LES BASIQUES AUTOUR DES NGS

MODULE 5 - DUBII 2018

Olivier Rué & Olivier Kirsh

PROGRAMME

- Objectifs de ce cours
- Jeu de données
- Environnement de travail
- Introduction
- Les principales ressources en ligne
- Les données issues du séquençage
- Le format FASTQ
- Le génome de référence
- Le format FASTA
- Le mapping
- Les annotations du génome de référence
- Le format GFF
- Autres indispensables

OBJECTIFS DE CE COURS

1. Connaître les ressources bioinformatiques en ligne pour traiter des données NGS
2. Savoir récupérer en local ce type de ressources
3. Découvrir les principaux formats de fichiers

JEU DE DONNÉES

Bacterial regulons in Escherichia coli with ChIP-seq and RNA-seq

Myers KS, Park DM, Beauchene NA, Kiley PJ. Defining bacterial regulons using ChIP-seq. Methods. 2015 Sep 15;86:80-8. doi: 10.1016/j.ymeth.2015.05.022. Epub 2015 May 29. Review.

PubMed PMID: 26032817

ENVIRONNEMENT DE TRAVAIL

Pour ce cours, vous travaillerez sur la machine locale qui vous est mise à disposition

Vous aurez besoin du terminal et d'un navigateur web

INTRODUCTION

Les essentiels d'un *pipeline* d'analyse de données NGS

- Ressources (données issues du séquençage, génome de référence, annotations, base de données...)
  - Centralisées
  - Pas toujours pour les organismes non modèles
  - Une liste exhaustive est disponible entre autres ici
- Outils
  - Certains incontournables au NGS
  - Dépendants de la thématique et/ou de l'analyse souhaitée
  - Beaucoup sont recensés ici : labworm.com
- Puissance de calcul (mémoire/coeurs) !
  - clusters de calculs ouverts à la communauté *[cf cours Unix 2]*
- Des connaissances en Unix (au moins)

LES PRINCIPALES RESSOURCES EN LIGNE

Vous aurez forcément besoin d'utiliser des ressources produites et expertisées par d'autres. Il existe une multitude de sites recensant des ressources bioinformatiques. Les incontournables et les plus généralistes sont :

- Nucleotide Sequence Databases :
  - NCBI (National Center for Biotechnology Information)
  - ENA (European Nucleotide Archive)
- Genome databases :
  - Ensembl (vertebrate genomes)
  - UCSC (Genome browser intégré)

Une classification plus exhaustive (séquences protéiques, domaines protéiques, structures protéiques 3D, voies métaboliques...) est disponible entre autres ici

NCBI

Le NCBI héberge et met en relation de nombreuses bases de données :

- Pubmed : publications et citations
- SRA : Sequence Read Archive (données de sortie de séquençage ou alignées vs référence)
- Genbank : collection de séquences ADN annotées publiques
- RefSeq : collection de séquences de génomes, de transcripts et de protéines, non redondantes
- dbSNP : collection de variants, microsatellites... humains
- GEO : collection de données liées à l'expression des gènes
- ...

Tout (ou presque) est interconnecté

ENA

L'ENA est l'équivalent européen du NCBI, hébergé à l'EMBL-EBI :

- Données de séquençage
- Assemblages de génomes annotés
- Annotations fonctionnelles de génomes
- ...

LES ACCESSION NUMBERS

Les ressources sont décrites par un Accession Number. C'est leur numéro unique qui permet de les retrouver.

La nomenclature change en fonction des ressources (NCBI, ENA...) mais le préfixe reste fixe :

- 'DR' pour DDBJ Sequence Read Archive (DRA)
- 'EGA' pour European Genome-phenome Archive (EGA)
- 'ER' pour ENA Read
- 'SR' pour NCBI Sequence Read Archive (SRA)

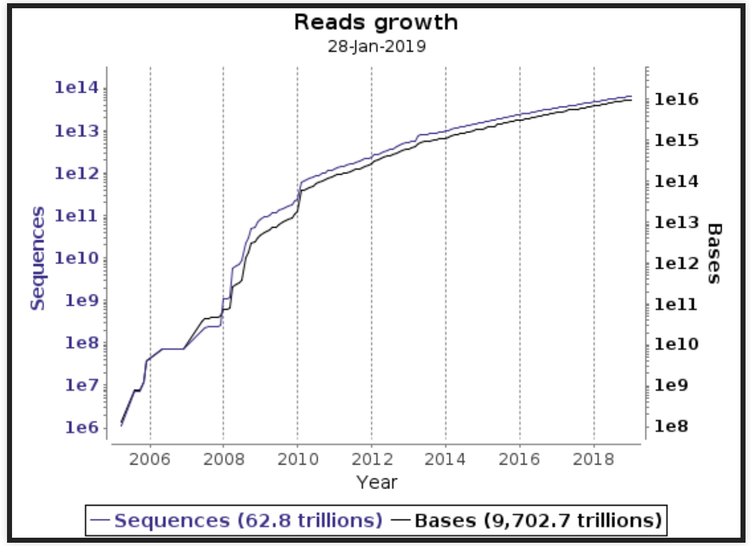
Par contre un même Accession Number peut être stocké sur plusieurs ressources !

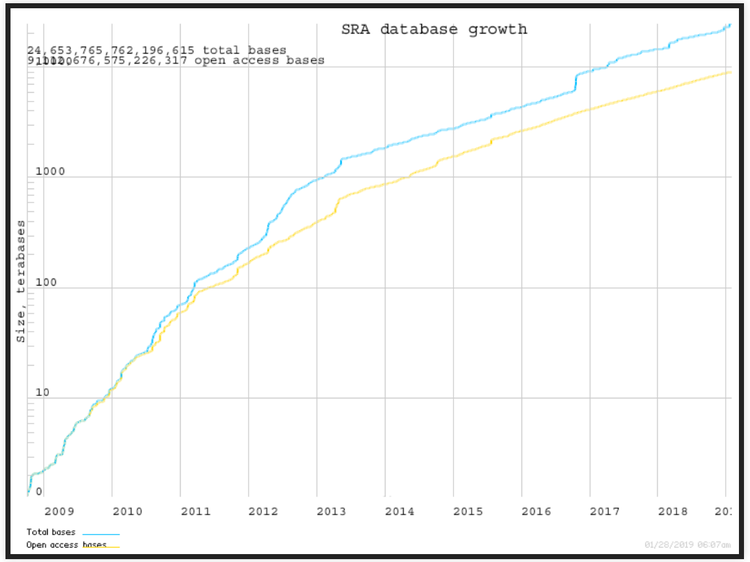
LES DONNÉES ISSUES DU SÉQUENÇAGE

LES PRINCIPALES RESSOURCES PUBLIQUES

La grande majorité des données brutes sont stockées soit au NCBI (Etats-Unis - NIH), soit à l'EBI (Angleterre - EMBL).

Depuis quelques temps, les reviewers exigent que vos données soient déposées et accessibles à la communauté.





EXERCICE N°1 : RETROUVER OÙ ONT ÉTÉ DÉPOSÉES LES DONNÉES DE SÉQUENÇAGE DE LA PUBLICATION

Une section est souvent dédiée à la mise à disposition des données en fin d'article

1. Dans quelle base de données ont été déposées les données brutes ?
2. Quel est leur numéro d'accession ?
3. Allez sur le site en question et entrez le numéro d'accession dans la barre de recherche. Combien d'échantillons sont disponibles ?
4. Cliquez sur la sous-série GSE41187 (analyse Chip-Seq) puis suivez le lien vers SRA
5. Combien de fichiers de séquences brutes pouvez-vous récupérer ?
6. Que pouvez-vous dire sur le séquençage du 1er run ? (SRX220453)
7. Pouvez-vous le télécharger via l'interface ?
8. Quelle solution vous est proposée ?

CORRECTION N°1

1. Dans quelle base de données ont été déposées les données brutes ? *GEO - NCBI*
2. Quel est leur numéro d'accession ? *GSE41195*
3. Allez sur le site en question et entrez le numéro d'accession dans la barre de recherche. Combien d'échantillons sont disponibles ? *48*
4. Cliquez sur la sous-série GSE41187 (analyse Chip-Seq) puis suivez le lien vers SRA
5. Combien de fichiers de séquences brutes pouvez-vous récupérer ? *9*
6. Que pouvez-vous dire sur le séquençage du 1er run ? (SRX220453) *Chip-Seq, Illumina GAIIx, reads sinlge-end, 878,466 reads*
7. Pouvez-vous le télécharger via l'interface ? *Non*
8. Quelle solution vous est proposée ? *SRA Toolkit*

SRA TOOLKIT

Suite d'utilitaires permettant, en ligne de commande, de récupérer des données du NCBI. L'outil est généralement accessible sur les clusters de bioinformatique. La documentation est indispensable pour l'utiliser (au début du moins).

Les principales commandes sont :

- prefetech : pour télécharger les fichiers sra
- fastq-dump : pour télécharger les fichiers au format fastq ou fasta

EXERCICE N°2 : UTILISER SRA TOOLKIT

1. Affichez l'aide de la commande fastq-dump dans le terminal
2. Télécharger le run SRR653522
3. Affichez les dix premières lignes du fichier
4. *Bonus : télécharger une liste de fichiers SRR*

CORRECTION N°2 : UTILISER SRA TOOLKIT

```
fastq-dump -h # Afficher l'aide
[orue@migale tmp]$ fastq-dump -h

Usage:
  fastq-dump [options] <path> [<path>...]
  fastq-dump [options] <accession>

</accession></path></path>
```

```
fastq-dump SRR653522 # Télécharger le run SRR653522
```

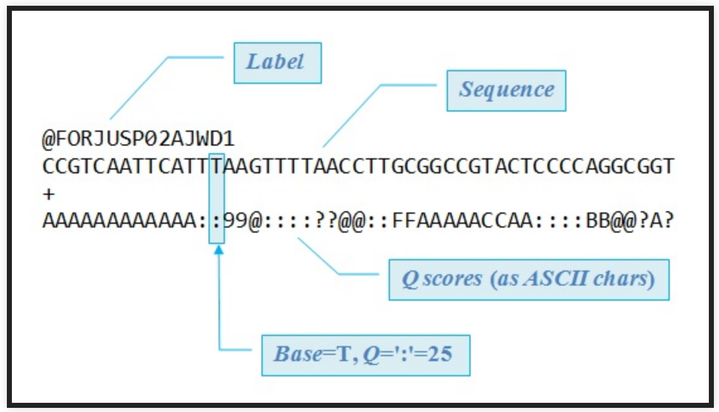
```
head SRR653522.fastq
@SRR653522.1 ILLUMINA05:2:1:0:2016 length=35
NCGTCGACCATTTCCAGCTCGCCGCCAACCGGAAC
+SRR653522.1 ILLUMINA05:2:1:0:2016 length=35
&<9<<<<<<<<<<<<9<<<<<<<<<2<<<<<<<<
@SRR653522.2 ILLUMINA05:2:1:0:69 length=35
NGGCTAATGGCATCTTTATGATTGTTAATAATAGA
+SRR653522.2 ILLUMINA05:2:1:0:69 length=35
&=AA=<2====<=??A=,==:=====9==.<=2=.
@SRR653522.3 ILLUMINA05:2:1:0:418 length=35
NAACGTACAAATACTGGGGTGATTTTGTTCGGATC
```

```
head SRR_accessions.txt # Il faut remplir le fichier avec les Accession Numbers trouvés sur la page du projet (SR
SRR576933
SRR576934
SRR576935
SRR576936
SRR576937
SRR576938
SRR653520
SRR653521
SRR653522

cat SRR_accessions.txt | xargs fastq-dump # Télécharger une liste de SRR
```

LE FORMAT FASTQ

C'est le format utilisé pour stocker des fragments séquencés (lectures ou reads)



- Fichier texte (.fastq)
- 4 lignes par séquence
- Hautement compressible (.fastq.gz)

```
gzip file.fastq # To compress
gzip -d file.fastq.gz # To uncompress
```

ENCODAGE DE LA QUALITÉ DE LA BASE SÉQUENCÉE

Un score Phred est attribué par le séquenceur à chaque base séquencée, quelle que soit la technologie utilisée. Elle représente la confiance accordée à la base.

Elle est encodée en ASCII pour tenir sur un seul caractère.

