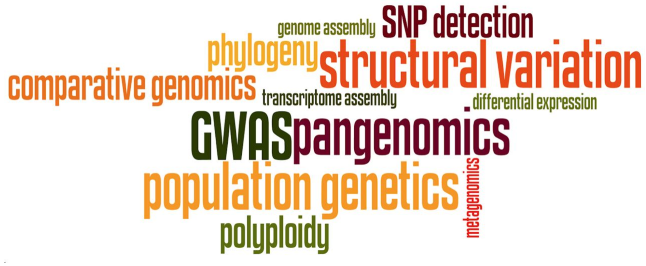
South Green bioinformatics platform

Session de formation 2019

SouthGreen
bioinformatics platform

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



Rice

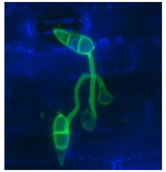
Banana

Palm

Sorghum

Coffee

Cassava


Magnaporthe

www.southgreen.fr

South Green 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 Fredéric

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

South Green
bioinformatics platform

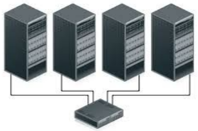
Workflow manager

TOGGLe
Toolbox for generic NGS analyses

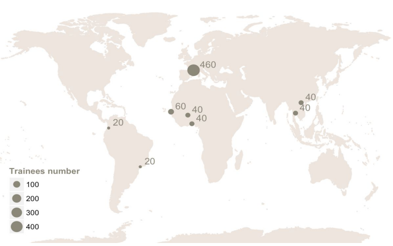
SNAKEMAKE

Galaxy

HPC and trainings….



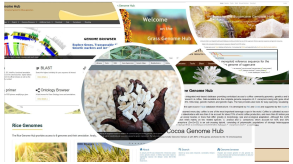
37 courses organized last 7 years



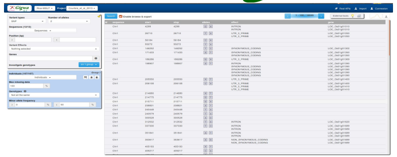
IRD
Institut de Recherche pour le Développement
FRANCE

cirad

Genome Hubs & Information System

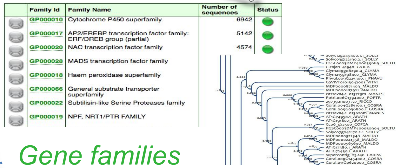


Gigwa



SNPs and Indels

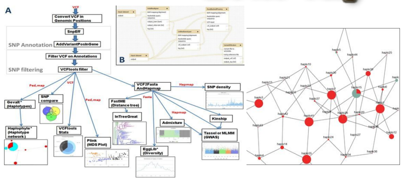
GreenPhyl



	Family Id	Family Name	Number of sequences	Status
	GP000010	Cytochrome P450 superfamily	6942	●
	GP000017	AP2/EREBP transcription factor family: ERF/DREB group (partial)	5142	●
	GP000020	NAC transcription factor family	4574	●
	GP000028	MADS transcription factor family		
	GP000018	Haem peroxidase superfamily		
	GP000066	General substrate transporter superfamily		
	GP000022	Subtilisin-like Serine Proteases family		
	GP000019	NPF, NRT1/PTR 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

SouthGreen bioinformatics platform

ABiMS

Erwan Corre

ASTRE

Etienne Loire
Julie Reveillaud

diade

Valentin Klein

TransVIH MI

Emmanuelle Beyne

IPME

Marie Simonin
Sébastien Cunnac

UMR QualiSud

Florentin Constancias

MiVEGEC

i-Trop bioinformatics

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

South Green bioinformatics platform

Modules de formation 2019

- Toutes nos formations :

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

- Topo & TP : Workflow managers

- Environnement de travail : Logiciels à installer

South Green bioinformatics platform

Workflow Manager

TOGGLe

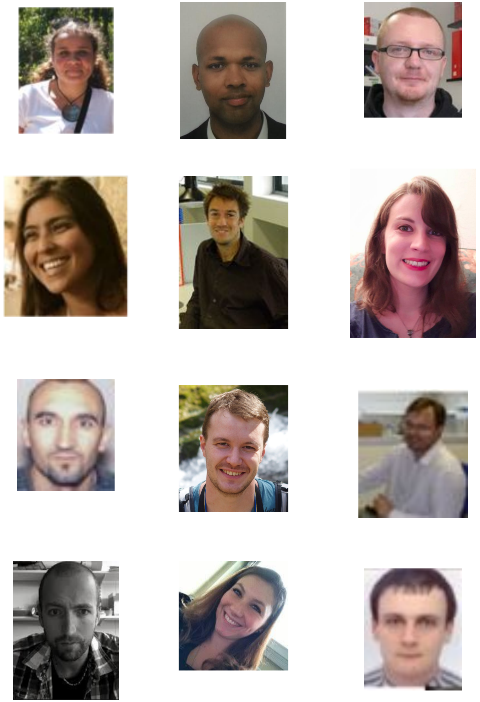
Galaxy PROJECT

www.southgreen.fr

Bioversity International · cirad · INRA SCIENCE & IMPACT · IRD Institut de Recherche pour le Développement FRANCE · Montpellier SupAgro

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**



Objectifs de la formation

objectifs:

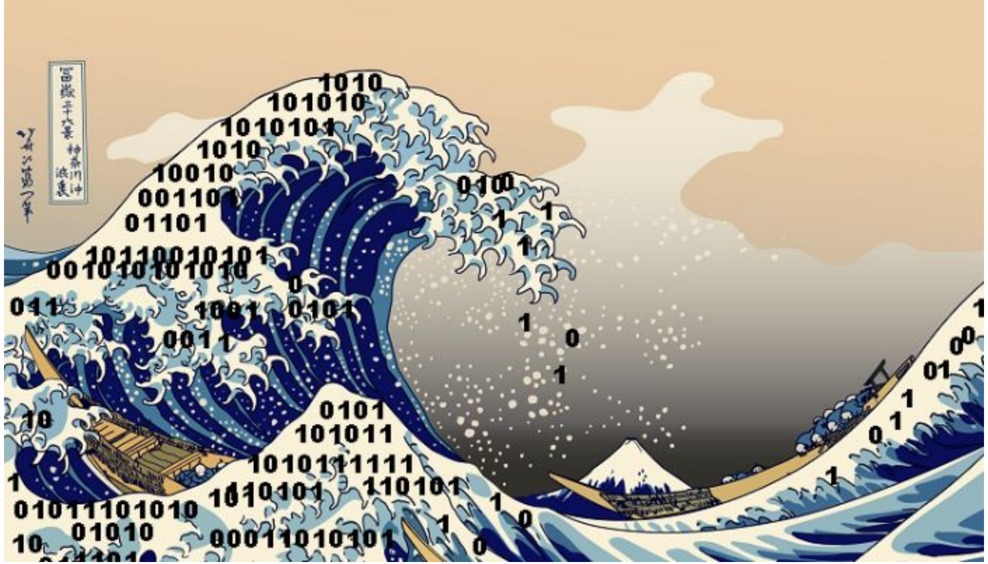
Utiliser les gestionnaires de Workflow de South Green afin de construire de manière automatique vos propres pipelines.

Applications

Tout savoir sur les 2 principaux gestionnaires de workflow South Green bioinformatics platform

- 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

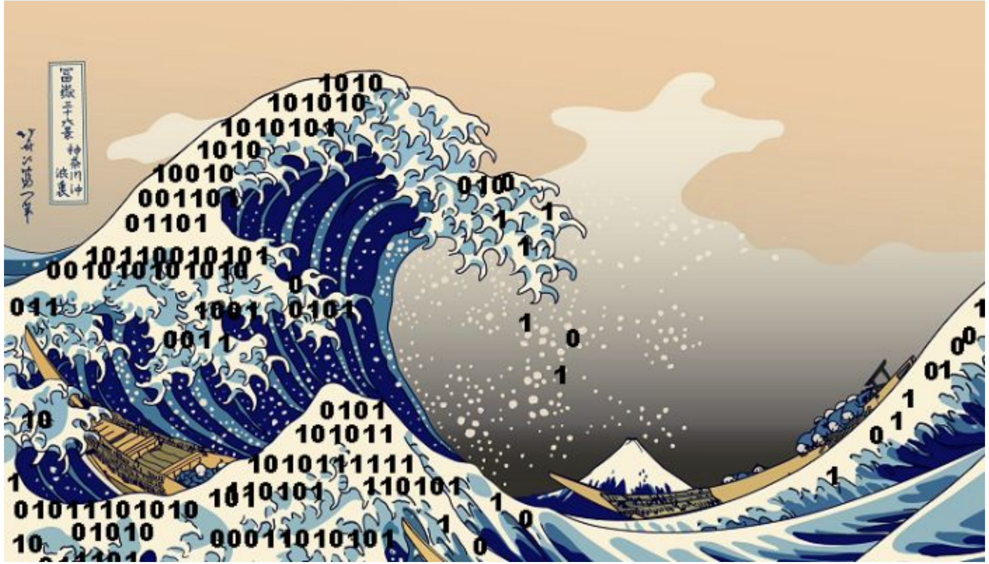
Pourquoi utiliser un gestionnaire de workflow?



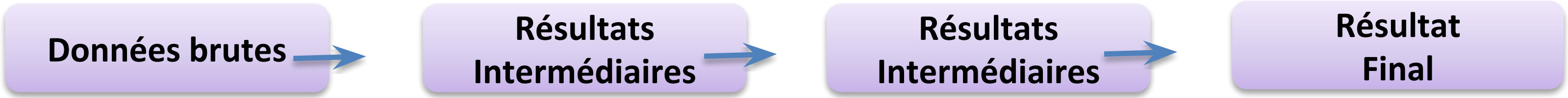
The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

HOW'S THE BIG DATA PROJECT COMING ALONG, HOSKINS?
© D.Fletcher for CloudTweaks.com

Pourquoi utiliser un gestionnaire de workflow?



Créer son propre pipeline via une méthode facile et conviviale



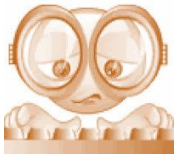
Données brutes → Résultats Intermédiaires → Résultats Intermédiaires → Résultat Final

Pourquoi utiliser un gestionnaire de workflow?

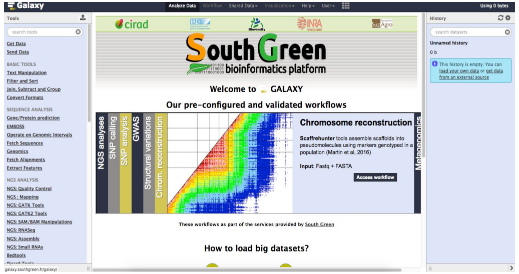
- 3 solutions proposées par South Green bioinformatics platform

GUI tools

CLI tools



Galaxy



South Green bioinformatics platform

Welcome to GALAXY

Our pre-configured and validated workflows

Chromosome reconstruction

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

Input: Fastq + FASTA

These workflows as part of the services provided by South Green

How to load big datasets?

