

Analyse technique de logiciel

Galaxy

Florian CARRE
Benjamin DARTIGUES
Sebastien BEAUQUIS
Guillaume BERNARD
Tom LESLUYES

Université de Bordeaux 1

30 Janvier 2012

Sommaire

1 NGS : Next Generation Sequencing

- Pyroséquençage
- Illumina

2 Utilisation

- Utilisation en ligne
- Utilisation en local

3 Fonctionnalités

- Présentation générale
- Ajout de plug-ins
- Workflow

Sommaire

1 NGS : Next Generation Sequencing

- Pyroséquençage
- Illumina

2 Utilisation

- Utilisation en ligne
- Utilisation en local

3 Fonctionnalités

- Présentation générale
- Ajout de plug-ins
- Workflow

Sommaire

1 NGS : Next Generation Sequencing

- Pyroséquençage
- Illumina

2 Utilisation

- Utilisation en ligne
- Utilisation en local

3 Fonctionnalités

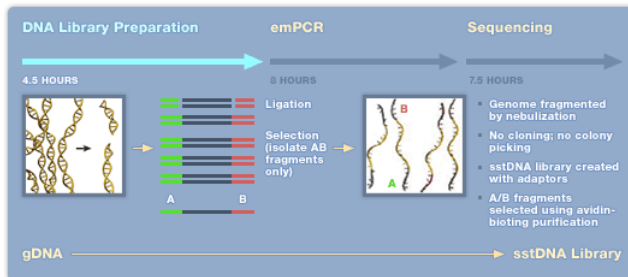
- Présentation générale
- Ajout de plug-ins
- Workflow

NGS

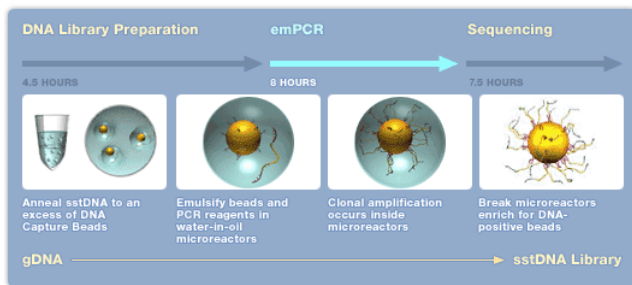
Présentation des techniques de NGS

- la technique 454 : le pyrosequençage
- la méthode de sequençage Illumina.

