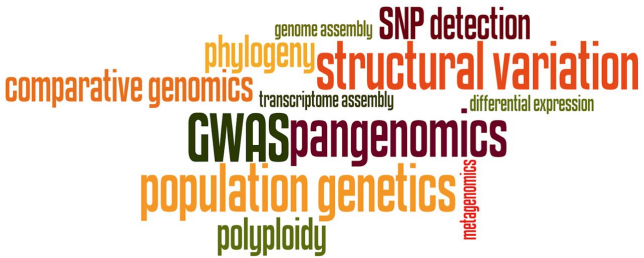
SouthGreen bioinformatics platform

Session de formation 2019

SouthGreen bioinformatics platform

Bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants and their pathogens




Rice

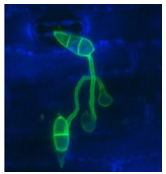

Banana


Palm


Sorghum


Coffee


Cassava


Magnaporthe

www.southgreen.fr

SouthGreen bioinformatics platform


Charte

diade

Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

IPME

Comte Aurore
Dereeper Alexis

diade IPME

Orjuela-Bouniol Julie

agap

Bocs Stephanie
De Lamotte Frédéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne

Bioversity International

Rouard Mathieu
Guignon Valentin
Catherine Breton

UMR BGPI

Mahé Frédéric
Ravel Sébastien

Intertryp

Sempere Guilhem

Bioversity International

cirad

IRD Institut de Recherche pour le Développement FRANCE

INRA SCIENCE & IMPACT

SouthGreen bioinformatics platform

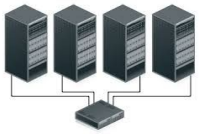
Workflow manager

TOGGLe
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

HPC and trainings….



IRD Institut de Recherche pour le Développement

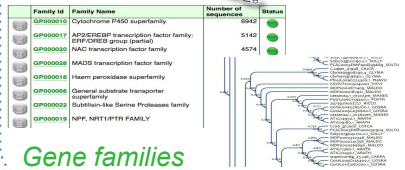
cirad

Genome Hubs & Information System

Gigwa

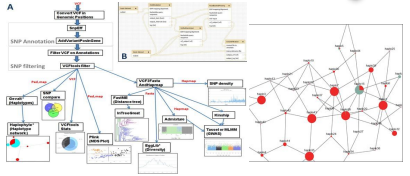
SNPs and Indels

GreenPhyl



Gene families

SNiPlay



https://github.com/SouthGreenPlatform

@green_bioinfo

The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

SouthGreen bioinformatics platform

ABiMS — Erwan Corre

IPME — Marie Simonin, Sébastien Cunnac

ASTRE (Animal · Santé, Territoires · Risques · Écosystèmes) — Etienne Loire, Julie Reveillaud

UMR QualiSud — Florentin Constancias

diade — Valentin Klein

MIVEGEC (Maladies Infectieuses et Vecteurs : Écologie, Génétique, Évolution et Contrôle) — Valérie Noël

TransVIH MI — Emmanuelle Beyne

i-Trop bioinformatics — Valérie Noël

And more collaborators !

South Green bioinformatics platform

18-19/03	Guide de survie à Linux - IRD
21/03	Initiation à l'utilisation du cluster CIRAD – CIRAD
22/03	Initiation à l'utilisation du cluster itrop - IRD
15-16/04	Initiation au gestionnaires de workflow SG & Gigwa – IRD
18-19/04	Guide du Jedi en Linux & bash - CIRAD
13–16/05	Python - IRD
17/05	Initiation aux analyses de données transcriptomiques – IRD
21/05	Utilisation avancée du cluster IRD – IRD
23-24/05	Initiation aux analyses de données métagénomiques – IRD
6/06	Manipulation de données et figures sous R – CIRAD
26-28/06	Assemblage et annotation de transcriptomes - IRD

Modules de formation 2019

- Toutes nos formations :
  https://southgreenplatform.github.io/trainings/

- Environnement de travail : **bioinfo-inter.ird.fr:8080**

Workflow Manager

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Objectifs de la formation

objectifs:

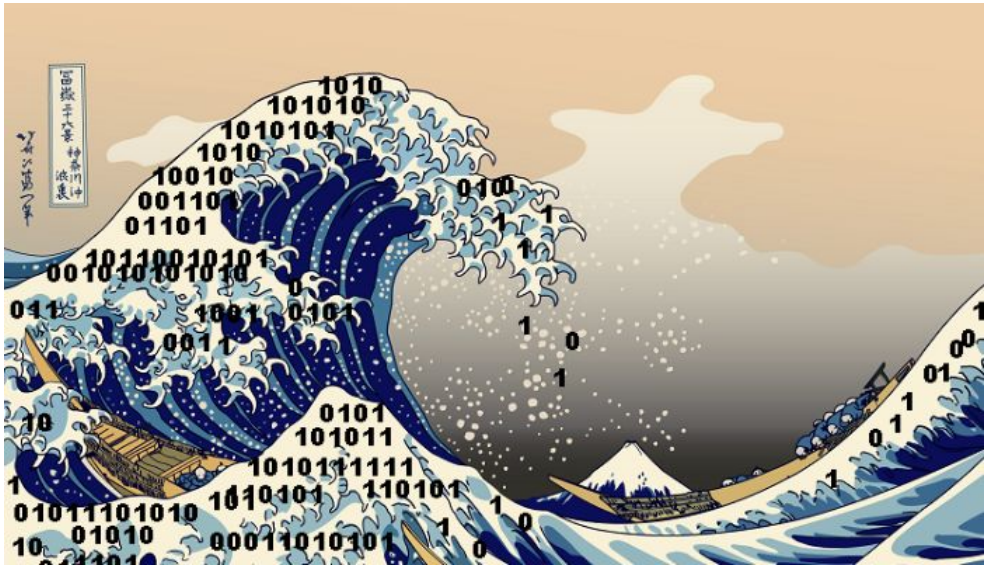
Utiliser les gestionnaires de Workflow de South Green afin de construire de manière automatique vos propres pipelines.

Applications

Tout savoir sur les 2 principaux gestionnaires de workflow

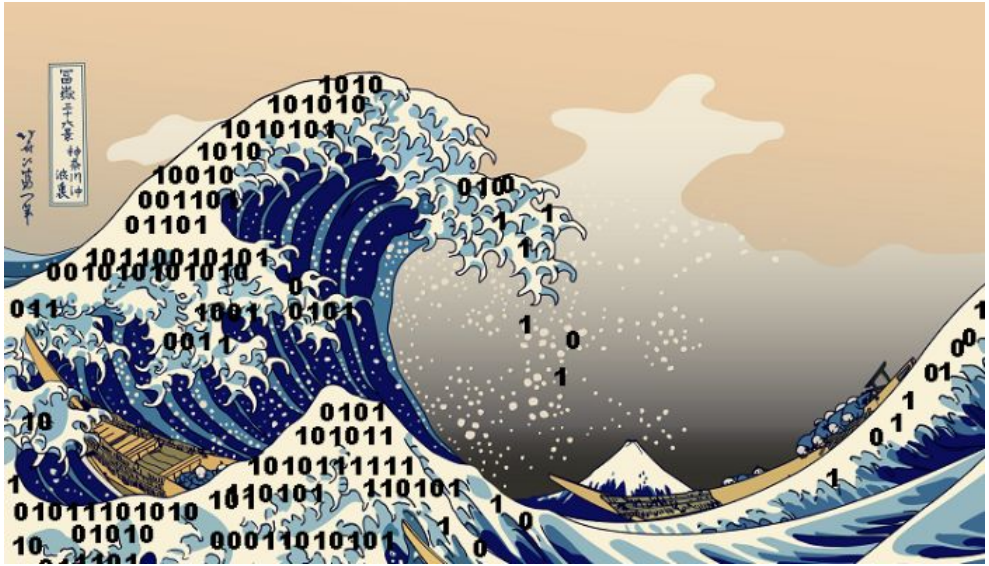
- Utiliser les outils
- Construire son propre workflow
- Pratiquer sur un même cas d'utilisation : Appel de SNPs à partir de reads Illumina de 3 échantillons

Pourquoi utiliser un gestionnaire de workflow?



The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

Pourquoi utiliser un gestionnaire de workflow?



Créer son propre pipeline via une méthode facile et conviviale

Données brutes → Résultats Intermédiaires → Résultats Intermédiaires → Résultat Final

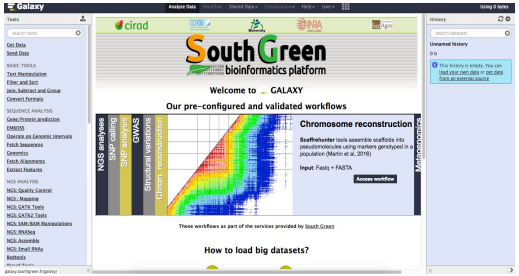
Pourquoi utiliser un gestionnaire de workflow?

- 3 solutions proposées par South Green bioinformatics platform

GUI tools

CLI tools

Galaxy



TOGGLE

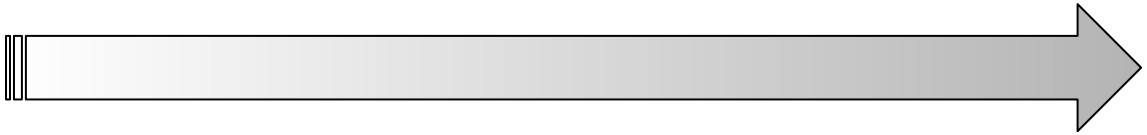
>_

Snakemake

>_

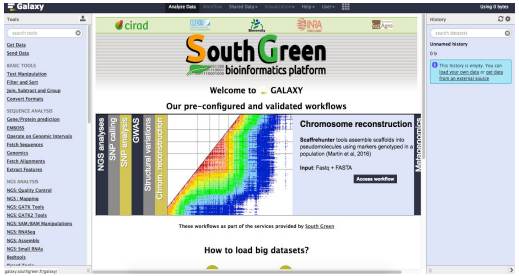
Pourquoi utiliser un gestionnaire de workflow?

GUI tools



CLI tools

Galaxy



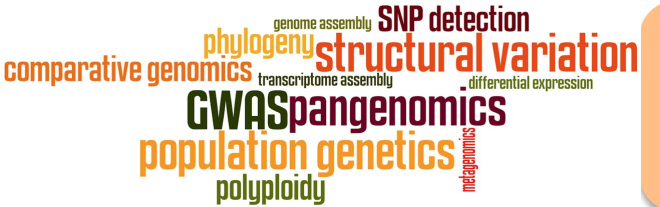
TOGGLE

>_

Snakemake

>_

Facilité d'utilisation
Bonne documentation



Facilité de développement

Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline et des données

Apporte un cadre robuste

✓ Vérifie le format des fichiers
Valide l'enchaînement des outils

\+ Automatisation de certaines étapes clefs
(ex : indexation de la référence)

Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline et des données

Apporte un cadre robuste

Reproductibilité & traçabilité

Sauvegarde des options, version des logiciels, partage des analyses

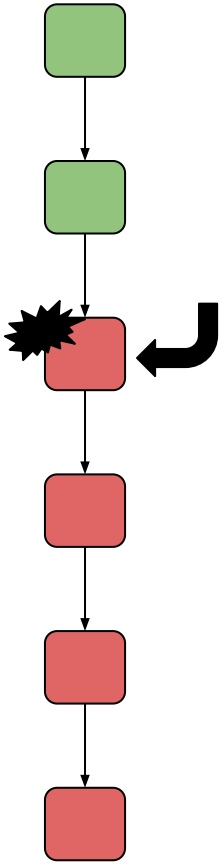
Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline et des données

Reproductibilité & traçabilité

Apporte un cadre robuste

Suivi des erreurs & reprise en cours



Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline et des données

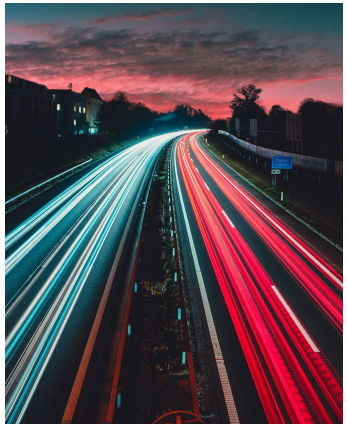
Apporte un cadre robuste

Reproductibilité & traçabilité

Analyse de gros jeu de données

Suivi des erreurs & reprise en cours

Pourquoi utiliser un gestionnaire de workflow?



Contrôle du pipeline et des données

Connection HPC Parallélisation

Apporte un cadre robuste

Reproductibilité & traçabilité

Analyse de gros jeu de données

Suivi des erreurs & reprise en cours

Pourquoi utiliser un gestionnaire de workflow?

	TOGGLE	Galaxy
Interface	Command line	GUI (Web interface)
Predefined Pipelines	SNP calling, RNASeq and WGS large scale ...	Metagenomics, RNASeq, SNP calling, post-analyses ...
Number of Samples	+++	++
Quota (related to infra)	Disk space “/data/projects”	IRD 100Go data Cirad 100Go => 300Go
Parallelization (related to infra conf)	IRD 300 cores Cirad 600 cores	IRD *16 cores / one node* Cirad *200 cores*
Number of tools available	++ (120)	++++ (5500 avail)
Post-analyses Graphical figures	Not yet	Yes

South Green bioinformatics platform

Introduction au gestionnaire de workflow:

Galaxy : Introduction

- Plateforme de fouille et de gestion de données
- Rendre la bioinformatique accessible sans compétence en programmation informatique

Instance IRD : http://bioinfo-inter.ird.fr:8080 → http://galaxy.ird.fr (bientôt)

Instance SouthGreen : http://galaxy.southgreen.fr/galaxy/

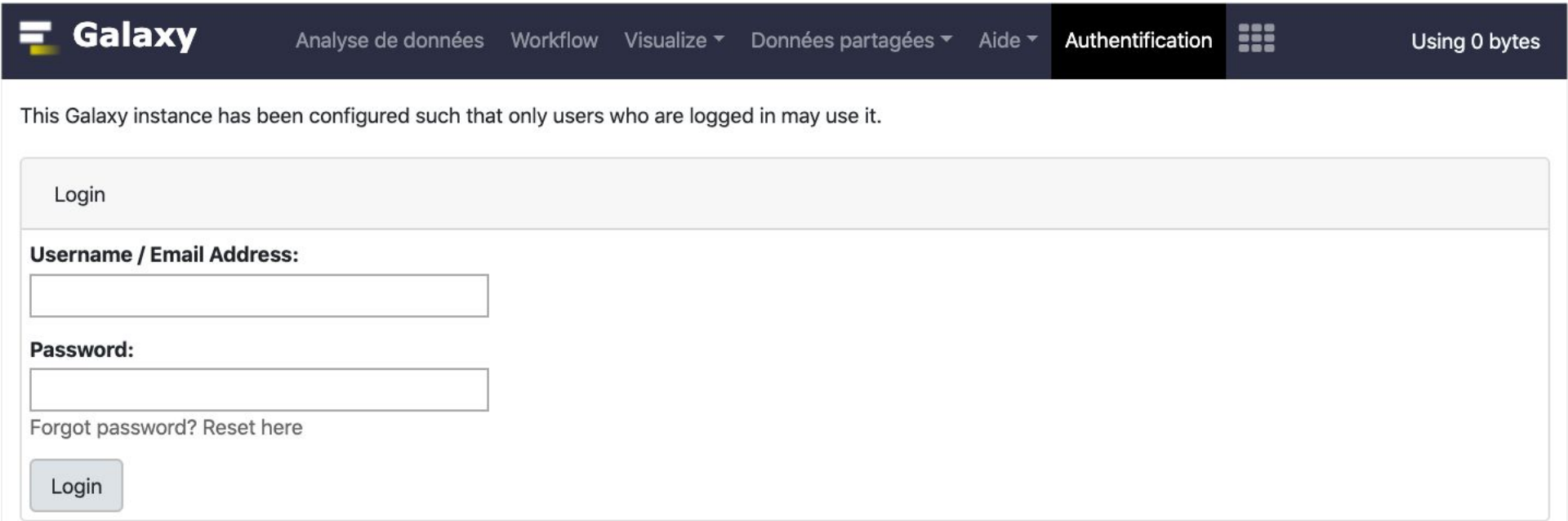
Galaxy : Connection

TP

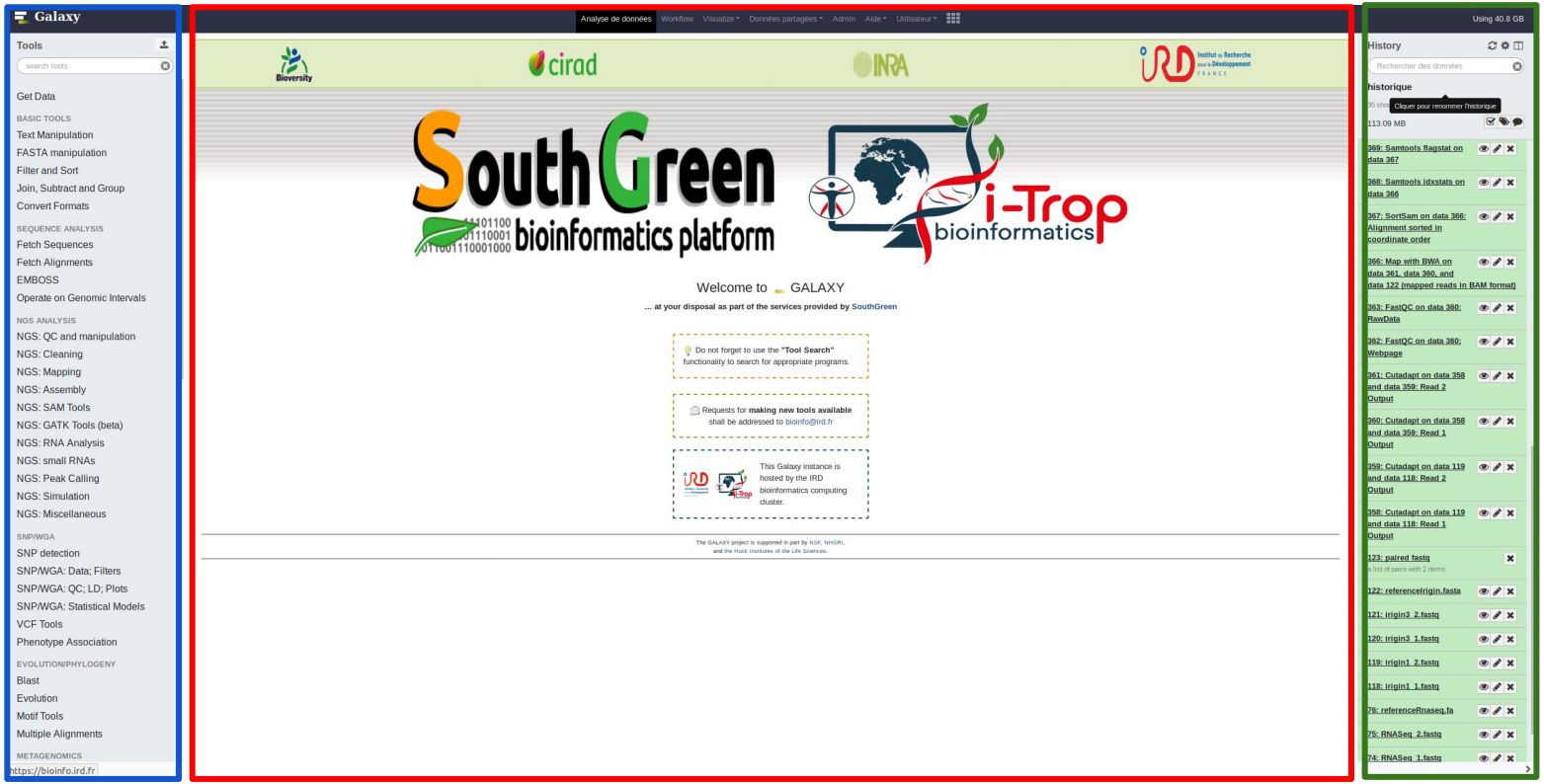
Connectez-vous sur la plateforme Galaxy IRD à l'adresse suivante :

