



Treatment of Raw NGS Data

**Cleaning, Formatting,
Assembly, Mapping**

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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- Perform a **mapping** of Illumina data upon a reference sequence
- 5- Clean a multiple **SAM** file

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



IRD
Institut de recherche
pour le développement

FASTQ file → TEXT file

STRUCTURE:

```
@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA
CGCCAAGAAGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT
+
cffffcfeffdeefffffffcfffffcfffffcafccffffdffffdfefeddf^eececfffdcbff
@HWUSI-EAS454_0006:1:37:16314:3410#CTTGTA
AGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGTGGTGGCCG
+
`bTbbcccccccccccYeedded`ceec]dddde^a`deeeeec\`dddcbaadadYd`]]Jc_`bc^`\\`
```

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



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

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT

+

cffffcfeffdeefffeffffcffffffffffffcffffdfffffafcfffdffffdfefeddf^eececffdfcbfffb

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



Sequence Name

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

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT

+

cffffcfeffdeeffffffcfffffcffffcffffdfffffafcfffdffffdfefeddf^eececffdfcbfffb

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Son nom doit etre unique dans le fichier



IRD
Institut de recherche
pour le développement

Sequence, IUPAC

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

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT

+

cffffcfeffdeeffffffcfffffcffffcffffdfffffafcfffdffffdfefeddf^eececffdfcbfffb

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT

+

cffffcfeffdeeffffffcfffffcfffffcafccffffdffffdfefeddf^eececffdfcbfffb

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



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

CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT

+

cffffcfeffdeefffcfffffffccfffefffffafcfffdffffdfefeddf^eececffdfcbfffb

Quality, ASCII

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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




SS
 XX ..
 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII ..
 JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ ..
 LLLLLLLLLLLLLLLLLLLL L L L L L L L L L L L ..
 !"#\$%&'()*+, -./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_ `abcdefghijklmnopqrstuvwxyz{|}-

33 59 64 73 104 126

@HWUSI-EAS454_0006:1:112:14105:5498#CTTGTA
 CGCCAAGAAGTGTAGCAAAACGGCAGAGCTCGTGGATTAAACAAACAGAGGATTCGGTGAGGATTGAGGGGGAGT
 +
 cffffcfeffdeefff ffffff cfffff fffffd ffffffa fc fffff d ffffff d fef eddf ^ eececff d fcbff

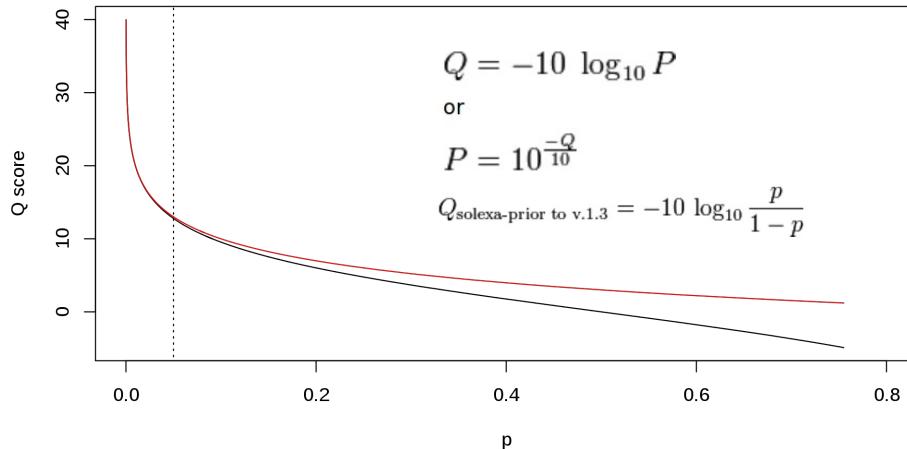
f → Quality of 38

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

Représente une qualité de 38

WHAT IS **QUALITY** ?

Quality value **Q** is an integer mapping of **p** (i.e., the probability that the corresponding base call is incorrect).



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



Picking Up DATA

Data Library "Formation"

Jeux de données utilisés dans les modules de formation de l'équipe ID. Chaque module est st

Name
<input type="checkbox"/> Annotation_gene ▾
<input type="checkbox"/> Annotation_transcriptome ▾
<input type="checkbox"/> Phylogenie ▾
<input type="checkbox"/> PreProcessing and Mapping ▾ ←
<input type="checkbox"/> SNP ▾
<input type="checkbox"/> V2 ▾

For selected items:

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Récupération des données



Picking Up DATA

Data Library "Formation"

Jeux de données utilisés dans les modules de formation de l'équipe ID. Chaque module est st

Name
<input type="checkbox"/> Annotation_gene ▾
<input type="checkbox"/> Annotation_transcriptome ▾
<input type="checkbox"/> Phylogenie ▾
<input type="checkbox"/> PreProcessing and Mapping ▾
<input type="checkbox"/> SNP ▾
<input type="checkbox"/> V2 ▾

