Ce TP a pour but de vous initier au traitement des données brutes NGS de manière à les mettre en forme pour les utiliser dans une analyse SNP

TD Goals

1- Understanding a **FASTQ** file

2- Cleaning Illumina/**FASTQ** dataset

3- Perform an **Assembly**

4- Perfom a **mapping** of Illumina data upon a reference sequence

5- Clean a multiple **SAM** file

FASTQ file → TEXT file

STRUCTURE:

```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA
CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGT
+
cfffcfeffdeefefffcfffffffcffeffffdffffafcfffffdffffdfefeddf^eececfffdfcbffb
@HWUSI-EAS454_0006:1:37:16314:3410#CTTGTA
AGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGTGGTGGCCG
+
`bTbbcccccceeeeeceeeeeccYeedded`ceec]dddde^a`deeeec\`dddcbaadadYd`]]Jc_^bc^^\
```


Le format utilisé couramment pour les données brutes est le FASTQ (Fasta + Quality)

```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA
CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGT
+
cfffcfeffdeefefffcffffffffcffeffffdffffafcffffffdffffdfefeddf^eececfffdfcbffb
```


Une seule séquence de type NGS est représentée en format FASTQ (Fasta + Quality) par 4 lignes différentes

Son nom doit etre unique dans le fichier

Sa séquence est au format en général ATGC mais peut etre aussi au format IUPAC, incluant les codes dégénérés YRM...

```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA
CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGT
+
cfffcfeffdeefefffcffffffffcffeffffdffffafcfffffdffffdfefeddf^eececfffdfcbffb
```


Cette ligne peut n'etre que le symbole '+' ou bien '+' suivi du nom de la sequence (ex +MASEQUENCE)

La qualité est codée en ASCII, manière informatique de coder des valeurs numériques sur un seul caractère

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.......................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.......................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```


L'echelle de codage dépend du type de codage. L'echelle a toujours 40 unites, mais son démarrage dépend du type de codage

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                        |    |        |                              |                   |
33                      59   64       73                            104                 126
```

```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGT

+
cfffcfeffdeefefffcfffffffcffeffffdffffafcfffffdffffdfefeddf^eececfffdfcbffb
```

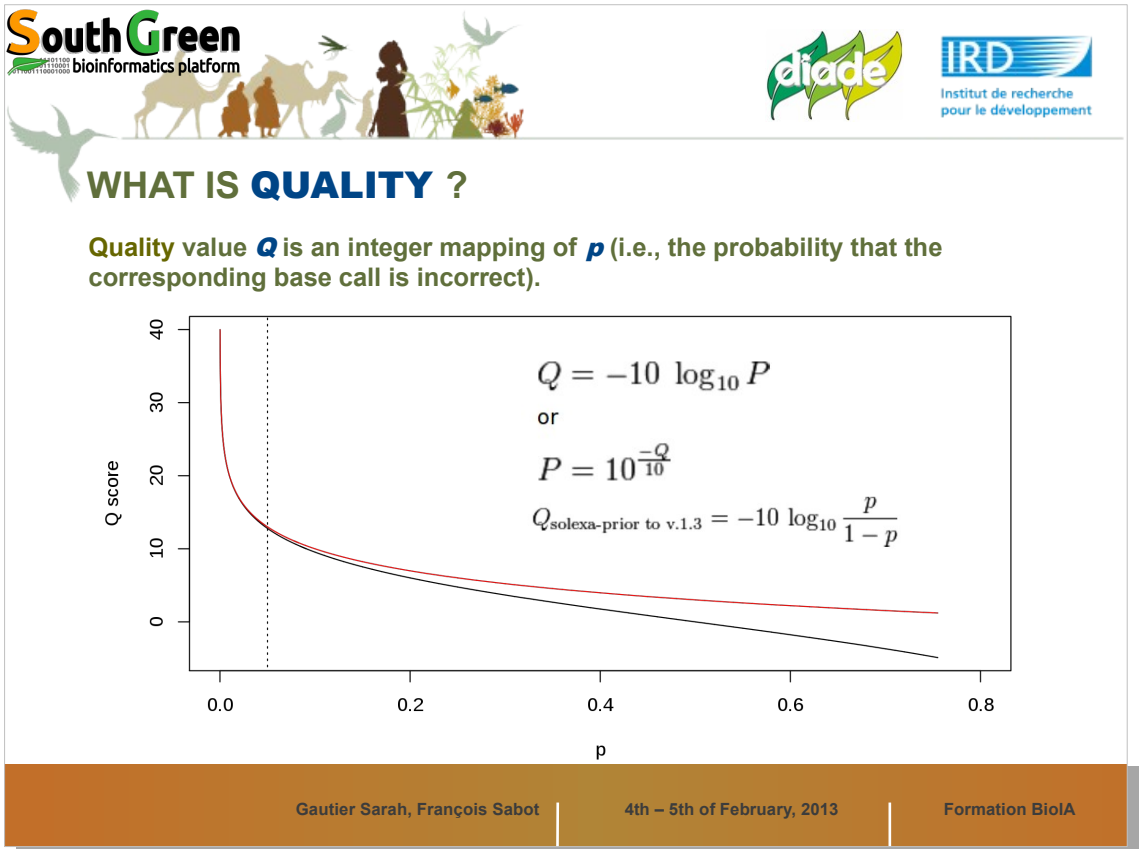

Ici, le 'f' Illumina

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                        |    |        |                              |                    |
33                      59   64       73                            104                  126
```

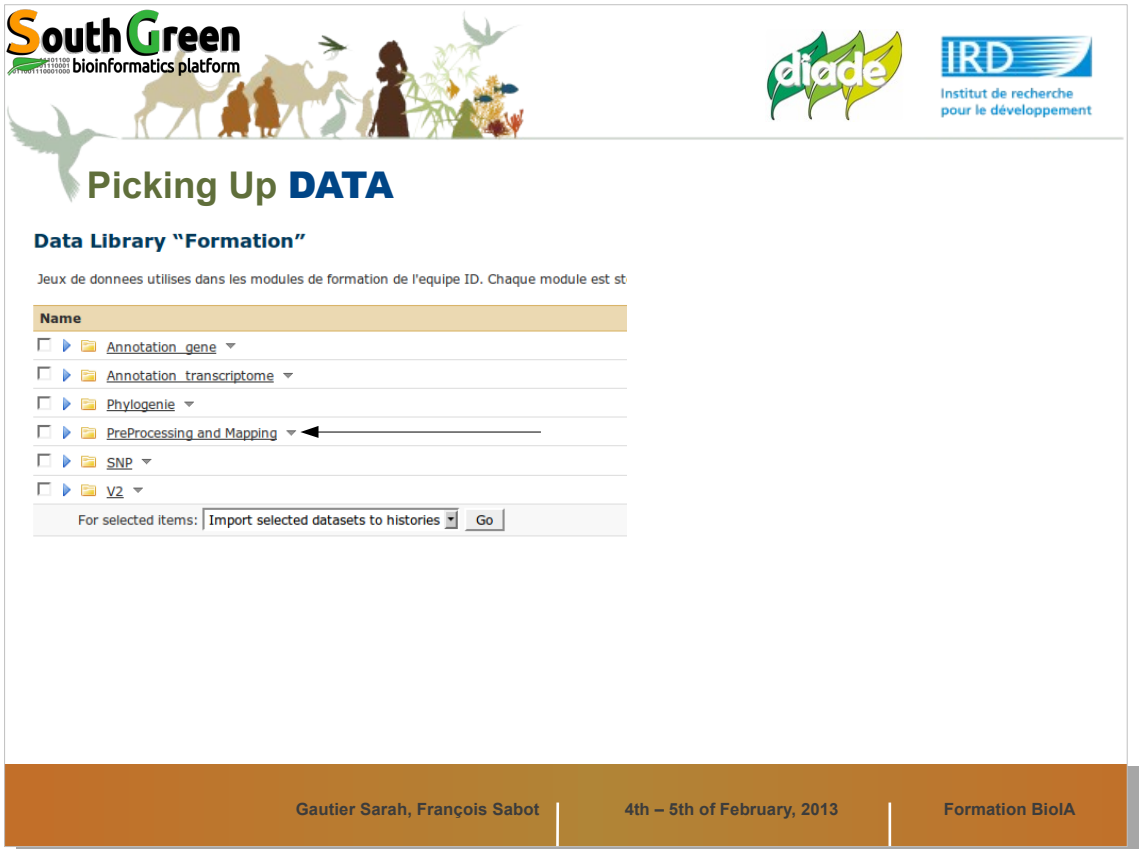
```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTTCGGTGAGGATTGAGGGGGAGT

+
cfffcfeffdeefefffcfffffffcffeffffdffffafcfffffdffffdfefeddf^eececfffdfcbffb
```