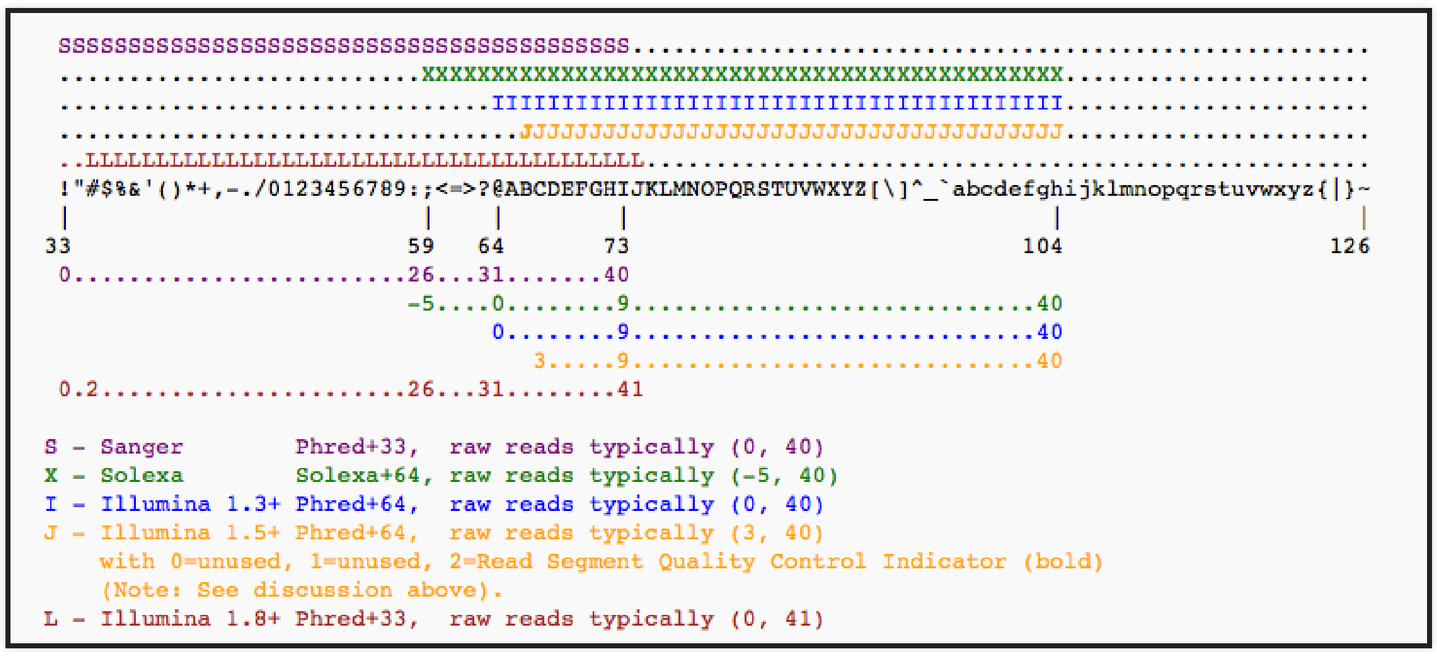
Les scores de qualité varient selon les technologies de séquençage ! Attention !

Quality score interpretation

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Materials from Wikepedia



EXERCICE N°3 : EXTRAIRE INFORMATIONS DU FASTQ

1. Combien de séquences y a-t-il dans le fichier SRR5344684_1.fastq ?

CORRECTION N°3 : EXTRAIRE INFORMATIONS DU FASTQ

```
wc -l SRR5344684_1.fastq # Puis diviser par 4
89629572 SRR5344684_1.fastq
cat SRR5344684_1.fastq | echo $((`wc -l`/4))
22407393
```

CONTRÔLE QUALITÉ DES FICHIERS FASTQ

FastQC est la référence pour analyser la qualité des lectures au format FASTQ.

MultiQC permet de synthétiser tous les rapports en un seul rapport.

Vous aurez accès à :

- Nombre de lectures
- Taille des lectures
- Qualité moyenne de la base séquencée à chaque position
- Présence d'adaptateurs de séquençage
- GC content
- ...

```
fastqc file.fastq.gz # Run fastqc on one compressed FASTQ file
# Open .html in a web browser
```

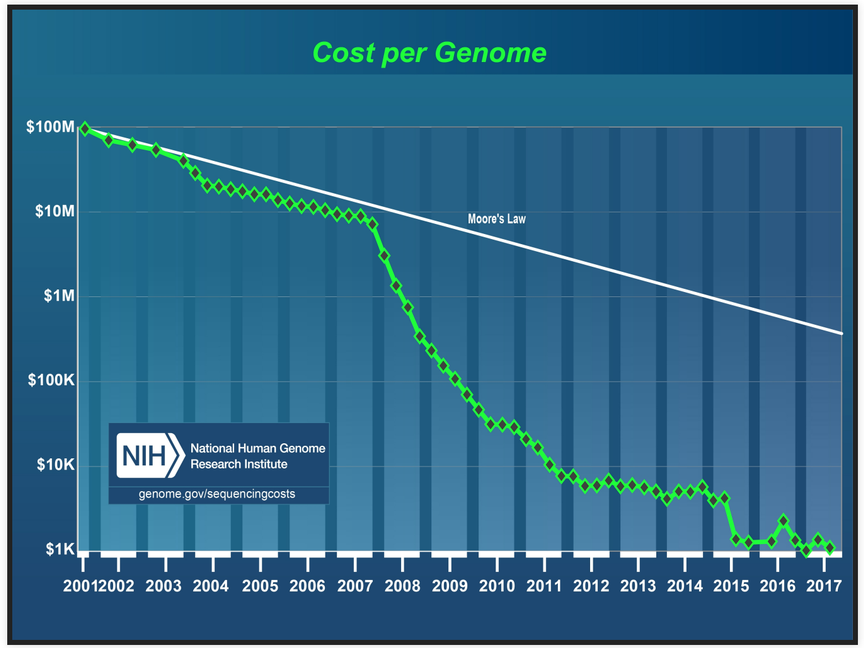
Vous le verrez en détail lors du cours Production de données #2

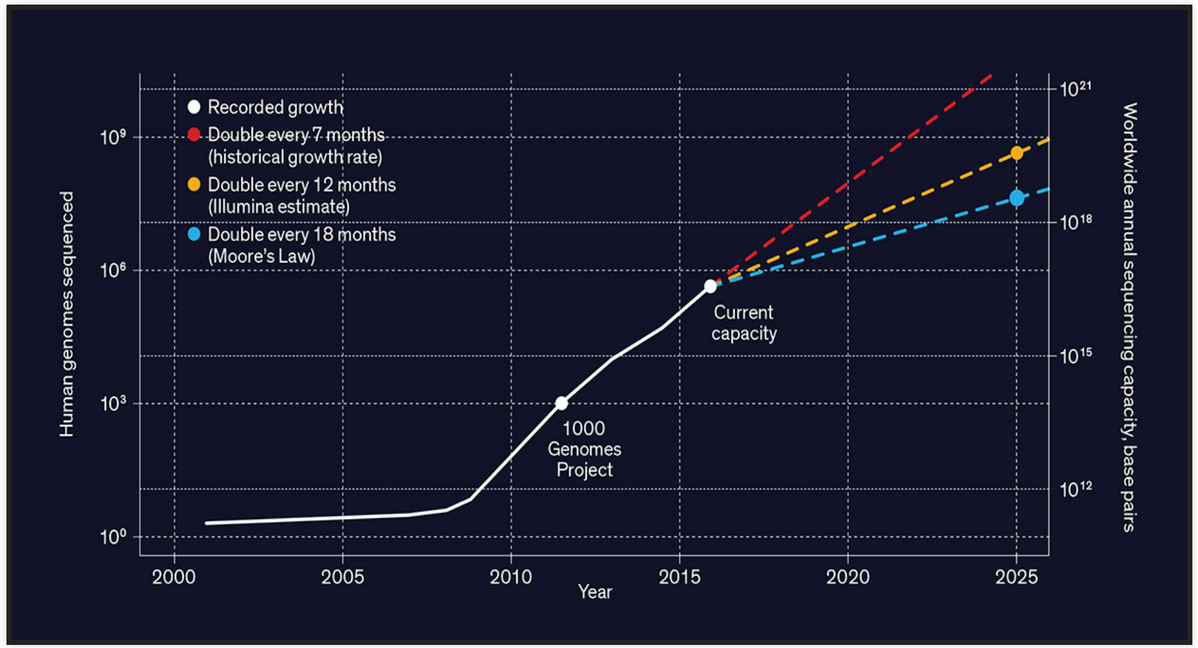
LE GÉNOME DE RÉFÉRENCE

La plupart des pipelines bioinformatiques reposent sur la présence d'une référence. Soit elle existe, soit il faut en construire une.

Attention, il existe des génomes complètement séquencés et très bien annotés (les organismes modèles), d'autres sont une succession de fragments représentant une sous-partie du génome.

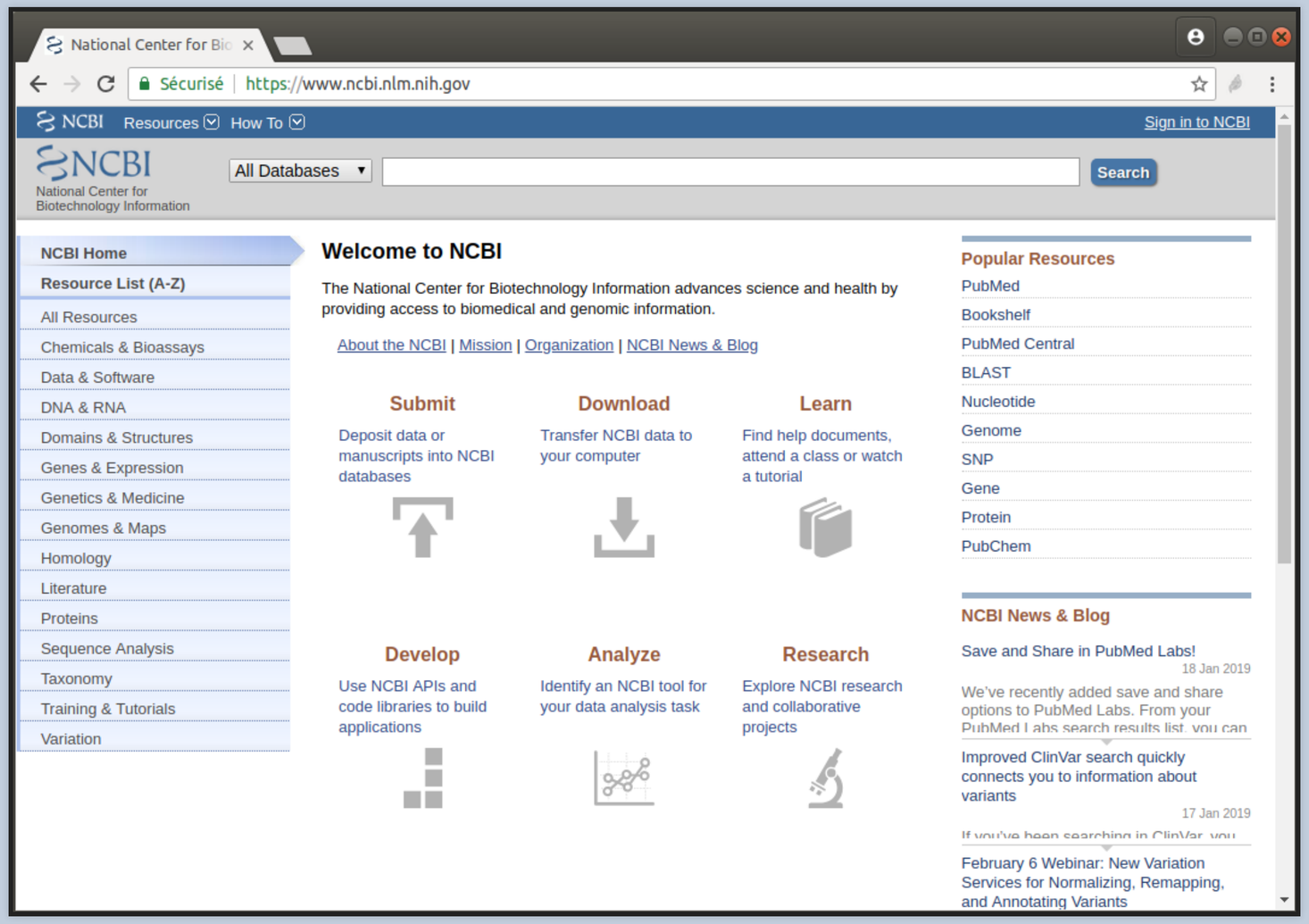
La croissance du nombre de génomes est exponentielle (exemple du génome humain) :

