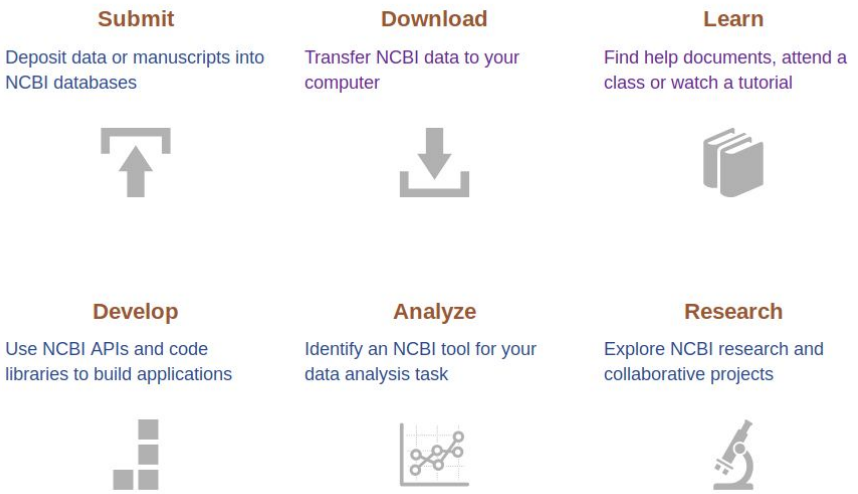
Méthodes et outils bioinformatiques pour l'analyse des données à haut débit

Olivier Rué & Olivier Kirsh

Slides complémentaires

NCBI : National Center for Biotechnology Information



Ressources à disposition (BD)

Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/)

SRA (https://www.ncbi.nlm.nih.gov/sra/)

GenBank (https://www.ncbi.nlm.nih.gov/genbank/) : DDBJ, ENA + NCBI data

RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) : non redondant

dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) BD variants humains

… (beaucoup)

Gros avantage : tout (ou presque) est interconnecté

Outils et APIs fournis par le NCBI

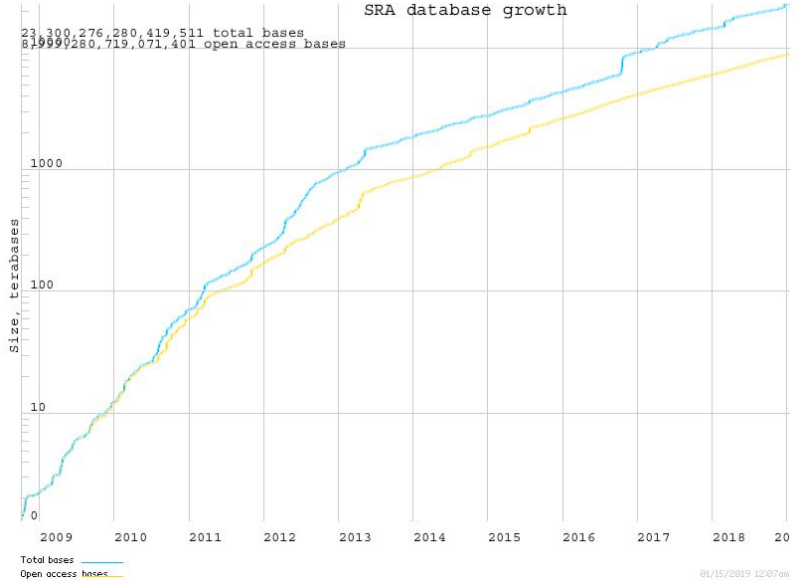
- Entrez Programming Utilities (E-utilities) https://www.ncbi.nlm.nih.gov/books/NBK25497/
  - Requêtage de la BD du NCBI via 9 E-utilities (ESearch, EFetch…)
- NCBI-Blast+
  - outil téléchargeable
  - API Rest (intégration dans des applications)
- Browsers
  - 1000 genomes Browser
- Une multitude d’autres !

Cas d’utilisation du NCBI

- Je veux utiliser des données publiques (des lectures, un génome, une annotation, une base de données de variants…)
- Informations sur un organisme -> NCBI taxonomy
- Bibliographie sur n’importe quoi -> PubMed par mot clé
- J’ai assemblé une séquence mais je ne sais pas ce que c’est -> NCBI Blast
- …
- il faut presenter, GEO, SRA et ENA
- sra toolkit aussi

SRA

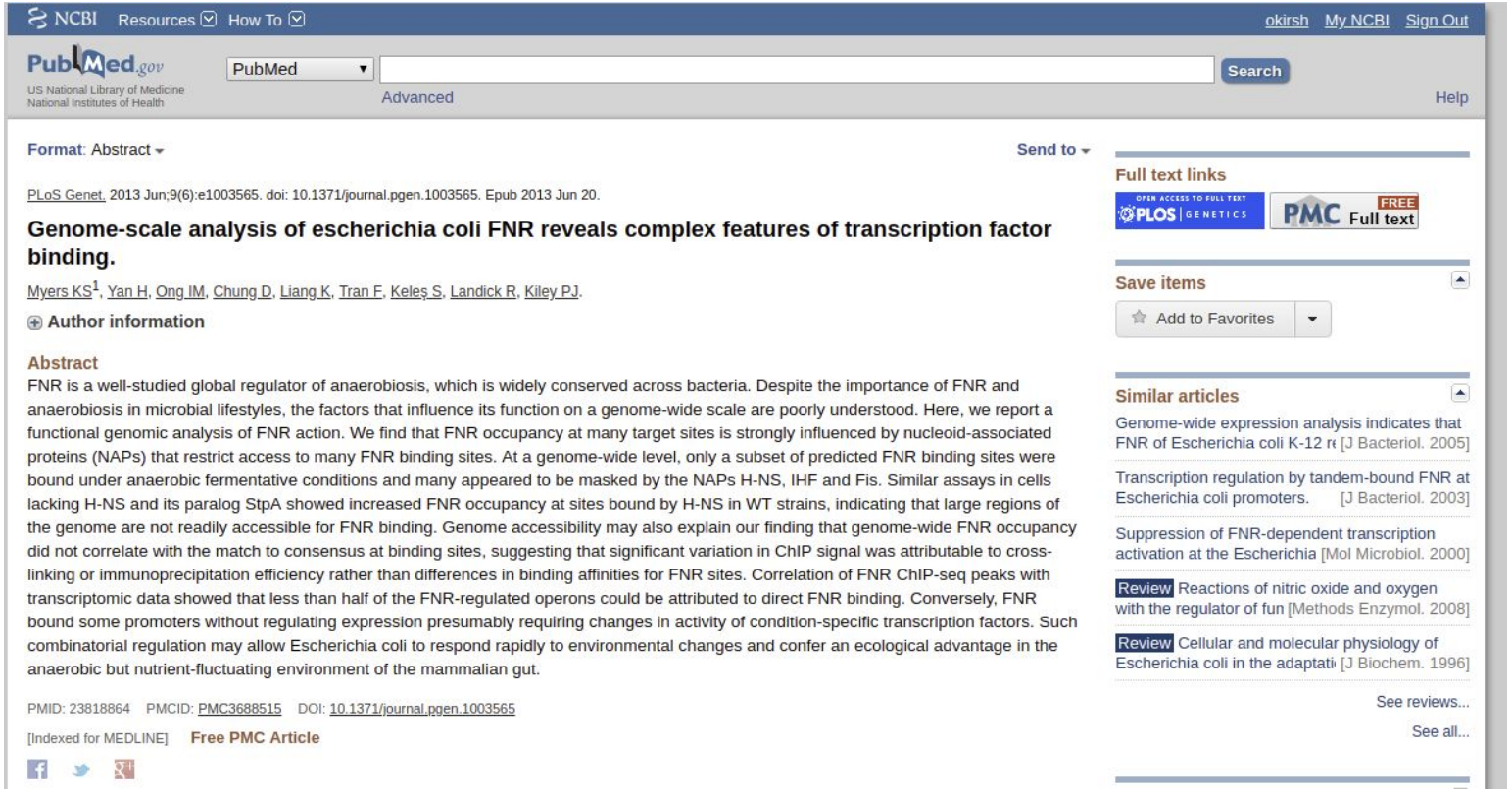
- Stockage de données de séquençage (brutes ou alignées)
- Possibilité de déposer vos séquences
- Possibilité d’avoir accès aux séquences
  - en ligne
  - via SRA-toolkit



Exercice 2 : récupérer des fichiers FASTQ

- Récupérer le FASTQ ayant pour RUNID : SRR126457
  - Quelle technologie de séquençage a permis de produire cet échantillon ?
  - Combien y a-t-il de lectures ?
  - Trouvez un moyen de télécharger ce fichier en utilisant l'interface (passer par Experiment)
- Récupérer le FASTQ via la ligne de commande
  - fastq-dump

How to retreive public datasets



NCBI Resources How To okirsh My NCBI Sign Out

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Search
Advanced Help

Format: Abstract Send to

PLoS Genet. 2013 Jun;9(6):e1003565. doi: 10.1371/journal.pgen.1003565. Epub 2013 Jun 20.

Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding.

Myers KS[1], Yan H, Ong IM, Chung D, Liang K, Tran F, Keleş S, Landick R, Kiley PJ.

Author information

Abstract

FNR is a well-studied global regulator of anaerobiosis, which is widely conserved across bacteria. Despite the importance of FNR and anaerobiosis in microbial lifestyles, the factors that influence its function on a genome-wide scale are poorly understood. Here, we report a functional genomic analysis of FNR action. We find that FNR occupancy at many target sites is strongly influenced by nucleoid-associated proteins (NAPs) that restrict access to many FNR binding sites. At a genome-wide level, only a subset of predicted FNR binding sites were bound under anaerobic fermentative conditions and many appeared to be masked by the NAPs H-NS, IHF and Fis. Similar assays in cells lacking H-NS and its paralog StpA showed increased FNR occupancy at sites bound by H-NS in WT strains, indicating that large regions of the genome are not readily accessible for FNR binding. Genome accessibility may also explain our finding that genome-wide FNR occupancy did not correlate with the match to consensus at binding sites, suggesting that significant variation in ChIP signal was attributable to cross-linking or immunoprecipitation efficiency rather than differences in binding affinities for FNR sites. Correlation of FNR ChIP-seq peaks with transcriptomic data showed that less than half of the FNR-regulated operons could be attributed to direct FNR binding. Conversely, FNR bound some promoters without regulating expression presumably requiring changes in activity of condition-specific transcription factors. Such combinatorial regulation may allow Escherichia coli to respond rapidly to environmental changes and confer an ecological advantage in the anaerobic but nutrient-fluctuating environment of the mammalian gut.

PMID: 23818864 PMCID: PMC3688515 DOI: 10.1371/journal.pgen.1003565

[Indexed for MEDLINE] Free PMC Article

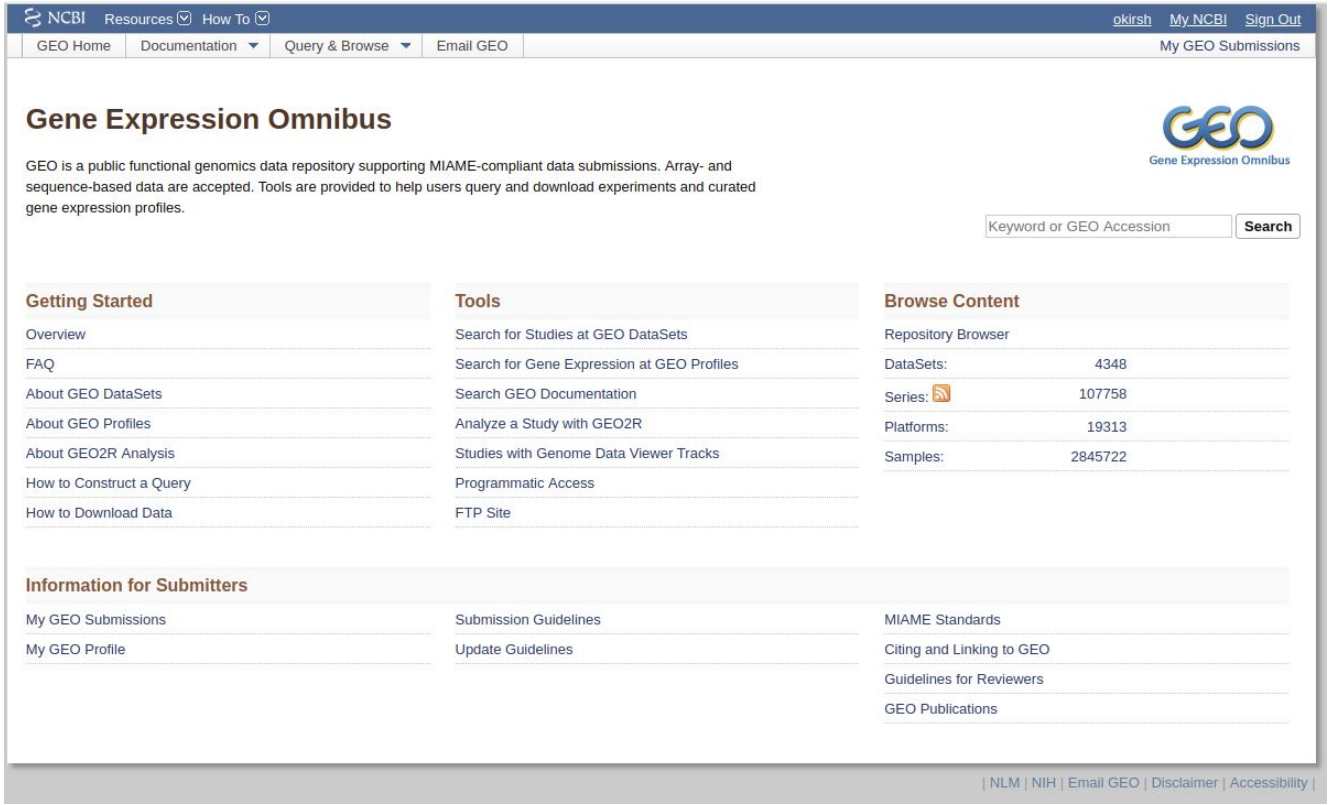
Full text links
PLOS GENETICS PMC FREE Full text