f → **Quality** of 38

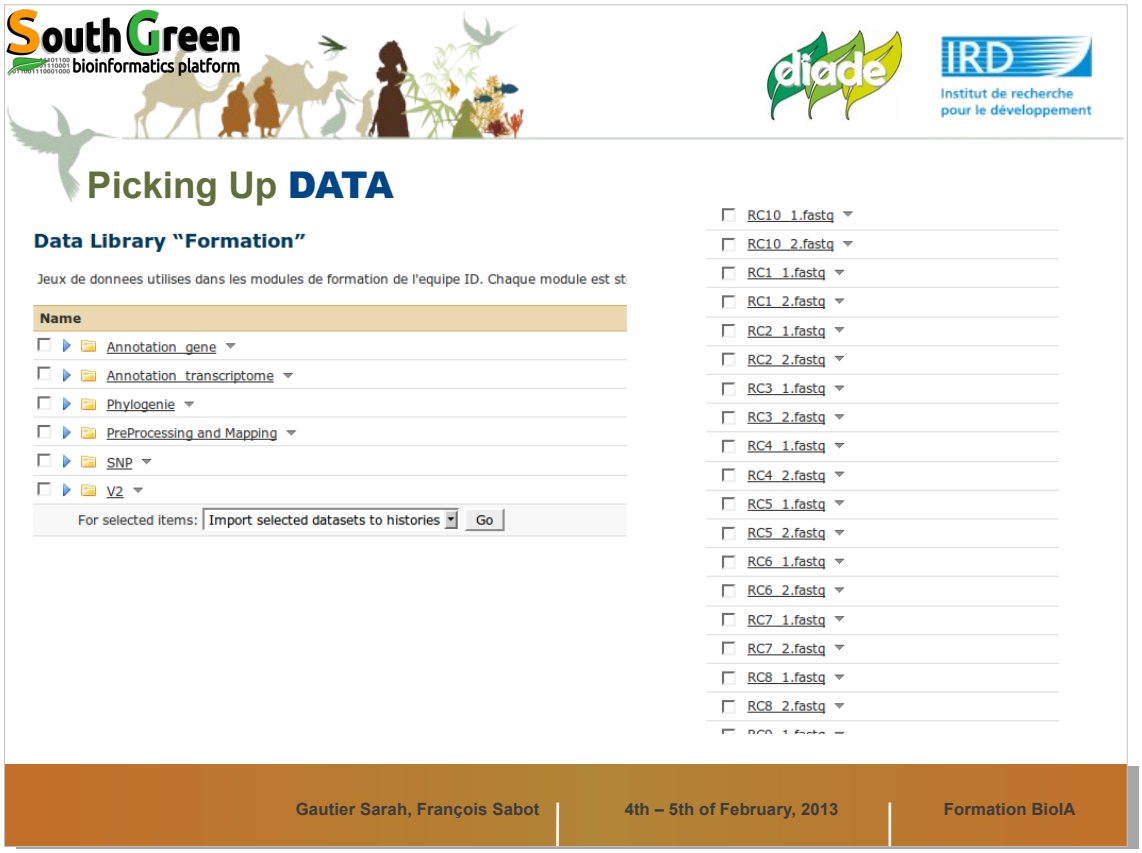
Represente une qualité de 38



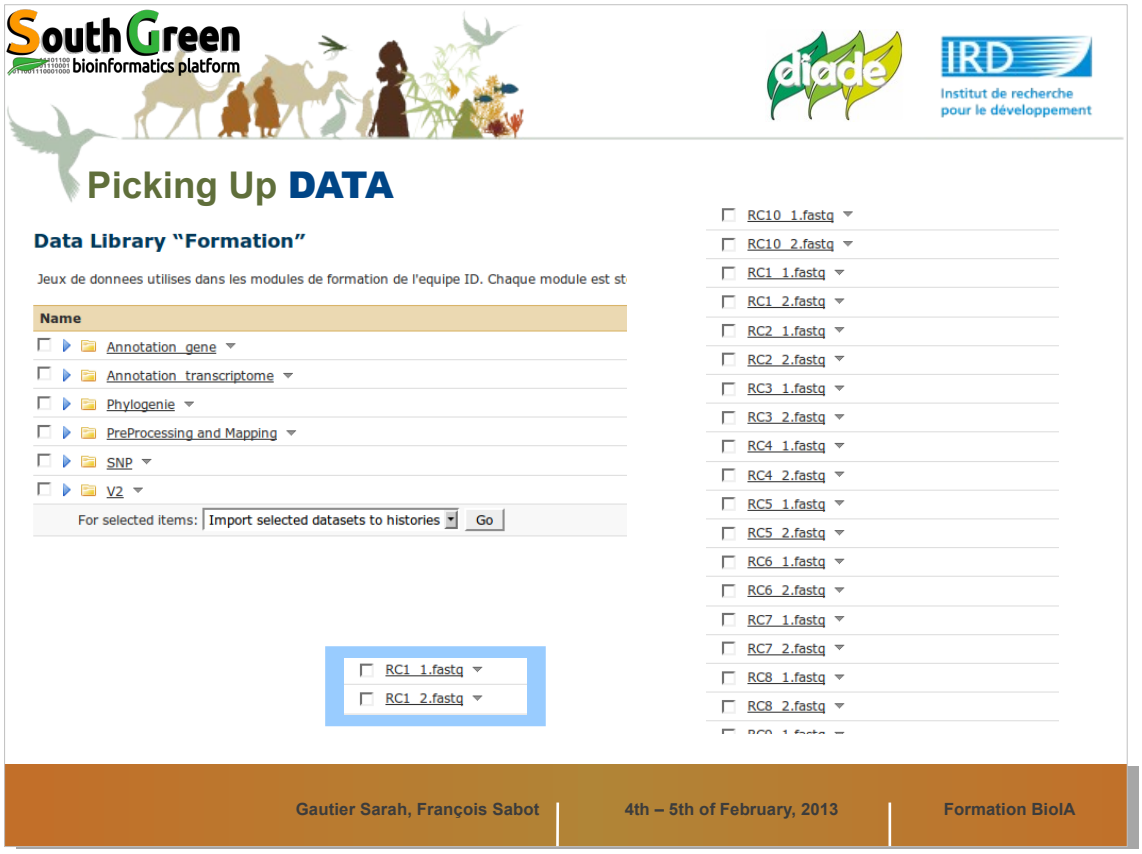
En fait, plus la qualité est proche de 40 (max théorique en séquencage), moins la base a des chances d'avoir été mal lue.



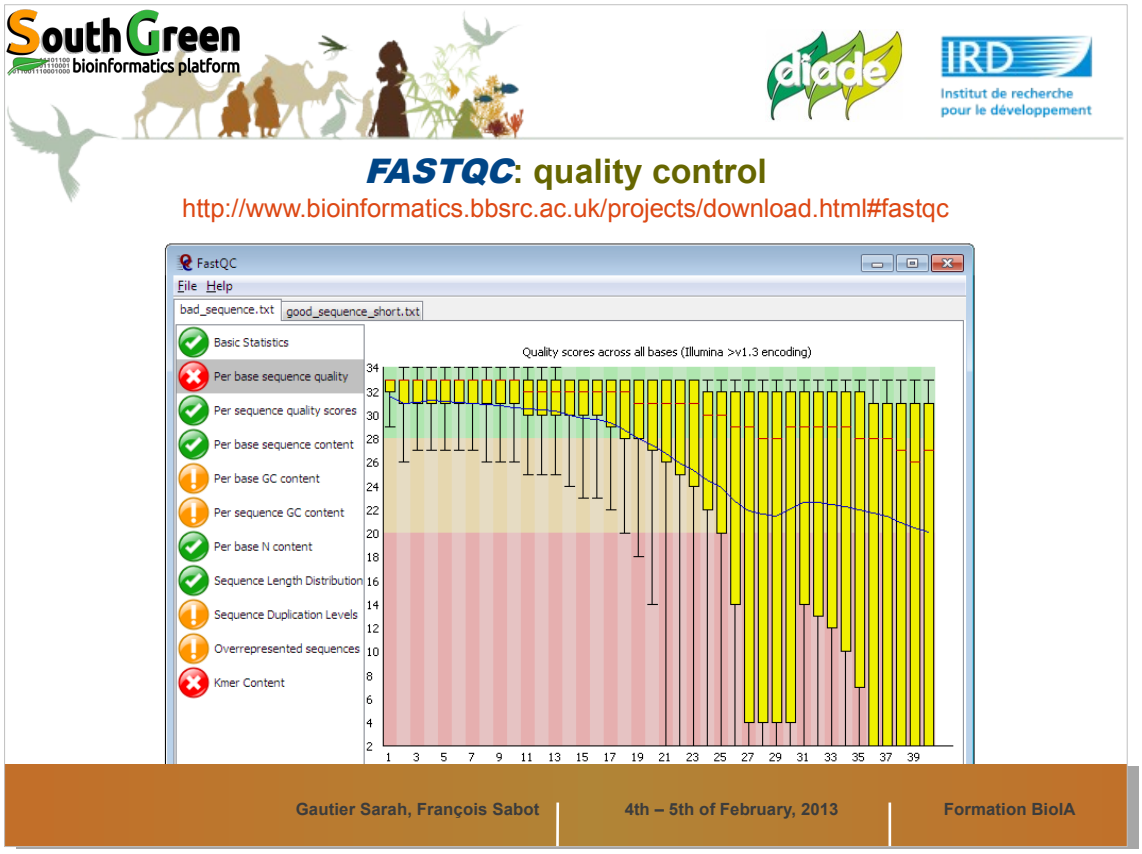
Récupération des données



Chaque traitement doit se faire individu par individu.
Ici, on traite deux fichiers par individus (séquence pairée)



Chaque groupe aura un individu spécifique assigné



FASTQC permet un contrôle rapide de la Qualité des données

FASTQC peut etre lancé dans le Galaxy SouthGreen, mais peut aussi etre telechargé indépendamment

FASTQC: quality control

NGS: Quality Control

- Control composition sequence Percentage and number of CG in a sequence
- Cutadapt Remove adapter sequences from Fastq/Fasta
- Filter FASTQ (ARCAD) on Length and Mean Quality
- FastQC quality control checks on raw sequence data

FastQC (version 1.0.0)

FASTQ reads:

9: Sam_from_BWA

Contaminants:

Selection is Optional

Two fields per line separated by a TAB: name DNA_sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Execute

FASTQC: quality control

FastQC Report

Thu 26 Jan 2012
input1.fastq

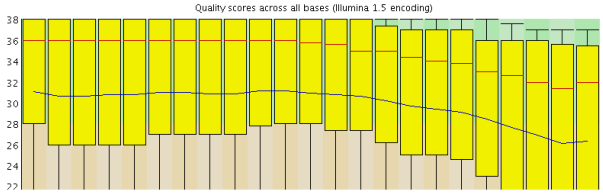
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