http://bioinfo-inter.ird.fr:8080/

Utilisez pour aujourd'hui le compte formationN/formationN



Interface principale

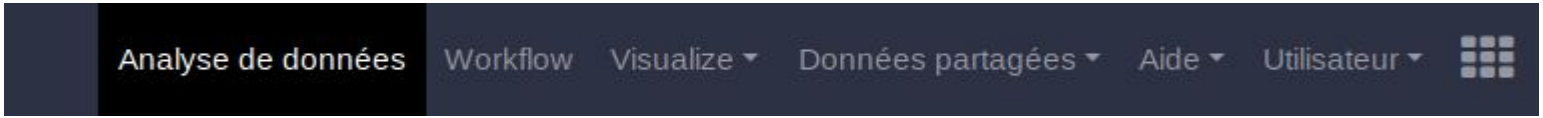


outils

Panel principal

historique

Panel supérieur



Lien	Utilisation
Analyse de données	Retour sur la page principale
Workflow	Accès aux workflows existant ou créé un nouveau workflow
Visualize	Environnement interactif de visualisation de données
Données partagées	Accès aux libraries, historiques et workflows publics ou partagés avec vous
Aide	Lien à l'aide de galaxy
Utilisateur	Préférence du compte et accès aux sauvegardes de l'utilisateur

Galaxy : Import de fichiers

→ **Il y a plusieurs solutions pour importer un fichier :**

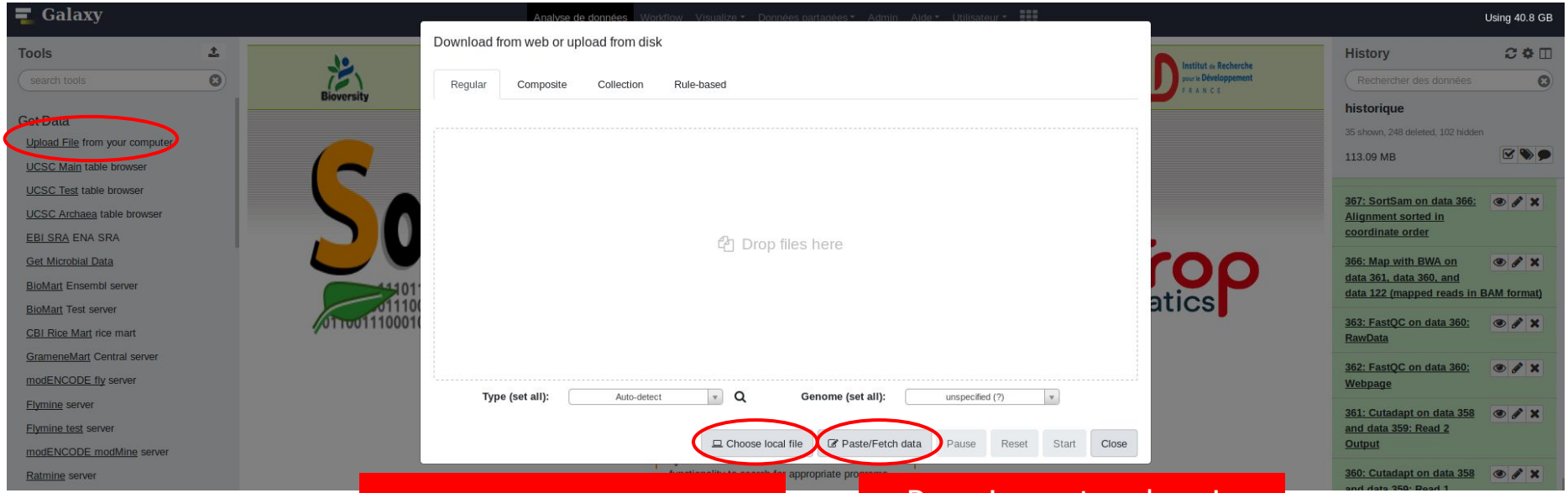
- *Importer un fichier **stocké localement** sur votre ordinateur en cliquant sur «choisissez un fichier »*
- *Importer un fichier à partir d'une **URL** en copiant l'adresse dans le cadre « URL/Text »*
- ***Copier** directement le texte du fichier dans le cadre « URL/Text »*
- *Importer un **fichier partagé** par une personne / public dans les "données partagées"*

→ **Pour les gros jeux de données**

- *Importer un fichier **déposé sur un serveur FTP***
- ***Créer un lien symbolique** vers un fichier sur un cluster (voir avec les admins)*

Galaxy : Import de fichiers

→ **Import depuis l'ordinateur ou un site externe**



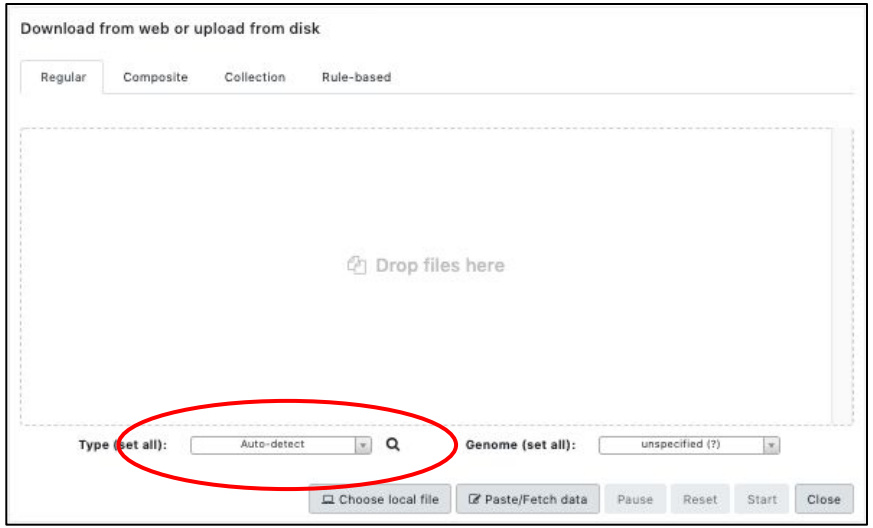
Pour importer depuis votre ordinateur

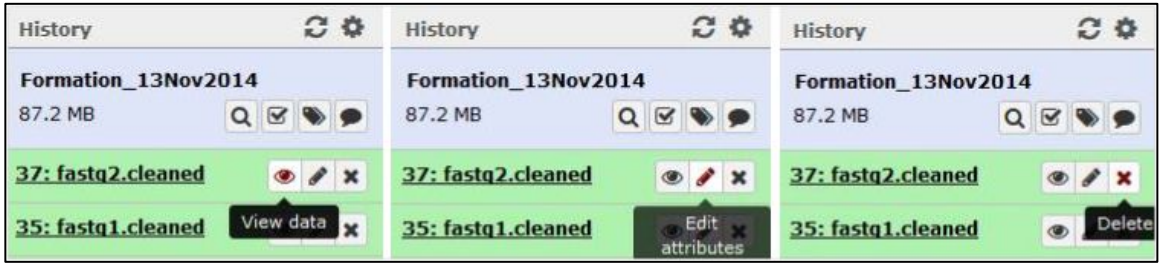
Pour importer depuis une autre page web, ou copier-coller

Galaxy : Import de fichiers

- Au chargement d'un fichier :
  - On peut choisir le type de fichier (txt, fasta, …)
  - Galaxy peut détecter le type automatiquement

- Pour changer le type d'un dataset après chargement:
  - *Edit Attributes → Datatype*





Galaxy : Import de fichiers

TP

→ Depuis la library partagée

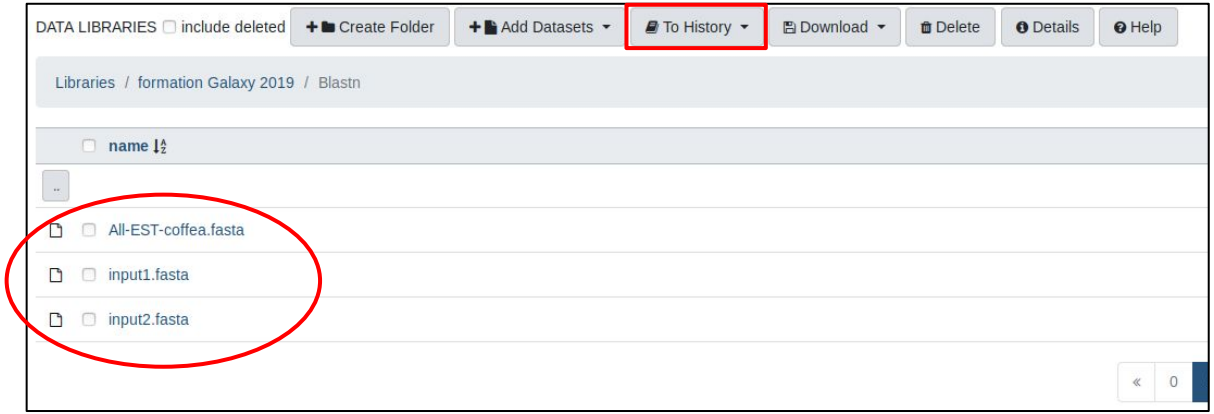
Accédez aux données partagées

(Données partagées → Bibliothèque de données → formation Galaxy 2019 → Blastn)

1) Cliquez sur la library
Formation Galaxy 2019
Blastn

2) Cochez les fichiers:
input1.fasta
input2.fasta
All-EST-coffea.fasta

3) Cliquez sur le bouton "To history" pour importer les données.
→ as Datasets



Suivi des imports sans l'historique:

Bleu : le job a été soumis

Jaune : le job est en cours de traitement

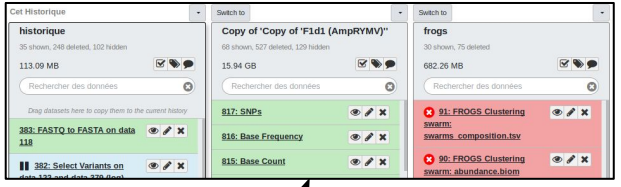
Vert : le job s'est terminé avec succès

Rouge : le job est en erreur

Galaxy : Historique

Vous pouvez avoir autant d'historiques que vous voulez et naviguer entre différents historiques.

- **1 historique = 1 analyse**
- Nommer les historique de façon reconnaissable



Saved Histories

search history names and tags

Advanced Search

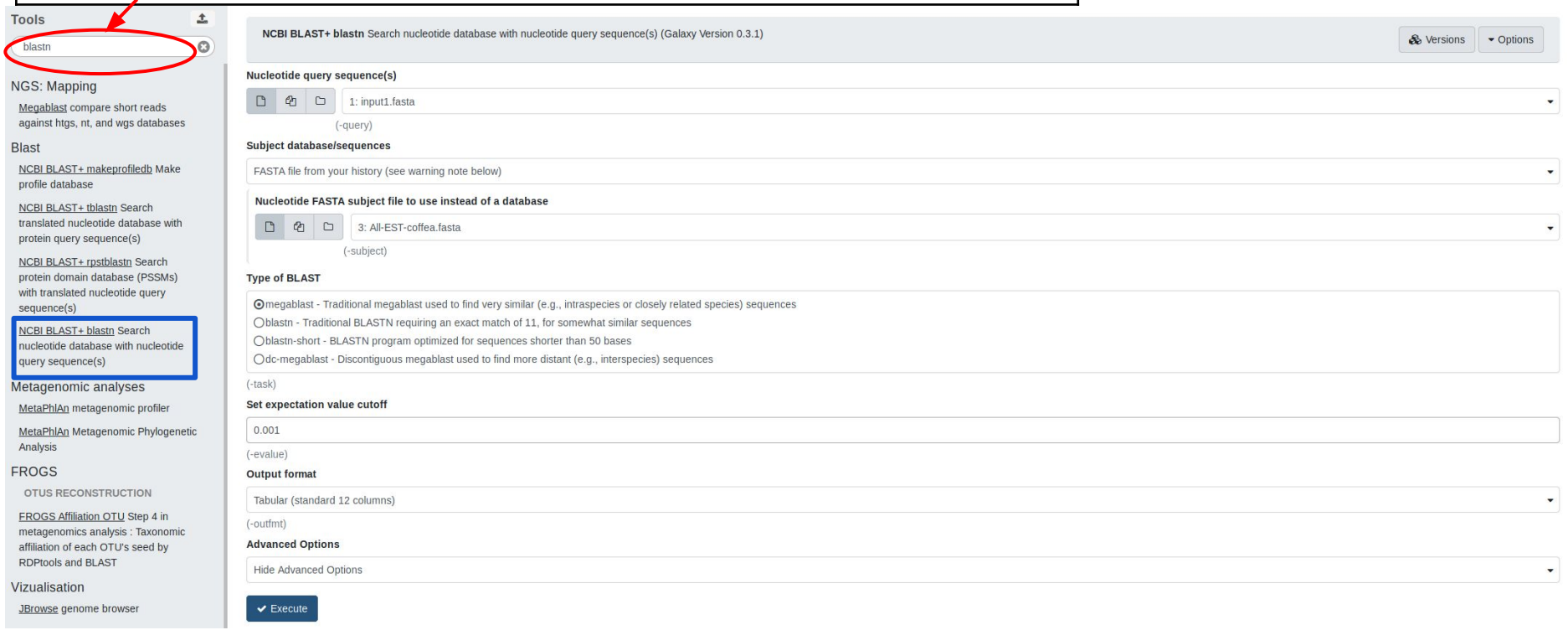
Name	Items	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated↑	Status
historique	383	23 2 248 102			113.1 MB	Mar 21, 2019	il y a 1 heure	current history
Copy of 'Copy of 'F1d1 (AmpRYMV)"	817	68 527 129			15.9 GB	Feb 21, 2019	il y a 2 jours	
frogs	105	21 9 75			682.3 MB	Sep 19, 2018	il y a 6 jours	
Unnamed history	8	2 6			73.2 MB	Mar 20, 2019	Mar 20, 2019	
RNASEQ	105	13 72 32			143.1 MB	Jan 15, 2019	Mar 19, 2019	
traceancestor	60	20 31 9			167.0 MB	Mar 04, 2019	Mar 13, 2019	
vana	849	42 216 138			18.3 GB	Mar 12, 2019	Mar 12, 2019	
KDEClassifier	105	4 82 31			291.6 MB	Mar 04, 2019	Mar 07, 2019	
benchmark calling variant	191	13 2 149 41		Shared, Accessible	14.6 MB	Jan 16, 2019	Feb 20, 2019	
Unnamed history	56	12 4 40			243.5 MB	Dec 13, 2017	Feb 06, 2019	
mapping	7	4 3 1			3.1 MB	Sep 26, 2018	Jan 03, 2019	
frogs 2	21	6 2 13			137.7 MB	Sep 19, 2018	Sep 21, 2018	

Galaxy : Outil

TP

Pour trouver facilement un outil vous pouvez taper son nom dans la case de recherche « search tools ».