For selected items: Import selected datasets to histories

- [RC10_1.fastq](#) ▾
- [RC10_2.fastq](#) ▾
- [RC1_1.fastq](#) ▾
- [RC1_2.fastq](#) ▾
- [RC2_1.fastq](#) ▾
- [RC2_2.fastq](#) ▾
- [RC3_1.fastq](#) ▾
- [RC3_2.fastq](#) ▾
- [RC4_1.fastq](#) ▾
- [RC4_2.fastq](#) ▾
- [RC5_1.fastq](#) ▾
- [RC5_2.fastq](#) ▾
- [RC6_1.fastq](#) ▾
- [RC6_2.fastq](#) ▾
- [RC7_1.fastq](#) ▾
- [RC7_2.fastq](#) ▾
- [RC8_1.fastq](#) ▾
- [RC8_2.fastq](#) ▾
- [RC9_1.fastq](#) ▾

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



Picking Up DATA

Data Library "Formation"

Jeux de données utilisés dans les modules de formation de l'équipe ID. Chaque module est structuré comme suit :

Name
<input type="checkbox"/> Annotation_gene ▾
<input type="checkbox"/> Annotation_transcriptome ▾
<input type="checkbox"/> Phylogenie ▾
<input type="checkbox"/> PreProcessing and Mapping ▾
<input type="checkbox"/> SNP ▾
<input type="checkbox"/> V2 ▾

For selected items: Import selected datasets to histories

<input type="checkbox"/> RC1_1.fastq ▾
<input type="checkbox"/> RC1_2.fastq ▾

[RC10_1.fastq](#) ▾

[RC10_2.fastq](#) ▾

[RC1_1.fastq](#) ▾

[RC1_2.fastq](#) ▾

[RC2_1.fastq](#) ▾

[RC2_2.fastq](#) ▾

[RC3_1.fastq](#) ▾

[RC3_2.fastq](#) ▾

[RC4_1.fastq](#) ▾

[RC4_2.fastq](#) ▾

[RC5_1.fastq](#) ▾

[RC5_2.fastq](#) ▾

[RC6_1.fastq](#) ▾

[RC6_2.fastq](#) ▾

[RC7_1.fastq](#) ▾

[RC7_2.fastq](#) ▾

[RC8_1.fastq](#) ▾

[RC8_2.fastq](#) ▾

[RC9_1.fastq](#) ▾

Gautier Sarah, François Sabot

4th – 5th of February, 2013

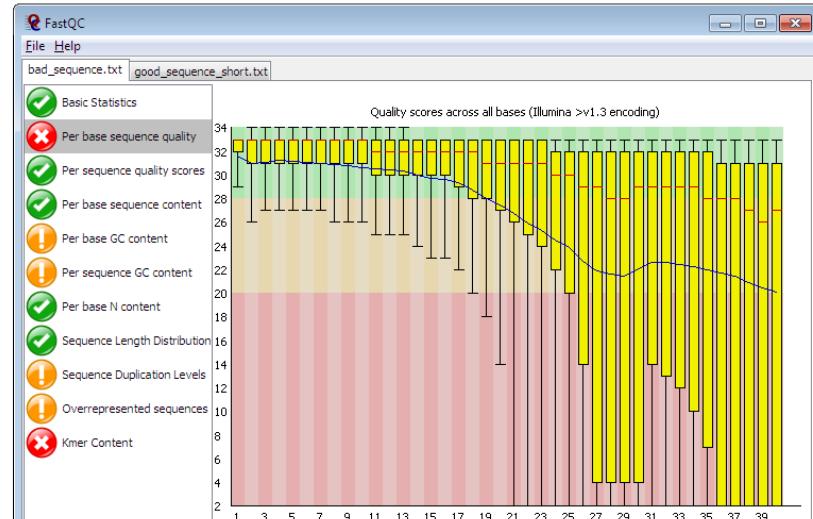
Formation BioIA

Chaque groupe aura un individu spécifique assigné



FASTQC: quality control

<http://www.bioinformatics.bbsrc.ac.uk/projects/download.html#fastqc>



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

FASTQC permet un contrôle rapide de la Qualité des données



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: quality control

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



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: quality control

FastQC (version 1.0.0)

