CHOLET Pierre

Assemblage d'un pipeline d'analyse bioinformatique pour l'analyse de bactéries pathogènes par séquençage à haut débit

2ème année DUT Génie Biologique

Option Bio-Informatique

Année 2014/2015

Groupe TPA

Maître de stage : BONNET R.
Tutrice de stage : POLONAIS V.

Antenne d'Aurillac, Département Génie biologique
IUT de Clermont-Ferrand, Université d'Auvergne

Plan

Introduction

Problématique

I. Pré-traitement

II. Typage de séquence

III. Détection des gènes et plasmides

IV. Assemblage et Phylogénie

V. Résultats

Conclusion

Introduction

C.H.U → Centre National de Référence (CNR)

de la **Résistance aux Antibiotiques**

Caractérisation de souches bactériennes **multirésistantes**

Détection des gènes de résistance et de virulence

→ enzymes de résistance impliquées (β-lactamases, carbapénémases..)

→ comportement et facteurs de virulence (adhésion, toxines..)

Classification phylogénique

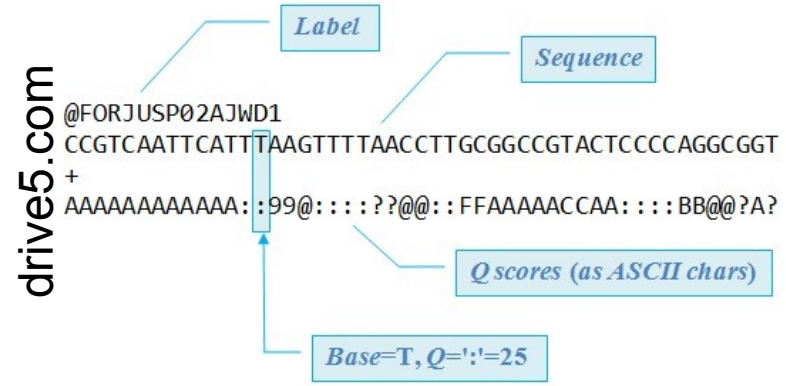
Problématique

- Comment organiser la surveillance de l'**émergence** et de la **diffusion** des **mécanismes de résistance aux antibiotiques** ?
- Comment permettre cette approche par des **moyens bioinformatiques** ?
- Les analyses bioinformatiques sont elles suffisamment **fiables** et représentent-elle un **avantage** comparé aux méthodes traditionnelles ?

