

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

Olivier Rué & Olivier Kirsh

Slides complémentaires

# NCBI : National Center for Biotechnology Information

## 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



# 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

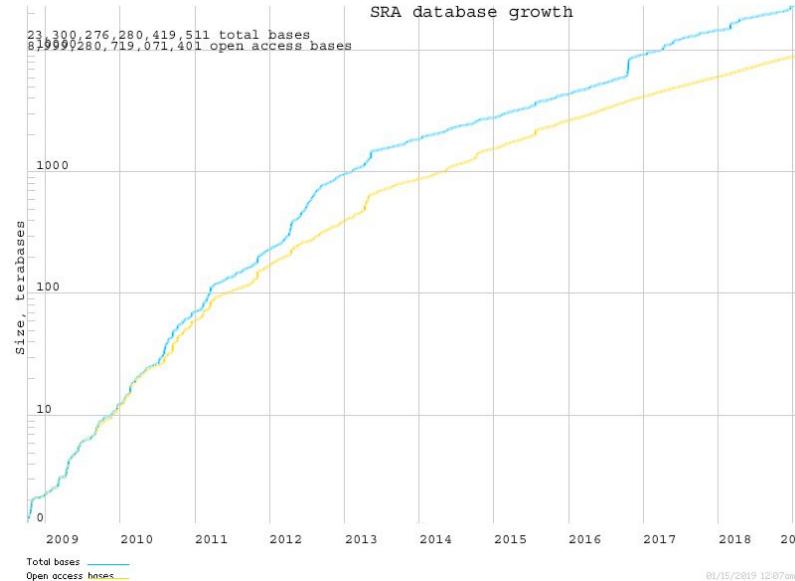
- Entrez Programming Utilities (E-utilities) <https://www.ncbi.nlm.nih.gov/books/NBK25497/>
  - Requêteage de la BD du NCBI via 9 E-utilities (ESearch, EFetch...)
- NCBI-Blast+
  - outil téléchargeable
  - API Rest (intégration dans des applications)
- Browsers
  - 1000 genomes Browser
- Une multitude 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 retrieve public datasets

NCBI Resources How To

PubMed Advanced Search 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<sup>1</sup>, Yan H, Ong JM, Chung D, Liang K, Tran F, Keles 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

OPEN ACCESS TO FULL TEXT PLOS GENETICS PMC Full text

Save items

Add to Favorites ▾

Similar articles

Genome-wide expression analysis indicates that FNR of *Escherichia coli* K-12 r<sup>t</sup> [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<sup>v</sup> [J Biochem. 1996]

See reviews... See all...

f t g

# NCBI - GEO DataSets: Fetch FASTQ

NCBI Resources How To

GEO Home Documentation Query & Browse Email GEO okirish My NCBI Sign Out My GEO Submissions

## Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

 Gene Expression Omnibus

Keyword or GEO Accession

Getting Started	Tools	Browse Content
Overview	Search for Studies at GEO DataSets	Repository Browser
FAQ	Search for Gene Expression at GEO Profiles	DataSets: 4348
About GEO DataSets	Search GEO Documentation	Series:  107758
About GEO Profiles	Analyze a Study with GEO2R	Platforms: 19313
About GEO2R Analysis	Studies with Genome Data Viewer Tracks	Samples: 2845722
How to Construct a Query	Programmatic Access	
How to Download Data	FTP Site	

### Information for Submitters

My GEO Submissions	Submission Guidelines	MIAME Standards
My GEO Profile	Update Guidelines	Citing and Linking to GEO
		Guidelines for Reviewers
		GEO Publications

| NLM | NIH | Email GEO | Disclaimer | Accessibility |

# NCBI - GEO DataSets: Fetch FASTQ

NCBI Resources How To

GEO DataSets GEO DataSets ▾ Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor bii Search

Create alert Advanced Help

Summary ▾ Send to: ▾

**Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding**

(Submitter supplied) This SuperSeries is composed of the SubSeries listed below.

Organism: **Escherichia coli** K-12; **Escherichia coli** str. K-12 substr. MG1655star; **Escherichia coli** str. K-12 substr. MG1655

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

Platforms: GPL14649 GPL16109 GPL8708 48 Samples

Download data: CALLS, GFF, PAIR, TXT, WIG

Series Accession: GSE41195 ID: 200041195

[PubMed](#) [Full text in PMC](#) [Similar studies](#) [Analyze with GEO2R](#)

**Related information**

PubMed  
Full text in PMC  
Similar studies  
BioProject  
Taxonomy

**Search details**

Genome-scale[All Fields] AND analysis[All Fields] AND ("escherichia coli"[MeSH Terms]  
OR "Escherichia coli"[Organism]  
OR escherichia coli[All Fields])

Search See more...

# NCBI - GEO DataSets: Fetch FASTQ

**NCBI** **GEO**  
Gene Expression Omnibus

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

Contact: okirsh [at] | My submissions [at] | Sign Out [at]

