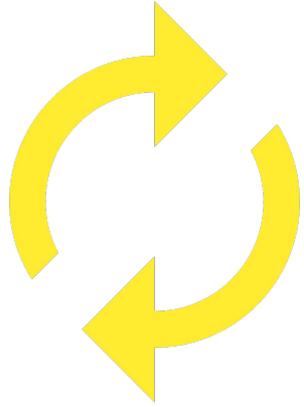


Session de formation 2019





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

genome assembly **SNP detection**
phylogeny **structural variation**
comparative genomics transcriptome assembly differential expression
GWAS **pan-genomics**
population genetics metagenomics
polyploidy



Rice



Banana



Palm



Sorghum



Coffee



Cassava



Magnaporthe



Larmande Pierre

Sabot François

Tando Ndomassi

Tranchant-Dubreuil

Christine



Comte Aurore

Dereeper Alexis



Orjuela-Bouniol Julie



Bocs Stephanie

De Lamotte Frédéric

Droc Gaetan

Dufayard Jean-François

Hamelin Chantal

Martin Guillaume

Pitollat Bertrand

Ruiz Manuel

Sarah Gautier

Summo Marilyne



Rouard Mathieu

Guignon Valentin

Catherine Breton



Mahé Frédéric

Ravel Sébastien



Sempere Guilhem



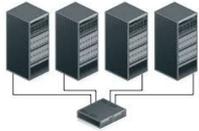
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

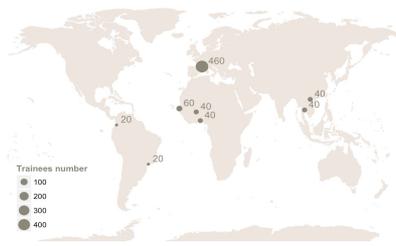
SNAKEMAKE

Galaxy

HPC and trainings....



37 courses organized last 7 years



Genome Hubs & Information System



Gigwa

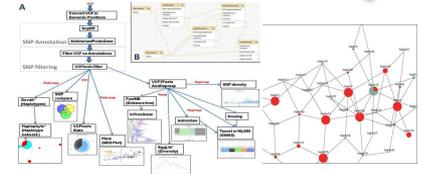
SNPs and Indels

GreenPhyl

Family Id	Family Name	Number of subsequence	Status
SP000013	Cytoschrome P450 superfamily	5542	●●●
SP000017	AP2/ERF1B transcription factor family: ERF/ERF1B group (partial)	5142	●●●
SP000005	NAC transcription factor family	4874	●●●
SP000028	MADS transcription factor family		
SP000018	Haem peroxidase superfamily		
SP000066	General substrate transporter superfamily		
SP000022	Subtilisin-like Serine Proteases family		
SP000019	NIP, NIP1/NIPR FAMILY		

Gene families

SNIPlay



<https://github.com/SouthGreenPlatform>



@green_bioinfo



Erwan Corre



Marie Simonin
Sébastien Cunnac



Etienne Loire
Julie Reveillaud



Florentin Constancias



Valentin Klein



Valérie Noël



Emmanuelle Beyne



And more collaborators !

- 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

TOGGLE

 **Galaxy**
PROJECT

www.southgreen.fr

<https://southgreenplatform.github.io/trainings>





objectifs:

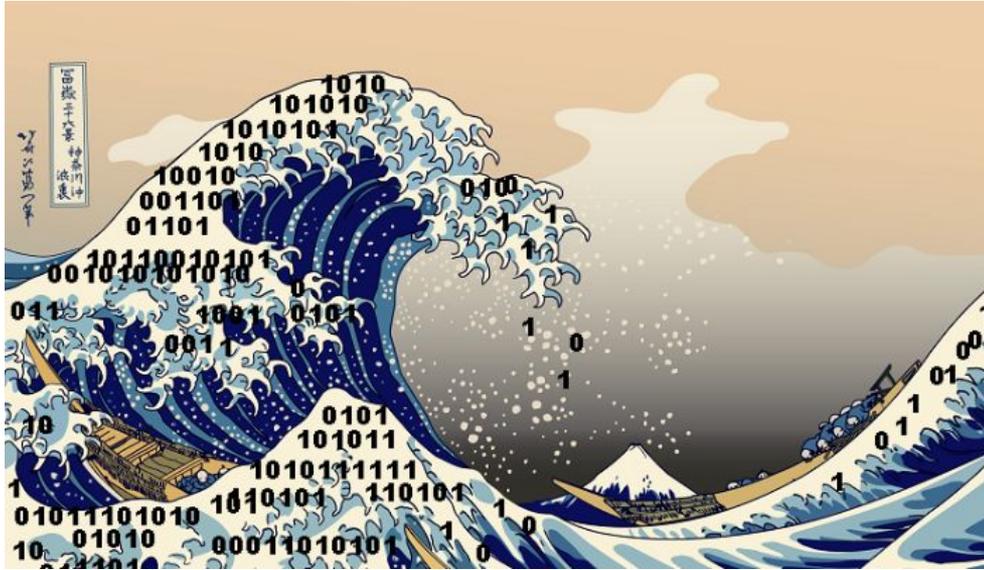
Utiliser les gestionnaires de Workflow de South Green afin de construire de manière automatique vos propres pipelines.

Applications

Tout savoir sur les 2 principaux gestionnaires de workflow

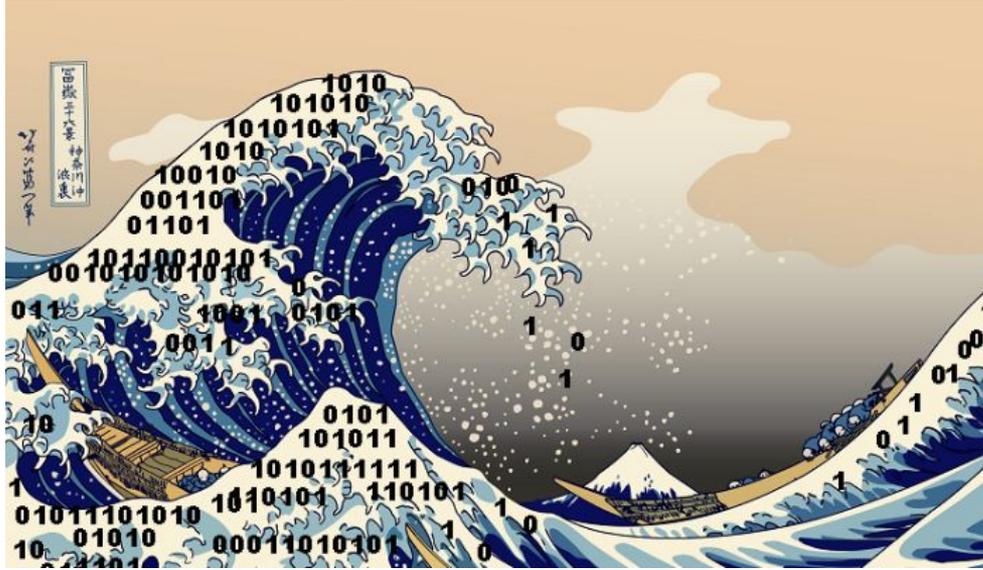


- Utiliser les outils
- Construire son propre workflow
- Pratiquer sur un même cas d'utilisation : Appel de SNPs à partir de reads Illumina de 3 échantillons



The Great Wave off Kanagawa, Hokusai @amitechsolutions.com





Créer son propre pipeline via une méthode facile et conviviale

Données brutes



Résultats
Intermédiaires



Résultats
Intermédiaires



Résultat
Final

- 3 solutions proposées par **SouthGreen** bioinformatics platform

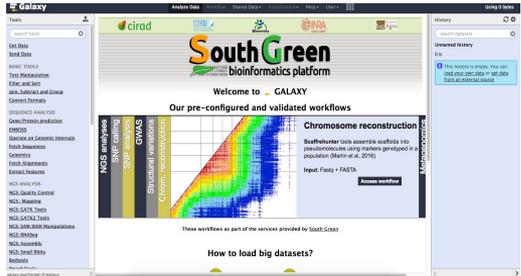
GUI tools



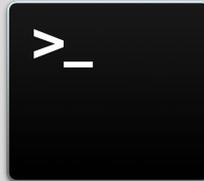
CLI tools



Galaxy



TOGLe



Snakemake



GUI tools



CLI tools



Facilité d'utilisation
Bonne documentation

genome assembly **SNP detection**
phylogeny **structural variation**
comparative genomics transcriptome assembly differential expression
GWAS pangenomics
population genetics metagenomics
polyploidy

Facilité de
développement

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)

Contrôle du pipeline
et des données

Reproductibilité
& traçabilité

**Apporte un cadre
robuste**

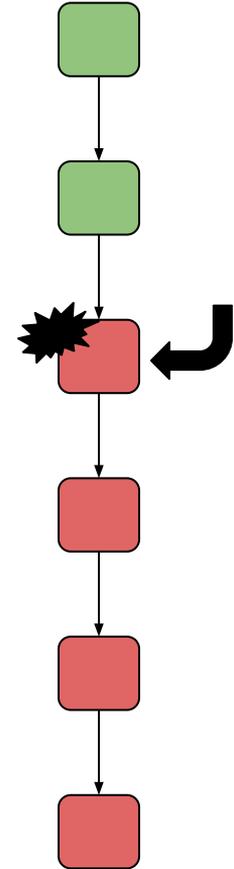
Sauvegarde des options, version des logiciels,
partage des analyses

Contrôle du pipeline
et des données

Reproductibilité
& traçabilité

Apporte un cadre
robuste

Suivi des erreurs
& reprise en cours



Contrôle du pipeline
et des données

Reproductibilité
& traçabilité

**Apporte un cadre
robuste**

Suivi des erreurs
& reprise en cours



Analyse de gros
jeu de données

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

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

Introduction au gestionnaire de
workflow:



- 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/>



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



This Galaxy instance has been configured such that only users who are logged in may use it.

