SouthGreen bioinformatics platform

# TOGGLe

***Toolbox for Generic NGS analyses***

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

www.southgreen.fr

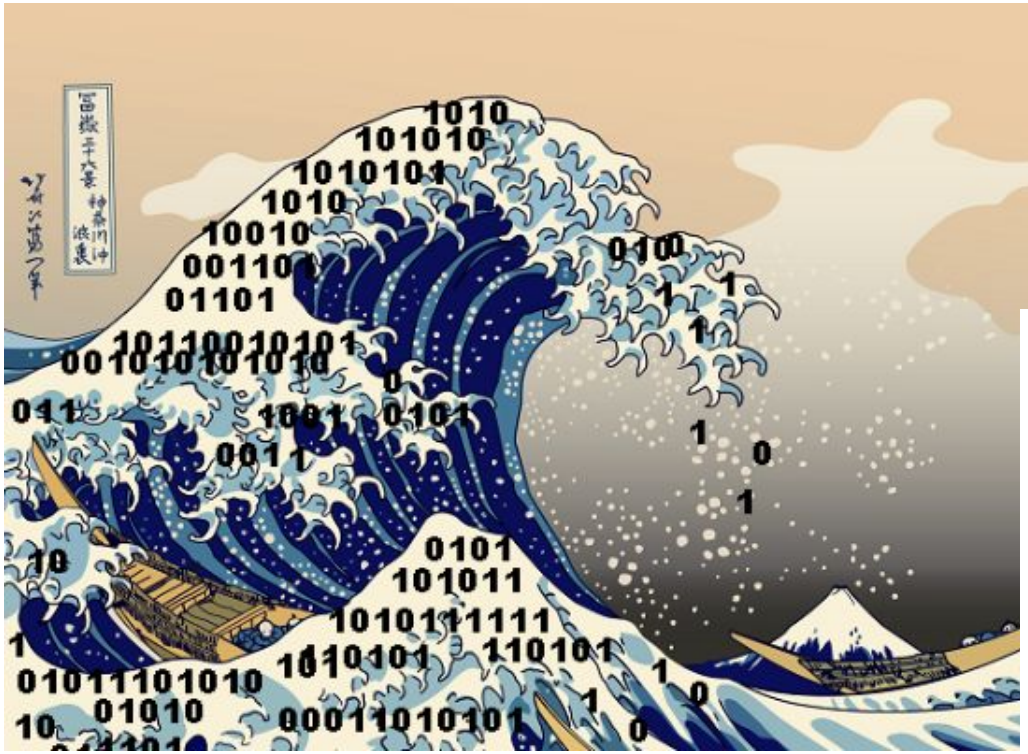
https://southgreenplatform.github.io/trainings