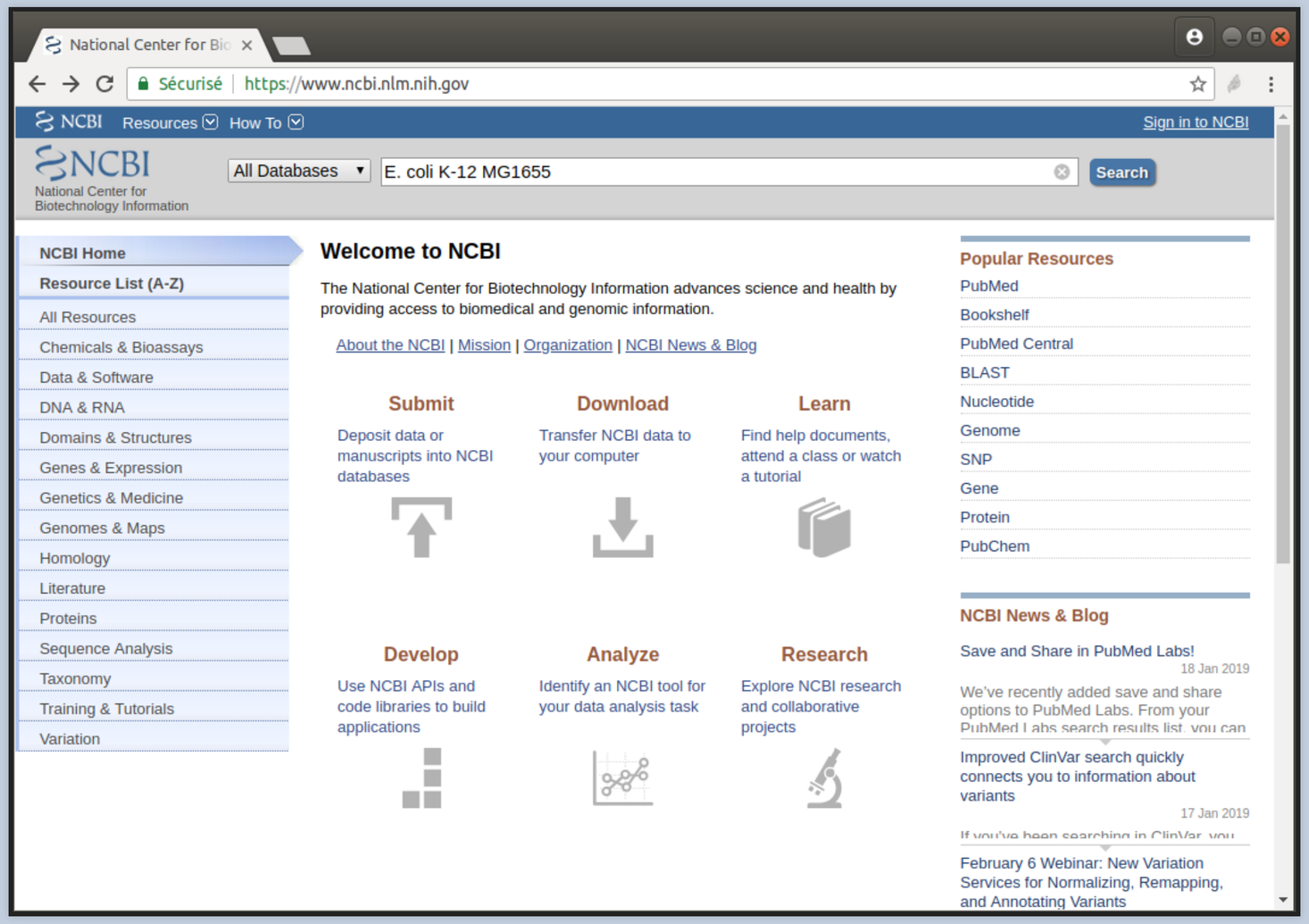


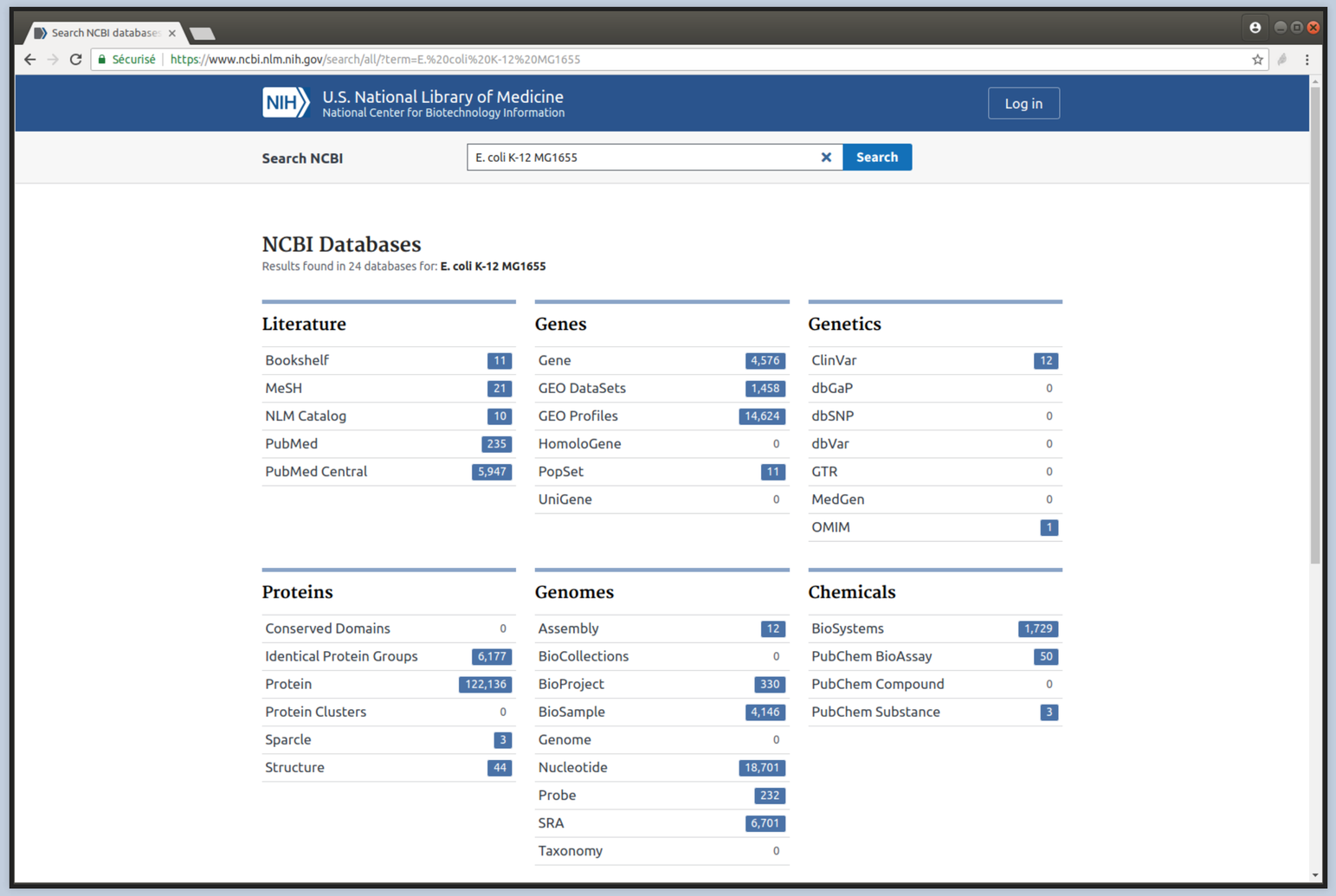


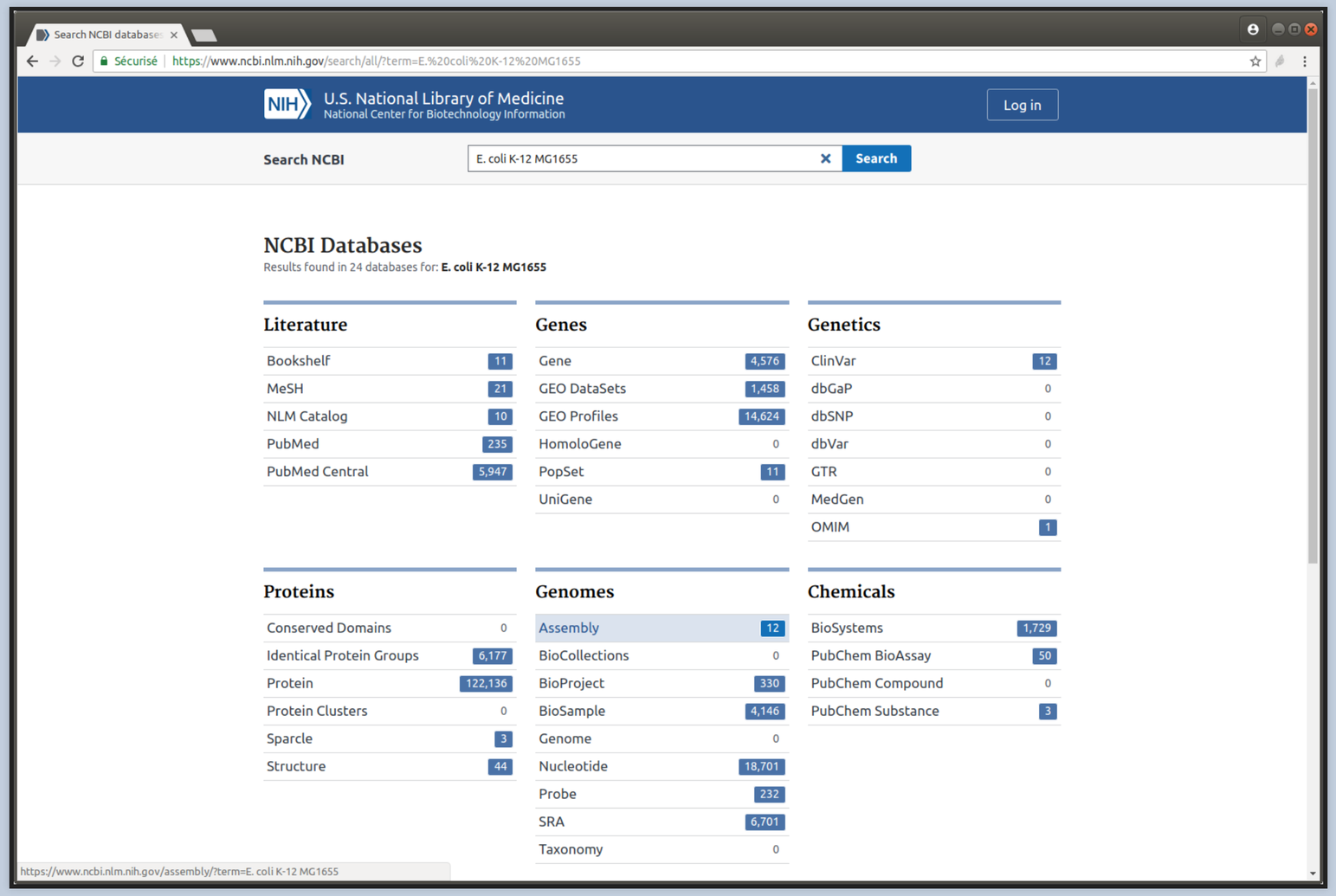
EXERCICE N°4 : RÉCUPÉRER LE GÉNOME UTILISÉ DANS LA PUBLICATION : E. COLI K-12 MG1655

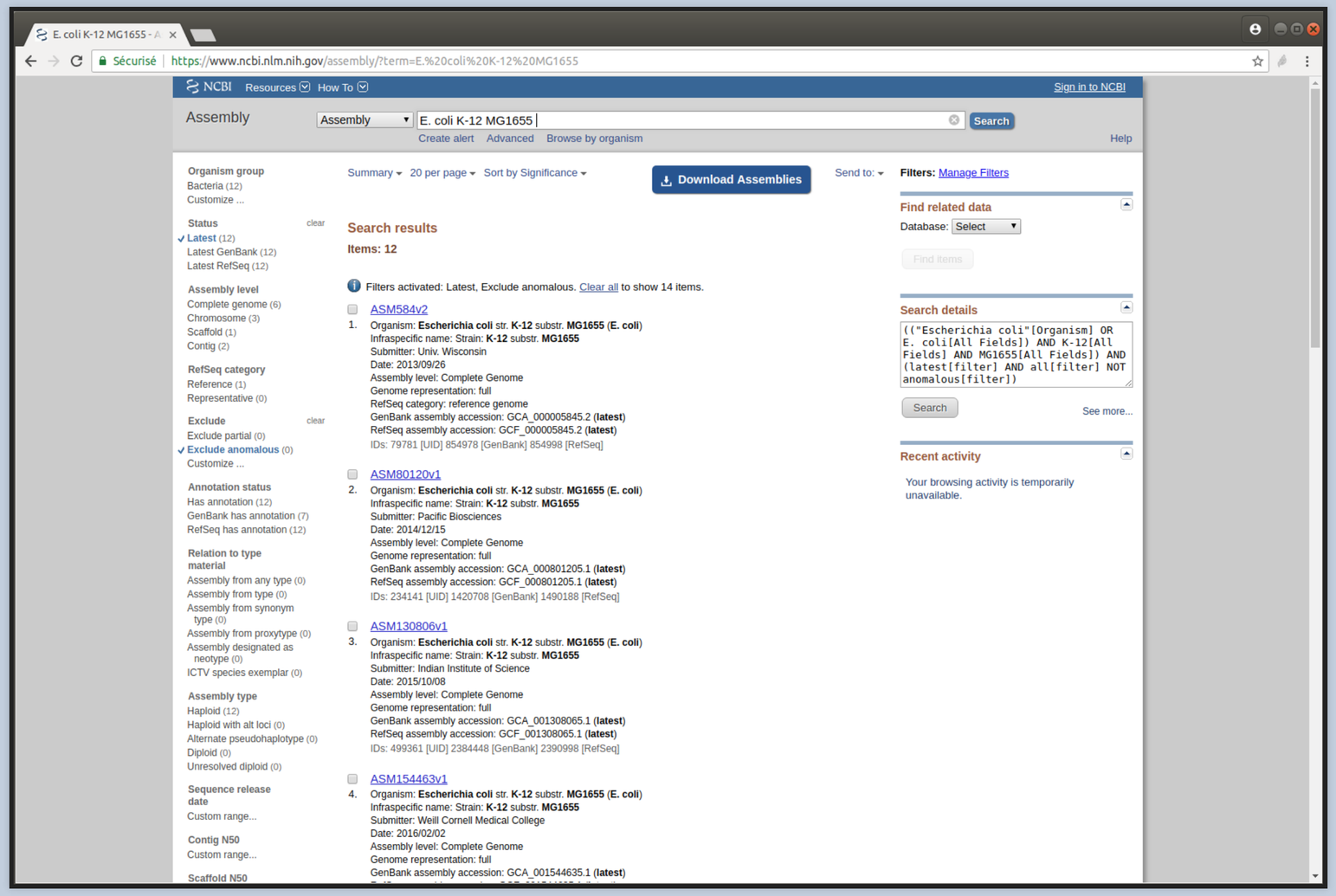
1. Recherchez cette référence sur le site du NCBI
2. Combien y a-t-il de versions de cet assemblage ?
3. Téléchargez la dernière version mise à jour
4. Affichez les dix premières lignes du fichier