Save items
Add to Favorites

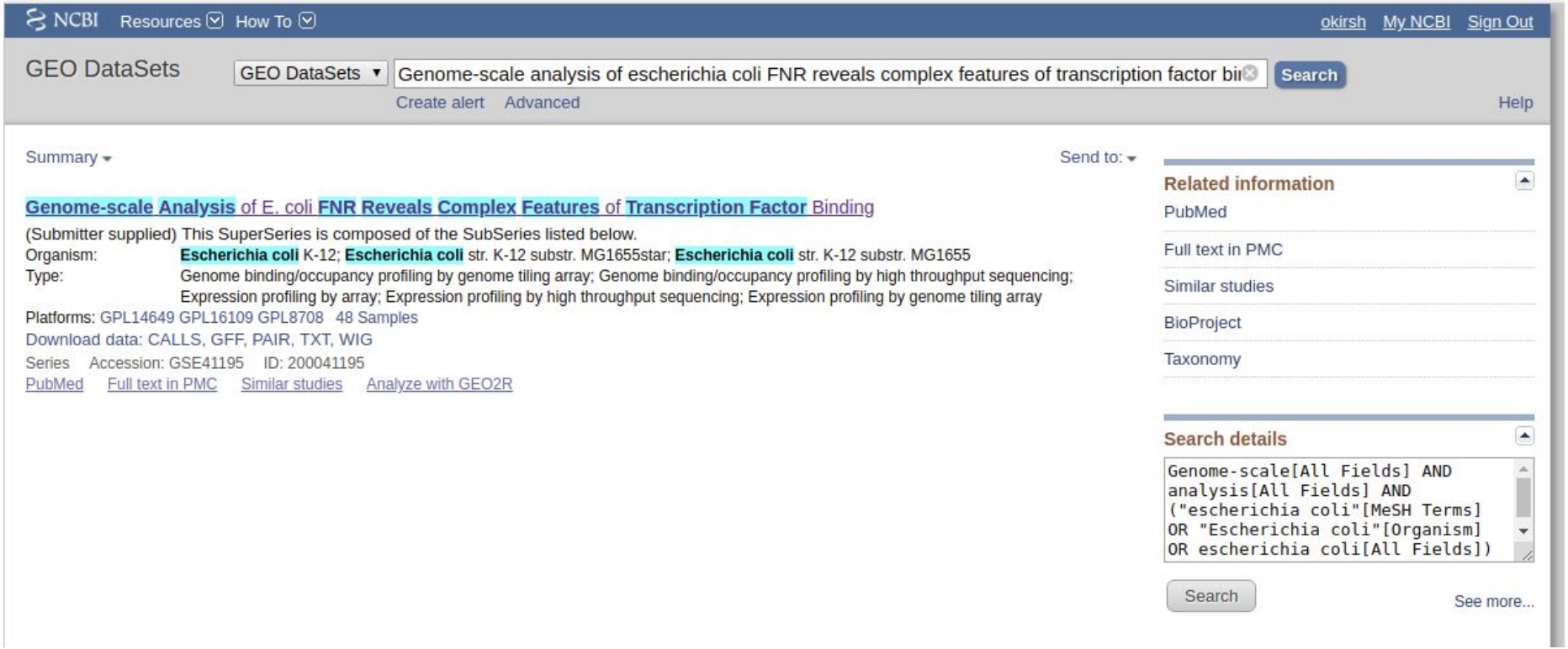
Similar articles
Genome-wide expression analysis indicates that FNR of Escherichia coli K-12 r [J Bacteriol. 2005]
Transcription regulation by tandem-bound FNR at Escherichia coli promoters. [J Bacteriol. 2003]
Suppression of FNR-dependent transcription activation at the Escherichia [Mol Microbiol. 2000]
Review Reactions of nitric oxide and oxygen with the regulator of fun [Methods Enzymol. 2008]
Review Cellular and molecular physiology of Escherichia coli in the adaptati [J Biochem. 1996]

See reviews...
See all...

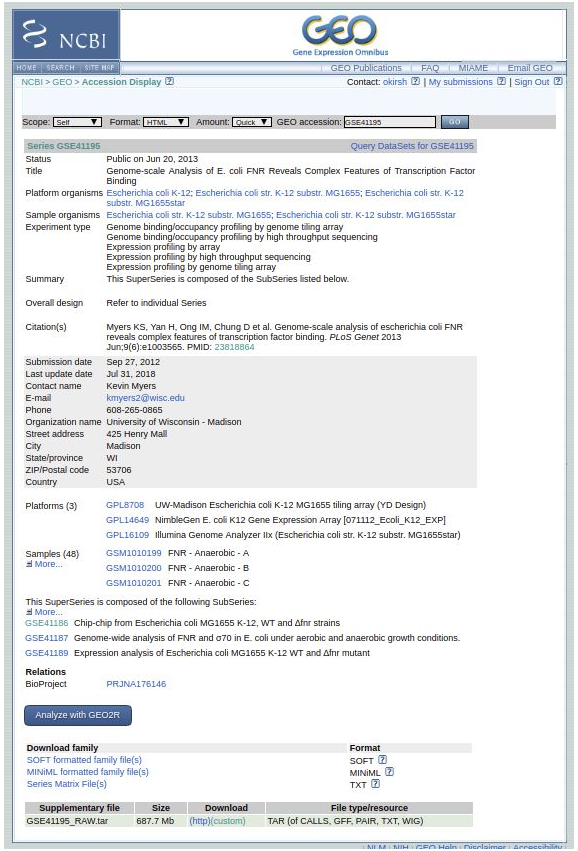
NCBI - GEO DataSets: Fetch FASTQ



NCBI - GEO DataSets: Fetch FASTQ



NCBI - GEO DataSets: Fetch FASTQ



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Contact: okirsh | My submissions | Sign Out

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE41195 | GO

Series GSE41195 — Query DataSets for GSE41195

Status	Public on Jun 20, 2013
Title	Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding
Platform organisms	Escherichia coli K-12; Escherichia coli str. K-12 substr. MG1655; Escherichia coli str. K-12 substr. MG1655star
Sample organisms	Escherichia coli str. K-12 substr. MG1655; Escherichia coli str. K-12 substr. MG1655star
Experiment type	Genome binding/occupancy profiling by genome tiling array Genome binding/occupancy profiling by high throughput sequencing Expression profiling by array Expression profiling by high throughput sequencing Expression profiling by genome tiling array
Summary	This SuperSeries is composed of the SubSeries listed below.
Overall design	Refer to individual Series
Citation(s)	Myers KS, Yan H, Ong IM, Chung D et al. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet* 2013 Jun;9(6):e1003565. PMID: 23818864
Submission date	Sep 27, 2012
Last update date	Jul 31, 2018
Contact name	Kevin Myers
E-mail	kmyers2@wisc.edu
Phone	608-265-0865
Organization name	University of Wisconsin - Madison
Street address	425 Henry Mall
City	Madison
State/province	WI
ZIP/Postal code	53706
Country	USA

Platforms (3) GPL8708 UW-Madison Escherichia coli K-12 MG1655 tiling array (YD Design)
GPL14649 NimbleGen E. coli K12 Gene Expression Array [071112_Ecoli_K12_EXP]
GPL16109 Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)

Samples (48) GSM1010199 FNR - Anaerobic - A
More...
GSM1010200 FNR - Anaerobic - B
GSM1010201 FNR - Anaerobic - C

This SuperSeries is composed of the following SubSeries:
More...
GSE41186 Chip-chip from Escherichia coli MG1655 K-12, WT and Δfnr strains
GSE41187 Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.
GSE41189 Expression analysis of Escherichia coli MG1655 K-12 WT and Δfnr mutant

Relations
BioProject PRJNA176146

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

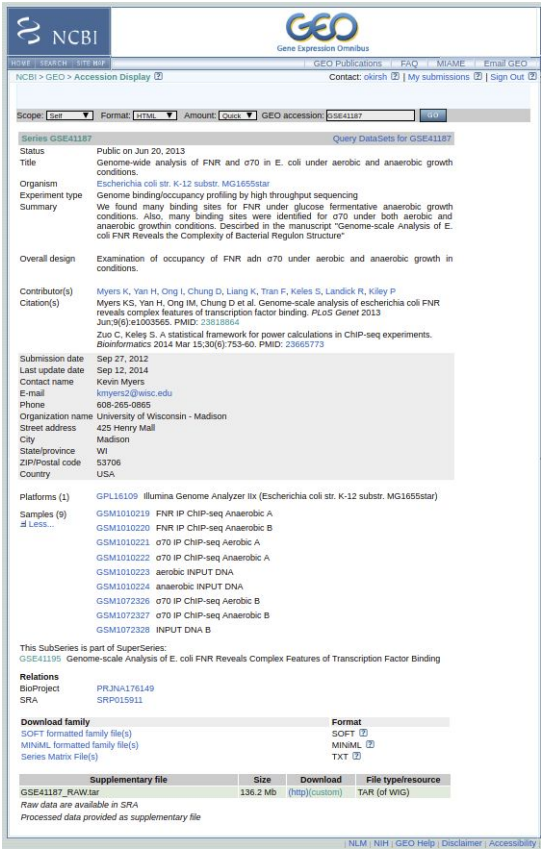
Supplementary file	Size	Download	File type/resource
GSE41195_RAW.tar	687.7 Mb	(http)(custom)	TAR (of CALLS, GFF, PAIR, TXT, WIG)

NLM | NIH | GEO Help | Disclaimer | Accessibility

Platforms (3) GPL8708 UW-Madison Escherichia coli K-12 MG1655 tiling array (YD Design)
GPL14649 NimbleGen E. coli K12 Gene Expression Array [071112_Ecoli_K12_EXP]
GPL16109 Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)

Samples (48)
Less...
GSM1010199 FNR - Anaerobic - A
GSM1010200 FNR - Anaerobic - B
GSM1010201 FNR - Anaerobic - C
GSM1010202 ß - Anaerobic - A
GSM1010203 ß - Anaerobic - B
GSM1010204 ß - Aerobic - A
GSM1010205 ß - Aerobic - B
GSM1010206 HNS - Anaerobic A
GSM1010207 HNS - Aerobic A
GSM1010208 IHF - Anaerobic A
GSM1010209 Δfnr - Anaeroibc
GSM1010210 Ptac::fnr - A - 4 µM IPTG
GSM1010211 Ptac::fnr - B - 4 µM IPTG
GSM1010212 Ptac::fnr - A - 8 µM IPTG
GSM1010213 Ptac::fnr - B - 8 µM IPTG
GSM1010214 Ptac::fnr - A - 16 µM IPTG
GSM1010215 Ptac::fnr - B - 16 µM IPTG
GSM1010216 Ptac::fnr - C - 16 µM IPTG
GSM1010217 FNR - Anaerobic - Affinity Purified - A
GSM1010218 FNR - Anaerobic - Affinity Purified - B
GSM1010219 FNR IP ChIP-seq Anaerobic A
GSM1010220 FNR IP ChIP-seq Anaerobic B
GSM1010221 σ70 IP ChIP-seq Aerobic A
GSM1010222 σ70 IP ChIP-seq Anaerobic A
GSM1010223 aerobic INPUT DNA
GSM1010224 anaerobic INPUT DNA
GSM1010240 Ecoli_wild-type_rep1_Anaerobic
GSM1010241 Ecoli_wild-type_rep2_Anaerobic
GSM1010242 Ecoli_dFNR_rep1_Anaerobic
GSM1010243 Ecoli_dFNR_rep2_Anaerobic
GSM1010244 Ecoli_wild-type_rep1_anaerobic
GSM1010245 Ecoli_wild-type_rep2_anaerobic
GSM1010246 Ecoli_dFNR_rep1_anaerobic
GSM1010247 Ecoli_dFNR_rep2_anaerobic
GSM1054383 FNR - ΔhnsΔstpA A
GSM1054384 FNR - ΔhnsΔstpA B
GSM1072326 σ70 IP ChIP-seq Aerobic B
GSM1072327 σ70 IP ChIP-seq Anaerobic B
GSM1072328 INPUT DNA B
GSM1124977 Ecoli_hns-stpA_repA_Aerobic_Tiling
GSM1124978 Ecoli_hns-stpA_repB_Aerobic_Tiling
GSM1124979 Ecoli_hns-stpA_repA_Anaerobic_Tiling
GSM1124980 Ecoli_hns-stpA_repB_Anaerobic_Tiling
GSM1126445 Ecoli_wild-type_repA_Aerobic
GSM1126446 Ecoli_wild-type_repB_Aerobic
GSM1143890 HNS - Anaerobic B
GSM1143891 HNS - Aerobic B
GSM1143892 IHF - Anaerobic B

This SuperSeries is composed of the following SubSeries:
More...
GSE41186 Chip-chip from Escherichia coli MG1655 K-12, WT and Δfnr strains
GSE41187 Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.
GSE41189 Expression analysis of Escherichia coli MG1655 K-12 WT and Δfnr mutant



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Contact: okirsh | My submissions | Sign Out

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE41187 | GO

Series GSE41187 — Query DataSets for GSE41187

Status	Public on Jun 20, 2013
Title	Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.
Organism	Escherichia coli str. K-12 substr. MG1655star
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	We found many binding sites for FNR under glucose fermentative anaerobic growth conditions. Also, many binding sites were identified for σ70 under both aerobic and anaerobic growthin conditions. Descirbed in the manuscript "Genome-scale Analysis of E. coli FNR Reveals the Complexity of Bacterial Regulon Structure"
Overall design	Examination of occupancy of FNR adn σ70 under aerobic and anaerobic growth in conditions.
Contributor(s)	Myers K, Yan H, Ong I, Chung D, Liang K, Tran F, Keles S, Landick R, Kiley P
Citation(s)	Myers KS, Yan H, Ong IM, Chung D et al. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet* 2013 Jun;9(6):e1003565. PMID: 23818864 Zuo C, Keles S. A statistical framework for power calculations in ChIP-seq experiments. Bioinformatics 2014 Mar 15;30(6):753-60. PMID: 23665773
Submission date	Sep 27, 2012
Last update date	Sep 12, 2014
Contact name	Kevin Myers
E-mail	kmyers2@wisc.edu
Phone	608-265-0865
Organization name	University of Wisconsin - Madison
Street address	425 Henry Mall
City	Madison
State/province	WI
ZIP/Postal code	53706
Country	USA

Platforms (1) GPL16109 Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)

Samples (9)
Less...
GSM1010219 FNR IP ChIP-seq Anaerobic A
GSM1010220 FNR IP ChIP-seq Anaerobic B
GSM1010221 σ70 IP ChIP-seq Aerobic A
GSM1010222 σ70 IP ChIP-seq Anaerobic A
GSM1010223 aerobic INPUT DNA
GSM1010224 anaerobic INPUT DNA
GSM1072326 σ70 IP ChIP-seq Aerobic B
GSM1072327 σ70 IP ChIP-seq Anaerobic B
GSM1072328 INPUT DNA B