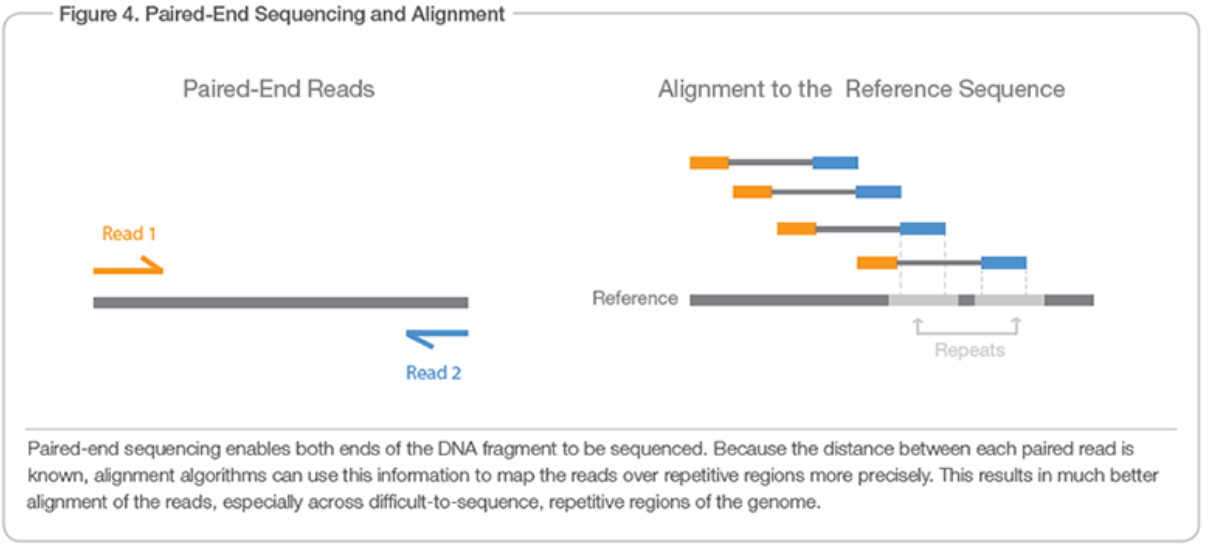
I. Pré-traitement : Principe

Séquençage Illumina en **paired-end**

Matériel : 2 fichiers fastq



drive5.com



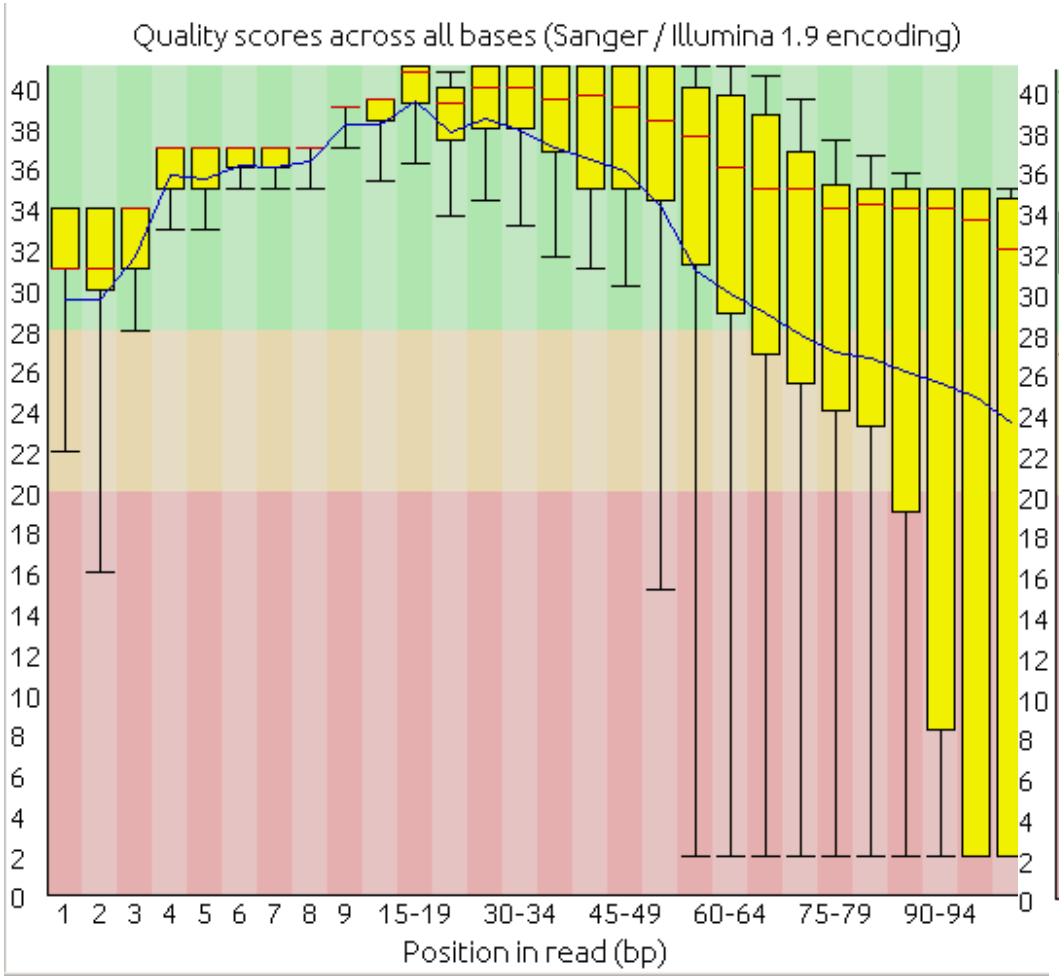
illumina.com/technology

Google Image

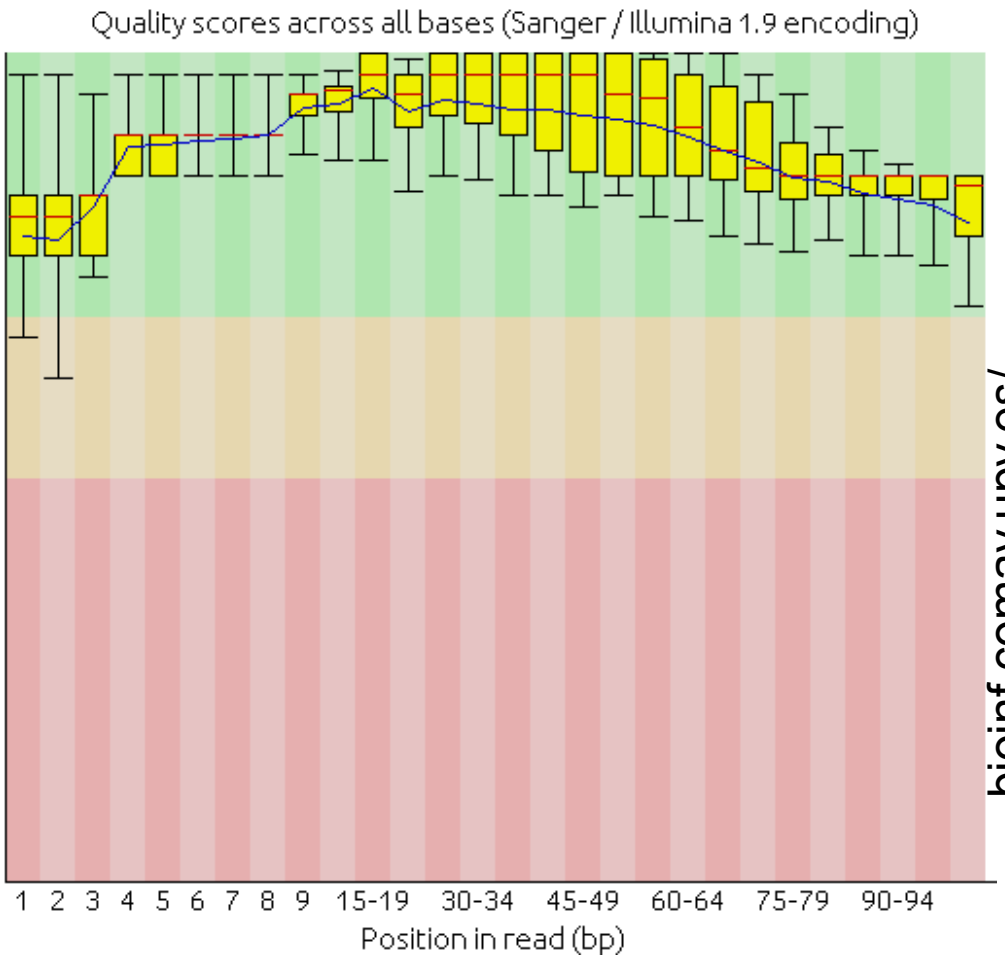
Trimming : réduire les séquences au seules régions de bonnes qualité et exclure les adaptateurs permettant le séquençage

I. Pré-traitement : Trimming

Score de qualité de chaque base d'une séquence arbitraire



Avant Trimming



Après

bioinf.comav.upv.es/

I. Pré-traitement : Procédé

2 outils : **Trim galore** et **Trimmomatic**

```
Fichier  Édition  Affichage  Rechercher  Terminal  Aide
  -h, --help            show this help message and exit
  --inf INPUT_FORWARD   Fichier "forward" du séquençage Paired-End
  --inr INPUT_REVERSE   Fichier "reverse" du séquençage Paired-End
  -s SCORE, --trm_score SCORE
                        Seuil de qualité minimale (défaut=[28])
  -t {galore,trimmomatic,skip}, --trimmer {galore,trimmomatic,skip}
                        Outil de trimming (défaut=trimmomatic)
  -n THREADS, --threads THREADS
                        Nombre de threads (processeurs) (défaut=1)
  --trm_stringence STRINGENCE
                        Stringence pour trim_galore
  --trm_clip            Trimmer 20 pb en 5' et 10 pb en 3'
  -l MIN_LEN, --trm_min_len MIN_LEN
                        Taille minimale pour retenir un read trimmé
                        (défaut=[100])
  --trm_bypass TRM_BYPASS [TRM_BYPASS ...]
                        Commande à entrer pour le trimming, entre quotes (ex:
                        "-q 28 --fastq") (Attention au trimmer choisi)
  -p {33,64}, --phred {33,64}
                        Encodage du score phred du fastq (défaut=33)
  --trm_adapter ADAPTATEUR
                        Adaptateur spécifique du séquençage pour
                        trim_galore
  --suffixe_forward SUFFIXE_F
```