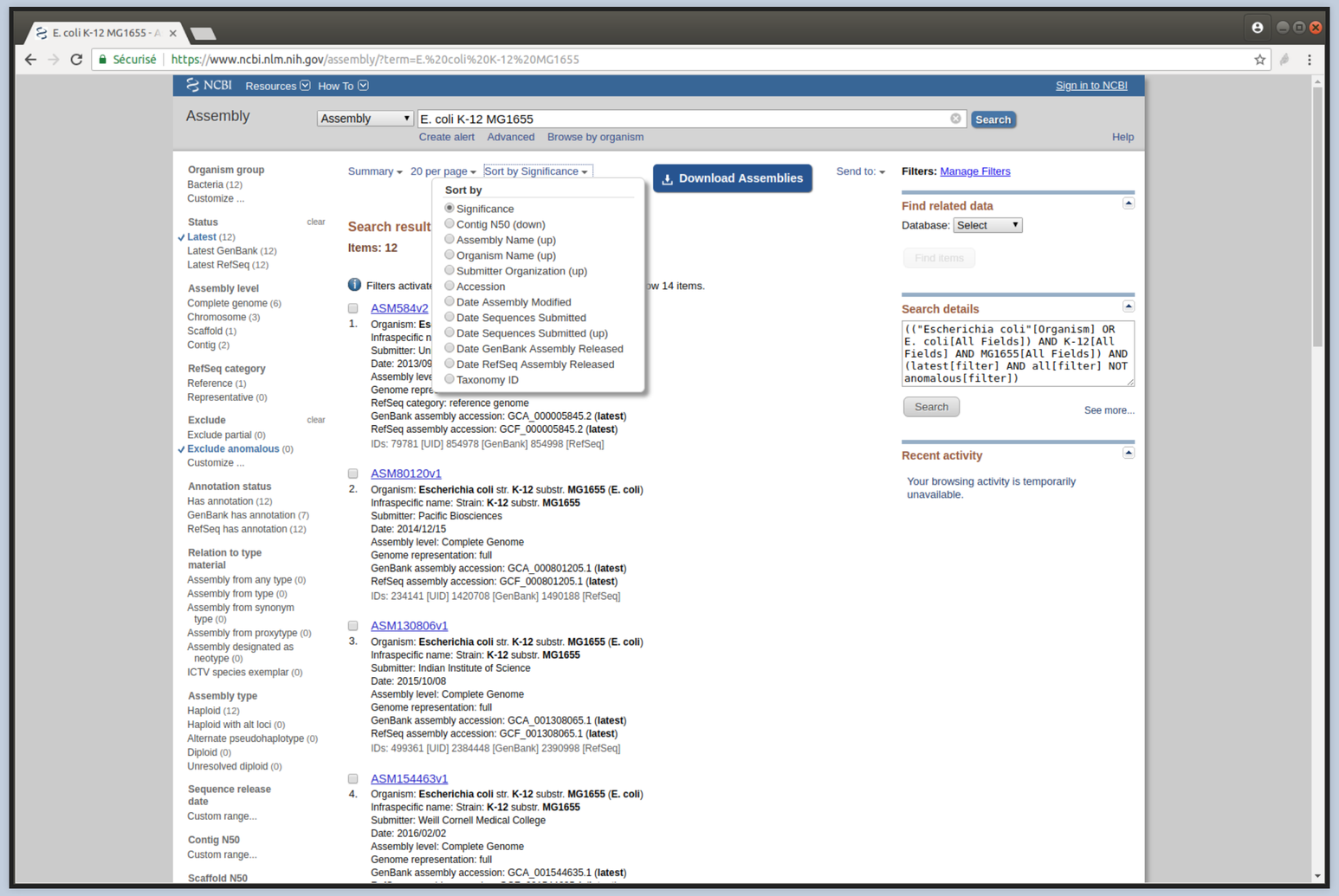

National Center for Bio
Sécurisé | https://www.ncbi.nlm.nih.gov
NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
Save and Share in PubMed Labs!
18 Jan 2019
We've recently added save and share options to PubMed Labs. From your PubMed Labs search results list, you can
Improved ClinVar search quickly connects you to information about variants
17 Jan 2019
If you've been searching in ClinVar, you
February 6 Webinar: New Variation Services for Normalizing, Remapping, and Annotating Variants

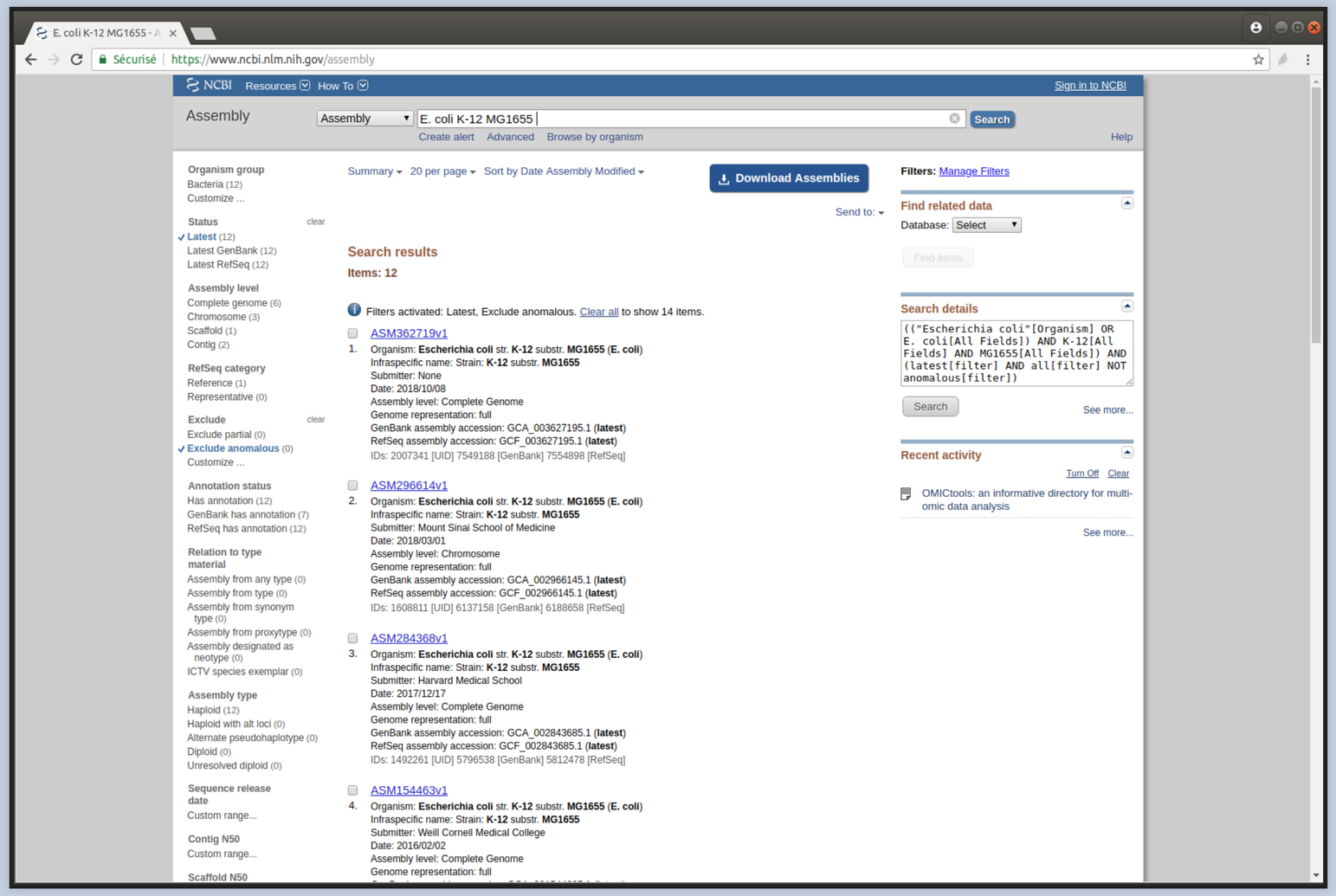

National Center for Bio
Sécurisé | https://www.ncbi.nlm.nih.gov
NCBI Resources How To Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
E. coli K-12 MG1655
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About the NCBI | Mission | Organization | NCBI News & Blog
Submit
Deposit data or manuscripts into NCBI databases
Download
Transfer NCBI data to your computer
Learn
Find help documents, attend a class or watch a tutorial
Develop
Use NCBI APIs and code libraries to build applications
Analyze
Identify an NCBI tool for your data analysis task
Research
Explore NCBI research and collaborative projects
Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem
NCBI News & Blog
Save and Share in PubMed Labs!
18 Jan 2019
We've recently added save and share options to PubMed Labs. From your
Improved ClinVar search quickly connects you to information about variants
17 Jan 2019
If you've been searching in ClinVar, you
February 6 Webinar: New Variation Services for Normalizing, Remapping, and Annotating Variants

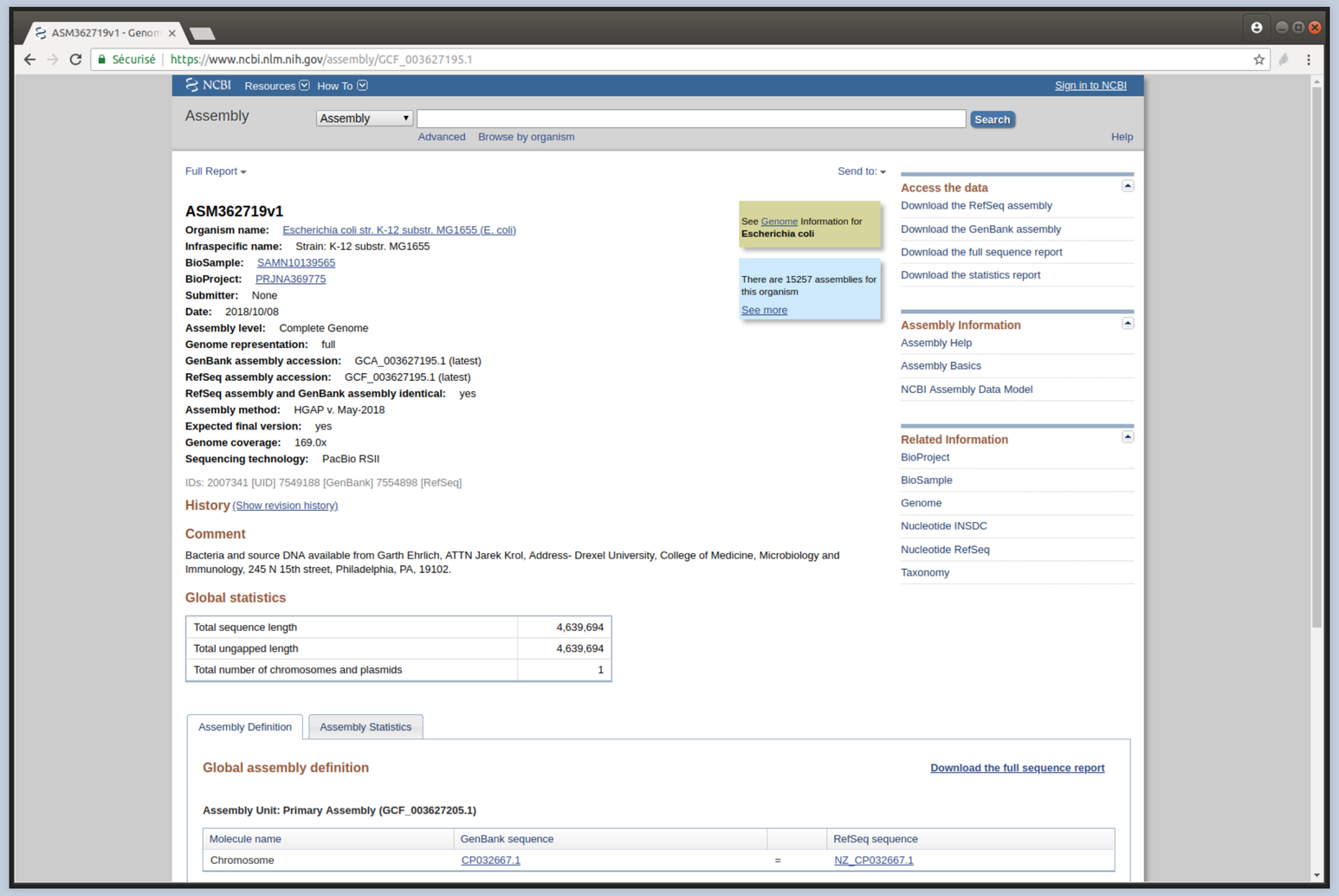

Search NCBI databases
Sécurisé | https://www.ncbi.nlm.nih.gov/search/all/?term=E.%20coli%20K-12%20MG1655
NIH U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
E. coli K-12 MG1655
Search
NCBI Databases
Results found in 24 databases for: E. coli K-12 MG1655
Literature
Bookshelf 11
MeSH 21
NLM Catalog 10
PubMed 235
PubMed Central 5,947
Genes
Gene 4,576
GEO DataSets 1,458
GEO Profiles 14,624
HomoloGene 0
PopSet 11
UniGene 0
Genetics
ClinVar 12
dbGaP 0
dbSNP 0
dbVar 0
GTR 0
MedGen 0
OMIM 1
Proteins
Conserved Domains 0
Identical Protein Groups 6,177
Protein 122,136
Protein Clusters 0
Sparcle 3
Structure 44
Genomes
Assembly 12
BioCollections 0
BioProject 330
BioSample 4,146
Genome 0
Nucleotide 18,701
Probe 232
SRA 6,701
Taxonomy 0
Chemicals
BioSystems 1,729
PubChem BioAssay 50
PubChem Compound 0
PubChem Substance 3

