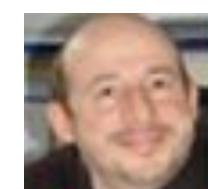
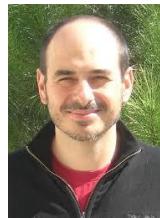


# Session de formation 2018



- 12 Mars**      Guide de survie à Linux : les commandes de base pour débuter sur un serveur linux
- 13 Mars**      Linux avancé : manipuler et filtrer des fichiers sans connaissance de programmation
- 15 Mars**      Initiation à l'utilisation du cluster bioinformatique itrop
- 22 Mars**      Initiation à git
- 23 Mars**      **Initiation aux gestionnaires de workflow South Green:  
Galaxy ou TOGGLE**
- 26 Mars**      Initiation aux analyses de données transcriptomiques





IRD

Institut de Recherche  
pour le Développement

# South Green

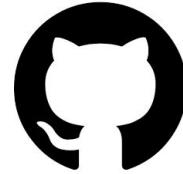
bioinformatics platform



plateau i-trop



[www.southgreen.fr](http://www.southgreen.fr)



<https://github.com/SouthGreenPlatform>



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

# Session de formation 2018



- Toutes nos formations :  
<https://southgreenplatform.github.io/trainings/>
- Topo & TP : [Workflow Managers](#)
- Environnement de travail : [Logiciels à installer](#)



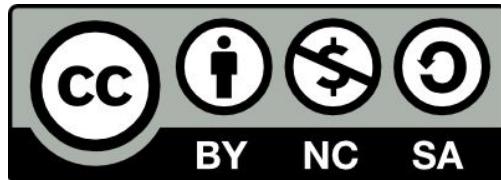
# Workflow Manager

TOGGLE

Galaxy

[www.southgreen.fr](http://www.southgreen.fr)