FASTQ reads:**Contaminants:**

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

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

FASTQC: quality control

FastQC Report

Thu 26 Jan 2012
input1.fastq

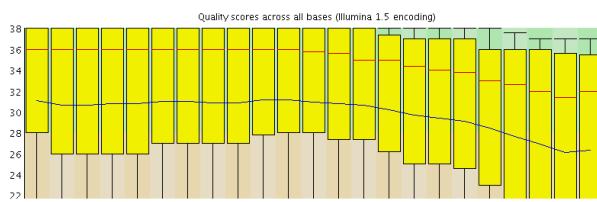
Summary

- ✓ 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



Gautier Sarah, François Sabot

4th – 5th of February, 2013

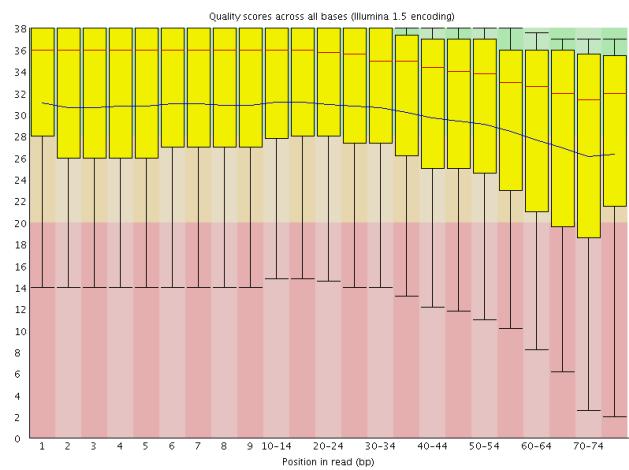
Formation BioIA



IRD
Institut de recherche
pour le développement

FASTQC: quality control

✓ Per base sequence quality



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

NGS: Quality Control

- Control composition sequence
Percentage and number of CG in a sequence
- Cutadapt Remove adapter sequences from Fastq/Fasta



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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.

The screenshot shows the SouthGreen bioinformatics platform interface. At the top, there are logos for SouthGreen, diade, and IRD. Below the logos, a banner says "Cleaning DATA". The main area is a configuration panel for the Cutadapt tool. The "Cutadapt" tab is selected. In the configuration panel, there are several settings:

- Fastq file to trim:** 7: input1.fastq
- 3' Adapters:** Add new 3' Adapters...
- 5' Adapters:** Add new 5' Adapters...
- 5' or 3' (Anywhere) Adapters:** Add new 5' or 3' (Anywhere) Adapters...
- Maximum error rate:** 0.1
- Match times:** 1
- Minimum overlap length:** 3
- Discard Trimmed Reads:** (checkbox)
- Minimum length:** 0
- Maximum length:** 0
- Quality cutoff:** 0
- Additional output options:** Default (radio button selected)

At the bottom of the configuration panel is an "Execute" button. Two arrows point from the text "7" and "20" to the "Minimum length" and "Quality cutoff" fields respectively.

At the bottom of the interface, there is an orange bar with the names "Gautier Sarah, François Sabot", the date "4th – 5th of February, 2013", and the text "Formation BioIA".

La similarité avec un adaptateur doit être 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 être de 20 minimum.



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

Cleaning DATA

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



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](#)

Cleaning DATA

Filter FASTQ (ARCAD) (version 1.0.0)

FASTQ File:

7: input1.fastq

Requires groomed data: if your data does not appear here try using the FASTQ groomer.

Minimum read length:

35

Mean of qualities:

30.0

Execute

**What it does ** Filter reads based on their length and their qualities mean values

System Message: WARNING/2 (<string>, line 3); [backlink](#)

Inline strong start-string without end-string.

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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

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.

Verifying DATA

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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.

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.

Verifying DATA

Separate Fastq pair and single (version 0.5)

forward fastq file input:**reverse fastq file input:****Execute**



IRD
Institut de recherche
pour le développement

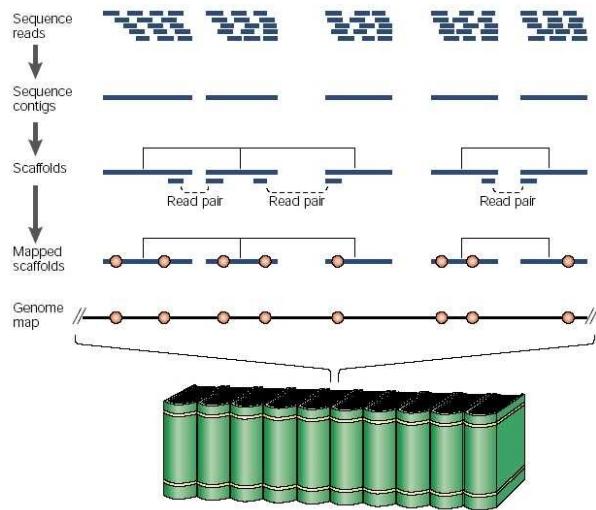
Your **Data** are ready for use

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Assembling DATA



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



Concatenate all files

UNTESTED TOOLS

Text Manipulation

- [Add column to an existing dataset](#)
- [Compute an expression on every row](#)
- [Concatenate datasets tail-to-head](#)

Concatenate datasets

Concatenate Dataset:



X History does not include a required format / build

Datasets

Dataset 1

Select:



[Remove Dataset 1](#)

[Add new Dataset](#)