Login

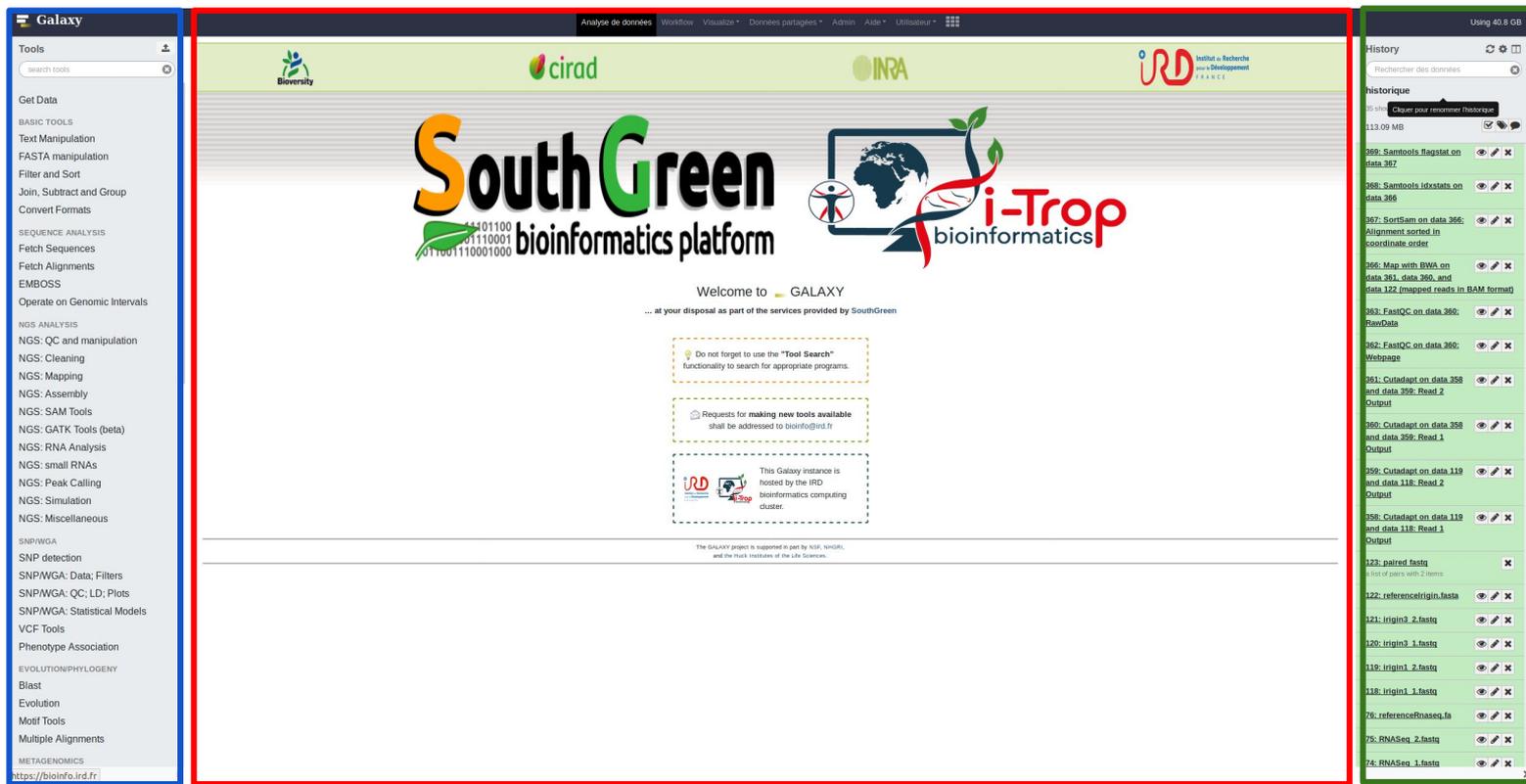
Username / Email Address:

Password:

Forgot password? [Reset here](#)

Login

Interface principale



The screenshot displays the main interface of the South Green bioinformatics platform. It is divided into three main sections:

- Left Panel (Outils):** A sidebar containing a search bar and a list of tool categories such as "Get Data", "SEQUENCE ANALYSIS", "NGS ANALYSIS", "SNP/WGA", "EVOLUTION/PHYLOGENY", and "METAGENOMICS".
- Central Panel (Panel principal):** The main dashboard area featuring logos for "Biodiversity", "cirad", "INRA", and "IRD Institut de Recherche pour le Développement FRANCE". It includes the "South Green bioinformatics platform" and "i-Trop bioinformatics" branding. A central message reads: "Welcome to GALAXY ... at your disposal as part of the services provided by SouthGreen". Below this are three informational boxes: one about "Tool Search" functionality, one about requests for new tools, and one stating the platform is hosted by the IRD cluster.
- Right Panel (historique):** A "History" sidebar showing a list of recent jobs with details like "369: Samtools flagstat on data 367" and "368: Samtools idxstats on data 368".

outils

Panel principal

historique

Panel supérieur



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

→ 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)

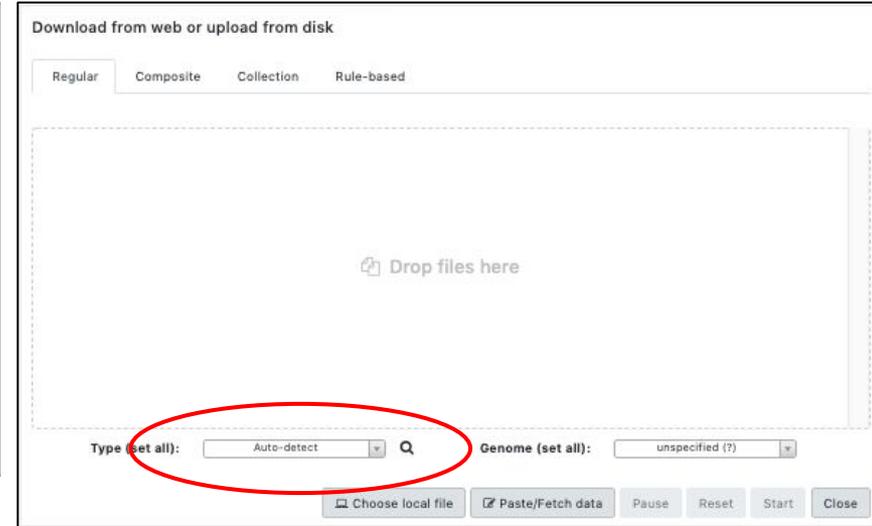
→ Import depuis l'ordinateur ou un site externe

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar is visible, with 'Upload File from your computer' circled in red. The main area shows a 'Download from web or upload from disk' dialog box. This dialog has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. It features a large dashed box with the text 'Drop files here'. Below this box are two dropdown menus: 'Type (set all):' set to 'Auto-detect' and 'Genome (set all):' set to 'unspecified (?)'. At the bottom of the dialog, the 'Choose local file' and 'Paste/Fetch data' buttons are circled in red. The background shows a Galaxy workflow with various tools and a history panel on the right.

Pour importer depuis
votre ordinateur

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

- 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*





→ 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:

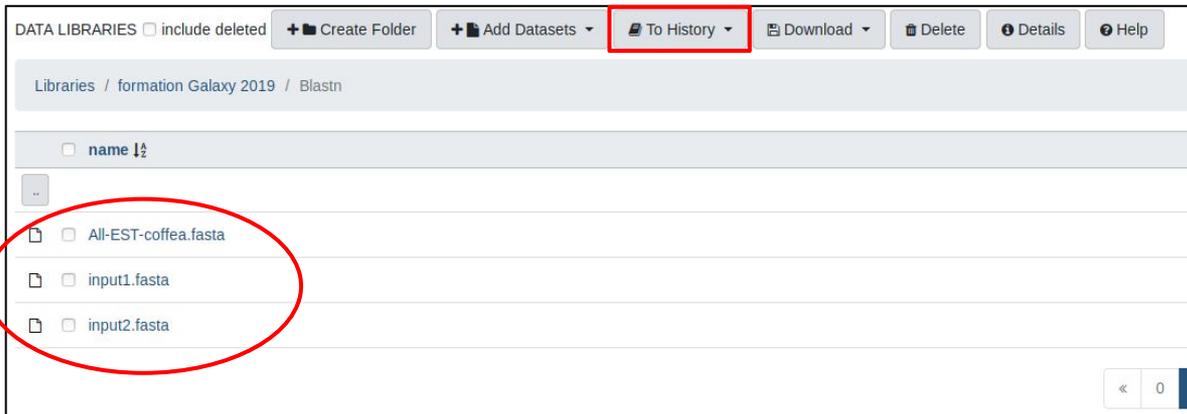
input1.fasta

input2.fasta

All-EST-coffee.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

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

Advanced Search

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





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

**Cherchez
BLASTN**

Tools

blastn

NGS: Mapping

Megablast compare short reads against htgs, nt, and wgs databases

Blast

NCBI BLAST+ makeprofiledb Make profile database

NCBI BLAST+ tblastn Search translated nucleotide database with protein query sequence(s)

NCBI BLAST+ rpstblastn Search protein domain database (PSSMs) with translated nucleotide query sequence(s)

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)

Metagenomic analyses

MetaPhlan metagenomic profiler

MetaPhlan Metagenomic Phylogenetic Analysis

FROGS

OTUS RECONSTRUCTION

FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

Vizualisation

JBrowse genome browser

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 0.3.1)

Versions Options

Nucleotide query sequence(s)

1: input1.fasta (-query)

Subject database/sequences

FASTA file from your history (see warning note below)

Nucleotide FASTA subject file to use instead of a database

3: All-EST-coffee.fasta (-subject)

Type of BLAST

- megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
- blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
- blastn-short - BLASTN program optimized for sequences shorter than 50 bases
- dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

(-task)

Set expectation value cutoff

0.001 (-evalue)

Output format

Tabular (standard 12 columns) (-outfmt)

Advanced Options

Hide Advanced Options

Execute

1) Lancez BLASTN avec les paramètres suivants :

- Query = input1.fasta
- Banque Fasta de l'historique = all-EST-cofea.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



Nom automatique

Modifier nom / extension

Voir les données

Supprimer

Relancer blastn avec les mêmes paramètres



Télécharger

```
34: blastn input1.fasta vs 'All-EST-coffea.fasta'
2,897 lines
format: tabular, database: ?
gi|33391745|gb|AY273814.1| gi|31580930
gi|33391745|gb|AY273814.1| gi|82472623
gi|33391745|gb|AY273814.1| gi|31585146
gi|33391745|gb|AY273814.1| gi|31575874
gi|33391745|gb|AY273814.1| gi|31118113
```

aperçu



34: blastn
input1.fasta vs 'All-EST-cofea.fasta'

This job is currently running

34: blastn input1.fasta vs 'All-EST-cofea.fasta'

2,897 lines
format: tabular, database: ?

1	2
gi 33391745 gb AY273814.1	gi 31580930
gi 33391745 gb AY273814.1	gi 82472623
gi 33391745 gb AY273814.1	gi 31585146
gi 33391745 gb AY273814.1	gi 31575874
gi 33391745 gb AY273814.1	gi 31118113

- 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
---------------------------	-------------------------------------	--------	-----	----	---	-----	-----	---	-----	-----	------

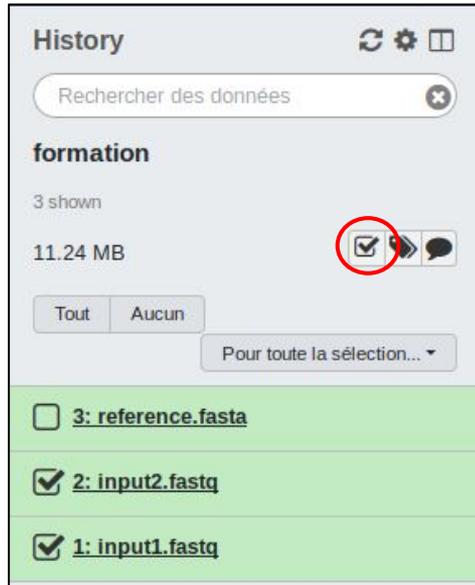
- 1) Lancez BLASTN pour obtenir un fichier tabulé
- 2) Lancez FILTER et ne gardez que les lignes ou le pourcentage d'identité est de plus de 98
- 3) Lancez CUT et ne gardez que les "subject ID" sur le fichier précédent

- 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

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



History

Rechercher des données

formation

3 shown

11.24 MB

Tout Aucun

Pour toute la sélection...