This SubSeries is part of SuperSeries:
GSE41195 Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

Relations
BioProject PRJNA176149
SRA SRP015911

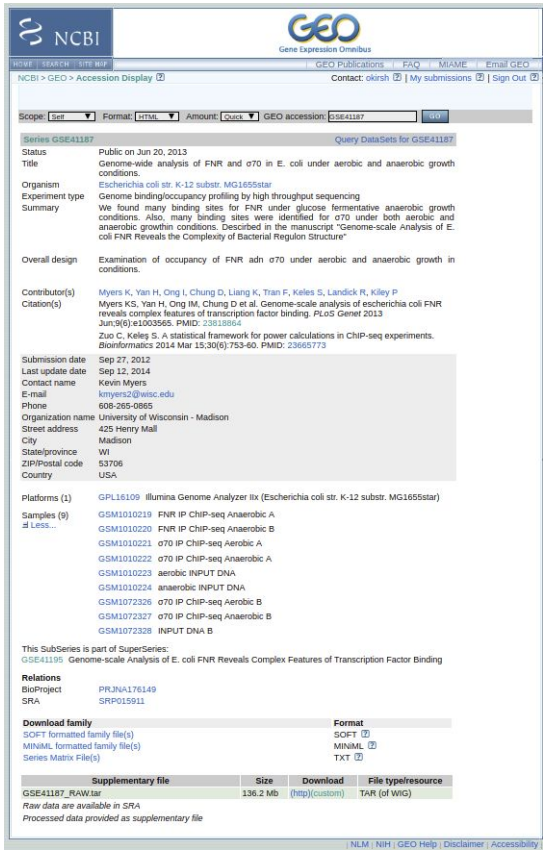
Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE41187_RAW.tar	136.2 Mb	(http)(custom)	TAR (of WIG)

Raw data are available in SRA
Processed data provided as supplementary file

NLM | NIH | GEO Help | Disclaimer | Accessibility

NCBI - GEO DataSets: Fetch FASTQ



NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > Accession Display

Contact: okirsh | My submissions | Sign Out

Scope: Self Format: HTML Amount: Quick GEO accession: GSE41187 Go

Series GSE41187 Query DataSets for GSE41187

Status	Public on Jun 20, 2013
Title	Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.
Organism	Escherichia coli str. K-12 substr. MG1655star
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	We found many binding sites for FNR under glucose fermentative anaerobic growth conditions. Also, many binding sites were identified for σ70 under both aerobic and anaerobic growthin conditions. Described in the manuscript "Genome-scale Analysis of E. coli FNR Reveals the Complexity of Bacterial Regulon Structure"
Overall design	Examination of occupancy of FNR adn σ70 under aerobic and anaerobic growth in conditions.
Contributor(s)	Myers K, Yan H, Ong I, Chung D, Liang K, Tran F, Keles S, Landick R, Kiley P
Citation(s)	Myers KS, Yan H, Ong IM, Chung D et al. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet* 2013 Jun;9(6):e1003565. PMID: 23818864 Zuo C, Keleş S. A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* 2014 Mar 15;30(6):753-60. PMID: 23665773
Submission date	Sep 27, 2012
Last update date	Sep 12, 2014
Contact name	Kevin Myers
E-mail	kmyers2@wisc.edu
Phone	608-265-0865
Organization name	University of Wisconsin - Madison
Street address	425 Henry Mall
City	Madison
State/province	WI
ZIP/Postal code	53706
Country	USA

Platforms (1) GPL16109 Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)

Samples (9) Less...

GSM1010219 FNR IP ChIP-seq Anaerobic A
GSM1010220 FNR IP ChIP-seq Anaerobic B
GSM1010221 σ70 IP ChIP-seq Aerobic A
GSM1010222 σ70 IP ChIP-seq Anaerobic A
GSM1010223 aerobic INPUT DNA
GSM1010224 anaerobic INPUT DNA
GSM1072326 σ70 IP ChIP-seq Aerobic B
GSM1072327 σ70 IP ChIP-seq Anaerobic B
GSM1072328 INPUT DNA B

This SubSeries is part of SuperSeries:
GSE41195 Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

Relations

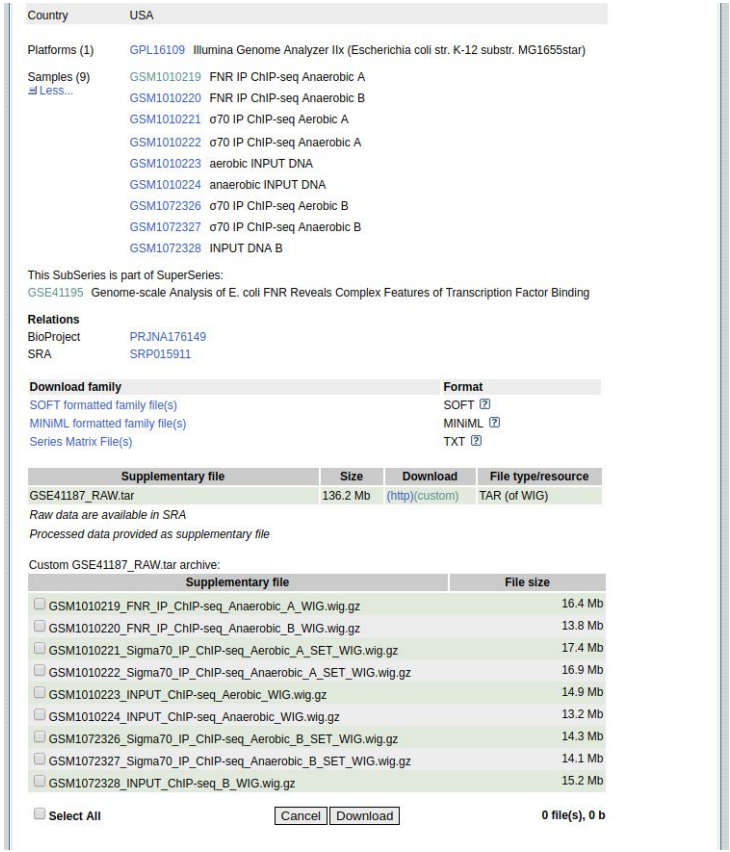
BioProject	PRJNA176149
SRA	SRP015911

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE41187_RAW.tar	136.2 Mb	(http)(custom)	TAR (of WIG)

Raw data are available in SRA
Processed data provided as supplementary file

NLM | NIH | GEO Help | Disclaimer | Accessibility



Country	USA

Platforms (1) GPL16109 Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)

Samples (9) Less...

GSM1010219 FNR IP ChIP-seq Anaerobic A
GSM1010220 FNR IP ChIP-seq Anaerobic B
GSM1010221 σ70 IP ChIP-seq Aerobic A
GSM1010222 σ70 IP ChIP-seq Anaerobic A
GSM1010223 aerobic INPUT DNA
GSM1010224 anaerobic INPUT DNA
GSM1072326 σ70 IP ChIP-seq Aerobic B
GSM1072327 σ70 IP ChIP-seq Anaerobic B
GSM1072328 INPUT DNA B

This SubSeries is part of SuperSeries:
GSE41195 Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

Relations

BioProject	PRJNA176149
SRA	SRP015911

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE41187_RAW.tar	136.2 Mb	(http)(custom)	TAR (of WIG)

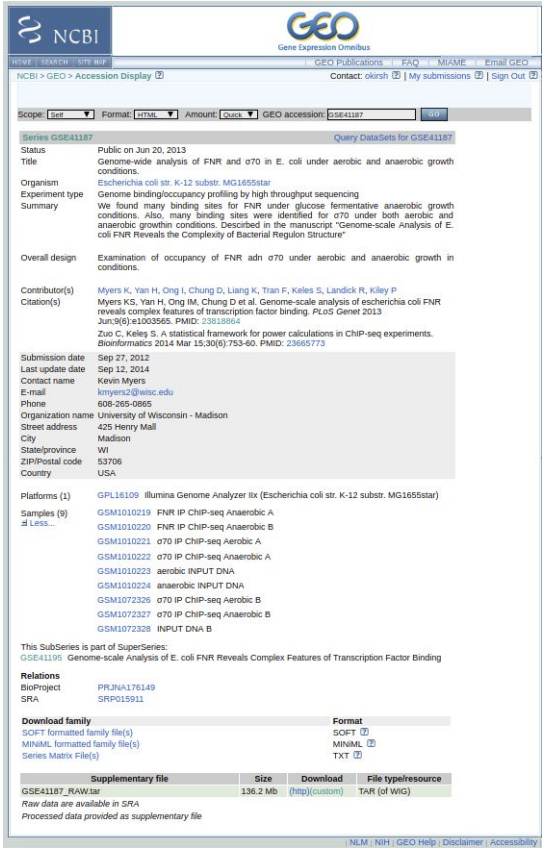
Raw data are available in SRA
Processed data provided as supplementary file

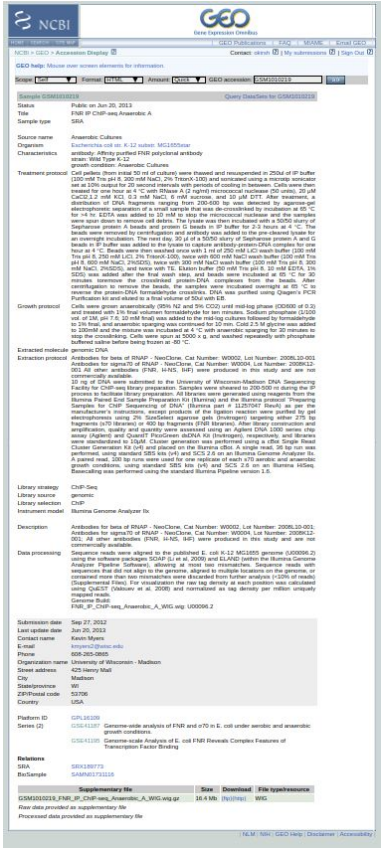
Custom GSE41187_RAW.tar archive:

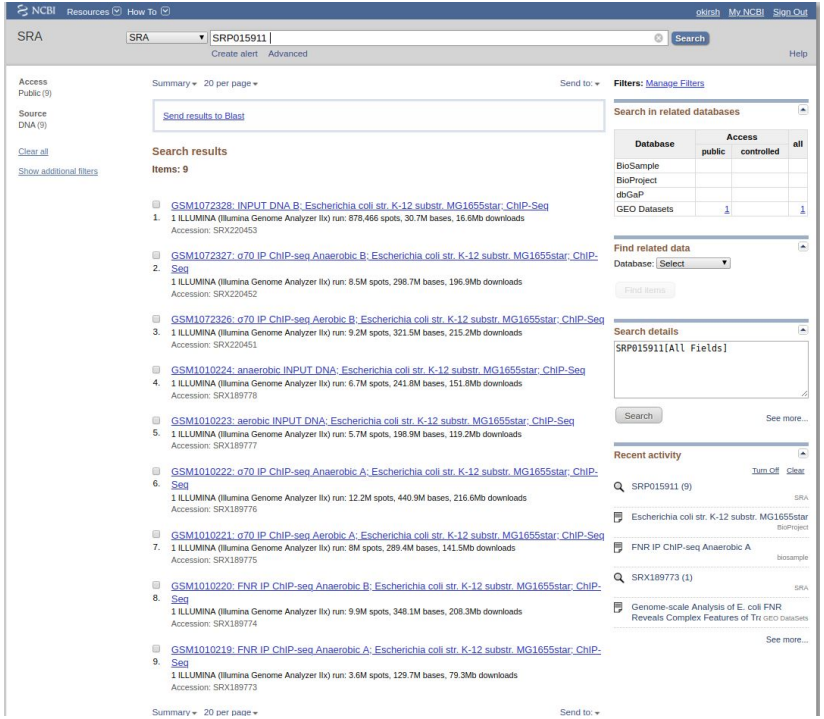
Supplementary file	File size
GSM1010219_FNR_IP_ChIP-seq_Anaerobic_A_WIG.wig.gz	16.4 Mb
GSM1010220_FNR_IP_ChIP-seq_Anaerobic_B_WIG.wig.gz	13.8 Mb
GSM1010221_Sigma70_IP_ChIP-seq_Aerobic_A_SET_WIG.wig.gz	17.4 Mb
GSM1010222_Sigma70_IP_ChIP-seq_Anaerobic_A_SET_WIG.wig.gz	16.9 Mb
GSM1010223_INPUT_ChIP-seq_Aerobic_WIG.wig.gz	14.9 Mb
GSM1010224_INPUT_ChIP-seq_Anaerobic_WIG.wig.gz	13.2 Mb
GSM1072326_Sigma70_IP_ChIP-seq_Aerobic_B_SET_WIG.wig.gz	14.3 Mb
GSM1072327_Sigma70_IP_ChIP-seq_Anaerobic_B_SET_WIG.wig.gz	14.1 Mb
GSM1072328_INPUT_ChIP-seq_B_WIG.wig.gz	15.2 Mb