[Execute](#)

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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



Untested Tools → NGS → Assembly → **Assemble with MIRA**

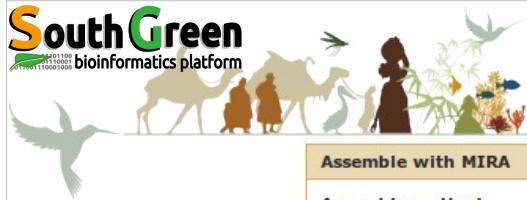
ASSEMBLY

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

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA





Assemble with MIRA

Assembly method: De novo
Mapping mode requires backbone/reference sequence(s)

Assembly type: EST (transcriptome)

Assembly quality grade: Normal

Backbones/reference chromosomes?: No
Required for mapping, optional for de novo assembly.

Sanger/Capillary reads?: No

454 reads?: No

Solexa/Illumina reads?: Yes

Execute

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

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



BLAST contigs upon reference database

TOOLS

- [Convert Formats](#)
- [Evolution](#)
- [Filter and Sort](#)
- [Gene/Protein prediction](#)
- [NGS: Quality Control](#)
- [NGS: Mapping](#)
- [NGS: SAM/BAM Manipulations](#)
- [NGS: SNP Detection](#)
- [Protein Structures](#)
- [Sequence comparisons](#)

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

Le BLAST permet de comparer des séquences requêtes à une base de données de séquences

BLAST contigs upon reference database

TOOLS

[Convert Formats](#)

[Evolution](#)

[Filter and Sort](#)

[Gene/Protein prediction](#)

[NGS: Quality Control](#)

[NGS: Mapping](#)

[NGS: SAM/BAM Manipulations](#)

[NGS: SNP Detection](#)

[Protein Structures](#)

[Sequence comparisons](#)

- [BLAST+ blastn \(MC\)](#) Search nucleotide database with nucleotide query sequence(s)
- [BLAST+ blastp \(MC\)](#) Search protein database with protein query sequence(s)
- [BLAST+ blastx \(MC\)](#) Search protein database with translated nucleotide query sequence(s)

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Ici nous utilisons un BLASTN, nucleique vs nucleique




BLAST contigs upon reference database

TOOLS

- [Convert Formats](#)
- [Evolution](#)
- [Filter and Sort](#)
- [Gene/Protein prediction](#)
- [NGS: Quality Control](#)
- [NGS: Mapping](#)
- [NGS: SAM/BAM Manipulations](#)
- [NGS: SNP Detection](#)
- [Protein Structures](#)
- [Sequence comparisons](#)

BLAST+ blastn (MC)

Nucleotide query sequence(s): 102: Gblocks on data 96

Subject database/sequences: BLAST Database

Nucleotide BLAST database: nt

Type of BLAST:

- megablast
- blastn
- blastn-short
- dc-megablast

Set expectation value cutoff: 0.001

Output format: Tabular (standard 12 columns)

Advanced Options: Hide Advanced Options

Execute

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

La base de données de référence est NT



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

1- Calcul of positions for each reads

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



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**

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



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**

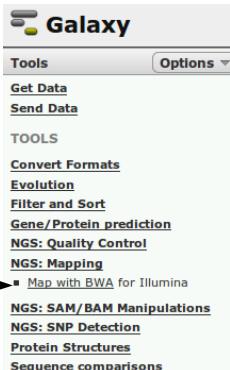
Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



Mapping



Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

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



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?:

Select a reference genome:

Is this library mate-paired?:

FASTQ file:

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:

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

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Un set de paramètres 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.



IRD
Institut de recherche
pour le développement

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

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA





Mapping

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

The screenshot shows the "Map with BWA for Illumina (version 1.2.3)" interface. The top navigation bar includes "File", "Edit", "View", "Help", and "About". Below the header, there's a "Search" field and a "Submit" button. The main area contains several configuration sections:

- Reference genome:** Set to "Human reference genome".
- Is this library mate-paired?**: Set to "Single-end".
- PELIB:** Set to "Read pairs".
- BWA parameters to set:** Includes fields for "Maximum edit distance (min=0)", "Fraction of missing alignments given 2% uniform base error rate (min=0)", "Maximum number of gap opens (min=0)", "Maximum number of gap extensions (min=0)", and "Number of first subsequences to take as seed (min=1)".
- Other parameters:** Includes "Disable long deletion within [value] bp towards the 2'-end (min=0)", "Minimum insertion/deletion within [value] bp towards the end (min=0)", "Number of first subsequences to take as seed (min=1)", "Gap extension penalty (min=0)", "Gap mismatch penalty (min=M)", "Gap open penalty (min=0)", and "Process with suboptimal alignments if there are no more than INT equally best hits. (min=R)".
- XA tag parameters:** Includes "Disable iterative search (min=N)", "Maximum number of alignments to output in the XA tag for reads paired properly (xmax/samp=0)", and "Maximum number of alignments to output in the XA tag for discordant read pairs (excluding singletons) (sample=0)".
- Sam parameters:** Includes "Maximum insert size for a read pair to be considered as being mapped properly (sample=0)", "Maximum occurrences of a read for pairing (sample=0)", and "Specify the read group for this SAM? (sample=0)".
- Output options:** Includes "Suppress the header in the output SAM file?" and "DARK produces SAM with several lines of header information".