Source personnelle à partir du pipeline

II. Typage de Séquence : MLST

Escherichia Coli : 8 gènes de ménage

- dinB
- pabB
- putP
- trpB
- icdA
- polB
- trpA
- uidA

La combinaison détermine un profil, le « Sequence Type »

N° Souche	ST	dinB	icdA	pabB	polB	putP	trpA	trpB	uidA
1ecoli	477	17	9	28	12	9	135	9	11
6coli	471	6	6	4	2	154	7	2	4
7coli	2	8	2	7	3	7	1	4	2
21coli	NF	30	45	33	37	27	34	24	-
22coli	43	9	1	15	7	4	9	6	9

Source personnelle à partir de SRST2

II. Typage de séquence : MLST

Nécessité d'une base de données

```
Fichier  Édition  Affichage  Rechercher  Terminal  Aide
                        Taux de couverture (en %) minimum pour rapporter un
                        allèle (défaut: 90)
  --srst_divergence SRST_DIVERGENCE
                        Taux de divergence (en %) maximum pour exclure un
                        allèle (défaut: 10)
  --mlst_db MLST_DB     Chemin vers la base de donnée MLST
  --mlst_definition MLST_DEFINITIONS
                        Chemin vers le fichier de définitions correspondant
                        à la base de donnée MLST
  --mlst_delimiter MLST_DELIMITEUR
                        Délimiteur gène/allèle du fichier de définitions
  --resist_db DETECTION_RESISTANCE_DB [DETECTION_RESISTANCE_DB ...]
                        Chemin(s) vers la(les) base(s) de données de gènes
                        de résistance (Fasta)
  --vir_db DETECTION_VIRULENCE_DB [DETECTION_VIRULENCE_DB ...]
                        Chemin(s) vers la(les) base(s) de données de gènes
                        de virulence (Fasta)
  --plsm_mlst PLASMIDES_MLST
                        Chemin(s) vers la(les) base(s) de données de MLST
                        pour les plasmides (Fasta)
  --plsm_def PLASMIDES_DEFINITIONS
                        Chemin(s) vers le(les) fichiers de profils MLST de
                        plasmides (Fasta)
  --plsm_del PLASMIDES_DELIMITEUR
```

Source personnelle à partir du pipeline

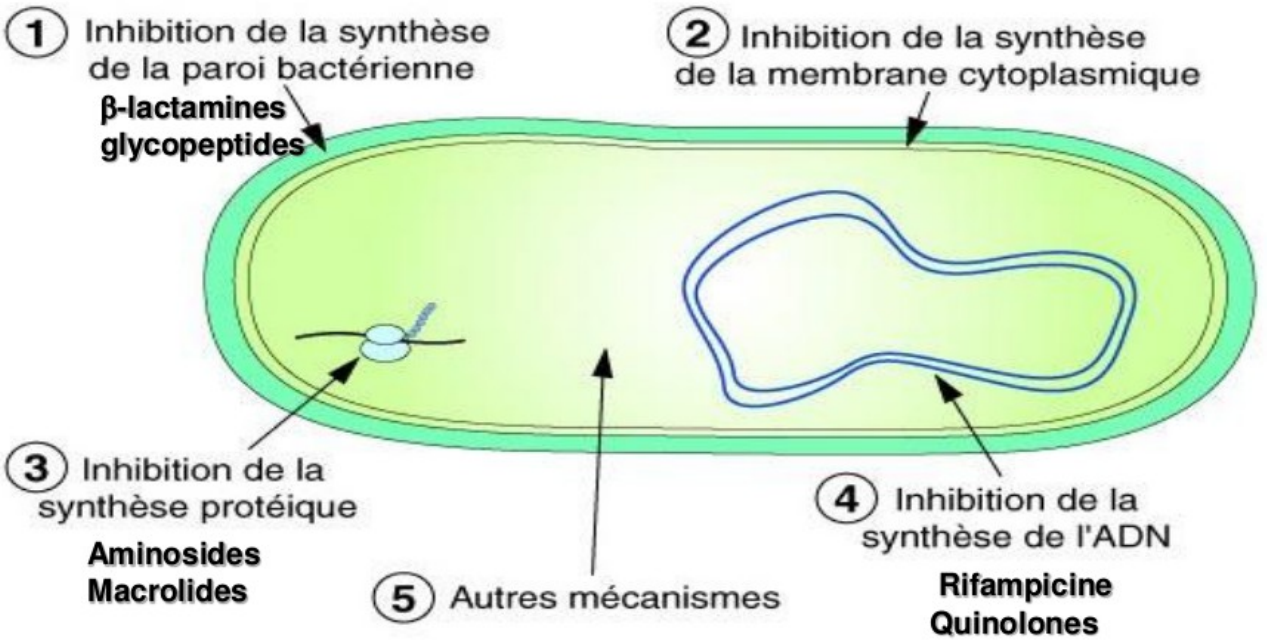
Les bases de données sont recensées sur pubmlst.org et sur le site de l'institut Pasteur

