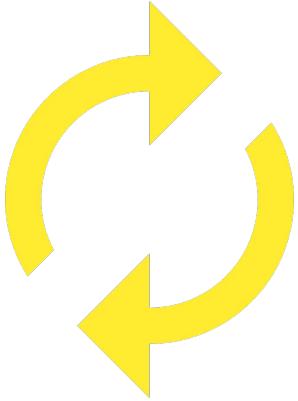




Session de formation 2019





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

comparative genomics
phylogenetics
GWAS
population genetics
polypliody

genome assembly
transcriptome assembly
metagenomics

SNP detection
structural variation
differential expression



Rice



Banana



Palm



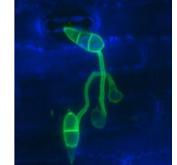
Sorghum



Coffee



Cassava



Magnaporthe

South Green

bioinformatics platform



Charte



Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant-Dubreuil
Christine



Comte Aurore
Dereeper Alexis



Orjuela-Bouniol Julie



Bocs Stephanie
De Lamotte Fredéric
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



South Green

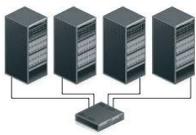
bioinformatics platform

Workflow manager

TOGGLE
Toolbox for generic NGS analyses



HPC and trainings....



Genome Hubs & Information System

Gigwa



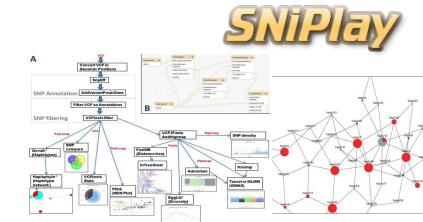
SNPs	Indels
100	100
200	200
300	300
400	400

SNPs and Indels

GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000010	Cytochrome P450 superfamily	6942	Green
GP000017	AP2/EREBP transcription factor family	5142	Green
GP000020	NAC transcription factor family	4574	Green
GP000028	MADS transcription factor family		
GP000018	Heme Peroxidase superfamily		
GP000021	General substrate transporter		
GP000022	Subtilisin-like Serine Proteases family		
GP000019	NPF/NRT1/PTP FAMILY		

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



Erwan Corre



Marie Simonin
Sébastien Cunnac



Etienne Loire
Julie Reveillaud



Florentin Constancias



Valentin Klein



Emmanuelle Beyne



Valérie Noël

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/>
- Topo & TP : Workflow managers
- Environnement de travail : Logiciels à installer

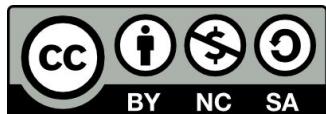


Workflow Manager

TOEGLe

= Galaxy
PROJECT

www.southgreen.fr



Formateurs

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



Objectifs de la formation

objectifs:



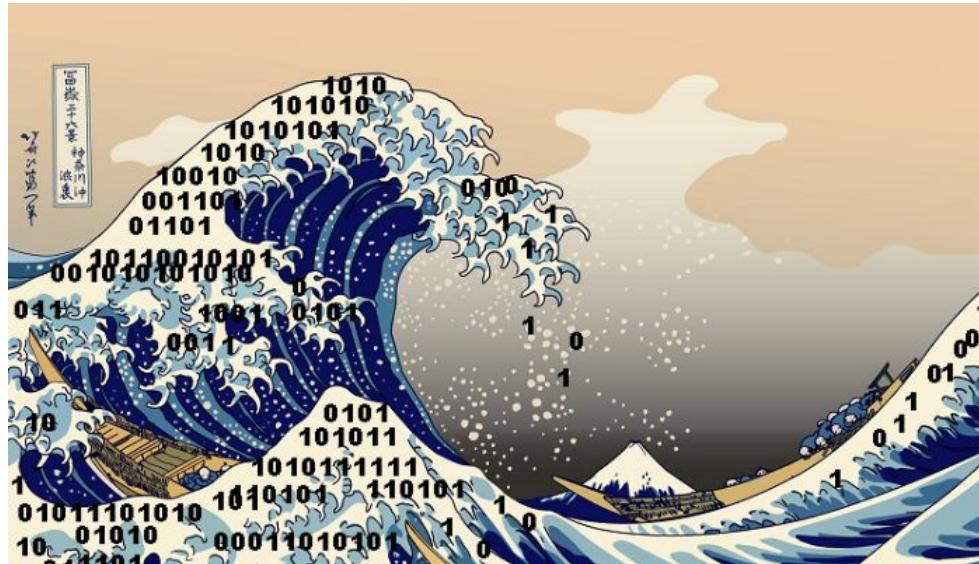
Utiliser les gestionnaires de Workflow de South Green afin de construire de manière automatique vos propres pipelines.

Applications

Tout savoir sur les 2 principaux gestionnaires de workflow

-  bioinformatics platform

Pourquoi utiliser un gestionnaire de workflow?