Cherchez BLASTN



Galaxy : Outil

TP

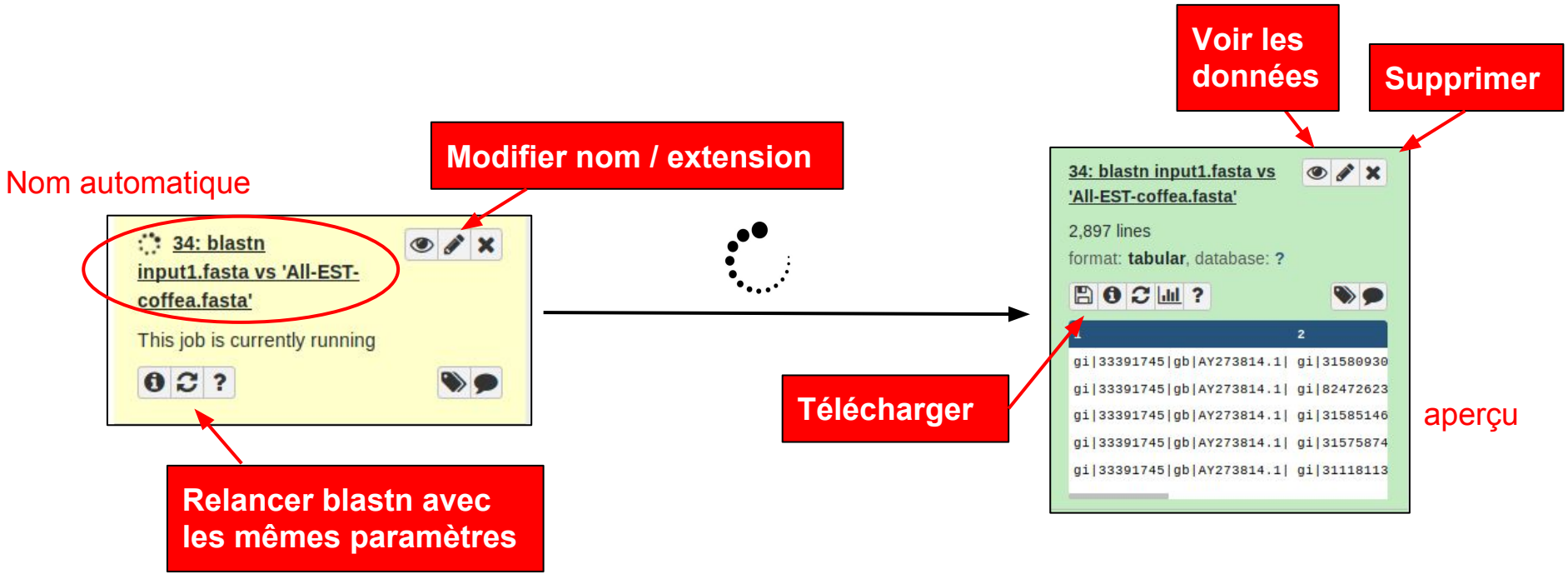
1) **Lancez BLASTN avec les paramètres suivants :**

- Query = input1.fasta
- Banque Fasta de l'historique = all-EST-coffea.fasta
- Output = fichier tabulé de 12 colonnes

Blast *(Basic local alignment Search tool) = permet de trouver des régions similaires entre 2 séquences de nucléotides (blastn) ou d'acides aminés*

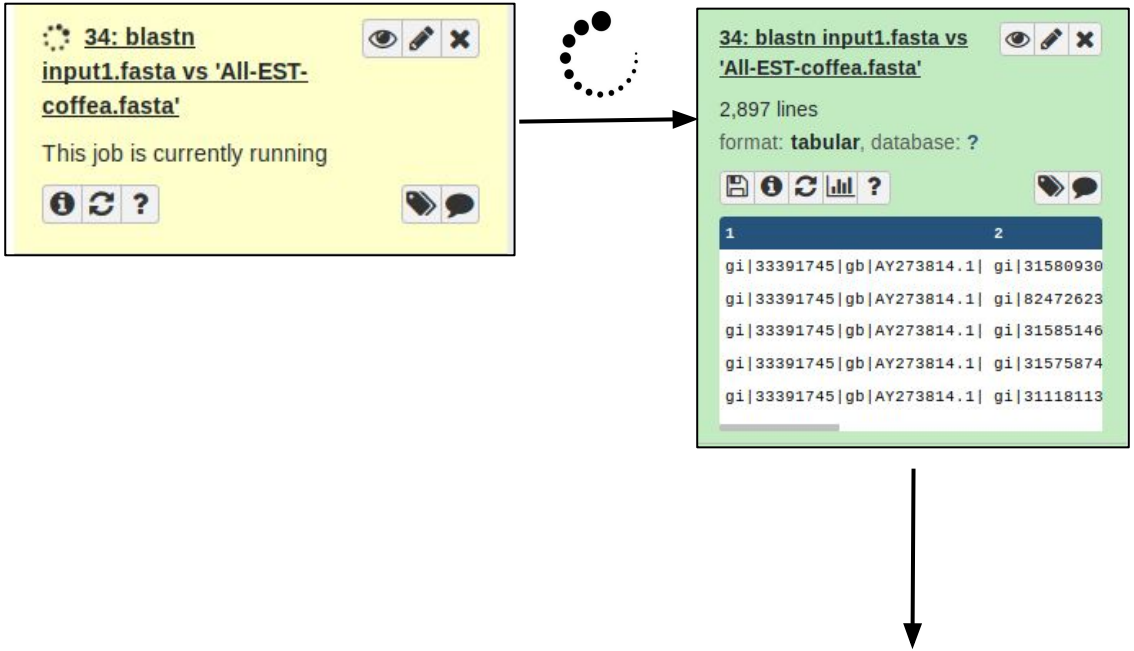
Galaxy : Outil

TP



Galaxy : Outil

TP



- Output tabular format (6 or 7):

1. query id
2. subject id
3. percent identity
4. alignment length
5. number of mismatches
6. number of gap openings
7. query start
8. query end
9. subject start
10. subject end
11. expect value
12. bit score

gi|33391745|gb|AY273814.1|	gi|315809301|gb|GW481759.1|GW481759	94.922	709	14	2	188	896	1	687	0.0	1128

Galaxy : Enchaîner les outils

TP

1) Lancez BLASTN pour obtenir un fichier tabulé

2) Lancez FILTER et ne gardez que les lignes ou le pourcentage d'identité est de plus de 98

3) Lancez CUT et ne gardez que les "subject ID" sur le fichier précédent

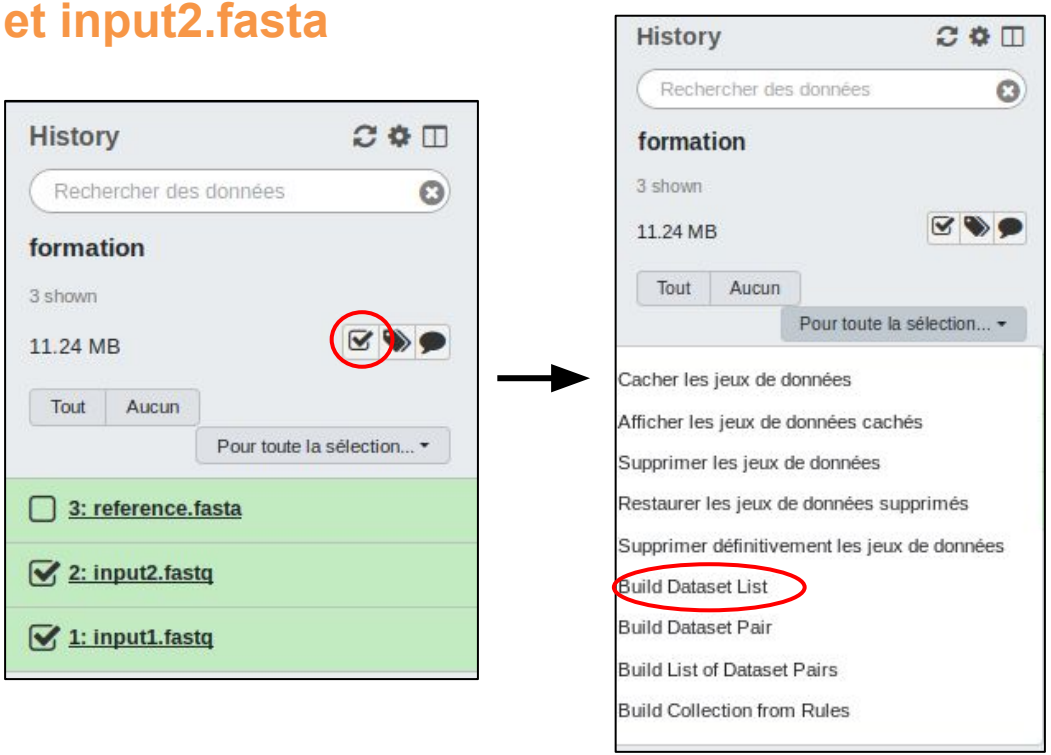
- Output tabular format (6 or 7):

1. query id
2. subject id
3. percent identity
4. alignment length
5. number of mismatches
6. number of gap openings
7. query start
8. query end
9. subject start
10. subject end
11. expect value
12. bit score

Galaxy : Collections

Collection = Permet d'effectuer une même analyse sur plusieurs échantillons

1) Créez une collection avec les deux jeux de séquences input1.fasta et input2.fasta



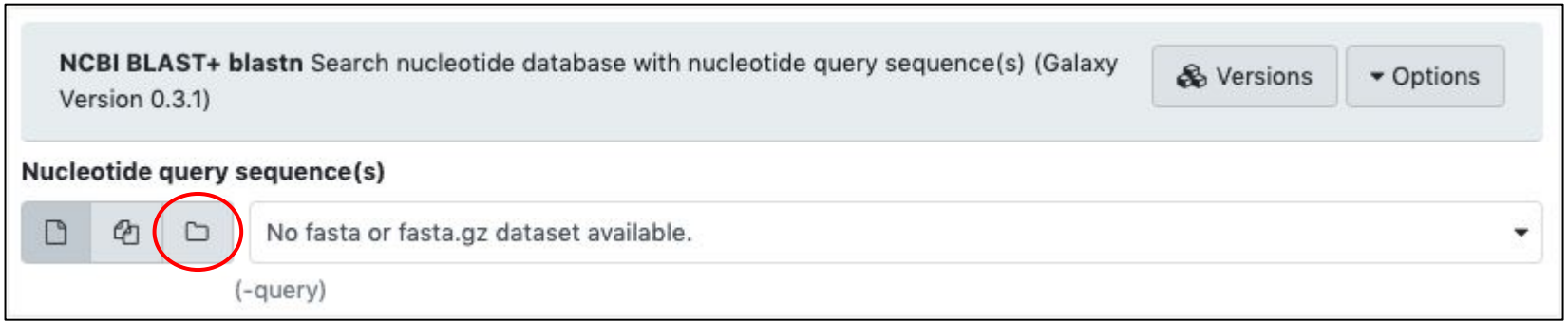
Galaxy : Collections

TP

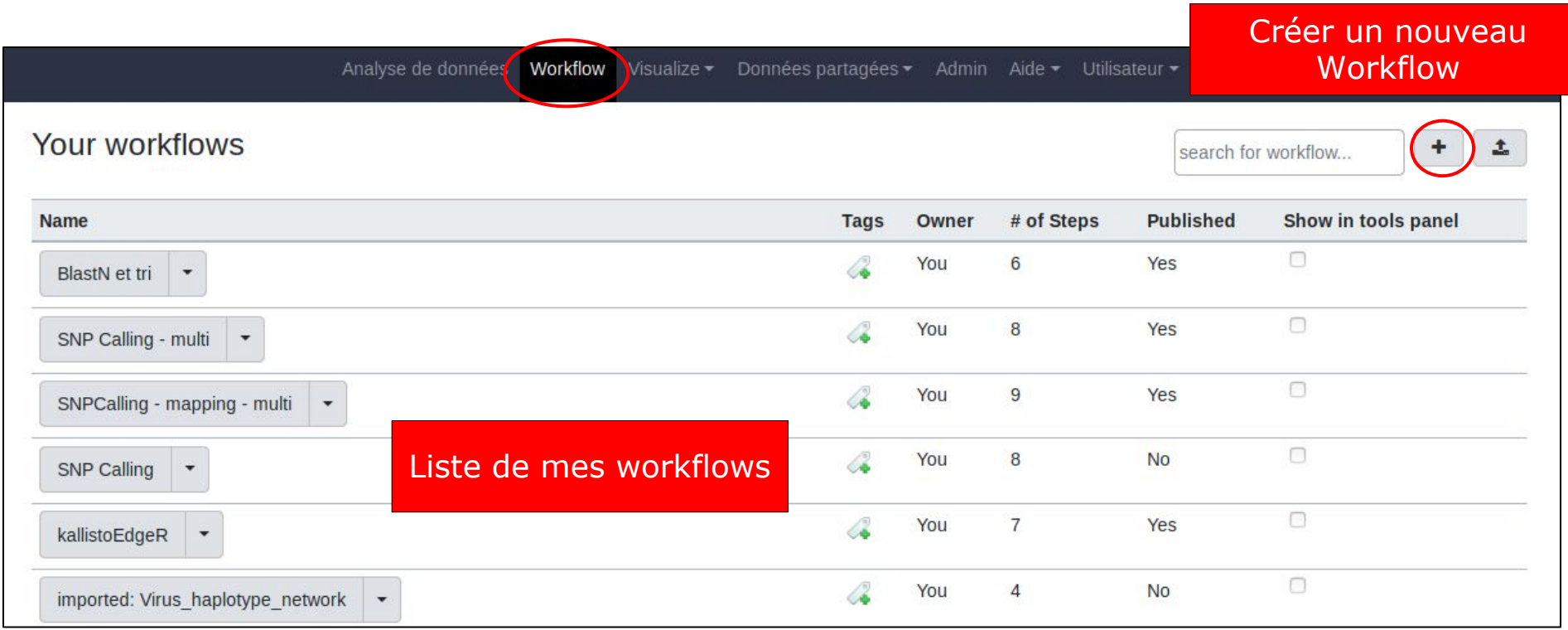
Collection = Permet d'effectuer une même analyse sur plusieurs échantillons

1) Créez une collection avec les deux jeux de séquences

2) Lancez blastn sur la collection et observez le résultat



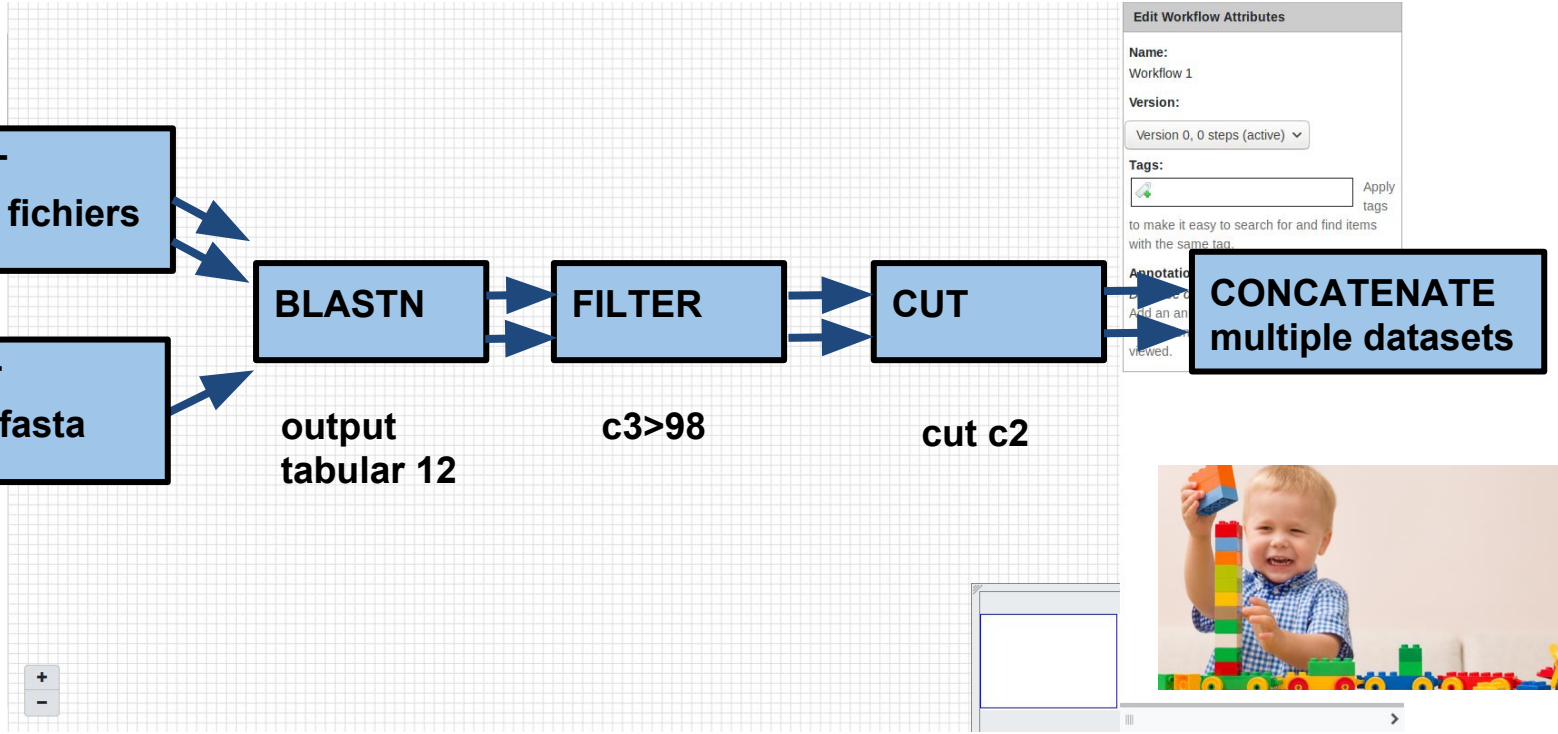
Galaxy : Workflows



Galaxy : Workflows

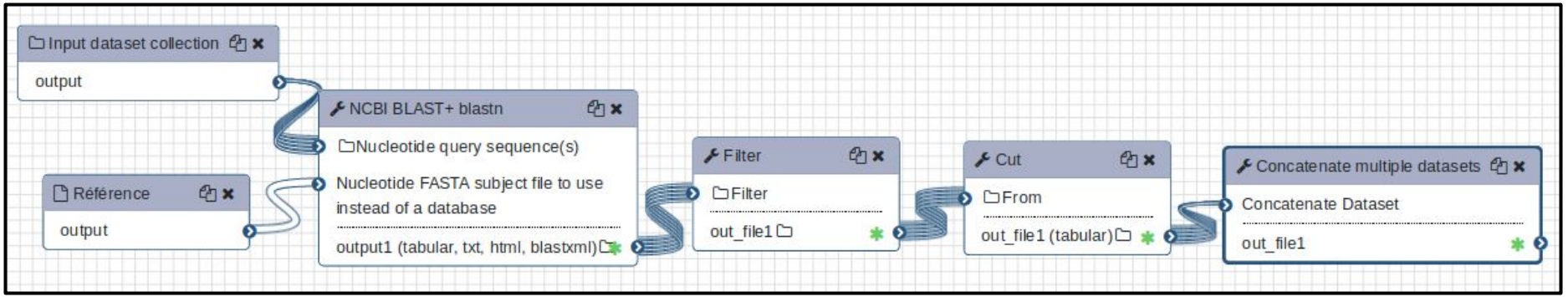
TP

Créez votre premier Workflow : BLASTN, paramétrez le et lancez le !