Scope: Set ▾ 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 organism: *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 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 2012  
 Last update date: Jul 2013  
 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)  
 GPL14449 NimbleGen E. coli K12 Gene Expression Array [071112\_Ecoli\_K12\_EXP]  
 GPL11609 illumina Genome Analyzer Iix (Escherichia coli str. K-12 substr. MG1655star)  
 Samples (48):  
 GSM1010199 FNR - Anaerobic - A  
 GSM1010200 FNR - Anaerobic - B  
 GSM1010201 FNR - Anaerobic - C  
 GSM1010202 B - Anaerobic - A  
 GSM1010203 B - Anaerobic - B  
 GSM1010204 B - Aerobic - A  
 GSM1010205 B - Aerobic - B  
 GSM1010206 HNS - Anaerobic A  
 GSM1010207 HNS - Aerobic A  
 GSM1010208 IHF - Anaerobic A  
 GSM1010209 Δfnr - Anaerobic  
 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 θ70 IP ChIP-seq Aerobic A  
 GSM1010222 θ70 IP ChIP-seq Aerobic A  
 GSM1010223 aerobic INPUT DNA  
 GSM1010224 anaerobic INPUT DNA  
 GSM1010240 Ecoli\_wild-type\_repl\_Anaerobic  
 GSM1010241 Ecoli\_wild-type\_repl\_Anaerobic  
 GSM1010242 Ecoli\_dFNR\_repl\_Anaerobic  
 GSM1010243 Ecoli\_dFNR\_repl2\_Anaerobic  
 GSM1010244 Ecoli\_wild-type\_repl1\_Anaerobic  
 GSM1010245 Ecoli\_wild-type\_repl2\_Anaerobic  
 GSM1010246 Ecoli\_dFNR\_repl1\_Anaerobic  
 GSM1010247 Ecoli\_dFNR\_repl2\_Anaerobic  
 GSM1054383 FNR - ΔfnsAtpA A  
 GSM1054384 FNR - ΔfnsAtpA B  
 GSM1072326 θ70 IP ChIP-seq Aerobic B  
 GSM1072327 θ70 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\_repb\_Aerobic\_Tiling  
 GSM1124980 Ecoli\_hns-stpA\_repb\_Aerobic\_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 θ70 in *E. coli* under aerobic and anaerobic growth conditions.  
 GSE41188 Expression analysis of *Escherichia coli* MG1655 K-12 WT and Δfnr mutant  
 Relations:  
 BioProject: PRJNA176146  
 Analyze with GEO2R

Download family: SOFT formatted family file(s) | MIML formatted family file(s) | Series Matrix File(s)

Format: SOFT ▾ MINML ▾ TXT ▾

Supplementary file: Size: 687.7 Mb Download: (http://custom) File type/resource: TAR (of WIG)

GSE41195\_RAW.tar

Platforms (3):  
 GPL8708 UW-Madison Escherichia coli K-12 MG1655 tiling array (YD Design)  
 GPL14449 NimbleGen E. coli K12 Gene Expression Array [071112\_Ecoli\_K12\_EXP]  
 GPL11609 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 B - Anaerobic - A  
 GSM1010203 B - Anaerobic - B  
 GSM1010204 B - Aerobic - A  
 GSM1010205 B - Aerobic - B  
 GSM1010206 HNS - Anaerobic A  
 GSM1010207 HNS - Aerobic A  
 GSM1010208 IHF - Anaerobic A  
 GSM1010209 Δfnr - Anaerobic  
 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 θ70 IP ChIP-seq Aerobic A  
 GSM1010222 θ70 IP ChIP-seq Aerobic A  
 GSM1010223 aerobic INPUT DNA  
 GSM1010224 anaerobic INPUT DNA  
 GSM1010240 Ecoli\_wild-type\_repl\_Anaerobic  
 GSM1010241 Ecoli\_wild-type\_repl\_Anaerobic  
 GSM1010242 Ecoli\_dFNR\_repl1\_Anaerobic  
 GSM1010243 Ecoli\_dFNR\_repl2\_Anaerobic  
 GSM1010244 Ecoli\_wild-type\_repl1\_Anaerobic  
 GSM1010245 Ecoli\_wild-type\_repl2\_Anaerobic  
 GSM1010246 Ecoli\_dFNR\_repl1\_Anaerobic  
 GSM1010247 Ecoli\_dFNR\_repl2\_Anaerobic  
 GSM1054383 FNR - ΔfnsAtpA A  
 GSM1054384 FNR - ΔfnsAtpA B  
 GSM1072326 θ70 IP ChIP-seq Aerobic B  
 GSM1072327 θ70 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\_repb\_Aerobic\_Tiling  
 GSM1124980 Ecoli\_hns-stpA\_repb\_Aerobic\_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 θ70 in *E. coli* under aerobic and anaerobic growth conditions.  
 GSE41188 Expression analysis of *Escherichia coli* MG1655 K-12 WT and Δfnr mutant

**NCBI** **GEO**  
Gene Expression Omnibus

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

Contact: okirsh [at] | My submissions [at] | Sign Out [at]

Scope: Set ▾ 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 θ70 in *E. coli* under aerobic and anaerobic growth conditions.  
 Organism: *Escherichia coli* str. K-12 substr. MG1655star  
 Experiment type: Genome-wide occupancy profiling by ChIP-seq under glucose fermentative anaerobic growth conditions. Also, many binding sites were identified for θ70 under both aerobic and anaerobic growth conditions. Described in the manuscript "Genome-scale analysis of *E. coli* FNR Reveals the Complexity of Bacterial Regulon Structure".  
 Summary: Overall design: Examination of occupancy of FNR and θ70 under aerobic and anaerobic growth conditions.  
 Contributors(s): Myers K, Van H, Ong I, Chang D, Liang K, Tran P, Kelles S, Landick R, Kelley P  
 Citation(s): Myers K, Van H, Ong I, Chang 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, Keleg S, A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* 2014 Mar;30(6):753-60. PMID: 23685773

Submission date: Sep 2012  
 Last update date: Sep 12, 2014  
 Contact name: Kevin Myers  
 E-mail: kmayers2@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): GPL8708 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 θ70 IP ChIP-seq Aerobic A  
 GSM1010222 θ70 IP ChIP-seq Aerobic A  
 GSM1010223 aerobic INPUT DNA  
 GSM1010224 anaerobic INPUT DNA  
 GSM1010225 θ70 IP ChIP-seq Aerobic A  
 GSM1010226 θ70 IP ChIP-seq Anaerobic A  
 GSM1010227 aerobic INPUT DNA  
 GSM1010228 θ70 IP ChIP-seq Aerobic B  
 GSM1010229 θ70 IP ChIP-seq Anaerobic B  
 GSM1010230 INPUT DNA B  
 GSM1010231 θ70 IP ChIP-seq Aerobic B  
 GSM1010232 θ70 IP ChIP-seq Anaerobic B  
 GSM1010233 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: PRJNA176146  
 SRA: SRP015911

Download family: SOFT formatted family file(s) | MIML formatted family file(s) | Series Matrix File(s)

Format: SOFT ▾ MINML ▾ TXT ▾

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

GSE41187\_RAW.tar

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

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

NCBI > GEO > Accession Display | Contact: clersh | My submissions | Sign Out | GEO accession: GSE41187 | DOWNLOAD

Scope: Sett | Format: HTML | Amount: Quick | GEO accession: GSE41187 | DOWNLOAD

Series GSE41187 | Query DataSets for GSE41187

Status: Public on Jun 20, 2013  
 Title: Genome-wide analysis of FNR and o70 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 o70 under both aerobic and anaerobic growth conditions. Described in the manuscript "Genome-scale Analysis of E. coli FNR Reveals the Complexity of Bacterial Regulation Structure".