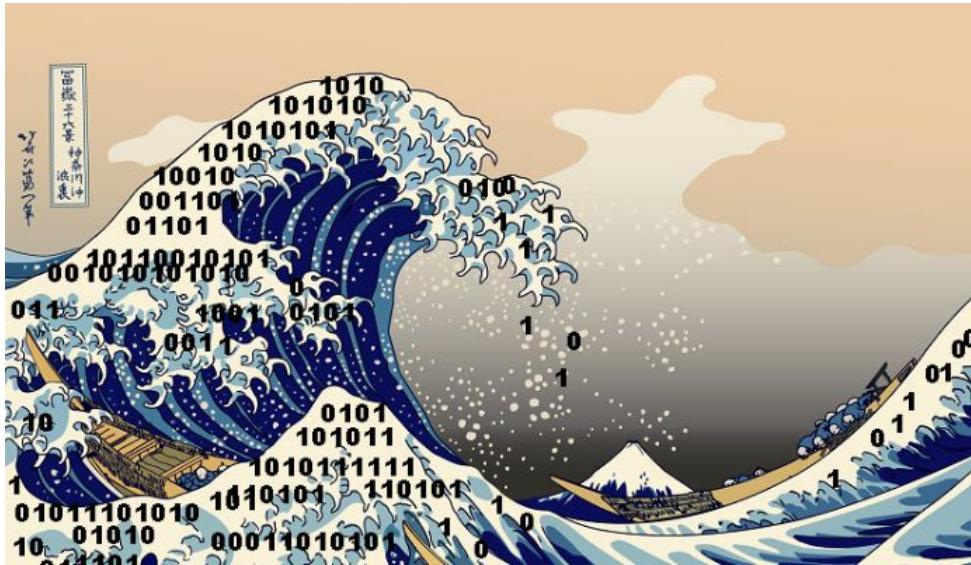


The Great Wave off Kanagawa, Hokusai @amitechsolutions.com



© D.Fletcher for CloudTweaks.com

Pourquoi utiliser un gestionnaire de workflow?



Créer son propre pipeline via une méthode facile et conviviale

Données brutes

Résultats
Intermédiaires

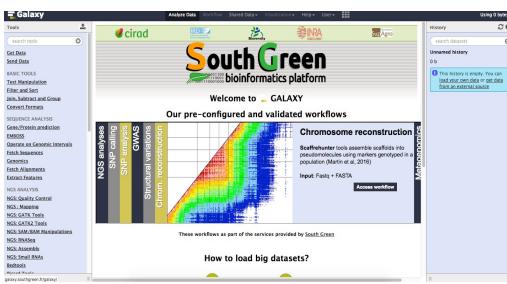
Résultats
Intermédiaires

Résultat
Final

- 3 solutions proposées par



GUI tools

The screenshot shows the South Green bioinformatics platform interface. It features a top navigation bar with links like "Get Data", "Send Data", "Tools", "Help", "About", "Text Manipulation", "Filter and Sort", "Data Analysis and Curate", "Genomic formats", "Bioinformatics and Disease", "EMBOSS", "Disease as Cancerous Tumors", "Disease Genes", "Generics", "Data Models", "Statistical Methods", "NGS Quality Control", "NGS Assembly", "NGS GATK Tools", "NGS GATK Scripts", "NGS SAMtools Scripts", "NGS BEDTOOLS", "NGS BEDTOOLS Scripts", "NGS Small RNA", "MicroRNA", and "MicroRNA Scripts". Below the navigation, there's a "Workflow" section titled "Our pre-configured and validated workflows" with options for "SNV calling", "SnpEff", "GWA", and "Chromosome reconstruction". A large heatmap visualization is shown under the "Chromosome reconstruction" section.

TOGGLE

```
$order
1=fastqc
2=cutadapt
3=bwa mem
4=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration

$cutadapt
-q 30
-m 35
$bwa mem
-n 5
...
$gatk
-q bioinfo.q
-d Y
```



CLI tools



Snakemake

```
rule final:
    input:
        expand('{OutDir}/{_final_data}/tmp/{_tmp_protein.fasta}', sep = '_')
    output:
        OutDir/report/report.html

rule report:
    threads: 1
    input:
        _tmp_assembly = OutDir/_report/data_report/Assembly_quality.csv
        params : _tmp_freeze_id
    output:
        OutDir/_report/report.html
    shell:
        -t
        module purge
        module load bioinfo/0.1.4.3
        Rscript ./markdown/render(*script/report.Rd*, output_file={out})
        ```

rule recuperation_stat:
 threads: 1
 input:
 _tmp_assembly = datadir,outdir = OutDir, lien = expand(OutDir*_4,
 params : _tmp_freeze_id)
 output:
 assembly = OutDir/_report/data_report/Assembly_quality.csv
 stat = OutDir/_report/statistics/statistics.csv
 script/qualityassembly.py -d {input.assembly} -o {output.assembly}
 script/RecoupinGFF.py -d {input.outdir}_4_mergedGFF -c {output.Annot}
        ```

rule renamerida:
    threads: 1
    input:
        _tmp_assembly = _final_data/sep/
        params : _tmp_freeze_id
    output:
        OutDir/_final_data/tmp/{_tmp_protein.fasta}
    shell:
        -t
        module purge
        script/renamedcards.py -d {input} -w {wildcards.sep}
        ````
```