**TOGGLe Team**
**toggle@ird.fr**

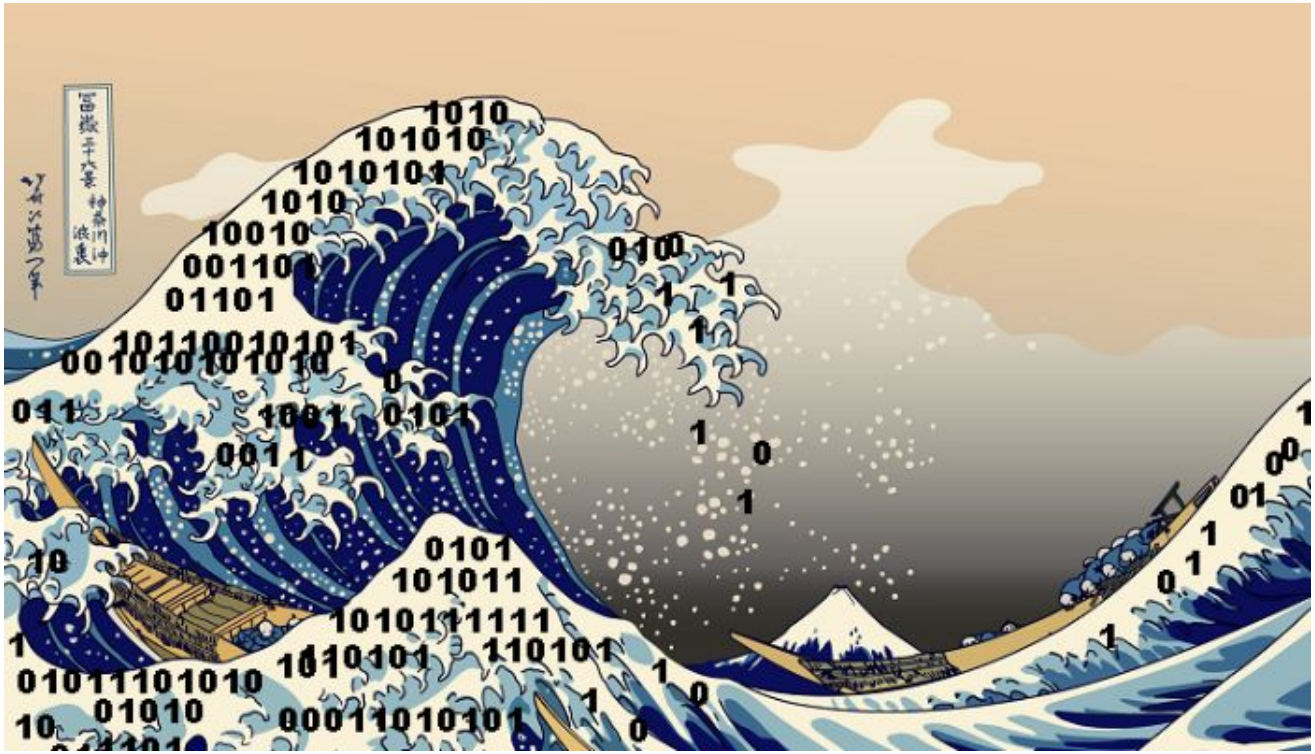
Bioversity International · cirad · INRA SCIENCE & IMPACT · IRD Institut de Recherche pour le Développement FRANCE · Montpellier SupAgro

# Why using workflow manager?

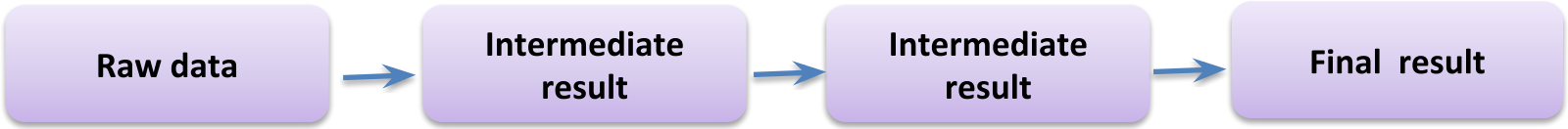


The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

# Why using workflow manager?




To create his own pipeline through an easy and user-friendly approach

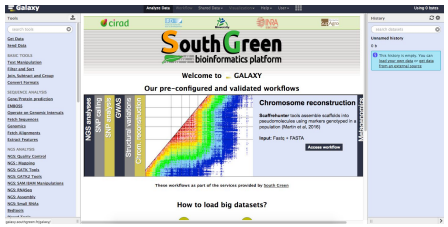


# Why using workflow manager?

- 3 solutions used and implemented by

**GUI tools** → **CLI tools**

Galaxy




Snakemake

TOGGLe

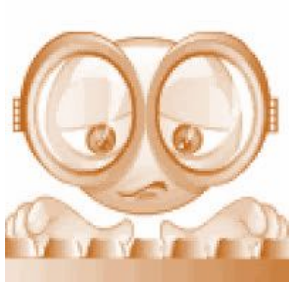
# Why using workflow manager?