Select All Cancel Download 0 file(s), 0 b

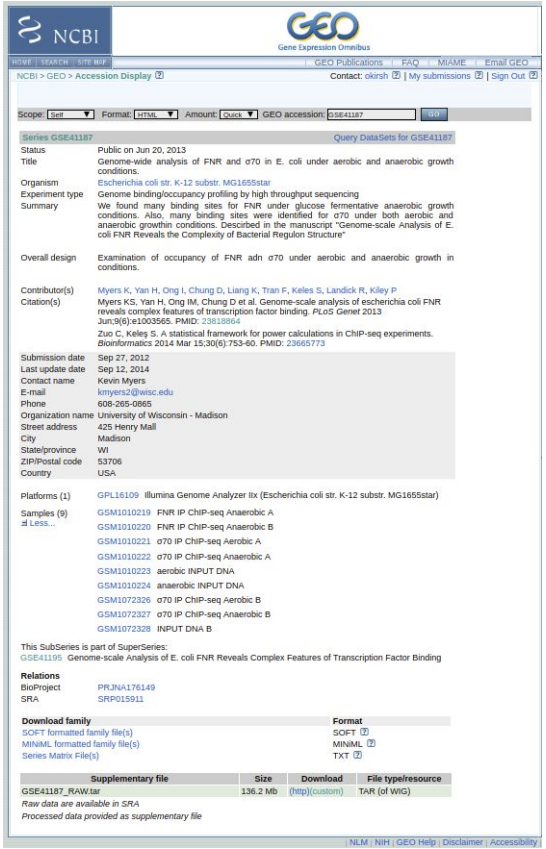
NCBI - GEO DataSets: Fetch FASTQ

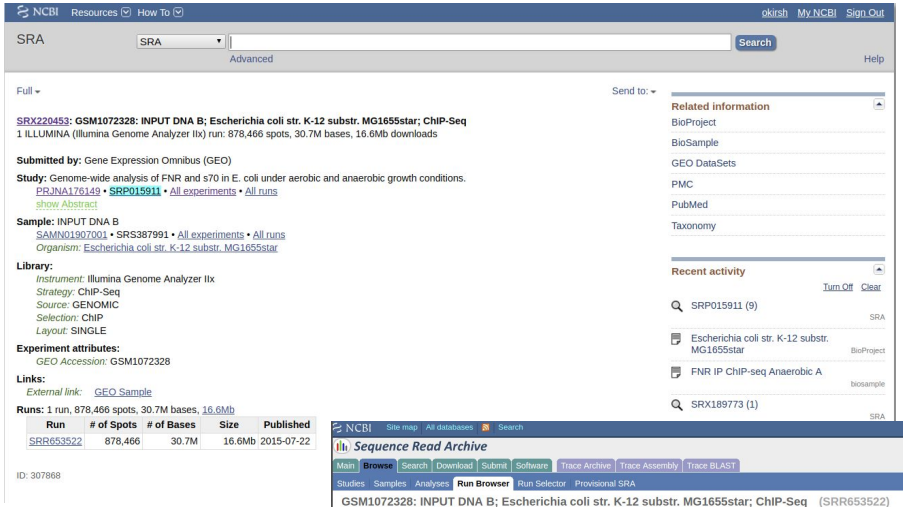


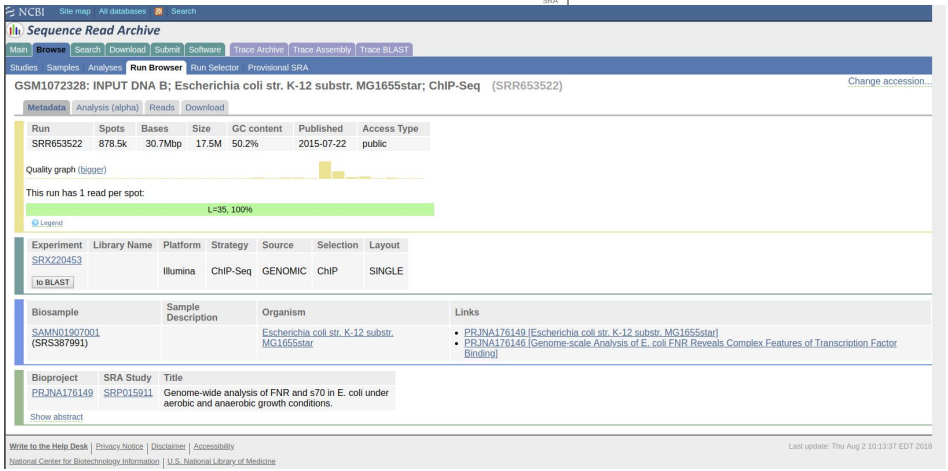




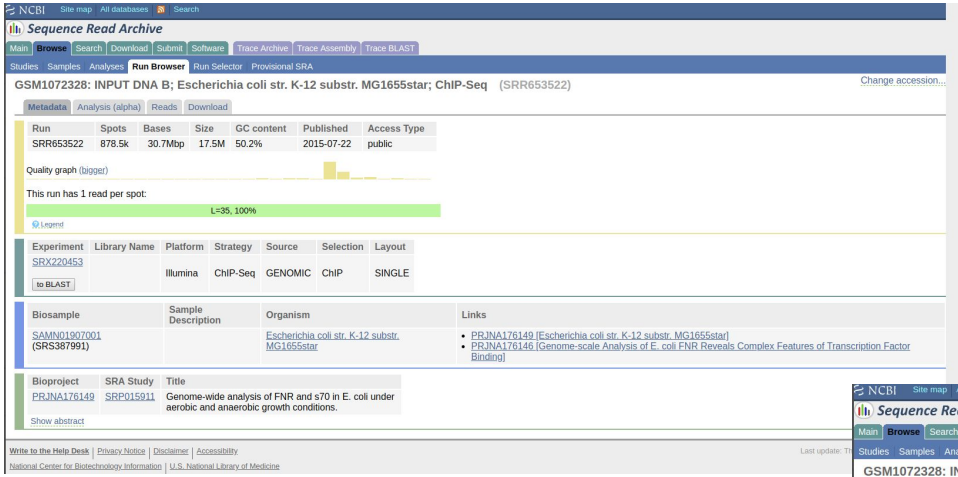
NCBI - GEO DataSets: Fetch FASTQ

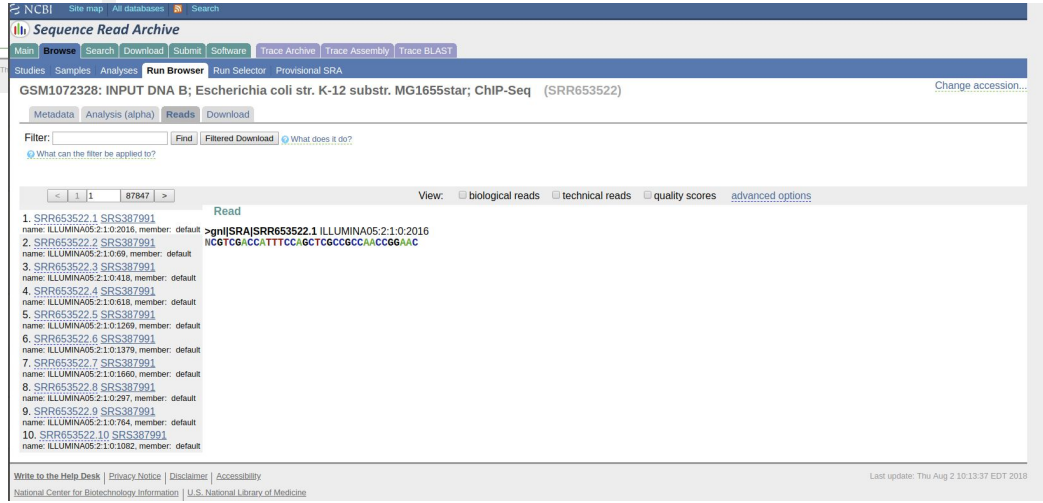




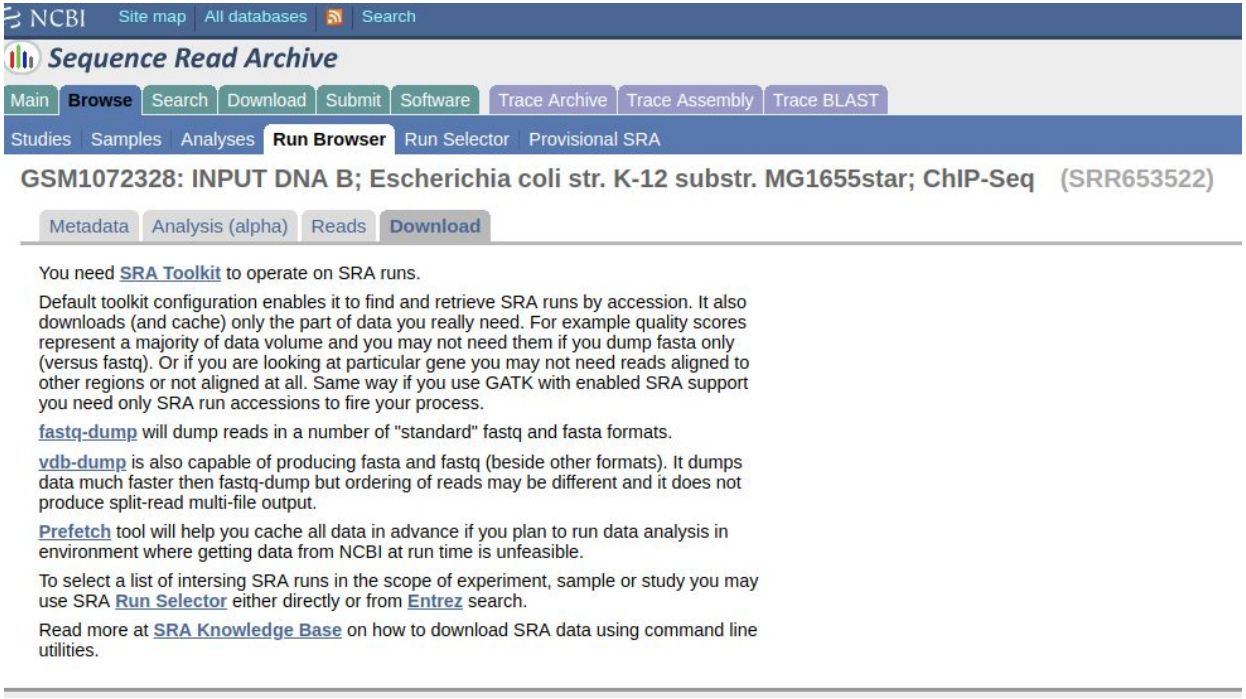


NCBI - SRA: Fetch FASTQ

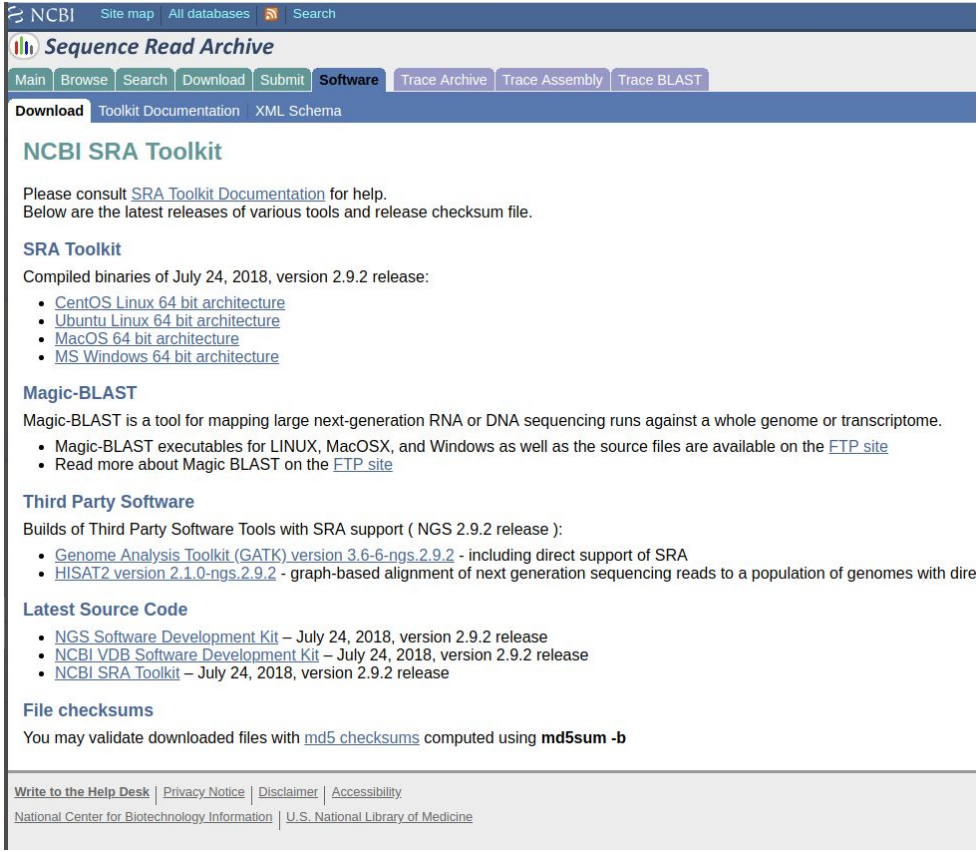


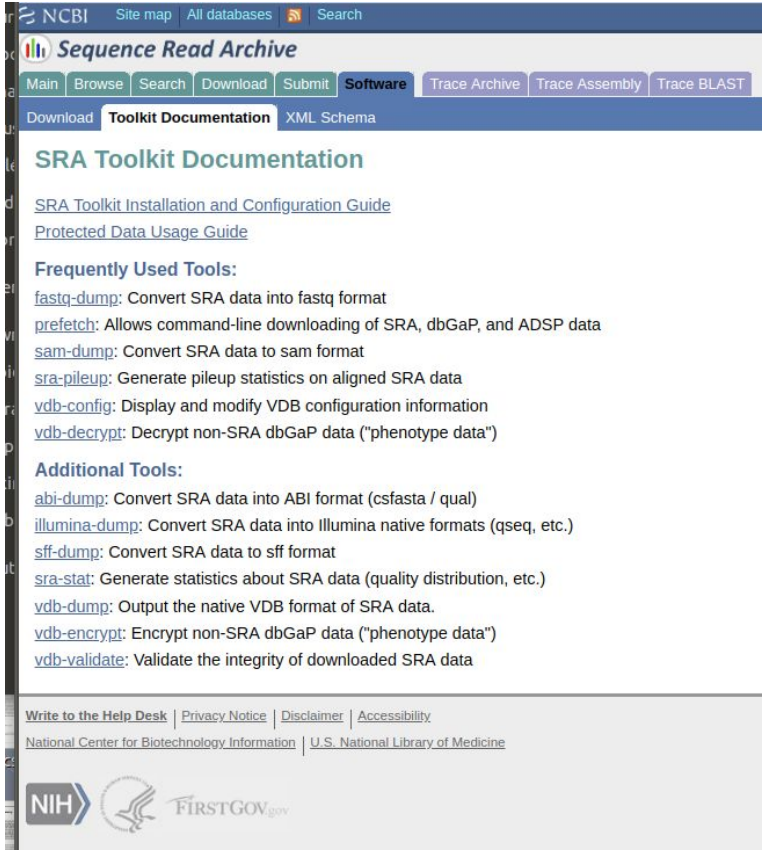


NCBI - SRA: Fetch FASTQ



NCBI - SRA Toolkit : Fetch FASTQ





NCBI - SRA Toolkit : prefetech

SRA Toolkit Documentation

Back to List of the Tools

Tool: prefetch

Usage:

```
prefetch [options] <path/SRA file | path/kart file> [<path/file> ...]
prefetch [options] <SRA accession>
prefetch [options] --list <kart_file>
```

Frequently Used Options:

General:

`-h | --help`	Displays ALL options, general usage, and version information.
`-V | --version`	Display the version of the program.