Galaxy : Workflows

TP



TIPS:

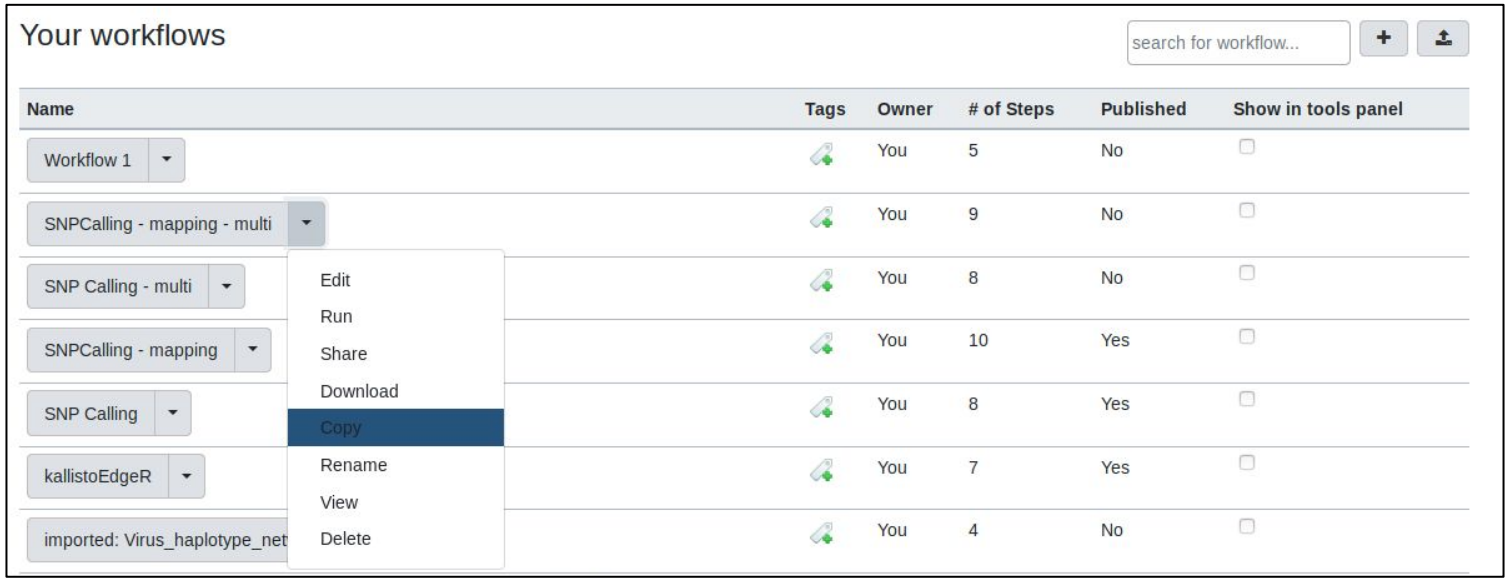
Cochez l'astérisque verte pour afficher votre résultat intermédiaire ou décochez la pour le cacher !

Galaxy : Workflows

Il est possible d'effectuer une copie d'un workflow

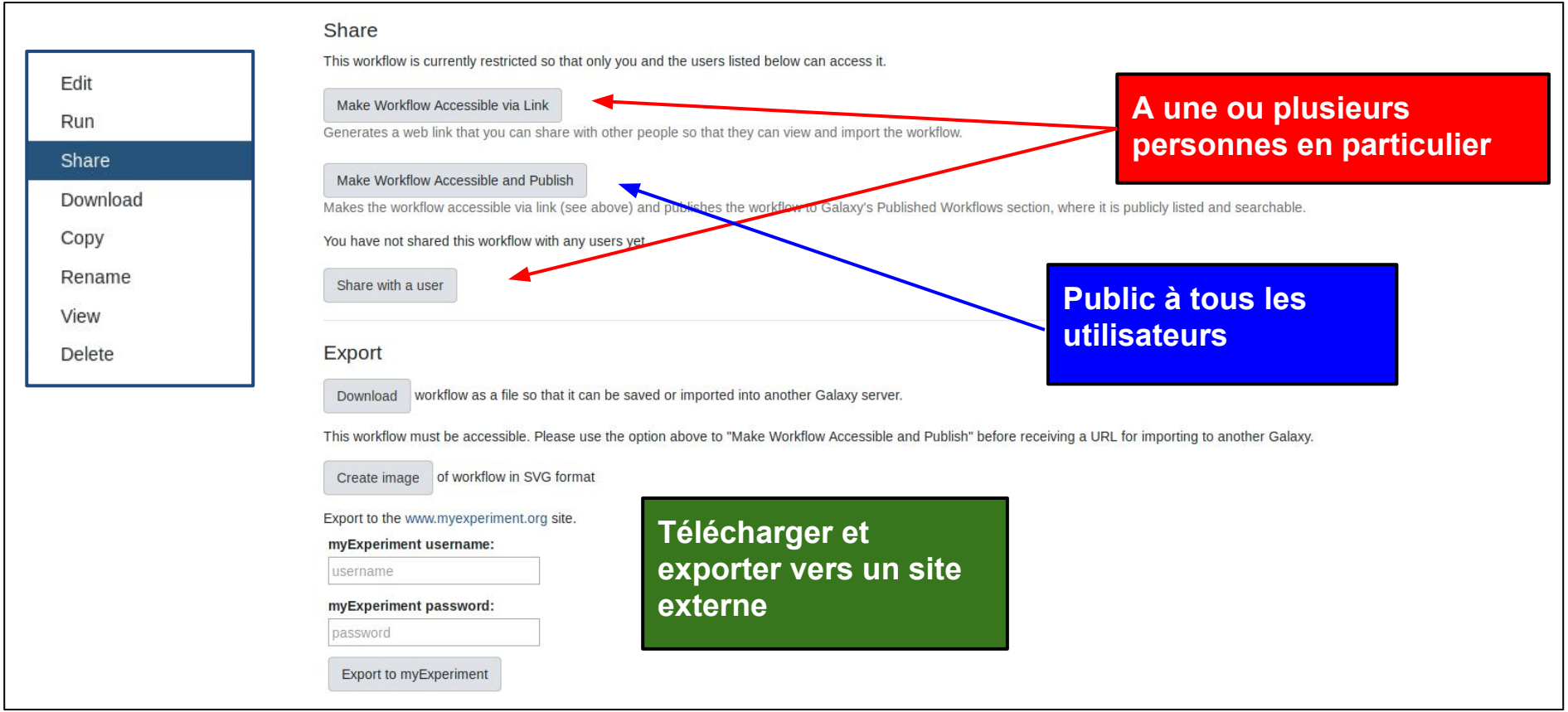
→ par exemple pour rajouter une brique ou modifier les paramètres

→ type de données en entrée (collection, paires, simple fichier)



Galaxy : Workflows

Partager un workflow (ou un historique)



SouthGreen bioinformatics platform

SNP-Calling avec:

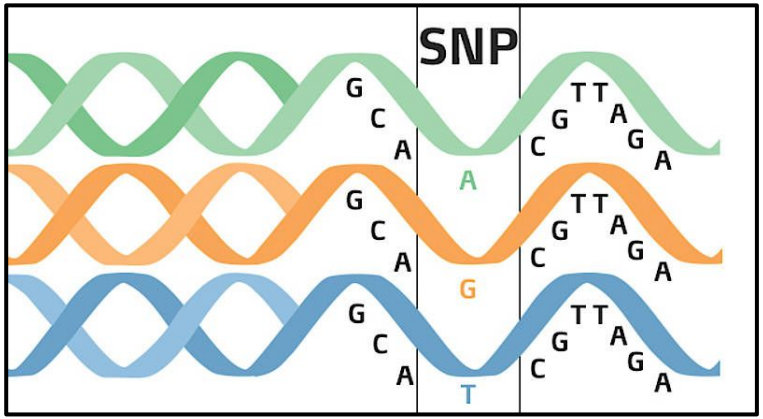
SNP-CALLING

Objectifs :

Avec le séquençage NGS on obtient de nombreux reads présentant des différences / mutations au niveau de certains nucléotides. Ce sont des SNP. Comment les détecter?

Pré-requis:

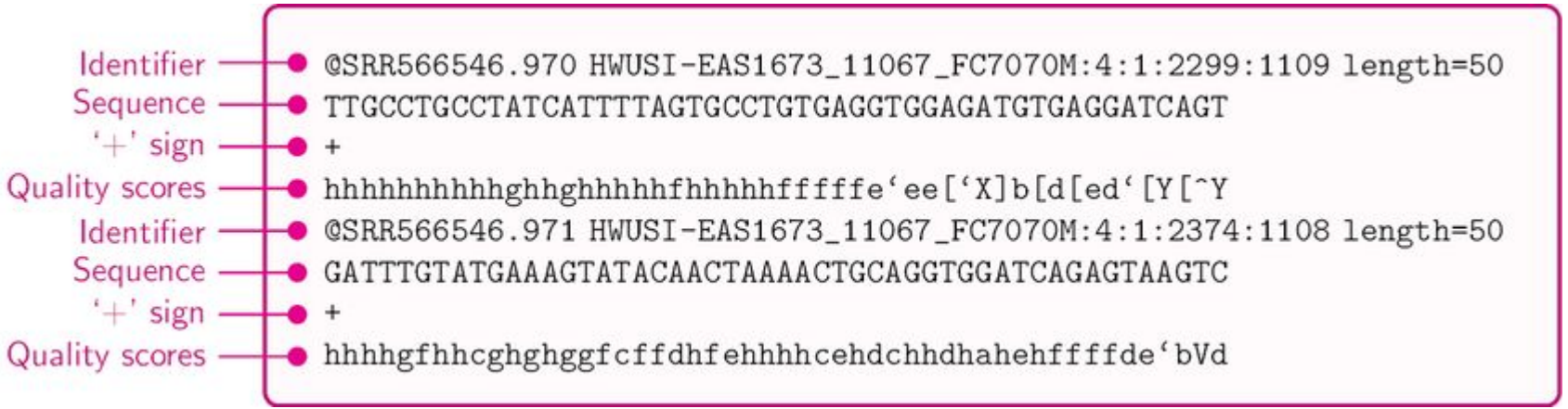
- Une séquence de référence (Fasta)
- Des reads (FastQ)



SNP-CALLING : introduction

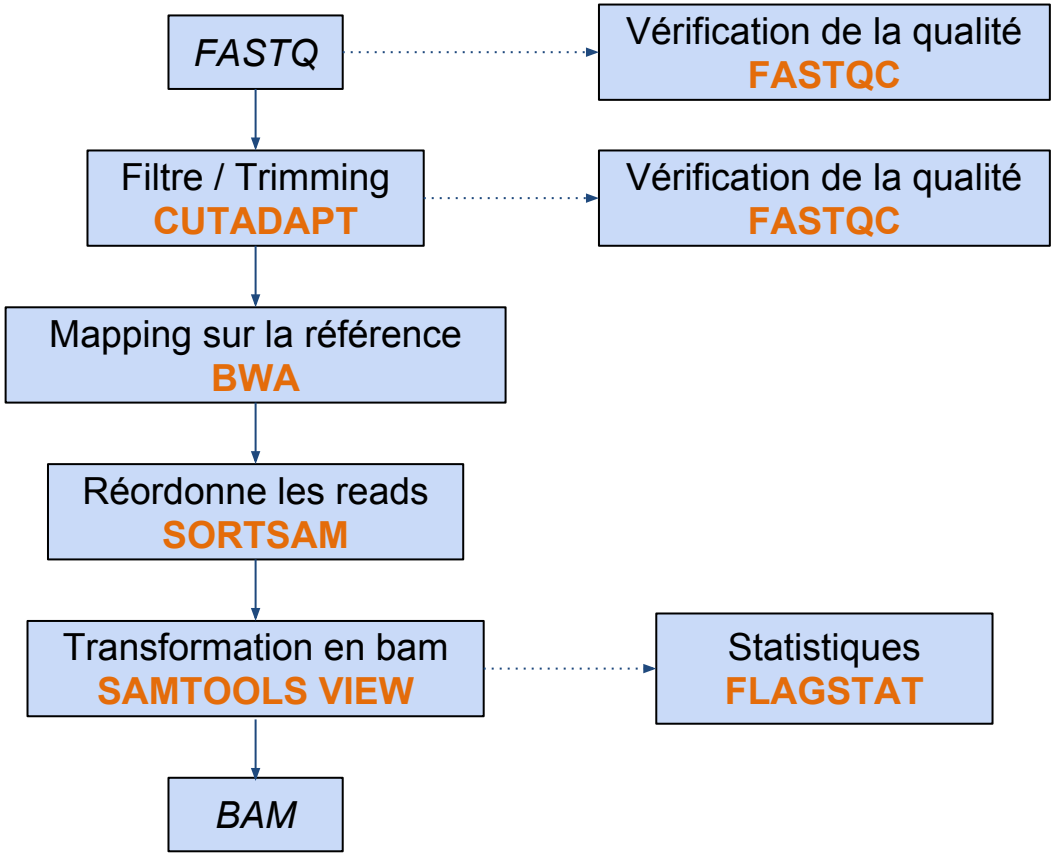
Le format FastQ

Format concis et compact qui stocke à la fois séquence et qualité de séquençage.