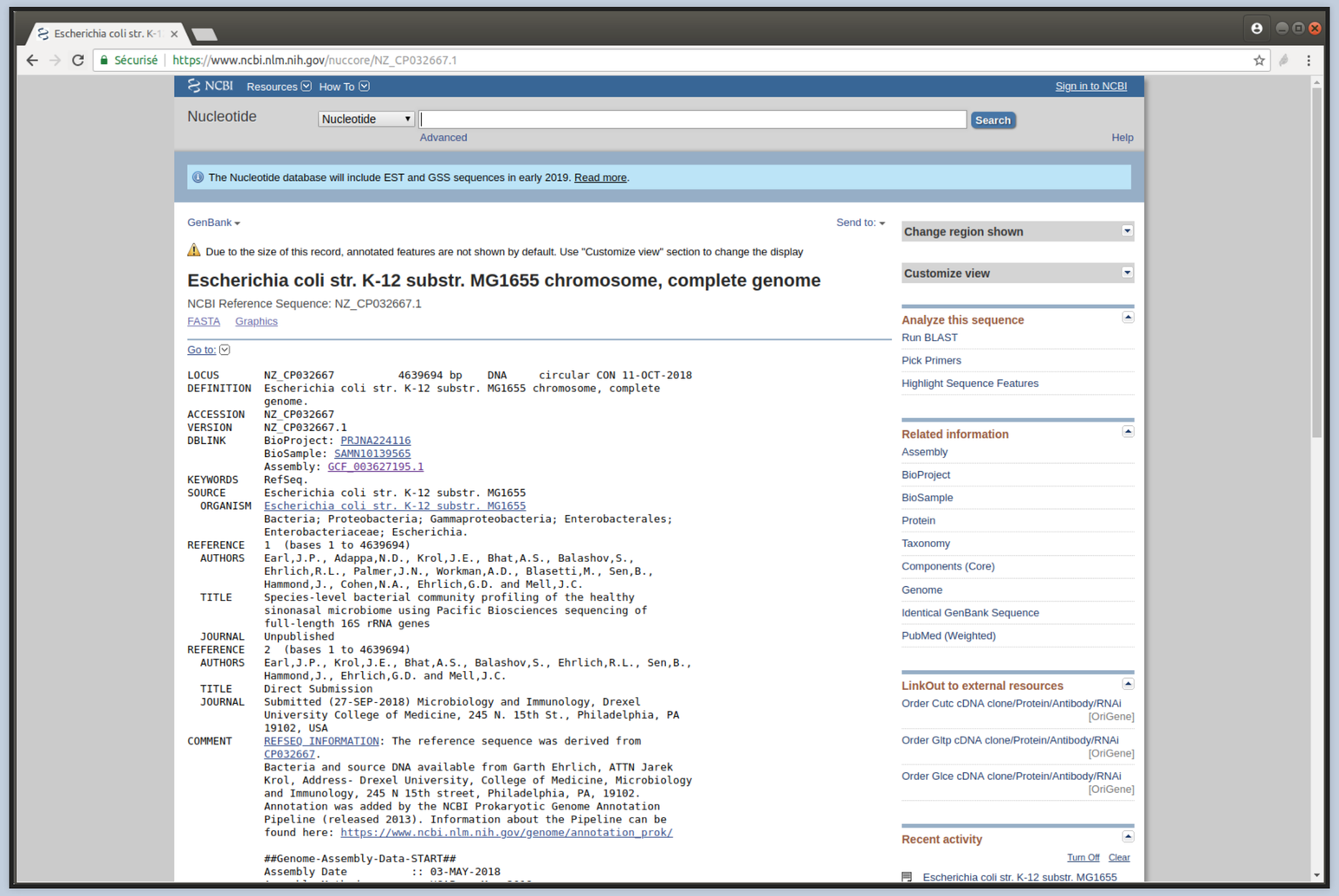

Search NCBI databases
Sécurisé | https://www.ncbi.nlm.nih.gov/search/all/?term=E.%20coli%20K-12%20MG1655
NIH U.S. National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
E. coli K-12 MG1655
Search
NCBI Databases
Results found in 24 databases for: E. coli K-12 MG1655
Literature
Bookshelf 11
MeSH 21
NLM Catalog 10
PubMed 235
PubMed Central 5,947
Genes
Gene 4,576
GEO DataSets 1,458
GEO Profiles 14,624
HomoloGene 0
PopSet 11
UniGene 0
Genetics
ClinVar 12
dbGaP 0
dbSNP 0
dbVar 0
GTR 0
MedGen 0
OMIM 1
Proteins
Conserved Domains 0
Identical Protein Groups 6,177
Protein 122,136
Protein Clusters 0
Sparcle 3
Structure 44
Genomes
Assembly 12
BioCollections 0
BioProject 330
BioSample 4,146
Genome 0
Nucleotide 18,701
Probe 232
SRA 6,701
Taxonomy 0
Chemicals
BioSystems 1,729
PubChem BioAssay 50
PubChem Compound 0
PubChem Substance 3
https://www.ncbi.nlm.nih.gov/assembly/?term=E. coli K-12 MG1655

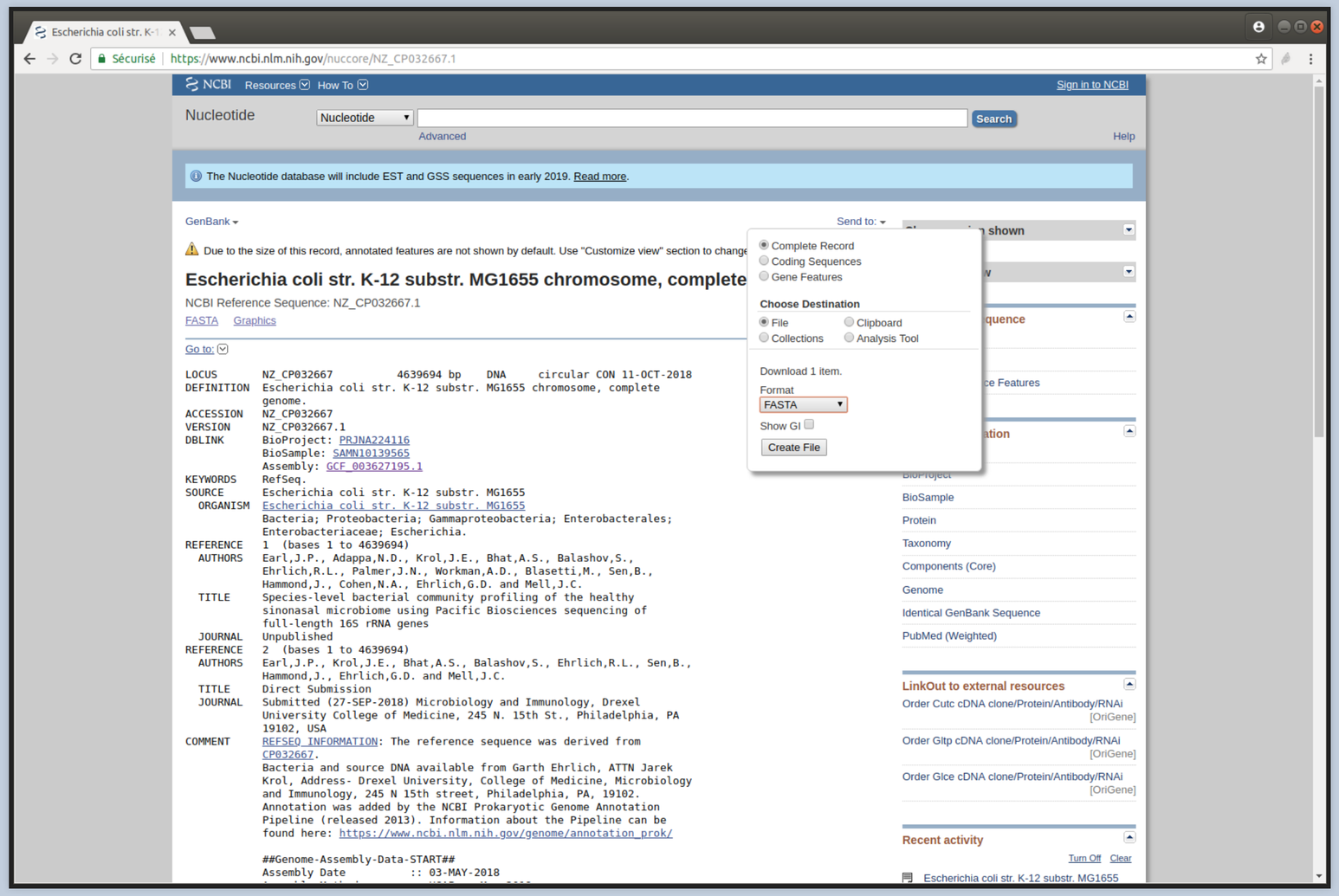

E. coli K-12 MG1655 - A
Sécurisé | https://www.ncbi.nlm.nih.gov/assembly/?term=E.%20coli%20K-12%20MG1655
NCBI Resources How To
Sign in to NCBI
Assembly
Assembly
E. coli K-12 MG1655
Search
Create alert Advanced Browse by organism
Help
Organism group
Bacteria (12)
Customize ...
Status clear
Latest (12)
Latest GenBank (12)
Latest RefSeq (12)
Assembly level
Complete genome (6)
Chromosome (3)
Scaffold (1)
Contig (2)
RefSeq category
Reference (1)
Representative (0)
Exclude clear
Exclude partial (0)
Exclude anomalous (0)
Customize ...
Annotation status
Has annotation (12)
GenBank has annotation (7)
RefSeq has annotation (12)
Relation to type material
Assembly from any type (0)
Assembly from type (0)
Assembly from synonym type (0)
Assembly from proxytype (0)
Assembly designated as neotype (0)
ICTV species exemplar (0)
Assembly type
Haploid (12)
Haploid with alt loci (0)
Alternate pseudohaplotype (0)
Diploid (0)
Unresolved diploid (0)
Sequence release date
Custom range...
Contig N50
Custom range...
Scaffold N50
Summary 20 per page Sort by Significance
Download Assemblies
Send to:
Search results
Items: 12
Filters activated: Latest, Exclude anomalous. Clear all to show 14 items.
ASM584v2
1. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Univ. Wisconsin
Date: 2013/09/26
Assembly level: Complete Genome
Genome representation: full
RefSeq category: reference genome
GenBank assembly accession: GCA_000005845.2 (latest)
RefSeq assembly accession: GCF_000005845.2 (latest)
IDs: 79781 [UID] 854978 [GenBank] 854998 [RefSeq]
ASM80120v1
2. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Pacific Biosciences
Date: 2014/12/15
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_000801205.1 (latest)
RefSeq assembly accession: GCF_000801205.1 (latest)
IDs: 234141 [UID] 1420708 [GenBank] 1490188 [RefSeq]
ASM130806v1
3. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Indian Institute of Science
Date: 2015/10/08
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_001308065.1 (latest)
RefSeq assembly accession: GCF_001308065.1 (latest)
IDs: 499361 [UID] 2384448 [GenBank] 2390998 [RefSeq]
ASM154463v1
4. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Weill Cornell Medical College
Date: 2016/02/02
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_001544635.1 (latest)
Filters: Manage Filters
Find related data
Database: Select
Find items
Search details
((("Escherichia coli"[Organism] OR E. coli[All Fields]) AND K-12[All Fields] AND MG1655[All Fields]) AND (latest[filter] AND all[filter] NOT anomalous[filter])
Search
See more...
Recent activity
Your browsing activity is temporarily unavailable.


E. coli K-12 MG1655 - A
Sécurisé | https://www.ncbi.nlm.nih.gov/assembly/?term=E.%20coli%20K-12%20MG1655
NCBI Resources How To
Sign in to NCBI
Assembly
Assembly
E. coli K-12 MG1655
Search
Create alert Advanced Browse by organism
Help
Organism group
Bacteria (12)
Customize ...
Status clear
Latest (12)
Latest GenBank (12)
Latest RefSeq (12)
Assembly level
Complete genome (6)
Chromosome (3)
Scaffold (1)
Contig (2)
RefSeq category
Reference (1)
Representative (0)
Exclude clear
Exclude partial (0)
Exclude anomalous (0)
Customize ...
Annotation status
Has annotation (12)
GenBank has annotation (7)
RefSeq has annotation (12)
Relation to type material
Assembly from any type (0)
Assembly from type (0)
Assembly from synonym type (0)
Assembly from proxytype (0)
Assembly designated as neotype (0)
ICTV species exemplar (0)
Assembly type
Haploid (12)
Haploid with alt loci (0)
Alternate pseudohaplotype (0)
Diploid (0)
Unresolved diploid (0)
Sequence release date
Custom range...
Contig N50
Custom range...
Scaffold N50
Summary 20 per page Sort by Significance
Download Assemblies
Send to:
Sort by
Significance
Contig N50 (down)
Assembly Name (up)
Organism Name (up)
Submitter Organization (up)
Accession
Date Assembly Modified
Date Sequences Submitted
Date Sequences Submitted (up)
Date GenBank Assembly Released
Date RefSeq Assembly Released
Taxonomy ID
Search result
Items: 12
Filters activate
ow 14 items.
ASM584v2
1. Organism: Es
Infraspecific n
Submitter: Un
Date: 2013/09
Assembly leve
Genome repre
RefSeq category: reference genome
GenBank assembly accession: GCA_000005845.2 (latest)
RefSeq assembly accession: GCF_000005845.2 (latest)
IDs: 79781 [UID] 854978 [GenBank] 854998 [RefSeq]
ASM80120v1
2. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Pacific Biosciences
Date: 2014/12/15
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_000801205.1 (latest)
RefSeq assembly accession: GCF_000801205.1 (latest)
IDs: 234141 [UID] 1420708 [GenBank] 1490188 [RefSeq]
ASM130806v1
3. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Indian Institute of Science
Date: 2015/10/08
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_001308065.1 (latest)
RefSeq assembly accession: GCF_001308065.1 (latest)
IDs: 499361 [UID] 2384448 [GenBank] 2390998 [RefSeq]
ASM154463v1
4. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Weill Cornell Medical College
Date: 2016/02/02
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_001544635.1 (latest)
Filters: Manage Filters
Find related data
Database: Select
Find items
Search details
(("Escherichia coli"[Organism] OR E. coli[All Fields]) AND K-12[All Fields] AND MG1655[All Fields]) AND (latest[filter] AND all[filter] NOT anomalous[filter])
Search
See more...
Recent activity
Your browsing activity is temporarily unavailable.