- 3: [reference.fasta](#)
- 2: [input2.fastq](#)
- 1: [input1.fastq](#)




History

Rechercher des données

formation

3 shown

11.24 MB

Tout Aucun

Pour toute la sélection...

- Cacher les jeux de données
- Afficher les jeux de données cachés
- Supprimer les jeux de données
- Restaurer les jeux de données supprimés
- Supprimer définitivement les jeux de données
- Build Dataset List**
- Build Dataset Pair
- Build List of Dataset Pairs
- Build Collection from Rules



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**

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Version 0.3.1) [Versions](#) [Options](#)

Nucleotide query sequence(s)

   No fasta or fasta.gz dataset available. 

(-query)

Créer un nouveau
Workflow

Analyse de données **Workflow** Visualize ▾ Données partagées ▾ Admin Aide ▾ Utilisateur ▾

Your workflows

search for workflow...

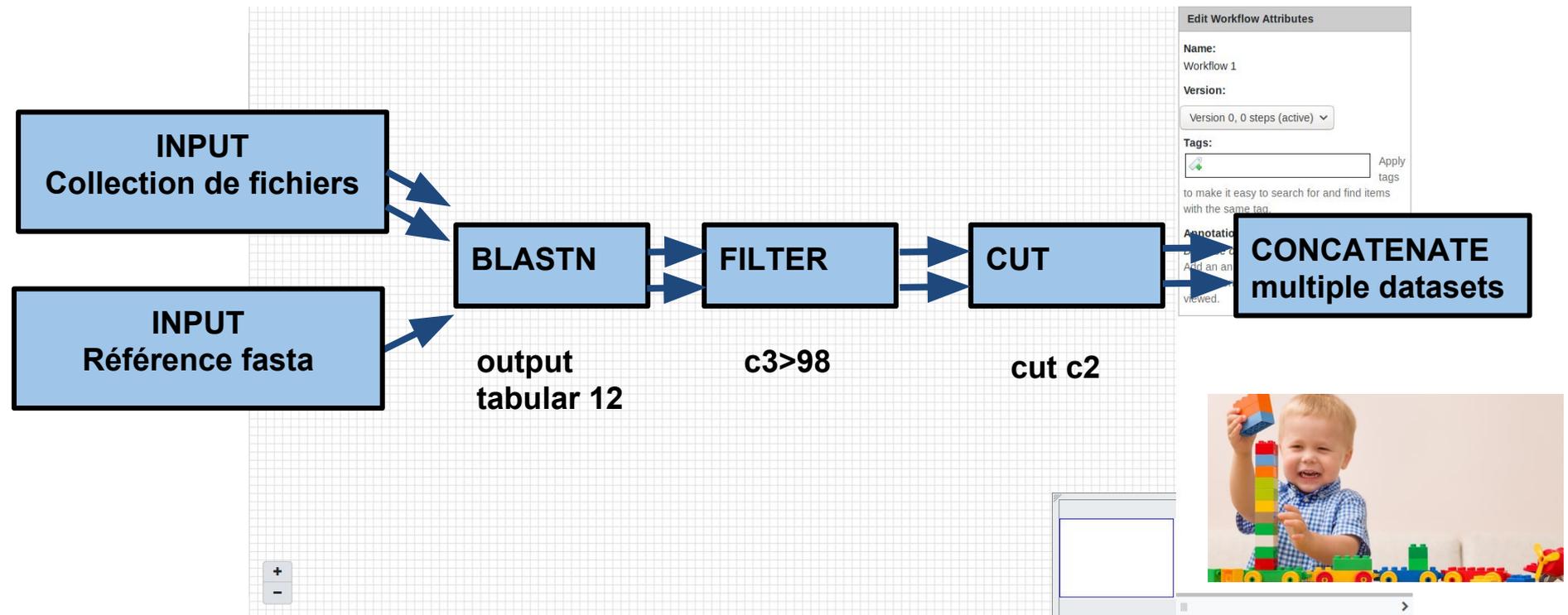


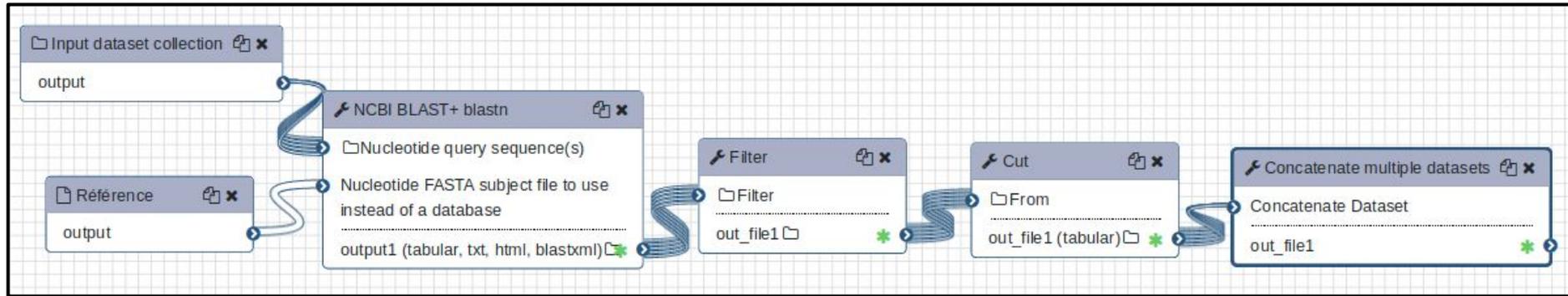
Name	Tags	Owner	# of Steps	Published	Show in tools panel
BlastN et tri ▾		You	6	Yes	<input type="checkbox"/>
SNP Calling - multi ▾		You	8	Yes	<input type="checkbox"/>
SNPCalling - mapping - multi ▾		You	9	Yes	<input type="checkbox"/>
SNP Calling ▾		You	8	No	<input type="checkbox"/>
kallistoEdgeR ▾		You	7	Yes	<input type="checkbox"/>
imported: Virus_haplotype_network ▾		You	4	No	<input type="checkbox"/>

Liste de mes workflows



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





TIPS:

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

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)

Your workflows

search for workflow... + ↓

Name	Tags	Owner	# of Steps	Published	Show in tools panel
Workflow 1		You	5	No	<input type="checkbox"/>
SNPCalling - mapping - multi		You	9	No	<input type="checkbox"/>
SNP Calling - multi		You	8	No	<input type="checkbox"/>
SNPCalling - mapping		You	10	Yes	<input type="checkbox"/>
SNP Calling		You	8	Yes	<input type="checkbox"/>
kallistoEdgeR		You	7	Yes	<input type="checkbox"/>
imported: Virus_haplotype_net		You	4	No	<input type="checkbox"/>

Edit

Run

Share

Download

Copy

Rename

View

Delete

Partager un workflow (ou un historique)

- Edit
- Run
- Share**
- Download
- Copy
- Rename
- View
- Delete

Share

This workflow is currently restricted so that only you and the users listed below can access it.

Make Workflow Accessible via Link

Generates a web link that you can share with other people so that they can view and import the workflow.

Make Workflow Accessible and Publish

Makes the workflow accessible via link (see above) and publishes the workflow to Galaxy's Published Workflows section, where it is publicly listed and searchable.

You have not shared this workflow with any users yet

Share with a user

A une ou plusieurs personnes en particulier

Public à tous les utilisateurs

Export

Download workflow as a file so that it can be saved or imported into another Galaxy server.

This workflow must be accessible. Please use the option above to "Make Workflow Accessible and Publish" before receiving a URL for importing to another Galaxy.

Create image of workflow in SVG format

Export to the www.myexperiment.org site.

myExperiment username:

myExperiment password:

Export to myExperiment

Télécharger et exporter vers un site externe

SNP-Calling avec:

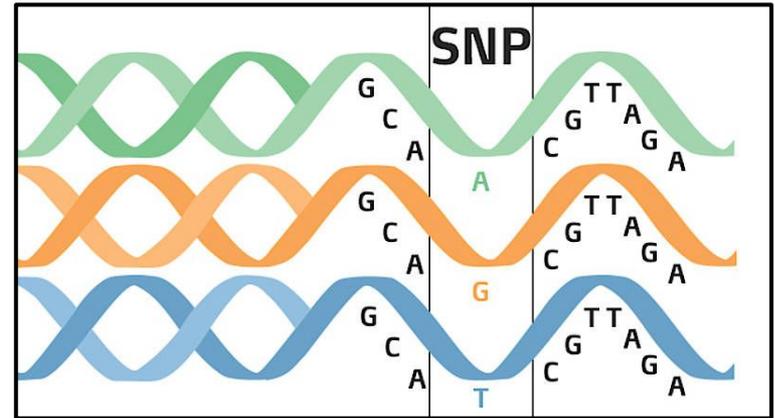


Objectifs :

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)