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

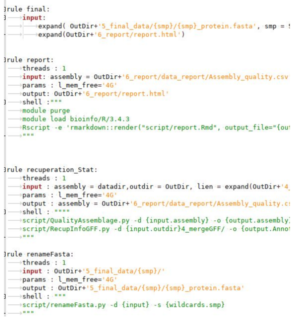
TOGGLe

```
$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
```

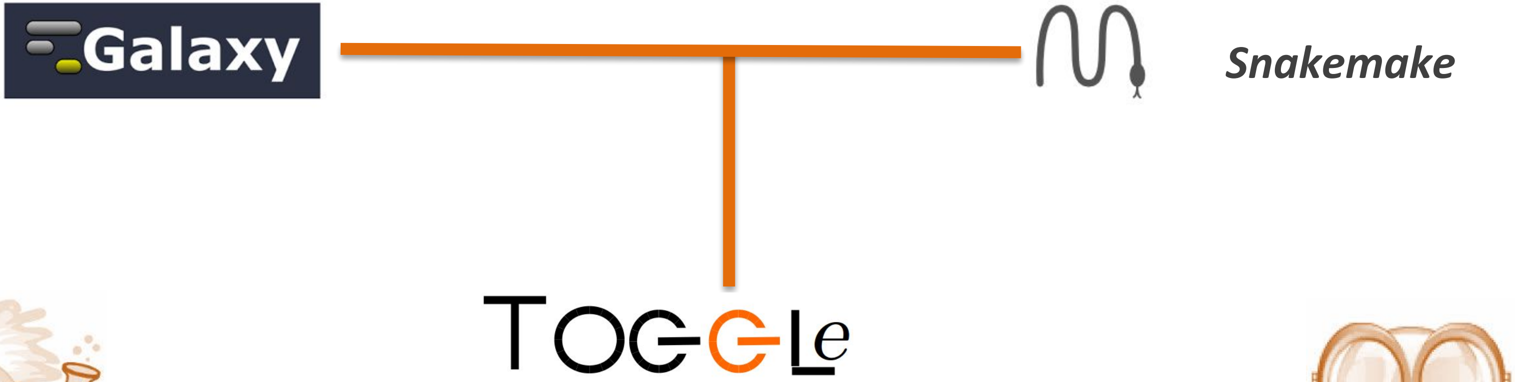
Snakemake



```
rule final:
    input:
        expand( OutDir+'5_final_data/{smp}/{smp}_protein.fasta', smp = 
        expand(OutDir+'6_report/report.html')


rule report:
    threads : 1
    input: assembly = OutDir+'6_report/data_report/Assembly_quality.csv
    params : l_mem_free='4G'
    output: OutDir+'6_report/report.html'
    shell :"""
    module purge
    module load bioinfo/R/3.4.3
    Rscript -e 'rmarkdown::render("script/report.Rmd", output_file="{ou
    """



rule recuperation_Stat:
    threads : 1
    input : assembly = datadir,outdir = OutDir, lien = expand(OutDir+'4_
    params : l_mem_free='4G'
    output : assembly = OutDir+'6_report/data_report/Assembly_quality.c
    shell : """"
    script/QualityAssemblage.py -d {input.assembly} -o {output.assembly
    script/RecupInfoGFF.py -d {input.outdir}4_mergeGFF/ -o {output.Anno
    """

rule renameFasta:
    threads : 1
    input : OutDir+'5_final_data/{smp}/'
    params : l_mem_free='4G'
    output : OutDir+'5_final_data/{smp}/{smp}_protein.fasta'
    shell : """
    script/renameFasta.py -d {input} -s {wildcards.smp}
    """
```

Pourquoi utiliser un gestionnaire de workflow?

- 3 solutions proposées par South Green bioinformatics platform



Galaxy

Snakemake

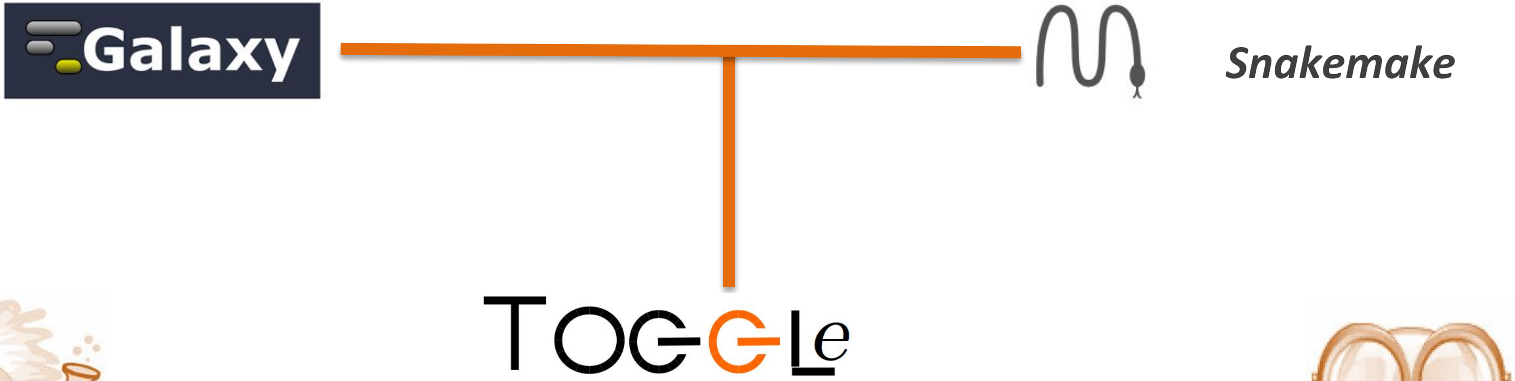
TOGGLe

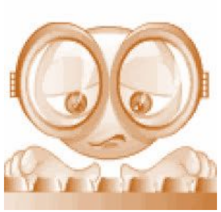
Facilité d'utilisation
Bonne documentation

Facilité de développement

Pourquoi utiliser un gestionnaire de workflow?