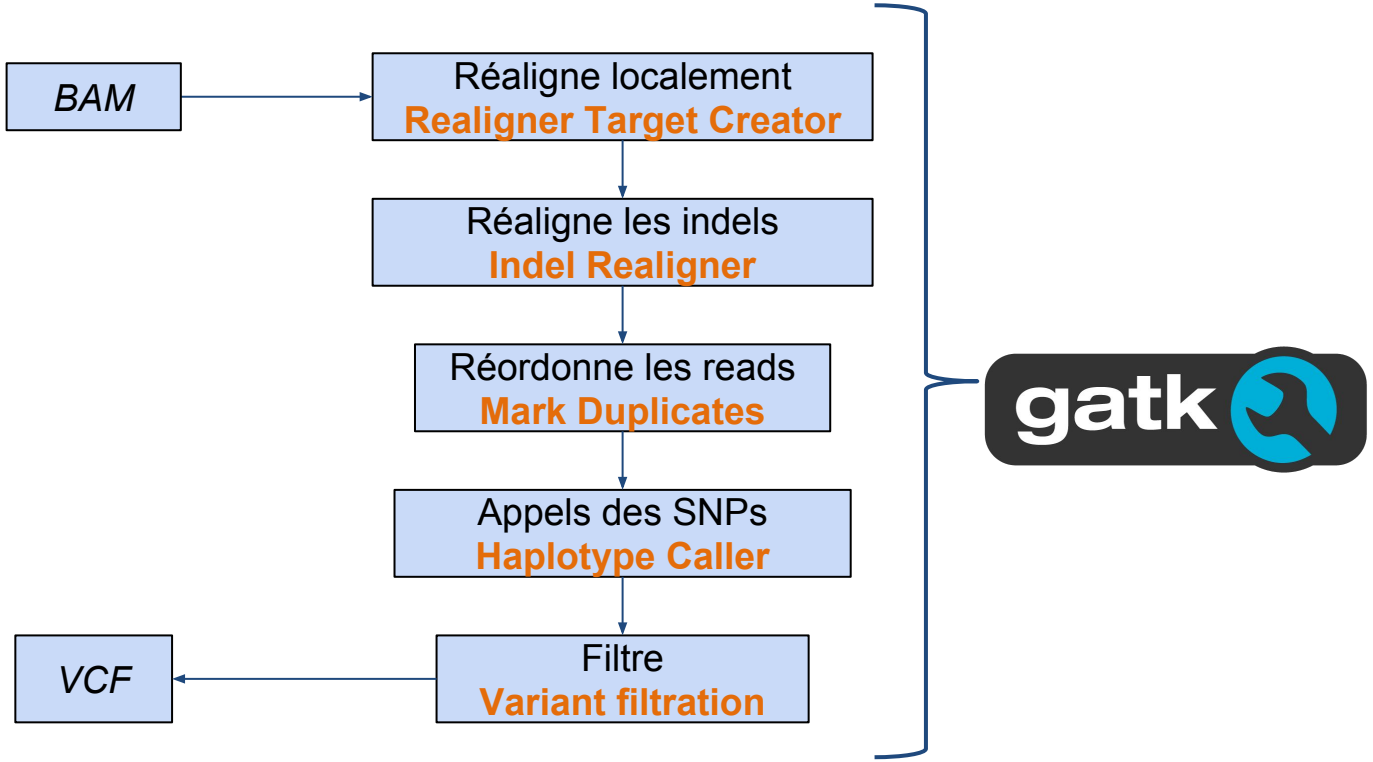
SNP-CALLING : introduction

1) Mapping



SNP-CALLING : introduction

2) SNP calling



SNP-CALLING : introduction

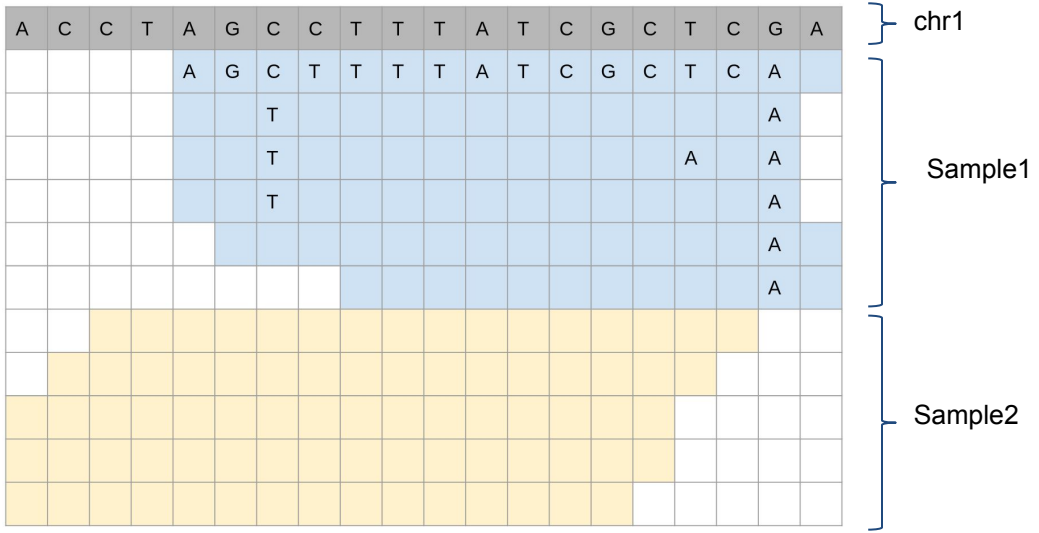
VCF : Description des variants par position + assignation génotypique