Data transfer:

`-f | --force <value>`	Force object download. One of: no, yes, all. no [default]: Skip download if the object if found and complete; yes: Download it even if it is found and is complete; all: Ignore lock files (stale locks or if it is currently being downloaded: use at your own risk!).
`--transport <value>`	Value one of: ascp (only), http (only), both (first try ascp, fallback to http). Default: both.
`-l | --list`	List the contents of a kart file.
`-s | --list-sizes`	List the content of kart file with target file sizes.
`-N | --min-size <size>`	Minimum file size to download in KB (inclusive).
`-X | --max-size <size>`	Maximum file size to download in KB (exclusive). Default: 20G.
`-o | --order <value>`	Kart prefetch order. One of: kart (in kart order), size (by file size: smallest first). default: size.
`-a | --ascp-path <ascp-binary|private-key-file>`	Path to ascp program and private key file (asperaweb_id_dsa.openssh).
`-p | --progress <value>`	Time period in minutes to display download progress (0: no progress). Default: 1.
`--option-file <file>`	Read more options and parameters from the file.

Use examples:

`prefetch cart_0.krt`
Download the files listed in the kart file.

`prefetch -l cart_0.krt`
Lists the contents of the kart file.

`prefetch -X 200G cart_0.krt`
Sets the maximum download file size to 200GB and downloads the files listed in the kart.

`prefetch -o kart cart_0.krt`
Downloads the contents in the order listed in the kart. Preferred for large run sets (example: 100+) where calculating the download sizes may cause a delay to the start of downloads.

`prefetch -a "/opt/aspera/bin/ascp|/opt/aspera/etc/asperaweb_id_dsa.openssh" SRR390728`
When the toolkit is unable to locate an installed version of Aspera, the location of ascp and ssh key (-a /opt/aspera/bin/ascp|/opt/aspera/bin/asperaweb_id_dsa.openssh") can be provided.

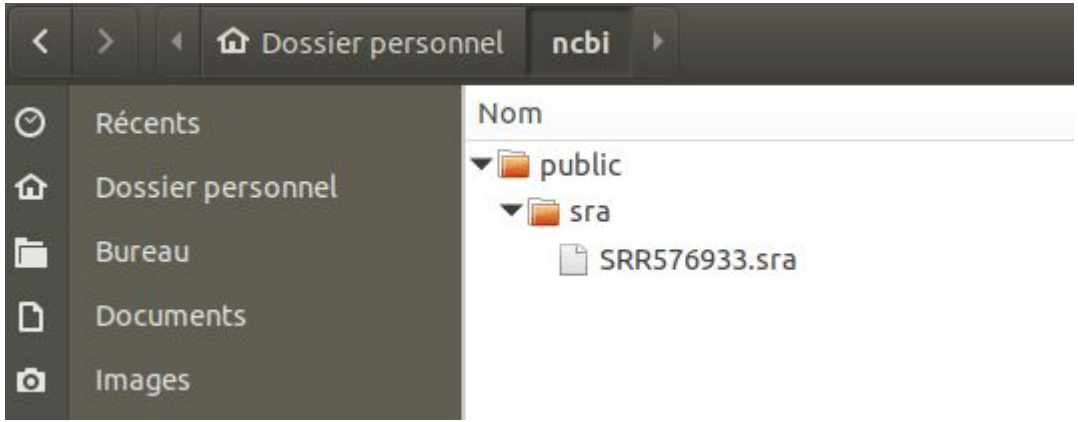
`prefetch -t ascp -a "/opt/aspera/bin/ascp|/opt/aspera/bin/asperaweb_id_dsa.openssh" --option-file file.txt`
Will force download to be only through aspera (-t ascp) and will prevent http download, default operation is to attempt ascp first and use http if Aspera is not found or fails. Will sequentially download the SRA data files and references required for a list of accessions in "file.txt". The format for "file.txt" is a newline-separated list of accessions: SRR# SRR# SRR# …

`prefetch ~/Downloads/SRR390728.sra`
If you have already downloaded an SRA datafile (example here: SRR390728.sra, present in the "~/Downloads" directory), this command will retrieve all of the reference sequences required to extract the data. This command is useful for resolving errors of the type "name not found while resolving tree" - meaning that a reference(s) is required, but cannot be located.

`prefetch -c SRR390728`
This command will check the availability of all needed reference sequences (-c) for a given accession.

NCBI - SRA Toolkit : prefetech

```
(ngs) olivier@dellolinux:~$ prefetch -c SRR576933

2019-01-21T15:42:26 prefetch.2.9.1: 1) Downloading 'SRR576933'...
2019-01-21T15:42:26 prefetch.2.9.1:  Downloading via https...
2019-01-21T15:42:56 prefetch.2.9.1: 1) 'SRR576933' was downloaded successfully
```



NCBI - SRA Toolkit : prefetech set up

Main | Browse | Search | Download | Submit | **Software** | Trace Archive | Trace Assembly | Trace BLAST

Download | **Toolkit Documentation** | XML Schema

SRA Toolkit Documentation

Back to List of the Tools

Tool: vdb-config

Usage:

```
vdb-config [options] [<query>]
```

Frequently Used Options:

-a | --all	print all information [default]
-p | --cfg	print current configuration
-f | --files	print user configuration files in use
-e | --env	print shell variables
-m | --modules	print external modules
-s | --set <name=value>	set configuration node value
--import <ngc-file>	import ngc file
-o | --output <x | n>	output type: one of (x n), where 'x' is xml (default), 'n' is native
--proxy <uri[:port]>	set HTTP proxy server configuration
--proxy-disable <yes | no>	enable/disable using HTTP proxy
--root	enforce configuration update while being run by superuser
-h | --help	Output brief explanation for the program
-V | --version	Display the version of the program then quit
--option-file <file>	Read more options and parameters from the file

Use examples:

`vdb-config -i`
Runs the configuration tool in interactive mode. Usage instructions with pictures are available on the SRA Toolkit Installation and Configuration Guide. VT* and recommended to enable color in the terminal. Users having display problems with the interactive tool are recommended to exit interactive mode by pressing th for configuration.

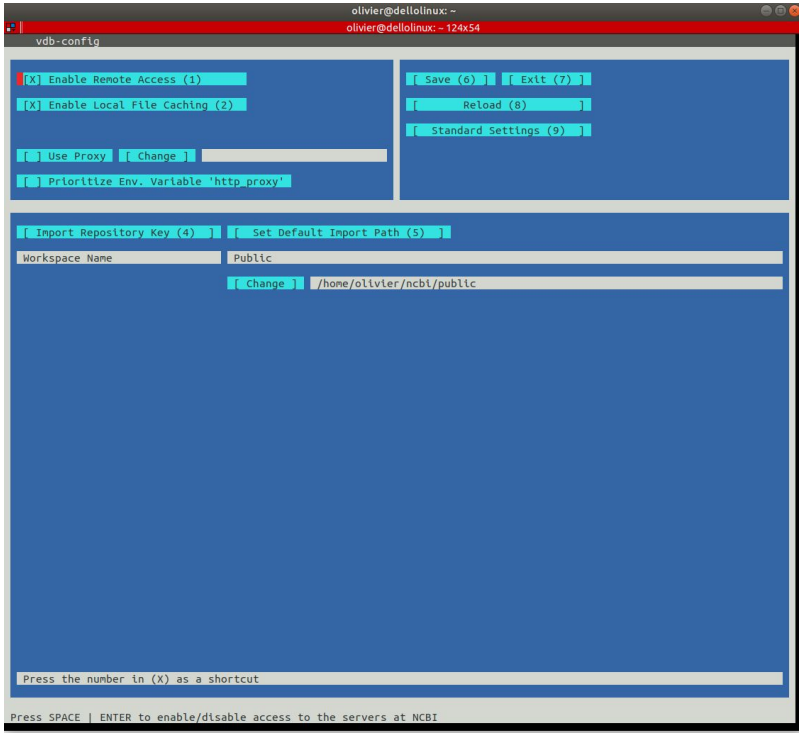
`vdb-config --interactive-mode textual`
Runs the configuration tool in textual mode. Common configuration commands are selected using the keyboard in this mode.

`vdb-config --import prj_1234.ngc`
Imports a dbGaP repository key (prj_1234.ngc) by command line.

`vdb-config --set repository/user/main/public/root=/path/to/new/location`
Moves the user repository storing reference sequence files to a different location.

vdb-config v2.5x

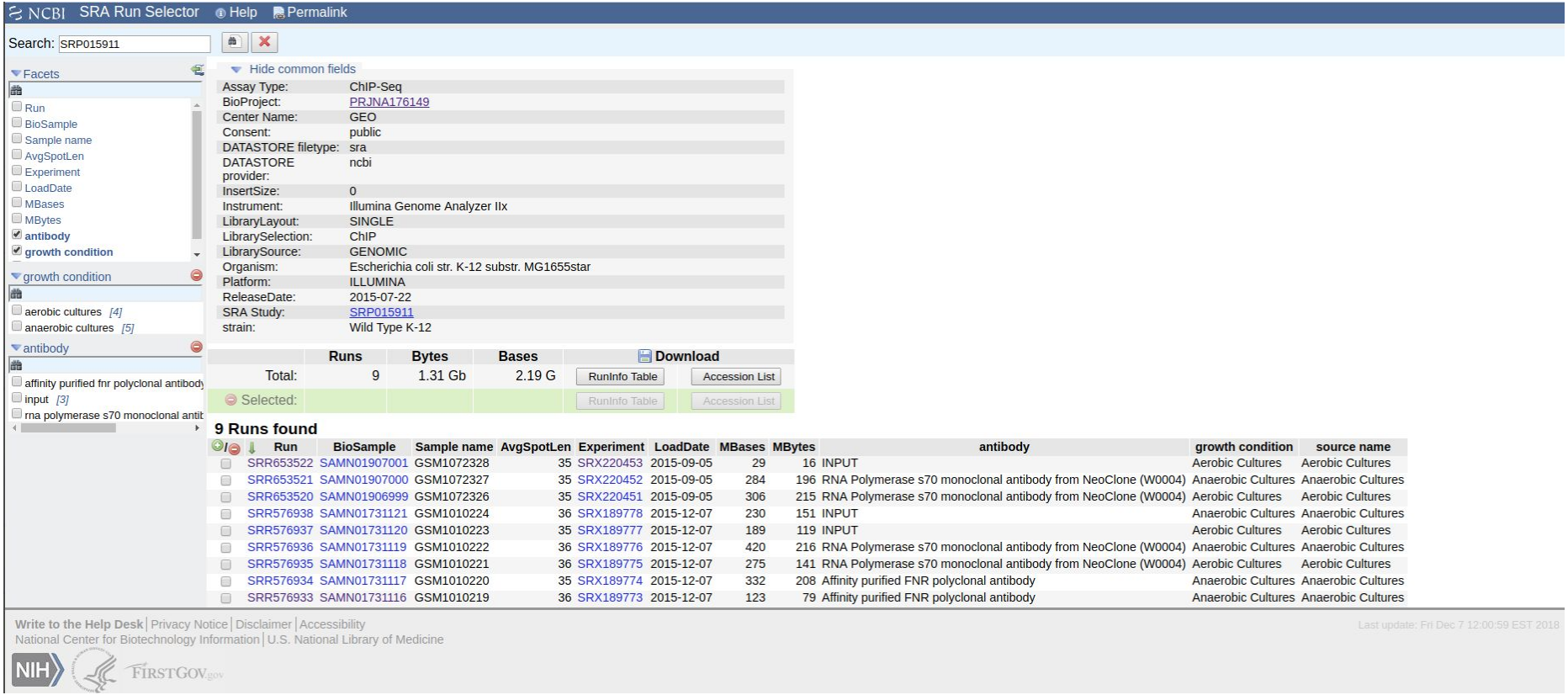
```
(ngs) olivier@dellolinux:~$
(ngs) olivier@dellolinux:~$ vdb-config -i
```



NCBI - SRA run selector: Fetch FASTQs

search for SRP, GSE, GSM, ...

NCBI - SRA run selector: Fetch FASTQs



Select anaerobic samples & IP FNR, then download “Runinfo Table” & “Accession List”

NCBI - SRA run selector : Fetch FASTQs

SraRunTable.txt

AvgSpotLen	BioSample	Experiment	LoadDate	MBases	MBytes	Run	SRA_Sample	Sample_Name	antibody	growth_condition	source_name	Assay_Type	BioProject	Center_Name	Consent	DATASTORE_filetype	DATASTORE_provider	InsertSize	Instrun
36	SAMN01731116	SRX189773	2015-12-07	123	79	SRR576933	SRS365655	GSM1010219	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures	ChIP-Seq	PRJNA176149	GEO	public	sra	ncbi	0	Illumina
35	SAMN01731117	SRX189774	2015-12-07	332	208	SRR576934	SRS365656	GSM1010220	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures	ChIP-Seq	PRJNA176149	GEO	public	sra	ncbi	0	Illumina

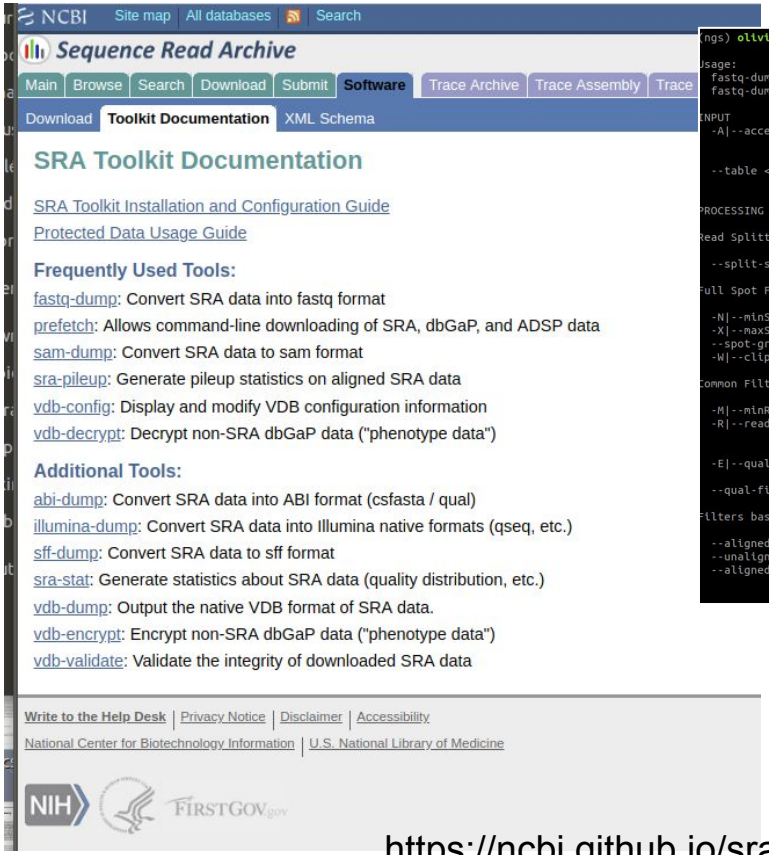
SRR_Acc_List.txt

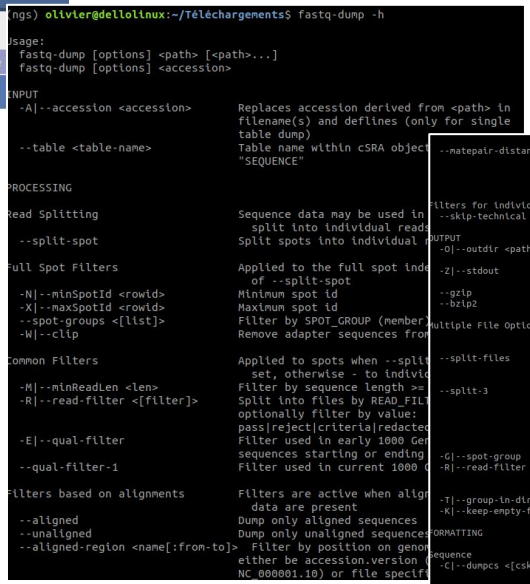
NCBI - SRA Tool Kit : prefetch more than 1 file