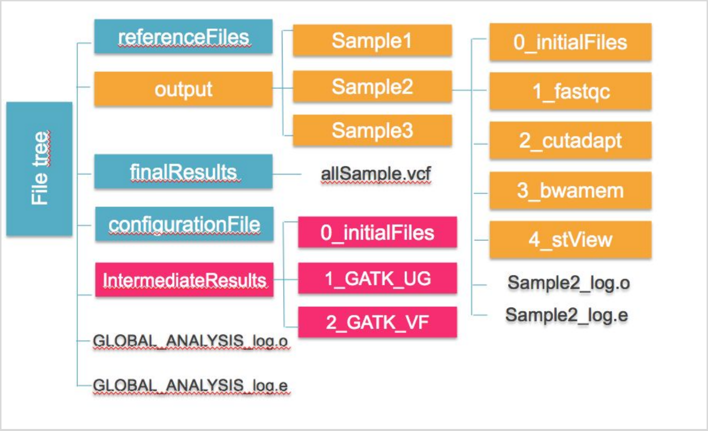
- 3 solutions proposées par South Green bioinformatics platform





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

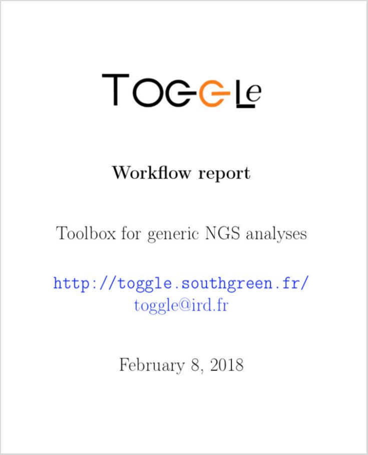
+

Automatisation de certaines étapes clefs

(ex : indexation de la référence)

Pourquoi utiliser un gestionnaire de workflow?

Contrôle du pipeline et des données



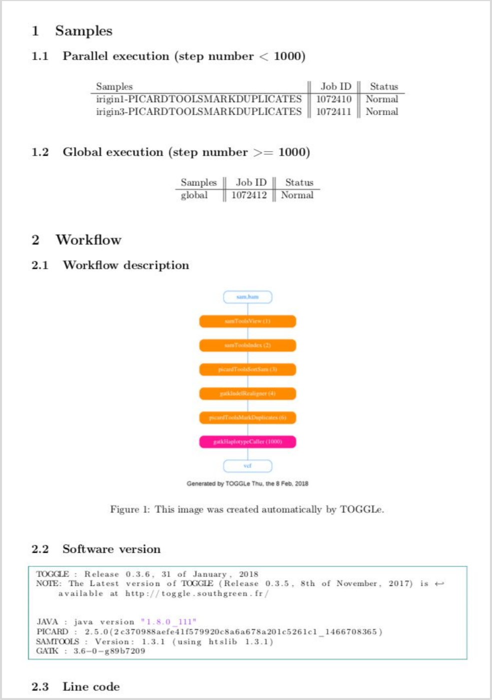
TOGGLe

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-PICARDTOOLSMARKDUPLICATES	1072410	Normal
irigin3-PICARDTOOLSMARKDUPLICATES	1072411	Normal

1.2 Global execution (step number >= 1000)

Samples	Job ID	Status
global	1072412	Normal

2 Workflow

2.1 Workflow description

sam,bam
samToolsView (1)
samToolsIndex (2)
picardToolsSortSam (3)
gatkIndelRealigner (4)
picardToolsMarkDuplicates (6)
gatkHaplotypeCaller (1000)
vcf
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 http://toggle.southgreen.fr/