Overall design: Examination of occupancy of FNR and o70 under aerobic and anaerobic growth 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 Jun 9(6):e1003565. PMID: 23813969  
 Zou C, Keleg 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: Keles S  
 E-mail: keles@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 o70 IP ChIP-seq Aerobic A  
 GSM1010222 o70 IP ChIP-seq Anaerobic A  
 GSM1010223 aerobic INPUT DNA  
 GSM1010224 anaerobic INPUT DNA  
 GSM1072326 o70 IP ChIP-seq Aerobic B  
 GSM1072327 o70 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:  
 SOFT formatted family file(s) | Format: SOFT |  
 MINIML formatted family file(s) | Format: MINIML |  
 Series Matrix File(s) | Format: 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

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)	GSM1010219 FNR IP ChIP-seq Anaerobic A GSM1010220 FNR IP ChIP-seq Anaerobic B GSM1010221 o70 IP ChIP-seq Aerobic A GSM1010222 o70 IP ChIP-seq Anaerobic A GSM1010223 aerobic INPUT DNA GSM1010224 anaerobic INPUT DNA GSM1072326 o70 IP ChIP-seq Aerobic B GSM1072327 o70 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:  
 SOFT formatted family file(s) | Format: SOFT |  
 MINIML formatted family file(s) | Format: MINIML |  
 Series Matrix File(s) | Format: 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

**Sample Sheets:**

- GSE41187\_RAW.txt (136.2 Mb, [HTTP](#))
- GSE41187\_A\_WIG.wig (16.4 Mb, [HTTP](#))

**Supplementary file:** GSE41187\_RAW.txt

**Raw data are available in SRA**

**Processed data provided as supplementary file**

**Platform:** GPL16109: illumina Genome Analyzer IIX (Escherichia coli str. K-12 substr. MG1655star)

**Samples:**

- GSM1010219: FNR IP ChIP-seq Anaerobic A
- GSM1010220: FNR IP ChIP-seq Anaerobic B
- GSM1010221: c70 IP ChIP-seq Aerobic A
- GSM1010222: c70 IP ChIP-seq Aerobic B
- GSM1010223: aerobic INPUT DNA
- GSM1010224: anaerobic INPUT DNA
- GSM1072326: c70 IP ChIP-seq Aerobic B
- GSM1072327: o70 IP ChIP-seq Anaerobic B
- GSM1072328: INPUT DNA B

**This SubSeries is part of SuperSeries:** GSE41187: Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

**Relations:**

- BioProject: PRJNA176149
- SRA: SRP015911

**Download family:**

- SOFT: formatted family file(s)
- MINIM: formatted family file(s)
- Series Matrix File(s)

**Format:**

- SOFT
- MINIM
- TXT

**Supplementary file:** GSE41187\_RAW.txt

**Size:** 136.2 Mb

**Download:** [HTTP](#)

**File type:** resource

**Sample Sheets:**

- GSE41187\_RAW.txt (136.2 Mb, [HTTP](#))
- GSE41187\_A\_WIG.wig (16.4 Mb, [HTTP](#))

**Supplementary file:** GSE41187\_RAW.txt

**Raw data are available in SRA**

**Processed data provided as supplementary file**

**Platform:** GPL16109: illumina Genome Analyzer IIX (Escherichia coli str. K-12 substr. MG1655star)

**Samples:**

- GSM1010219: FNR IP ChIP-seq Anaerobic A
- GSM1010220: FNR IP ChIP-seq Anaerobic B
- GSM1010221: c70 IP ChIP-seq Aerobic A
- GSM1010222: c70 IP ChIP-seq Aerobic B
- GSM1010223: aerobic INPUT DNA
- GSM1010224: anaerobic INPUT DNA
- GSM1072326: c70 IP ChIP-seq Aerobic B
- GSM1072327: o70 IP ChIP-seq Anaerobic B
- GSM1072328: INPUT DNA B

**This SubSeries is part of SuperSeries:** GSE41187: Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

**Relations:**

- BioProject: PRJNA176149
- SRA: SRP015911

**Download family:**

- SOFT: formatted family file(s)
- MINIM: formatted family file(s)
- Series Matrix File(s)