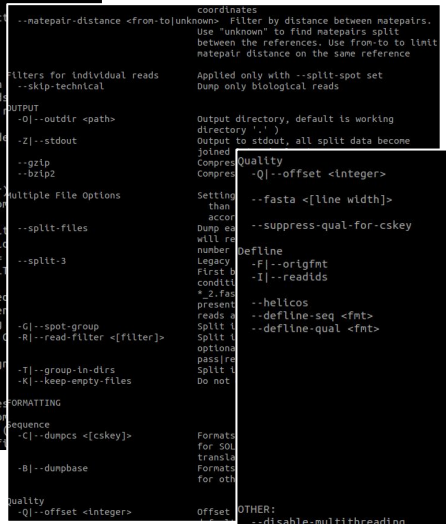
```
(ngs) olivier@dellolinux:~/Téléchargements$ cat SRR_Acc_List.txt | xargs -n 1 prefetch

2019-01-21T15:54:19 prefetch.2.9.1: 1) Downloading 'SRR576934'...
2019-01-21T15:54:19 prefetch.2.9.1:  Downloading via https...
2019-01-21T15:55:10 prefetch.2.9.1: 1) 'SRR576934' was downloaded successfully

2019-01-21T15:55:11 prefetch.2.9.1: 1) 'SRR576933' is found locally
```

Magic!!

NCBI - SRA Tool Kit : SRA to FASTQ, fastq-dump







https://ncbi.github.io/sra-tools/fastq-dump.html

NCBI - SRA Tool Kit : SRA to FASTQ, fastq-dump

```
(ngs) olivier@dellolinux:~/ncbi/public/sra$ fastq-dump *.sra
Read 3603544 spots for SRR576933.sra
Written 3603544 spots for SRR576933.sra
Read 9946604 spots for SRR576934.sra
Written 9946604 spots for SRR576934.sra
Read 13550148 spots total
Written 13550148 spots total
```

```
(ngs) olivier@dellolinux:~/ncbi/public/sra$ wc -l *.fastq
  14414176 SRR576933.fastq
  39786416 SRR576934.fastq
  54200592 total
```

```
(ngs) olivier@dellolinux:~/ncbi/public/sra$ ls -ltrh
total 2,8G
-rw-rw-r-- 1 olivier olivier  80M janv. 21 16:42 SRR576933.sra
-rw-rw-r-- 1 olivier olivier 209M janv. 21 16:55 SRR576934.sra
-rw-rw-r-- 1 olivier olivier 1,9G janv. 21 17:10 SRR576934.fastq
-rw-rw-r-- 1 olivier olivier 693M janv. 21 17:10 SRR576933.fastq
```

```
(ngs) olivier@dellolinux:~/ncbi/public/sra$ head *.fastq
==> SRR576933.fastq <==
@SRR576933.1 HWUSI-EAS1789_0000:2:20:1269:14140 length=36
AAGCATGGAATAACCGCCTGGTGAATGCTCGCCATA
+SRR576933.1 HWUSI-EAS1789_0000:2:20:1269:14140 length=36
EDEA=EEEEEBFBDFDEBDAD=DDBAC5CCEEE:E@
@SRR576933.2 HWUSI-EAS1789_0000:2:20:1270:19579 length=36
TGGAGGCTGACCACGATAAGCTGCCGCTGGTGGTGC
+SRR576933.2 HWUSI-EAS1789_0000:2:20:1270:19579 length=36
EDF:D?=DEEE?EEE5DDDAEB:ECEBBEA>AA956
@SRR576933.3 HWUSI-EAS1789_0000:2:20:1270:17351 length=36
AGTGCGATGCCGTTCACCCGGTTTTCTTTATCATTA

==> SRR576934.fastq <==
@SRR576934.1 HWUSI-EAS787_0001:1:1:16343:952 length=35
NGATCAAAAACAAAAAATATTTCACTCGANAGGAG
+SRR576934.1 HWUSI-EAS787_0001:1:1:16343:952 length=35
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR576934.2 HWUSI-EAS787_0001:1:1:17465:951 length=35
NTGACGTAAAATCTTGTAGAGGCCAACGCNCATAA
+SRR576934.2 HWUSI-EAS787_0001:1:1:17465:951 length=35
%/2//77787@CCC@@@5@@@@@@@%%%%%%%%%%
@SRR576934.3 HWUSI-EAS787_0001:1:1:16452:962 length=35
```

2 Runs found

Run	BioSample	Sample name	AvgSpotLen	Experiment	LoadDate	MBases	MBytes	antibody	growth condition	source name
SRR576934	SAMN01731117	GSM1010220	35	SRX189774	2015-12-07	332	208	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures
SRR576933	SAMN01731116	GSM1010219	36	SRX189773	2015-12-07	123	79	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures

NCBI - SRA Tool Kit : SRA to FASTQ, fastq-dump

Tool: fastq-dump

Download ZIP File | Download TAR Ball | View On GitHub

Usage

fastq-dump [options] < path/file > [< path/file > ...]
fastq-dump [options] < accession >

Frequently Used Options

General

-h | --help	Displays ALL options, general usage, and version information.
-V | --version	Display the version of the program.

Data Formatting

--split-files	Dump each read into separate file. Files will receive suffix corresponding to read number.
--split-spot	Split spots into individual reads.
--fasta < [line width] >	FASTA only, no qualities. Optional line wrap width (set to zero for no wrapping).
-I | --readids	Append read id after spot id as 'accession.spot.readid' on defline.
-F | --origfmt	Defline contains only original sequence name.
-C | --dumpcs < [cskey] >	Formats sequence using color space (default for SOLiD). "cskey" may be specified for translation.
-B | --dumpbase	Formats sequence using base space (default for other than SOLiD).
-Q | --offset < integer >	Offset to use for ASCII quality scores. Default is 33 ("!").

Filtering

-N | --minSpotId < rowid >	Minimum spot id to be dumped. Use with "X" to dump a range.
-X | --maxSpotId < rowid >	Maximum spot id to be dumped. Use with "N" to dump a range.
-M | --minReadLen < len >	Filter by sequence length >= < len >
--skip-technical	Dump only biological reads.
--aligned	Dump only aligned sequences. Aligned datasets only; see sra-stat.
--unaligned	Dump only unaligned sequences. Will dump all for unaligned datasets.

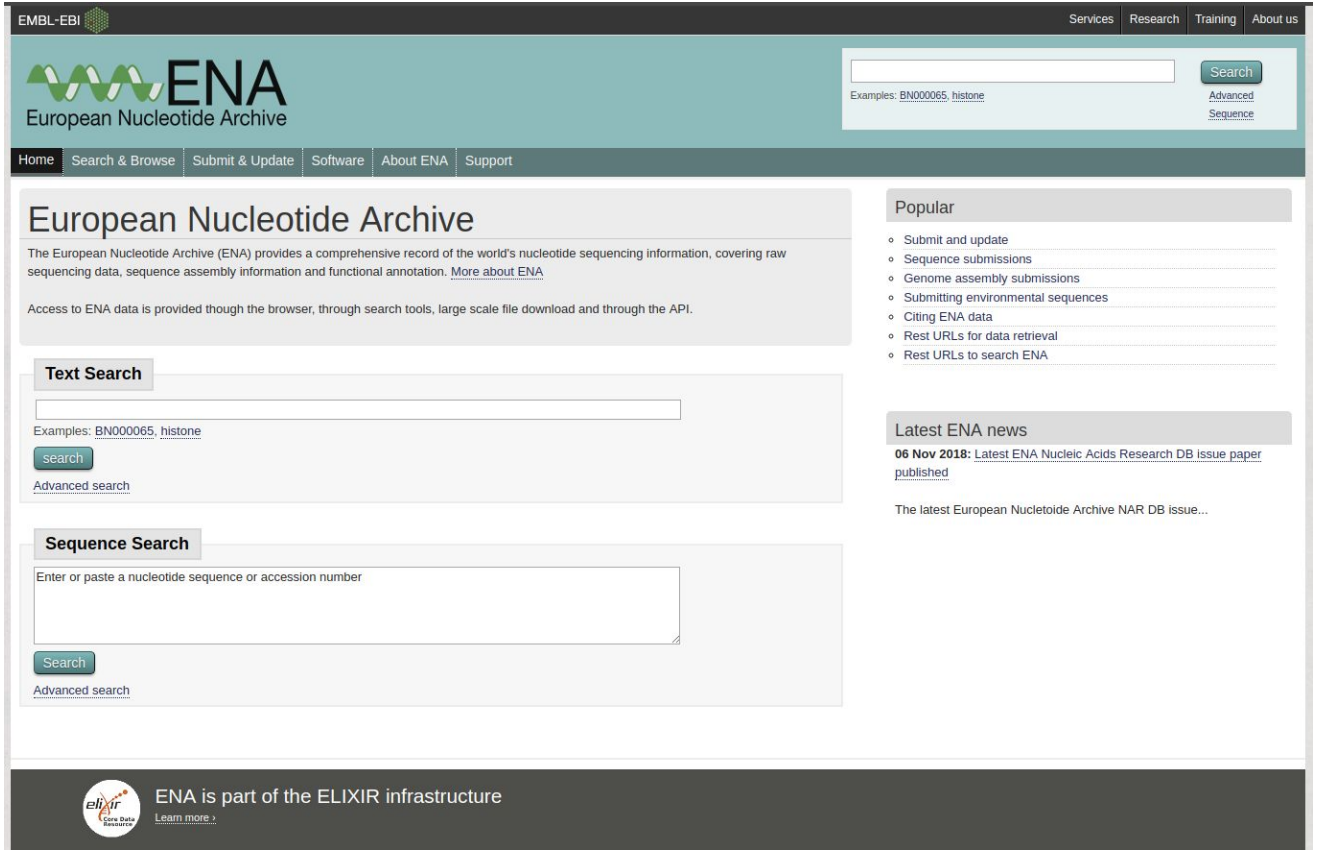
Workflow and piping

-O | --outdir < path >	Output directory, default is current working directory ('.').
-Z | --stdout	Output to stdout, all split data become joined into single stream.
--gzip	Compress output using gzip.
--bzip2	Compress output using bzip2.

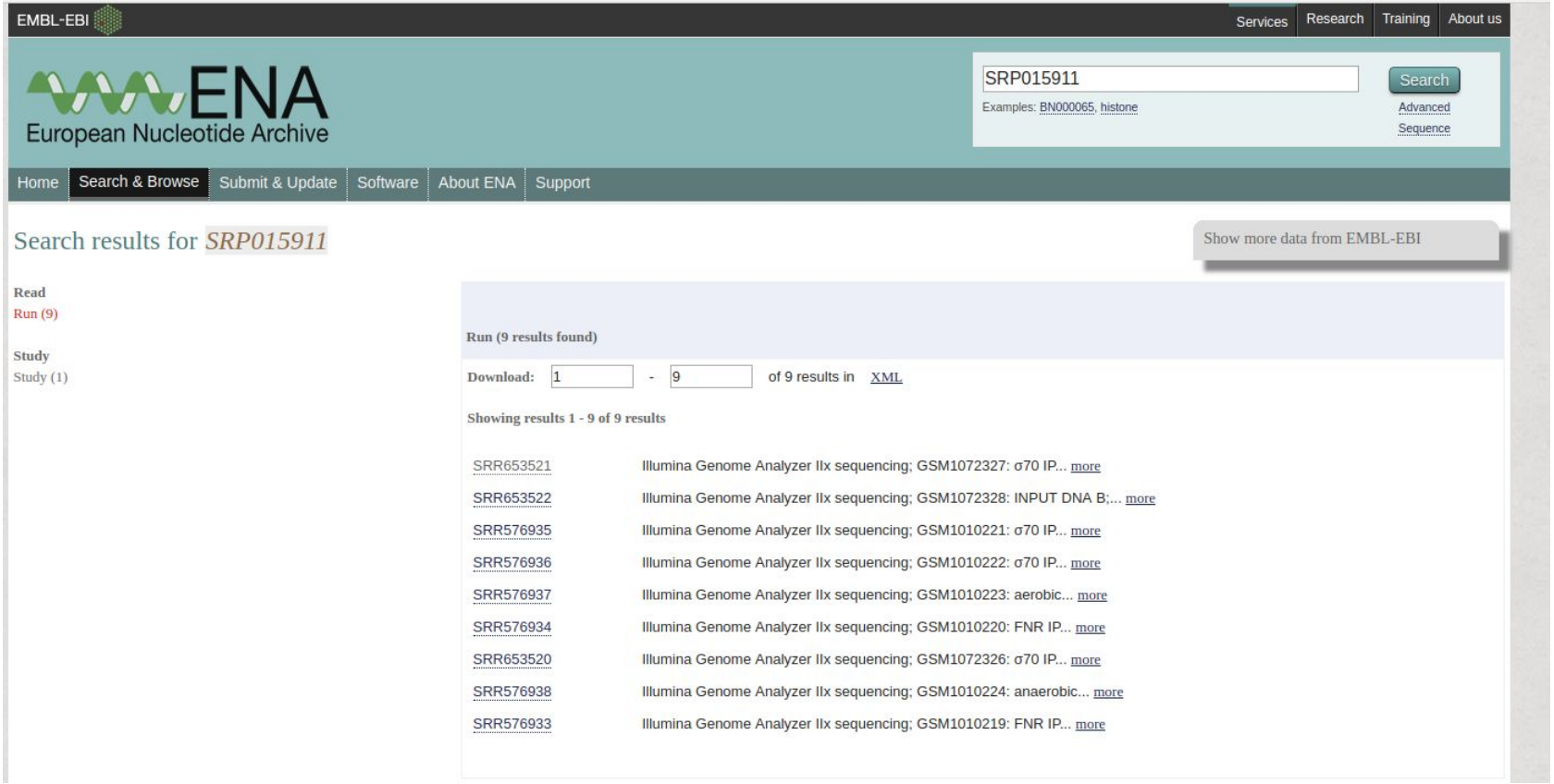
Attention “RTFM”