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



## The objectif!

Using South Green Workflow Managers to build automatically your own pipelines



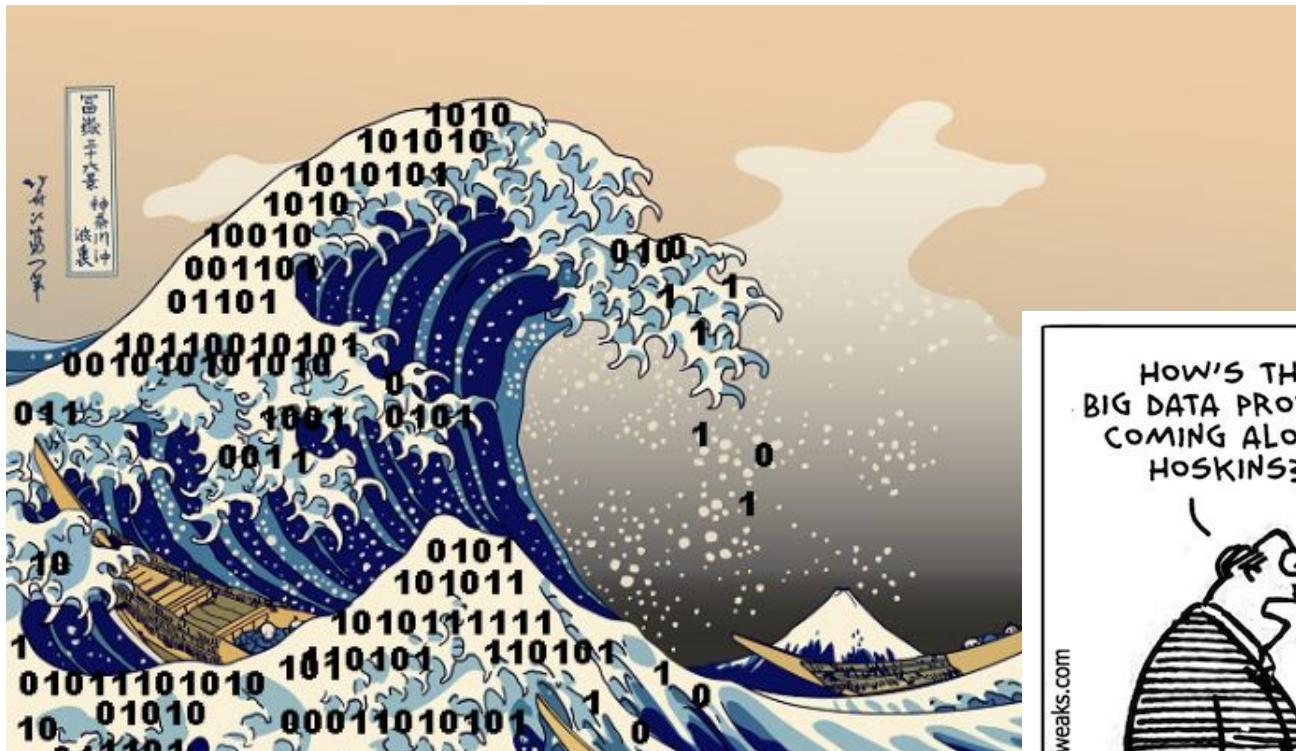
## Applications

Know all about our 2 main workflow managers developed by

- Use these tools
- Build your own workflow
- Common use case : SNP calling from reads (illumina) of 3 samples



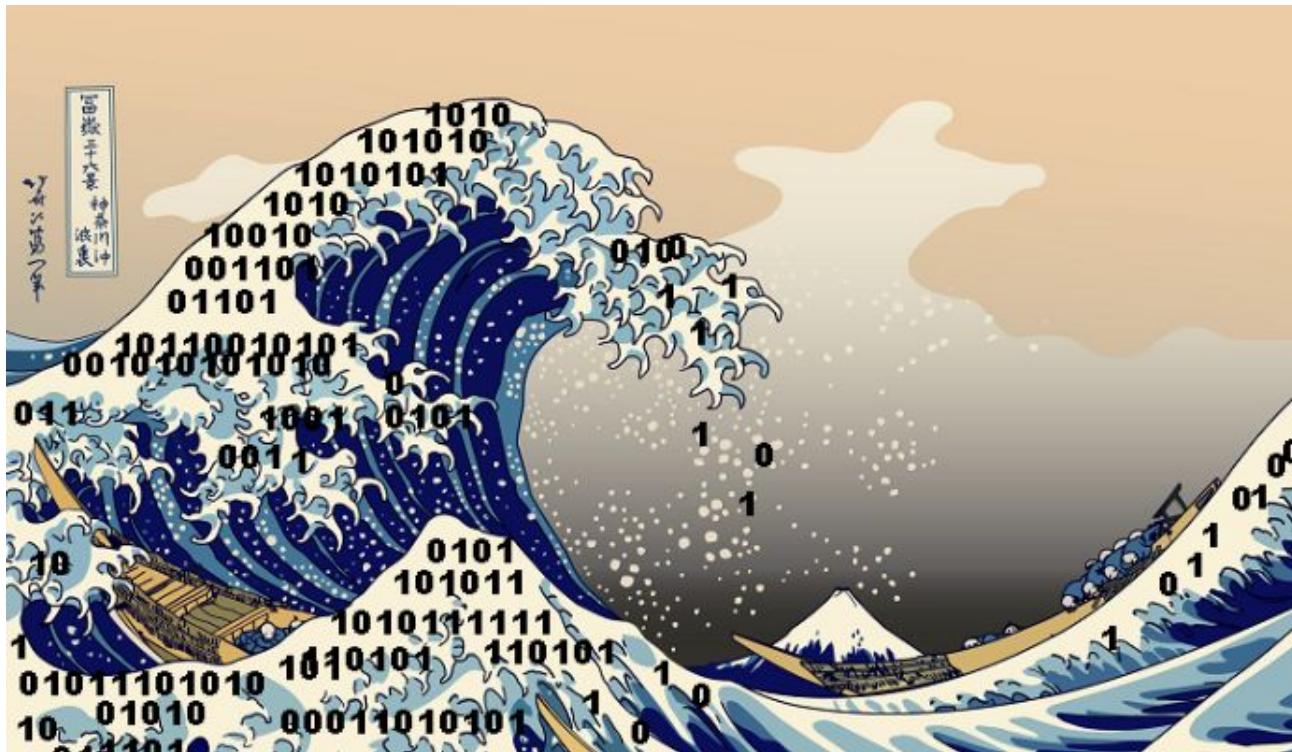
# Why using a workflow manager?