**Format:**

- SOFT
- MINIM
- TXT

**Supplementary file:** GSE41187\_RAW.txt

**Size:** 136.2 Mb

**Download:** [HTTP](#)

**File type:** resource

**Search results:**

Items: 9

- GSM1072328: INPUT DNA B: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 878,468 spots, 30.7M bases, 16.6Mb downloads
- Accession: SRX202453
- GSM1072327: o70 IP ChIP-seq Anaerobic B: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 2 Seg
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 8.5M spots, 298.7M bases, 196.9Mb downloads
- Accession: SRX202452
- GSM1072326: c70 IP ChIP-seq Aerobic B: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 3 ILLUMINA (Illumina Genome Analyzer IIx) run: 9.2M spots, 321.5M bases, 212.2Mb downloads
- Accession: SRX202451
- GSM1010224: anaerobic INPUT DNA: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 4 ILLUMINA (Illumina Genome Analyzer IIx) run: 6.7M spots, 241.8M bases, 151.8Mb downloads
- Accession: SRX189778
- GSM1010223: aerobic INPUT DNA: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 5 ILLUMINA (Illumina Genome Analyzer IIx) run: 5.7M spots, 198.9M bases, 119.2Mb downloads
- Accession: SRX189777
- GSM1010222: o70 IP ChIP-seq Anaerobic A: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 6 Seg
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 12.2M spots, 440.9M bases, 216.6Mb downloads
- Accession: SRX189776
- GSM1010221: o70 IP ChIP-seq Aerobic A: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 7 Seg
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 8M spots, 289.4M bases, 141.5Mb downloads
- Accession: SRX189775
- GSM1010220: FNR IP ChIP-seq Anaerobic B: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 8 Seg
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 9.9M spots, 348.1M bases, 208.3Mb downloads
- Accession: SRX189774
- GSM1010219: FNR IP ChIP-seq Anaerobic C: Escherichia coli str. K-12 substr. MG1655star: ChIP-Seq
- 9 Seg
- 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 3.6M spots, 129.7M bases, 79.3Mb downloads
- Accession: SRX189773

**Filters:** Manage Filters

**Search in related databases:**

Database	public	controlled	all
BioSample	1	1	1
BioProject	1	1	1
dbGaP	1	1	1
GEO Datasets	1	1	1

**Find related data:**

Database: Select

**Recent activity:**

- SRP015911 (9)
- Escherichia coli str. K-12 substr. MG1655star BioProject
- FNR IP ChIP-seq Anaerobic A biosample
- Genome-scale Analysis of E. coli FNR Reveals Complex Features of Tr GEO Datasets

# NCBI - GEO DataSets: Fetch FASTQ

**GEO**  
Gene Expression Omnibus

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

NCBI > GEO > Accession Display [?] Contact: okirsh [?] My submissions [?] Sign Out [?]

Scope: Spec Format: HTML Amount: Once GEO accession: GSE41187 [?] 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: Genomic 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 growth conditions. Described in the manuscript "Genome-scale Analysis of E. coli FNR Reveals the Complexity of Bacterial Regulation Networks".

Overall design: Examination of occupancy of FNR and 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, Riley 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: 23736773  
Zuo C, Keles S. A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* 2014 Mar;30(6):753-60. PMID: 23665773

Submission date: Sep 27, 2012  
Last update date: Sep 12, 2014  
Contact name: Kevin Myers  
E-mail: kmyers@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):  
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 Aerobic B  
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:  
GSE41187 Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

Relations:  
BioProject: PRJNA176149  
SRA: SRP015911

Download family:  
SOFT formatted family file(s)  
MINIM, formatted family file(s)  
Series Matrix File(s)

Supplementary file: GSE41187\_RAW.tar Size: 136.2 Mb Download: (http://custom) File type/resource: TAR (of WIG)

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

NLM | NIH | GEO Help | Disclaimer | Accessibility

NCBI Resources How To

SRA SRA Advanced Search Help

SRA20453: GSM1072328: INPUT DNA B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq 1 ILLUMINA (Illumina Genome Analyzer IIx) run: 878,466 spots, 30.7M bases, 16.6Mb downloads

Submitted by: Gene Expression Omnibus (GEO)  
Study: Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.  
PRJNA176149 • SRP015911 • All experiments • All runs show Abstract

Sample: INPUT DNA B  
SAMN01907001 • SRS387991 • All experiments • All runs  
Organism: Escherichia coli str. K-12 substr. MG1655star

Library: Illumina Genome Analyzer IIx  
Strategy: ChIP-Seq  
Source: GENOMIC  
Selection: ChIP  
Layout: SINGLE

Experiment attributes:  
GEO Accession: GSM1072328

Links:  
External link: GEO Sample

Runs: 1 run, 878,466 spots, 30.7M bases, 16.6Mb  
Run # of Spots # of Bases Size Published  
SRR653522 878,466 30.7M 16.6Mb 2015-07-22

ID: 307868

Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST Studies Samples Analyses Run Browser Run Selector Provisional SRA

GSM1072328: INPUT DNA B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq (SRR653522) Change accession:

Metadata Analysis (alpha) Reads Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR653522	878.5k	30.7Mbp	17.5M	50.2%	2015-07-22	public

Quality graph (http://)

This run has 1 read per spot:  
L=35, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX220453		Illumina	ChIP-Seq	GENOMIC	ChIP	SINGLE

To BLAST

Biosample	Sample Description	Organism	Links
SAMN01907001 (SRS387991)	Escherichia coli str. K-12 substr. MG1655star		<ul style="list-style-type: none"><li>PRJNA176149 [Escherichia coli str. K-12 substr. MG1655star]</li><li>PRJNA176146 [Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding]</li></ul>

Bioproject SRA Study Title  
PRJNA176149 SRP015911 Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.

Show abstract

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility  
National Center for Biotechnology Information | U.S. National Library of Medicine

Last update: Thu Aug 2 10:37 EDT 2018

# NCBI - SRA: Fetch FASTQ

NCBI Site map All databases Search

Sequence Read Archive

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

Studies Samples Analyses Run Browser Run Selector Provisional SRA

GSM1072328: INPUT DNA B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq (SRR653522)

Change accession...

Metadata Analysis (alpha) Reads Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR653522	878.5k	30.7Mbp	17.5M	50.2%	2015-07-22	public

Quality graph (bigger)

This run has 1 read per spot:  
L=35, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX220453		Illumina	ChIP-Seq	GENOMIC	ChIP	SINGLE

To BLAST

Biosample Sample Description Organism Links

SAMN01907001 Escherichia coli str. K-12 substr. MG1655star (SR5387991)

- PRJNA176149 [Escherichia coli str. K-12 substr. MG1655star]
- PRJNA176146 [Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding]

Bioproject SRA Study Title

PRJNA176149 SRP015911 Genome-wide analysis of FNR and s70 in E. coli under aerobic and anaerobic growth conditions.