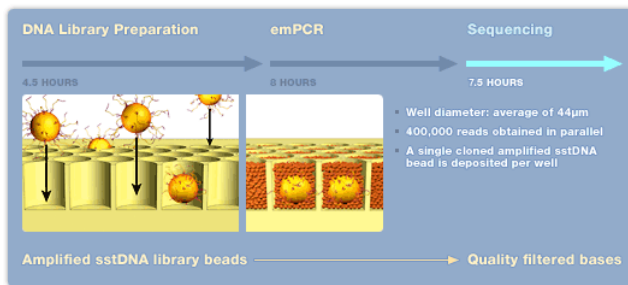
Le pyroséquençage

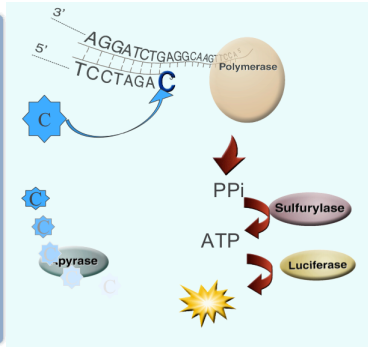


Le pyroséquençage

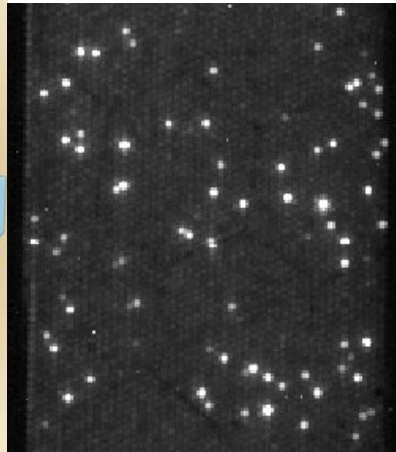
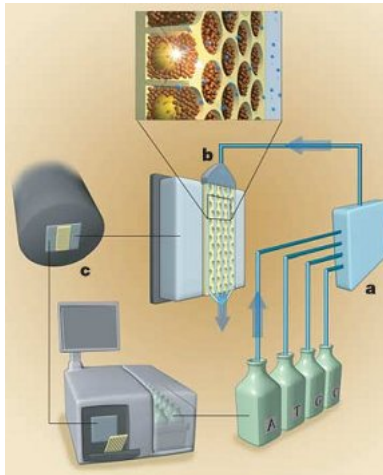


Le pyroséquençage





Le pyroséquençage



Illumina

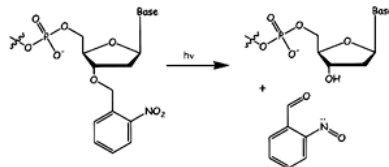


Illustration de la déprotection d'un nucléotide protégé en 3'-O par le groupement 2 nitrophénil après illumination au UV >30 nm

De Sanger à Illumina

CRT : Cycle Reversible Termination



Trois types d'utilisation différentes

- Utilisation en ligne.
- Utilisation en local.
- Utilisation sur un "cloud".

Avantages et inconvénient

Avantages de l'utilisation en ligne

- Aucune installation nécessaire.
- Interface ergonomique.
- Nombreux "tutoriels" et aides.

Inconvénient de l'utilisation en ligne

- Connexion internet requise.

Interface graphique

Liste d'outils :

Chargement de données
Manipulation de fichiers
Analyses (GWAS, Analyse
NGS...)

Interface de chargement des outils

Ici nous avons sélectionné
l'outil : Get Data > Upload File

Votre espace :

Fichiers chargés dans Galaxy
Historique des analyses
Fichiers résultats générés

The screenshot shows the Galaxy web interface. On the left, a sidebar lists various tools under the 'Tools' tab. A blue arrow points from the 'Liste d'outils' text to the 'Tools' sidebar. In the center, the 'Upload File (version 1.1.3)' tool is selected, and a blue arrow points from the 'Interface de chargement des outils' text to this tool. The tool interface shows options for 'File Format' (Auto-detect), 'File' (with a 'Browse...' button), and 'URL/Text'. A blue arrow points from the 'Votre espace' text to the 'History' tab on the right, which shows a list of uploaded files, including '1: ConsensusCh-2.fa'.

Galaxy Analyze Data Workflow Upload Data Visualization Help User

Tools Options ▾

search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Human Genome Variation
Genome Diversity
EMBOSS

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Browse...

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
Please create or log in to a Galaxy account to view files uploaded via FTP.		
This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at main.g2.bx.psu.edu using your Galaxy credentials (email address and password).		

Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
Click to Search or Select

History Options ▾

0 bytes
1: ConsensusCh-2.fa 0 0

Page d'aide d'un des outils de Galaxy

The screenshot shows the Galaxy web interface. At the top is a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Help, and User. On the left is a sidebar menu with various tool categories like Tools, Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, and Metagenomic analyses. The main content area displays the help page for the BED format tool. It includes a warning that the tool only works with BED files containing 3 or 6 fields. It explains that the BED format represents a single gene with information about exons, coding sequence, and UTRs. An example shows how two BED lines are converted into a single line describing a gene. It also lists the fields returned by the tool and provides a detailed list of the BED format fields and their meanings.

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Tools Options

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Gene BED To Exon/Intron /Codon BED expander
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses

Warning: THIS TOOL WORKS ONLY ON BED FILES THAT CONTAINS 3 OR 6 FIELDS (SEE EXAMPLE AND ABOUT FORMATS BELOW). THE OUTPUT WILL BE EMPTY IF APPLIED TO A BED FILE WITH 3 OR 6 FIELDS.

