Project S. 2: NGS - Instructions

Ahmed Sayadi

asayadi@univ-catholyon.fr

FISE2 – 2025/2026

Introduction

- Les avancées technologiques autour des données de séquençage ont dépassé toutes les attentes ces dernières décennies.

- Le développement de nouvelles stratégies de séquençage, accompagné d'une amélioration continue des logiciels bioinformatiques, est prometteur et inarrêtable.

- Une bonne maîtrise des méthodes bioinformatiques et une interprétation rigoureuse des résultats sont essentielles pour exploiter pleinement ces données,

Introduction

- Un des objectifs de ces travaux pratiques est de se familiariser avec les outils et méthodes bioinformatiques couramment utilisés pour analyser des données de séquençage. Pour cela, nous travaillerons avec des données réelles issues de génomique et d'expression génique, en appliquant des méthodes de pointe.

- Vous commencerez par choisir un article scientifique récent, puis vous réanalyserez les données de l'étude de manière similaire aux auteurs, afin de réévaluer leurs conclusions biologiques.

Introduction

- La première tâche consistera à **planifier les analyses que vous devrez réaliser**.
- Vous **adapterez les analyses en fonction des outils disponibles** et de vos propres intérêts. Ainsi, même en partant des mêmes données, chacun pourra emprunter un chemin différent et obtenir des résultats distincts.

- **Créez votre propre « pipleline »**, écrivez vos scripts en fonction de vos objectifs spécifiques, et intégrez les vérifications nécessaires pour vous assurer que chaque étape fonctionne comme prévu.

Paper

RESEARCH ARTICLE Open Access

RNA-seq and Tn-seq reveal fitness determinants of vancomycin-resistant *Enterococcus faecium* during growth in human serum

Xinglin Zhang[1,2], Vincent de Maat[2], Ana M. Guzmán Prieto[2], Tomasz K. Prajsnar[3], Jumamurat R. Bayjanov[2], Mark de Been[2], Malbert R. C. Rogers[2], Marc J. M. Bonten[2], Stéphane Mesnage[3], Rob J. L. Willems[2] and Willem van Schaik[2,4*]

Abstract

Background: The Gram-positive bacterium *Enterococcus faecium* is a commensal of the human gastrointestinal tract and a frequent cause of bloodstream infections in hospitalized patients. The mechanisms by which *E. faecium* can survive and grow in blood during an infection have not yet been characterized. Here, we identify genes that contribute to growth of *E. faecium* in human serum through transcriptome profiling (RNA-seq) and a high-throughput transposon mutant library sequencing approach (Tn-seq).

Results: We first sequenced the genome of *E. faecium* E745, a vancomycin-resistant clinical isolate, using a combination of short- and long read sequencing, revealing a 2,765,010 nt chromosome and 6 plasmids, with sizes ranging between 9.3 kbp and 223.7 kbp. We then compared the transcriptome of *E. faecium* E745 during exponential growth in rich medium and in human serum by RNA-seq. This analysis revealed that 27.8% of genes on the *E. faecium* E745 genome were differentially expressed in these two conditions. A gene cluster with a role in purine biosynthesis was among the most upregulated genes in *E. faecium* E745 upon growth in serum. The *E. faecium* E745 transposon mutant library was then used to identify genes that were specifically required for growth of *E. faecium* in serum. Genes involved in de novo nucleotide biosynthesis (including *pyrK_2, pyrF, purD, purH*) and a gene encoding a phosphotransferase system subunit (*manY_2*) were thus identified to be contributing to *E. faecium* growth in human serum. Transposon mutants in *pyrK_2, pyrF, purD, purH* and *manY_2* were isolated from the library and their impaired growth in human serum was confirmed. In addition, the *pyrK_2* and *manY_2* mutants were tested for their virulence in an intravenous zebrafish infection model and exhibited significantly attenuated virulence compared to *E. faecium* E745.

Conclusions: Genes involved in carbohydrate metabolism and nucleotide biosynthesis of *E. faecium* are essential for growth in human serum and contribute to the pathogenesis of this organism. These genes may serve as targets for the development of novel anti-infectives for the treatment of *E. faecium* bloodstream infections.

