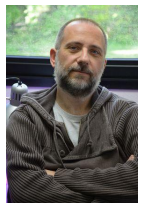
# Session de formation 2018

South Green bioinformatics platform

| | |
|---|---|
| **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 |

Bioversity International






Institut de Recherche pour le Développement

diade IPME









# SouthGreen bioinformatics platform






agap UMR BGPI Biologie et Génétique des Interactions Plante-Parasite LSTM Laboratoire des Symbioses Tropicales et Méditerranéennes

INRA SCIENCE & IMPACT cirad



Bioversity International

Institut de Recherche pour le Développement

SouthGreen bioinformatics platform

diade IPME

UMI233 TransVIH-MI

MiVEGEC

Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle

**plateau i-trop**

agap UMR BGPI Biologie et Génétique des Interactions Plante-Parasite LSTM Laboratoire des Symbioses Tropicales et Méditerranéennes

INRA SCIENCE & IMPACT

cirad

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 : **Linux For Dummies**
- Environnement de travail : **Logiciels à installer**

# Workflow Manager

TOGGLE

Galaxy

www.southgreen.fr

https://southgreenplatform.github.io/trainings



Bioversity International

cirad

INRA SCIENCE & IMPACT

IRD Institut de Recherche pour le Développement FRANCE

Montpellier SupAgro

# Objectifs du module

## The objectif!

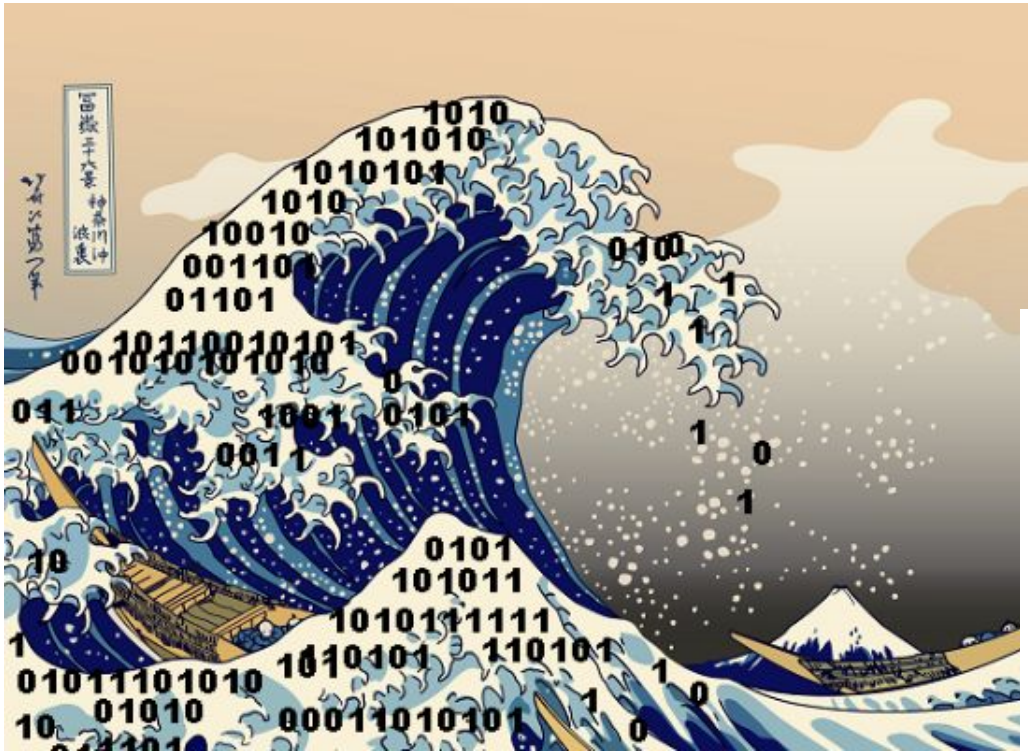
**Utiliser des gestionnaires de workflow pour lancer vos pipelines d'analyse de manière automatique**

## Applications

Connaître les 2 principaux gestionnaires de workflow développés par la plateforme : Galaxy et TOGGLe

