

TP1

Objectif

Stocker et Trier un jeu de données de séquençage

Données

Format fasta

Taille des lectures : 100 bp

TP1

reads.fasta



Lectures au format fasta



LECTURE + FILTRE + TRI

*Seuls les lectures qui apparaissent
au moins 2 fois sont conservé
Une seule instance est conservée*



Sort_reads.txt

```
ATGAGACACAGATAGAGACACAGATATAGAGGACACAGGGATA
CACAGGATATACCACAGAGGATTTAGGAGAGACACAGGAGGAA
CTGGAGAGACAGAGGAAAAGATACACGAGGATTAGGACACGAG
TCTATTATATATAGGCCAGATTAGGACACAGGGAGGAGAGAAT
```

Lectures triées par ordre alphabétique

TP1

Matériel

- Programme generateFastaFile
 - Usage : generateFastaFile <genome_size> <coverage>
 - Génère un fichier **reads.fasta**

Mbp



TP1 : déroulement

1. Charger et compiler le programme **generateFastaReads**
`gcc -O3 -o generateFastaReads generateFastaReads.c`
2. Ecrire le programme tp1
en entrée : `reads.fasta` en sortie : `sort_reads.txt`
Format de `sort_reads.txt` = lectures sans commentaires
3. Tester le programme sur des petits jeux de données
4. Mesurer les performances du programme avec les paramètres suivants
 - Taille génome : 1 Mbp, 3 Mbp, 10 Mbp
 - Couverture : 10, 30, 100

9 situations

TP1

A rendre avant le **15/11/2019**

1. Le code source avec (beaucoup de) commentaires
2. Explications sur la stratégie employée
3. Rapport sur le comportement du programme en fonction de la taille des données
 - Temps d'exécution
 - Le nombre de lectures sélectionnées

Envoyez à lavenier@irisa.fr :

- Code source (en C)
- Le rapport (format pdf)