Show abstract

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility

National Center for Biotechnology Information | U.S. National Library of Medicine

NCBI Site map All databases Search

Sequence Read Archive

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

Studies Samples Analyses Run Browser Run Selector Provisional SRA

GSM1072328: INPUT DNA B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq (SRR653522)

Change accession...

Metadata Analysis (alpha) Reads Download

Filter: Find Filtered Download What does it do? What can the filter be applied to?

< 1 1 > 87847 Read View: biological reads technical reads quality scores advanced options

1. SRR653522\_1 SRS387991  
name: ILLUMINA05.2:1:0:2016 member: default  
2. SRR653522\_2 SRS387991  
name: ILLUMINA05.2:1:0:69 member: default  
3. SRR653522\_3 SRS387991  
name: ILLUMINA05.2:1:0:48 member: default  
4. SRR653522\_4 SRS387991  
name: ILLUMINA05.2:1:0:618 member: default  
5. SRR653522\_5 SRS387991  
name: ILLUMINA05.2:1:0:126 member: default  
6. SRR653522\_6 SRS387991  
name: ILLUMINA05.2:1:0:137 member: default  
7. SRR653522\_7 SRS387991  
name: ILLUMINA05.2:1:0:1869 member: default  
8. SRR653522\_8 SRS387991  
name: ILLUMINA05.2:1:0:297 member: default  
9. SRR653522\_9 SRS387991  
name: ILLUMINA05.2:1:0:764 member: default  
10. SRR653522\_10 SRS387991  
name: ILLUMINA05.2:1:0:882 member: default

>gnl|SRA|SRR653522\_1 ILLUMINA05.2:1:0:2016  
NCGTCACCATTTCCAGCTCGCCGCCAACCGGAAAC

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility

National Center for Biotechnology Information | U.S. National Library of Medicine

Last update: Thu Aug 2 10:13:37 EDT 2018

# NCBI - SRA: Fetch FASTQ

NCBI Site map All databases Search

 Sequence Read Archive

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

Studies Samples Analyses Run Browser Run Selector Provisional SRA

GSM1072328: INPUT DNA B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq (SRR653522)

Metadata Analysis (alpha) Reads Download

You need [SRA Toolkit](#) to operate on SRA runs.

Default toolkit configuration enables it to find and retrieve SRA runs by accession. It also downloads (and cache) only the part of data you really need. For example quality scores represent a majority of data volume and you may not need them if you dump fasta only (versus fastq). Or if you are looking at particular gene you may not need reads aligned to other regions or not aligned at all. Same way if you use GATK with enabled SRA support you need only SRA run accessions to fire your process.

[fastq-dump](#) will dump reads in a number of "standard" fastq and fasta formats.

[vdb-dump](#) is also capable of producing fasta and fastq (beside other formats). It dumps data much faster than fastq-dump but ordering of reads may be different and it does not produce split-read multi-file output.

[Prefetch](#) tool will help you cache all data in advance if you plan to run data analysis in environment where getting data from NCBI at run time is unfeasible.

To select a list of interspersed SRA runs in the scope of experiment, sample or study you may use SRA [Run Selector](#) either directly or from [Entrez](#) search.

Read more at [SRA Knowledge Base](#) on how to download SRA data using command line utilities.

# NCBI - SRA Toolkit : Fetch FASTQ

NCBI Site map All databases Search

Sequence Read Archive

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

Download Toolkit Documentation XML Schema

## NCBI SRA Toolkit

Please consult [SRA Toolkit Documentation](#) for help.  
Below are the latest releases of various tools and release checksum file.

### SRA Toolkit

Compiled binaries of July 24, 2018, version 2.9.2 release:

- [CentOS Linux 64 bit architecture](#)
- [Ubuntu Linux 64 bit architecture](#)
- [MacOS 64 bit architecture](#)
- [MS Windows 64 bit architecture](#)

### Magic-BLAST

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.

- Magic-BLAST executables for LINUX, Mac OSX, and Windows as well as the source files are available on the [FTP site](#)
- Read more about Magic BLAST on the [FTP site](#)

### Third Party Software

Builds of Third Party Software Tools with SRA support ( NGS 2.9.2 release ):

- [Genome Analysis Toolkit \(GATK\) version 3.6-6/ngs.2.9.2](#) - including direct support of SRA
- [HISAT2 version 2.1.0/ngs.2.9.2](#) - graph-based alignment of next generation sequencing reads to a population of genomes with dire

### Latest Source Code

- [NGS Software Development Kit](#) – July 24, 2018, version 2.9.2 release
- [NCBI VDB Software Development Kit](#) – July 24, 2018, version 2.9.2 release
- [NCBI SRA Toolkit](#) – July 24, 2018, version 2.9.2 release

### File checksums

You may validate downloaded files with [md5 checksums](#) computed using `md5sum -b`

NCBI Site map All databases Search

Sequence Read Archive

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

Download Toolkit Documentation XML Schema

## SRA Toolkit Documentation

[SRA Toolkit Installation and Configuration Guide](#)

[Protected Data Usage Guide](#)

### Frequently Used Tools:

- [fastq-dump](#): Convert SRA data into fastq format  
[prefetch](#): Allows command-line downloading of SRA, dbGaP, and ADSP data  
[sam-dump](#): Convert SRA data to sam format  
[sra-pileup](#): Generate pileup statistics on aligned SRA data  
[vdb-config](#): Display and modify VDB configuration information  
[vdb-decrypt](#): Decrypt non-SRA dbGaP data ("phenotype data")

### Additional Tools:

- [abi-dump](#): Convert SRA data into ABI format (csfasta / qual)  
[illumina-dump](#): Convert SRA data into Illumina native formats (qseq, etc.)  
[sff-dump](#): Convert SRA data to sff format  
[sra-stat](#): Generate statistics about SRA data (quality distribution, etc.)  
[vdb-dump](#): Output the native VDB format of SRA data.  
[vdb-encrypt](#): Encrypt non-SRA dbGaP data ("phenotype data")  
[vdb-validate](#): Validate the integrity of downloaded SRA data

