

Analyse technique de logiciel Galaxy

Florian CARRE
Benjamin DARTIGUES
Sebastien BEAUQUIS
Guillaume BERNARD
Tom LESLUYES

Université de Bordeaux 1

30 Janvier 2012

Sommaire

- 1 Etat de l'art
 - Les logiciels de workflow
- 2 NGS : Next Generation Sequencing
 - Pyroséquençage
 - Illumina
- 3 Utilisation
 - Utilisation en ligne
 - Utilisation en local
- 4 Fonctionnalités
 - Présentation générale
 - Ajout de plug-ins
 - Workflow

Sommaire

- 1 Etat de l'art
 - Les logiciels de workflow
- 2 NGS : Next Generation Sequencing
 - Pyroséquençage
 - Illumina
- 3 Utilisation
 - Utilisation en ligne
 - Utilisation en local
- 4 Fonctionnalités
 - Présentation générale
 - Ajout de plug-ins
 - Workflow

Sommaire

- 1 Etat de l'art
 - Les logiciels de workflow
- 2 NGS : Next Generation Sequencing
 - Pyroséquençage
 - Illumina
- 3 Utilisation
 - Utilisation en ligne
 - Utilisation en local
- 4 Fonctionnalités
 - Présentation générale
 - Ajout de plug-ins
 - Workflow

Sommaire

- 1 Etat de l'art
 - Les logiciels de workflow
- 2 NGS : Next Generation Sequencing
 - Pyroséquençage
 - Illumina
- 3 Utilisation
 - Utilisation en ligne
 - Utilisation en local
- 4 Fonctionnalités
 - Présentation générale
 - Ajout de plug-ins
 - Workflow

les logiciels de workflow

Présentation générales

On distingue deux types de logiciels de workflow :

- les logiciels de workflow 1
- les logiciels de workflow 2

○○○○○
○○○○○
○○○○○○○
○○○○○
○○○○

les logiciels de workflow 1

Ergatis, BIOMOBY

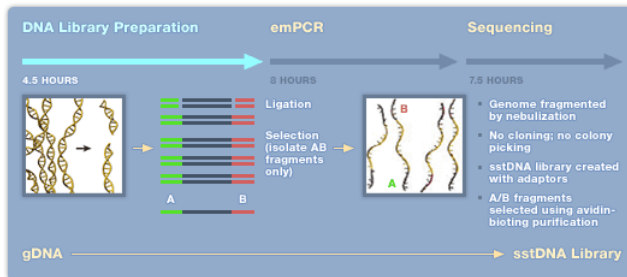
- Ergatis
- BIOMOBY

NGS

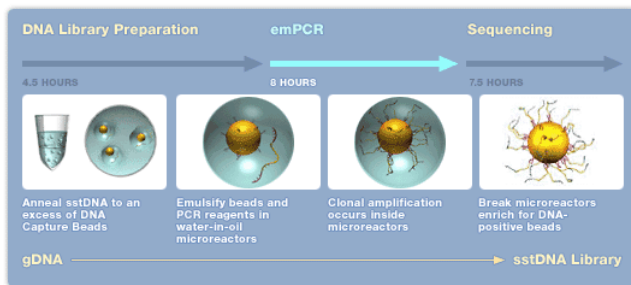
Présentation des techniques de NGS

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

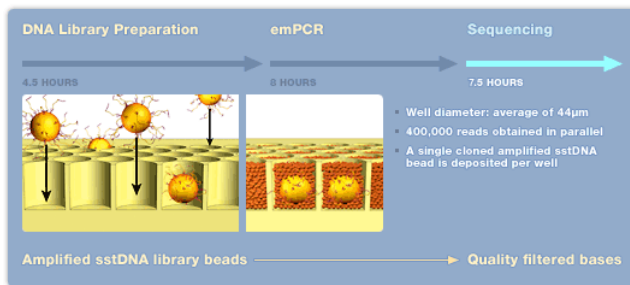
Le pyroséquençage



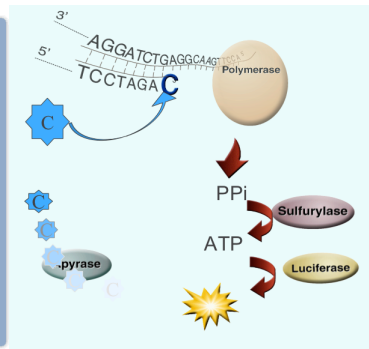
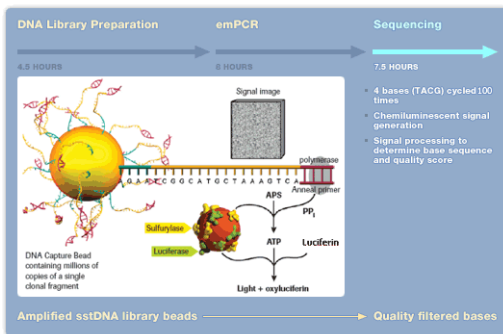
Le pyroséquençage



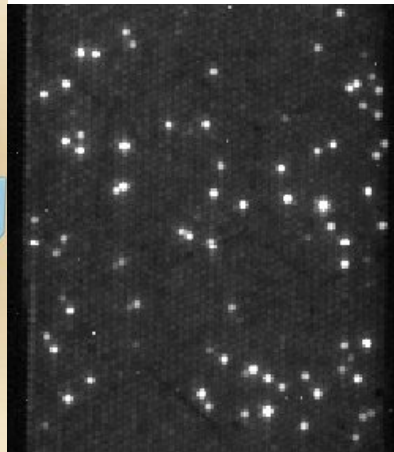
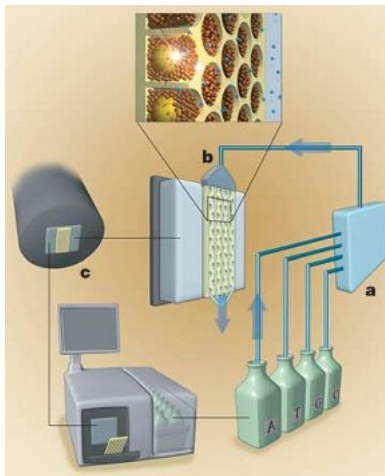
Le pyroséquençage