- 3 solutions proposées par

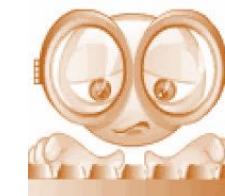


*Snakemake*

TOGGLE



Facilité d'utilisation  
Bonne documentation



Facilité de  
développement

- 3 solutions proposées par



*Snakemake*

TOGGLE



SNP detection

**population genetics**

**GWAS**

transcriptome assembly

genome assembly

transcriptomics

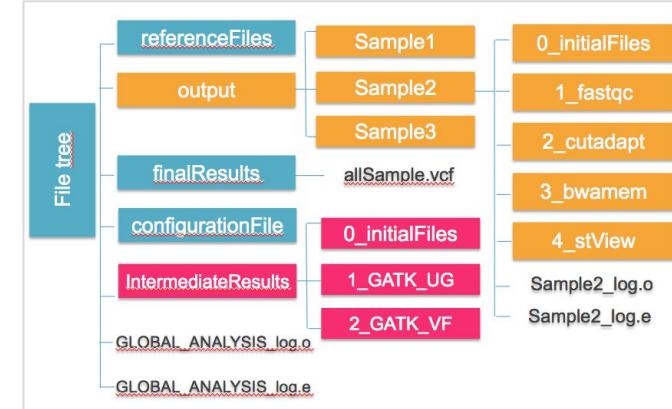
**structural variant detection**

phylogeny

differential expression

# 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

Reproductibilité  
& traçabilité

Apporte un cadre  
robuste

Sauvegarde des options, version des logiciels,  
partage des analyses

**TOGGLE**

Workflow report

Toolbox for generic NGS analyses

<http://toggle.southgreen.fr/>  
toggle@ird.fr

February 8, 2018

## 1 Samples

### 1.1 Parallel execution (step number < 1000)

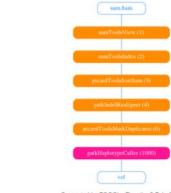
Samples	Job ID	Status
irigin1-PICARDTOOLSMARKDUPPLICATES	1072410	Normal
irigin3-PICARDTOOLSMARKDUPPLICATES	1072411	Normal

### 1.2 Global execution (step number >= 1000)

Samples	Job ID	Status
global	1072412	Normal

## 2 Workflow

### 2.1 Workflow description



Generated by TOGGLE Thu, the 8 Feb. 2018

Figure 1: This image was created automatically by TOGGLE.

### 2.2 Software version

TOGGLE : Release 0.3.6 , 31 of January , 2018
NOTE: The Latest version of TOGGLE (Release 0.3.5 , 8th of November , 2017) is ↵ available at <a href="http://toggle.southgreen.fr/">http://toggle.southgreen.fr/</a>
JAVA : java version "1.8.0_111"
PICARD : 2.5.0(2c370988a4d11579920c8a6a7678a201c5261c1_1466798365)
SAMTOOLS : Version: 1.3.1 (using htseq 1.3.1)
GAIK : 3.6-0-g89b7205

### 2.3 Line code

# Pourquoi utiliser un gestionnaire de workflow?

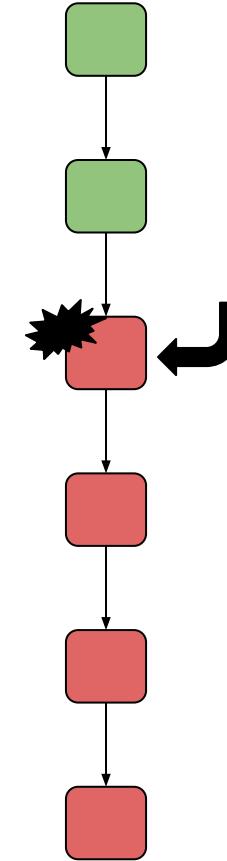
-add / -add:	use if you want to add new samples to an already run analysis.
-rerun / -rerun:	use if you want to re-run samples that have encountered error previously.

Contrôle du pipeline  
et des données

Apporte un cadre  
robuste

Reproductibilité  
& traçabilité

Suivi des erreurs  
& reprise en cours



# Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline  
et des données

Apporte un cadre  
robuste

Reproductibilité  
& traçabilité

Suivi des erreurs  
& reprise en cours



Analyse de gros  
jeu de données

# Pourquoi utiliser un gestionnaire de workflow?

Reproductibilité  
& traçabilité

Contrôle du pipeline  
et des données

Apporte un cadre  
robuste

Suivi des erreurs  
& reprise en cours

Connection HPC  
Parallélisation

Analyse de gros  
jeu de données





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

# TOGGLE

Toolbox for Generic NGS analyses

A framework to quickly build pipelines  
and to perform large-scale NGS analysis



<https://github.com/SouthGreenPlatform/TOGGLE>



# What is TOGGLE ?

- An alternate solution between GUI tools (Galaxy) & CLI tools (SnakeMake)
- Numerous tools integrated to perform post analysis
- Targets both biologists & bioinformaticians