JAVA : java version "1.8.0_111"
PICARD : 2.5.0(2c370988aefe41f579920c8a6a678a201c5261c1_1466708365)
SAMTOOLS : Version: 1.3.1 (using htslib 1.3.1)
GATK : 3.6-0-g89b7209
```

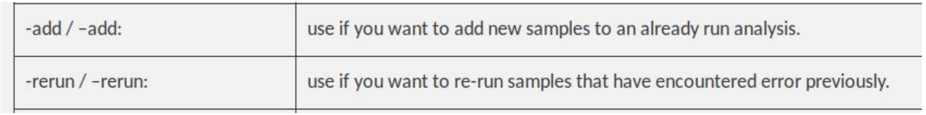
2.3 Line code

Reproductibilité & traçabilité

Apporte un cadre robuste

Sauvegarde des options, version des logiciels, partage des analyses

Pourquoi utiliser un gestionnaire de workflow?



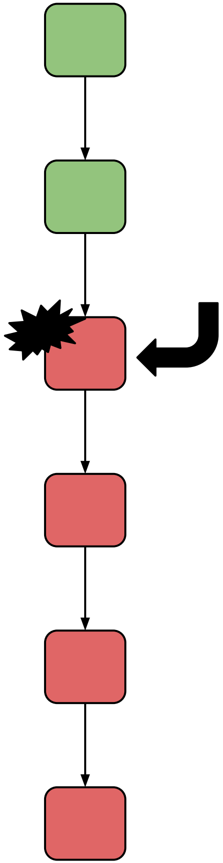
-add / –add:	use if you want to add new samples to an already run analysis.
-rerun / –rerun:	use if you want to re-run samples that have encountered error previously.

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

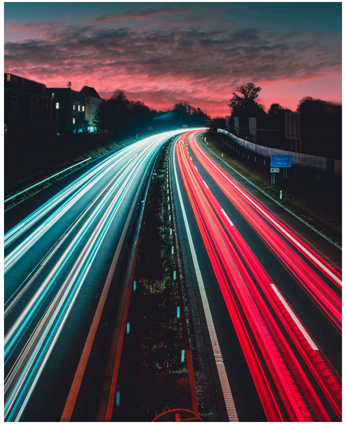
Apporte un cadre robuste

Reproductibilité & traçabilité

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

TOGGLe

Toolbox for Generic NGS analyses

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

htps://github.com/SouthGreenPlatorm/TOGGLE

Bioversity International   cirad   INRA SCIENCE & IMPACT   IRD Institut de Recherche pour le Développement FRANCE

What is TOGGLe ?

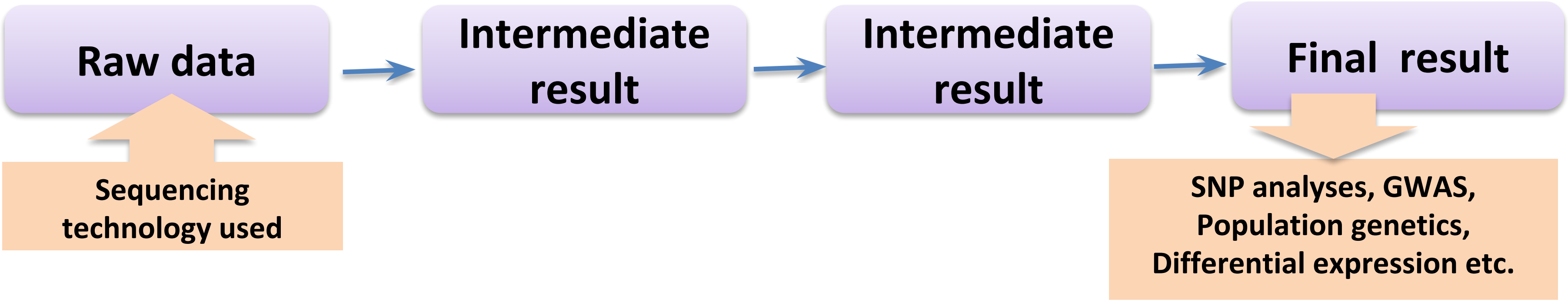
- An alternate solution between GUI tools (Galaxy) & CLI tools (SnakeMake)
- Numerous tools integrated to perform post analysis
- Targets both biologists & bionformaticians

19 modules, 120 functions 120 open-source tools

PROGRAMMING REPUBLIC OF PERL

What is TOGGLe ?

19 modules, 120 functions
40 open-source tools

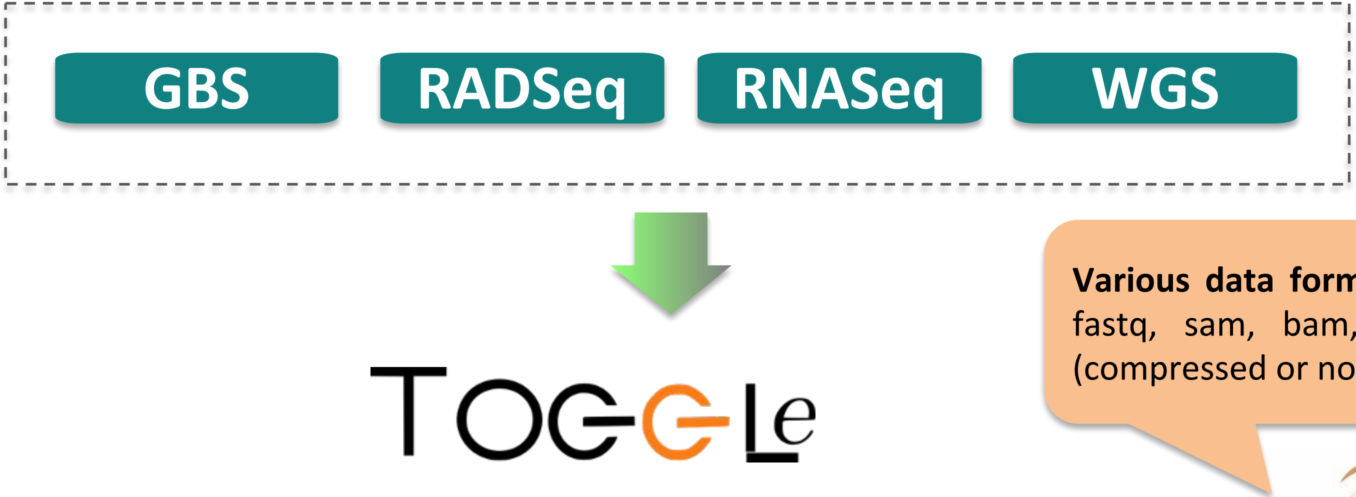


Raw data → Intermediate result → Intermediate result → Final result