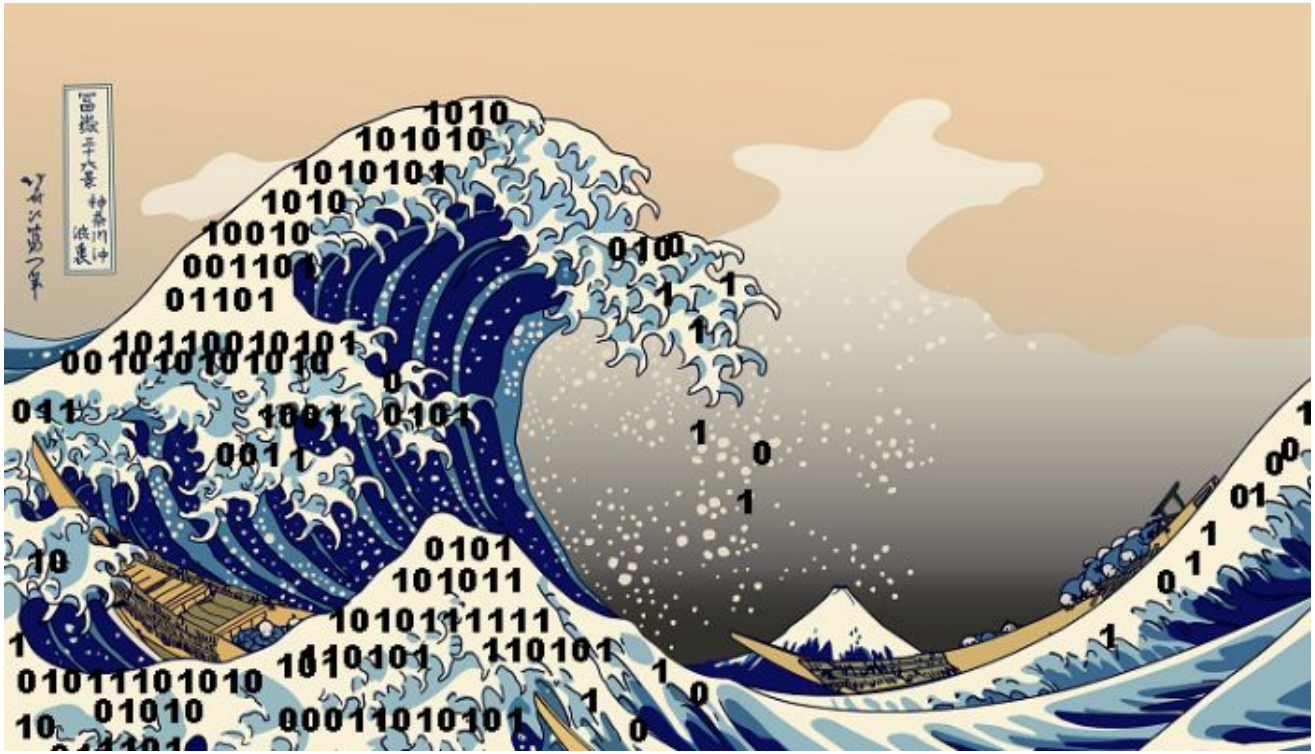
- Prise en main des outils
- Définir son propre workflow
- Mise en application commune : détection des SNPs en partant de reads illumina générés sur 3 individus