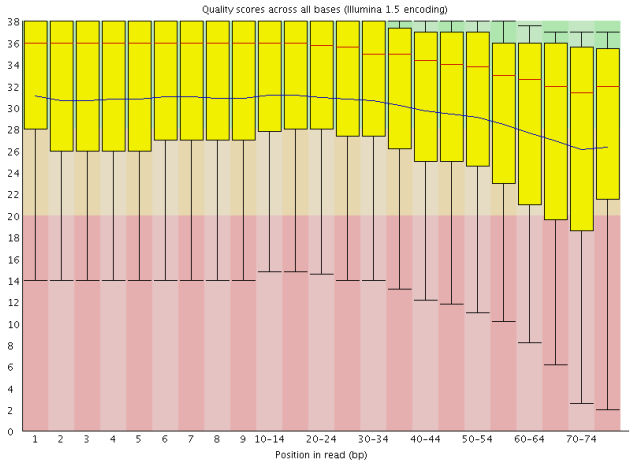
Measure	Value
Filename	input1.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	16322
Filtered Sequences	0
Sequence length	75-76
%GC	49

Per base sequence quality

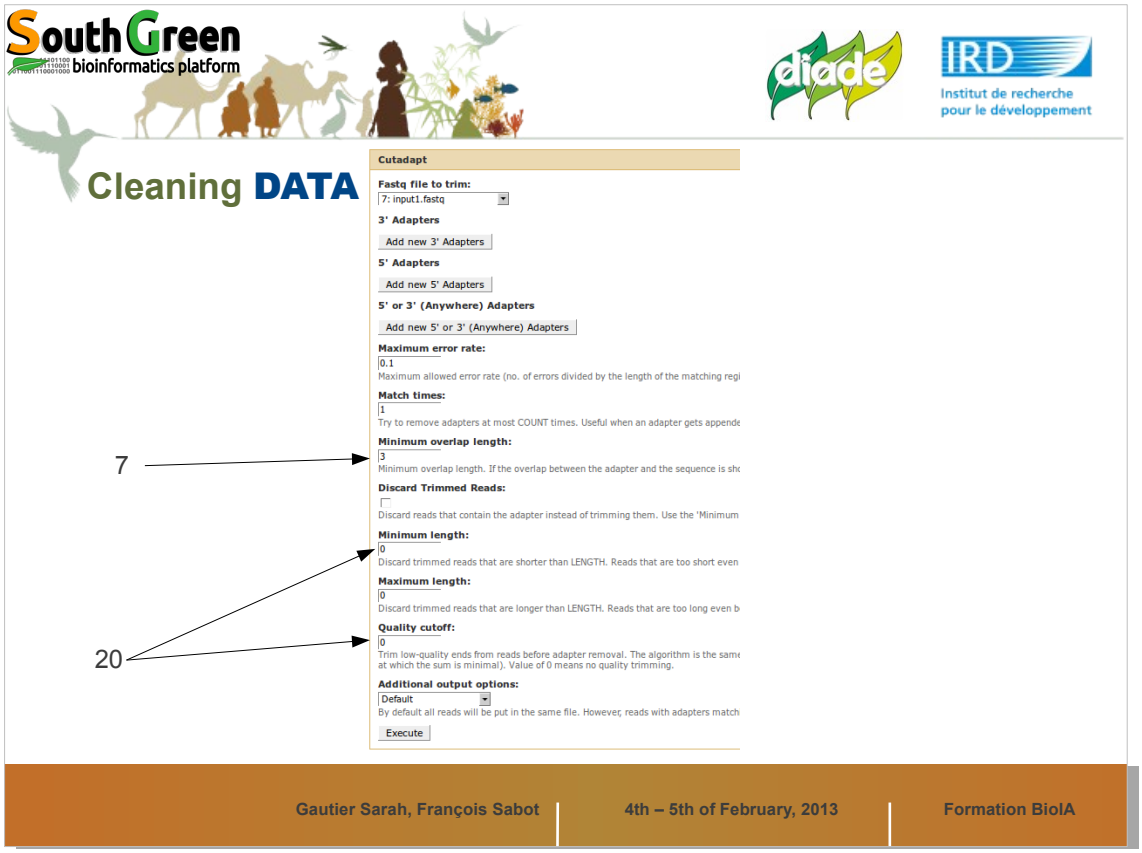


FASTQC: quality control

Per base sequence quality



CutAdapt a la fois nettoie les données des adaptateurs restants, mais permet aussi d'éliminer et couper les séquences de manière à n'avoir que des bases de haute qualité. Cela permet d'éviter les faux positifs plus tard.



La similarité avec un adaptateur doit etre de 7 bases minimum, avec une erreur de 10%. En dessous de 7 bases, trop de séquences de bonne qualité sont éliminées. Au delà, trop d'adaptateurs restent dans les données.

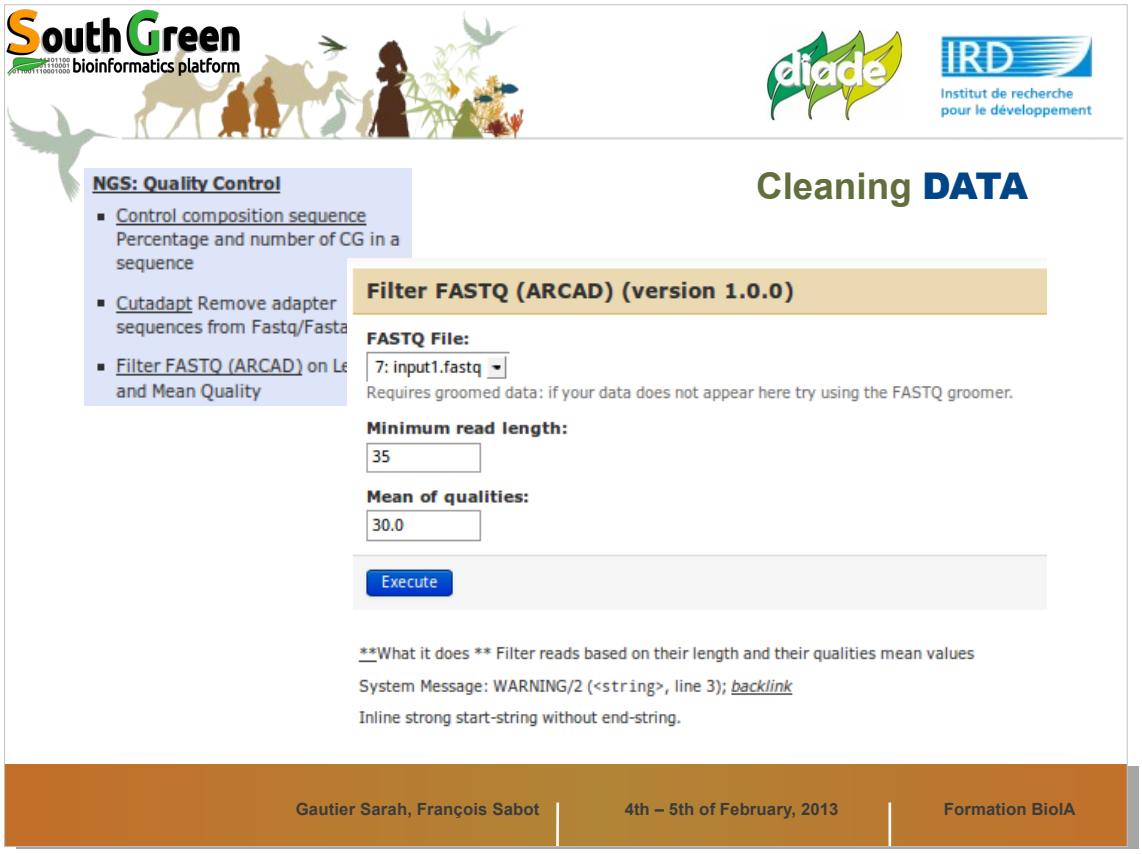
La taille minimale à conserver est de 20.

La qualité de toutes les bases restantes doit etre de 20 minimum.

Cleaning **DATA**

NGS: Quality Control

- Control composition sequence
  Percentage and number of CG in a sequence
- Cutadapt Remove adapter sequences from Fastq/Fasta
- Filter FASTQ (ARCAD) on Length and Mean Quality



FilsterFastq Arcad va encore nettoyer en eliminant les seéquences de moins de 35 bases, et qui ont une qualité MOYENNE de moins de 30.

Il faut vérifier maintenant que les données sont bien encore en paires, ie que chaque séquence forward ai bien une sequence reverse associée. Sinon, le mapping ne pourra pas se faire.

Verifying **DATA**

NGS: Quality Control

- Control composition sequence Percentage and number of CG in a sequence
- Cutadapt Remove adapter sequences from Fastq/Fasta
- Filter FASTQ (ARCAD) on Length and Mean Quality
- FastQC quality control checks on raw sequence data
- Separate Fastq pair and single Filter FASTQ file on unpaired read.

Separate Fastq pair and single (version 0.5)

forward fastq file input:

7: input1.fastq

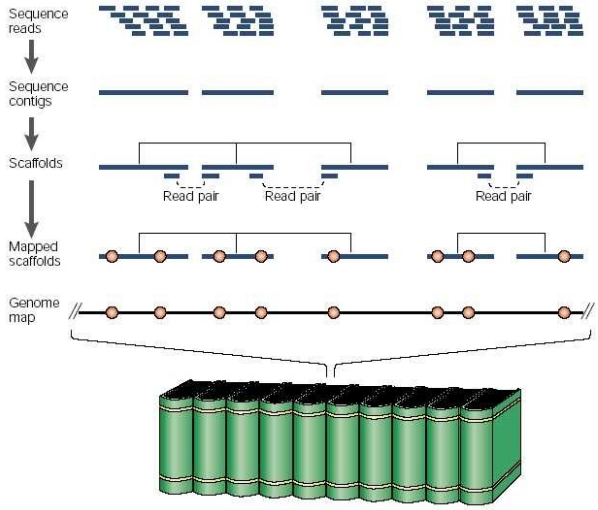
reverse fastq file input:

7: input1.fastq

Execute

Your **Data** are ready for use

Assembling DATA

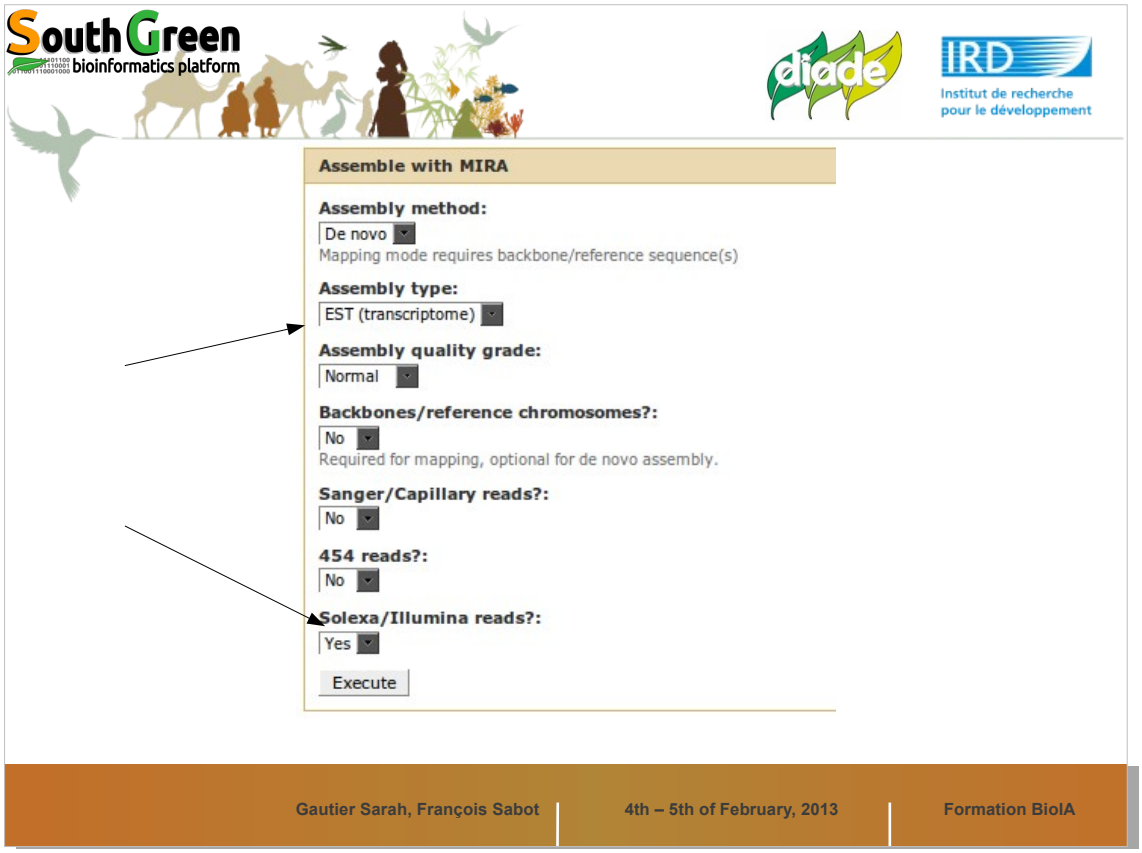


L'assembleur MIRA demande à ce que les données soient réunies dans un seul fichier d'entrée

Untested Tools → NGS → Assembly → Assemble with MIRA

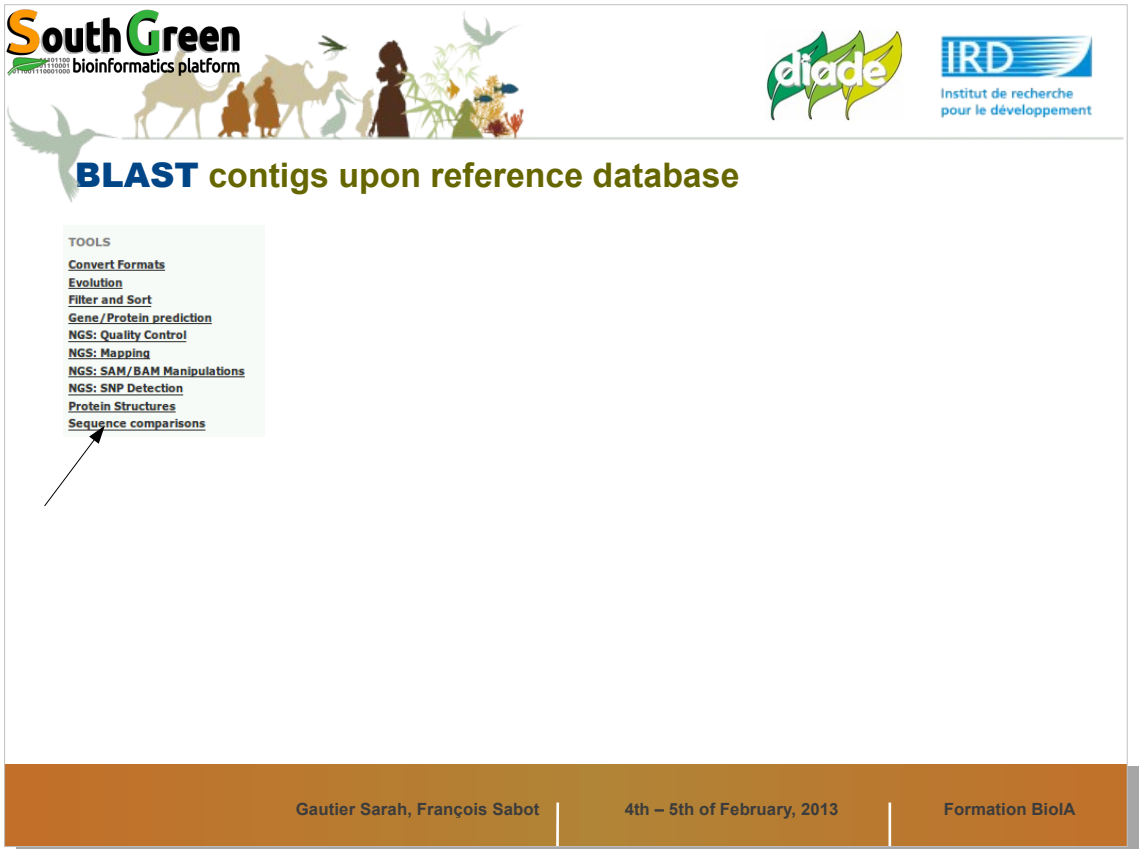
ASSEMBLY

- Assemble with MIRA Takes Sanger, Roche, and Illumina data

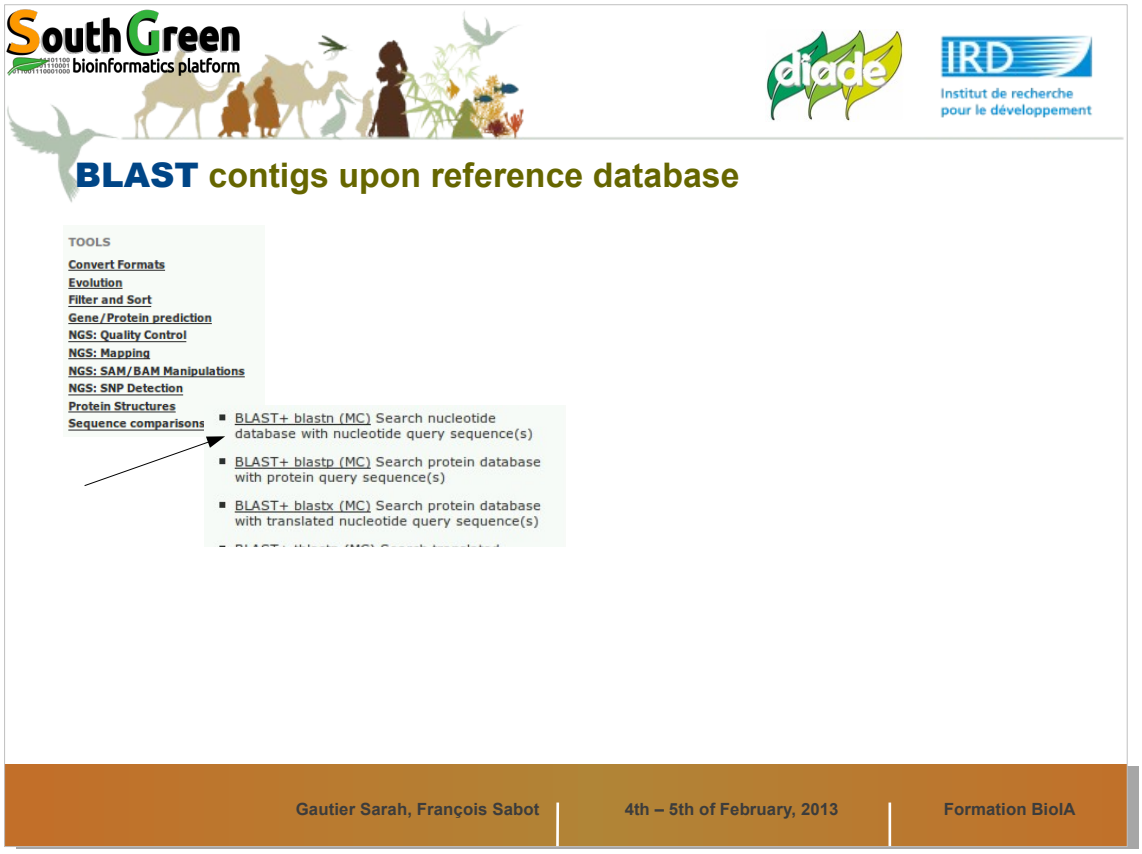


Il faut bien annoncé que nous travaillons sur un set de données d'expression, sinon, le logiciel tentera de relier tous les contigs entre eux. Or, un gene exprimé sous forme ARN est physiquement indépendant d'un autre.