# NCBI - SRA Toolkit : prefetch

## 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 is 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.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# ...
```

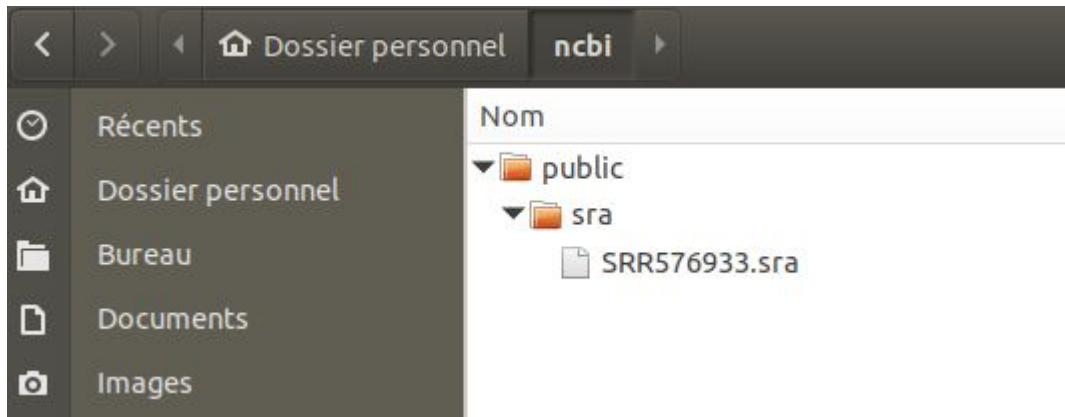
```
prefetch ~/Downloads/SRR390728.sra
If you have already downloaded an SRA datafile (example here: SRR390728.sra, present in the "~/Downloads" directory), this command will retrieve all of the reference sequences required to extract the data. This command is useful for resolving errors of the type "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

```
2019-01-21T15:42:11 prefetch.2.9.1: 1) SRR576934 ts found locally
(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



## 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 <url[: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 the key.

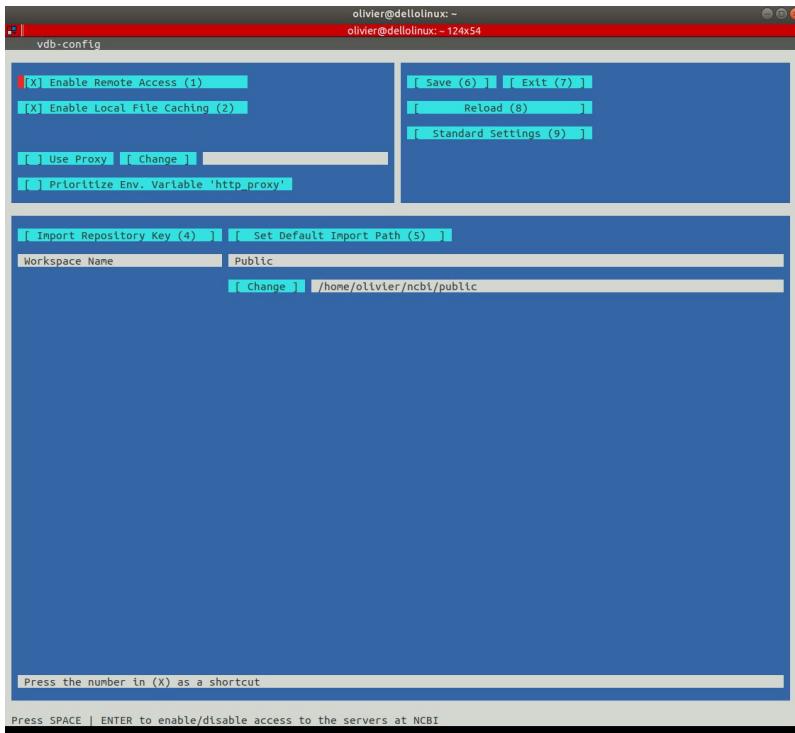
vdb-config --interactive-mode textual  
Runs the configuration tool in textual mode. Common configuration commands are selected using the keyboard in this mode.

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

NCBI SRA Run Selector [Help](#)

Search:   

Nothing found

[Write to the Help Desk](#) | [Privacy Notice](#) | [Disclaimer](#) | [Accessibility](#)  
National Center for Biotechnology Information | U.S. National Library of Medicine

search for SRP, GSE, GSM, ...

# NCBI - SRA run selector: Fetch FASTQs

NCBI SRA Run Selector | Help | Permalink

Search: SRP015911

Facets

- Run
- BioSample
- Sample name
- AvgSpotLen
- Experiment
- LoadDate
- MBases
- MBytes
- antibody
- growth condition

growth condition

- aerobic cultures [4]
- anaerobic cultures [5]

antibody

- affinity purified fnr polyclonal antibody
- input [3]
- RNA Polymerase s70 monoclonal antibody

Hide common fields

Assay Type:	ChIP-Seq
BioProject:	<a href="#">PRJNA176149</a>
Center Name:	GEO
Consent:	public
DATASTORE filetype:	sra
DATASTORE provider:	ncbi
InsertSize:	0
Instrument:	Illumina Genome Analyzer IIx
LibraryLayout:	SINGLE
LibrarySelection:	ChIP
LibrarySource:	GENOMIC
Organism:	Escherichia coli str. K-12 substr. MG1655star
Platform:	ILLUMINA
ReleaseDate:	2015-07-22
SRA Study:	<a href="#">SRP015911</a>
strain:	Wild Type K-12

Total: 9 Bytes: 1.31 Gb Bases: 2.19 G

Download

Selected:

Run BioSample Sample name AvgSpotLen Experiment LoadDate MBases MBytes antibody growth condition source name

Run	BioSample	Sample name	AvgSpotLen	Experiment	LoadDate	MBases	MBytes	antibody	growth condition	source name
SRR653522	SAMN01907001	GSM1072328	35	SRX220453	2015-09-05	29	16	INPUT	Aerobic Cultures	Aerobic Cultures
SRR653521	SAMN01907000	GSM1072327	35	SRX220452	2015-09-05	284	196	RNA Polymerase s70 monoclonal antibody from NeoClone (W0004)	Anaerobic Cultures	Anaerobic Cultures
SRR653520	SAMN01906999	GSM1072326	35	SRX220451	2015-09-05	306	215	RNA Polymerase s70 monoclonal antibody from NeoClone (W0004)	Aerobic Cultures	Aerobic Cultures
SRR576938	SAMN01731121	GSM1010224	36	SRX189778	2015-12-07	230	151	INPUT	Anaerobic Cultures	Anaerobic Cultures
SRR576937	SAMN01731120	GSM1010223	35	SRX189777	2015-12-07	189	119	INPUT	Aerobic Cultures	Aerobic Cultures
SRR576936	SAMN01731119	GSM1010222	36	SRX189776	2015-12-07	420	216	RNA Polymerase s70 monoclonal antibody from NeoClone (W0004)	Anaerobic Cultures	Anaerobic Cultures
SRR576935	SAMN01731118	GSM1010221	36	SRX189775	2015-12-07	275	141	RNA Polymerase s70 monoclonal antibody from NeoClone (W0004)	Aerobic Cultures	Aerobic Cultures
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

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility  
National Center for Biotechnology Information | U.S. National Library of Medicine

Last update: Fri Dec 7 12:00:59 EST 2018

NIH FIRST GOV.gov

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

# NCBI - SRA run selector : Fetch FASTQs

SraRunTable.txt

AvsSpotLen	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	Instru
36	SAMN01731116	SRX189773	2015-12-07	123	79	SRR576933	SR365655	GSM1010219	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures	ChIP-Seq	PRJNA176149	GEO	public	sra	ncbi	0	Illumini
35	SAMN01731117	SRX189774	2015-12-07	332	208	SRR576934	SR365656	GSM1010220	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures	ChIP-Seq	PRJNA176149	GEO	public	sra	ncbi	0	Illumini

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

NCBI Site map All databases Search

Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace

Download Toolkit Documentation XML Schema

## SRA Toolkit Documentation

[SRA Toolkit Installation and Configuration Guide](#)

[Protected Data Usage Guide](#)

### Frequently Used Tools:

**fastq-dump**: Convert SRA data into fastq format

**prefetch**: Allows command-line downloading of SRA, dbGaP, and ADSP data

**sam-dump**: Convert SRA data to sam format

**sra-pileup**: Generate pileup statistics on aligned SRA data

**vdb-config**: Display and modify VDB configuration information

**vdb-decrypt**: Decrypt non-SRA dbGaP data ("phenotype data")

### Additional Tools:

**abi-dump**: Convert SRA data into ABI format (csfasta / qual)

**illumina-dump**: Convert SRA data into Illumina native formats (qseq, etc.)

**sff-dump**: Convert SRA data to sff format

**sra-stat**: Generate statistics about SRA data (quality distribution, etc.)

**vdb-dump**: Output the native VDB format of SRA data.

**vdb-encrypt**: Encrypt non-SRA dbGaP data ("phenotype data")

**vdb-validate**: Validate the integrity of downloaded SRA data

[Write to the Help Desk](#) | [Privacy Notice](#) | [Disclaimer](#) | [Accessibility](#)

National Center for Biotechnology Information | U.S. National Library of Medicine

NIH FIRSTGOV.gov