III. Détection : Gènes

Antibiotiques

Mécanisme d'action

inhibition de la synthèse de la paroi bactérienne :
sur les bactéries en phase de croissance - bactéricidie.



Molina, Jean-Michel ; L'essentiel en pratique sur les β-lactamines

III. Détection : Gènes

Résistance

Enzymes inactivatrices d'antibiotiques

- β-lactamines
- Aminosides
- Sulfamides
- Cyclines
- Phénicolés
- Macrolides

Virulence

Au sein d'**îlots de pathogénicité**

- Adhésion
- Capture du fer
- Capsule
- Systèmes de sécrétion et toxines

III. Détection : Gènes

2 paramètres :

- Bases de données de gènes de résistance
- Bases de données de gènes de virulence

```
Fichier  Édition  Affichage  Rechercher  Terminal  Aide
                        Taux de divergence (en %) maximum pour exclure un
                        allèle (défaut: 10)
  --mlst_db MLST_DB     Chemin vers la base de donnée MLST
  --mlst_definition MLST_DEFINITIONS
                        Chemin vers le fichier de définitions correspondant
                        à la base de donnée MLST
  --mlst_delimiter MLST_DELIMITEUR
                        Délimiteur gène/allèle du fichier de définitions
  --resist_db DETECTION_RESISTANCE_DB [DETECTION_RESISTANCE_DB ...]
                        Chemin(s) vers la(les) base(s) de données de gènes
                        de résistance (Fasta)
  --vir_db DETECTION_VIRULENCE_DB [DETECTION_VIRULENCE_DB ...]
                        Chemin(s) vers la(les) base(s) de données de gènes
                        de virulence (Fasta)
  --plsm_mlst PLASMIDES_MLST
                        Chemin(s) vers la(les) base(s) de données de MLST
                        pour les plasmides (Fasta)
  --plsm_def PLASMIDES_DEFINITIONS
                        Chemin(s) vers le(les) fichiers de profils MLST de
                        plasmides (Fasta)
  --plsm_del PLASMIDES_DELIMITEUR
                        Délimiteur gène/allèle des fichiers de défnitions
                        de MLST des plasmides
  --plsm_db PLASMIDES_DB [PLASMIDES_DB ...]
```

Source personnelle à partir du pipeline

III. Détection : Plasmides

Propagation de gènes de résistance par **transfert horizontal**

Possède une **origine de réplication**

2 mêmes origines de réplication → **type d'incompatibilité**

2 types : **conjugables** ou **mobilisables**

Transfert assuré par le système de sécrétion de type IV

Synthétisé seulement par les plasmides conjugables

IV. Assemblage

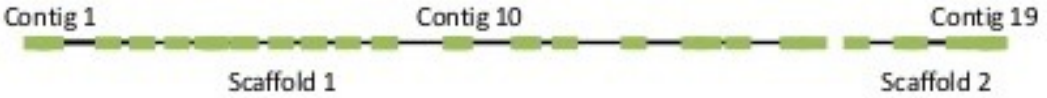
Assembly

Genome

Fragment

Sequence

Assemble



Inspiration: http://goo.gl/VOZVg

CENTER FOR MICROBIAL COMMUNITIES | AALBORG UNIVERSITY

slideshare.net
Mads Albertsen - Genomics and Metagenomics

IV. Phylogénie

Nécessite un génome de référence :