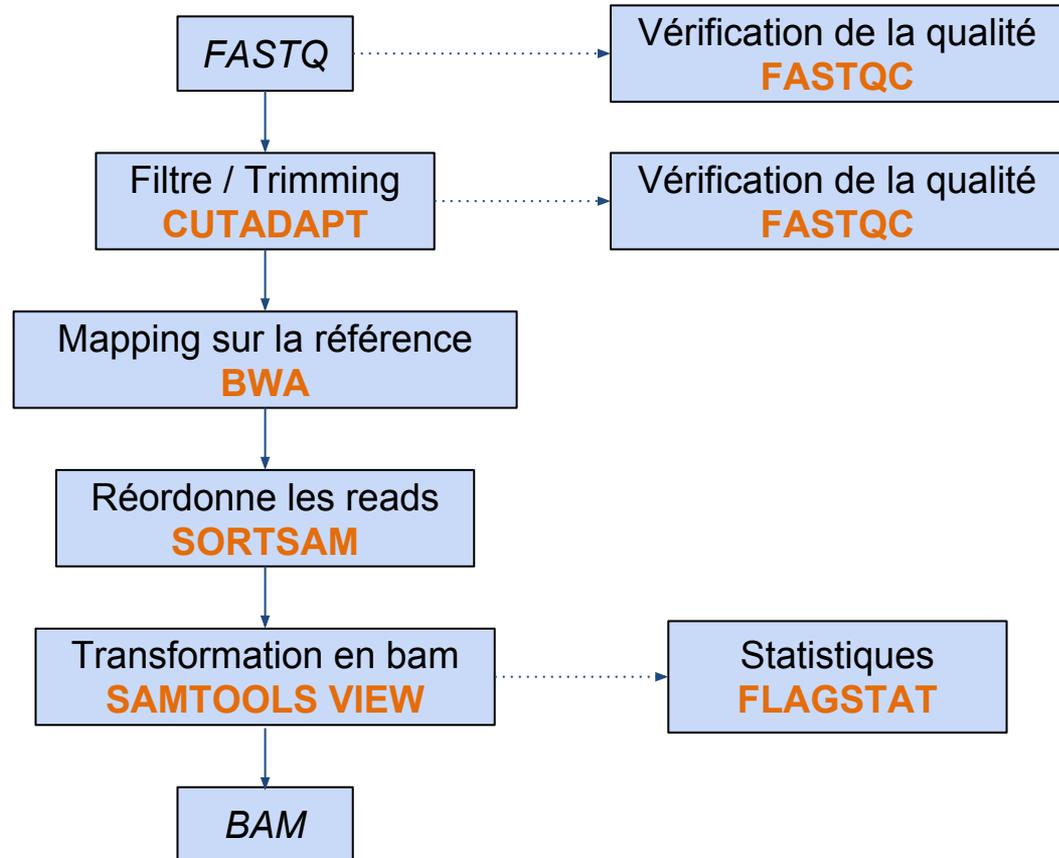


Le format FastQ

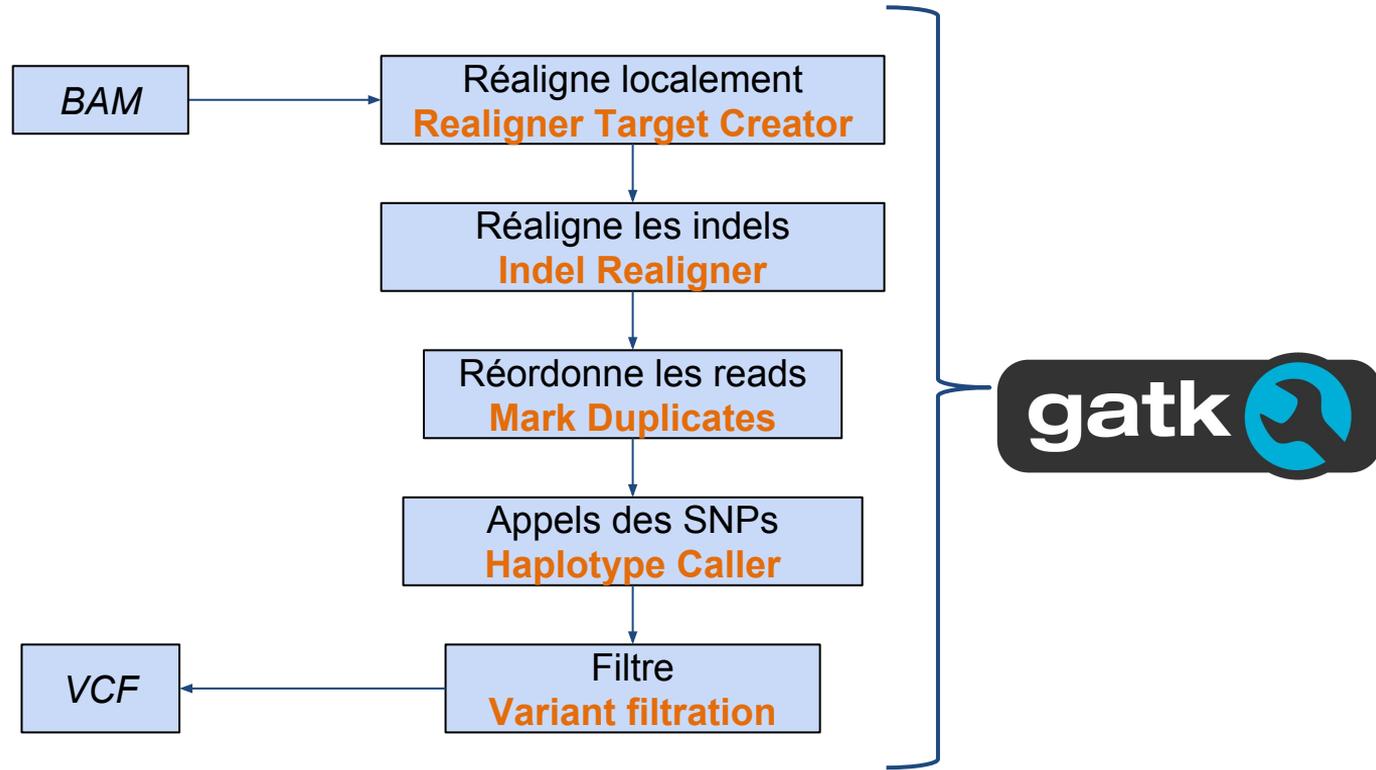
Format concis et compact qui stocke à la fois séquence et qualité de séquençage.

Identifiant	●	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Séquence	●	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	●	+
Qualité scores	●	hhhhhhhhhhghhghhhhhfhhhhhhfffffe'ee['X]b[d[ed'[Y[~Y
Identifiant	●	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Séquence	●	GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	●	+
Qualité scores	●	hhhhgfhhcghghggfcffdhf ehhhhcehdchhdhahehffffde'bVd