```
(ng) olivter@deLLlinux:~/Téléchargements$ fastq-dump -h
```

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

INPUT  
-A|--accession <accession> Replaces accession derived from <path> in  
filename(s) and defines (only for single  
table dump)  
Table name within cSRA object  
"SEQUENCE"

PROCESSING  
Read Splitting  
--split-spot Split spots into individual reads  
Full Spot Filters  
-N|--minSpotId <rowid> Applied to the full spot index  
-X|--maxSpotId <rowid> of --split-spot  
-S|--spot-groups <[list]> Minimum spot id  
-W|--clip Maximum spot id  
Filter by SPOT\_GROUP (member  
of list)  
Remove adapter sequences from  
the sequence

Common Filters  
-M|--minReadLen <len> Applied to spots when --split-  
set, otherwise - to individual  
spots. Filtered by sequence length >=

-R|--read-filter <[filter]> Spots listed by SPOT\_ID\_FILT  
optionally filter by value:  
pass/reject/criteria/edacted

-E|--qual-filter Filter used in early 1000 Gen  
sequences starting or ending

--qual-filter-1 Filter used in current 1000 Gen

Filters based on alignments  
--aligned Dump only aligned sequences  
--unaligned Dump only unaligned sequences  
--aligned-region <name[:from-to]> Filter by position on genome  
either be accession.version  
NC\_000001.10 or file specific

OPTIONS  
-m|--matepair-distance <from-to> Filter by distance between matepairs,  
use "unknown" to find matepairs split  
between the references. Use From-to to limit  
matepair distance on the same reference

Filters for individual reads  
--skip-technical

OUTPUT  
-O|--outdir <path> coordinates  
-Z|--stdout OUTPUT  
--gztp Multiple File Options  
-z|--gzip

Output directory, default is working  
directory '.' )  
Output to stdout, all split data become  
joined

Compress  
-C|--compress Quality  
-Q|--offset <integer>  
Setting  
than  
will re  
number  
define  
-F|--origfmt  
-I|--readids  
Legacy  
First  
present  
+ 2. fast  
reads a  
Split t  
-F|--helicos  
-D|--define-seq <fmt>  
-D|--define-qual <fmt>

Quality  
-Q|--offset <integer>  
Setting  
than  
will re  
number  
define  
-F|--origfmt  
-I|--readids  
Legacy  
First  
present  
+ 2. fast  
reads a  
Split t  
-F|--helicos  
-D|--define-seq <fmt>  
-D|--define-qual <fmt>

FORMATTING  
-G|--spot-group  
-R|--read-filter <[filter]>  
-T|--group-in-dir  
-K|--keep-empty-files  
FORMATTING  
sequence  
<C>--dumps <[cskey]>  
-B|--dumpbase  
quality  
-Q|--offset <integer>  
offset  
OTHER:  
-d|--disable-multithreading  
-h|--help  
-v|--version  
-L|--log-level <level>  
-v|--verbose  
--ncbi\_error\_report  
--legacy-report  
fastq-dump : 2.9.1 ( 2.9.1-1 )

# 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
AAGCATGGATAACCGCCTGGTGAATGCTGCCATA
+SRR576933.1 HWUSI-EAS1789_0000:2:20:1269:14140 length=36
EDEA=EEEEEEBFBDFDEBDAD=DDBAC5CCEEE:E@
@SRR576933.2 HWUSI-EAS1789_0000:2:20:1270:19579 length=36
TGGAGGCTGACCACGATAAGCTGCCCTGGTGC
+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
AGTGCATGCCGTTCACCCGGTTCTTTATCATTA

==> SRR576934.fastq <==
@SRR576934.1 HWUSI-EAS787_0001:1:1:16343:952 length=35
NGATCAAAACAAAAAATATTTCACTCGANAGGAG
+SRR576934.1 HWUSI-EAS787_0001:1:1:16343:952 length=35
%%%%%%%%%%%%%%%
@SRR576934.2 HWUSI-EAS787_0001:1:1:17465:951 length=35
NTGACGTAAATCTTAGGGCCAACGCNCATAA
+SRR576934.2 HWUSI-EAS787_0001:1:1:17465:951 length=35
%//77787@CCC@@@0@0000@0@%%%%%%%%
@SRR576934.3 HWUSI-EAS787_0001:1:1:16452:962 length=35
AGCTCTACACTTCATGTTATTTCTCATNATTC
```