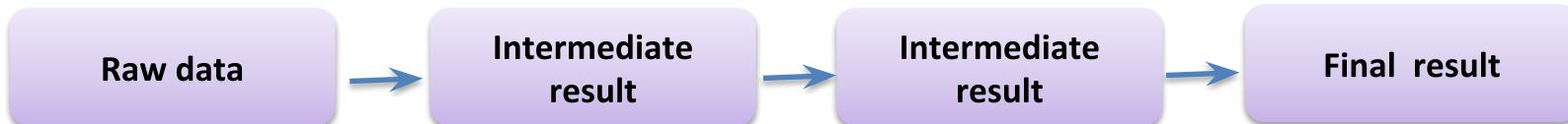
The Great Wave off Kanagawa, Hokusai      @amitechsolutions.com



# Why using a workflow manager?



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



- 3 solutions used and implemented by

GUI tools



CLI tools

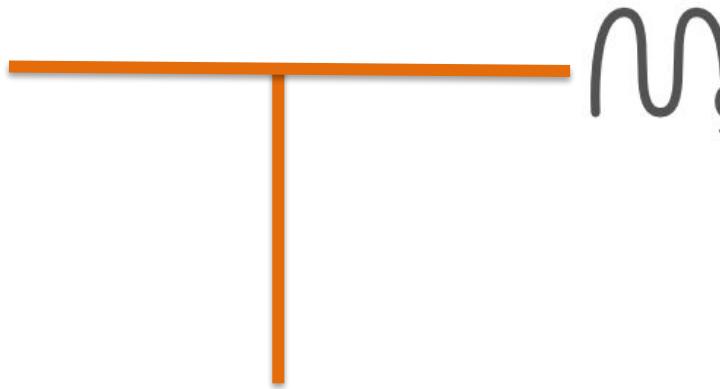


TOGGLE

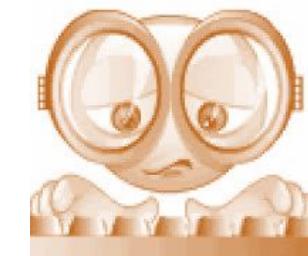


# Why using workflow manager?

- 3 solutions used and implemented by



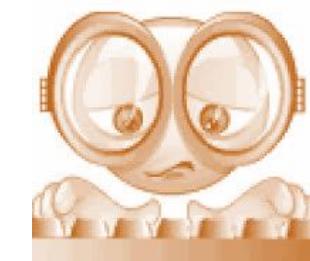
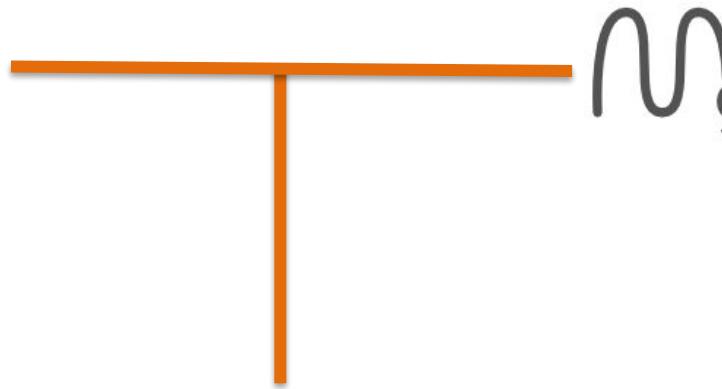
*Snakemake*