Keywords: *Enterococcus faecium*, Transcriptome, Transposon mutant library screening, Nucleotide biosynthesis, Carbohydrate metabolism, Virulence, Zebrafish

Introduction

- Comme mentionné précédemment, vous allez utiliser les données réelles de l'article que vous avez choisis. Toutes ces données sont disponibles dans l'Archive de Lecture de Séquences (https://www.ncbi.nlm.nih.gov/sra/).
- Recherchez dans le papier le numéro d'accession utilisé pour soumettre les données de séquençage.
- Enfin, il est crucial dans tout projet scientifique de documenter les analyses réalisées. Pour cela, vous commencerez par créer un compte de dépôt **GitHub** où vous enregistrerez les codes et méthodes utilisés.

Project Planning

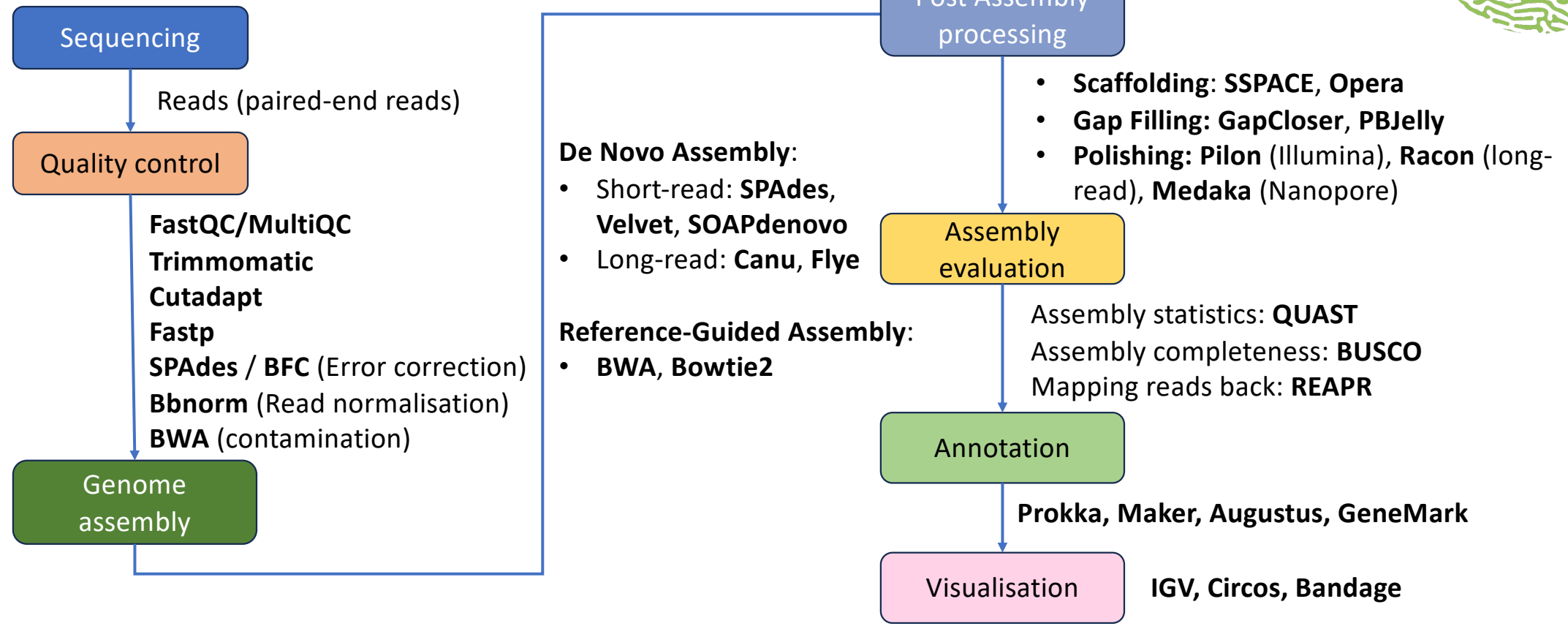
La première étape de ces travaux pratiques sera donc de **créer un plan de projet**. Celui-ci doit aborder au moins les points suivants :

- **Quel est l'objectif de votre projet ?** Quelles questions souhaitez-vous explorer à travers votre recherche ?
- **Quelles analyses effectuerez-vous pour répondre à ces questions ?** Dans quel ordre ? Quels logiciels utiliserez-vous ? Y a-t-il des étapes plus longues ?
- **Quels types de données allez-vous manipuler ? Comment allez-vous organiser vos données ?**
- Ce plan vous guidera tout au long du projet et facilitera la gestion de vos analyses.