- 3 solutions used and implemented by

Galaxy

*Snakemake*

TOGGLE

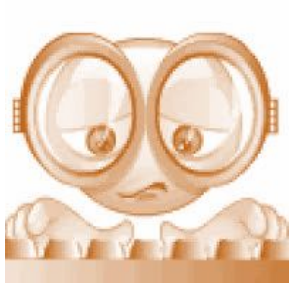


- Targets both biologists & bionformaticians

# Why using workflow manager?

- 3 solutions used and implemented by

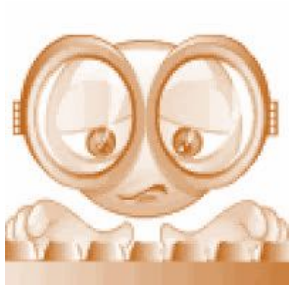
**Ease of use**
**Well-documented manual &workflow examples**



**Ease of development & evolution**

# Why using workflow manager?

- 3 solutions used and implemented by



SNP detection
GWAS
population genetics
transcriptome assembly
genome assembly
transcriptomics
structural variant detection
phylogeny
differential expression

# Why using TOGGLE?

**Pipeline & data sanity controls**

## A robust bioinformatics framework

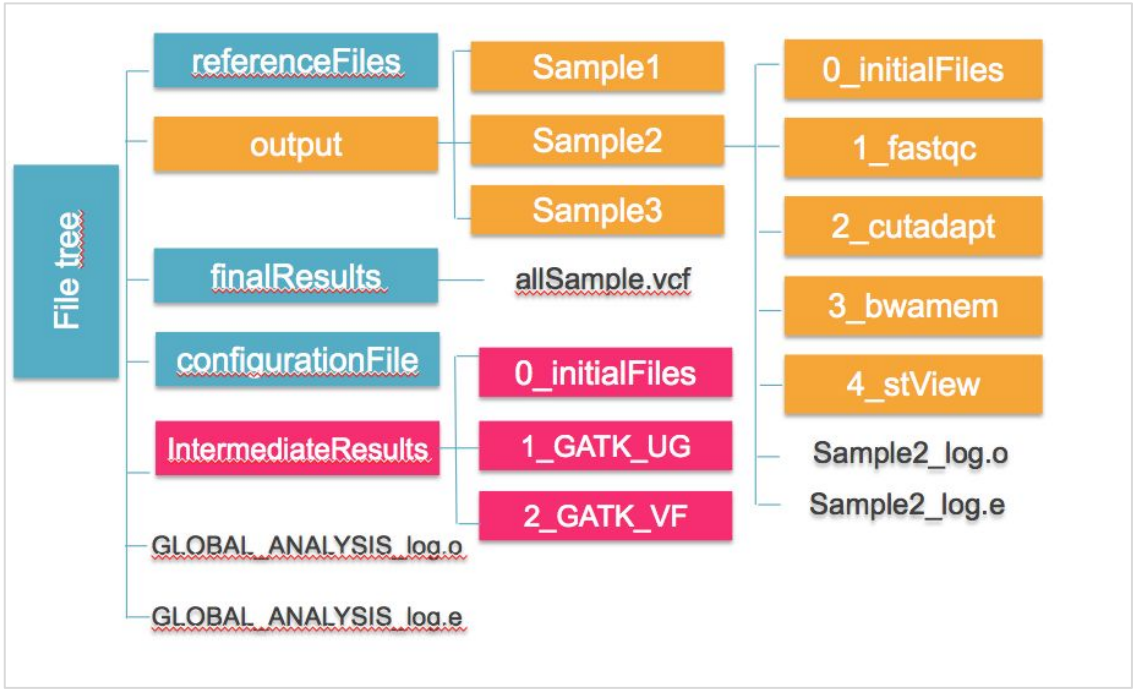
File format & content
Pipeline content

Missing but requested steps for ensuring the pipeline running

# Why using workflow manager?

**Pipeline & data sanity controls**

**Reproducibility & Traceability**



# Why using workflow manager?

Pipeline & data sanity controls

**A robust bioinformatics framework**

Reproducibility & Traceability

Error tracking & reentrancy

# Why using workflow manager?

Pipeline & data sanity controls


?