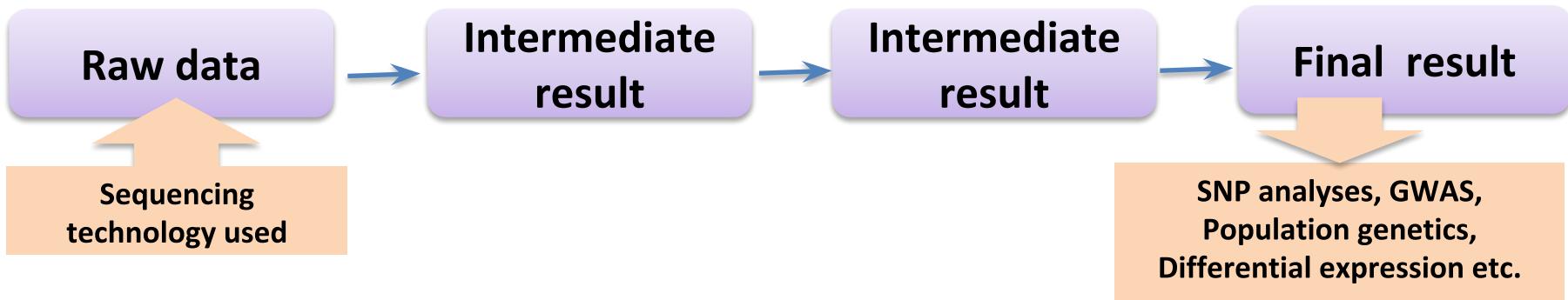
19 modules, 120 functions 120  
open-source tools