Sequencing technology used

SNP analyses, GWAS, Population genetics, Differential expression etc.

What is TOGGLe ?



GBS

RADSeq

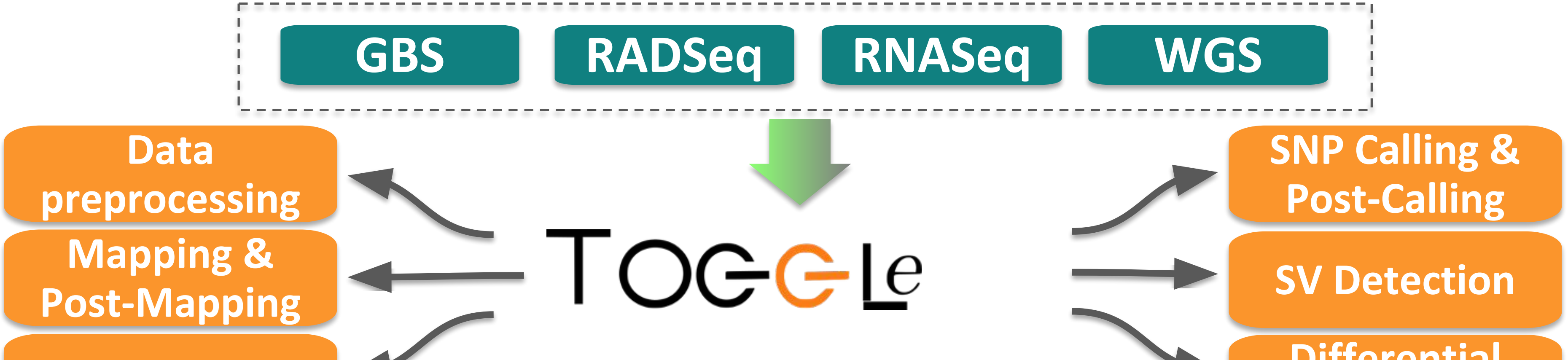
RNASeq

WGS

TOGGLe

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

Use TOGGLe for ?



GBS

RADSeq

RNASeq

WGS

TOGGLe

Data preprocessing

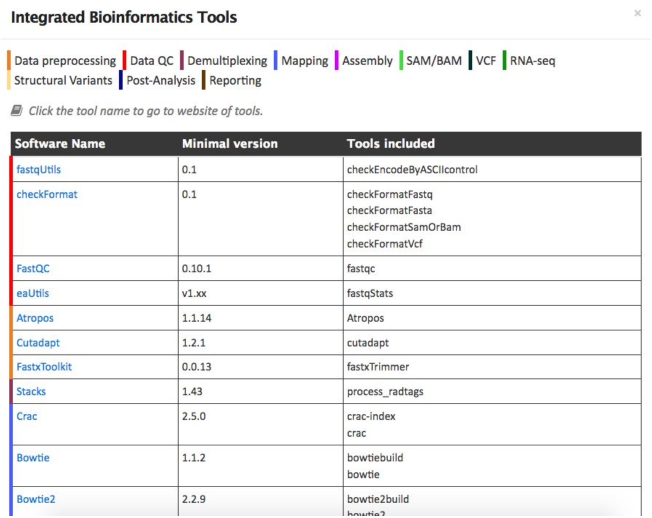
Mapping & Post-Mapping

Assembly

SNP Calling & Post-Calling

SV Detection

Differential Expression



Integrated Bioinformatics Tools

Data preprocessing | Data QC | Demultiplexing | Mapping | Assembly | SAM/BAM | VCF | RNA-seq | Structural Variants | Post-Analysis | Reporting

Click the tool name to go to website of tools.

Software Name	Minimal version	Tools included
fastqUtils	0.1	checkEncodeByASCIIcontrol
checkFormat	0.1	checkFormatFastq checkFormatFasta checkFormatSamOrBam checkFormatVcf
FastQC	0.10.1	fastqc
eaUtils	v1.xx	fastqStats
Atropos	1.1.14	Atropos
Cutadapt	1.2.1	cutadapt
FastxToolkit	0.0.13	fastxTrimmer
Stacks	1.43	process_radtags
Crac	2.5.0	crac-index crac
Bowtie	1.1.2	bowtiebuild bowtie
Bowtie2	2.2.9	bowtie2build bowtie2

South Green bioinformatics platform

Tools includes

Data preprocessing

Atropos, Stacks
FastQC, Cutadapt
FASTX-Toolkit

SV Detection

DuplicationDetector,
BreakDancer, Pindel

Differential Expression

Htseq-count, cufflinks

Mapping

Bowtie, Bowtie2
BWA, crac
Tophat2

TOGGLe

https://toggle.southgreen.fr

SNP Calling

SAMtools, GATK
VarScan, SNPEff

Post Mapping

NGSUtils, picardTools
SAMtools, GATK

Assembly

Trinity
TGI-CL
Abyss