Project Planning

- Commencez par **lire l'article**, il est essentiel de l'avoir lu et compris.
- Il est primordial de définir dès le début les questions auxquelles vous souhaitez répondre.
- Assurez-vous de bien connaître les données accessibles, telles que le **type de données** (RNA-seq, séquençage de génome entier, lectures courtes/longues, etc.), la source biologique (tissu, échantillon, etc.), et les conditions de génération (type de librairies, adaptateurs spécifiques, éventuels problèmes, etc.). Ces détails influenceront le **choix des analyses** et **logiciels**.
- Une méthode intuitive consiste à dessiner un schéma reliant les données d'entrée et de sortie pour chaque étape, les analyses à réaliser et les outils à utiliser, comme illustré dans la Figure suivante.

Project Planning



Project Data

- Les données avec lesquelles vous allez travailler proviennent du **NCBI Sequence Read Archive (SRA)**.
- Les fichiers que vous avez ont été tronqués afin de limiter leur taille et, par conséquent, le temps de calcul.
- Les fichiers sont nommés selon leur identifiant **SRA**, (voir l'article).
- Ces étapes vous aideront à bien comprendre vos données.
  - Accédez à la page du **SRA** : http://www.ncbi.nlm.nih.gov/sra.
  - Recherchez l'identifiant **SRA** de vos échantillons (https://www.ncbi.nlm.nih.gov/sra/ERX1864560).
  - Sous "**Study**", cliquez sur **All Runs** pour voir les ensembles de données associés.
  - Sous "**Select**", cliquer sur "**Metadata**" pour téléchargez **SRArunTable** afain d'obtenir les métadonnées.

Project Data

Voici quelques recommandations pour créer une **table de métadonnées** pour vos fichiers de séquence :

- **Une ligne par échantillon, une colonne par variable** : Par exemple, toutes les informations sur ERR00001 doivent figurer sur une seule ligne, avec des colonnes pour les variables : identifiant_SRA, tissu, stade, type_de_séquençage, etc.
- **Ne fusionnez pas les cellules !**
- **Évitez les espaces !** Utilisez plutôt '-' ou '_'.
- **Soyez cohérent** dans vos formats.
- **Utilisez des noms explicites** pour les variables.
- **Enregistrez les données en fichiers texte** au format .tsv (valeurs séparées par des tabulations) ou .csv (valeurs séparées par des virgules).

Project Data

Voici quelques conseils pour organiser votre répertoire de travail :

- **Séparez les données et le code** : Gardez les fichiers code dans un dossier séparé des petits fichiers de données (résultats, figures, textes).
- **Noms uniques et informatifs pour les fichiers de données** : Donnez des noms clairs aux fichiers pour les identifier facilement.
- **Numérotez les dossiers/fichiers** pour les organiser facilement par ordre de création.

Project Data

- Utiliser des liens symboliques qui créent des raccourcis vers les fichiers d'origine.

```
$ ln -s /path/to/original/file /path/to/new/file
```

- **Compressez les fichiers de données**, surtout les gros fichiers (ex : FASTQ, SAM), car l'espace de stockage est limité. La compression n'affecte pas la plupart des outils, et les fichiers peuvent être décompressés à la volée si nécessaire.

```
# Uncompressing and piping
$ zcat sampleA.fastq.gz | wc -l
```

```
genome_analyses/
├── analyses
│   ├── 01_preprocessing
│   │   ├── trimming_software
│   │   │   ├── data/trimmed_data/ERR00001.trimmed.fastq.gz
│   │   │   └── data/trimmed_data/ERR00002.trimmed.fastq.gz
│   │   ├── fastqc_raw
│   │   │   └── fastqc_report.txt
│   │   └── fastqc_trim
│   │       └── fastqc_report.txt
│   ├── 02_genome_assembly
│   │   ├── assembly_softwareA_settingsX
│   │   └── assembly_softwareB_settingsX
│   └── 03_structural_annotation
│       └── annotation_software_settingsX
├── code
│   ├── 0_download_data.sh
│   ├── 1_preprocessing.sh
│   ├── 2_genome_assembly.sh
│   └── 3_structural_annotation.sh
├── data
│   ├── metadata
│   │   └── sample_information.csv
│   ├── raw_data
│   │   ├── ERR000002.fastq.gz
│   │   └── ERR00001.fastq.gz
│   └── trimmed_data
│       ├── ERR00001.trimmed.fastq.gz
│       └── ERR00002.trimmed.fastq.gz
```