# What is TOGGLE ?



19 modules, 120 functions  
40 open-source tools



# What is TOGGLE ?

GBS

RADSeq

RNASeq

WGS

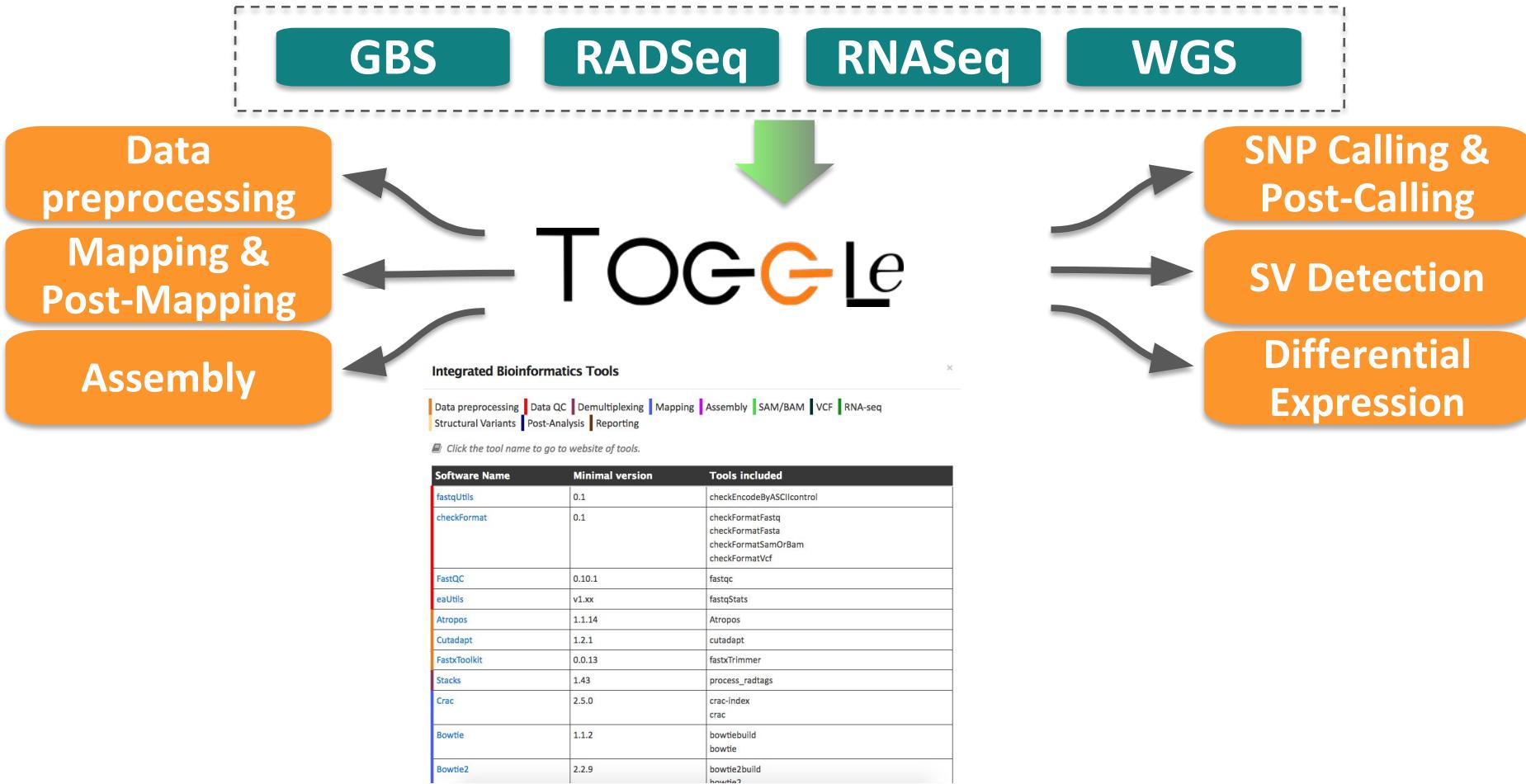


TOGGLE

Various data format : fasta,  
fastq, sam, bam, bed, vcf  
(compressed or not)



# Use TOGGLE for ?



# Tools includes

## Data preprocessing

**Atropos, Stacks**  
**FastQC, Cutadapt**  
**FASTX-Toolkit**

## Mapping

**Bowtie, Bowtie2**  
**BWA, crac**  
**Tophat2**

## Post Mapping

**NGSUtils, picardTools**  
**SAMtools, GATK**

## SV Detection

**DuplicationDetector,**  
**BreakDancer, Pindel**

## Differential Expression

**Htseq-count, cufflinks**

## SNP Calling

**SAMtools, GATK**  
**VarScan, SNPEff**

## Assembly

**Trinity**  
**TGI-CL**  
**Abyss**

## Post Calling

**BEDtools, plink, sNMF**  
**FastME, plink, readseq**

## Long Reads

**nanoplot, wtdbg2, minimap2, flye**

TOGGLE

<https://toggle.southgreen.fr>

# How to perform an analysis with TOGGLE ?

A command-line based pipeline framework



A single command line

```
toggleGenerator.pl -d DIR -o DIR -c FILE
```

# What does TOGGLE need to run ?

```
toggleGenerator.pl -d|--directory DIR -c|--config FILE -o|--outputdir DIR [-r|--reference FILE] [-k|--keyfile FILE] [-g|--gff FILE] [-nocheck|--nocheckFastq] [--help|-h]
```

## Required named arguments:

-d / --directory DIR	a folder with raw data to be treated (FASTA, FASTQ, SAM, BAM, BED, GFF, VCF)
-c / --config FILE	it is the <i>software.config.txt</i> file but it can be any text file (Unix format).
-o / --outputdir DIR	the current version of TOGGLE will not modify the initial data folder but will create an output directory with all analyses in. This module must be empty (TOGGLE will stop if not).

## Optional named arguments:

-r / --reference FILE	a reference FASTA file to be used. (1)
-g / --gff FILE	a GFF file to be used for some tools . Be careful the gff name must be different than the FASTA.
-k / --keyfile FILE	a keyfile use for demultiplexing step.
-add / --add	use if you want to add new samples to an already run analysis.
-rerun / --rerun:	use if you want to re-run samples that have encountered error previously.
-nocheck	by default checks if given formats for input files are correct. This option allows to skip this step.
-report / --report	generate pdf report ( <a href="#">more info</a> )
-h / --help	show help message and exit

# What does TOGGLE need to run ?

- An input directory (with fastq, sam/bam, vcf files)
- The name of output directory used to store the data generated by the analyses
- A unique and simple configuration file to design the pipeline and define software parameters.
- Optional arguments : reference file, annotation...

# A simple configuration file ...

## \$order

```
1=fastqc
2=cutadapt
3=bwa mem
4=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration
```

## \$cutadapt

```
-q 30
-m 35
```

## \$bwa mem

```
-n 5
```

```
...
```

## \$sge

```
-q bioinfo.q
-b Y
```

# To create your own workflow

```
$order
1=fastqc
2=cutadapt
3=bwa mem
4=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration
```

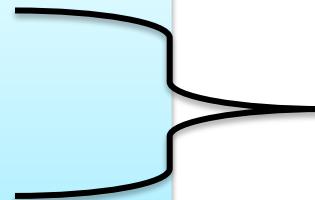
## Create your own workflow

- The workflow order
- The list of softwares to run

One line = the step followed by the software's name

# To create your own workflow

\$order  
1=fastqc  
2=cutadapt  
3=bwa mem  
4=samToolsView  
1000=gatkHaplotypeCaller  
1001=gatkVariantFiltration