Post Calling

BEDtools, plink, sNMF
FastME, plink, readseq

Long Reads

nanoplot, wtdbg2, minimap2, flye

How to perform an analysis with TOGGLe ?

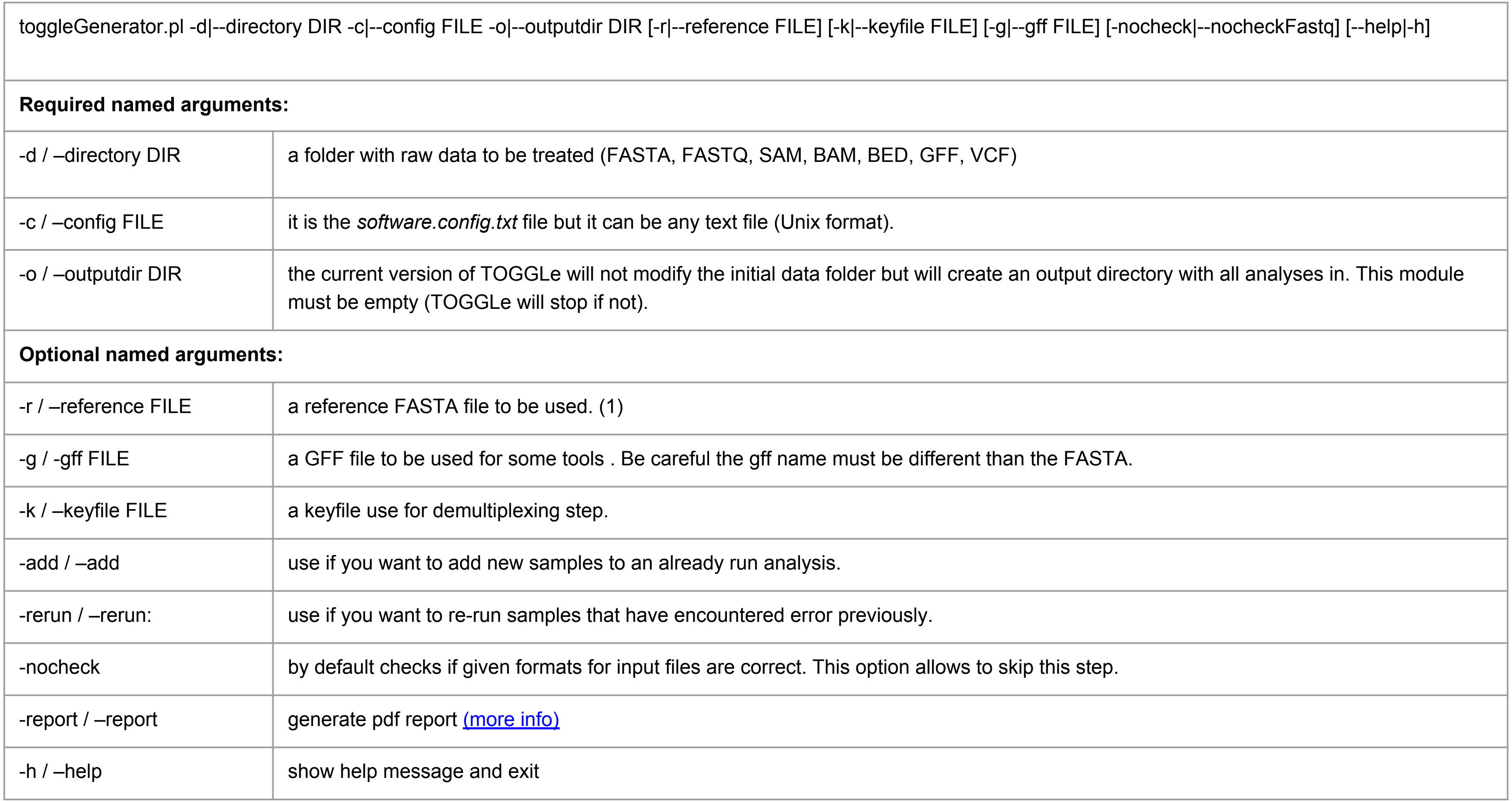
A command-line based pipeline framework

↓

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 *software.config.txt* 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 (more info)
-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

Create your own workflow

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

- The workflow order
- The list of softwares to run

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

To create your own workflow

Create your own workflow

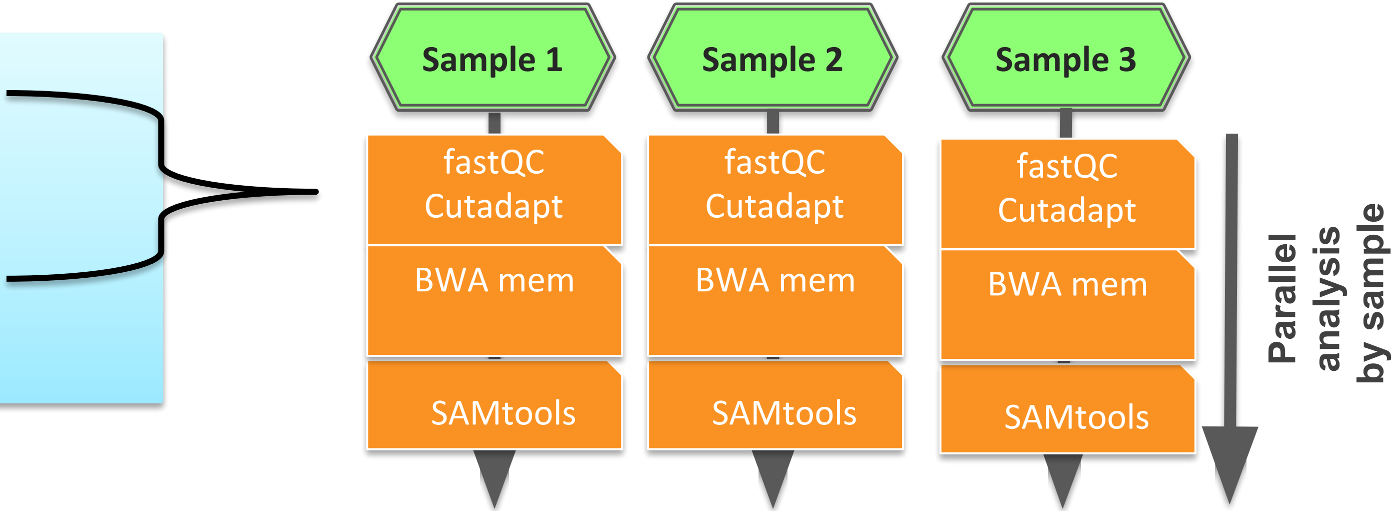
```
$order
1=fastqc
2=cutadapt
3=bwa mem
4=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration
```

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

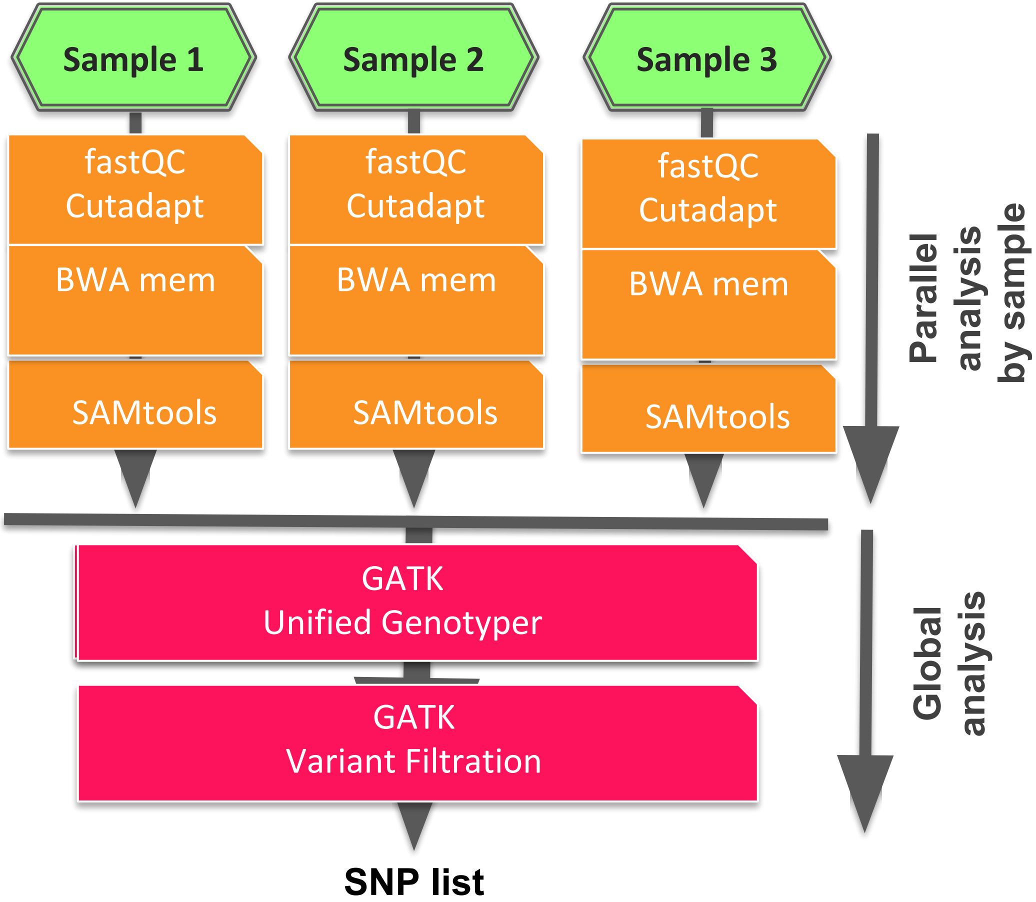


To create your own workflow

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

Step number >= 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

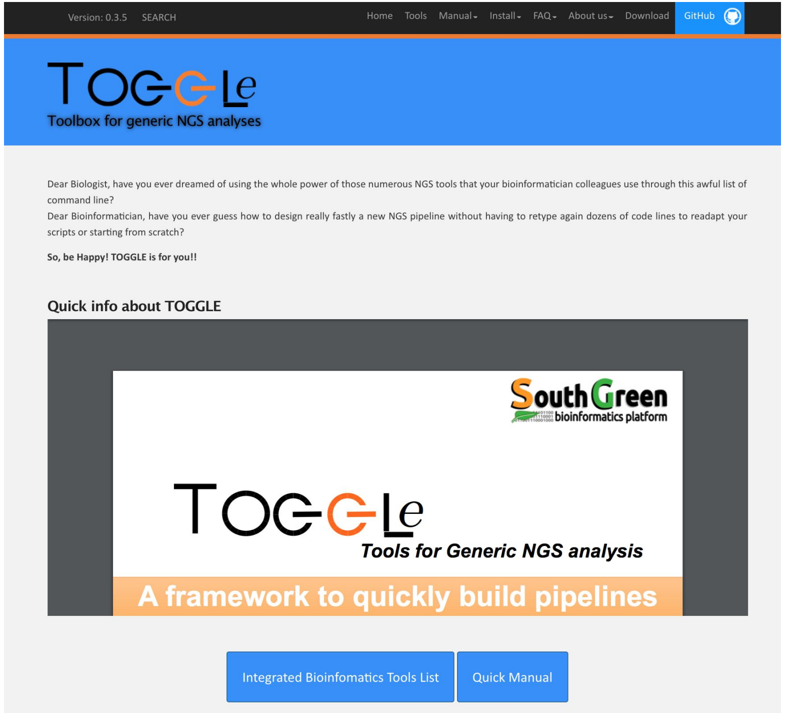
LSF, MPRUN, SLURM, SGE

More info on website

http://toggle.southgreen.fr/

User Manuals

Screencast



Dear Biologist, have you ever dreamed of using the whole power of those numerous NGS tools that your bioinformatician colleagues use through this awful list of command line?
Dear Bioinformatician, have you ever guess how to design really fastly a new NGS pipeline without having to retype again dozens of code lines to readapt your scripts or starting from scratch?