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF  ALT     QUAL FILTER INFO                              FORMAT      Sample1         Sample2         Sample3
2      4370    rs6057    G    A       29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51  1|0:48:8:51,51  1/1:43:5:.,.
2      7330    .         T    A       3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50  0|1:3:5:65,3    0/0:41:3
2      110696  rs6055    A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27  2|1:2:0:18,2    2/2:35:4
2      130237  .         T    .       47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60  0|0:48:4:56,51  0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50   PASS   NS=2;DP=9;AA=G                   GT:GQ:DP    0/1:35:4        0/2:17:2        1/1:40:3
```

SNP-CALLING : introduction

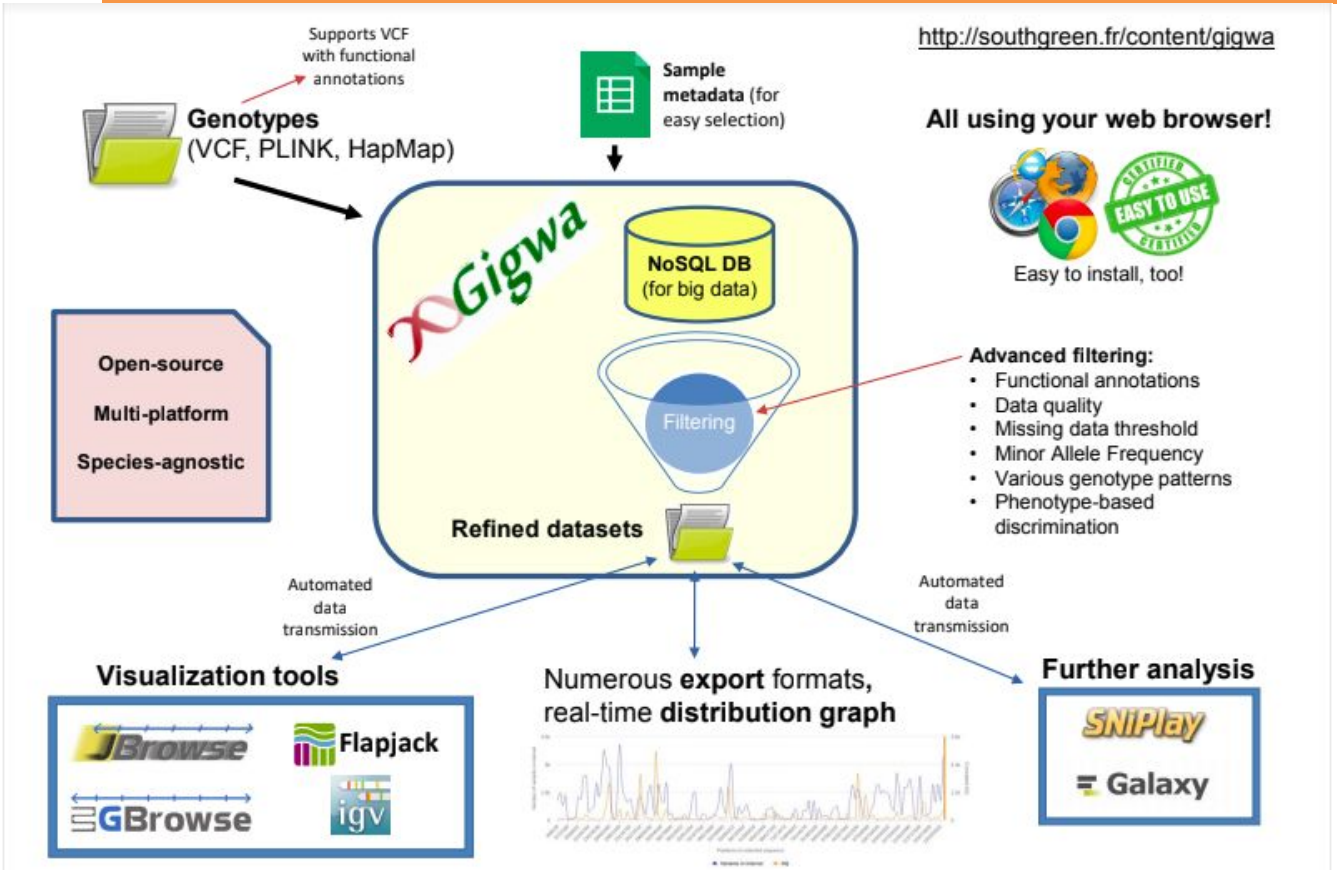
adapté de http://genoweb.toulouse.inra.fr/~formation/16_SGS-SNP/DOC/SGS_SNP_slides_Mars2017.pdf

VCF : Description des variants par position + assignation génotypique



#CHR	POS	ID	REF	ALT	QUAL	FILTER	[INFOS]	FORMAT	Sample1	Sample2
chr1	7	.	C	T	247.82	.	[INFO]	GT/AD/DP/GQ/PL	0/1:2,3:5:9.2:20,0,15	./.
chr1	19	.	G	A	124.34	.	[INFO]	GT/AD/DP/GQ/PL	0/0:5,0:5:20.2:0,42,94	./.

Gigwa : Genotype investigator for genome wide analyses



Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. *Gigwa-Genotype investigator for genome-wide analyses.* Gigascience [Internet] 2016;5:25. doi: 10.1186/s13742-016-0131-8. **(Version 2 publication currently in press)**

Galaxy : Import de fichiers

TP

→ Depuis la library partagée

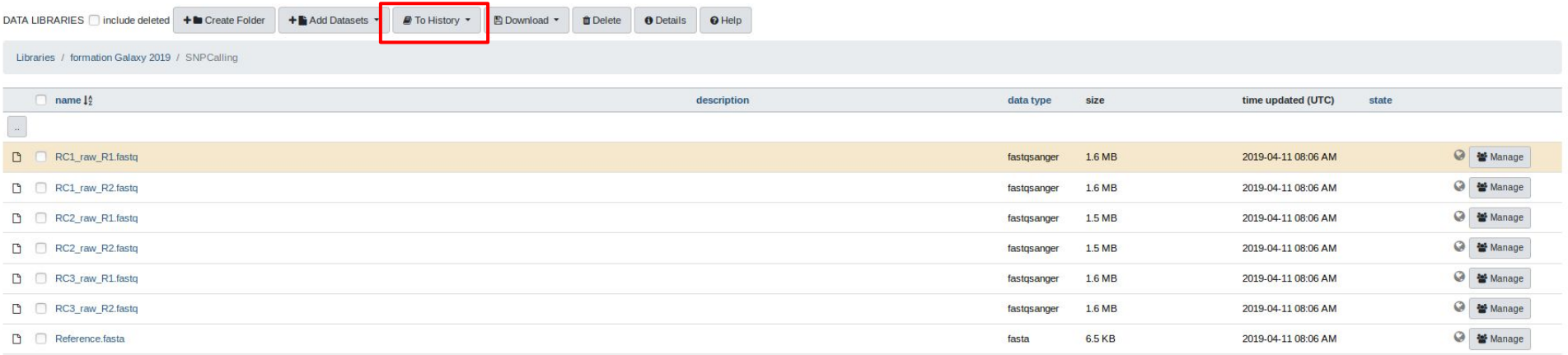
Accédez aux données partagées

(Données partagées → Bibliothèque de données → formation Galaxy 2019 → SNPCalling)

1) Cliquez sur la library **Formation Galaxy 2019 SNPCalling**

2) Cochez les fichiers:
RCX_raw_RX.fastq
Reference.fasta

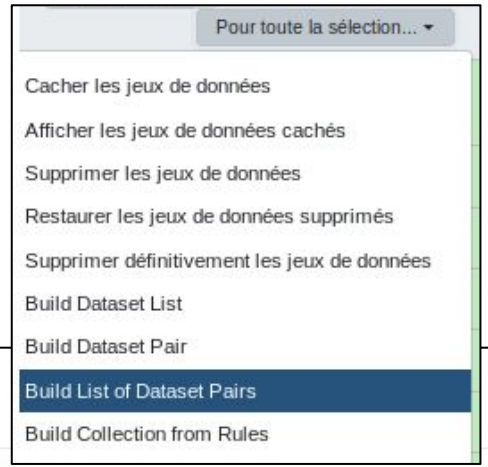
3) Cliquez sur le bouton “To history” pour importer les données.



SNP-CALLING : Workflows

TP

1) Créer une collection pour des données pairées



Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help. ×

3 unpaired forward - (3 filtered out) | Choose filters Clear filters Auto-pair | **3 unpaired reverse** - (3 filtered out)

R1 | R2

RC1_raw_R1.fastq	Pair these datasets	RC1_raw_R2.fastq
RC2_raw_R1.fastq	Pair these datasets	RC2_raw_R2.fastq
RC3_raw_R1.fastq	Pair these datasets	RC3_raw_R2.fastq

SNP-CALLING : Workflows

TP

1) FastQC : Read Quality reports

FastQC Read Quality reports (Galaxy Version 0.72) — Options

Short read data from your current history

6: input

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Adapter list

Nothing selected

list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)

Submodule and Limit specifing file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter

Disable grouping of bases for reads >50bp

Yes | No

Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

Lower limit on the length of the sequence to be shown in the report

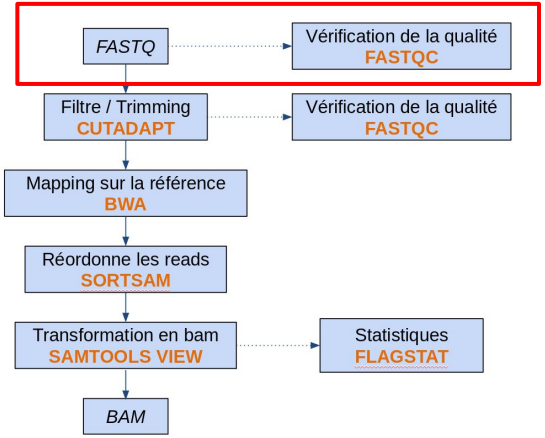
As long as you set this to a value greater or equal to your longest read length then this will be the sequence length used to create your read groups. This can be useful for making directly comaparable statistics from datasets with somewhat variable read lengths. (--min_length)

length of Kmer to look for

7

note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)

Execute

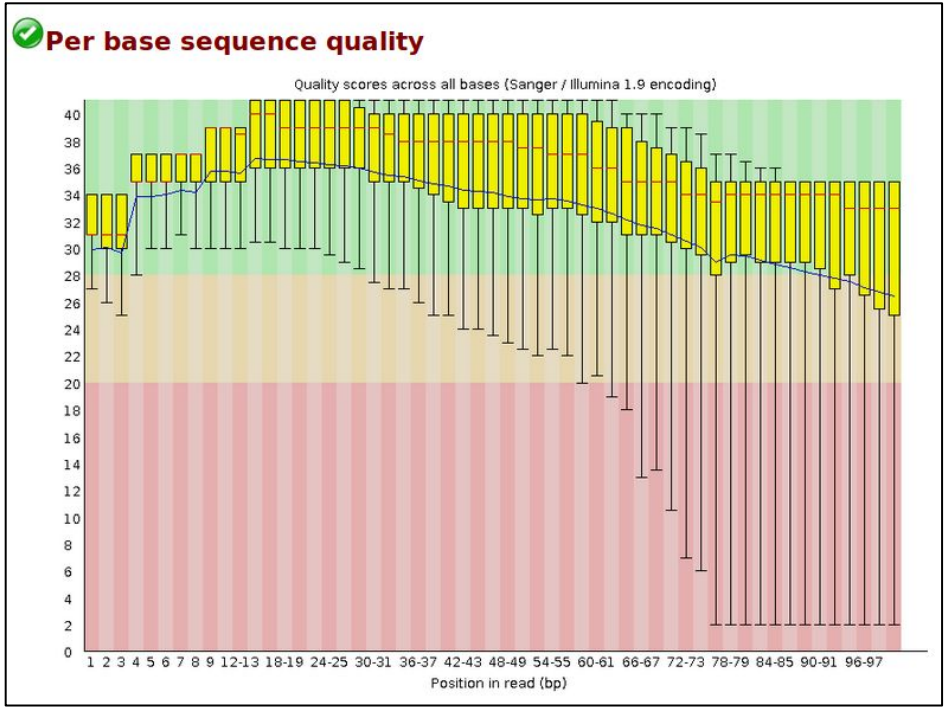


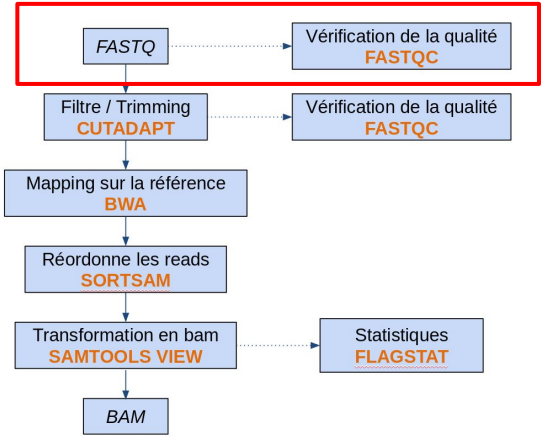
SNP-CALLING : Workflows

TP

1) FastQC : Read Quality reports

Produit 2 datasets en sortie : Rawdata & **webpage**



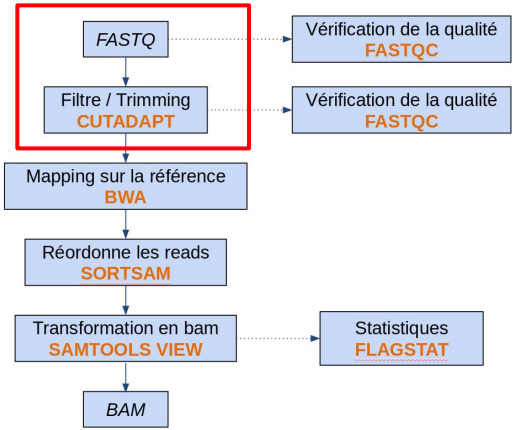


Explication détaillée de toutes les sorties:
https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

SNP-CALLING : Workflows

TP

2) Cutadapt : Remove adapter sequences from Fastq/Fasta



Paramètres :

Adapter Option:
Minimum overlap length = 7

Filter Option:
minimum-length=35

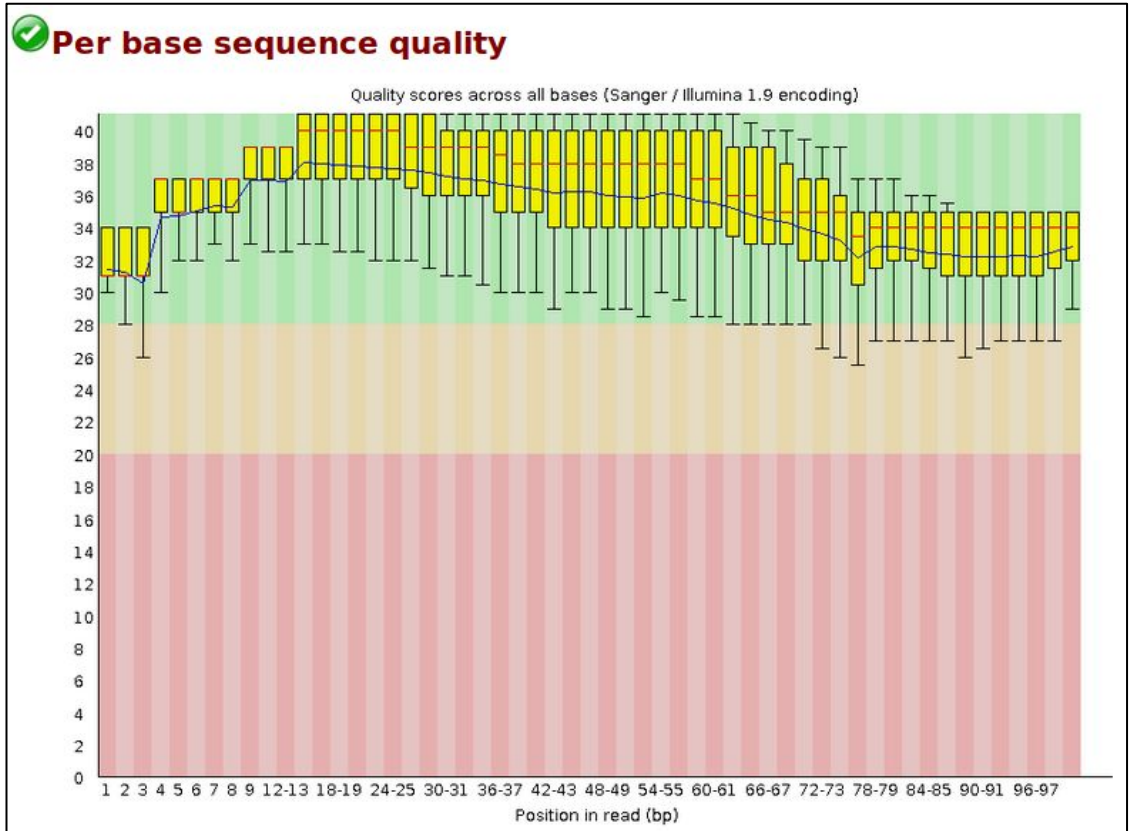
Read Modification Options:
Quality cutoff=20,20

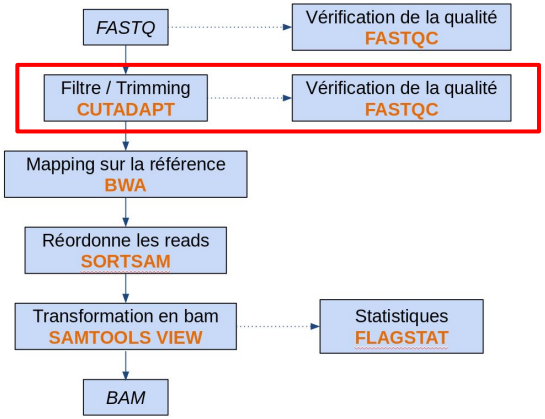
→ reads fastq filtrés en fonction de la qualité

SNP-CALLING : Workflows

TP

3) FastQC : Read Quality reports





SNP-CALLING : Workflows

TP

4) BWA map short reads against reference genome

Will you select a reference genome from your history or use a built-in index?

Use a genome from history and build index

Built-ins were indexed using default options. See `Indexes` section of help below

Use the following dataset as the reference sequence

7: Reference.fasta

You can upload a FASTA sequence to the history and use it as reference

Select input type

Paired fastq

Select between fastq and bam datasets and between paired and single end data

Select first set of reads

23: Cutadapt on collection 8: Read 1 Output

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Specify dataset with forward reads

Select second set of reads

24: Cutadapt on collection 8: Read 2 Output

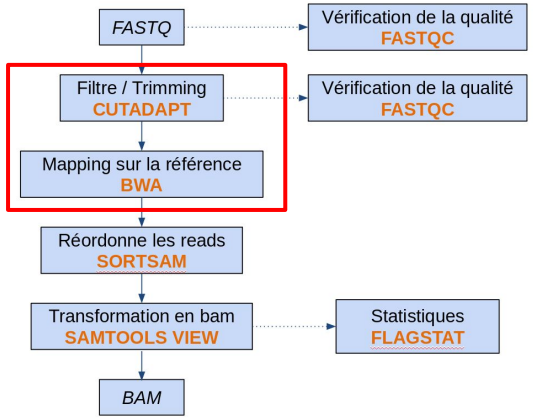
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Specify dataset with reverse reads

Set advanced paired end options?

Do not set

Provides additional controls



→ Collection de 3 bam

SNP-CALLING : Workflows

TP

Map with BWA on collection 23 and collection 24 (mapped reads in BAM format)

a list with 3 items

Add tags

RC1_raw

929.9 KB

format: **bam**, génome de référence: **?**

Reference genome size is 6632 bytes, generating BWA index with is algorithm
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forwa

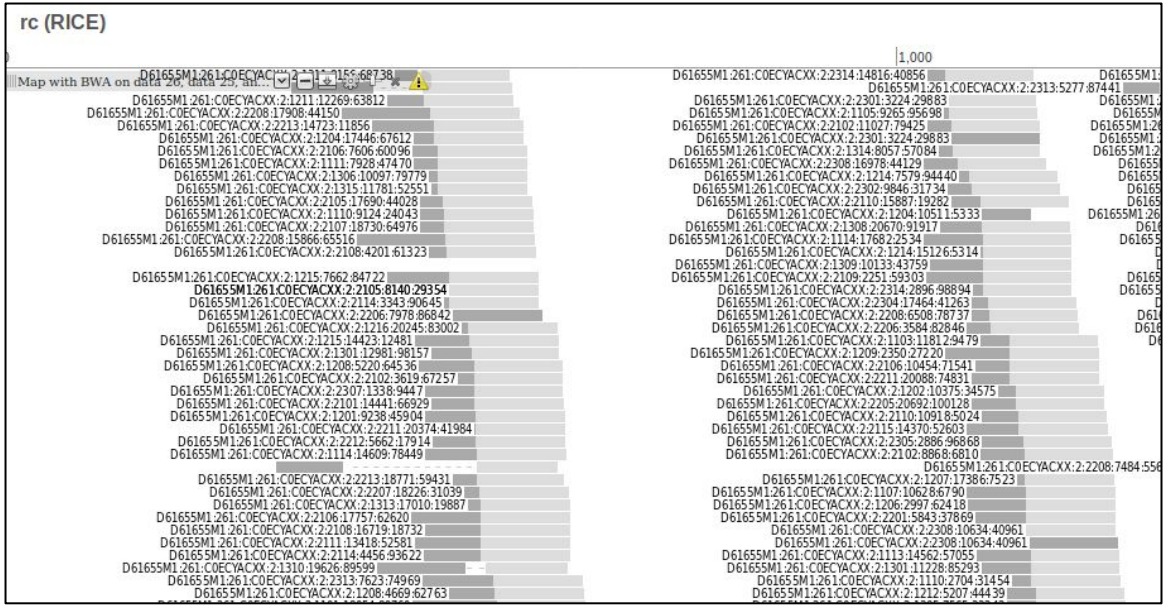
display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io

`Binary bam alignments file`

RC3_raw

RC2_raw

Visualisation du mapping dans Trackster



SNP-CALLING : Workflows

TP

Données statistiques du mapping dans bam.iobio.io

Map with BWA on collection 23 and collection 24 (mapped reads in BAM format)
a list with 3 items

Add tags

RC1_raw

929.9 KB

format: **bam**, génome de référence: **?**

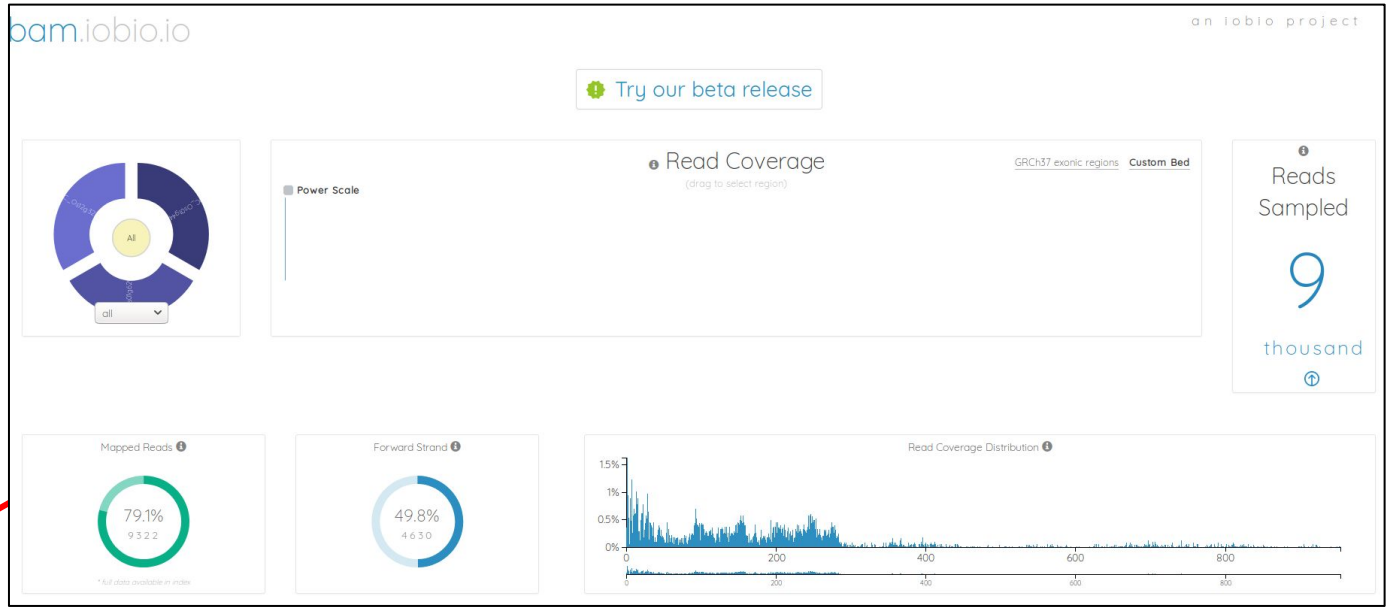
Reference genome size is 6632 bytes, generating BWA index with is algorithm
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forwa

display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io

Binary bam alignments file

RC3_raw

RC2_raw



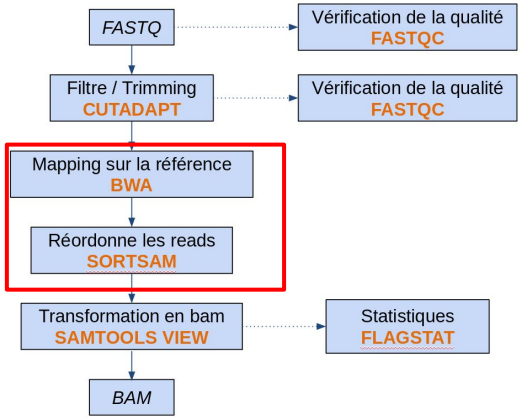
SNP-CALLING : Workflows

TP

5) SortSAM : sort SAM/BAM dataset

Paramètres :

Sort order = coordinate
Select validation Stringency = SIlent



SortSam sort SAM/BAM dataset (Galaxy Version 2.18.2.0)

Select SAM/BAM dataset or dataset collection

47: Map with BWA on collection 23 and collection 24 (mapped reads in BAM format)

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

If empty, upload or import a SAM/BAM dataset

Sort order

- (•) Coordinate
- ( ) Queryname

SORT_ORDER; default=coordinate. Selecting Queryname will output SAM file, as Galaxy does not support BAM files that are not coordinate sorted.

Select validation stringency

Silent

Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

✔ Execute

→ Collection de 3 bam ordonnés

SNP-CALLING : Workflows

TP

6) Samtools view : reformat, filter, or subsample

SAM/BAM/CRAM data set

52: (hidden) SortSam on data 48: Alignment sorted in coordinate order

Output type

BAM (-b)

Select output type. In case of counts only the total number of alignments is returned. All filters are taken into account (-b/-C/-c)

Filter alignment

Yes

Filter by regions

No

Filter by readgroup

No

Filter by quality

Skip alignments with MAPQ smaller than INT. (-q)

Filter by library

Only output alignments in library STR (-l)

Filter by number of CIGAR bases consuming query sequence

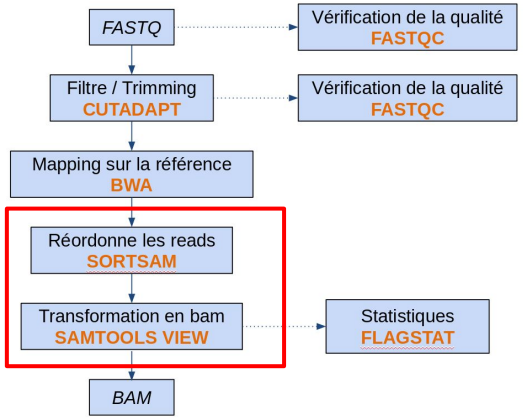
Only output alignments with number of CIGAR bases consuming query sequence greater than or equal INT. (-m)

Require that these flags are set

Select/Unselect all

× read is mapped in a proper pair

(-f)



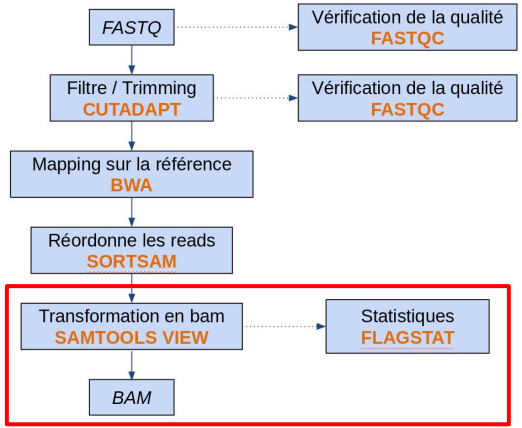
Paramètres :
Output Type = BAM
Require that these flags are set = read is mapped in a proper pair

→ Collection de 3 bam finaux

SNP-CALLING : Workflows

TP

7) Samtools Flagstat : tabulate descriptive stats for BAM dataset

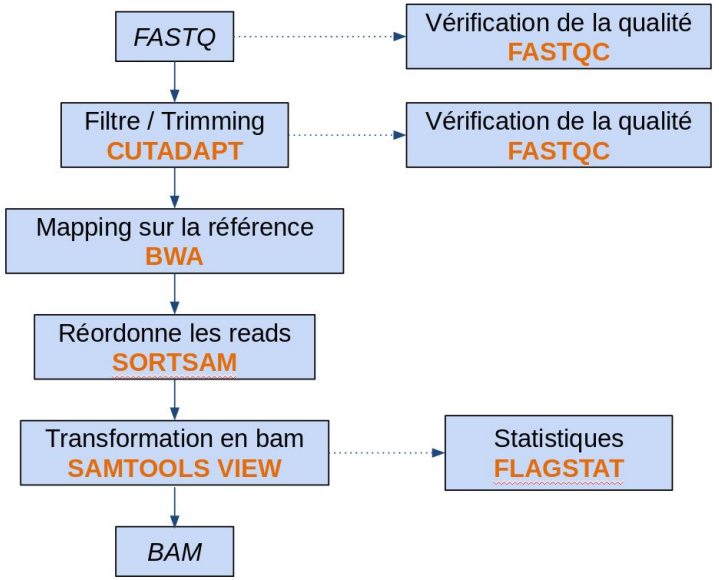



```
11812 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
9322 + 0 mapped (78.92% : N/A)
11812 + 0 paired in sequencing
5906 + 0 read1
5906 + 0 read2
9078 + 0 properly paired (76.85% : N/A)
9292 + 0 with itself and mate mapped
30 + 0 singletons (0.25% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