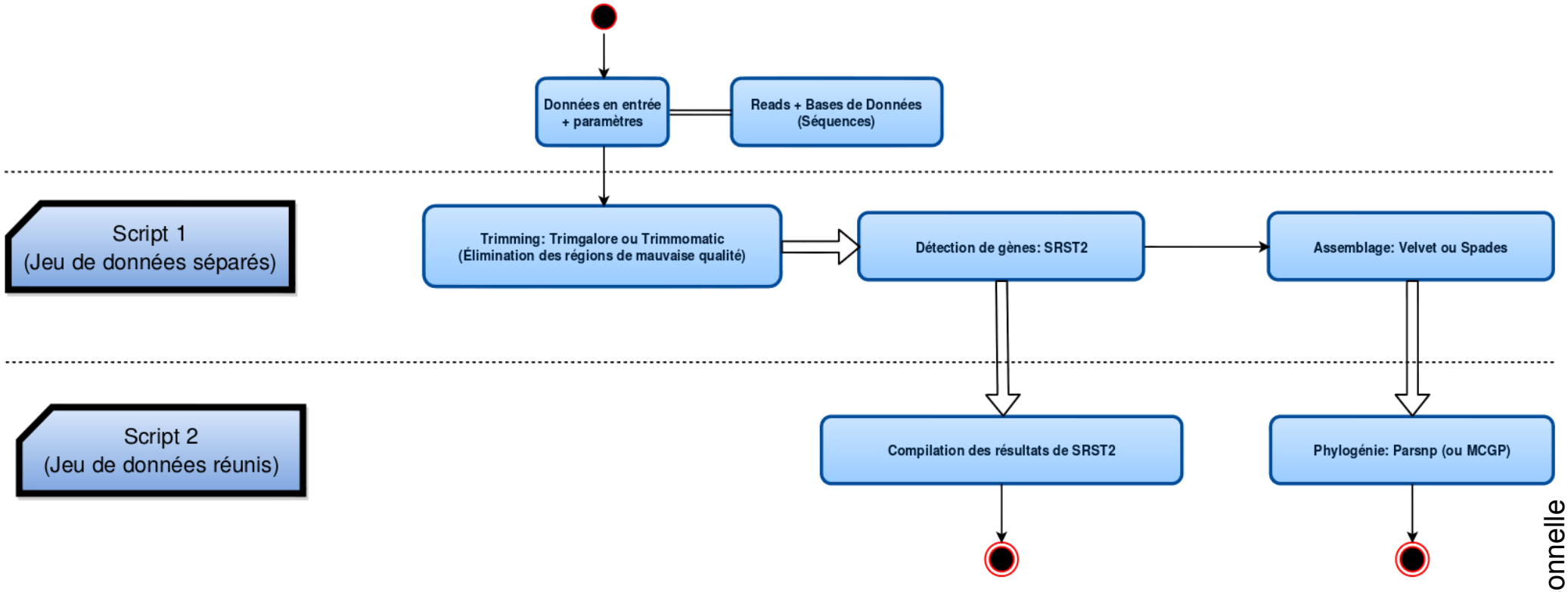
- Génome parmi ceux soumis à l'analyse
- Génome de référence au format fasta
- Génome annoté au format Genbank

```
Fichier  Édition  Affichage  Rechercher  Terminal  Aide
2ème partie du pipeline d'analyse de génomes bactériens, compilation et
formattage des résultats et phylogénie.

optional arguments:
  -h, --help            show this help message and exit
  --gb_ref GENBANK_REF  Génome de référence annoté au format genbank
  --fasta_ref FASTA_REF
                        Génome de référence au format fasta (inutile si le
                        genebank a été spécifié)
  --genome_type {contigs,scaffolds}
                        Type des fichiers du génome utilisé
  --no_recombination    Ne pas prendre en compte les recombinaisons (omettre
                        l'option -x de parsnp)
  --skip {compilation,phylogenie}
                        Étape à ne pas exécuter {compilation|phylogenie}
  --no_verbose          Désactive le mode verbose (informations
                        supplémentaires)
  --keep_files          Empêche la suppression des fichiers annexes et
                        rapports
  --clear_all           Attention: cette option annule la compilation de
                        résultats et la phylogénie mais supprime TOUS les
                        fichiers issus du pipeline (incluant tous les
                        résultats, veillez à les récupérer avant)
```