1) Mapping

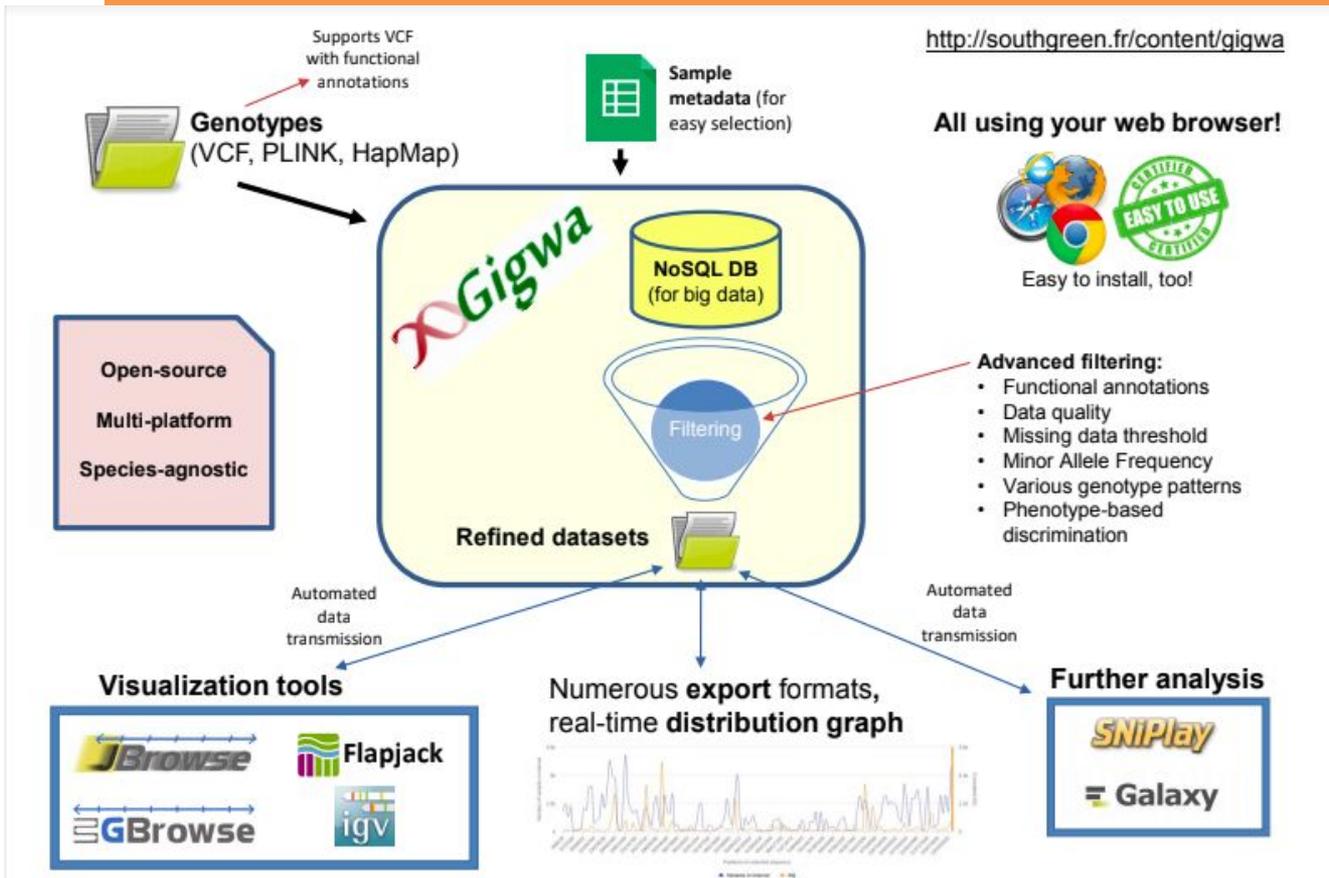


2) SNP calling



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
```



→ 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.

DATA LIBRARIES include deleted + Create Folder + Add Datasets To History Download Delete Details Help

Libraries / formation Galaxy 2019 / SNPCalling

<input type="checkbox"/> name ↕	description	data type	size	time updated (UTC)	state
<input type="checkbox"/> RC1_raw_R1.fastq		fastqsanger	1.6 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> RC1_raw_R2.fastq		fastqsanger	1.6 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> RC2_raw_R1.fastq		fastqsanger	1.5 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> RC2_raw_R2.fastq		fastqsanger	1.5 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> RC3_raw_R1.fastq		fastqsanger	1.6 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> RC3_raw_R2.fastq		fastqsanger	1.6 MB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage
<input type="checkbox"/> Reference.fasta		fasta	6.5 KB	2019-04-11 08:06 AM	<input type="checkbox"/> Manage



1) Créer une collection pour des données pairées

Pour toute la sélection... ▾

- Cacher les jeux de données
- Afficher les jeux de données cachés
- Supprimer les jeux de données
- Restaurer les jeux de données supprimés
- Supprimer définitivement les jeux de données
- Build Dataset List
- Build Dataset Pair
- Build List of Dataset Pairs**
- Build Collection from Rules

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

3 unpaired reverse - (3 filtered out)

R1

Auto-pair

R2

- RC1_raw_R1.fastq
- RC2_raw_R1.fastq
- RC3_raw_R1.fastq

- Pair these datasets
- Pair these datasets
- Pair these datasets

- RC1_raw_R2.fastq
- RC2_raw_R2.fastq
- RC3_raw_R2.fastq



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 CAAGCAGAAGACGGCATA CGA

Adapter list

Nothing selected

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

Submodule and Limit specifying 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

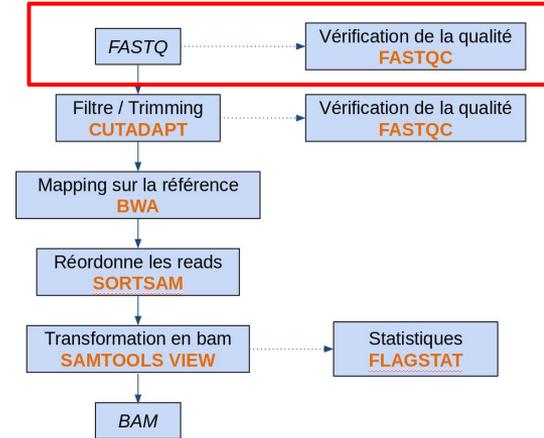
Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)

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 comparable statistics from datasets with somewhat variable read lengths. (--min_length)

length of Kmer to look for

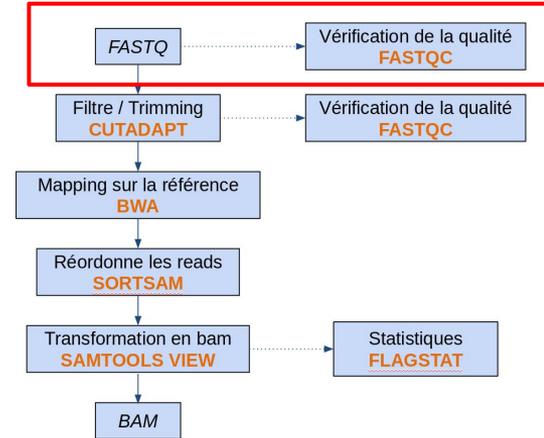
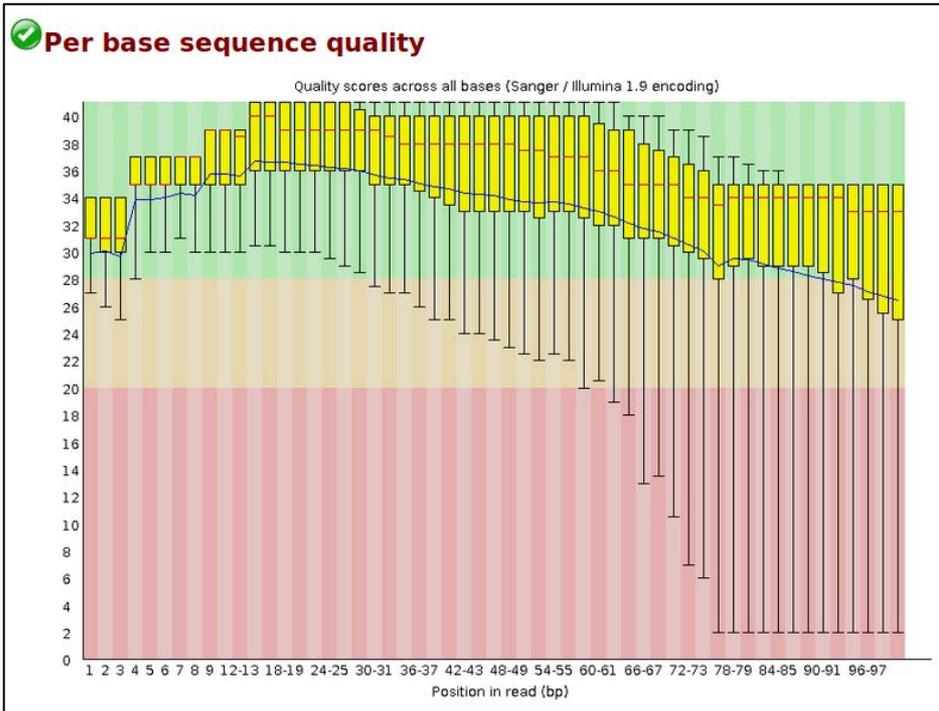
note: the Kmer test is disabled and needs to be enabled using a custom Submodule and limits file (--kmers)





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





2) Cutadapt : Remove adapter sequences from Fastq/Fasta

Paramètres :

Adapter Option:

Minimum overlap length = 7

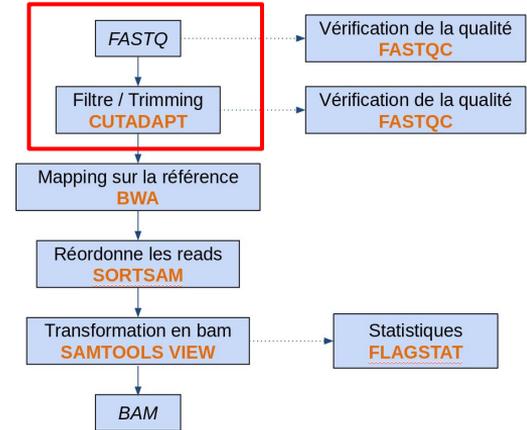
Filter Option:

minimum-length=35

Read Modification Options:

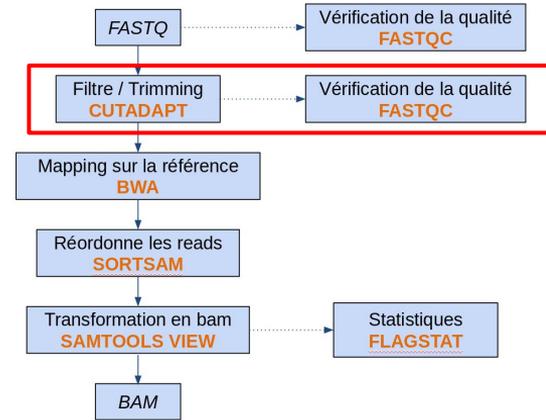
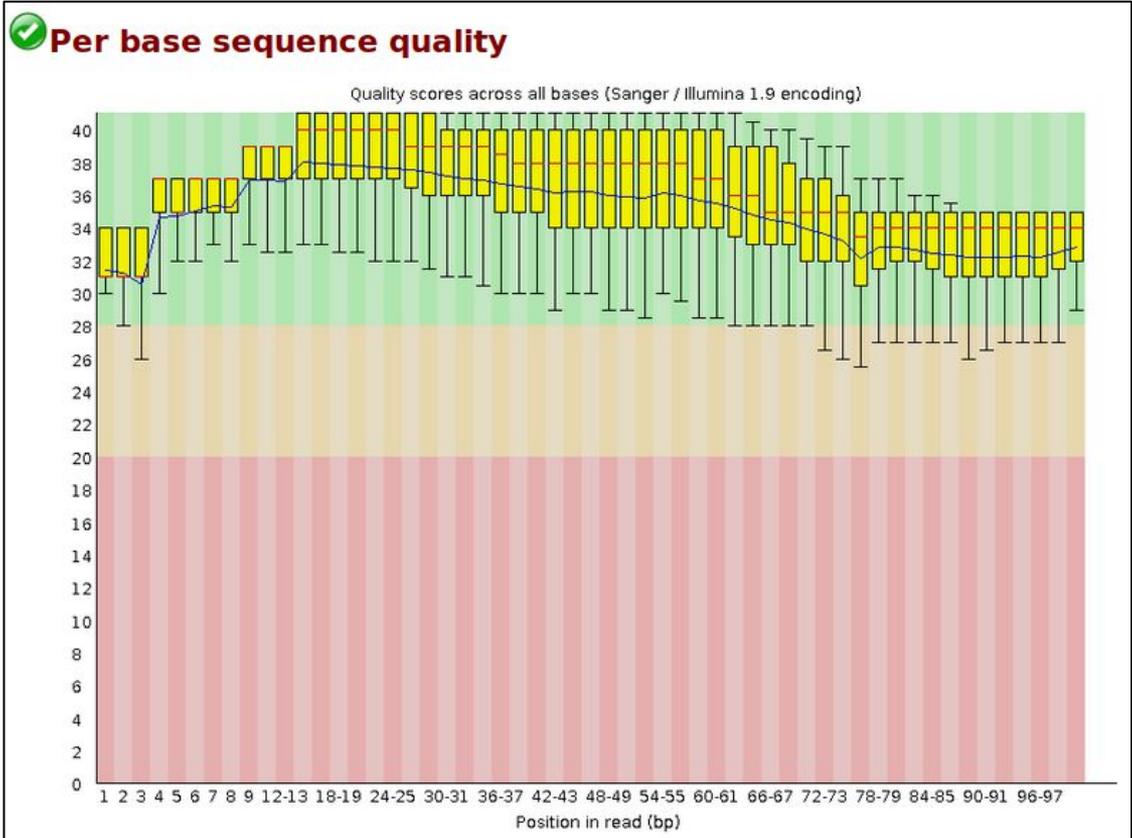
Quality cutoff=20,20

→ reads fastq filtrés en fonction de la qualité





3) FastQC : Read Quality reports





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

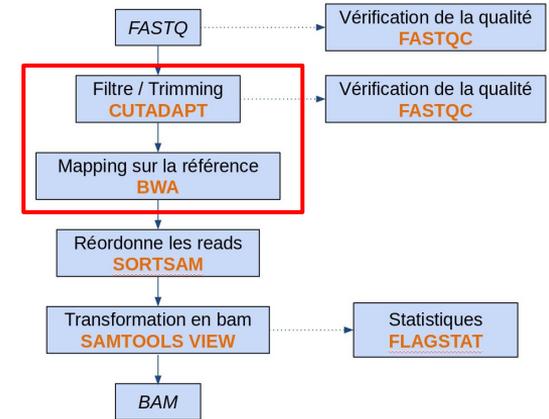
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Specify dataset with reverse reads

Set advanced paired end options?

Do not set

Provides additional controls



→ Collection de 3 bam



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 elapsed.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forward indices...

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

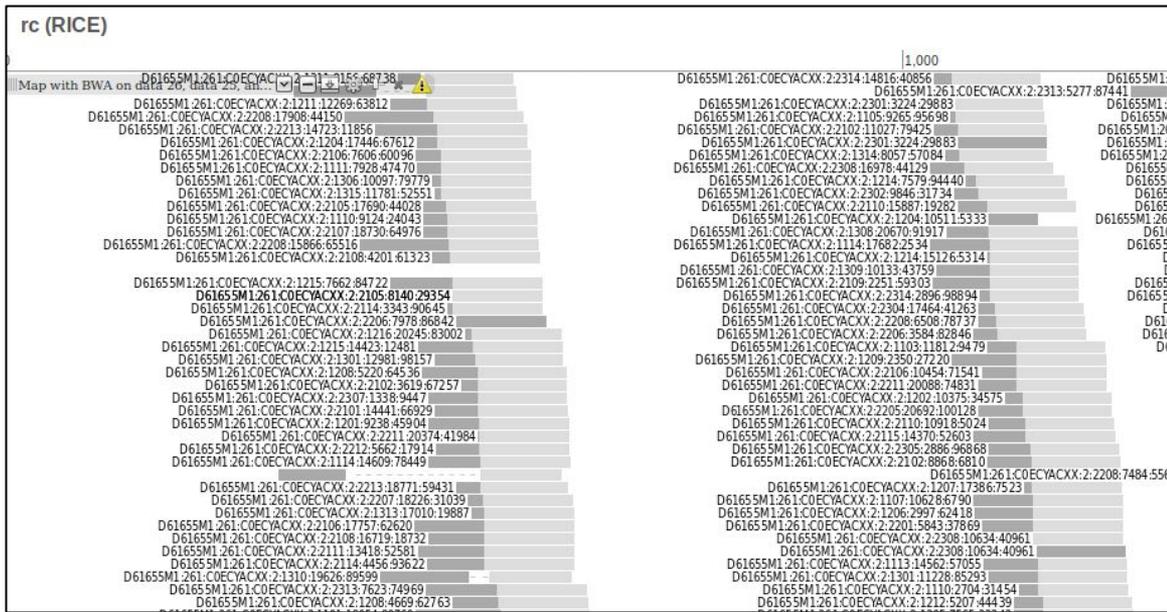
Binary bam alignments file

RC3_raw

RC2_raw

Visualisation du mapping dans Trackster

Trackster
Fast, interactive visualization for large, NGS/HTS datasets using only a web browser.





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

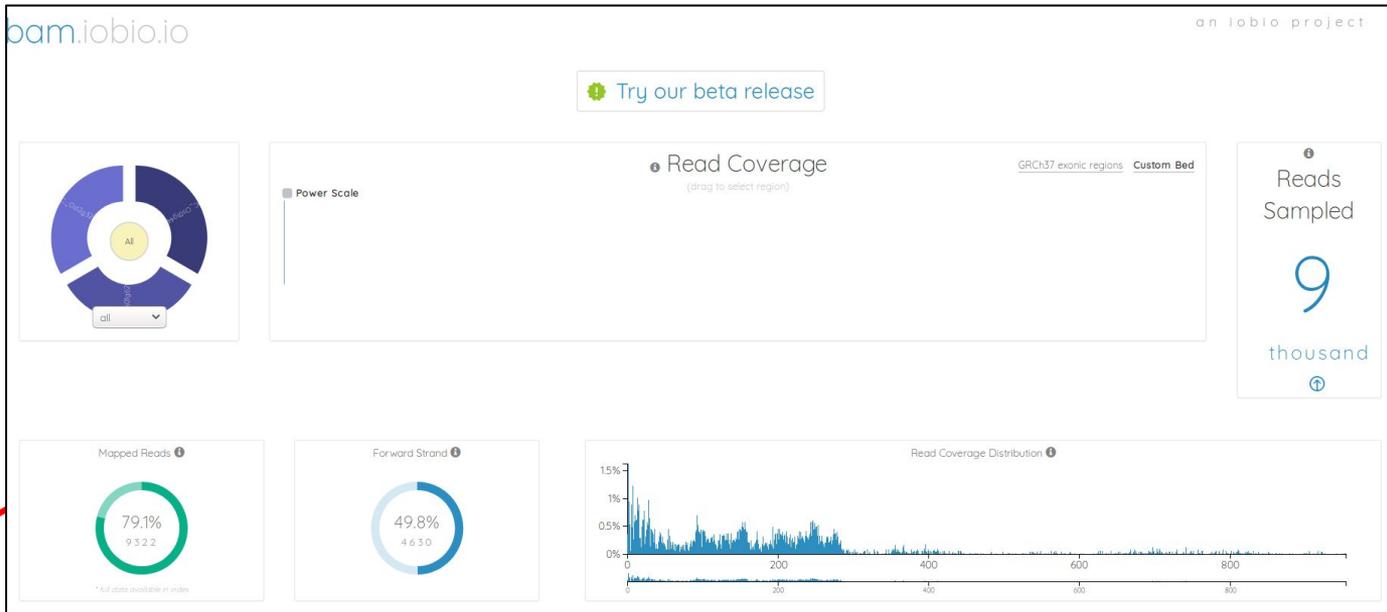
      

display with IGV local
display in IGB View
display at bam.iobio bam.iobio.io

Binary bam alignments file

RC3_raw  

RC2_raw  



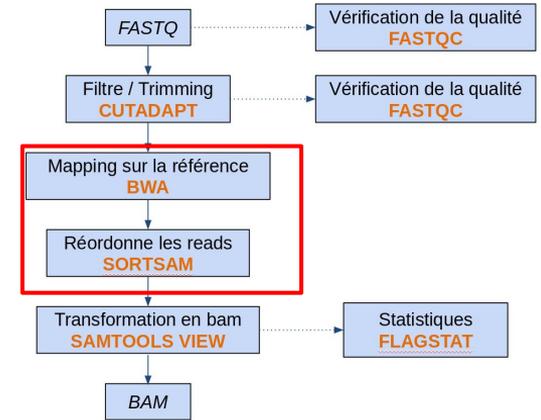


5) SortSAM : sort SAM/BAM dataset

Paramètres :

Sort order = coordinate

Select validation Stringency = Silent



→ Collection de 3 bam ordonnés

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



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

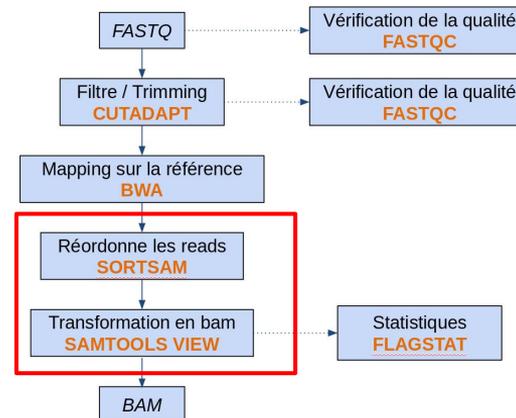
Only output alignments with number of CIGAR bases consuming query sequence greater than or equal INT. (-m)

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



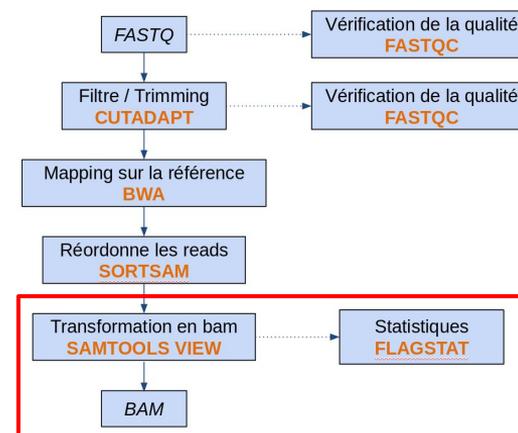
7) Samtools Flagstat : tabulate descriptive stats for BAM dataset

BAM File to report statistics of

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

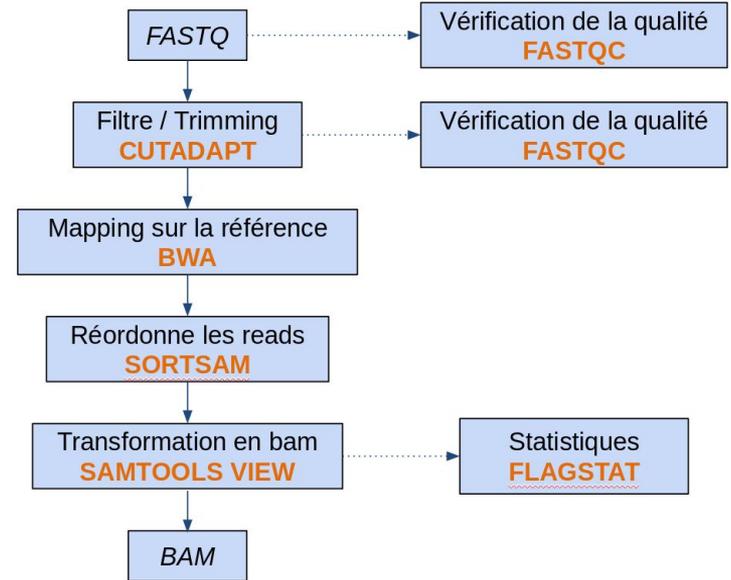
Execute

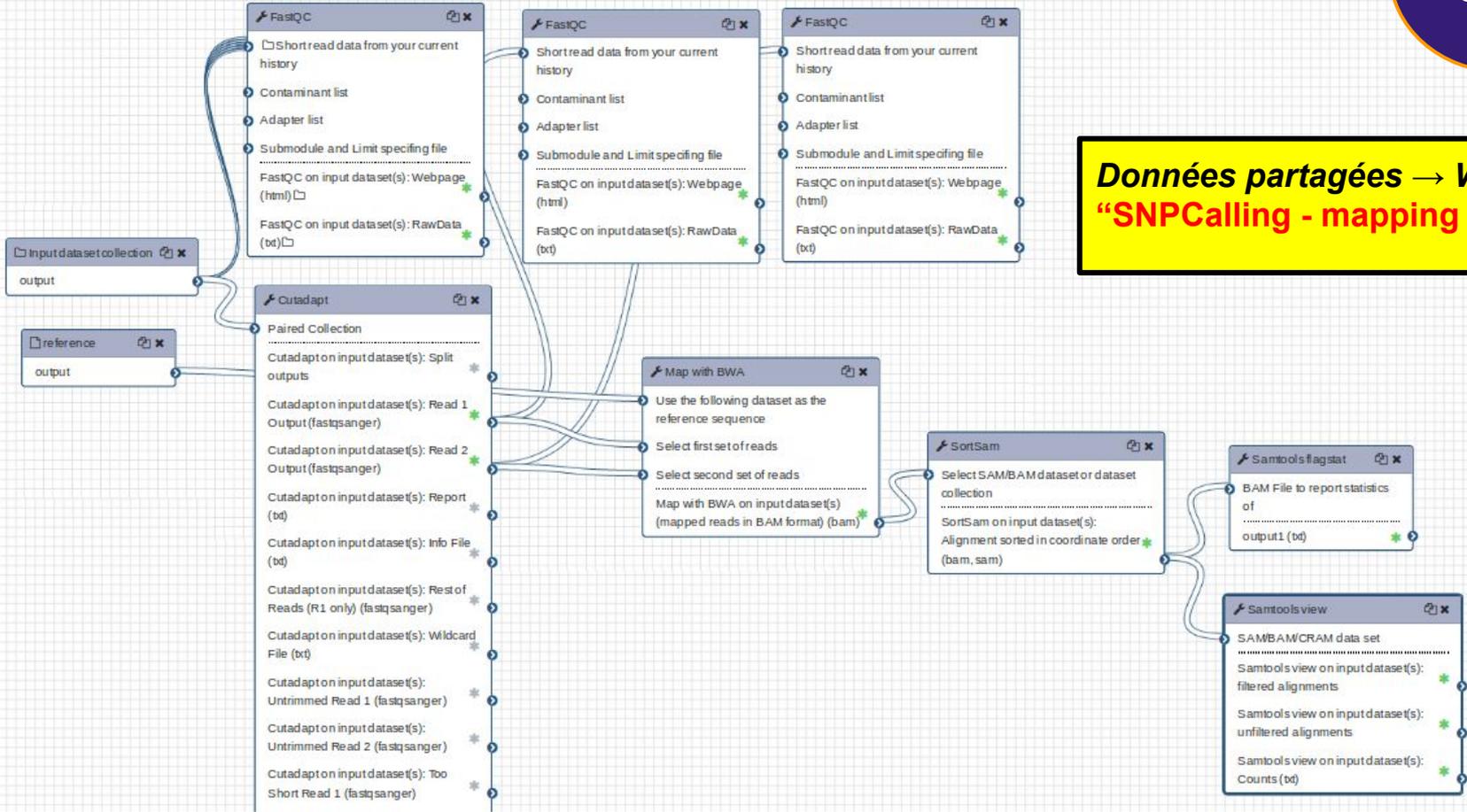
```
11812 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
9322 + 0 mapped (78.92% : N/A)
11812 + 0 paired in sequencing
5906 + 0 read1
5906 + 0 read2
9078 + 0 properly paired (76.85% : N/A)
9292 + 0 with itself and mate mapped
30 + 0 singletons (0.25%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