**A robust bioinformatics framework**

Reproducibility & Traceability

Large numbers of sample analyzed

Error tracking & reentrancy

# Why using workflow manager?

Pipeline & data sanity controls

HPC & Parallel execution

**A robust bioinformatics framework**

Reproducibility & Traceability

Large numbers of sample analyzed

Error tracking & reentrancy

# Workflow “Détection de SNPs”

| | 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** | 1 to 10000 | 1 to 50 |
| **Quota (related to infra)** | Disk space “/data/projects” 500Go to 1T | IRD 100Go data<br>Cirad 100Go => 300Go |
| **Parallelization (related to infra conf)** | IRD 300 cores<br>Cirad 600 cores | IRD *16 cores / one node*<br>Cirad *200 cores* |
| **Number of tools available** | 120 | 500 installed (total : 5500) |
| **Post-analyses Graphical figures** | No | Yes |

South Green bioinformatics platform

# What's TOGGLe ?

TOGGLe

# What is TOGGLe ?

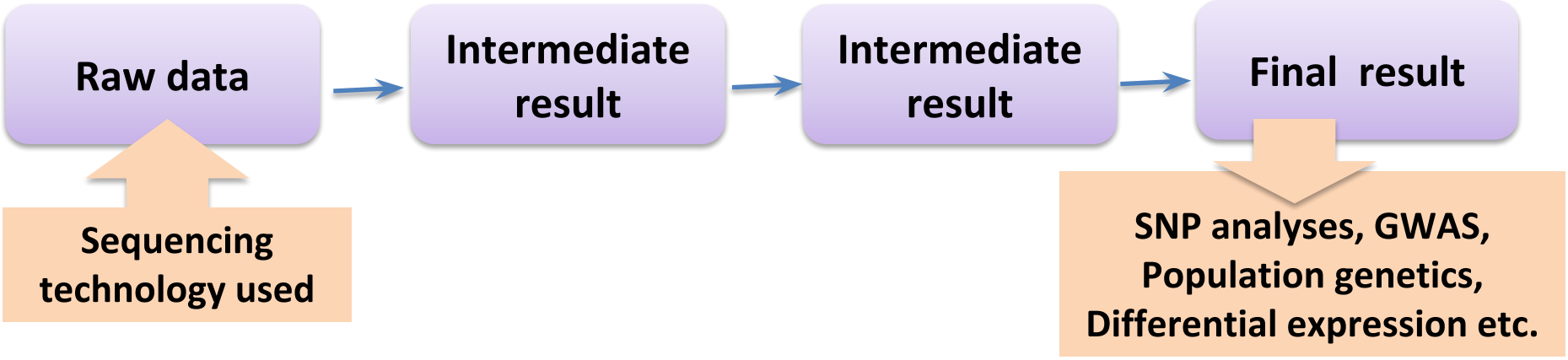
- A toolbox to perform large-scale NGS analyses

19 modules, 120 functions
120 open-source tools

# What is TOGGLe ?

- A toolbox to perform large-scale NGS analyses

19 modules, 120 functions
120 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 ?

GBS | RADSeq | RNASeq | WGS

Data preprocessing

Mapping & Post-Mapping

Assembly

TOGGLe

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<br>checkFormatFasta<br>checkFormatSamOrBam<br>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<br>crac |
| Bowtie | 1.1.2 | bowtiebuild<br>bowtie |
| Bowtie2 | 2.2.9 | bowtie2build<br>bowtie2 |