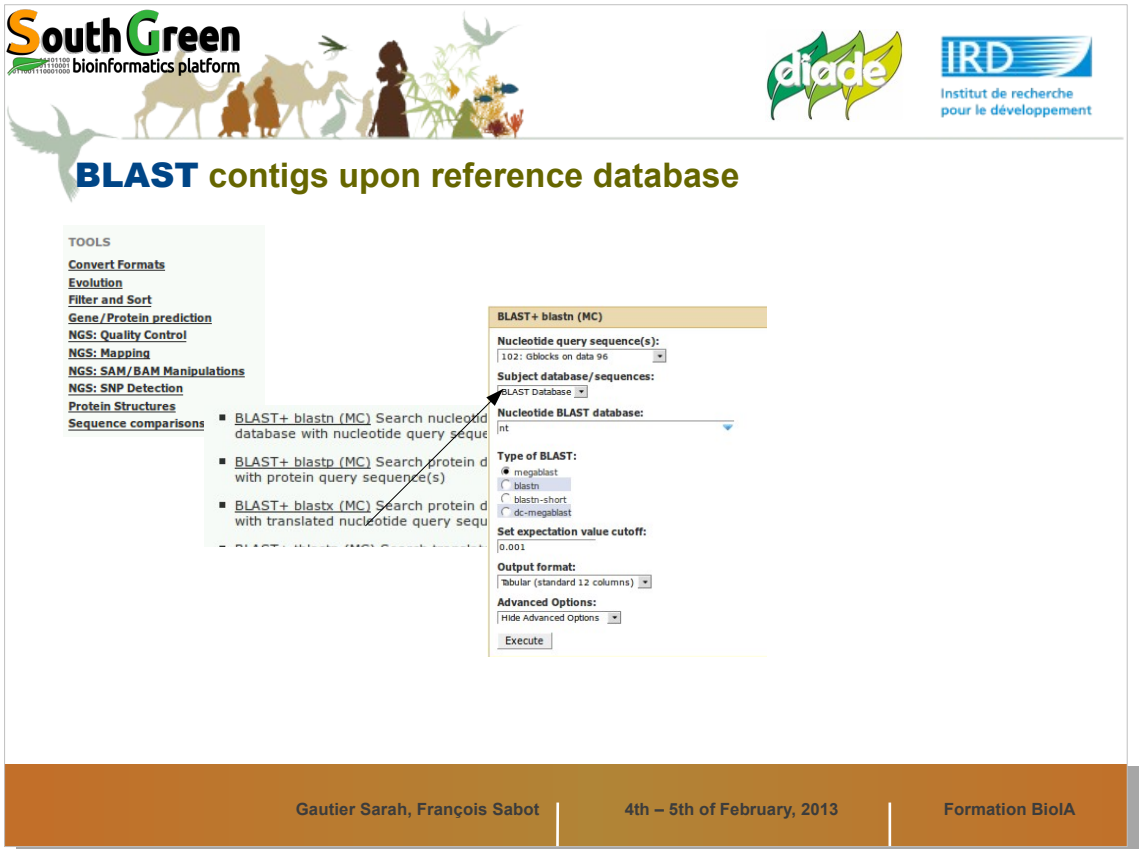
Il faut aussi signaler au logiciel que les données sont de type Illumina, pour permettre d'optimiser l'assemblage



Le BLAST permet de comparer des sequences requetes à une base de données de séquences



Ici nous utilisons un BLASTN, nucleique vs nucleique



La base de données de référence est NT

Nous entrons au coeur des analyses NGS pour les SNP, le mapping, ou comment positionner les sequences sur une reference.

Mapping: Place 'pair-ended' reads upon a reference

1- Calcul of positions for each reads

2- Relationship between each member of the pair

Mapping: Place 'pair-ended' reads upon a reference

1- Calcul of positions for each reads

2- Relationship between each member of the pair

3- Selection of the most probable respecting specific conditions

Mapping: Place 'pair-ended' reads upon a reference

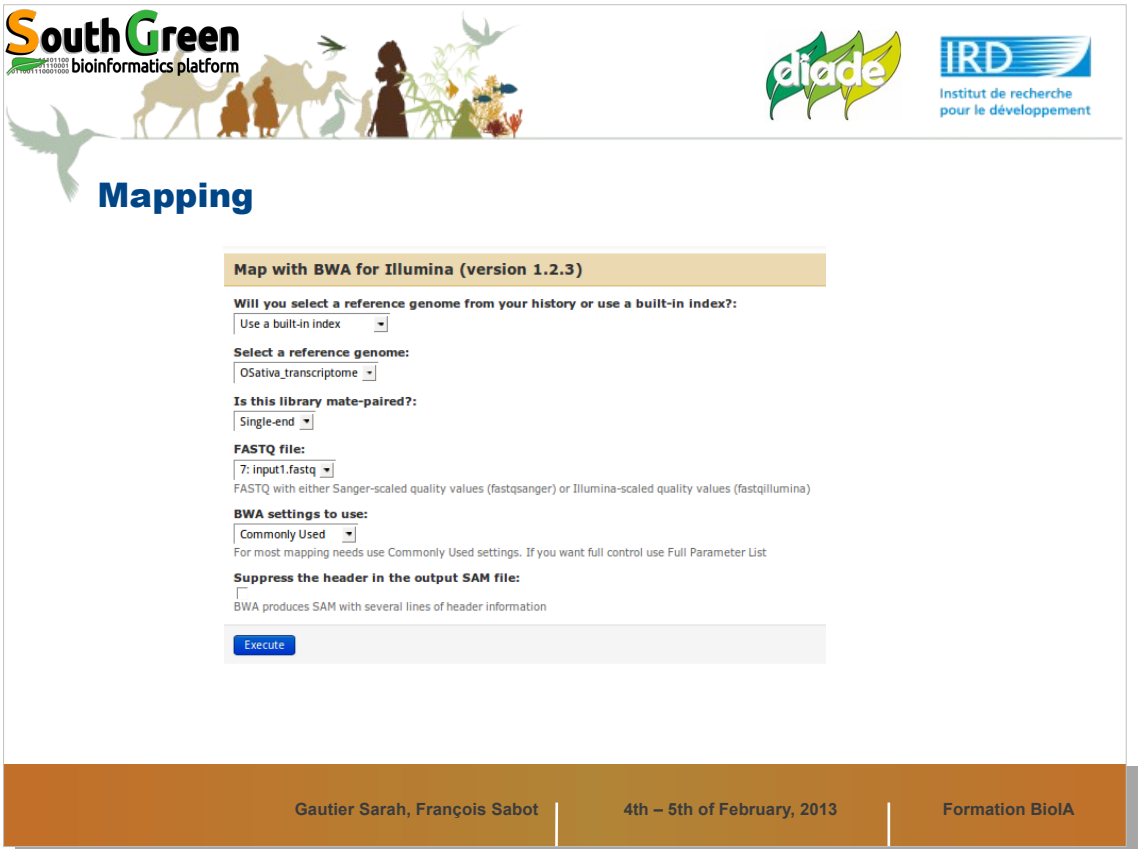
1- Calcul of positions for each reads

2- Relationship between each member of the pair

3- Selection of the most probable respecting specific conditions

4- SAM output format editing

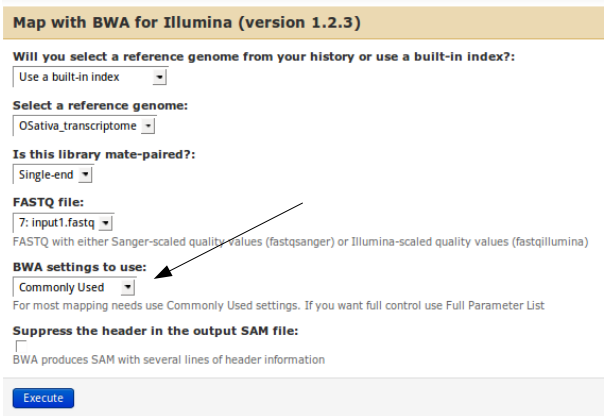
Nous allons utiliser BWA (Burrows-Wheeler Algoritm, Li et al 2009) pour 'mapper' nos séquences pairées et singles sur la référence.



Un set de parametres classiques est pré-défini. Ce set permet de travailler avec a peu pres toutes les données sans trop de soucis.

Il peut etre optimisé pour une analyse spécifique.

Mapping


Map with BWA for Illumina (version 1.2.3)
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Select a reference genome:
OSativa_transcriptome
Is this library mate-paired?:
Single-end
FASTQ file:
7: input1.fastq
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)
BWA settings to use:
Commonly Used
For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List
Suppress the header in the output SAM file:
BWA produces SAM with several lines of header information
Execute

Il existe beaucoup de parametres modifiables. Tous ces parametres sont décrits et expliqués en détails sur la page de téléchargementdu logiciel BWA, http://bio-bwa.sourceforge.net/

Reference From History:
Shared data/Data Library/Formation/SNP/reference.fas.txt

Library:
Paired-end or Single-end, depending

FASTQ files:
From your History (cleaned data)

BWA setting to use:
Commonly Used

Do not select "*Suppress the header in the output SAM file*"

Click Execute

Output file in SAM (Sequence Alignment/Map)

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read `r001/1` and `r001/2` constitute a read pair; `r003` is a chimeric read; `r004` represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT.............TCAGC
-r003                            ttagctTAGGC
-r001/2                                       CAGCGCCAT
```

The corresponding SAM format is:

```
@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         *
```


Le format SAM est un format normalisé, tabulé et de type .txt, et aura toujours la meme structure. Le détail du format est disponible ici http://samtools.sourceforge.net/SAM1.pdf

Les analyses suivantes demandent un fichier de mapping trié par coordonnées de la référence, et non pas dans l'ordre d'entrée des séquences.

Ici, en plus, nous demandons une sortie en BAM, forme binaire (compressé) du SAM. Ce fichier n'est plus lisible par l'homme.

Il est important d'enlever les duplicats techniques (meme couple de séquences lues deux fois) pour eviter les surestimations de profondeur lors de l'appel des SNP.

NGS: SAM/BAM Manipulations

SAM TOOLS

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- rmdup remove PCR duplicates

Removing technical duplicates

rmdup

BAM File:

Is data paired-end:

BAM is paired-end

Treat as single-end:

(-S)

Execute

NGS: SAM/BAM Manipulations

SAM TOOLS

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- rmdup remove PCR duplicates

Removing technical duplicates

rmdup

BAM File:

Is data paired-end:

BAM is paired-end

Treat as single-end:

(-S)

Execute

Depending the dataset

NGS: SAM/BAM Manipulations

SAM TOOLS

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- rmdup remove PCR duplicates

Removing technical duplicates

rmdup

BAM File:

Is data paired-end:

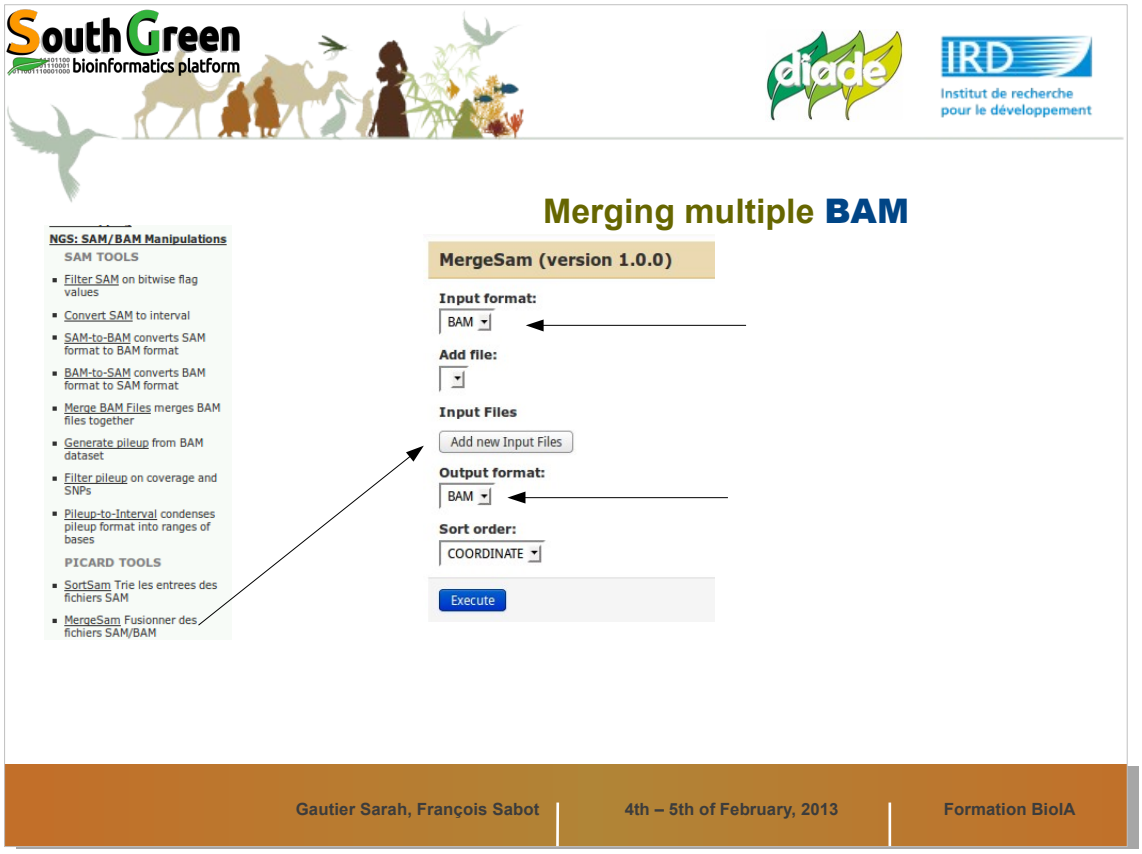
BAM is paired-end

Treat as single-end:

(-S)

Execute

Depending the dataset

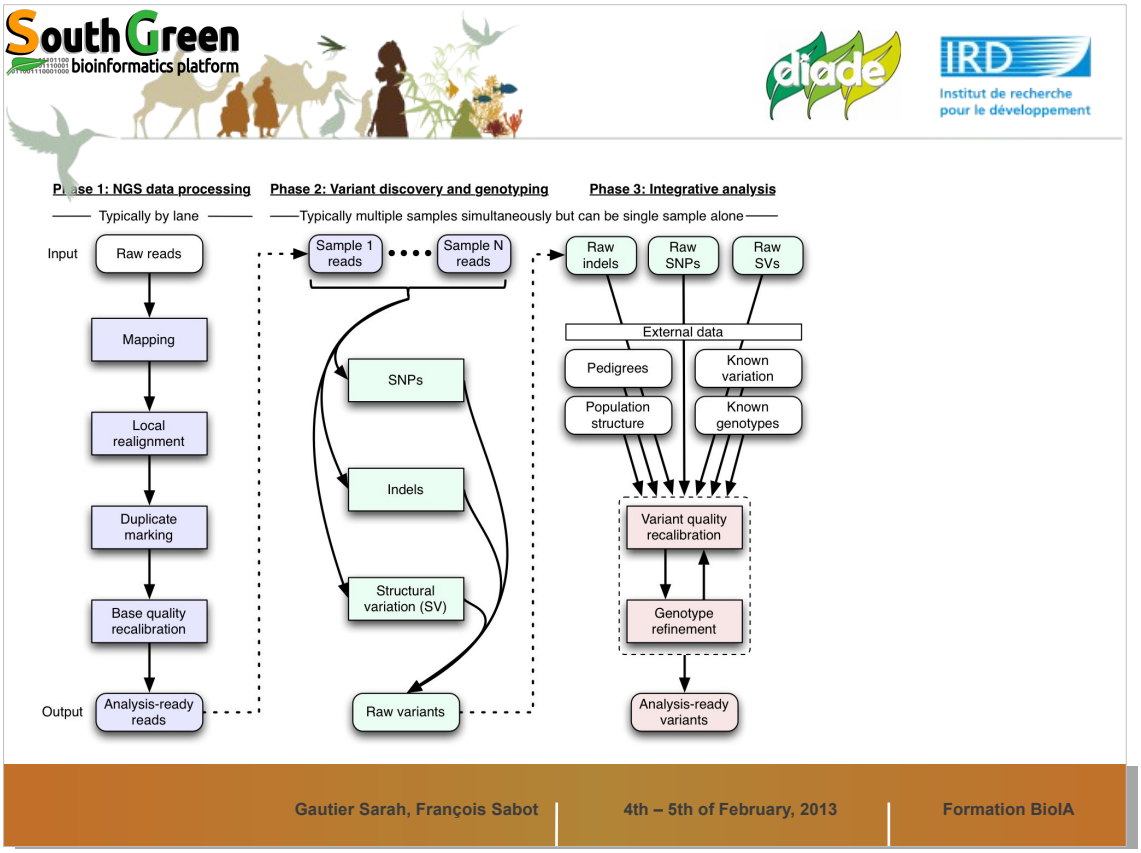


Il est possible de grouper des fichiers BAM issus de mapping multiples (plusieurs individus ou conditions expérimentales) sur la meme reference.

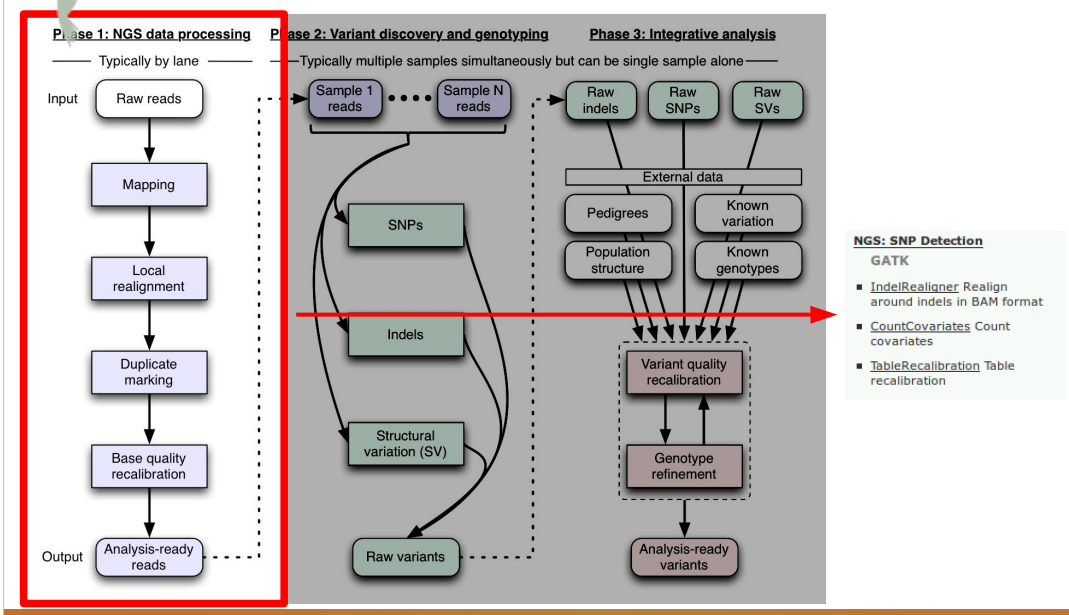
En premier lieu nous allons assembler les données paires et singles.