Il existe beaucoup de paramètres modifiables. Tous ces paramètres sont décrits et expliqués en détails sur la page de téléchargement du 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**GATAGCTGTGCTAGTAGGCAGTCAGGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002          aaaAGATAA*GGATA
+r003          gcctaAGCTAA
+r004          ATAGCT.....TCAGC
-r003          tttagctTGGCC
-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 AAAAGATAAAGGATA *
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 9N = 7 -39 CAGCGCCAT *
```

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

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





Galaxy

SortSam (version 1.0.0)

Tools

- NGS: Quality Control**
- NGS: Mapping**
- 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
 - [rmdupe](#) remove PCR duplicates
 - [coverage_samtools](#) provides simple stats on BAM files
 - [Depth](#) provides depth for each bases from a bam alignment
 - [flagstat](#) provides simple stats on BAM files
- PICARD TOOLS**
 - [SortSam](#) Trie les entrées des fichiers SAM

Input format: SAM

SAM: 9:Sam_from_BWA

Output format: BAM

Sort order: COORDINATE

Execute

Sorting SAM by coordinate

Please cite:

TODO

Overview

Description plus complet

System Message: WARNING/2 (<string>, li
Title underline too short.
Description plus complete de MonProg
.....
TODO

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

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.

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
- [rmdupe](#) remove PCR duplicates

Removing technical **duplicates**

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA

Il est important d'enlever les duplicates techniques (même couple de séquences lues deux fois) pour éviter 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
- [rmdupe](#) remove PCR duplicates

Removing technical duplicates

rmdupe

BAM File:

Is data paired-end:
 BAM is paired-end

Treat as single-end:

(-S)

Execute

Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA



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
- [rmduo](#) remove PCR duplicates

Removing technical duplicates

rmduo

BAM File:

Is data paired-end:
 BAM is paired-end

Treat as single-end:

(-S)

Execute

Depending the dataset





Merging multiple BAM

MergeSam (version 1.0.0)

Input format: BAM

Add file:

Input Files: Add new Input Files

Output format: BAM

Sort order: COORDINATE

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