# 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***

# 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. |
| -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

TOGGLE

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

**Create your own workflow**

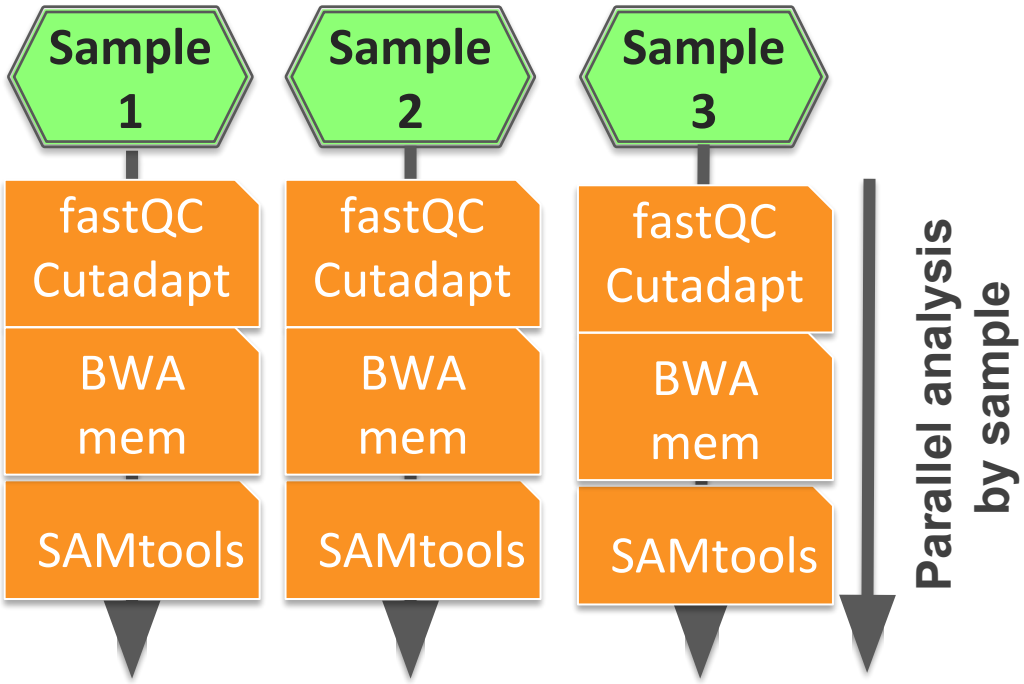
**$order**
1=fastqc
2=cutadapt
3=bwa mem
4=samToolsView
1000=gatkHaplotypeCaller
1001=gatkVariantFiltration