What it does

BED format can be used to represent a single gene in just one line, which contains the information about exons, coding sequence location (CDS), and positions of untranslated regions (UTRs). This tool *unpacks* this information by converting a single line describing a gene into a collection of lines representing individual exons, introns, UTRs, etc.

Example

Extracting Coding Exons + UTR Exons from the following two BED lines:

```
chr7 127475281 127491632 NM_000230 0 + 127486022 127488767 0 3 29,172,3225, 6,10713,13126
chr7 127486011 127488900 D49487 0 + 127486022 127488767 0 2 155,490, 6,2399
```

will return:

```
chr7 127475281 127475310 NM_000230 0 +
chr7 127485994 127486166 NM_000230 0 +
chr7 127488407 127491632 NM_000230 0 +
chr7 127486011 127486166 D49487 0 +
chr7 127488410 127488900 D49487 0 +
```

About formats

BED format Browser Extensible Data format was designed at UCSC for displaying data tracks in the Genome Browser. It has three required fields and additional optional ones. In the specific case of this tool the following fields must be present:

1. chrom - The name of the chromosome (e.g. chr1, chrX_random).
2. chromStart - The starting position in the chromosome. (The first base in a chromosome is numbered 0.)
3. chromEnd - The ending position in the chromosome, plus 1 (i.e., a half-open interval).
4. name - The name of the BED line.
5. score - A score between 0 and 1000.
6. strand - Defines the strand - either '+' or '-'.
7. thickStart - The starting position where the feature is drawn thickly at the Genome Browser.
8. thickEnd - The ending position where the feature is drawn thickly at the Genome Browser.
9. reserved - This should always be set to zero.
10. blockCount - The number of blocks (exons) in the BED line.
11. blockLines - A comma-separated list of the block lines. The number of items in this list should correspond to blockCount.
12. blockStarts - A comma-separated list of block starts. All of the blockStart positions should be calculated relative to chr.

Avantages et Inconvénient

Avantages de l'utilisation en local

- Ne nécessite pas de connexion internet.
- Possibilité de modifier les paramètres des plug-ins.
- Possibilité d'ajouter des plug-ins.
- Vitesse d'analyse.
- Conservation des données sensibles.

Inconvénient de l'utilisation en local

- Installation nécessaire.

Téléchargement du code source

- Récupération de la dernière version de Galaxy depuis Bitbucket.
- Téléchargement du répertoire Galaxy à partir de Mercurial.

Récupération à partir de Mercurial

Commandes pour copier le répertoire

```
mkdir Galaxy  
cd Galaxy  
hg clone https ://bitbucket.org/galaxy/galaxy-dist/
```

Commandes pour effectuer les mises à jour

```
hg incoming  
hg pull -u
```

Exécution du serveur local

Commandes pour exécuter le script *run.sh*

sh run.sh # commande de base

sudo run.sh # commande administrateur

Adresse du serveur de Galaxy

serving on http://127.0.0.1:8080

Exécution du serveur local

The screenshot shows the Galaxy web interface in a browser window. The address bar displays `http://127.0.0.1:8080/`. The interface includes a top navigation bar with links like 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. A left sidebar lists various tools under the 'Tools' section, including 'Get Data', 'Send Data', 'Encode Tools', 'Lift-Shift', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Fetch Alignments', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Wesley Analysis', 'Graph/Display Data', 'Regional Variation', 'Multiple regression', 'Multivariate Analysis', 'Evolution', 'Motif Tools', 'Multiple Alignments', 'Metagenomic analysis', 'FASTA manipulation', 'NCBI BLAST+', 'NGS: QC and manipulation', 'NGS: Picard (beta)', 'NGS: Mapping', 'NGS: Indel Analysis', 'NGS: RNA Analysis', 'NGS: SAM Tools', 'NGS: GATK Tools (beta)', 'NGS: Peak Calling', 'NGS: Simulation', 'SNP/WGA: Data, Filters', 'SNP/WGA: QC, LD, Plots', 'SNP/WGA: Statistical Models', 'Human Genome Variation', 'Genome Diversity', 'VCF Tools', and 'Misc Tools'. The main content area features a green status bar with a checkmark and the text 'Hello world! It's running...'. Below this is a message: 'To customize this page edit `static/newhome.html`'. A central graphic asks 'WWFMSD? grow nooxily appendages...' and includes a diagram of a workflow. The bottom of the main area states: 'This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.' The right sidebar shows a 'History' section with 'Options...' and 'Unnamed history' (0 bytes). A message box indicates: 'Your history is empty. Click "Get Data" on the left pane to start.'