Source personnelle à partir du pipeline

V. Résultats

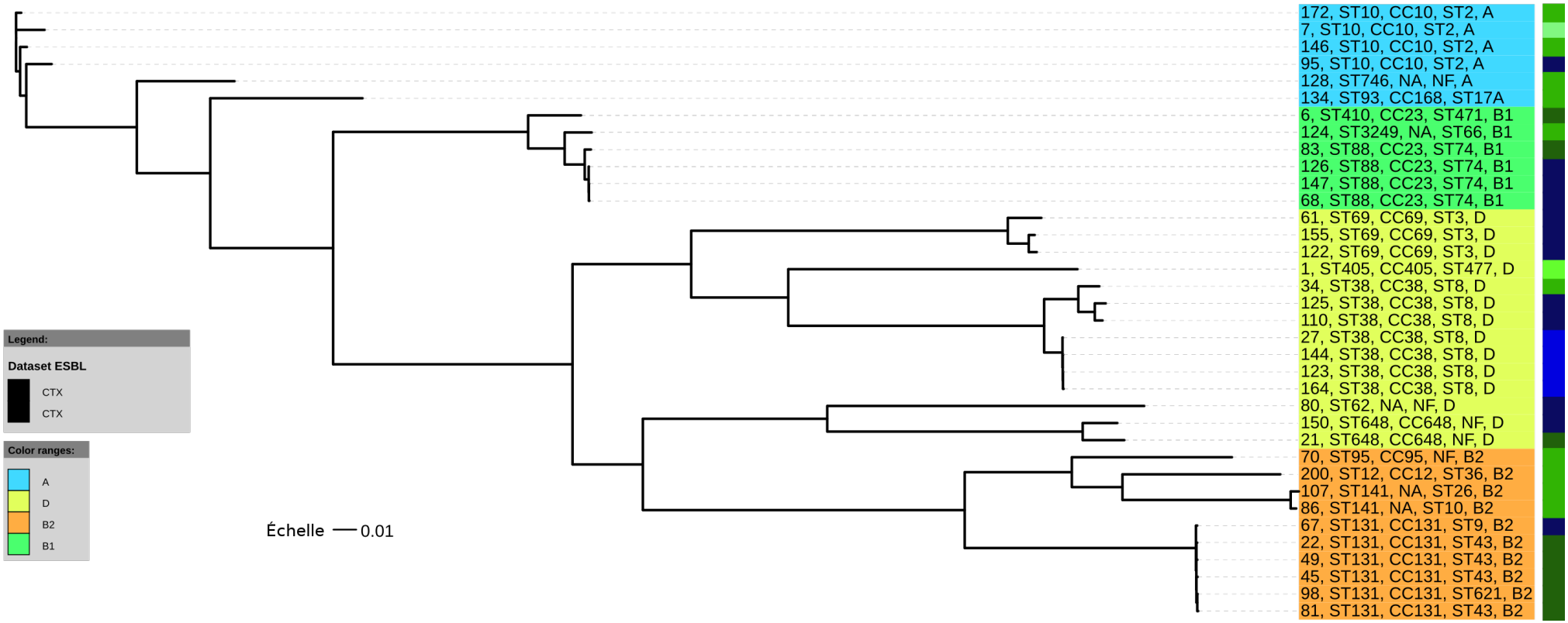


Source personnelle

V. Résultats

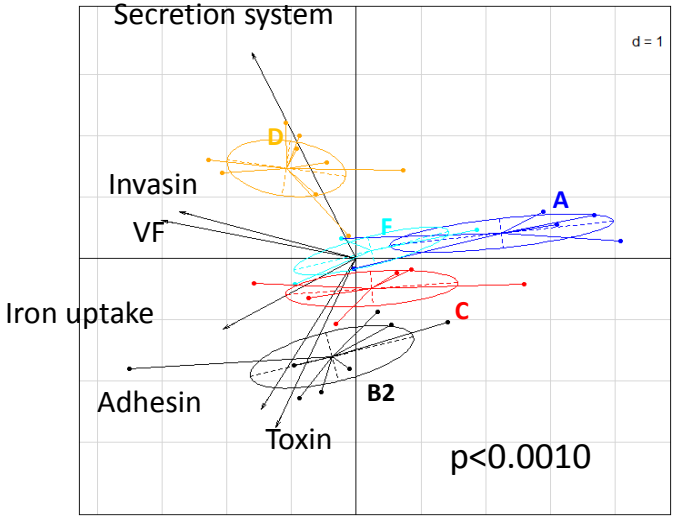
L'ensemble des résultats des analyses sont synthétisés