So, be Happy! TOGGLE is for you!!

Quick info about TOGGLE

South Green bioinformatics platform
TOGGLe
Tools for Generic NGS analysis
A framework to quickly build pipelines
Integrated Bioinfomatics Tools List
Quick Manual

Developer manual

Pre-defined Workflow files

https://github.com/SouthGreenPlatform/TOGGLE

TOGGLe 's team

Christine Tranchant-Dubreuil
christine.tranchant@ird.fr

Julie Orjuela-Bouniol
julie.orjuela@ird.fr

François Sabot
francois.sabot@ird.fr

Sébastien Ravel
sebastien.ravel@cirad.fr

Alexis Dereeper
alexis.dereeper@ird.fr

Ndomassi Tando
ndomassi.tando@ird.fr

Valérie Noel
valerie.noel@ird.fr

Valentin Klein
valentin.klein@ird.fr

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

I-Trop : @ItropBioinfo

http://southgreen.fr/content/gigwa

Genotypes
(VCF, PLINK, HapMap)

Supports VCF with functional annotations

Sample metadata
(for easy selection)

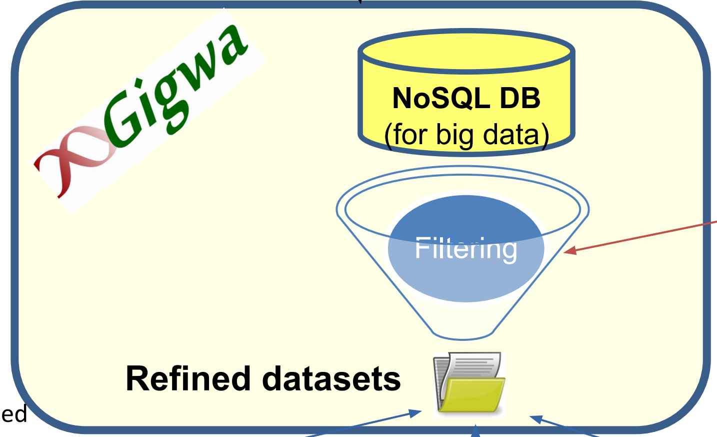
All using your web browser!

Easy to install, too!

Open-source

Multi-platform

Species-agnostic



Advanced filtering:

- Functional annotations
- Data quality
- Missing data threshold
- Minor Allele Frequency
- Various genotype patterns
- Phenotype-based discrimination

Automated data transmission

Automated data transmission

Visualization tools

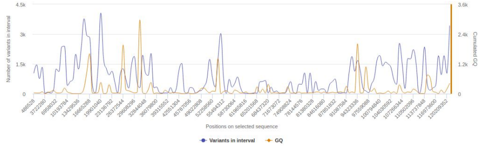
JBrowse

Flapjack

GBrowse

igv

Numerous **export** formats, real-time **distribution graph**



Further analysis

SNiPlay

Galaxy

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

Merci pour votre attention !