Fonctionnalités

Très nombreuses

- Présentation générale
- Ajout de plug-ins
- Workflow

Fonctionnalités

Très nombreuses

- Présentation générale
- Ajout de plug-ins
- Workflow

Présentation générale

Pré-traitement

- Manipulation de fichiers
 - ouverture de fichiers volumineux
 - ajout/suppression de lignes
 - concaténation, filtrage, intersection
 - etc ...
- Opérations sur les données
 - addition, soustraction, moyenne, calcul de taille de séquences
 - conversion, formatage
 - etc ...

Présentation générale

Pré-traitement

- Manipulation de fichiers
 - ouverture de fichiers volumineux
 - ajout/suppression de lignes
 - concaténation, filtrage, intersection
 - etc ...
- Opérations sur les données
 - addition, soustraction, moyenne, calcul de taille de séquences
 - conversion, formatage
 - etc ...

Présentation générale

Traitement

- Analyse de séquences
 - calcul de corrélation
 - recherche d'orthologues
 - utilisation des outils d'EMBOSS
 - etc ...
- Visualisation des données
 - alignements multiples
 - distribution de données (histogramme, scatterplot)
 - arbres phylogéniques
 - etc ...

Présentation générale

Traitement

- Analyse de séquences
 - calcul de corrélation
 - recherche d'orthologues
 - utilisation des outils d'EMBOSS
 - etc ...
- Visualisation des données
 - alignements multiples
 - distribution de données (histogramme, scatterplot)
 - arbres phylogéniques
 - etc ...

Ajout de plug-ins

- Instance locale
- Langages interprétés
- Langages compilés

Ajout de plug-ins

- Instance locale
- Langages interprétés
- Langages compilés



Calcul du GC% à l'aide d'un script Perl

```
1  #!/usr/bin/perl -w
2  open (IN, "<$ARGV[0]>");
3  open (OUT, ">$ARGV[1]>");
4  while (<IN>) {
5      chop;
6      if (m/^(>)/) {
7          s/^(>)/;
8          if ($> 1) {
9              print OUT sprintf("%.3f", $gc/$length) . "\n";
10         }
11         $gc = 0;
12         $length = 0;
13     } else {
14         ++$gc while m/[gc]/ig;
15         $length += length $_;
16     }
17 }
18 print OUT sprintf("%.3f", $gc/$length) . "\n";
19 close( IN );
20 close( OUT );
```

tool_conf.xml

```
1 <section name="MyTools" id="mTools">  
2   <tool file="myTools/toolExample.xml" />  
3 </section>
```

toolExample.xml

```
1 <tool id="fa_gc_content_1" name="Compute GC content">
2   <description>for each sequence in a file</description>
3   <command interpreter="perl">toolExample.pl $input $output</command>
4   <inputs>
5     <param format="fasta" name="input" type="data" label="Source file"/>
6   </inputs>
7   <outputs>
8     <data format="tabular" name="output" />
9   </outputs>
10  <tests>
11    <test>
12      <param name="input" value="fa_gc_content_input.fa"/>
13      <output name="out_file1" file="fa_gc_content_output.txt"/>
14    </test>
15  </tests>
16  <help>
17    This tool computes GC content from a FASTA file.
18  </help>
19 </tool>
```


Outil implémenté

MyTools

- Compute GC content for each sequence in a file

FIGURE: Ajout du script dans la liste d'outils

Compute GC content (version 1.0.0)

Source file:

1: Sequence.fasta ↕

Execute

FIGURE: Fichier d'entrée






2: Compute GC content on data 1	
1 line	
format: tabular, database: ?	
  	