1) CONSTRUISEZ LE WORKFLOW MAPPING

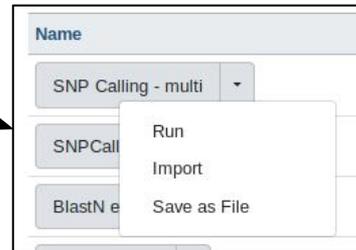
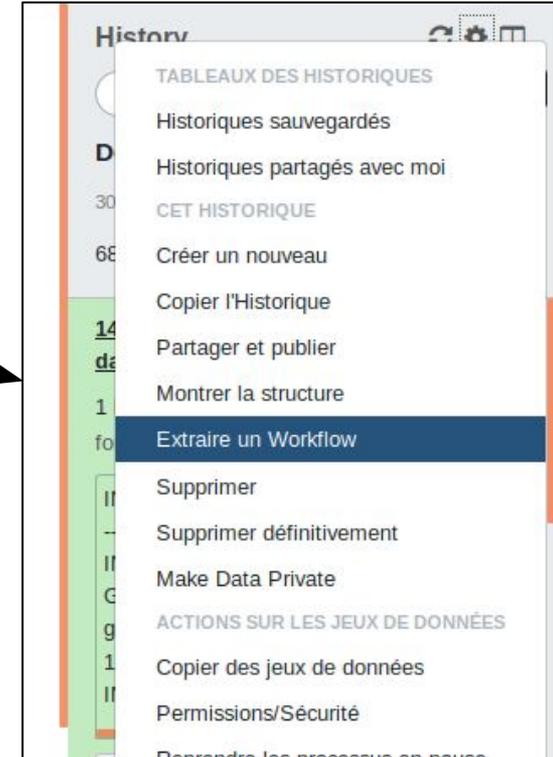




Données partagées → WF
“SNPCalling - mapping - multi”

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 !