E. coli K-12 MG1655 - A
Sécurisé | https://www.ncbi.nlm.nih.gov/assembly
NCBI Resources How To
Sign in to NCBI
Assembly
Assembly
E. coli K-12 MG1655
Search
Create alert Advanced Browse by organism
Help
Organism group
Bacteria (12)
Customize ...
Status clear
Latest (12)
Latest GenBank (12)
Latest RefSeq (12)
Assembly level
Complete genome (6)
Chromosome (3)
Scaffold (1)
Contig (2)
RefSeq category
Reference (1)
Representative (0)
Exclude clear
Exclude partial (0)
Exclude anomalous (0)
Customize ...
Annotation status
Has annotation (12)
GenBank has annotation (7)
RefSeq has annotation (12)
Relation to type material
Assembly from any type (0)
Assembly from type (0)
Assembly from synonym type (0)
Assembly from proxytype (0)
Assembly designated as neotype (0)
ICTV species exemplar (0)
Assembly type
Haploid (12)
Haploid with alt loci (0)
Alternate pseudohaplotype (0)
Diploid (0)
Unresolved diploid (0)
Sequence release date
Custom range...
Contig N50
Custom range...
Scaffold N50
Summary 20 per page Sort by Date Assembly Modified
Download Assemblies
Send to:
Search results
Items: 12
Filters activated: Latest, Exclude anomalous. Clear all to show 14 items.
ASM362719v1
1. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: None
Date: 2018/10/08
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_003627195.1 (latest)
RefSeq assembly accession: GCF_003627195.1 (latest)
IDs: 2007341 [UID] 7549188 [GenBank] 7554898 [RefSeq]
ASM296614v1
2. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Mount Sinai School of Medicine
Date: 2018/03/01
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_002966145.1 (latest)
RefSeq assembly accession: GCF_002966145.1 (latest)
IDs: 1608811 [UID] 6137158 [GenBank] 6188658 [RefSeq]
ASM284368v1
3. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Harvard Medical School
Date: 2017/12/17
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_002843685.1 (latest)
RefSeq assembly accession: GCF_002843685.1 (latest)
IDs: 1492261 [UID] 5796538 [GenBank] 5812478 [RefSeq]
ASM154463v1
4. Organism: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
Submitter: Weill Cornell Medical College
Date: 2016/02/02
Assembly level: Complete Genome
Genome representation: full
Filters: Manage Filters
Find related data
Database: Select
Find items
Search details
((("Escherichia coli"[Organism] OR E. coli[All Fields]) AND K-12[All Fields] AND MG1655[All Fields]) AND (latest[filter] AND all[filter] NOT anomalous[filter])
Search
See more...
Recent activity
Turn Off Clear
OMICtools: an informative directory for multi-omic data analysis
See more...


ASM362719v1 - Genom
Sécurisé | https://www.ncbi.nlm.nih.gov/assembly/GCF_003627195.1
NCBI Resources How To Sign in to NCBI
Assembly
Assembly Search
Advanced Browse by organism Help
Full Report Send to:
ASM362719v1
Organism name: Escherichia coli str. K-12 substr. MG1655 (E. coli)
Infraspecific name: Strain: K-12 substr. MG1655
BioSample: SAMN10139565
BioProject: PRJNA369775
Submitter: None
Date: 2018/10/08
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_003627195.1 (latest)
RefSeq assembly accession: GCF_003627195.1 (latest)
RefSeq assembly and GenBank assembly identical: yes
Assembly method: HGAP v. May-2018
Expected final version: yes
Genome coverage: 169.0x
Sequencing technology: PacBio RSII
IDs: 2007341 [UID] 7549188 [GenBank] 7554898 [RefSeq]
See Genome Information for Escherichia coli
There are 15257 assemblies for this organism
See more
Access the data
Download the RefSeq assembly
Download the GenBank assembly
Download the full sequence report
Download the statistics report
Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model
Related Information
BioProject
BioSample
Genome
Nucleotide INSDC
Nucleotide RefSeq
Taxonomy
History (Show revision history)
Comment
Bacteria and source DNA available from Garth Ehrlich, ATTN Jarek Krol, Address- Drexel University, College of Medicine, Microbiology and Immunology, 245 N 15th street, Philadelphia, PA, 19102.
Global statistics
Total sequence length 4,639,694
Total ungapped length 4,639,694
Total number of chromosomes and plasmids 1
Assembly Definition
Assembly Statistics
Global assembly definition
Download the full sequence report
Assembly Unit: Primary Assembly (GCF_003627205.1)
Molecule name GenBank sequence RefSeq sequence
Chromosome CP032667.1 = NZ_CP032667.1


Escherichia coli str. K-1
Sécurisé | https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP032667.1
NCBI Resources How To Sign in to NCBI
Nucleotide
Nucleotide Search
Advanced Help
The Nucleotide database will include EST and GSS sequences in early 2019. Read more.
GenBank Send to:
Due to the size of this record, annotated features are not shown by default. Use "Customize view" section to change the display
Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome
NCBI Reference Sequence: NZ_CP032667.1
FASTA Graphics
Go to:
LOCUS NZ_CP032667 4639694 bp DNA circular CON 11-OCT-2018
DEFINITION Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome.
ACCESSION NZ_CP032667
VERSION NZ_CP032667.1
DBLINK BioProject: PRJNA224116
BioSample: SAMN10139565
Assembly: GCF_003627195.1
KEYWORDS RefSeq.
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia.
REFERENCE 1 (bases 1 to 4639694)
AUTHORS Earl,J.P., Adappa,N.D., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Palmer,J.N., Workman,A.D., Blasetti,M., Sen,B., Hammond,J., Cohen,N.A., Ehrlich,G.D. and Mell,J.C.
TITLE Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 4639694)
AUTHORS Earl,J.P., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Sen,B., Hammond,J., Ehrlich,G.D. and Mell,J.C.
TITLE Direct Submission
JOURNAL Submitted (27-SEP-2018) Microbiology and Immunology, Drexel University College of Medicine, 245 N. 15th St., Philadelphia, PA 19102, USA
COMMENT REFSEQ_INFORMATION: The reference sequence was derived from CP032667.
Bacteria and source DNA available from Garth Ehrlich, ATTN Jarek Krol, Address- Drexel University, College of Medicine, Microbiology and Immunology, 245 N 15th street, Philadelphia, PA, 19102.
Annotation was added by the NCBI Prokaryotic Genome Annotation Pipeline (released 2013). Information about the Pipeline can be found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
##Genome-Assembly-Data-START##
Assembly Date :: 03-MAY-2018
Change region shown
Customize view
Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Related information
Assembly
BioProject
BioSample
Protein
Taxonomy
Components (Core)
Genome
Identical GenBank Sequence
PubMed (Weighted)
LinkOut to external resources
Order Cutc cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Gltp cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Glce cDNA clone/Protein/Antibody/RNAi [OriGene]
Recent activity
Turn Off Clear
Escherichia coli str. K-12 substr. MG1655


Escherichia coli str. K-1
Sécurisé | https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP032667.1
NCBI Resources How To Sign in to NCBI
Nucleotide
Nucleotide Search
Advanced Help
The Nucleotide database will include EST and GSS sequences in early 2019. Read more.
GenBank
Send to:
Due to the size of this record, annotated features are not shown by default. Use "Customize view" section to change
Escherichia coli str. K-12 substr. MG1655 chromosome, complete
NCBI Reference Sequence: NZ_CP032667.1
FASTA Graphics
Go to:
LOCUS NZ_CP032667 4639694 bp DNA circular CON 11-OCT-2018
DEFINITION Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome.
ACCESSION NZ_CP032667
VERSION NZ_CP032667.1
DBLINK BioProject: PRJNA224116
BioSample: SAMN10139565
Assembly: GCF_003627195.1
KEYWORDS RefSeq.
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia.
REFERENCE 1 (bases 1 to 4639694)
AUTHORS Earl,J.P., Adappa,N.D., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Palmer,J.N., Workman,A.D., Blasetti,M., Sen,B., Hammond,J., Cohen,N.A., Ehrlich,G.D. and Mell,J.C.
TITLE Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 4639694)
AUTHORS Earl,J.P., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Sen,B., Hammond,J., Ehrlich,G.D. and Mell,J.C.
TITLE Direct Submission
JOURNAL Submitted (27-SEP-2018) Microbiology and Immunology, Drexel University College of Medicine, 245 N. 15th St., Philadelphia, PA 19102, USA
COMMENT REFSEQ INFORMATION: The reference sequence was derived from CP032667.
Bacteria and source DNA available from Garth Ehrlich, ATTN Jarek Krol, Address- Drexel University, College of Medicine, Microbiology and Immunology, 245 N 15th street, Philadelphia, PA, 19102.
Annotation was added by the NCBI Prokaryotic Genome Annotation Pipeline (released 2013). Information about the Pipeline can be found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
##Genome-Assembly-Data-START##
Assembly Date :: 03-MAY-2018
Complete Record
Coding Sequences
Gene Features
Choose Destination
File
Clipboard
Collections
Analysis Tool
Download 1 item.
Format
FASTA
Show GI
Create File
BioSample
Protein
Taxonomy
Components (Core)
Genome
Identical GenBank Sequence
PubMed (Weighted)
LinkOut to external resources
Order Cutc cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Gltp cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Glce cDNA clone/Protein/Antibody/RNAi [OriGene]
Recent activity
Turn Off Clear
Escherichia coli str. K-12 substr. MG1655

```
head sequence.fasta
>NZ_CP032667.1 Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome
GTCCCGGCCCCGGCAGAACCGACCTATCGTTCTAACGTAAACGTCAAACACACGTTTGATAACTTCGTTG
AAGGTAAATCTAACCAACTGGCGCGCGCGGCGGCTCGCCAGGTGGCGGATAACCCTGGCGGTGCCTATAA
CCCGTTGTTCCTTTATGGCGGCACGGGTCTGGGTAAAACTCACCTGCTGCATGCGGTGGGTAACGGCATT
ATGGCGCGCAAGCCGAATGCCAAAGTGGTTTATATGCACTCCGAGCGCTTTGTTCAGGACATGGTTAAAG
CCCTGCAAAACAACGCGATCGAAGAGTTTAAACGCTACTACCGTTCCGTAGATGCACTGCTGATCGACGA
TATTCAGTTTTTTGCTAATAAAGAACGATCTCAGGAAGAGTTTTTCCACACCTTCAACGCCCTGCTGGAA
GGTAATCAACAGATCATTCTCACCTCGGATCGCTATCCGAAAGAGATCAACGGCGTTGAGGATCGTTTGA
AATCCCGCTTCGGTTGGGGACTGACTGTGGCGATCGAACCGCCAGAGCTGGAAACCCGTGTGGCGATCCT
GATGAAAAAGGCCGACGAAAACGACATTCGTTTGCCGGGCGAAGTGGCGTTCTTTATCGCCAAGCGTCTA
```

LE FORMAT FASTA

C'est un fichier texte contenant des informations sur une séquence (nucléique ou protéique).

FASTA Format

- simple format used by almost all programs
- >header line with a [return] at end
- Sequence (no specific requirements for line length, characters, etc)

```
>URO1 uro1 seq  Length: 2018  November 9, 2000 11:50  Type: N  Check: 3854 ..
CGCAGAAAGAGGAGGCGCTTGCCTTCAGCTTGTGGGAAATCCCGAAGATGGCCAAAGACA
ACTCAACTGTTCGTTGCTTCCAGGGCCTGCTGATTTTTGGAAATGTGATTATTGGTTGTT
GCGGCATTGCCCTGACTGCGGAGTGCATCTTCTTTGTATCTGACCAACACAGCCTCTACC
CACTGCTTGAAGCCACCGACAACGATGACATCTATGGGGCTGCCTGGATCGGCATATTTG
TGGGCATCTGCCTCTTCTGCCTGTCTGTTCTAGGCATTGTAGGCATCATGAAGTCCAGCA
GGAAAATTCTTCTGGCGTATTTCATTCTGATGTTTATAGTATATGCCTTTGAAGTGGCAT
CTTGTATCACAGCAGCAACACAACAAGACTTTTTCACACCCAACCTCTTCCTGAAGCAGA
TGCTAGAGAGGTACCAAAACAACAGCCCTCCAAACAATGATGACCAGTGGAAAAACAATG
GAGTCACCAAAACCTGGGACAGGCTCATGCTCCAGGACAATTGCTGTGGCGTAAATGGTC
CATCAGACTGGCAAAAATACACATCTGCCTTCCGGACTGAGAATAATGATGCTGACTATC
CCTGGCCTCGTCAATGCTGTGTTATGAACAATCTTAAAGAACCTCTCAACCTGGAGGCTT
```

Des blocs de 60 lettres sont souvent utilisés.