- Targets both biologists & bioinformaticians

# Why using a workflow manager?

- 3 solutions used and implemented by

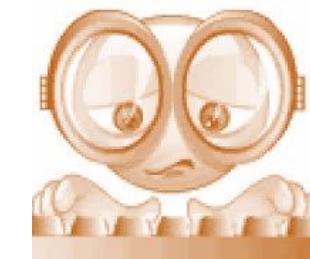
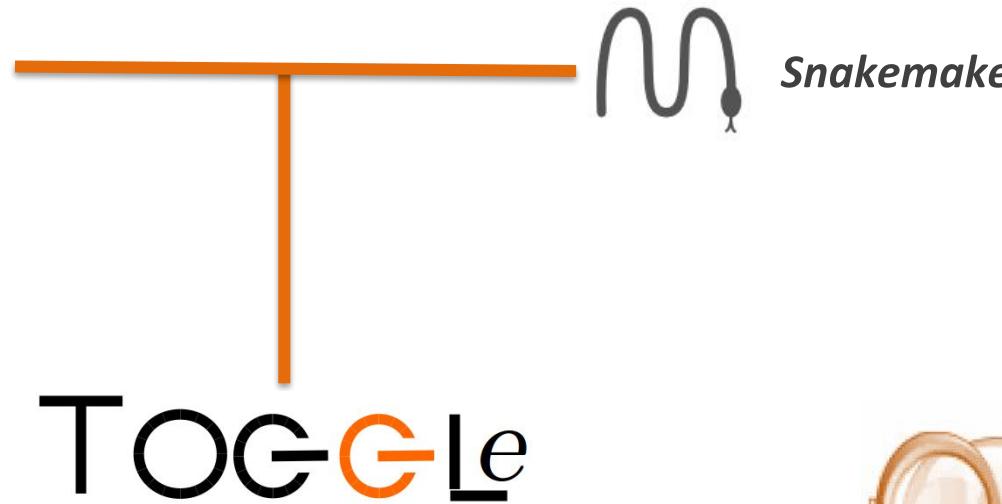