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

Tools

search tools

Get Data

BASIC TOOLS

Text Manipulation

FASTA manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

SEQUENCE ANALYSIS

Fetch Sequences

Fetch Alignments

EMBOSS

Operate on Genomic Intervals

NGS ANALYSIS

NGS: QC and manipulation

NGS: Cleaning

NGS: Mapping

NGS: Assembly

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: RNA Analysis

NGS: small RNAs

NGS: Peak Calling

NGS: Simulation

Workflow: SNP Calling - multi Run workflow

History Options

Send results to a new history

Yes No

1: Input dataset collection

227: MarkDuplicates on collection 194: MarkDuplicates BAM output

2: reference

161: reference.fasta

3: Realigner Target Creator (Galaxy Version 2.8.1)

4: Indel Realigner (Galaxy Version 2.8.1)

5: MarkDuplicates (Galaxy Version 2.18.2.0)

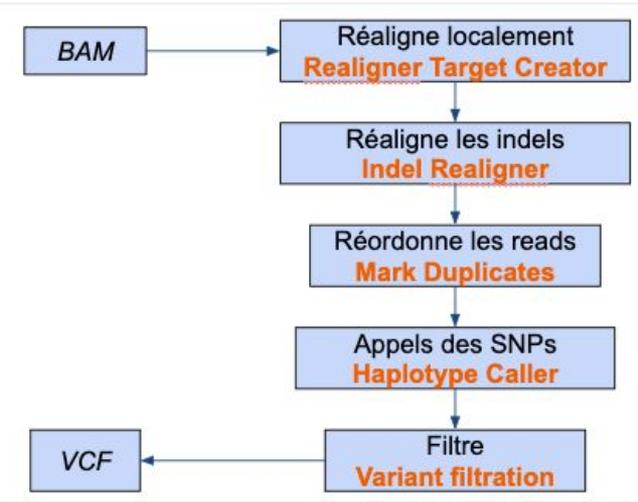
6: Haplotype Caller (Galaxy Version 2.8.2)

Covariates table recalibration file

No gatk_report dataset available.

The input covariates table file which enables on-the-fly base quality score recalibration. Enables on-the-fly recalibrate of base qualities. The covariates tables are produced by the BaseQualityScoreRecalibrator tool. Please be aware that one should only run recalibration with the covariates file created on the same input bam(s) (-BQSR,--BQSR <recal_file>)

Choose the source for the reference list





```
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared a
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for e
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for e
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=LOC_Os01g44110.1,length=2608>
##contig=<ID=LOC_Os01g62920.1,length=2879>
##contig=<ID=LOC_Os12g32240.1,length=1088>
##reference=file:///scratch2/galaxy/galaxy-19.01/galaxy/database/tmp/tmp-gatk-kyTSPp/gatk_input.fasta
#CHROM
LOC_Os01g44110.1
LOC_Os01g62920.1
LOC_Os12g32240.1
LOC_Os12g32240.1
LOC_Os12g32240.1
LOC_Os12g32240.1
LOC_Os12g32240.1
LOC_Os12g32240.1
```

History

Rechercher des données

Donnees alexis

30 shown, 61 deleted, 90 hidden

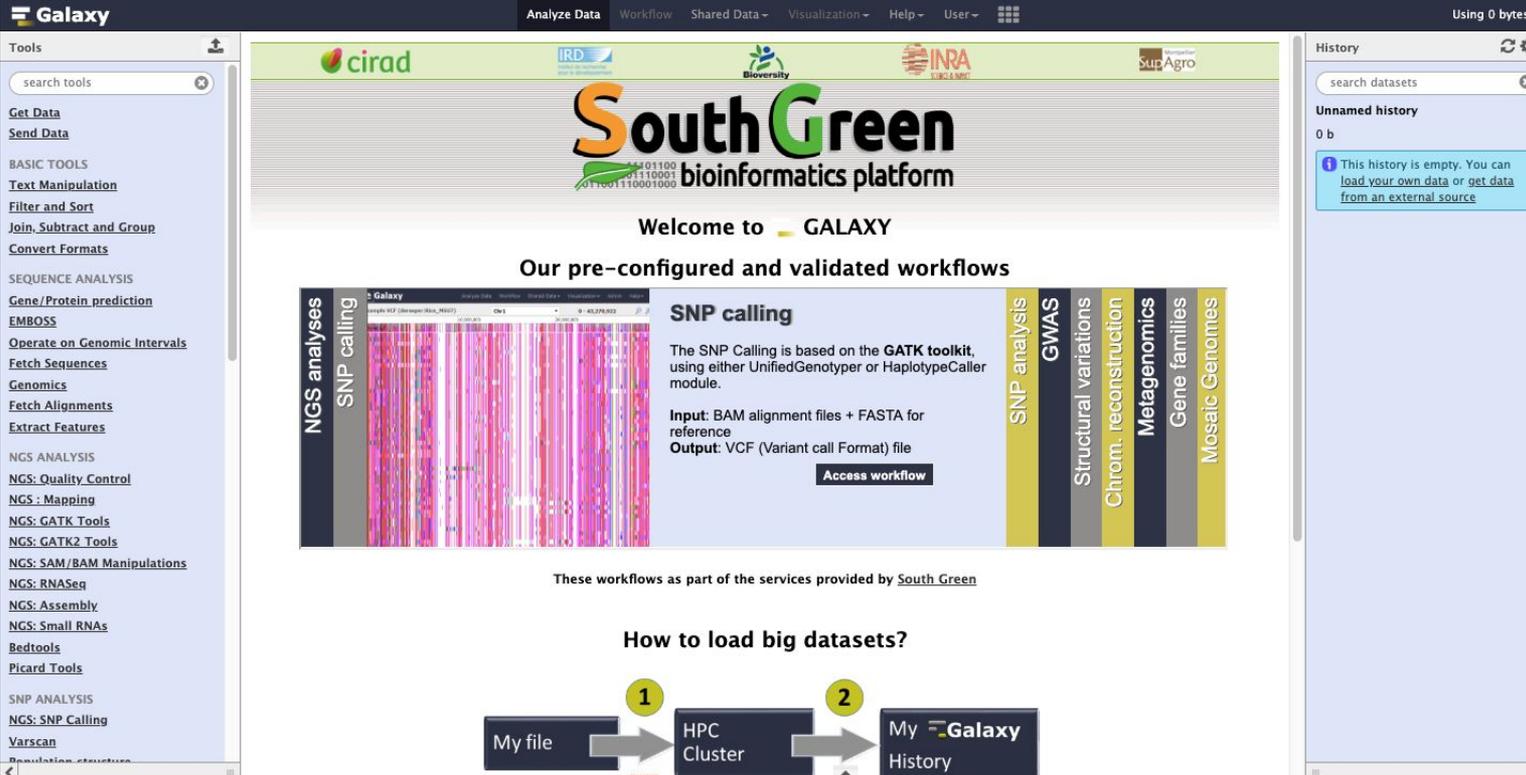
68.87 MB

- 144: Variant Filtration on data 7 and data 141 (log)**
👁️ ✎️ ✕
- 143: Variant Filtration on data 7 and data 141 (Variant File)**
👁️ ✎️ ✕
- 142: Haplotype Caller on data 7, data 140, and others (log)**
👁️ ✎️ ✕
- 141: Haplotype Caller on data 7, data 140, and others (VCF)**
👁️ ✎️ ✕
- 134: MarkDuplicates on collection 125: MarkDuplicates BAM output**
✕

a list with 3 items
- 133: MarkDuplicates on collection 125: MarkDuplicate metrics**
✕

a list with 3 items

> 9 workflows préconfigurés et validés par la plateforme



The screenshot displays the Galaxy web interface. At the top, there are navigation tabs: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The main content area features the SouthGreen logo and a 'Welcome to GALAXY' message. Below this, a section titled 'Our pre-configured and validated workflows' lists several categories: NGS analyses, SNP calling, SNP analysis, GWAS, Structural variations, Chrom. reconstruction, Metagenomics, Gene families, and Mosaic Genomes. The 'SNP calling' workflow is highlighted, with a description: 'The SNP Calling is based on the GATK toolkit, using either UnifiedGenotyper or HaplotypeCaller module.' It specifies 'Input: BAM alignment files + FASTA for reference' and 'Output: VCF (Variant call Format) file'. A 'Access workflow' button is visible. Below the workflow list, a diagram titled 'How to load big datasets?' shows a flow: 'My file' (with a '1' in a circle) points to 'HPC Cluster', which then points to 'My Galaxy History' (with a '2' in a circle).

Tools

search tools

Get Data
Send Data

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

SEQUENCE ANALYSIS
Gene/Protein prediction
EMBOSS
Operate on Genomic Intervals
Fetch Sequences
Genomics
Fetch Alignments
Extract Features

NGS ANALYSIS
NGS: Quality Control
NGS: Mapping
NGS: GATK Tools
NGS: GATK2 Tools
NGS: SAM/BAM Manipulations
NGS: RNASeq
NGS: Assembly
NGS: Small RNAs
Bedtools
Picard Tools
SNP ANALYSIS
NGS: SNP Calling
Varscan

Analyze Data Workflow Shared Data Visualization Help User

Using 0 bytes

History
search datasets

Unnamed history
0 b

This history is empty. You can load your own data or get data from an external source

NGS analyses
SNP calling

SNP analysis
GWAS
Structural variations
Chrom. reconstruction
Metagenomics
Gene families
Mosaic Genomes

Welcome to GALAXY

Our pre-configured and validated workflows

SNP calling

The SNP Calling is based on the GATK toolkit, using either UnifiedGenotyper or HaplotypeCaller module.