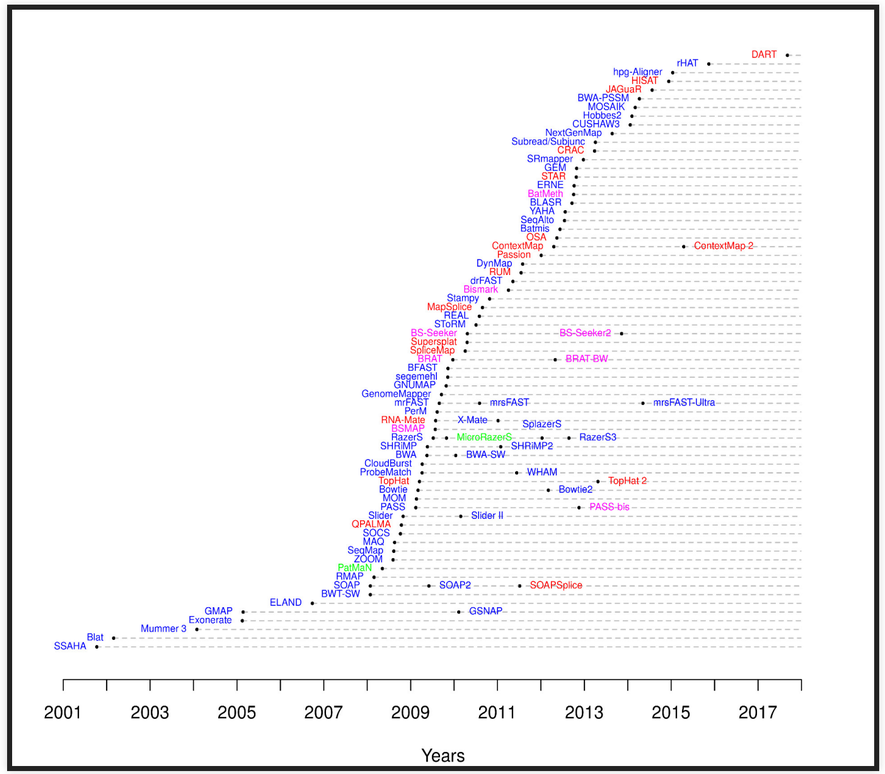
L'extension peut être .fa, .fasta, .fna...

LE MAPPING

LES ALGORITHMES DE MAPPING

Les algorithmes de mapping doivent s'adapter à la quantité et à la longueur des lectures issues du séquençage.

	Algorithm	Gapped	Quality-aware	Colorspace aware
BLAST	Hash table	Y	N	N
BLAT/SSHA2	Hash table	N	N	N
MAQ	Spaced seed	N	N	N
RMAP	Spaced seed	N	Y	N
ZOOM	Spaced seed	N	--	N
SOAP	Spaced seed	N	N	N
Eland	Spaced seed	N	N	N
SHRiMP	Q-gram/multi-seed	Y	Y	Y
BFAST	Q-gram/multi-seed	Y	Y	Y
Novoalign	Multi-seed + Vectorized SW	Y	Y	Y
clcBio	Multi-seed + Vectorized SW	Y	Y	Y
MUMmer	Tries	Y	N	N
OASIS	Tries	Y	--	--
VMATCH	Tries	Y	--	--
BWA/BWA-SW	Tries	Y	Y	Y
BOWTIE	Tries	Y	Y	Y
SOAP2	Tries	Y	N	N
Saruman	Exact (GPU)	Y	--	N



Le choix de l'outil et le paramétrage est dépendant des données (nature, longueur, quantité)

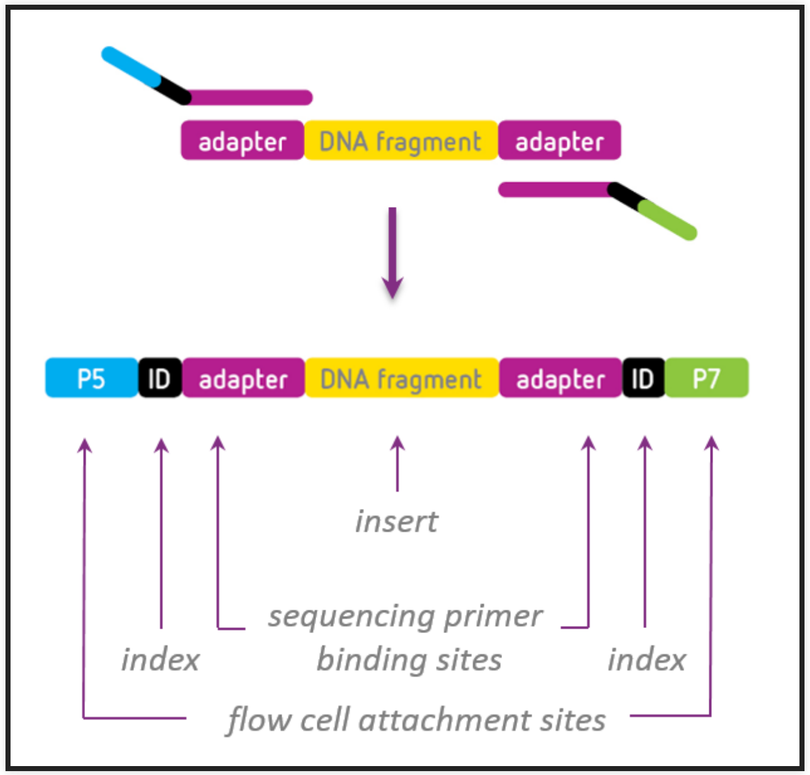
Quelques exemples parmi les plus utilisés :

- Génomique :
  - bowtie2 / bwa : lectures courtes <1 kb
- Transcriptomique (alignement splicé):
  - STAR
  - TopHat
  - HiSat2
- Long reads:
  - STAR
  - Encore une question de recherche !

AVANT DE MAPPER LES READS... ON NETTOIE

Il faut nettoyer les reads pour enlever tout ce qui n'est pas 'biologique' et de 'mauvaise qualité' :

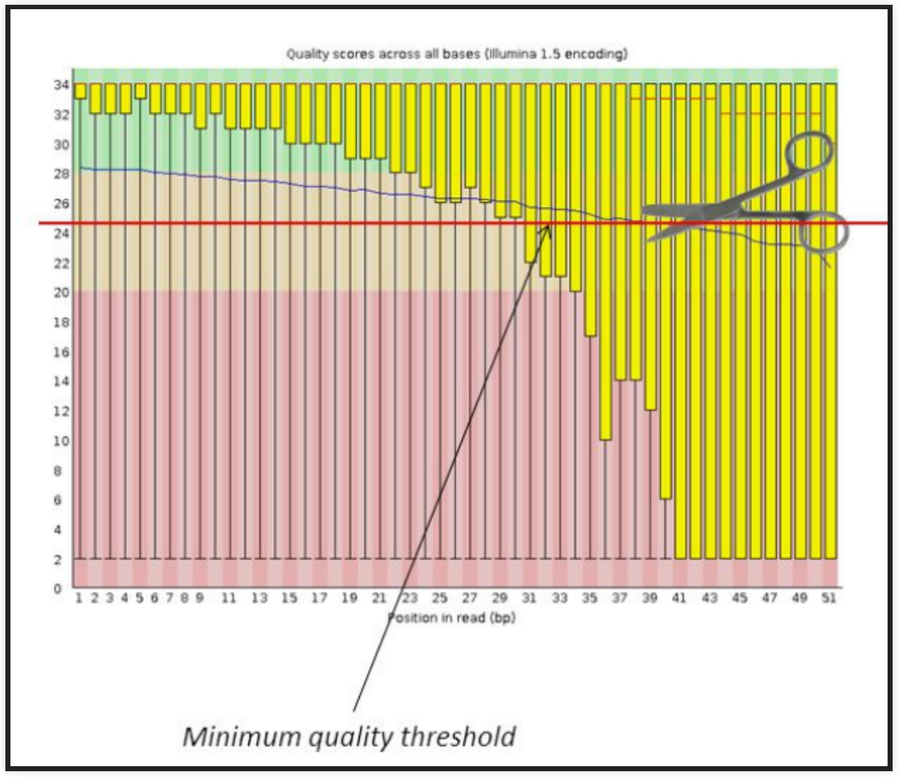
- Enlever les adaptateurs de séquençage, les barcodes... (cutadapt...)
- Trimmer les bases de mauvaise qualité (sickle, trimmomatic...)



AVANT DE MAPPER LES READS... ON NETTOIE

Puis filtrer/nettoyer/pre-processer selon différents critères :

- Enlever les lectures contenant des bases ambigues (pas toujours supportées par les outils)
- Supprimer les lectures trop petites
- Supprimer les lectures avec une complexité trop faible
- Merger les paires chevauchantes (pear, vsearch...)
- ...



AVANT DE MAPPER LES READS... ON NETTOIE

Chaque jeu de données est différent ! Il faut mesurer les effets de chaque étape (nombre de reads perdus, taille...) en faisant des graphiques, des tableaux...

C'est l'expérience qui guidera ensuite vos choix.

LE FORMAT SAM / BAM

Le format SAM a été élaboré lors du projet des 1000 génomes humains.

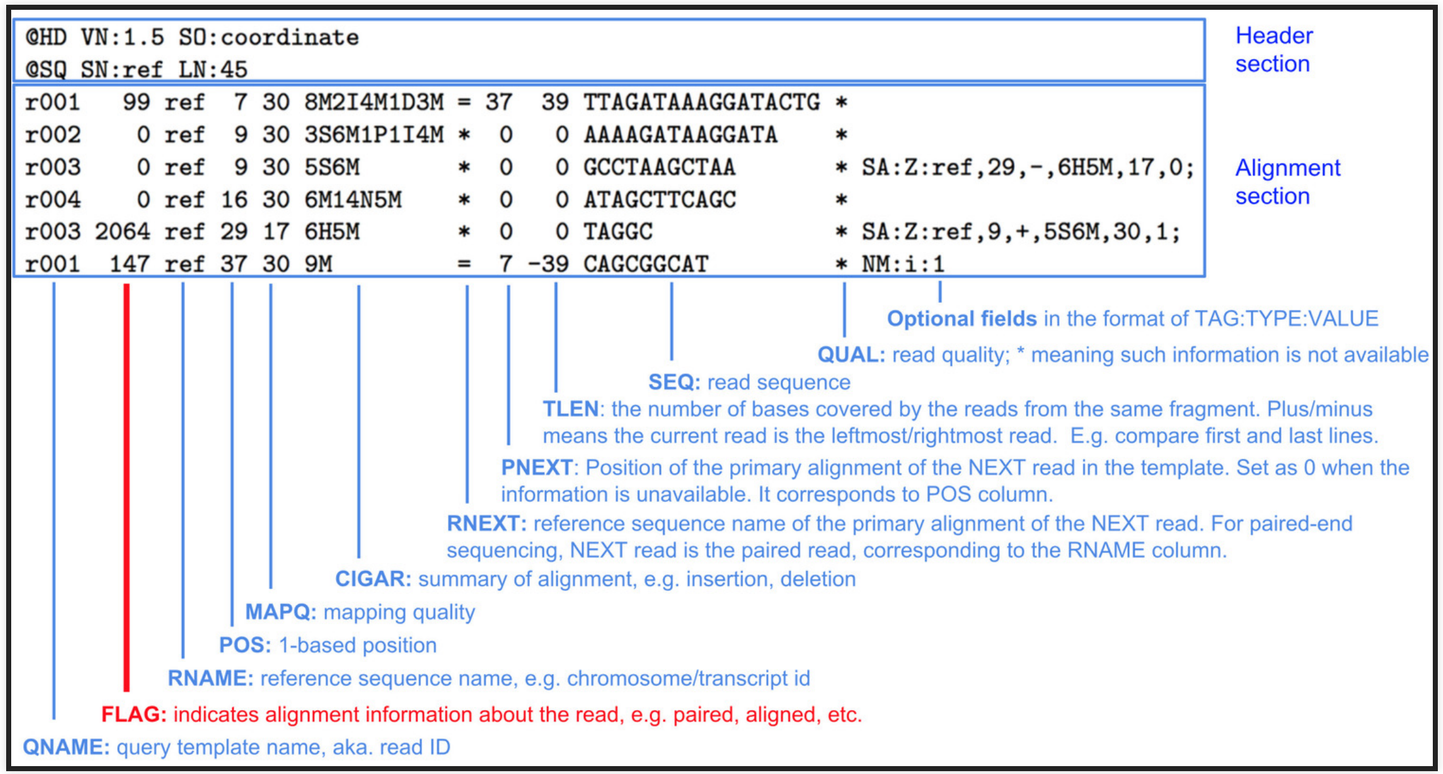
Il permet de stocker toutes les informations utiles des lectures alignées sur une référence.

Tous les outils de mapping génèrent du SAM.

Hautement compressible, indexable pour un accès rapide à l'information.

Le format SAM est un fichier texte. Sa version compressée est un BAM. Sa version ultra compressée est un CRAM.

LES INFORMATIONS DANS LE SAM



Documentation : SAM flags

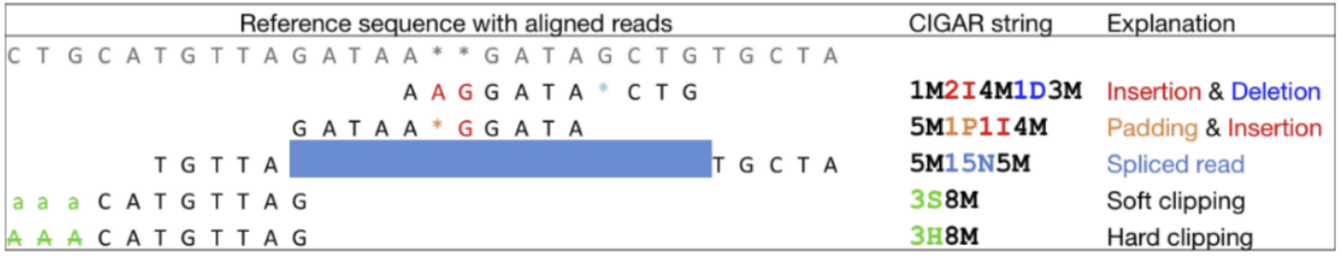
LES INFORMATIONS DANS LE SAM

`CIGAR` stands for *Concise Idiosyncratic Gapped Alignment Report*. This sixth field of a SAM file contains a so-called CIGAR string indicating which operations were necessary to map the read to the reference sequence at that particular locus.