**Ease of use**  
**Well-documented**  
**manual & workflow**  
**examples**

**Ease of development**  
**&**  
**evolution**

# Why using a workflow manager?

- 3 solutions used and implemented by



**GWAS**  
transcriptome assembly  
**structural variant detection**  
**phylogeny**

**SNP detection**  
**population genetics**  
genome assembly  
transcriptomics  
differential expression

# Why using a workflow manager?

Pipeline & data  
sanity controls

## A robust bioinformatics framework



File format & content  
Pipeline content

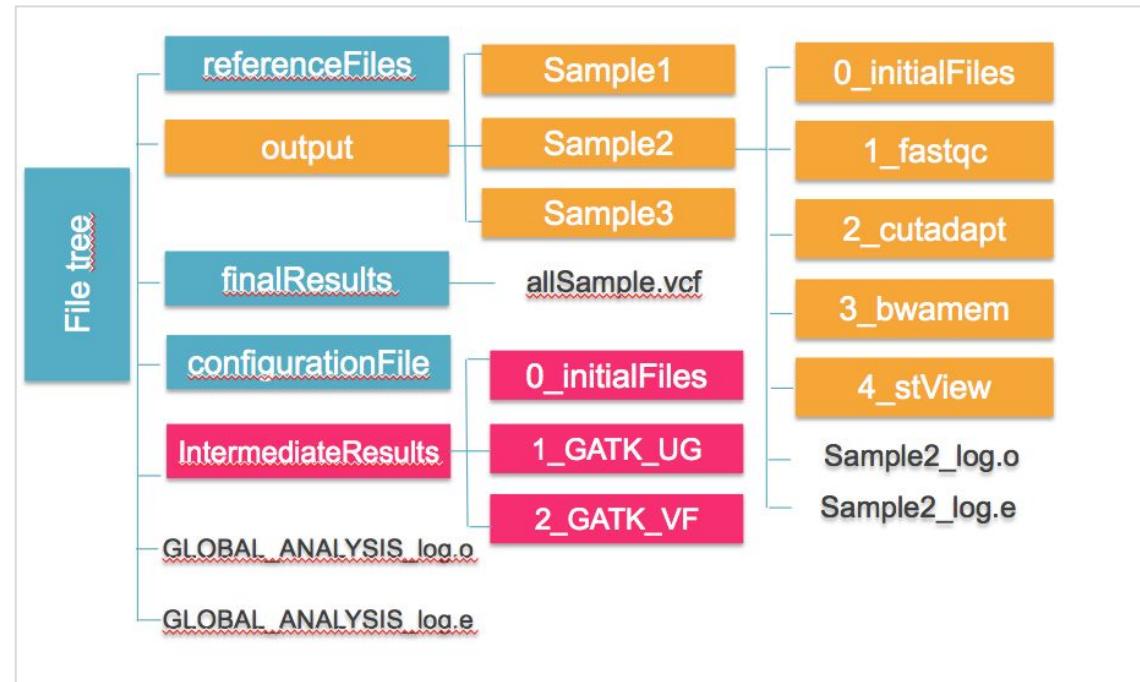


Missing but requested steps for ensuring  
the pipeline running

# Why using a workflow manager?

Pipeline & data  
sanity controls

Reproducibility  
& Traceability



# Why using a workflow manager?

Pipeline & data  
sanity controls

A robust bioinformatics  
framework

Reproducibility  
& Traceability

Error tracking &  
reentrancy