Check if you have Single or Paired End data

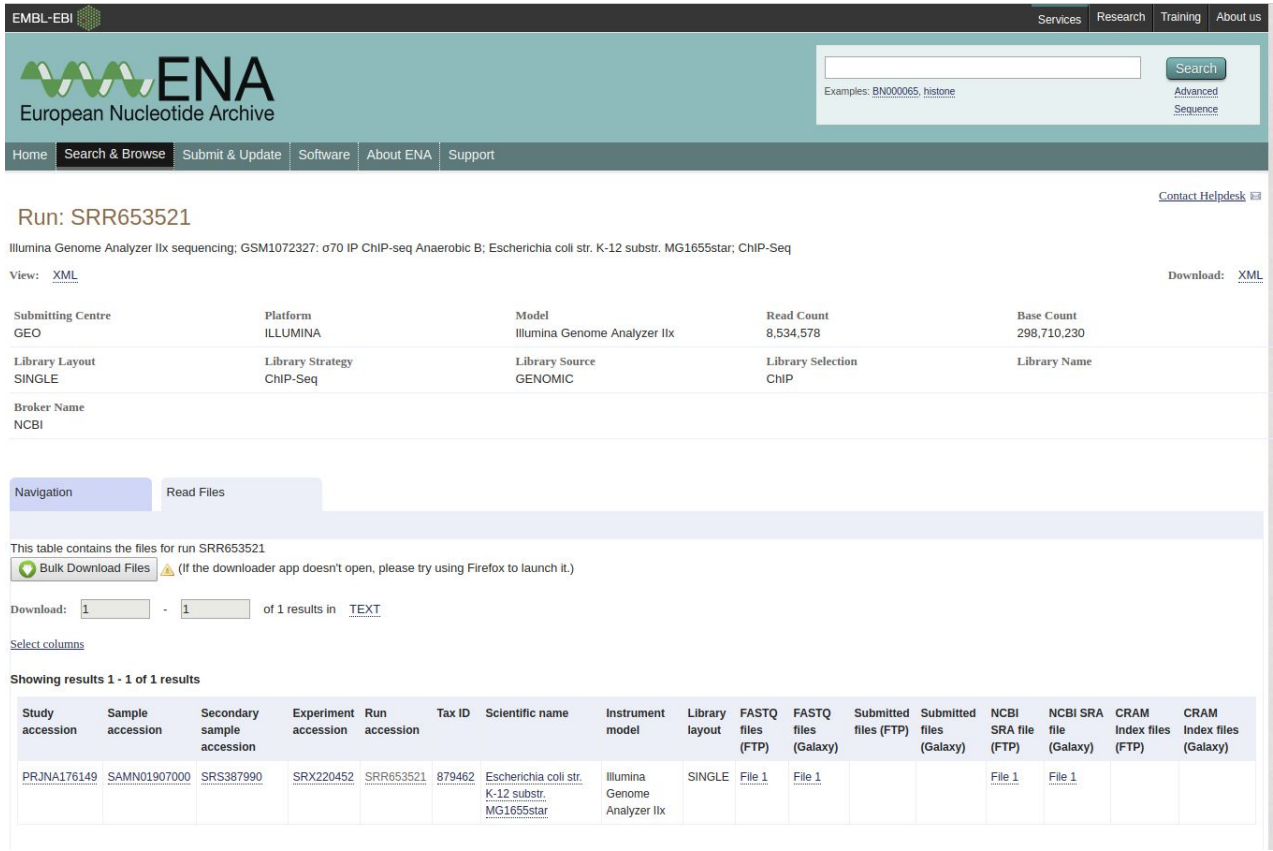
Alternative Data base in case of shut down!



ENA: European Nucleotide Archive



ENA: European Nucleotide Archive



FASTQ : Boites à outils

- FASTX ToolKit (http://hannonlab.cshl.edu/fastx_toolkit/)
- Trim_Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
- Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)

Renommer, trimmer, filtrer et plus encore!

https://en.wikipedia.org/wiki/FASTQ_format

FASTQ : Boites à outils….

- Cutadapt (https://cutadapt.readthedocs.io/en/latest/guide.html)
- prinseq(http://prinseq.sourceforge.net/manual.html)
- tally (https://www.ebi.ac.uk/~stijn/reaper/tally.html)
- reaper (http://www.ebi.ac.uk/~stijn/reaper/reaper.html)
- seqtk (https://github.com/lh3/seqtk)
- deconseq (http://deconseq.sourceforge.net/)
- SeqPrep (https://github.com/jstjohn/SeqPrep)

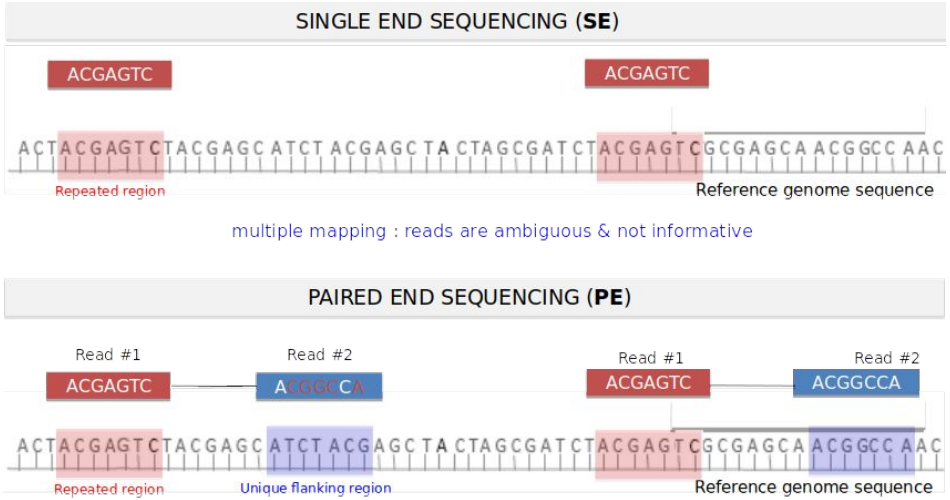
Suite et boites à outils généralistes

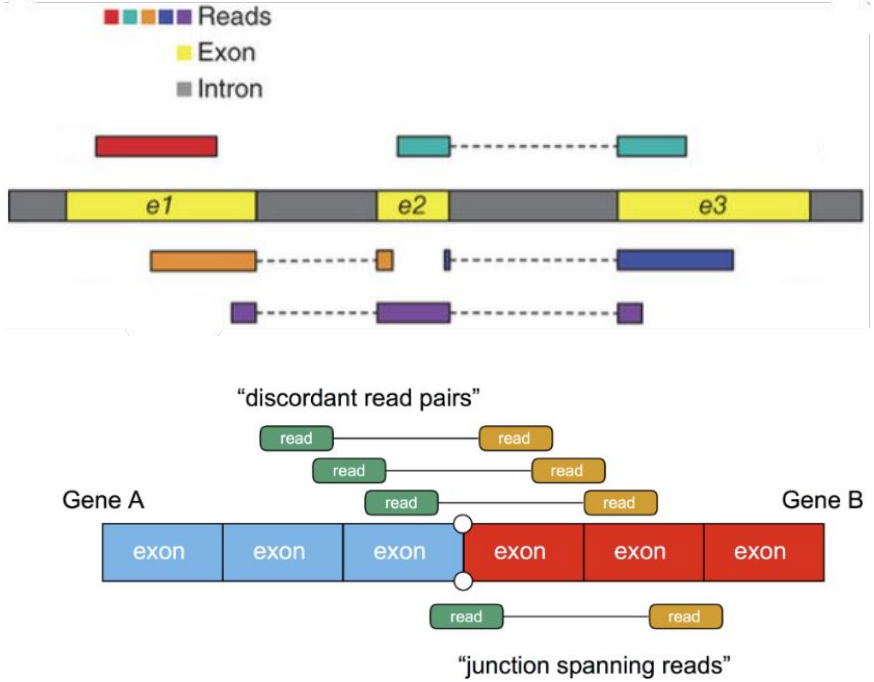
- GATK (https://software.broadinstitute.org/gatk/)
- Picard (https://broadinstitute.github.io/picard/)
- ….

https://en.wikipedia.org/wiki/FASTQ_format

Alignement des séquences : FASTQ -> SAM/BAM

= assigner des coordonnées chromosomiques et décrire comment celles ci sont déterminées





SAM/BAM : Sequence Alignment/Map

1.Header

```
@HD         VN:1.0      SO:unsorted
@SQ         SN:chr1     LN:197195432
@SQ         SN:chr2     LN:181748087
@SQ         SN:chr3     LN:159599783
@PG         ID:Bismark  VN:v0.14.4   CL:"bismark -N 1
/Users/Mezger/methylome/genome-mm9/ --path_to_bowtie /Users/bowtie2-2.1.0 --
non_directional --un -o /Volumes/align-mm9/1M -1 c10t_1M_R1.fastq -2 c10t_1M_R2.fastq"
```

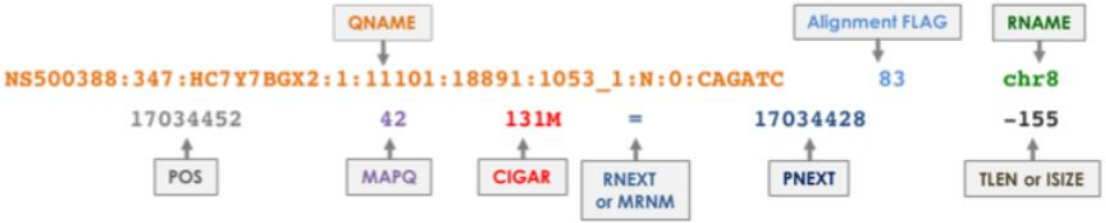
```
NS500388:347:HC7Y7BGX2:1:11101:18891:1053_1:N:0:CAGATC          9          chr8
          17034452    42          131M        =          17034428    -155
          TTAATTCACATCCATATACTAAAATCATAAAATACCTACAACCTAAAACTTAATCCTTAAACATCTTTCTCACA
ATAACTACTCCTTTAATCTTAACCAAATCTTCCTACTCCTATAATTTCCCTAAACTA
          EEA</EEEEEEEAEA<E/EEEEE/EAEE/<AEEAEEEAEA6EEEAA<<EEE<<EEEEEAE<EAAEEEEEEE6EE
EEEAEEEEEEEEEEEAAEEEAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEA/EEE
NM:i:23     MD:Z:2G0G13G2G1G6G0G2G3G6G0G0G4G0G6G17G2G7G19G5G2G9G2G0
          XM:Z:..hh.............h..x.h......hh..h...x......xhh....hh......h.........
........h..x.......h..................x.....x..h.........h..x     XR:Z:CT     XG:Z:GA
```

2.Body (reads alignment per se)

- metadata, at the beginning of the file
- each row starts with '@' :
  - **@HD** : VN : format version, SO: sorting order alignment
  - **@SQ**: names of all reference chromosomes used and their length
  - **@PG**: used mapping programs and command lines
  - **@RG** (readgroup): sample information = flow cell, lane, population, status…

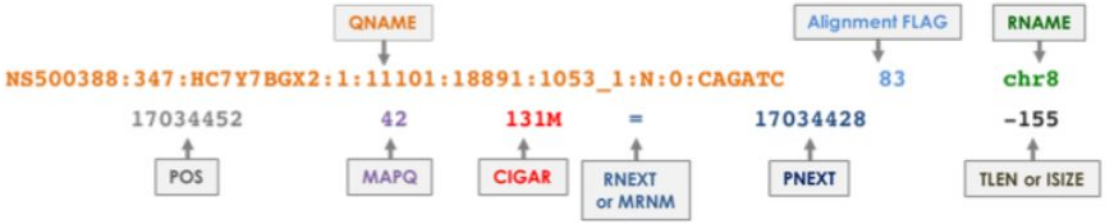
COLUMN	NAME	CONTENT	
1	QNAME	NS500388:347:HC7Y7BGX2:1:1110…	Mandatory
2	FLAG	83	
3	RNAME	chr8	
4	POS	17034452	
5	MAPQ	42	
6	CIGAR	131M	
7	RNEXT	=	
8	PNEXT	17034428	
9	TLEN	-155	
10	SEQ	TTAATTCACATCCATATAC …	
11	QUAL	EEA</EEEEEEEAEA<E	
12	NM-tag	NM:i:23	Optional
13	MD-tag	MD:Z:2G0G13G2G1G6 …	
14	XM-tag	XM:Z:..hh..... …	
15	XR-tag	XR:Z:CT	
16	XG-tag	XG:Z:GA	

SAM/BAM : Sequence Alignment/Map



1	QNAME	ID of the reads
2	FLAG	Alignment flag (Combination of bitwise FLAGs)
3	RNAME	Reference name (Chromosome name)
4	POS	Position in reference (5' position)
5	MAPQ	Mapping quality
6	CIGAR	Alignment description in CIGAR format (M:match ; I:insertion ; D:deletion ; S:substitution)
7	RNEXT or MRNM	Name of reference sequence where mate's alignment occurs. Set to = if the mate's reference is the same as this alignment, or * if there is no mate
8	PNEXT	Left most position of where the next alignment in this gap maps to the reference
9	TLEN or ISIZE	Total length or Insert size (for PE)

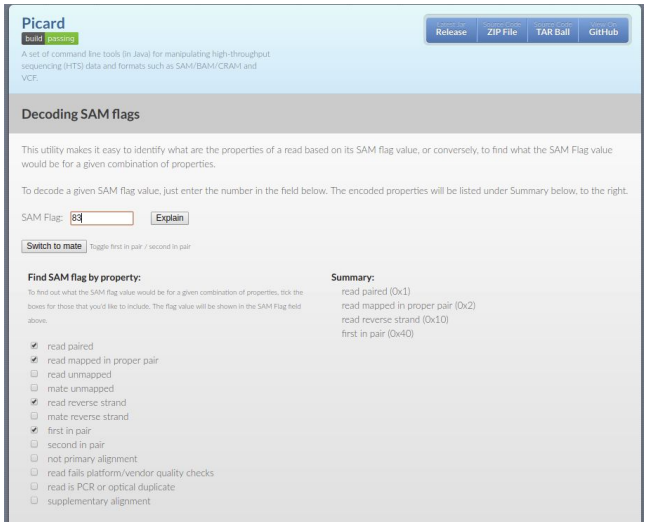
SAM/BAM : Sequence Alignment/Map



1	QNAME	ID of the reads
2	FLAG	Alignment flag (Combination of bitwise FLAGs)

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

https://broadinstitute.github.io/picard/explain-flags.html



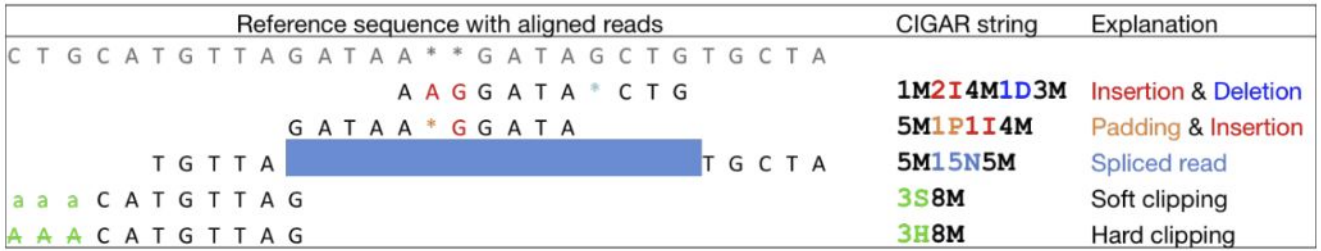
SAM/BAM : Sequence Alignment/Map

`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.