Partagez votre historique avec les autres apprenants, puis récupérez les fichiers BAM manquants pour les fusionner.

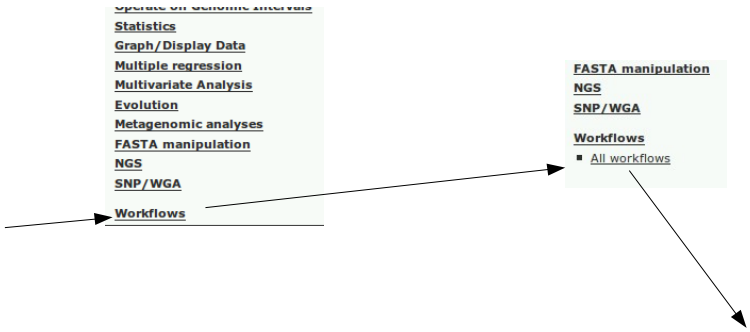
Le ReadGroup est une information rajoutée à chacune des séquences dans le fichier SAM de manière a associer une sequence à l'individu d'origine. Sans cette information il devient impossible de lier une variation à un individu dans une analyse de variabilité multi-individu.



La création d'un fichier BAM nettoyé ne constitue que l'etape preliminaire à l'analyse SNP...


Phase 1: NGS data processing
Phase 2: Variant discovery and genotyping
Phase 3: Integrative analysis
Typically by lane
Typically multiple samples simultaneously but can be single sample alone
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Sample 1 reads
Sample N reads
SNPs
Indels
Structural variation (SV)
Raw variants
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants
NGS: SNP Detection
GATK
IndelRealigner Realign around indels in BAM format
CountCovariates Count covariates
TableRecalibration Table recalibration

Workflow: How to avoid to re-do it by hand



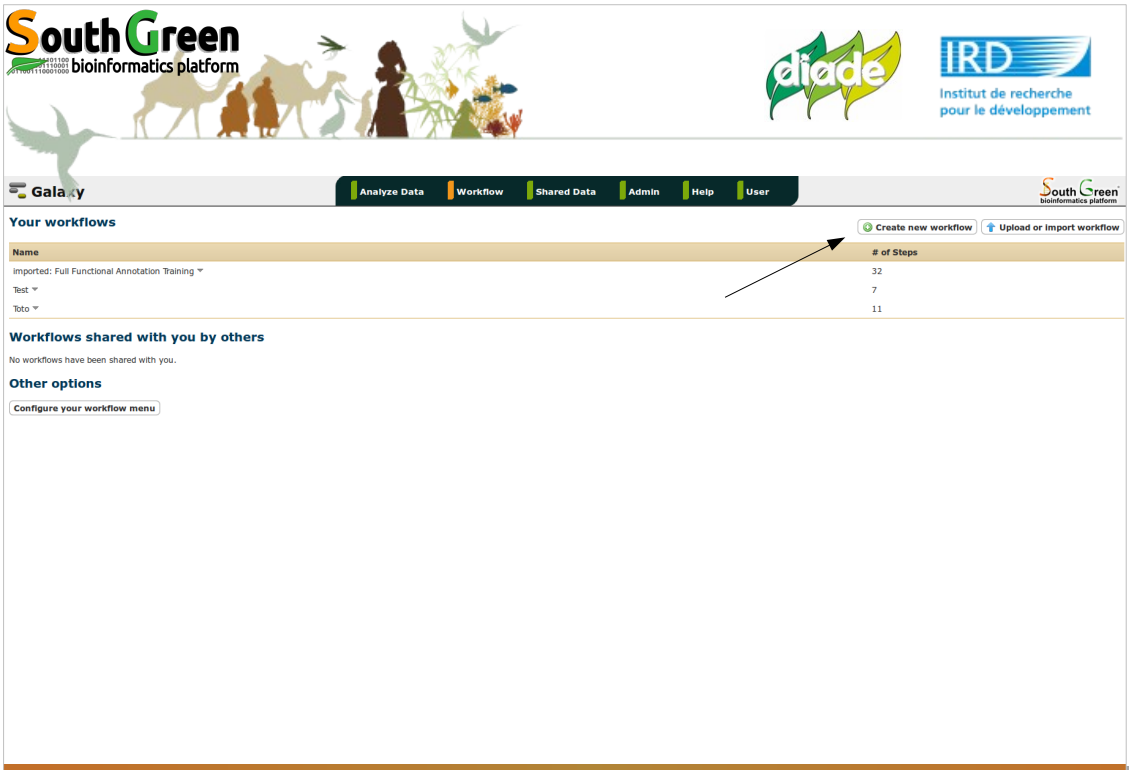
Your workflows

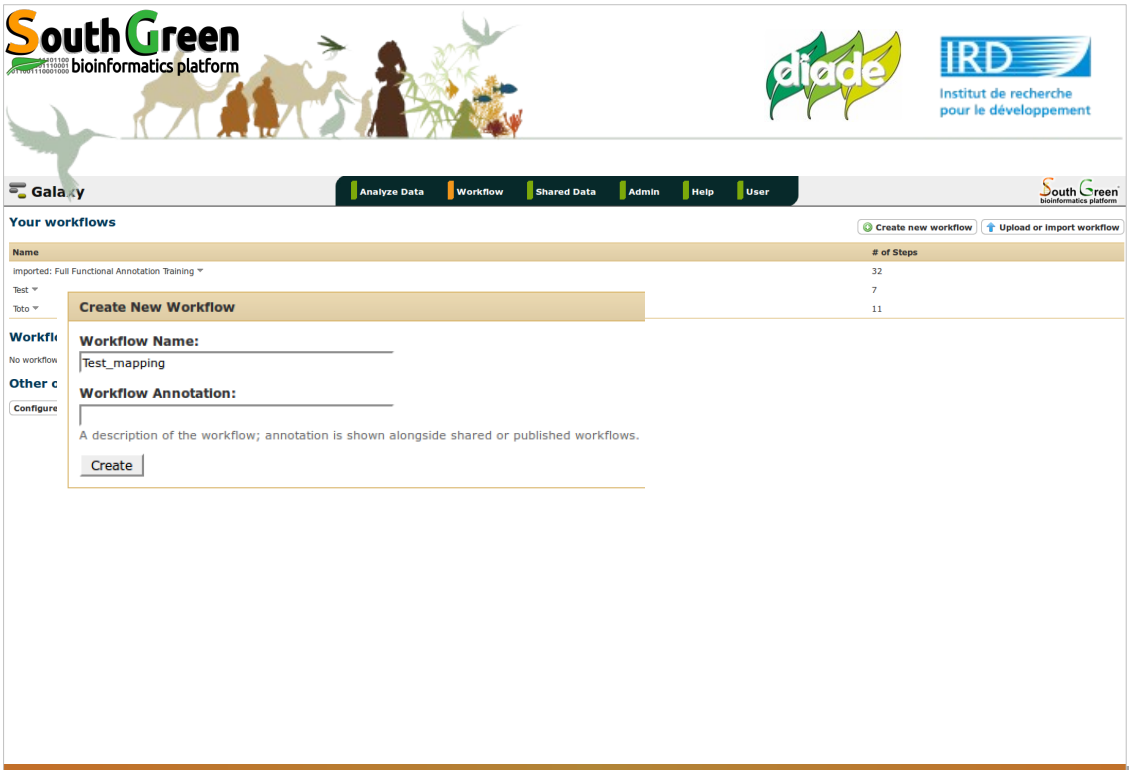
Switch to workflow management view

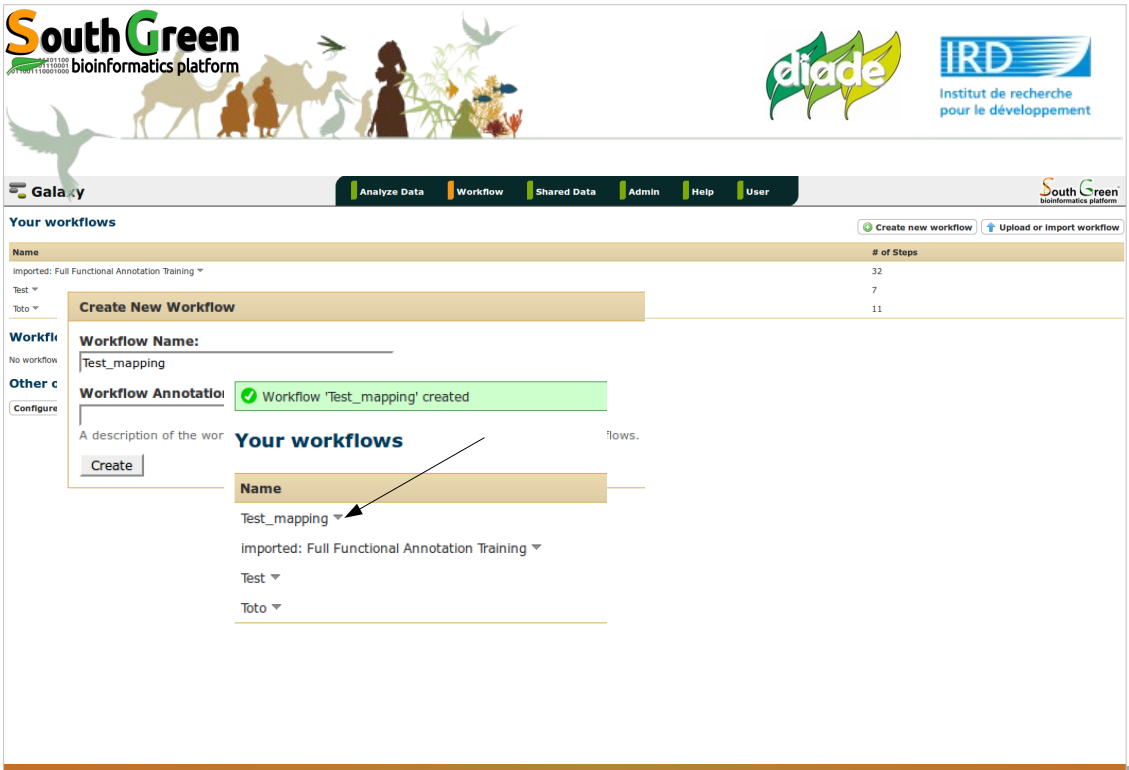
Name	# of Steps
imported: Full Functional Annotation Training	32
Test	7
Toto	11

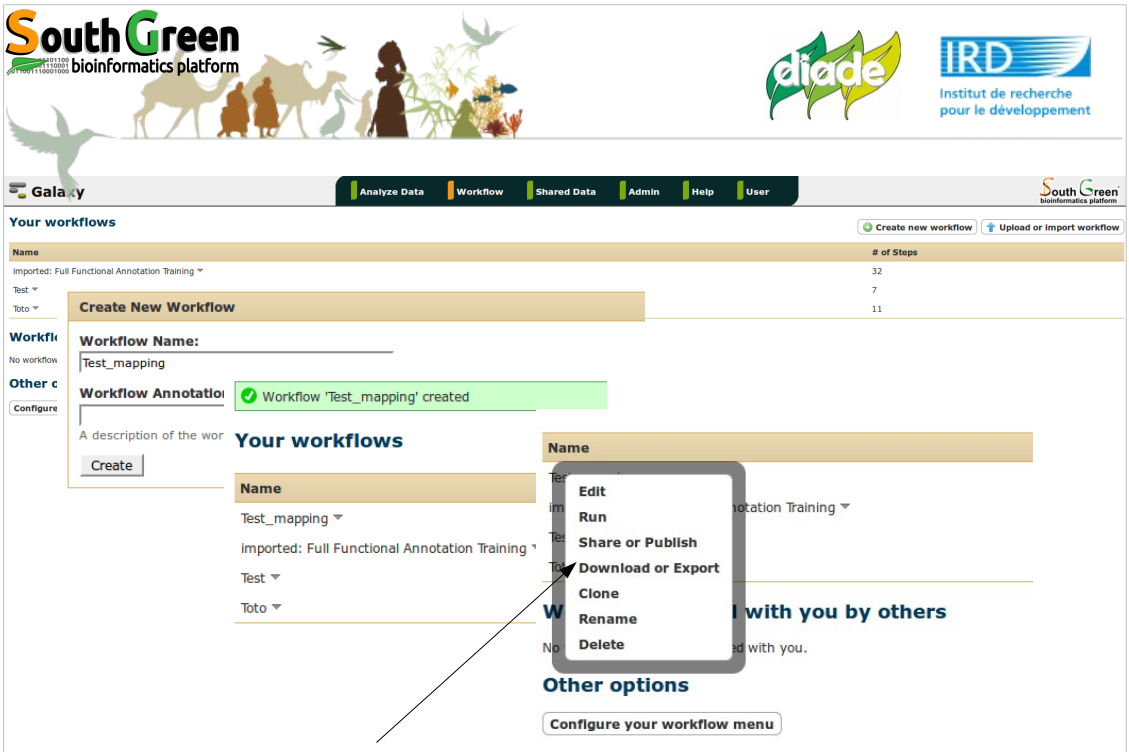
Workflows shared with you by others

No workflows have been shared with you.

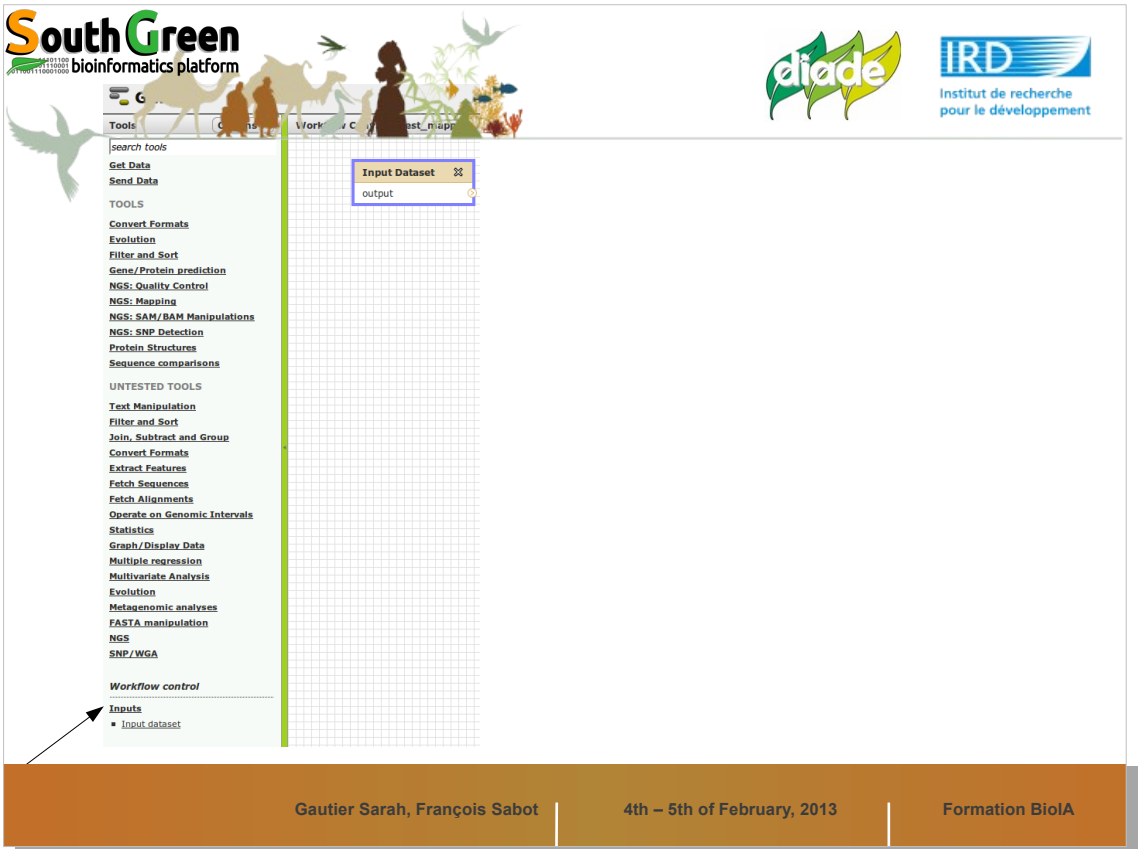

SouthGreen
bioinformatics platform
IRD
Institut de recherche
pour le développement
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
imported: Full Functional Annotation Training
32
Test
7
Toto
11
Workflows shared with you by others
No workflows have been shared with you.
Other options
Configure your workflow menu


SouthGreen
bioinformatics platform