# Why using a workflow manager?

Pipeline & data  
sanity controls



A robust bioinformatics  
framework

Reproducibility  
& Traceability

Large numbers  
of sample  
analyzed

Error tracking &  
reentrancy

# Why using a workflow manager?

Pipeline & data  
sanity controls

HPC & Parallel  
execution

A robust bioinformatics  
framework

Reproducibility  
& Traceability

Large numbers  
of sample  
analyzed

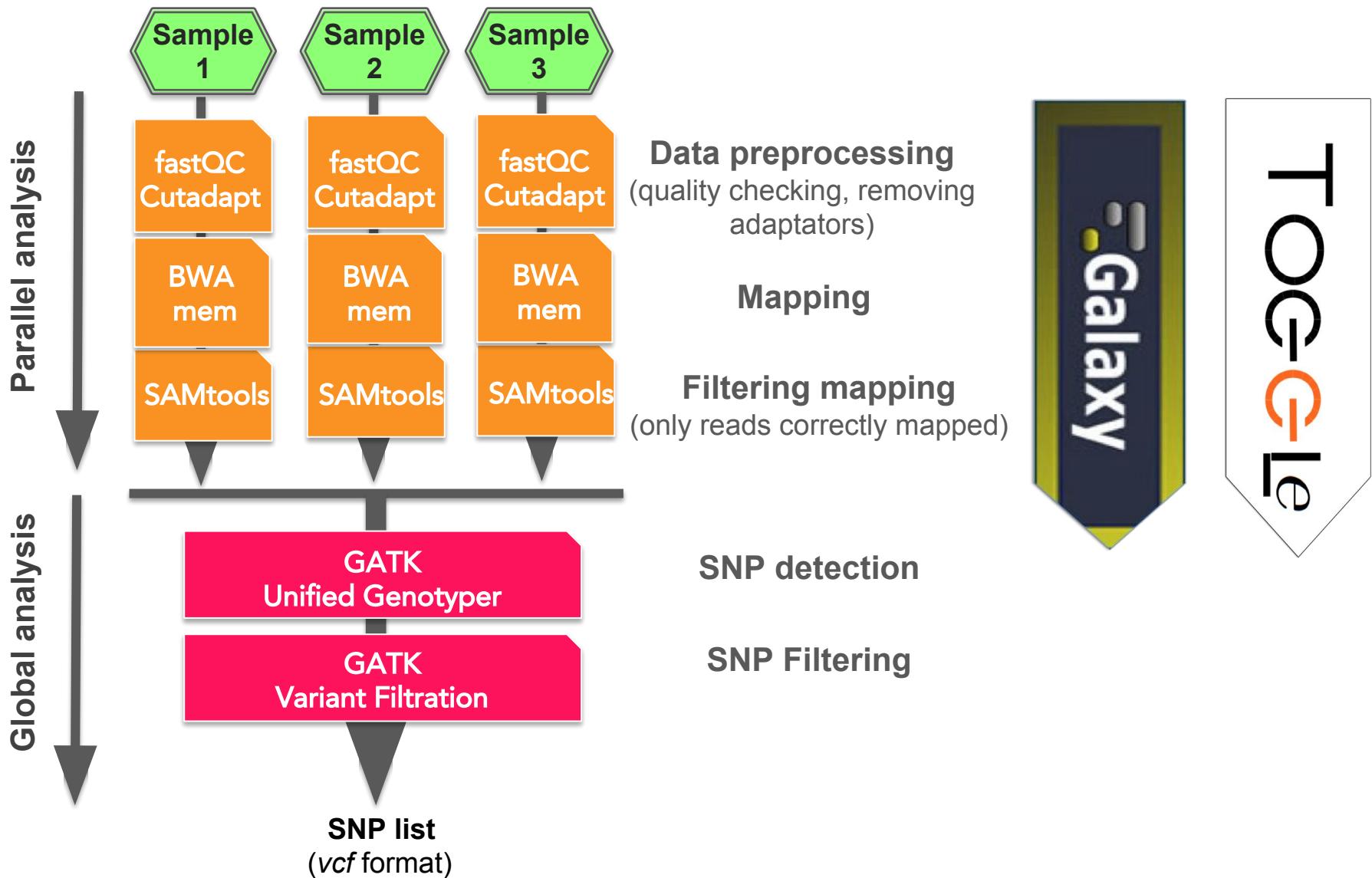
Error tracking &  
reentrancy

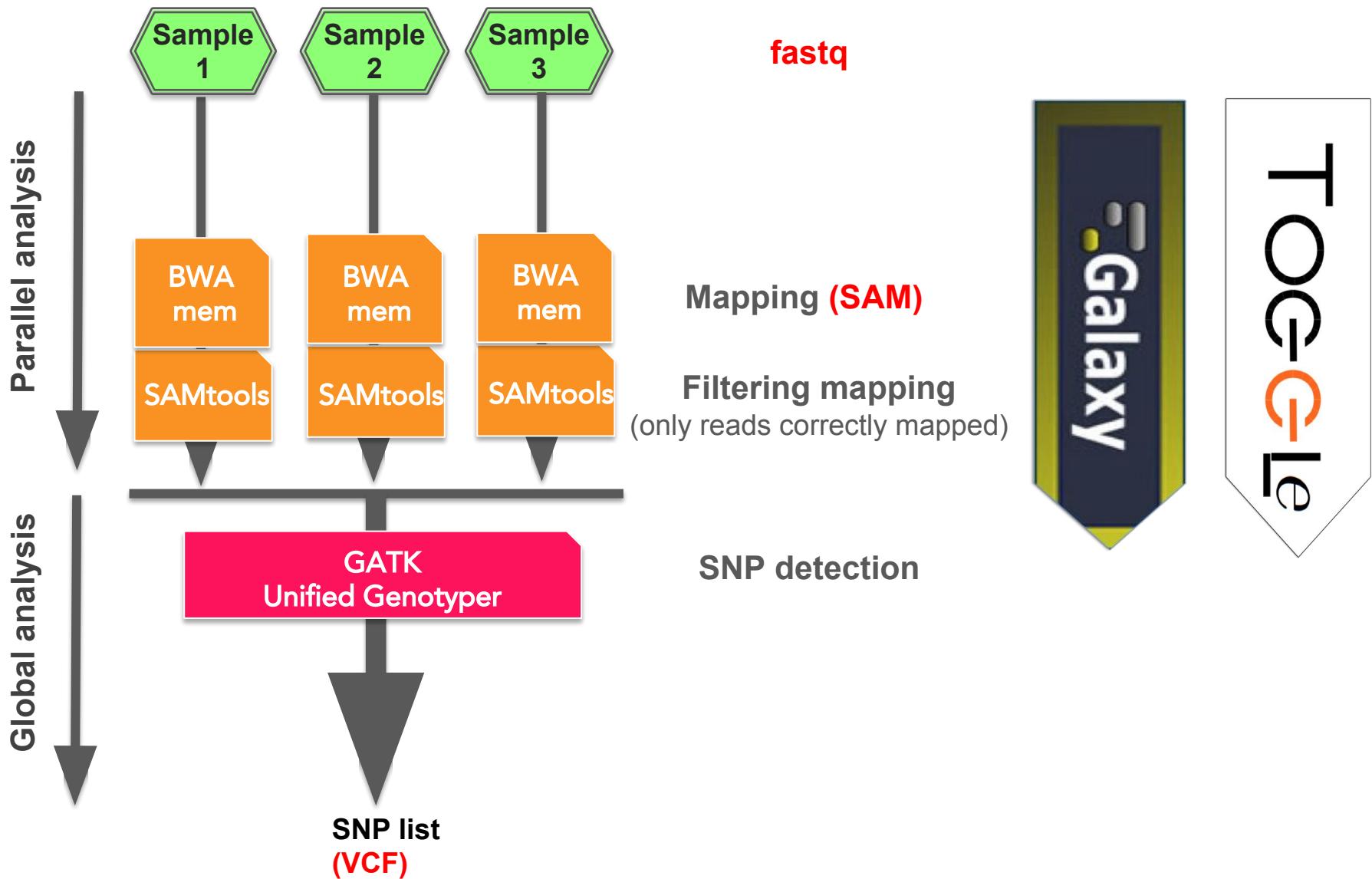


Interface	Command line	GUI (Web interface)
Predefined Pipelines	SNP calling, RNASeq and WGS large scale	Metagenomics, RNASeq, SNP calling, post-analyses
Number of Samples	1 to 10000	1 to 50
Quota (related to infra)	Disk space “/data/projects”	IRD 100Go data Cirad 100Go => 300Go
Parallelization (related to infra conf)	IRD 300 cores Cirad 600 cores	IRD 16 cores / one node Cirad 200 cores
Number of tools available	120	500 installed (total : 5500)
Post-analyses Graphical figures	Not yet	Yes



# Use case : Workflow “SNPs calling”





# Format FASTQ

Seq 1 @HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593I Sequencing info  
GAGAAGTTAACAGCTGGTATTATTTGTAAACATI  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:215:593I  
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhEI

Seq 2 @HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:234:551I Nucleotide sequence  
TGGGACTTTATCTGGAGGAGTGTGGAAAGCCATTI  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:234:551I  
hhI

Seq 3 @HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:338:194I  
TGGTTTATGCAGAAATTCTAGAATAAGGGTAACCTTI Quality score in ASCII  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:338:194I  
hhI

Seq 4 @HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:363:717I  
TCTCAGAAAATTGTTGTGATGTGTGTATTCAACTATI  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:363:717I  
hhI