## Create your own workflow

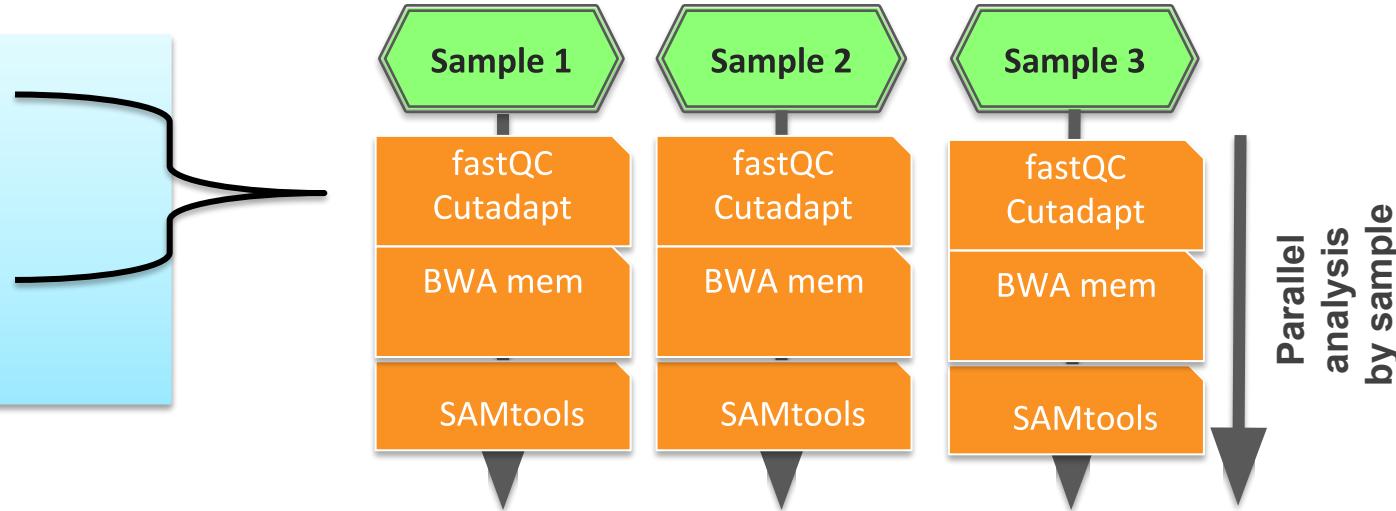
Step number < 1000

Parallel analysis by sample

# To create your own workflow

\$order

1=fastqc  
2=cutadapt  
3=bwa mem  
4=samToolsView  
1000=gatkHaplotypeCaller  
1001=gatkVariantFiltration