The following operations are defined in CIGAR format (also see figure below):

- **M** - Alignment (can be a sequence match or mismatch!)
- **I** - Insertion in the read compared to the reference
- **D** - Deletion in the read compared to the reference
- **N** - Skipped region from the reference. For mRNA-to-genome alignments, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- **S** - Soft clipping (clipped sequences are present in read); S may only have H operations between them and the ends of the string
- **H** - Hard clipping (clipped sequences are NOT present in the alignment record); can only be present as the first and/or last operation
- **P** - Padding (silent deletion from padded reference)
- **=** - Sequence match (not widely used)
- **X** - Sequence mismatch (not widely used)

The sum of lengths of the **M**, **I**, **S**, **=**, **X** operations must equal the length of the read. Here are some examples:



MANIPULER LES FORMATS SAM / BAM

En complément du format, la suite d'outils SAMTOOLS a été développée pour manipuler ces formats.

Fonctionnalités :

- Compression / Décompression
- Tri / Indexage
- Merge
- Statistiques
- ...

EXERCICE N°5 : MANIPULER DES FICHIERS SAM/BAM

1. Visualiser le contenu d'un fichier BAM
2. Obtenir des statistiques sur le fichier BAM
3. Trier le fichier BAM
4. Indexer le fichier BAM

Visualiser le contenu d'un fichier BAM

```
samtools view # Affiche l'aide
samtools view file.bam | less # Ne pas oublier | less
```

Obtenir des statistiques sur le fichier BAM

```
samtools flafstat # Affiche l'aide
samtools flagstat file.bam
```

Trier le fichier BAM

```
samtools sort # Affiche l'aide
samtools sort -o file.sorted.bam file.bam
```

Indexer le fichier BAM

```
samtools index # Affiche l'aide
samtools index file.sorted.bam
```

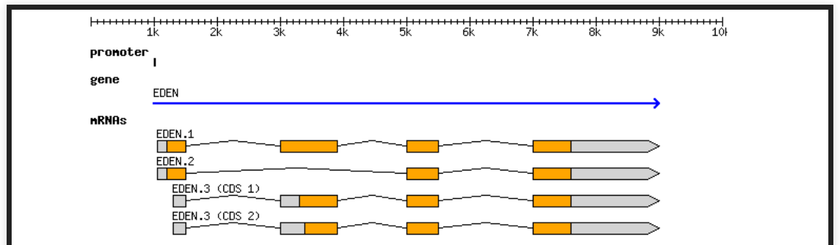
LES ANNOTATIONS DU GÉNOME DE RÉFÉRENCE

Le génome de référence donne uniquement des informations sur la séquence. Toutes les informations recensées ici (Sequence Ontology) (intron, exon, promoter, rRNA, CDS, centromere...) sont stockées ailleurs, dans un fichier d'annotation, qui indique leurs positions et leurs caractéristiques.

LE FORMAT GFF

Le format Gene Feature Format(GFF/ GFF2/GFF3) est un fichier texte où chaque ligne correspond à une annotation ou une feature. Les formats GFF, GFF2 et GFF3 sont légèrement différents.

La séquence peut être ajoutée en fin de GFF3




```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

Exemple d'un GFF3

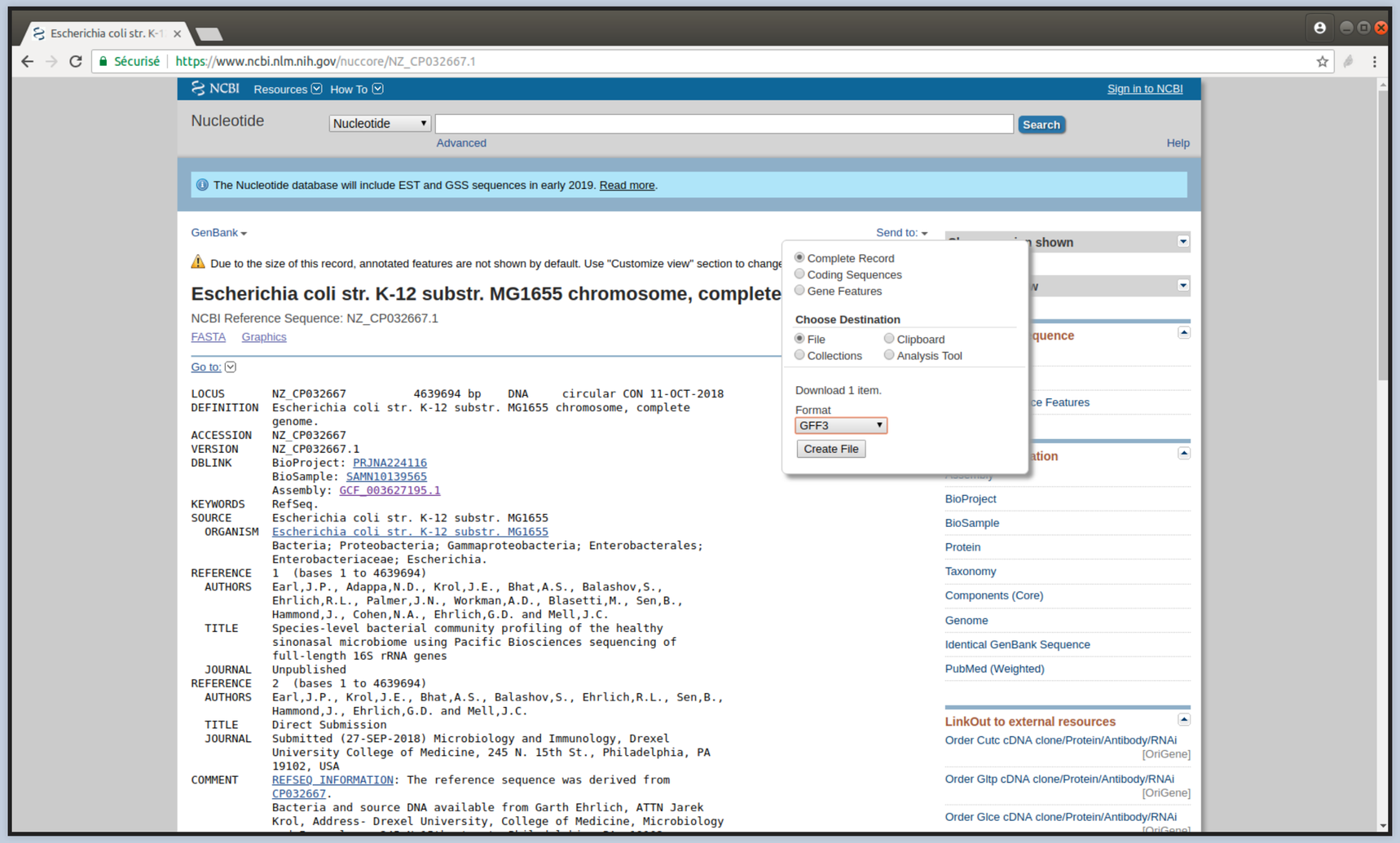
Il existe d'autres formats d'annotation, notamment Genbank, spécifique au NCBI, qui contient séquences et annotations.

```
misc_feature    447283..484667
                /note="Putative prophage; Putative lysogenic prophage
                region. Best Hits not completed. Putative attL/R; PHAGE01"
gene            complement(447387..448511)
                /locus_tag="BA0427"
                /db_xref="GeneID:1087434"
                /db_xref="prophinder:34501"
                /db_xref="sherw:116"
                /db_xref="phage_finder:16305"
CDS             complement(447387..448511)
                /locus_tag="BA0427"
                /note="identified by match to PFAM protein family HMM
                PF00589"
                /codon_start=1
                /transl_table=11
                /product="prophage LambdaBa04, site-specific recombinase,
                phage integrase family"
                /protein_id="NP_842969.1"
                /db_xref="GI:30260592"
                /db_xref="GeneID:1087434"
                /db_xref="aclame:protein:vir:6181"
                /translation="MKGYFRKRGEKWSFTIDIGKDPITGKRKQKTASGFKTKKEAERA
                CNELIHQFNTGSLVDDKNFTLSEYLQEWLENTAKQRVRETTFTNYKRAINSRIIPVLG
                SHKLKDLKPLHGQRFVKSLIDEGLSPAYIEYIFIVLKGSLEDAVRWELLFKNPFQHVE
                IPRPRKVVNSTWSIEETKKFLNRTKFENVIYYHLFLLALNTGMRRGEILGLKWKNFDL
                NEGKISVTETLIYDENGFRFTEPKTHGSKRLISIDQNLCKEFKSYKAKQNEFKLLFGQ
                SYEDNDLVFAKETGQPILPRTMTTTFNQFIKKADVPQIRFHDLRHTHATILLKLGINP
                KIVSERLGHSSIKTTLDTYSHVTIDMQESAVLKLSEALKS"
                /db_hit="ID=aclame:protein:vir:6181#Eval=6e-77#bits=281"
```

Exemple d'un fichier Genbank

EXERCICE N°6 : RÉCUPÉRER L'ANNOTATION DU GÉNOME E. COLI K-12 MG1655 AU NCBI

1. Récupérer l'annotation du génome E. coli K-12 MG1655 au NCBI au format GFF3
2. Afficher les 10 premières lignes du fichier GFF3


Escherichia coli str. K-1
Sécurisé | https://www.ncbi.nlm.nih.gov/nuccore/NZ_CP032667.1
NCBI Resources How To
Sign in to NCBI
Nucleotide
Nucleotide
Search
Advanced
Help
The Nucleotide database will include EST and GSS sequences in early 2019. Read more.
GenBank
Send to:
Due to the size of this record, annotated features are not shown by default. Use "Customize view" section to change
Escherichia coli str. K-12 substr. MG1655 chromosome, complete
NCBI Reference Sequence: NZ_CP032667.1
FASTA Graphics
Go to:
LOCUS NZ_CP032667 4639694 bp DNA circular CON 11-OCT-2018
DEFINITION Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome.
ACCESSION NZ_CP032667
VERSION NZ_CP032667.1
DBLINK BioProject: PRJNA224116
BioSample: SAMN10139565
Assembly: GCF_003627195.1
KEYWORDS RefSeq.
SOURCE Escherichia coli str. K-12 substr. MG1655
ORGANISM Escherichia coli str. K-12 substr. MG1655
Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia.
REFERENCE 1 (bases 1 to 4639694)
AUTHORS Earl,J.P., Adappa,N.D., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Palmer,J.N., Workman,A.D., Blasetti,M., Sen,B., Hammond,J., Cohen,N.A., Ehrlich,G.D. and Mell,J.C.
TITLE Species-level bacterial community profiling of the healthy sinonasal microbiome using Pacific Biosciences sequencing of full-length 16S rRNA genes
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 4639694)
AUTHORS Earl,J.P., Krol,J.E., Bhat,A.S., Balashov,S., Ehrlich,R.L., Sen,B., Hammond,J., Ehrlich,G.D. and Mell,J.C.
TITLE Direct Submission
JOURNAL Submitted (27-SEP-2018) Microbiology and Immunology, Drexel University College of Medicine, 245 N. 15th St., Philadelphia, PA 19102, USA
COMMENT REFSEQ INFORMATION: The reference sequence was derived from CP032667.
Bacteria and source DNA available from Garth Ehrlich, ATTN Jarek Krol, Address- Drexel University, College of Medicine, Microbiology
Complete Record
Coding Sequences
Gene Features
Choose Destination
File
Clipboard
Collections
Analysis Tool
Download 1 item.
Format
GFF3
Create File
BioProject
BioSample
Protein
Taxonomy
Components (Core)
Genome
Identical GenBank Sequence
PubMed (Weighted)
LinkOut to external resources
Order Cutc cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Gltp cDNA clone/Protein/Antibody/RNAi [OriGene]
Order Glce cDNA clone/Protein/Antibody/RNAi [OriGene]

```
head sequence.gff3
##sequence-region NZ_CP032667.1 1 4639694
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=511145
NZ_CP032667.1   RefSeq  region  1       4639694 .       +       .       ID=NZ_CP032667.1:1..4639694;Dbxref=taxon:
NZ_CP032667.1   RefSeq  gene    1052    2152    .       +       .       ID=gene-D8B36_RS00010;Name=D8B36_RS00010;
NZ_CP032667.1   Protein Homology        CDS     1052    2152    .       +       0       ID=cds-WP_000673464.1;Par
NZ_CP032667.1   RefSeq  gene    2152    3225    .       +       .       ID=gene-D8B36_RS00015;Name=recF;gbkey=Gen
NZ_CP032667.1   Protein Homology        CDS     2152    3225    .       +       0       ID=cds-WP_000060112.1;Par
NZ_CP032667.1   RefSeq  gene    3254    5668    .       +       .       ID=gene-D8B36_RS00020;Name=D8B36_RS00020;
NZ_CP032667.1   Protein Homology        CDS     3254    5668    .       +       0       ID=cds-WP_000072067.1;Par
NZ_CP032667.1   RefSeq  gene    5899    6306    .       +       .       ID=gene-D8B36_RS00025;Name=D8B36_RS00025;
```

AUTRES INDISPENSABLES

Formats

- Format BED
- Format BedGraph
- Format BigWig
- Format VCF
- ...

Outils

- BEDtools pour manipuler les fichiers génomiques (intersections...)
- IGV pour visualiser ces fichiers le long d'un génome
- ...

Librairies

- BioPython
- BioPerl
- ...