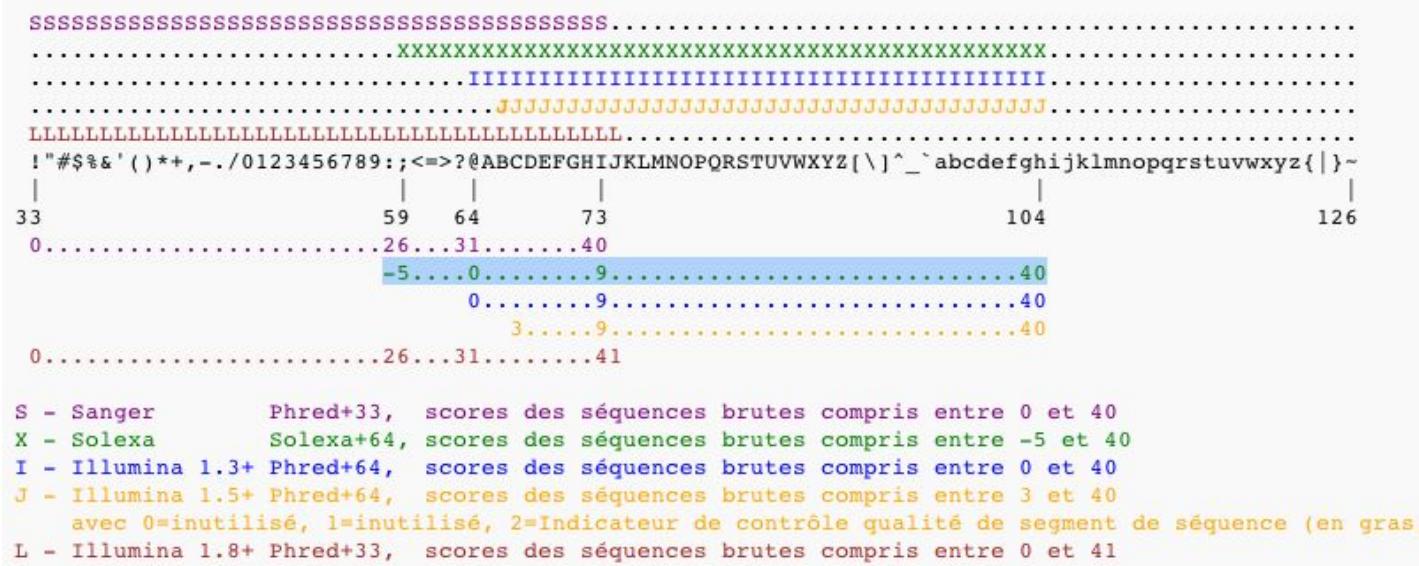
Seq 5 @HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:208:209I  
TTGATTTAACTCTGACAAAATAAACAAAAGTCTTAGGI  
+HWI-EAS236\_3\_FC\_20BTNAAXX:2:1:208:209I  
hhGh9I quality score are encoded with a single ASCII character for brevity

1 séquence = 4 lignes

- @ identifiant séquence
- séquence
- + nom séquence name (optionel).
- Qualité de la séquence (un caractère / base)

# Format FASTQ

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée



Toujours utilisé l'encodage sanger !

**SAM format :** <http://samtools.sourceforge.net/samtools.shtml>

Col	Name	Description
1	<b>QNAME</b>	Query NAME of the read or the read pair
2	<b>FLAG</b>	bitwise FLAG (pairing, strand, mate strand, etc.)
3	<b>RNAME</b>	Reference sequence NAME
4	<b>POS</b>	1-based leftmost POSition of clipped alignment
5	<b>MAPQ</b>	MAPping Quality (Phred-scaled)
6	<b>CIGAR</b>	extended CIGAR string (operations: MIDNSHP)
7	<b>NRNM</b>	Mate Reference NaMe (`=' if same as RNAME)
8	<b>MPOS</b>	1-based leftmost Mate POSition
9	<b>ISIZE</b>	inferred Insert SIZE
10	<b>SEQ</b>	query SEQuence on the reference
11	<b>QUAL</b>	query QUALity (ASCII-33)

```

@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *

```

# Format VCF (Variant Call Format)

```

##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##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=A,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=GT,Number=1>Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1>Type=Integer,Description="Genotype Quality">
##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 NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 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
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3' '

```

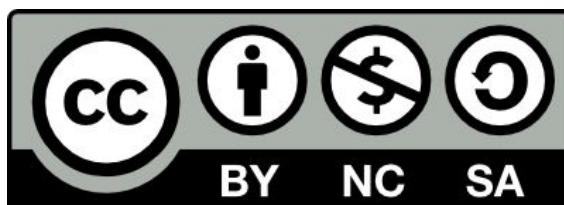
- **Variation 1 :** a good SNP
- **Variation 2 :** a possible SNP that has been filtered out because its quality is below 10
- **Variation 3 :** a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error)
- **Variation 4 :** a site that is called monomorphic reference (i.e. with no alternate alleles)
- **Variation 5 :** a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T).

# Formateurs itrop / South Green

- Alexis Dereeper
- Sébastien Ravel
- Christine Tranchant-Dubreuil



# 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/>