 	
1	
0.476	

FIGURE: Résultat

Exemple de workflow

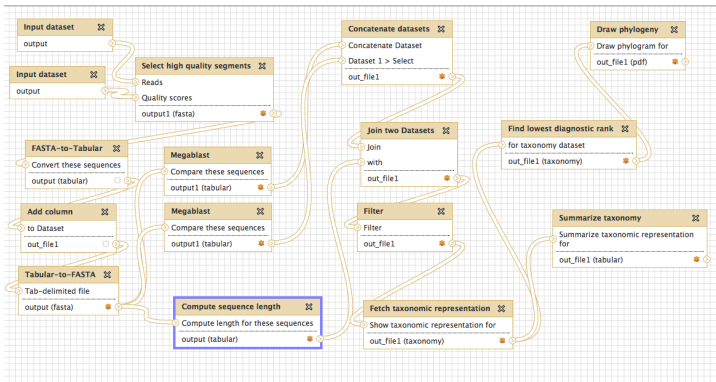


FIGURE: Workflow de métagénomique

Lancement de workflow

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 1%

Tools Options ▾

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Convert Formats
- FASTA manipulation
- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- Human Genome Variation
- Genome Diversity
- EMBOSS
- NGS TOOLBOX BETA
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Indel Analysis

Running workflow "Imported: metagenomic analysis" Expand All Collapse

Generic workflow for performing a metagenomic analysis on NGS data.

Step 1: Input dataset
454 Reads

reads 34: Tabular-to-FASTA on data 33
type to filter

Step 2: Input dataset
454 Quality Dataset

qualities 2: Trip A Left Side QV
type to filter

Step 3: Select high quality segments
Here we select segments of reads with contiguous high quality bases above threshold phred score of 20

Step 4: FASTA-to-Tabular
Convert to tabular format so that column for additional metadata can be added

Step 5: Add column
Add column for storing where the data came from (in the case of this data, this column indicates which trip it came from)

Step 6: Tabular-to-FASTA
Convert back to FASTA format for more analysis

Step 7: Megablast

Step 8: Megablast

History Options ▾

Unnamed history 2.6 Gb

- 44: Draw phylogeny on data 42
- 43: Summarize taxonomy on data 41
- 52: Find lowest diagnostic rank on data 41
- 41: Fetch taxonomic representation on data 40
- 34: Tabular-to-FASTA on data 33
- 15: Summarize taxonomic on data 13

15: Summarize taxonomic on data 13
15,719 lines
format: tabular, database: ?
info: SKIPPED line 29909: Too few fields
SKIPPED line 30627: Too few fields
SKIPPED line 77747: Too few fields
SKIPPED line 77748: Too few fields
SKIPPED line 77749: Too few fields
SKIPPED line 77750: Too few fields
SKIPPED line 77751: Too few fields
SKIPPED !!

1	2	3
root	NYLL	6465
root	root	2243054
spargkington	156763980	11
spargkington	156763615	2

FIGURE: Lancement du workflow

Résultat du workflow

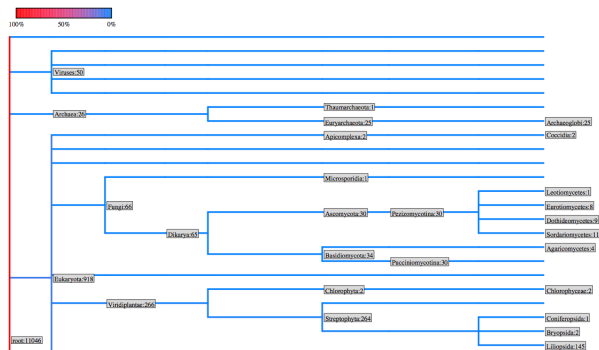


FIGURE: Résultat du workflow