PICARD TOOLS

- SortSam Trie les entrées des fichiers SAM
- MergeSam Fusionner des fichiers SAM/BAM

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

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

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.

NGS SNP Detection

GATK

- [IndelRealigner](#) Realign around indels in BAM format
- [CountCovariates](#) Count covariates
- [TableRecalibration](#) Table recalibration
- [UnifiedGenotyper](#) Unified Genotyper

VARSCAN

- [Pileup2snp](#) Identify SNPs from a pileup file

SNIPLAY UTILITIES

- [SamToFastaAlignments](#) Create a collection of FASTA alignments from a SAM output and detects intragroup SNP
- [AddReadGroupIntoSam](#) Add read group into a sam alignment

Adding readgroups

AddReadGroupIntoSam

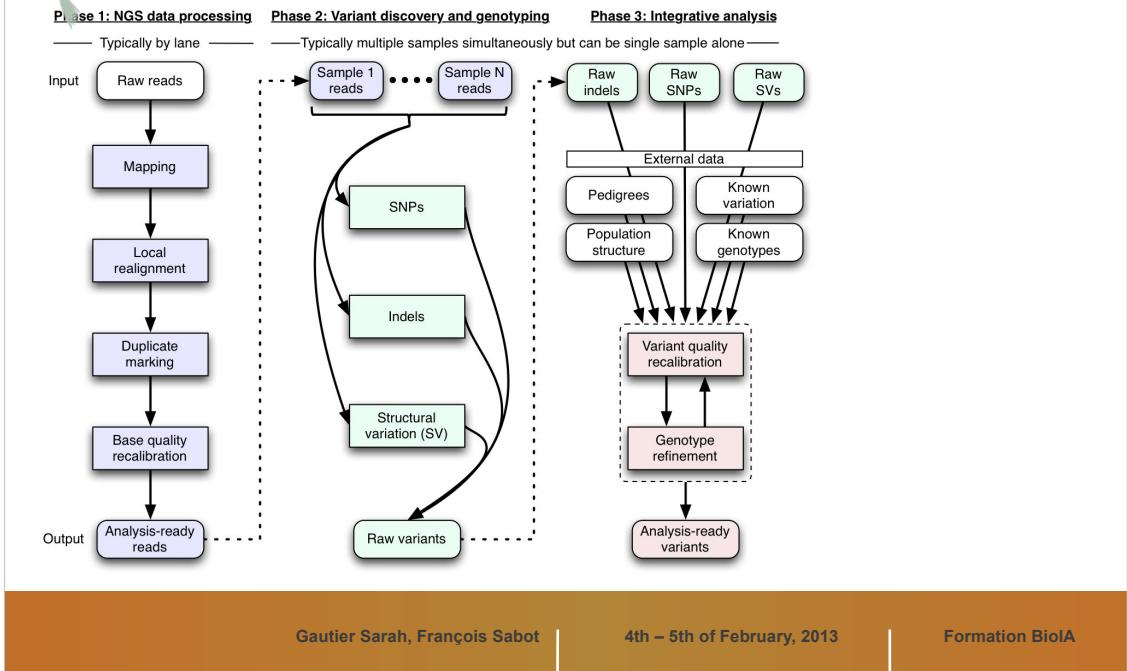
SAM mapping/alignment:

read group sample name:
 RC1, RC2...

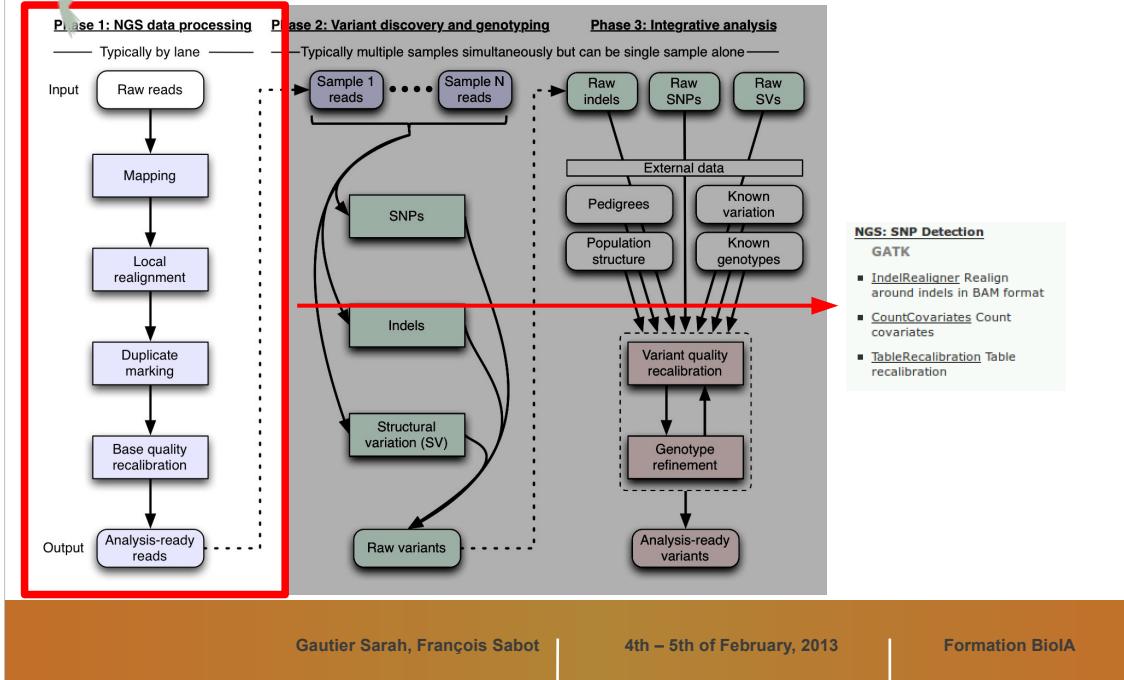
Platform (Illumina/Solid):

Program encapsulated in Galaxy and developed by Southgreen
 SNIPlay utilities

Le ReadGroup est une information rajoutée à chacune des séquences dans le fichier SAM de manière à 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 préliminaire à l'analyse SNP...



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

Operate on genomic analyses

- [Statistics](#)
- [Graph/Display Data](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Metagenomic analyses](#)
- [FASTA manipulation](#)
- [NGS](#)
- [SNP/WGA](#)
- [Workflows](#)

FASTA manipulation

- [NGS](#)
- [SNP/WGA](#)

Workflows

- [All workflows](#)

Your workflows

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.

Switch to workflow management view

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

The image shows the SouthGreen bioinformatics platform interface. At the top, there is a decorative banner featuring a camel, people, and tropical foliage. To the right are logos for diade and IRD.

The main navigation bar includes links for Galaxy, Analyze Data, Workflow, Shared Data, Admin, Help, and User. A search bar is located above the "Workflow" link.

Your workflows

Imported: Full Functional Annotation Training

Name	# of Steps
Test	32
Toto	7
	11

Workflows shared with you by others

No workflows have been shared with you.

Other options

Configure your workflow menu

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

An arrow points from the text "Create new workflow" in the top right corner of the workflow table towards the "Create new workflow" button in the bottom right corner of the interface.