# Why using workflow manager?

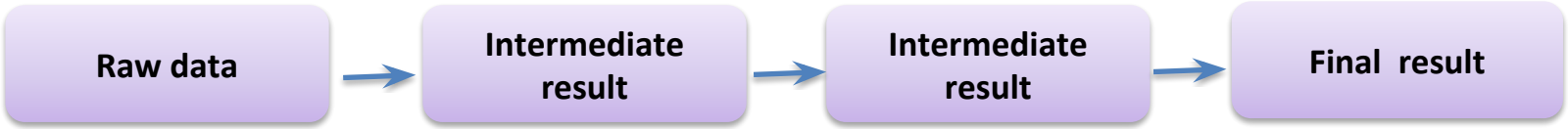


The Great Wave off Kanagawa, Hokusa @amitechsolutions.com

# Why using workflow manager?




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

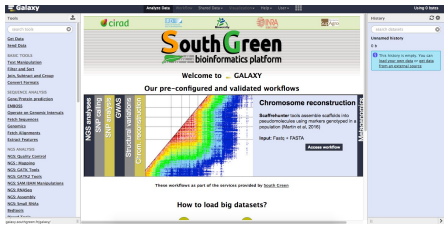


# Why using workflow manager?

- 3 solutions used and implemented by

**GUI tools** → **CLI tools**

Galaxy

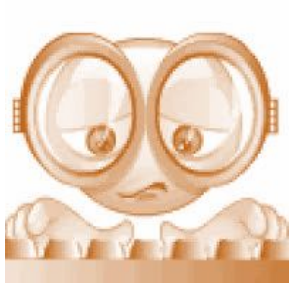



Snakemake

TOGGLe

# Why using workflow manager?

- 3 solutions used and implemented by

- Targets both biologists & bionformaticians

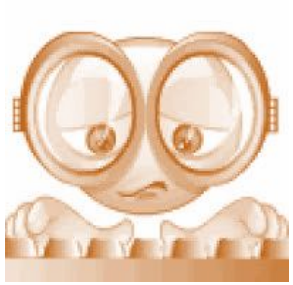
# Why using workflow manager?

- 3 solutions used and implemented by

Galaxy

Snakemake

TOGGLE



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

**Ease of development & evolution**

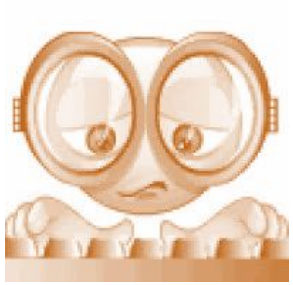
# Why using workflow manager?