IRD
Institut de recherche
pour le développement
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
imported: Full Functional Annotation Training
32
Test
7
Toto
11
Create New Workflow
Workflow Name:
Test_mapping
Workflow Annotation:
A description of the workflow; annotation is shown alongside shared or published workflows.
Create

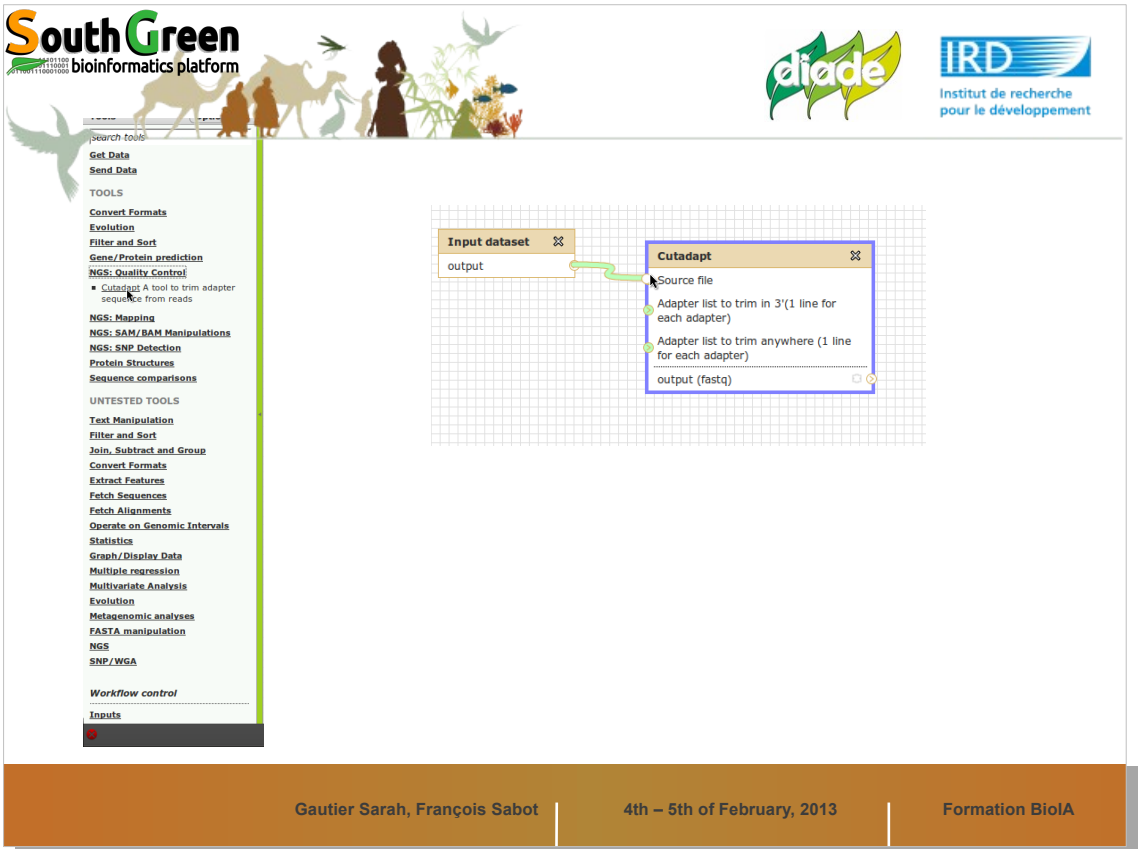

SouthGreen
bioinformatics platform
IRD
Institut de recherche pour le développement
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
imported: Full Functional Annotation Training
32
Test
7
Toto
11
Create New Workflow
Workflow Name:
Test_mapping
Workflow Annotation
Workflow 'Test_mapping' created
A description of the wor
Create
Your workflows
Name
Test_mapping
imported: Full Functional Annotation Training
Test
Toto


SouthGreen
bioinformatics platform
IRD
Institut de recherche pour le développement
Galaxy
Analyze Data
Workflow
Shared Data
Admin
Help
User
Your workflows
Create new workflow
Upload or import workflow
Name
# of Steps
imported: Full Functional Annotation Training
32
Test
7
Toto
11
Create New Workflow
Workflow Name:
Test_mapping
Workflow Annotation
Workflow 'Test_mapping' created
A description of the wor
Create
Your workflows
Name
Test_mapping
imported: Full Functional Annotation Training
Test
Toto
Edit
Run
Share or Publish
Download or Export
Clone
Rename
Delete
Other options
Configure your workflow menu


SouthGreen
bioinformatics platform
Options
IRD
Institut de recherche
pour le développement



Plusieurs *input datasets* peuvent etre utilisés dans un meme workflow.



Si vos données peuvent 'entrer' dans la brique suivante (format cohérent), le trait devient vert

Il devient possible donc d'obtenir des workflows plus ou moins complexes.

SouthGreen bioinformatics platform

IRD Institut de recherche pour le développement