SNP-CALLING : Workflows

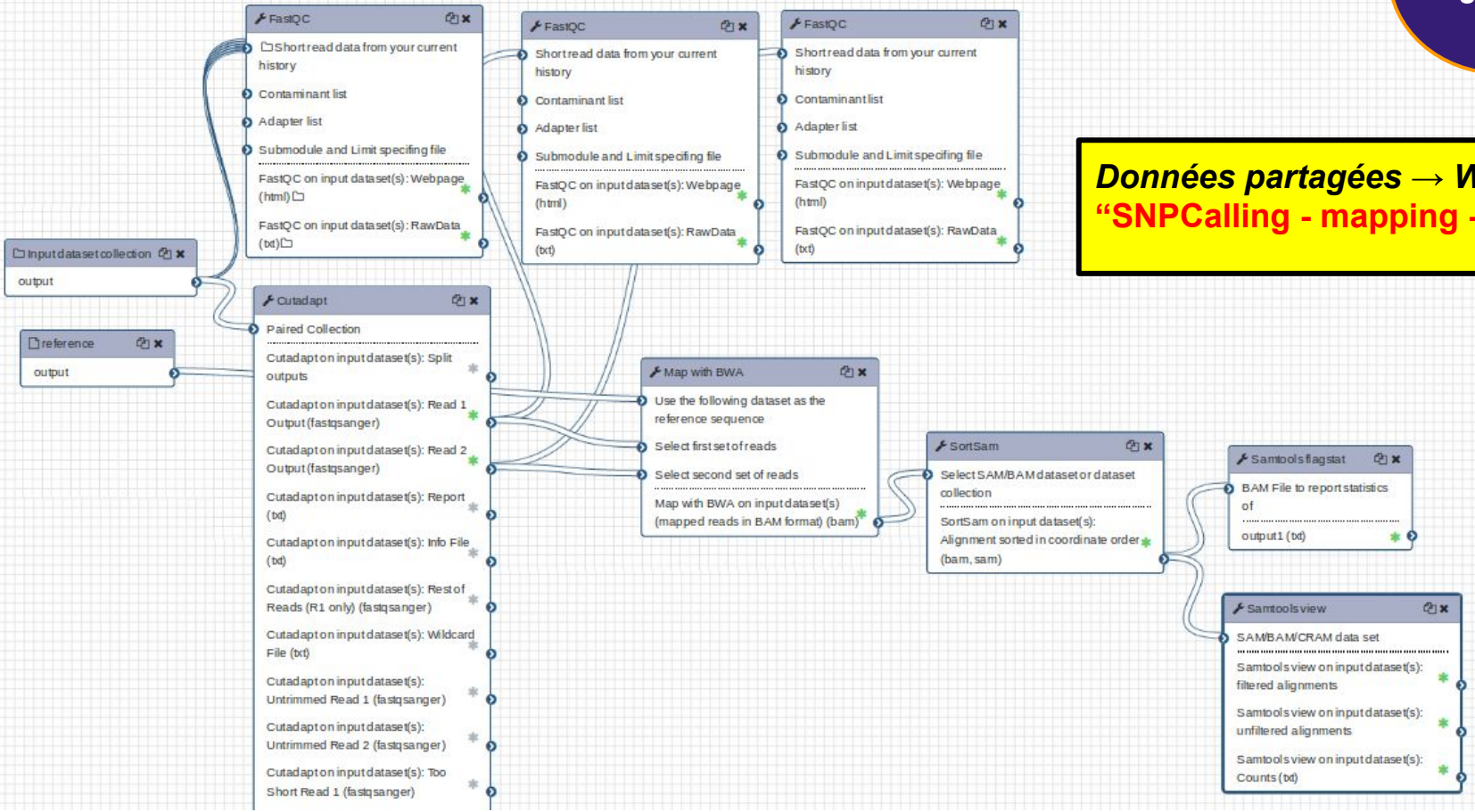
TP

1) CONSTRUISEZ LE WORKFLOW MAPPING



SNP-CALLING : Workflows

TP

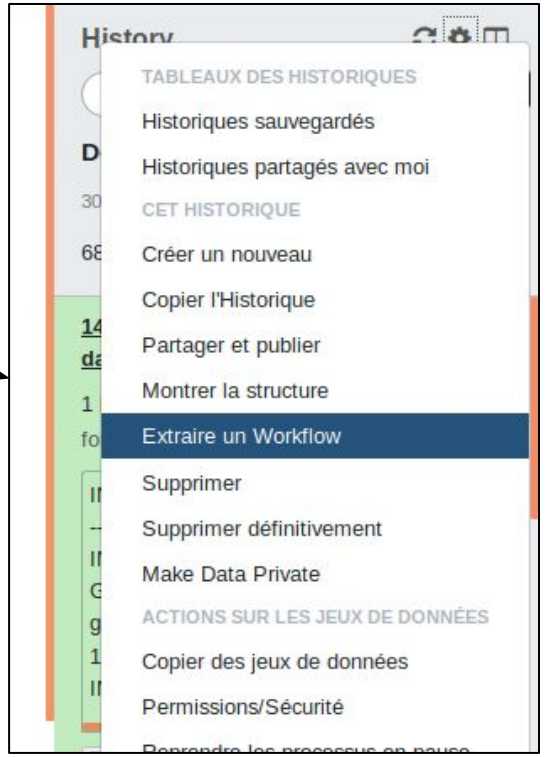


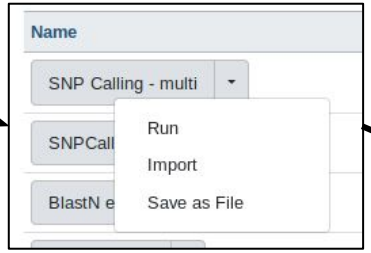
Données partagées → WF
“SNPCalling - mapping - multi”

Galaxy : Workflows

Pour lancer un WORKFLOW on peut :

- **L'extraire de l'historique** après avoir lancé tous les outils
- **Le construire manuellement** avec le canva
- **L'importer** depuis les données partagées

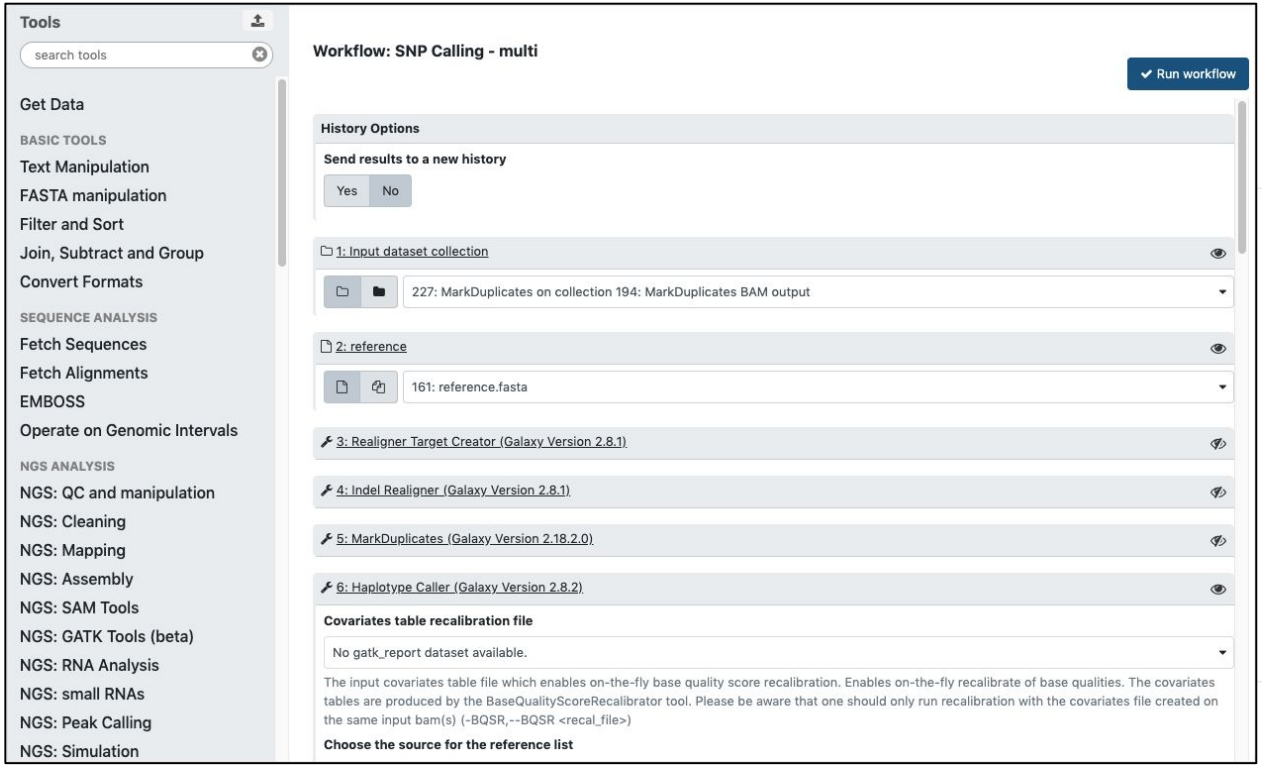


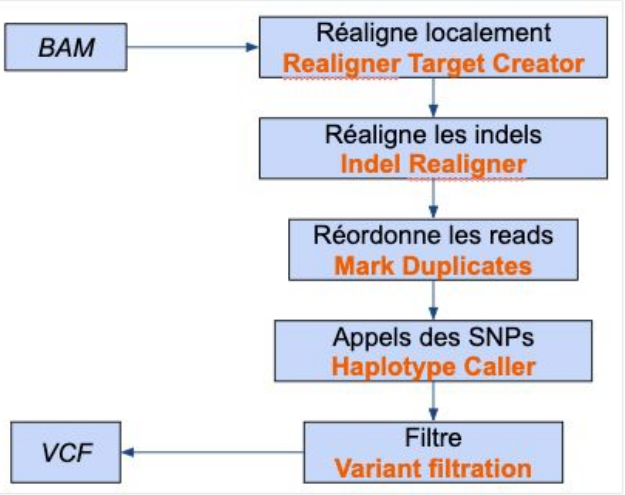


Dans vos Workflows personnels !

SNP-CALLING : Workflows

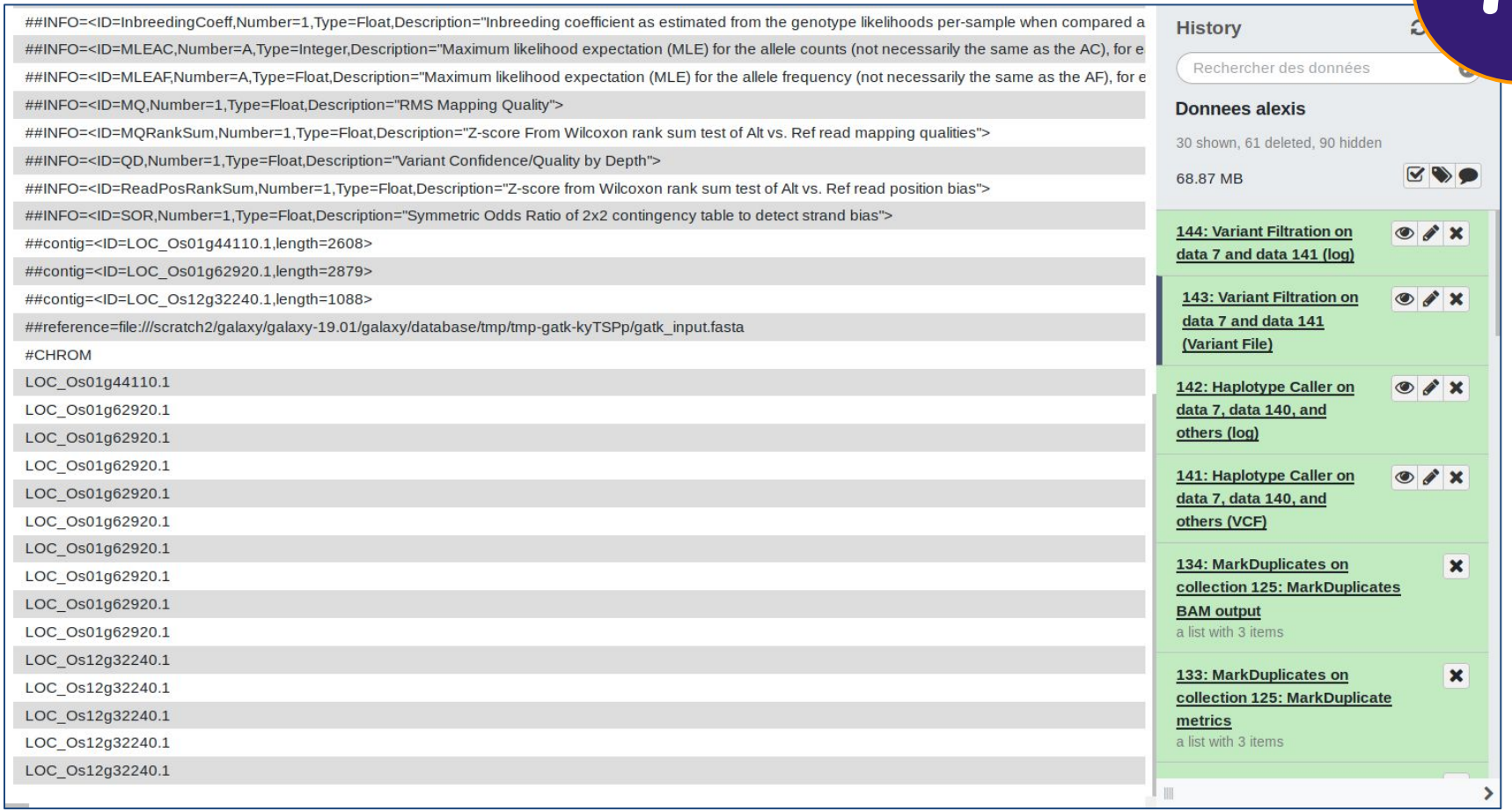
1) Télécharger le workflow “SNPCalling - multi” depuis les données partagées
2) Le lancer sur le BAM obtenu précédemment





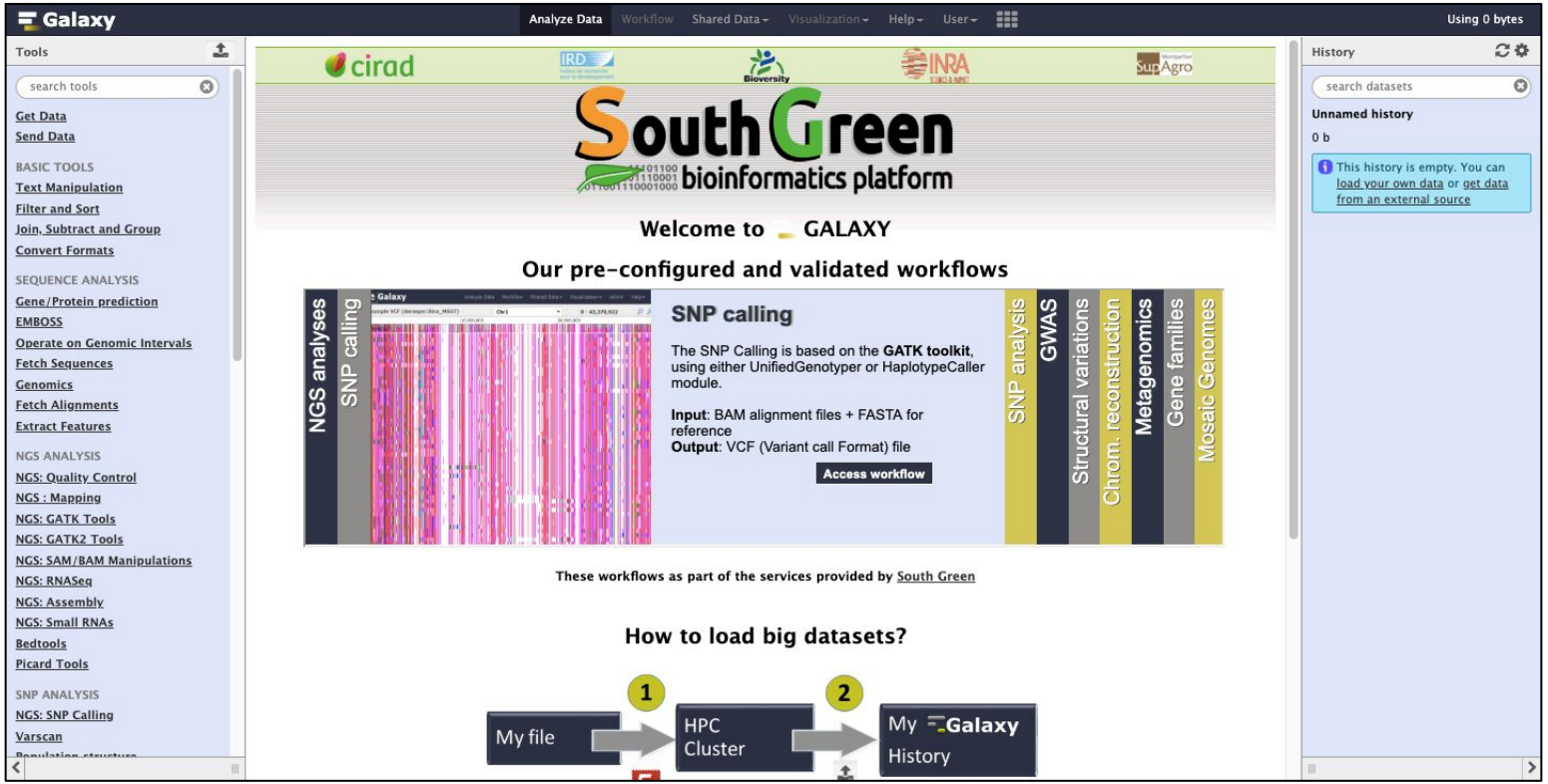
SNP-CALLING : Workflows

TP



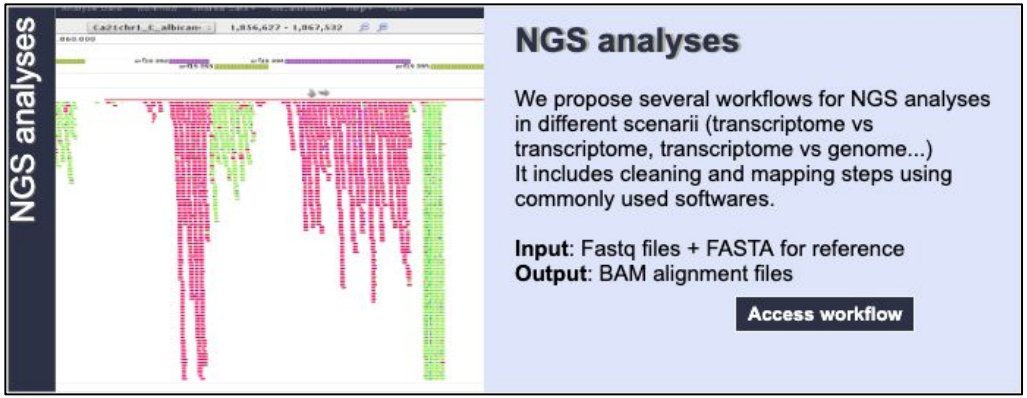
Aperçu des Workflows de SouthGreen

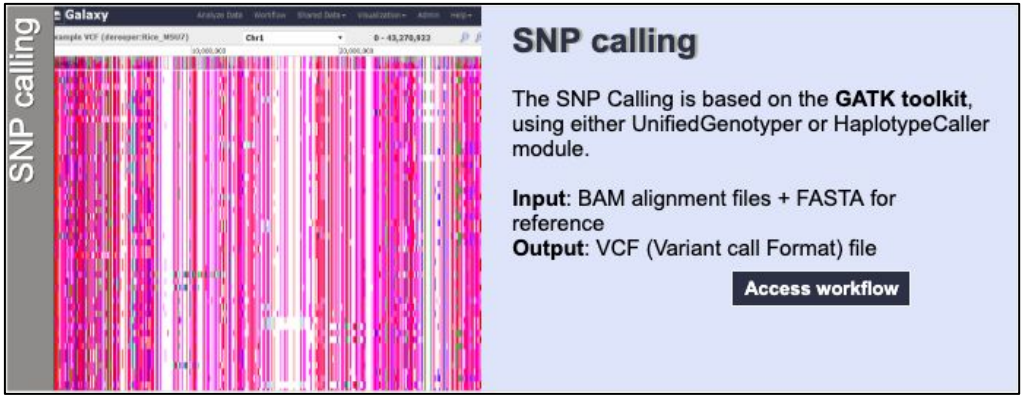
> 9 workflows préconfigurés et validés par la plateforme