- 3 solutions used and implemented by

Galaxy

Snakemake

TOGGLE



# Why using TOGGLE?

**Pipeline & data sanity controls**

## A robust bioinformatics framework

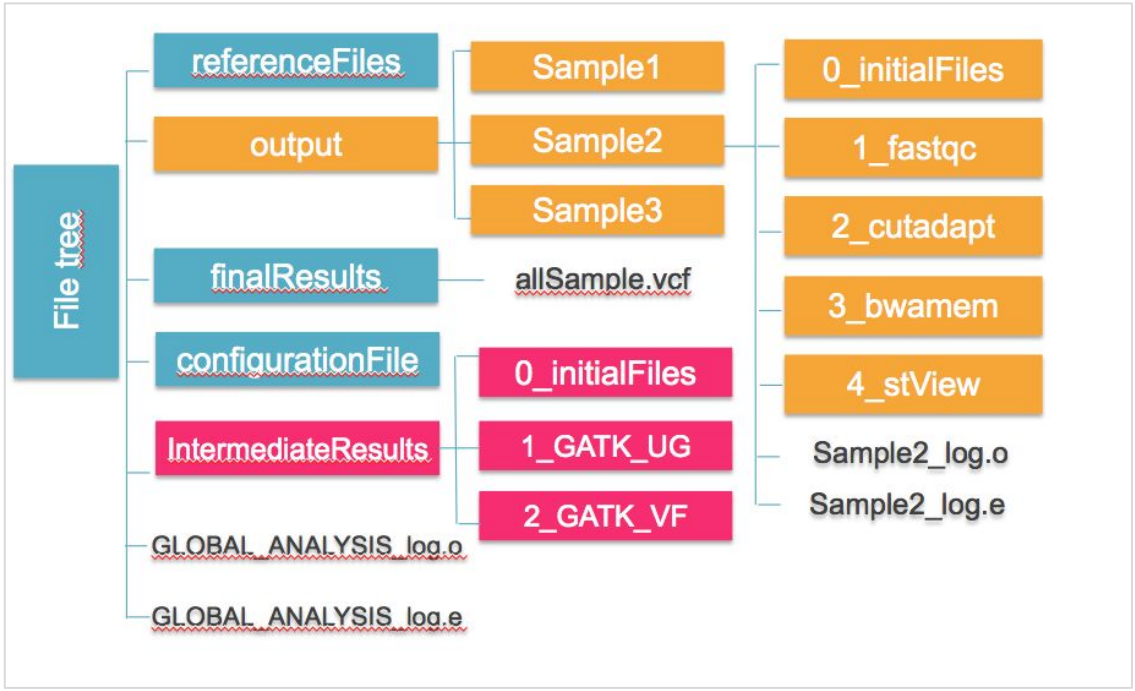
File format & content
Pipeline content

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

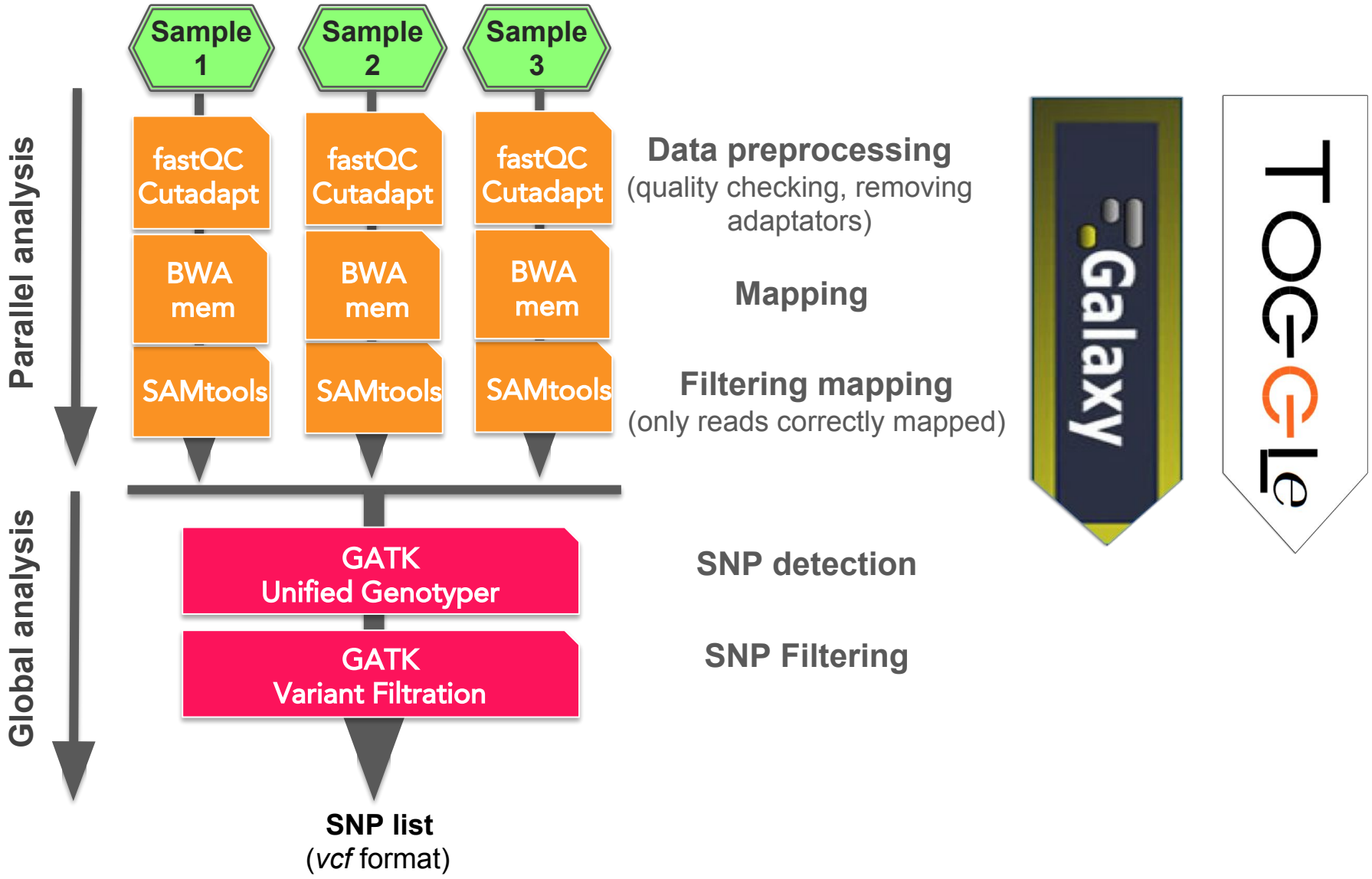
Error tracking & reentrancy

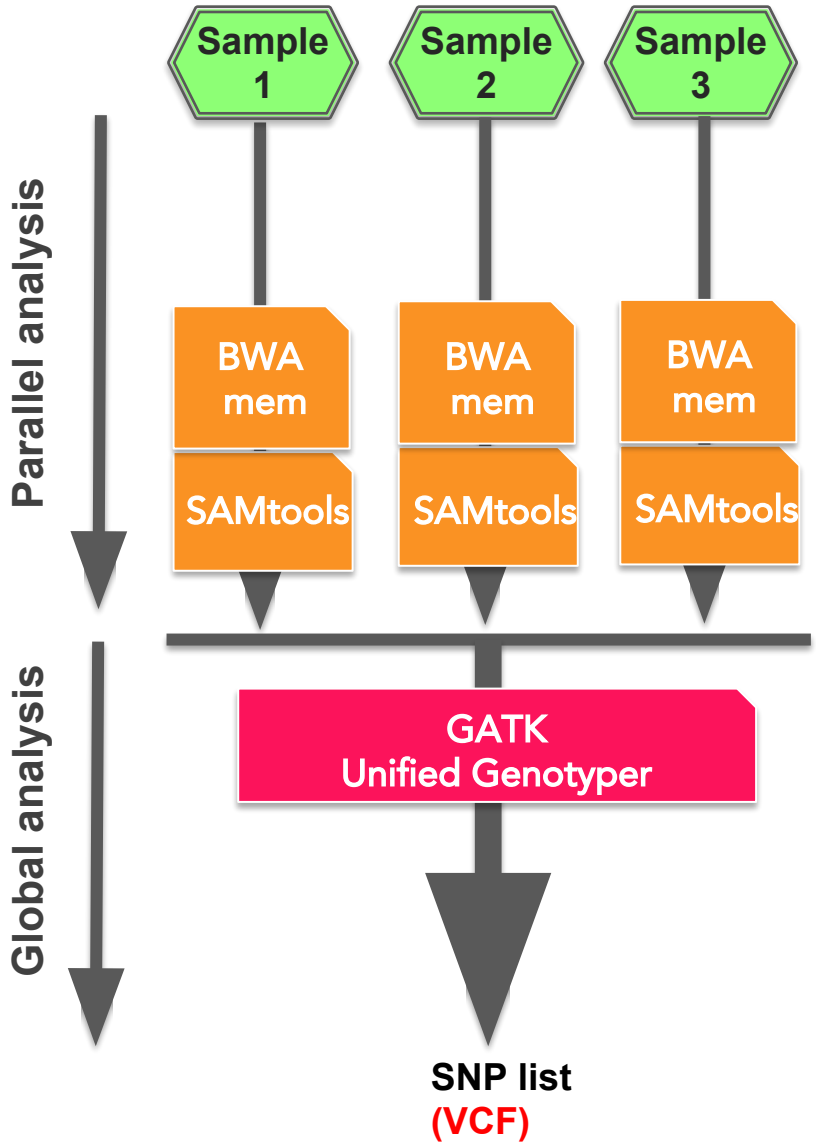
# Workflow “Détection de SNPs”

| | 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