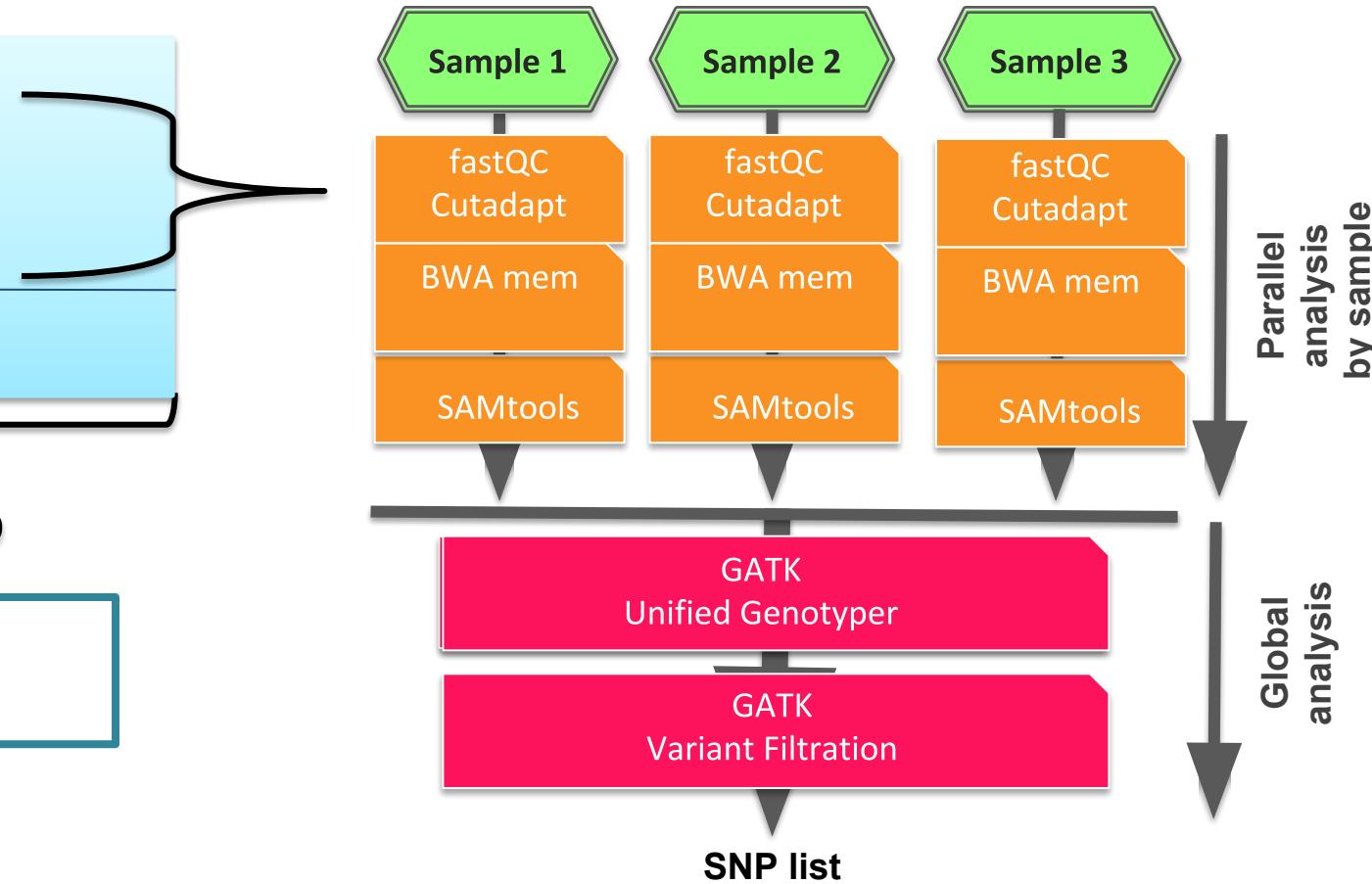
# To create your own workflow

\$order

1=fastqc  
2=cutadapt  
3=bwa mem  
4=samToolsView  
1000=gatkHaplotypeCaller  
1001=gatkVariantFiltration

Step number  $\geq 1000$

Global analysis  
(all samples)



# Software parameters

## \$order

```
1=fastqc
2=cutadapt
3=bwa mem
4=picardToolsSortSam
5=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration
```

## \$cutadapt

```
-q 30
-m 35
```

## \$bwa mem

```
-n 5
```

```
...
```

## \$sge

```
-q bioinfo.q
-b Y
```

## Software parameters

One tag per software (\$softwareName)  
followed by the list of options

# Scheduleurs usage

## \$order

1=fastqc  
2=cutadapt  
3=bwa mem  
4=picardToolsSortSam  
5=samToolsView  
1000=gatkHaplotypeCaller  
1001=gatkVariantFiltration

## \$cutadapt

-q 30  
-m 35

## \$bwa mem

-n 5  
...

## \$sge

-q bioinfo.q  
-b Y

## Job schedulers

LSF, MPRUN, SLURM, SGE



# More info on website

**<http://toggle.southgreen.fr/>**



The screenshot shows the TOGGLE website homepage. At the top, there is a dark header bar with the text "Version: 0.3.5" and "SEARCH". To the right are links for "Home", "Tools", "Manual", "Install", "FAQ", "About us", "Download", "GitHub", and a GitHub icon. Below the header is a blue banner with the TOGGLE logo and the text "Toolbox for generic NGS analyses". The main content area has a white background. It contains two paragraphs of text addressed to biologists and bioinformaticians, followed by the message "So, be Happy! TOGGLE is for you!!". Below this is a section titled "Quick info about TOGGLE" with a dark gray background. It features the SouthGreen logo at the top, followed by the TOGGLE logo and the text "Tools for Generic NGS analysis". A large orange banner at the bottom of this section contains the text "A framework to quickly build pipelines". At the very bottom of the page are two blue buttons labeled "Integrated Bioinformatics Tools List" and "Quick Manual".

User Manuals

Screencast

Developer manual

Pre-defined Workflow files



**<https://github.com/SouthGreenPlatform/TOGGLE>**

# TOGGLE 's team



Christine Tranchant-Dubreuil  
[christine.tranchant@ird.fr](mailto:christine.tranchant@ird.fr)



Julie Orjuela-Bouniol  
[julie.orjuela@ird.fr](mailto:julie.orjuela@ird.fr)



François Sabot  
[francois.sabot@ird.fr](mailto:francois.sabot@ird.fr)



Sébastien Ravel  
[sébastien.ravel@cirad.fr](mailto:sébastien.ravel@cirad.fr)



Alexis Dereeper  
[alexis.dereeper@ird.fr](mailto:alexis.dereeper@ird.fr)



Ndomassi Tando  
[ndomassi.tando@ird.fr](mailto:ndomassi.tando@ird.fr)



Valérie Noel  
[valerie.noel@ird.fr](mailto:valerie.noel@ird.fr)



Valentin Klein  
[valentin.klein@ird.fr](mailto:valentin.klein@ird.fr)



[toggle@ird.fr](mailto:toggle@ird.fr)

## Former collaborators

- Cécile Monat (UMR DIADE - IRD)
- Gautier Sarah (UMR AGAP - CIRAD)
- Abdoulaye Diallo (UMR DIADE - IRD)
- Laura Helou (UMR DIADE - IRD)
- Souhila Amazougarene (UMR DIADE - IRD)
- Mawussé Agbessi (UMR DIADE - IRD)
- Enrique Ortega-Abboud (UMR AGAP - CIRAD)
- Cédric Farcy (UMR AGAP - CIRAD)
- Maryline Summo (UMR AGAP - CIRAD)
- Ayité Kougbeadjo (UMR DIADE - IRD)