Aperçu des Workflows de SouthGreen

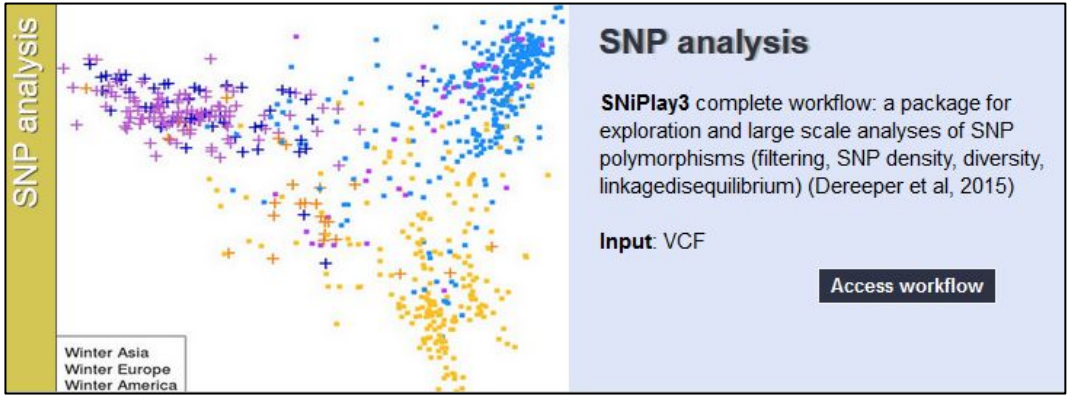
> 9 workflows préconfigurés et validés par la plateforme





Aperçu des Workflows de SouthGreen

> 9 workflows préconfigurés et validés par la plateforme

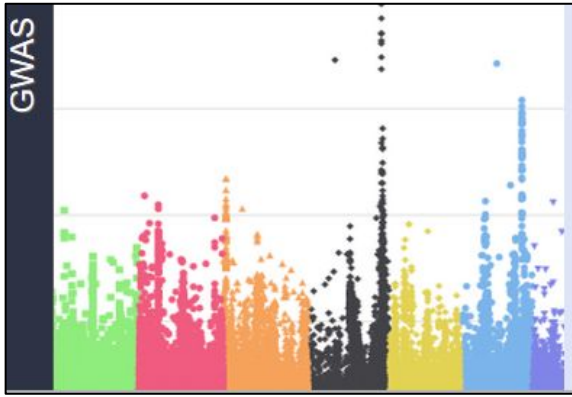


SNP analysis

SNiPlay3 complete workflow: a package for exploration and large scale analyses of SNP polymorphisms (filtering, SNP density, diversity, linkagedisequilibrium) (Dereeper et al, 2015)

Input: VCF

Access workflow



GWAS

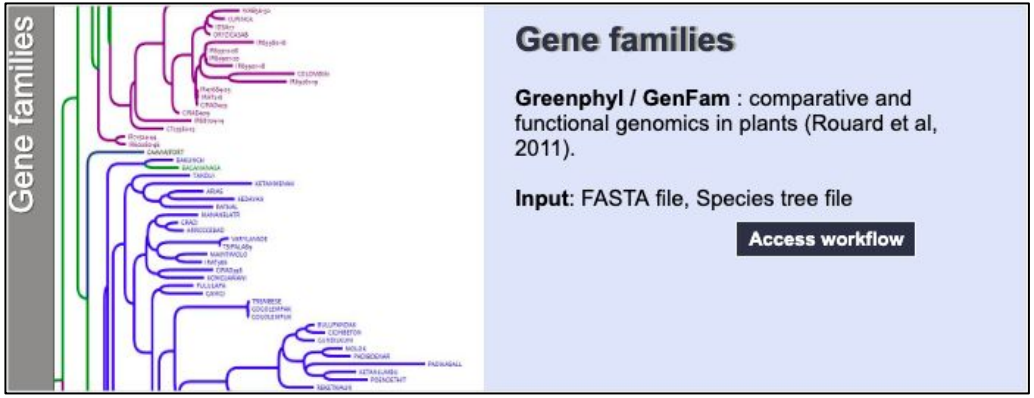
SNiPlay3 GWAS workflow: Tassel-based GWAS workflow (GLM model) including population structure and correction for structure (Dereeper et al, 2015)

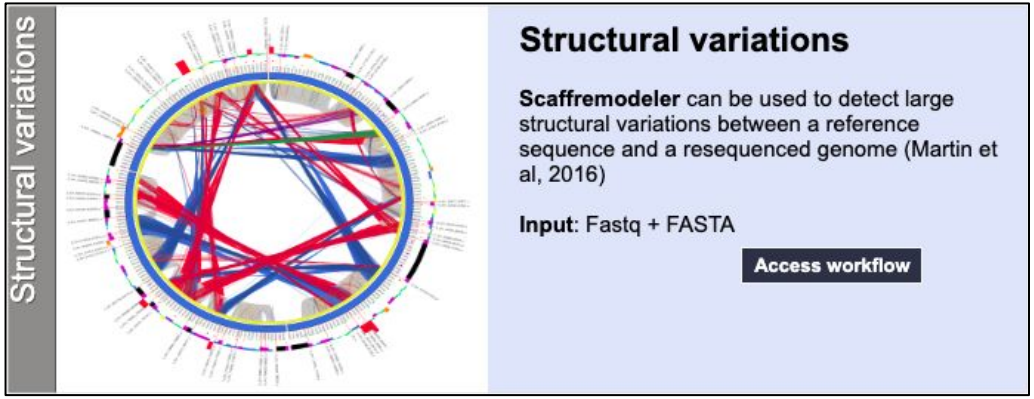
Input: VCF + Phenotypic tabulated file

Access workflow

Aperçu des Workflows de SouthGreen

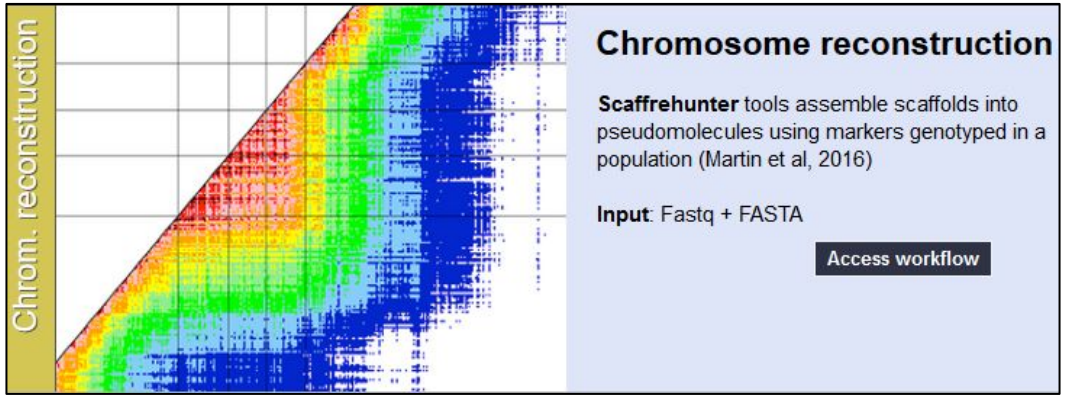
> 9 workflows préconfigurés et validés par la plateforme





Aperçu des Workflows de SouthGreen

> 9 workflows préconfigurés et validés par la plateforme



Chromosome reconstruction

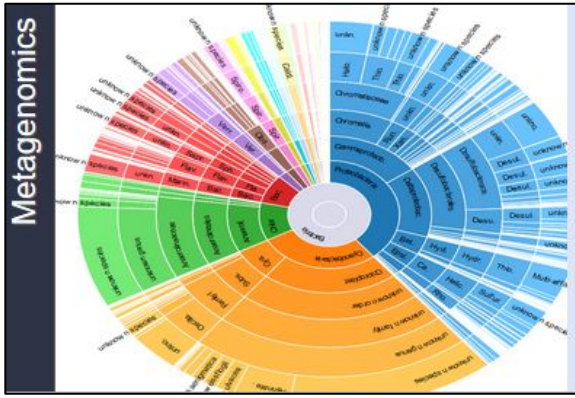
Scaffrehunter tools assemble scaffolds into pseudomolecules using markers genotyped in a population (Martin et al, 2016)

Input: Fastq + FASTA

Access workflow

GenPhySE
Génétique Physiologie et Systèmes d'Elevage

geno toul bioinfo



Metagenomics

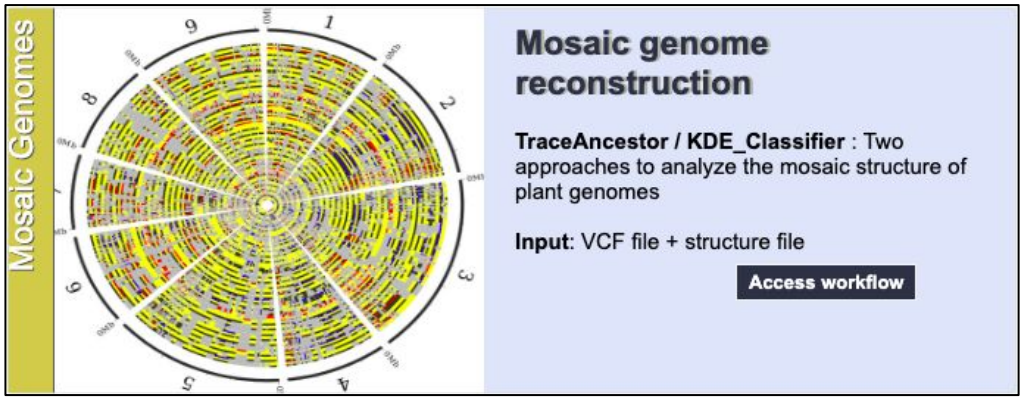
FROGS: Find Rapidly OTU with Galaxy Solution (Pascal et al, 2015)

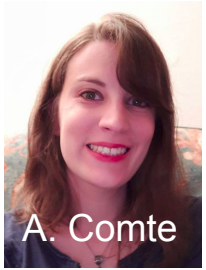
Input: Fastq files

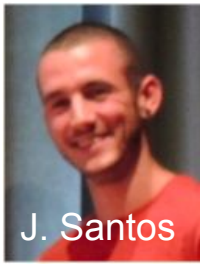
Access workflow

Aperçu des Workflows de SouthGreen

> 9 workflows préconfigurés et validés par la plateforme



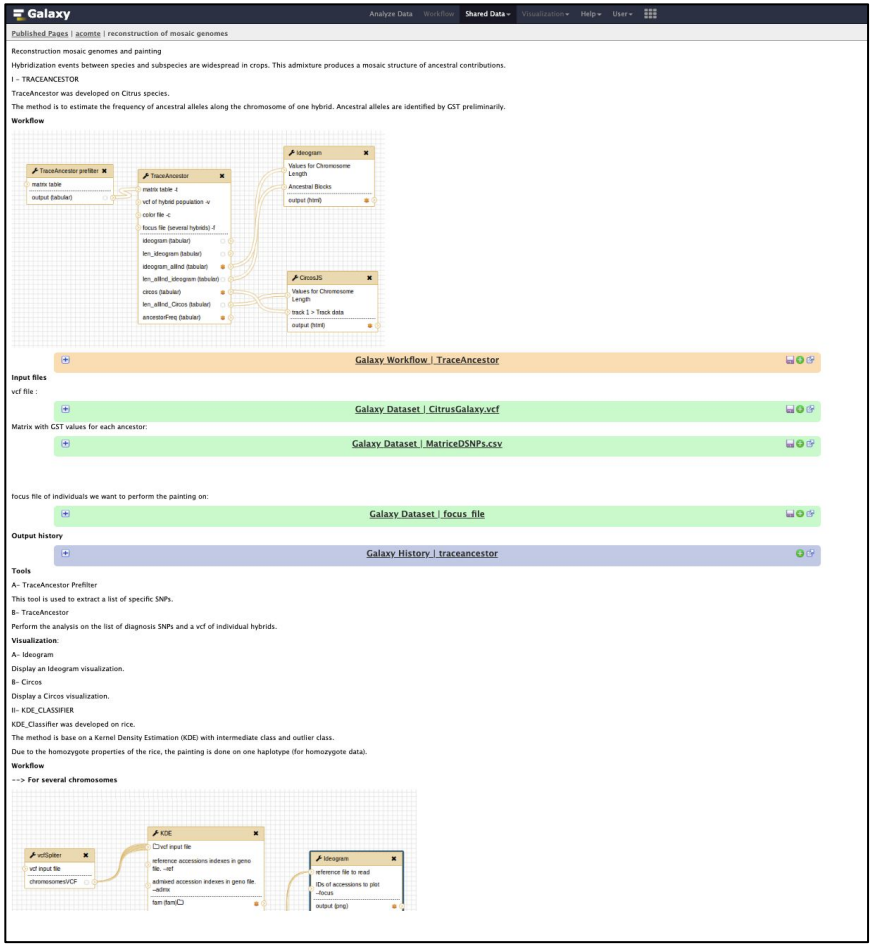




Aperçu des Workflows de SouthGreen

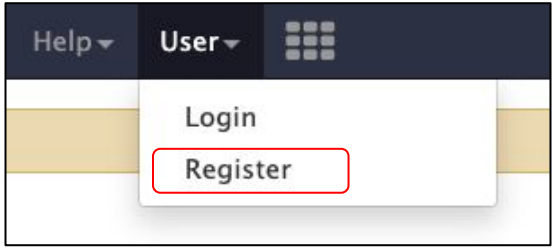
Une documentation est disponible dans les Galaxy pages

> Workflow / Jeux de données / Historique d'analyses



Galaxy : Bonnes pratiques

Galaxy CIRAD : http://galaxy.southgreen.fr/galaxy/



Comment créer un compte:
Directement sur Galaxy : http://galaxy.southgreen.fr/galaxy/user/create
Contactez : admin.bioinfo@cirad.fr pour augmenter l'espace alloué.

Pour tout problème ou demande (briques…): Contactez : admin.bioinfo@cirad.fr

---

Galaxy IRD : http://bioinfo-inter.ird.fr:8080/

Comment créer un compte:

Formulaire disponible sur le site web du plateau: https://bioinfo.ird.fr/index.php/platform/galaxy-account.

- La durée d'un compte est de 3 ans renouvelable sur demande au plateau bioinformatique.
- Quota utilisateur à fixer lors de la création du compte

Pour tout problème ou demande (briques…): contactez : bioinfo@ird.fr

Galaxy : Bonnes pratiques

Bonnes pratiques:

- Pensez à supprimer vos données / historique après analyse
  → galaxy n'est pas une plateforme de stockage

- Connaissez bien vos données et vos objectifs
  → configurations / paramètres

Galaxy : Bonnes pratiques

Comment citer Galaxy?

“The authors acknowledge the South Green Platform (http://www.southgreen.fr) for providing the galaxy instance (http://bioinfo-inter.ird.fr:8080/ or http://galaxy.southgreen.fr/galaxy/) that have contributed to the research results reported within this paper.”

→ N’oubliez pas de citer aussi les outils utilisés !

Comment citer les clusters?

"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: http://bioinfo.ird.fr/ "

“The authors acknowledge the CIRAD UMR-AGAP HPC (South Green Platform) at CIRAD montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.southgreen.fr”

Formateurs

- Christine Tranchant-Dubreuil
- **Sebastien Ravel**
- Alexis Dereeper
- **Jean-François Dufayard**
- Ndomassi Tando
- Bertrand Pitollat
- **François Sabot**
- **Julie Orjuela-Bouniol**
- Gautier Sarah
- **Aurore Comte**
- **Marilyne Summo**
- **Guilhem Sempere**
- **Emmanuelle Beyne**













SUIVEZ NOUS SUR TWITTER !

South Green : @green_bioinfo

I-Trop : @ItropBioinfo

Galaxy Project : @galaxyproject

Merci pour votre attention !