# Mise en application: description du workflow “Détection de SNPs”


Sample 1
Sample 2
Sample 3
fastQC Cutadapt
fastQC Cutadapt
fastQC Cutadapt
BWA mem
BWA mem
BWA mem
SAMtools
SAMtools
SAMtools
GATK Unified Genotyper
GATK Variant Filtration
Parallel analysis
Global analysis
Data preprocessing
(quality checking, removing adaptators)
Mapping
Filtering mapping
(only reads correctly mapped)
SNP detection
SNP Filtering
SNP list
(vcf format)
Galaxy
TOGGLE




fastq

Mapping (SAM)

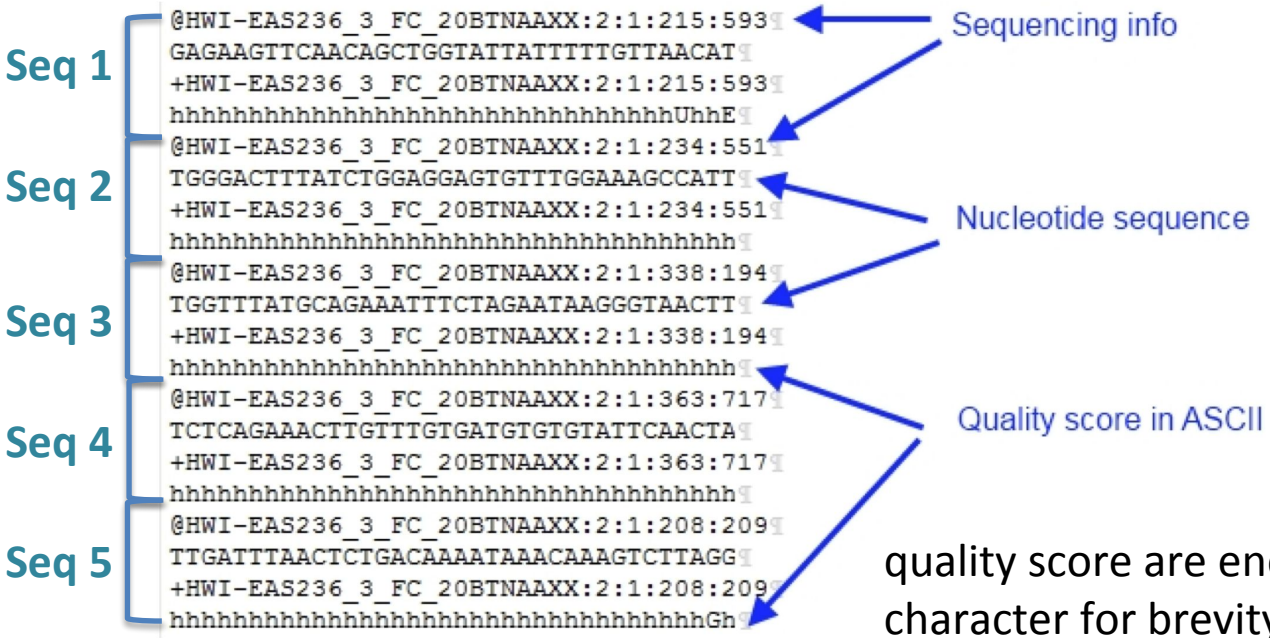
Filtering mapping
(only reads correctly mapped)