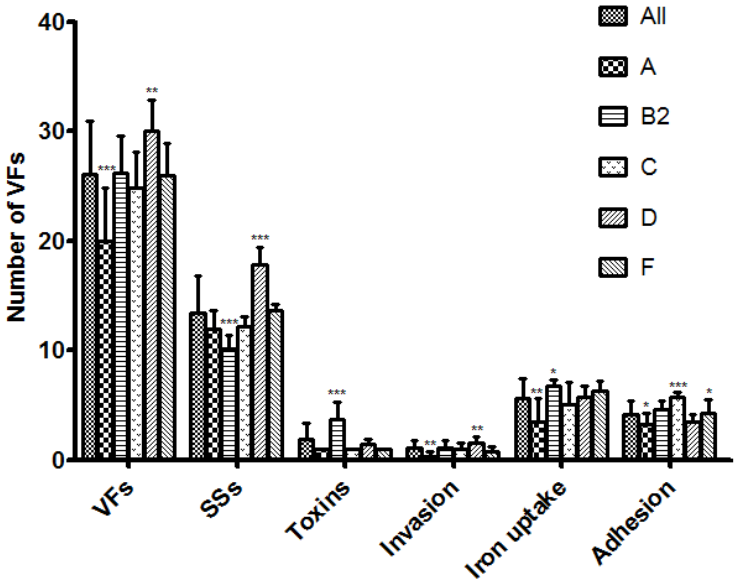
Phylogénie



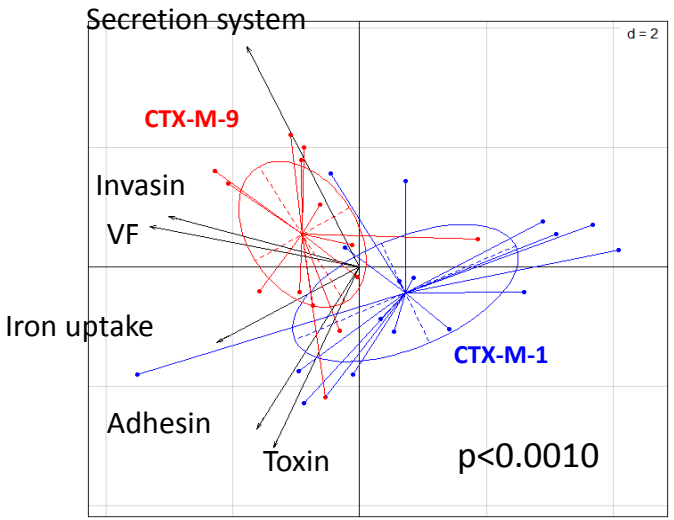
V. Résultats

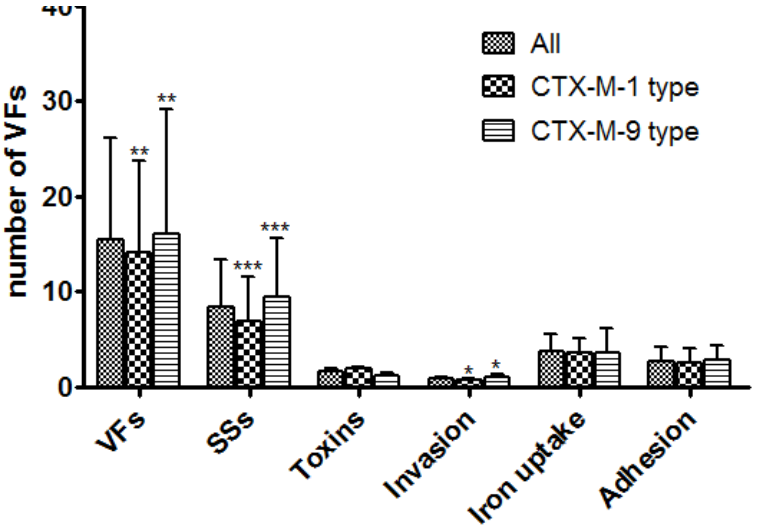
E. coli phylogroup





CTX-M group





Conclusion

2 scripts en python ; environ 1350 lignes.

Trimming : Trimmomatic / Trim Galore

Détection et **typage** : SRST2

Assemblage : Spades ; **Phylogénie** : Parsnp

Caractérisation de souches bactériennes multirésistantes

Gain de ressources et de temps en terme d'analyses

Vérification manuelle nécessaire dans certain cas

Remerciements

En remerciant M. Bonnet et son équipe.

Merci de votre attention.

Des questions ?