Le pyroséquençage



Le pyroséquençage





Illumina

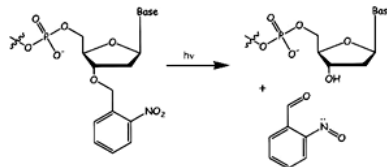


Illustration de la déprotection d'un nucléotide protégé en 3'-O par le groupement 2 nitrophényle après illumination au UV >30 nm

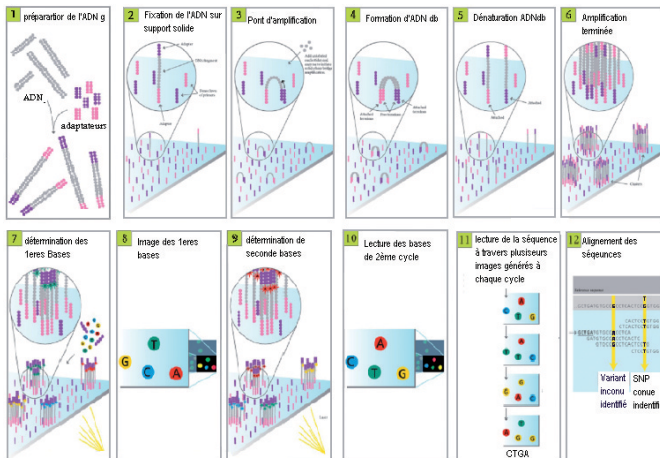
De Sanger à Illumina

CRT : Cycle Reversible Termination



Illumina

Illumina



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. The 'Get Data' category is expanded, and 'Upload File' is selected. The main panel displays the 'Upload File (version 1.1.3)' tool interface. It includes a 'File Format' dropdown set to 'Auto-detect', a 'File' input field with a 'Browse...' button, and a 'URL/Text' input area. A tip mentions that files larger than 2GB may fail due to browser limitations. Below the input fields, there is a section for 'Files uploaded via FTP' with a table showing columns for File, Size, and Date. The table is currently empty. At the bottom, there are checkboxes for 'Convert spaces to tabs' and a 'Genome' dropdown menu. On the right side of the interface, a 'History' panel shows a list of uploaded files, with '1: ConsensusCh-2.fa' selected.



Page d'aide d'un des outils de 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](#)

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

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

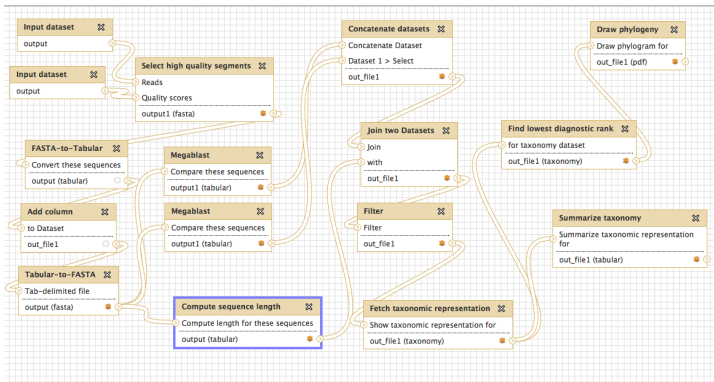


FIGURE: Workflow de métagénomique

Lancement de workflow

The screenshot displays the Galaxy web interface. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The main content area is titled 'Running workflow "Imported: metagenomic analysis"' and shows a generic workflow for performing a metagenomic analysis on NGS data. The workflow consists of eight steps:

- Step 1: Input dataset** (454 Reads): A dropdown menu for 'reads' is set to '34: Tabular-to-FASTA on data 33'.
- Step 2: Input dataset** (454 Quality Dataset): A dropdown menu for 'qualities' is set to '2: Trip A Left Side QV'.
- Step 3: Select high quality segments**: A description states 'Here we select segments of reads with contiguous high quality bases above threshold phred score of 20'.
- Step 4: FASTA-to-Tabular**: A description states 'Convert to tabular format so that column for additional metadata can be added'.
- Step 5: Add column**: A description states '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**: A description states 'Convert back to FASTA format for more analysis'.
- Step 7: Megablast**
- Step 8: Megablast**

The right sidebar shows the 'History' section with a list of datasets and workflows. The bottom of the interface shows a table with columns 1, 2, and 3, containing data for 'root', 'superkingdom', and 'eukaryotid'.

FIGURE: Lancement du workflow

Résultat du workflow

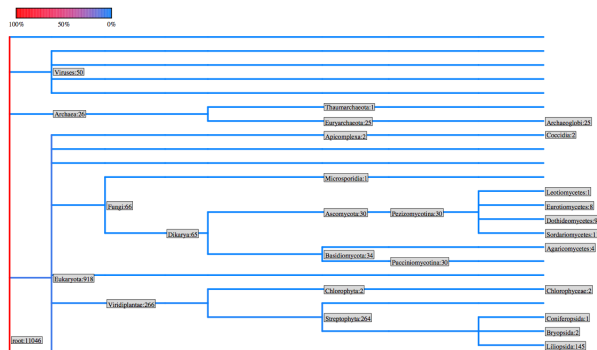


FIGURE: Résultat du workflow

Merci de votre attention