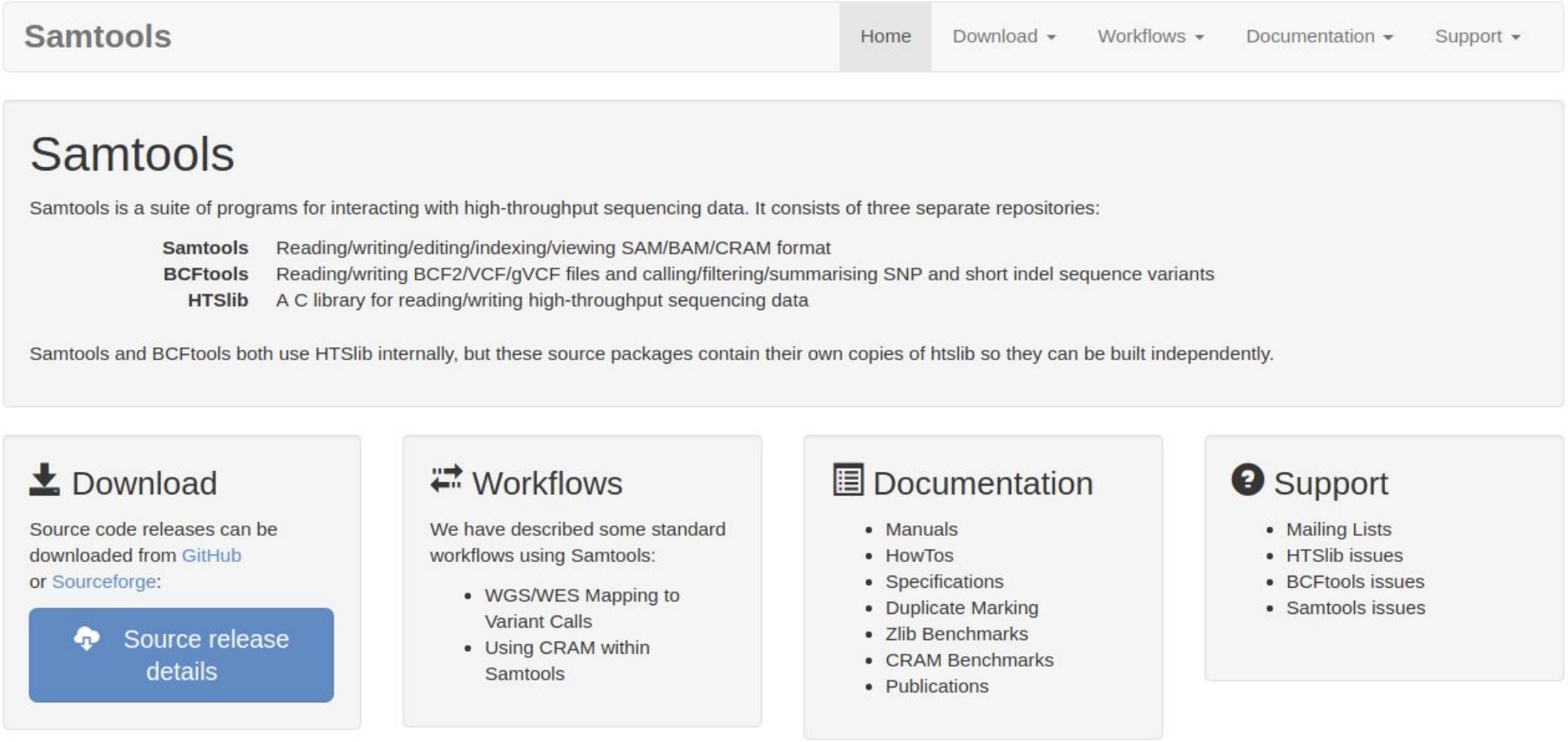
The following operations are defined in CIGAR format (also see figure below):

- **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:



Outils de manipulation des fichiers SAM/BAM



SAMTOOLS

http://www.htslib.org/doc/#manual-pages

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     fqidx          index/extract FASTQ
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags
     markdup        mark duplicates

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
     merge          merge sorted alignments
     mpileup        multi-way pileup
     sort           sort alignment file
     split          splits a file by read group
     quickcheck     quickly check if SAM/BAM/CRAM file appears intact
     fastq          converts a BAM to a FASTQ
     fasta          converts a BAM to a FASTA

  -- Statistics
     bedcov         read depth per BED region
     depth          compute the depth
     flagstat       simple stats
     idxstats       BAM index stats
     phase          phase heterozygotes
     stats          generate stats (former bamcheck)

  -- Viewing
     flags          explain BAM flags
     tview          text alignment viewer
     view           SAM<->BAM<->CRAM conversion
     depad          convert padded BAM to unpadded BAM
```

SAMTOOLS basic operation

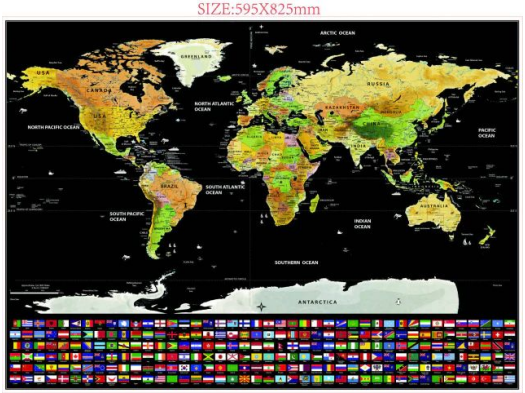
```
(ngs) olivier@dellolinux:~$ samtools view

Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]

Options:
  -b       output BAM
  -C       output CRAM (requires -T)
  -1       use fast BAM compression (implies -b)
  -u       uncompressed BAM output (implies -b)
  -h       include header in SAM output
  -H       print SAM header only (no alignments)
  -c       print only the count of matching records
  -o FILE  output file name [stdout]
  -U FILE  output reads not selected by filters to FILE [null]
  -t FILE  FILE listing reference names and lengths (see long help) [null]
  -L FILE  only include reads overlapping this BED FILE [null]
  -r STR   only include reads in read group STR [null]
  -R FILE  only include reads with read group listed in FILE [null]
  -q INT   only include reads with mapping quality >= INT [0]
  -l STR   only include reads in library STR [null]
  -m INT   only include reads with number of CIGAR operations consuming
           query sequence >= INT [0]
  -f INT   only include reads with all  of the FLAGs in INT present [0]
  -F INT   only include reads with none of the FLAGS in INT present [0]
  -G INT   only EXCLUDE reads with all  of the FLAGs in INT present [0]
  -s FLOAT subsample reads (given INT.FRAC option value, 0.FRAC is the
           fraction of templates/read pairs to keep; INT part sets seed)
  -M       use the multi-region iterator (increases the speed, removes
           duplicates and outputs the reads as they are ordered in the file)
  -x STR   read tag to strip (repeatable) [null]
  -B       collapse the backward CIGAR operation
  -?       print long help, including note about region specification
  -S       ignored (input format is auto-detected)
      --input-fmt-option OPT[=VAL]
               Specify a single input file format option in the form
               of OPTION or OPTION=VALUE
  -O, --output-fmt FORMAT[,OPT[=VAL]]...
               Specify output format (SAM, BAM, CRAM)
      --output-fmt-option OPT[=VAL]
               Specify a single output file format option in the form
               of OPTION or OPTION=VALUE
  -T, --reference FILE
               Reference sequence FASTA FILE [null]
  -@, --threads INT
               Number of additional threads to use [0]
```

```
(ngs) olivier@dellolinux:~$ samtools index
Usage: samtools index [-bc] [-m INT] <in.bam> [out.index]
Options:
  -b       Generate BAI-format index for BAM files [default]
  -c       Generate CSI-format index for BAM files
  -m INT   Set minimum interval size for CSI indices to 2^INT [14]
  -@ INT   Sets the number of threads [none]
```

```
(ngs) olivier@dellolinux:~$ samtools sort
Usage: samtools sort [options...] [in.bam]
Options:
  -l INT     Set compression level, from 0 (uncompressed) to 9 (best)
  -m INT     Set maximum memory per thread; suffix K/M/G recognized [768M]
  -n         Sort by read name
  -t TAG     Sort by value of TAG. Uses position as secondary index (or read name if -n is set)
  -o FILE    Write final output to FILE rather than standard output
  -T PREFIX  Write temporary files to PREFIX.nnnn.bam
      --input-fmt-option OPT[=VAL]
               Specify a single input file format option in the form
               of OPTION or OPTION=VALUE
  -O, --output-fmt FORMAT[,OPT[=VAL]]...
               Specify output format (SAM, BAM, CRAM)
      --output-fmt-option OPT[=VAL]
               Specify a single output file format option in the form
               of OPTION or OPTION=VALUE
      --reference FILE
               Reference sequence FASTA FILE [null]
  -@, --threads INT
               Number of additional threads to use [0]
```

Génomes de références??

1800



2019

Mus musculus

- mm7 (Aug. 2005)
- mm8 (Feb. 2006)
- mm9 (July 2007)
- mm10 (Dec. 2011)

Homo sapiens

- NCBI34 - hg16 (July 2004)
- NCBI35 - hg17 (July 2003)
- NCBI 36 - hg18 (Mar. 2006)
- GRGh 37 - hg19 (Feb 2009)
- GRGh38 - hg38 (Dec 2013)

Comment le choisir?

- Idéalement le plus proche de celui de nos échantillons
- En fonction de l'objectif de votre analyse

FASTA file

voir slides de cours

Génomes de références et annotation (GTF/GFF)

Gene Feature Format (GFF/ GFF2/GFF3) (http://gmod.org/wiki/GFF3)

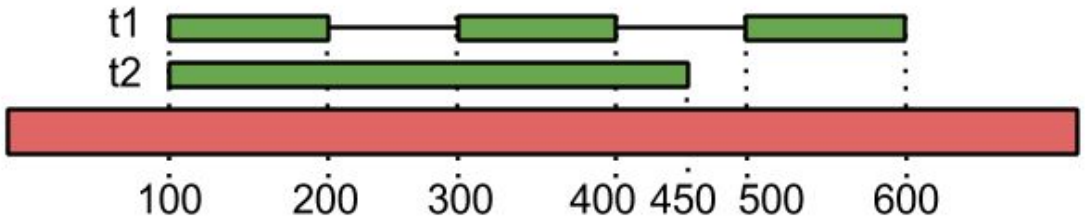
The GFF format is a flat tab-delimited file, each line of which corresponds to an annotation, or feature.

```
ctg123 example gene             1050 9000 . + . ID=EDEN;Name=EDEN;Note=protein kinase
ctg123 example mRNA             1050 9000 . + . ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1
ctg123 example five_prime_UTR   1050 1200 . + . Parent=EDEN.1
ctg123 example CDS              1201 1500 . + 0 Parent=EDEN.1
ctg123 example CDS              3000 3902 . + 0 Parent=EDEN.1
ctg123 example CDS              5000 5500 . + 0 Parent=EDEN.1
ctg123 example CDS              7000 7608 . + 0 Parent=EDEN.1
ctg123 example three_prime_UTR  7609 9000 . + . Parent=EDEN.1
```

Génomes de références et annotation (GTF/GFF)

Gene Transfer Format (GTF / GFF2.2) (http://mblab.wustl.edu/GTF2.html)

```
chr20   example   exon   100   200   .   +   .   gene_id "g1"; transcript_id "t1";
chr20   example   exon   300   400   .   +   .   gene_id "g1"; transcript_id "t1";
chr20   example   exon   500   600   .   +   .   gene_id "g1"; transcript_id "t1";
chr20   example   exon   100   450   .   +   .   gene_id "g1"; transcript_id "t2";
```



Génomes de références et annotation (GTF/GFF)

Gene Transfer Format (GTF / GFF2.2) (http://mblab.wustl.edu/GTF2.html)

Here is an example in which the "exon" feature is used. It is a 5 exon gene with 3 translated exons.

```
AB000381 Twinscan  exon         150   200   .   +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  exon         300   401   .   +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  exon         501   650   .   +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  exon         700   800   .   +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  exon         900  1000   .   +   .  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "AB000381.000"; transcript_id "AB000381.000.1
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "AB000381.000"; transcript_id "AB000381.000.1
```

Résumé des différents format

https://www.france-bioinformatique.fr/sites/default/files/formats.pdf

https://genome.ucsc.edu/ENCODE/fileFormats.html