 Galaxy Analyze Data Workflow Shared Data Admin Help User 

Your workflows

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

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

Galaxy

Your workflows

Create New Workflow

Workflow Name: Test_mapping

Workflow Annotation: Workflow 'Test_mapping' created

A description of the workflow.

Name

Test_mapping ▾

Imported: Full Functional Annotation Training ▾

Test ▾

Toto ▾

Your workflows

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA

SouthGreen bioinformatics platform



diade

IRD Institut de recherche pour le développement

Galaxy

Your workflows

Create New Workflow

Workflow Name: Test_mapping

Workflow Annotation: Workflow 'Test_mapping' created

A description of the workflow

Name

Test_mapping
Imported: Full Functional Annotation Training
Test
Toto

Workflow Annotations

Share with others

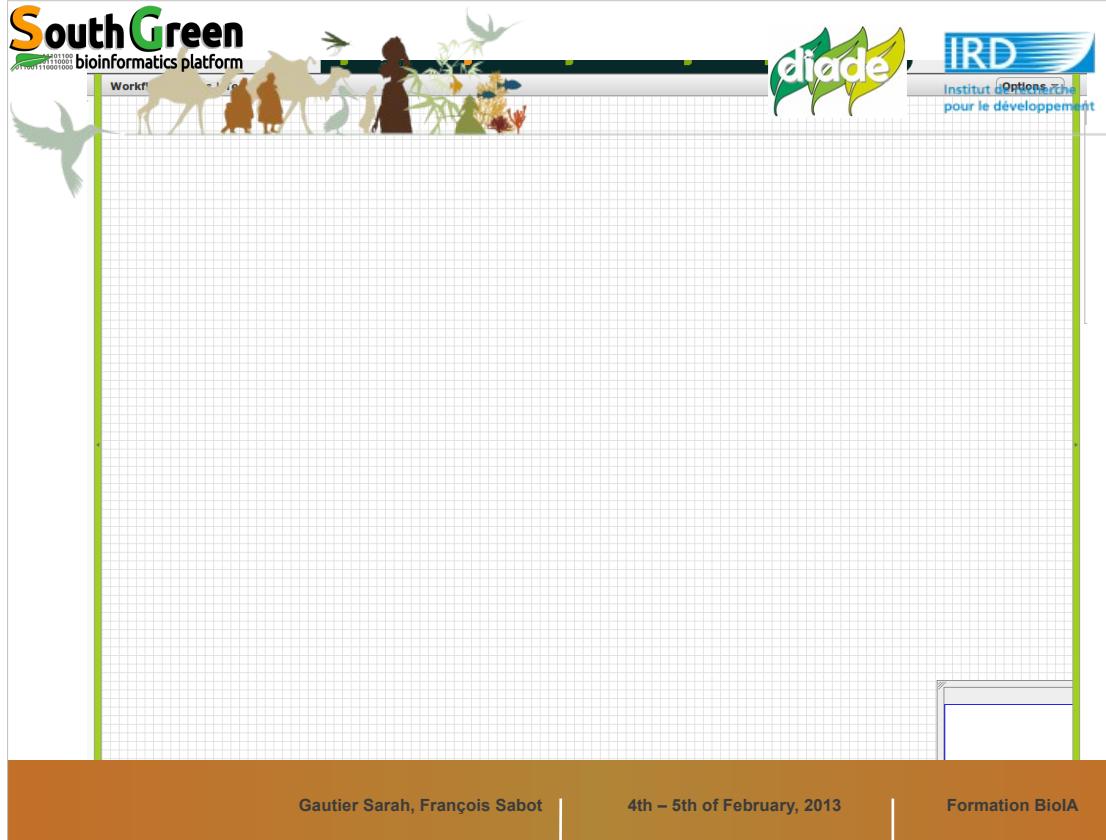
Other options

Configure your workflow menu

Workflow Actions

- Test
- Import
- Run
- Share or Publish
- Download or Export
- Clone
- Rename
- Delete

Gautier Sarah, François Sabot | 4th – 5th of February, 2013 | Formation BioIA



The screenshot shows the SouthGreen bioinformatics platform interface. At the top left is the logo "SouthGreen bioinformatics platform". On the right are logos for "diade" and "IRD Institut de recherche pour le développement". The main area is a workflow editor. A sidebar on the left lists various tools categorized under "TOOLS" and "UNTESTED TOOLS". In the center, a large grid workspace contains several nodes. One node is highlighted with a blue border and has the label "Input Dataset" above it. An arrow points from the bottom left towards this highlighted node. At the bottom of the screen, there is an orange footer bar with the text "Gautier Sarah, François Sabot", "4th – 5th of February, 2013", and "Formation BioIA".

Plusieurs *input datasets* peuvent être utilisés dans un même workflow.

The screenshot shows the SouthGreen bioinformatics platform interface. At the top left is the logo for "SouthGreen bioinformatics platform" featuring a stylized green and orange design. To the right are logos for "diade" and "IRD Institut de recherche pour le développement".

The main area displays a workflow diagram. A blue box labeled "Input dataset" has a green arrow pointing to a yellow box labeled "Cutadapt". The "Cutadapt" box contains the following text:

```

Input dataset ×
output → Cutadapt ×
Source file
Adapter list to trim in 3'(1 line for
each adapter)
Adapter list to trim anywhere (1 line
for each adapter)
output (fastq)

```

On the left side, there is a vertical sidebar with a green header containing the following sections and tools:

- TOOLS**
 - Convert Formats
 - Evolution
 - Filter and Sort
 - Gene/Protein prediction
 - NGS: Quality Control**
 - Cutadapt: A tool to trim adapter sequence from reads
 - NGS: Mapping
 - NGS: SAM/BAM Manipulations
 - NGS: SNP Detection
 - Protein Structures
 - Sequence comparisons
- UNTESTED TOOLS**
 - Text Manipulation
 - Filter and Sort
 - Join, Subtract and Group
 - Convert Formats
 - Extract Features
 - Fetch Sequences
 - Fetch Alignments
 - Operate on Genomic Intervals
 - Statistics
 - Graph/Display Data
 - Multiple regression
 - Multivariate Analysis
 - Evolution
 - Metagenomic analyses
 - FASTA manipulation
 - NGS
 - SNP/WGA
- Workflow control**
- Inputs**

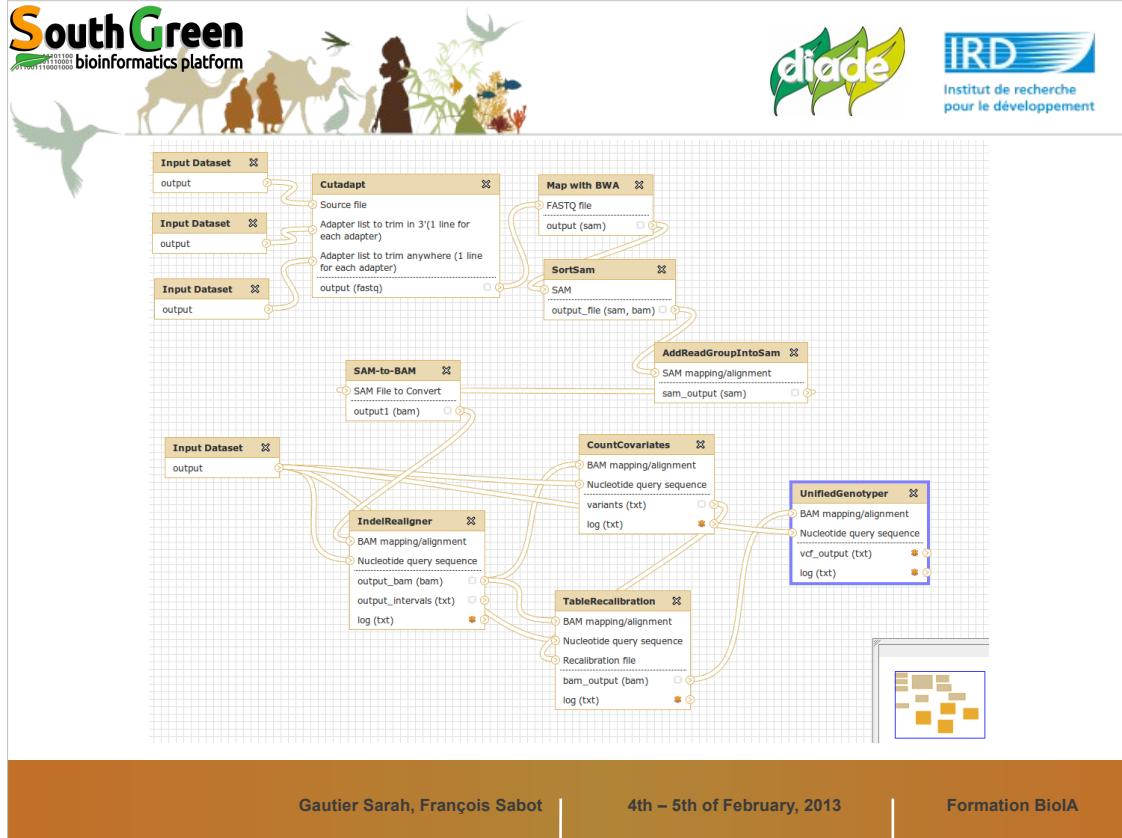
At the bottom of the interface, there is a footer bar with the following information:

- Gautier Sarah, François Sabot
- 4th – 5th of February, 2013
- Formation BioIA

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.



Gautier Sarah, François Sabot

4th – 5th of February, 2013

Formation BioIA