SNP detection

Galaxy

TOGGLE

# Format FASTQ



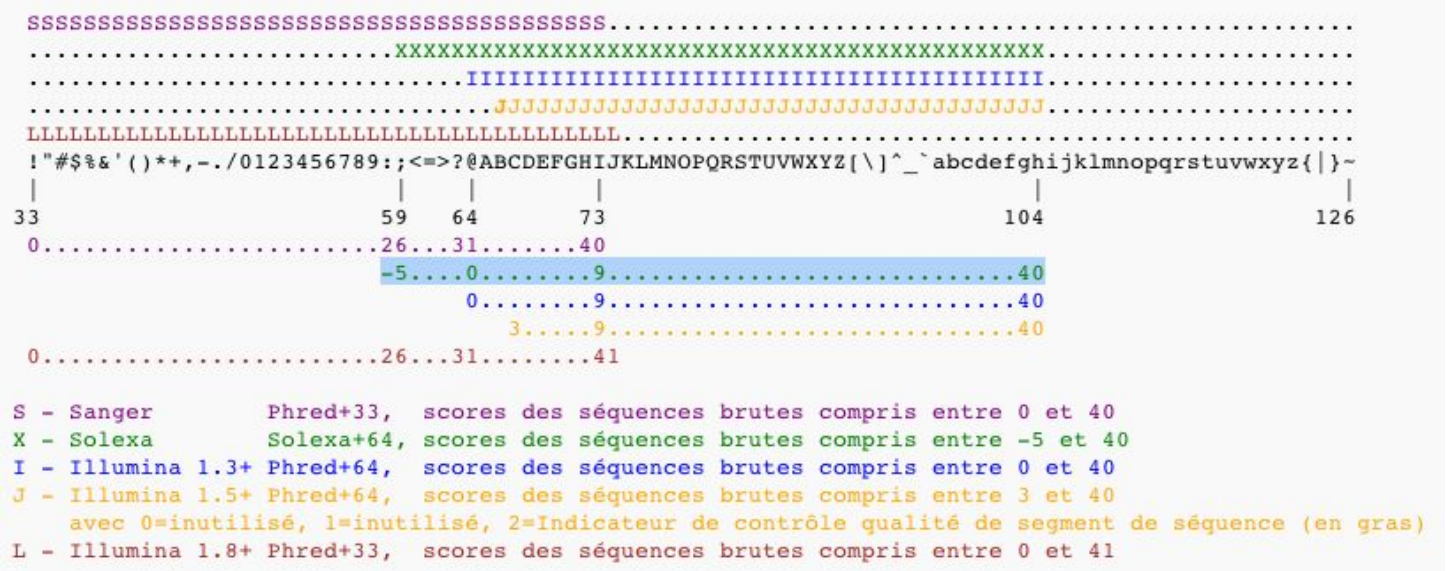

quality score are encoded with a single ASCII character for brevity

**1 séquence = 4 lignes**

- @ identifiant séquence
- sequence
- + 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 !**

# Format SAM (Sequence Alignment Map) / BAM

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

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