# TOGGLE : Bonnes pratiques

## Comment citer TOGGLE?

Tranchant-Dubreuil, C., Ravel, S., Monat, C., Sarah, G., Diallo, A., Helou, L., ... Sabot, F. (2018). [TOGGLE, a flexible framework for easily building complex workflows and performing robust large-scale NGS analyses](#). *BioRxiv*. <https://doi.org/10.1101/245480>

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

## Comment citer les clusters?

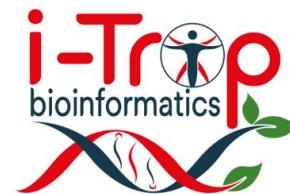
"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://bioinfo.ird.fr/> "

"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>"

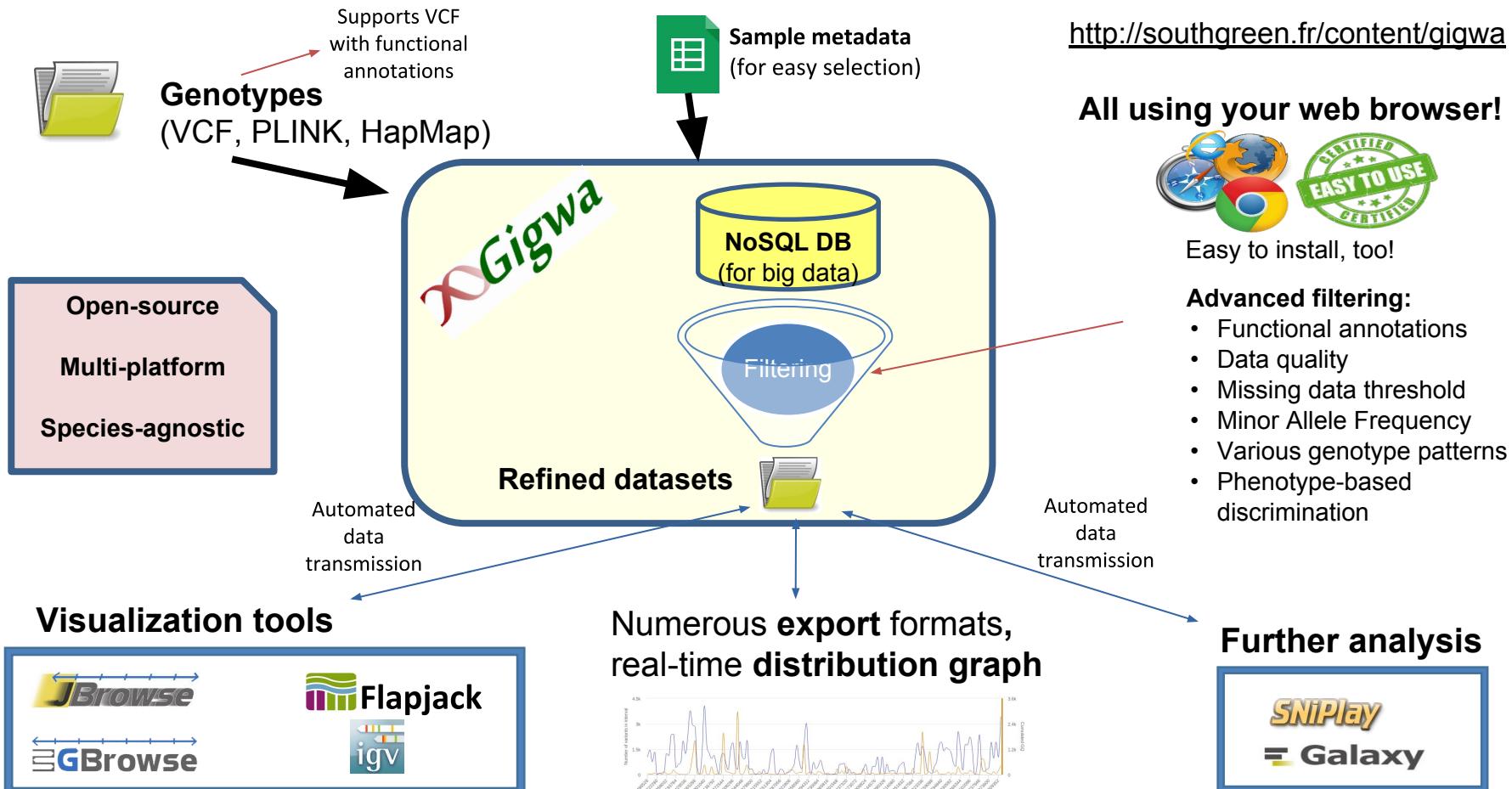
SUIVEZ NOUS SUR TWITTER !



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



i-Trop : [@ItropBioinfo](https://twitter.com/ItropBioinfo)



Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, Larmande P. *Gigwa-Genotype investigator for genome-wide analyses.*

# 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/>