Step number < 1OOO

**Parallel analysis by sample**

# To create your own workflow

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

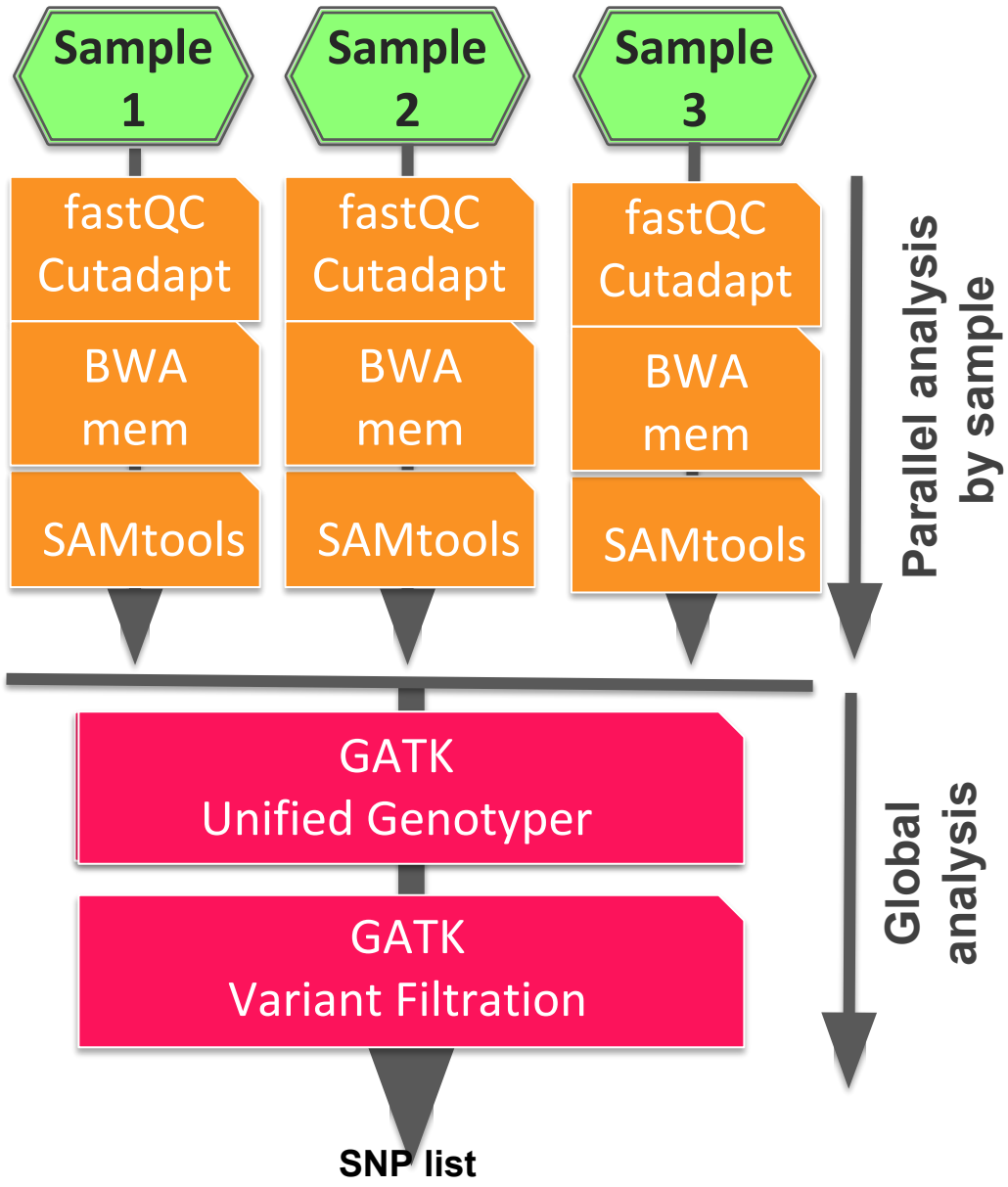


# To create your own workflow

**$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

Developer manual

Screencast

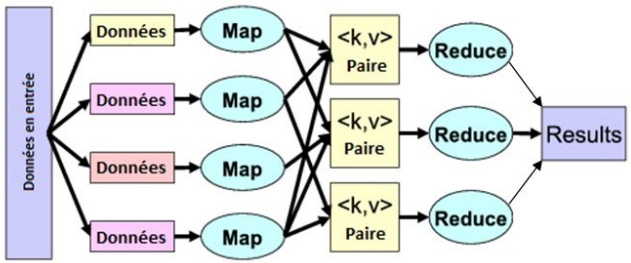
Pre-defined Workflow files


https://github.com/SouthGreenPlatform/TOGGLE

# What's next ?

- **New data analysis** : metagenomics, pacbio assembly, pangenome

- **New features** : complete reentrancy, non-sequential pipelines, embarrasingly-parallel approaches




# TOGGLe 's team

IRD Institut de Recherche pour le Développement FRANCE

**UMR DIADE**

**François Sabot**

**Christine Tranchant-Dubreuil**

**Ndomassi Tando**

**Alexis Dereeper**

Cécile Monat

Mawussé Agbessi

Souhila Amazougarene

Abdoulaye Diallo

Laura Helou

Ayité Kougbeadjo

cirad

**UMR BGPI**

**Sébastien Ravel**

ADN Agence d'information génétique

**Julie Orjuela-Bouniol**

cirad

INRA SCIENCE & IMPACT

**UMR AGAP**

Cédric Farcy

Enrique Ortega-Abboud

Gautier Sarah

Maryline Summo

SouthGreen bioinformatics platform