" label="Fasta file" format="fasta" help="format : fasta" />
        <param name="wordsize" type="integer" label="Wordsize (W)" value="7" help="Set word size for primers hash (nucleotide positions). Longer word size decreases hash collision rate, but increases memory usage. Also no mismatches are allowed within word size near 'inner' boundary of primers unless one uses discontiguous words, and no gaps are ever allowed in that region." />
        <param name="wordcnt" type="integer" label="Use ## discontinuos words (F)" value="1" help="Set discontiguous word count for primers hash (1 means 'use contiguous words'). Discontiguous words increase number of hash tables and decrease 'effective' word size (thus increasing hash collision rate), so make search significantly slower, but increase sencitivity by allowing mismatches within word size. Reasonable values are 1 (contiguous words) and 3." />
        <param name="margin" type="integer" label="Margin (M)" value="50" help="Set maximal allowed deviation of hit product size from expected STS size." />
        <param name="sts_size_lo" type="integer" label="Set default sts lower size (D)" value="100" help="Set ddefault STS size range - values used for STSs that have no size associated in file." />
        <param name="sts_size_hi" type="integer" label="Set default sts higher size (D)" value="400" help="Set ddefault STS size range - values used for STSs that have no size associated in file." />
        <param name="max_mismatch" type="integer" label="Max mismatches allowed (N)" value="0" help="Set maximal number of mismatches allowed in primer-to-sequence alignment (per primer!)." />
        <param name="max_gap" type="integer" label="Max indels allowed (G)" value="0" help="Set maximal number of gaps allowed in primer-to-sequence alignment (per primer!)." />
        <param name="output_format" type="select" help="Output formats">
            <label>Set output format (T)</label>
            <option value="1">classic, range (pos1..pos2)</option>
            <option value="2">classic, midpoint</option>
            <option value="3" selected="true">tabular</option>
            <option value="4">tabular with alignment in comments (slow)</option>
        </param>
    </inputs>

    <outputs>
        <data name="output" format="tabular" />
    </outputs>

    <help>
Full documentation: ftp://ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR/README.txt
```

Bonus

- Comment un outil arrive dans Galaxy ?

```
[lecorguille@n0 ~]$ e-PCR --help
e-PCR: invalid option -- -
usage: [-hV] [posix-options] stsfile [fasta ...]
[compat-options]
where posix-options are:
        -m ##     Margin (default 50)
        -w ##     Wordsize  (default 7)
        -n ##     Max mismatches allowed (default
        -g ##     Max indels allowed (default 0)
        -f ##     Use ## discontiguos words, slow
                          ##>1
        -o ##     Set output file
```




```
<tool id="abims_epcr" name="e-PCR">
    <!-- author : lecorguille@sb-roscoff.fr -->
    <!-- date : 11-05-12 -->
    <description>e-PCR parses stsfile in unists format, then reads nucleotide sequence data in FASTA format from files listed in commandline if
    any, or from stdin otherwise. For input sequences e-PCR finds matches and prints output in one of three formats.</description>

    <command>e-PCR -w $wordsize -f $wordcnt -m $margin -d$sts_size_lo-$sts_size_hi -n $max_mismatch -g $max_gap -t $output_format $infile_stsfile
    $infile_fasta > $output</command>

    <inputs>
        <param name="infile_stsfile" type="data" label="STS file" format="tabular" help="format : tabular" />
        <param name="infile_fasta" type="data" label="Fasta file" format="fasta" help="format : fasta" />
        <param name="wordsize" type="integer" label="Wordsize (W)" value="7" help="Set word size for primers hash (nucleotide positions).
        Longer word size decreases hash collision rate, but increases memory usage. Also no mismatches are allowed within word size near
        'inner' boundary of primers unless one uses discontiguous words, and no gaps are ever allowed in that region." />
        <param name="wordcnt" type="integer" label="Use ## discontinuos words (F)" value="1" help="Set discontiguous word count for primers
        hash (1 means 'use contiguous words'). Discontiguous words increase number of hash tables and decrease 'effective' word size (thus
        increasing hash collision rate), so make search significantly slower, but increase sencitivity by allowing mismatches within word
        size. Reasonable values are 1 (contiguous words) and 3." />
        <param name="margin" type="integer" label="Margin (M)" value="50" help="Set maximal allowed deviation of hit product size from
        expected STS size." />
        <param name="sts_size_lo" type="integer" label="Set default sts lower size (D)" value="100" help="Set ddefault STS size range - values
        used for STSs that have no size associated in file." />
        <param name="sts_size_hi" type="integer" label="Set default sts higher size (D)" value="400" help="Set ddefault STS size range -
        values used for STSs that have no size associated in file." />
        <param name="max_mismatch" type="integer" label="Max mismatches allowed (N)" value="0" help="Set maximal number of mismatches allowed
        in primer-to-sequence alignment (per primer!)." />
```

CONCLUSION

Conclusion

- Comment un outil arrive dans Galaxy ?

  - Tooshed
  - Description maison

  - Une seul adresse :

    => support.abims@sb-roscoff.fr