Input: BAM alignment files + FASTA for reference
Output: VCF (Variant call Format) file

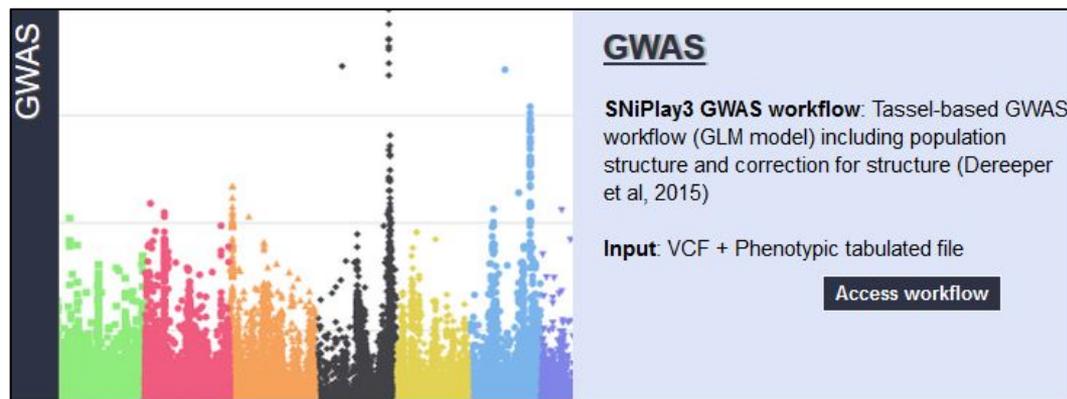
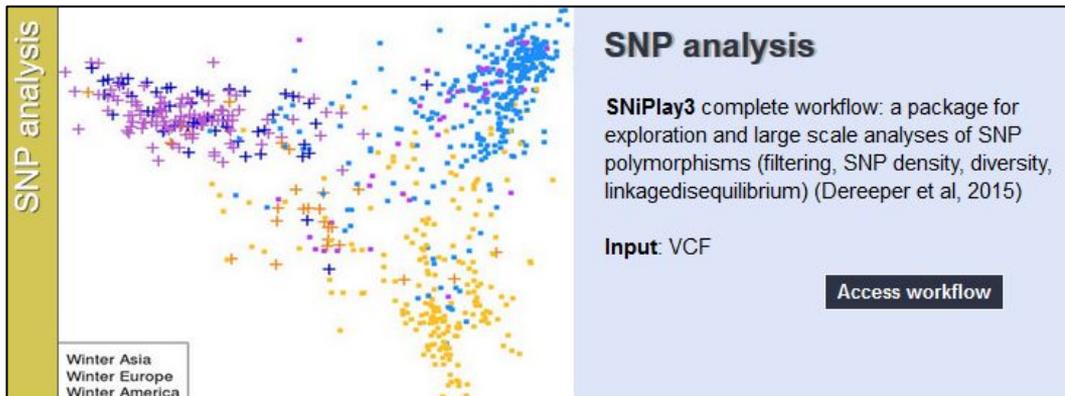
Access workflow

These workflows as part of the services provided by South Green

How to load big datasets?

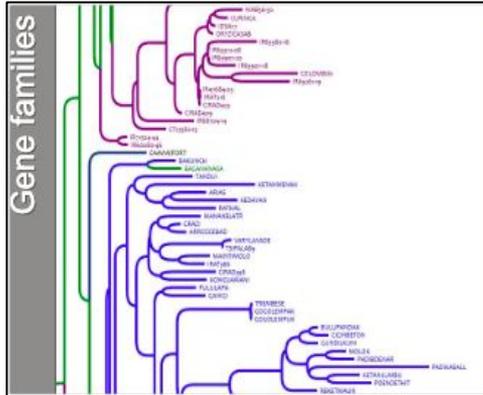
1 My file → HPC Cluster → 2 My Galaxy History

> 9 workflows préconfigurés et validés par la plateforme



> 9 workflows préconfigurés et validés par la plateforme

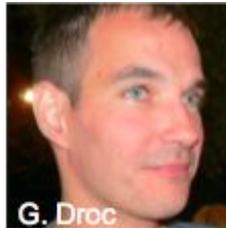
Gene families



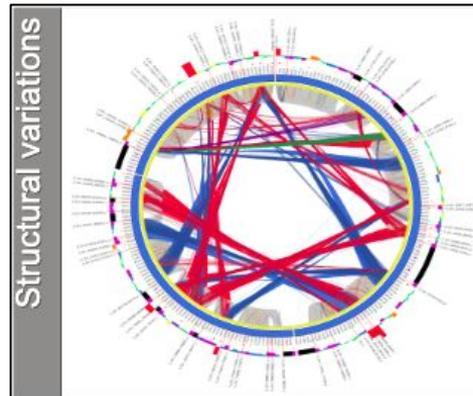
GreenPhyl / GenFam : comparative and functional genomics in plants (Rouard et al, 2011).

Input: FASTA file, Species tree file

[Access workflow](#)



Structural variations



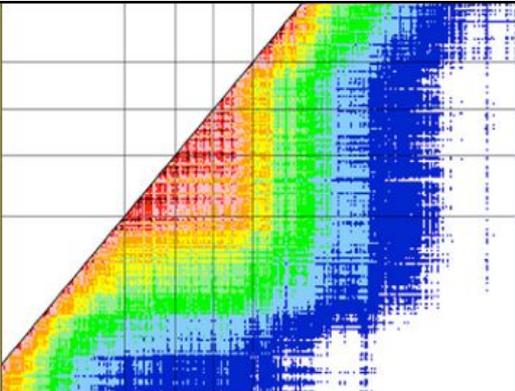
Scaffremodeler can be used to detect large structural variations between a reference sequence and a resequenced genome (Martin et al, 2016)

Input: Fastq + FASTA

[Access workflow](#)

> 9 workflows préconfigurés et validés par la plateforme

Chrom. reconstruction

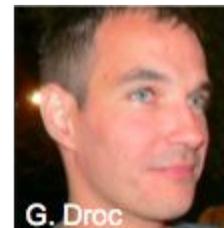


Chromosome reconstruction

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

Input: Fastq + FASTA

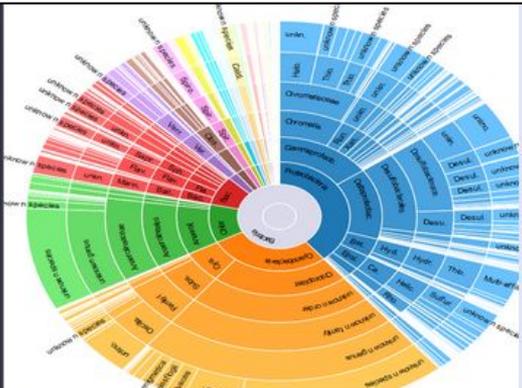
[Access workflow](#)



GenPhySE
 Génétique Physiologie et Systèmes d'Élevage



Metagenomics



Metagenomics

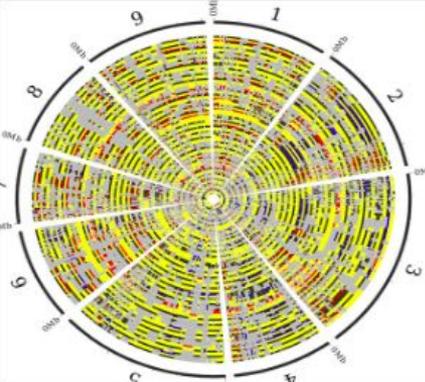
FROGS: Find Rapidly OTU with Galaxy Solution (Pascal et al, 2015)

Input: Fastq files

[Access workflow](#)

> 9 workflows préconfigurés et validés par la plateforme

Mosaic Genomes



Mosaic genome reconstruction

TraceAncestor / KDE_Classifier : Two approaches to analyze the mosaic structure of plant genomes

Input: VCF file + structure file

[Access workflow](#)



Une documentation est disponible dans les Galaxy pages

> Workflow / Jeux de données / Historique d'analyses



Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Pages | accounts | reconstruction of mosaic genomes

Reconstruction mosaic genomes and painting
Hybridization events between species and subspecies are widespread in crops. This admixture produces a mosaic structure of ancestral contributions.

1 - TRACEANCESTOR
TraceAncestor was developed on Citrus species.
The method is to estimate the frequency of ancestral alleles along the chromosome of one hybrid. Ancestral alleles are identified by GST preliminarily.

Workflow

TraceAncestor Pfilter (x)
main file
output (tabular)

TraceAncestor (x)
matrix table .i
vcf of hybrid population v
color file c
focus file (several hybrids).f
ideogram (tabular)
im_ideogram (tabular)
ideogram_allind (tabular)
im_allind_ideogram (tabular)
circos (tabular)
im_allind_Circos (tabular)
ancestralFreq (tabular)

Ideogram (x)
Values for Chromosome Length
Ancestral Blocks
output (img)

Circos (x)
Values for Chromosome Length
track 1 - Track data
output (img)

Galaxy_Workflow | TraceAncestor

Input files

vcf file :
Galaxy_Dataset | CitrusGalaxy.vcf

Matrix with GST values for each ancestor:
Galaxy_Dataset | MatriceDSNPs.csv

focus file of individuals we want to perform the painting on:
Galaxy_Dataset | focus_file

Output history
Galaxy_History | traceancestor

Tools

A- TraceAncestor Pfilter
This tool is used to extract a list of specific SNPs.

B- TraceAncestor
Perform the analysis on the list of diagnosis SNPs and a vcf of individual hybrids.

Visualization:

A- Ideogram
Display an Ideogram visualization.

B- Circos
Display a Circos visualization.

B- KDE_CLASSIFIER

KDE_Classifier was developed on rice.
The method is based on a Kernel Density Estimation (KDE) with intermediate class and outlier class.
Due to the homozygote properties of the rice, the painting is done on one haplotype (for homozygote data).

Workflow
--> For several chromosomes

KDE (x)
chr1 input file
reference accessions indexes in gene file .ref
ancestral accession indexes in gene file .altref
turn (bin)

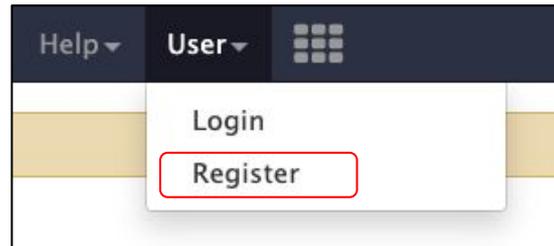
Ideogram (x)
reference file to read
ids of accessions to plot
chr1
output (img)

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

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

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?

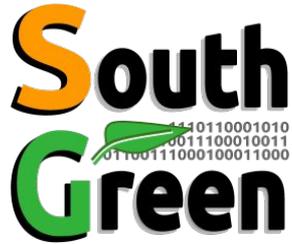
“The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://bioinfo.ird.fr/> ”

“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**





South Green : [@green_bioinfo](https://twitter.com/green_bioinfo)



I-Trop : [@ltropBioinfo](https://twitter.com/ltropBioinfo)



Galaxy Project : [@galaxyproject](https://twitter.com/galaxyproject)

Merci pour votre attention !



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:

<http://creativecommons.org/licenses/by-nc-sa/4.0/>