## 2 Runs found

	Run	BioSample	Sample name	AvgSpotLen	Experiment	LoadDate	MBases	MBytes	antibody	growth condition	source name
<input checked="" type="checkbox"/>	SRR576934	SAMN01731117	GSM1010220	35	SRX189774	2015-12-07	332	208	Affinity purified FNR polyclonal antibody	Anaerobic Cultures	Anaerobic Cultures
<input checked="" type="checkbox"/>	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

Usage

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

Frequently Used Options

General

- h | --help  
-V | --version

Data Formatting

- split-files  
--split-spot  
--fasta < [line width] >  
-I | --readids  
-F | --origfmt  
-C | --dumpcs < [cskey] >  
-B | --dumpbase  
-Q | --offset < integer >

Dump each read into separate file. Files will receive suffix corresponding to read number.  
Split spots into individual reads.  
FASTA only, no qualities. Optional line wrap width (set to zero for no wrapping).  
Append read id after spot id as 'accession.spot.readid' on define.  
Define contains only original sequence name.  
Formats sequence using color space (default for SOLID). "cskey" may be specified for translation.  
Formats sequence using base space (default for other than SOLID).  
Offset to use for ASCII quality scores. Default is 33 ("!").

Filtering

- N | --minSpotId < rowid >  
-X | --maxSpotId < rowid >  
-M | --minReadLen < len >  
--skip-technical  
--aligned  
--unaligned

Minimum spot id to be dumped. Use with "X" to dump a range.  
Maximum spot id to be dumped. Use with "N" to dump a range.  
Filter by sequence length  $\geq$  len  
Dump only biological reads.  
Dump only aligned sequences. Aligned datasets only; see sra-stat.  
Dump only unaligned sequences. Will dump all for unaligned datasets.

Workflow and piping

- o | --outdir < path >  
-z | --stdout  
--gzip  
--bzip2

Output directory, default is current working directory ('').  
Output to stdout, all split data become joined into single stream.  
Compress output using gzip.  
Compress output using bzip2.

Attention “RTFM”



Check if you have Single or Paired End data

# Alternative Data base in case of shut down!

The screenshot shows the ENA homepage with a dark header bar containing the EMBL-EBI logo, navigation links for Services, Research, Training, and About us, and a search bar with examples BN000065, histone, and buttons for Search, Advanced, and Sequence.

The main content area has a teal header "European Nucleotide Archive". Below it, a text block states: "The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)".

Below this, a "Text Search" section contains a search input field with examples BN000065, histone, a "search" button, and a link to "Advanced search".

A "Sequence Search" section contains a text input field with placeholder "Enter or paste a nucleotide sequence or accession number", a "Search" button, and a link to "Advanced search".

To the right, a "Popular" sidebar lists links for: Submit and update, Sequence submissions, Genome assembly submissions, Submitting environmental sequences, Citing ENA data, Rest URLs for data retrieval, and Rest URLs to search ENA.

A "Latest ENA news" sidebar displays a single item: "06 Nov 2018: [Latest ENA Nucleic Acids Research DB issue paper published](#)". A link below it says "The latest European Nucleotide Archive NAR DB issue...".

At the bottom, a dark footer bar features the ELIXIR Core Data Resource logo and the text "ENA is part of the ELIXIR infrastructure" with a "Learn more" link.

# ENA: European Nucleotide Archive

The screenshot shows the ENA homepage with a search bar at the top containing the identifier "SRP015911". Below the search bar are links for "Advanced Sequence". The main content area displays search results for "SRP015911", with a "Show more data from EMBL-EBI" button. On the left, there are filters for "Read" (Run (9)) and "Study" (Study (1)). The search results table lists nine entries, each with a run ID and a brief description of the sequencing experiment.

EMBL-EBI 

Services Research Training About us

**ENA**  
European Nucleotide Archive

SRP015911  
Examples: [BN000065](#), [histone](#)

Search  
[Advanced](#)  
[Sequence](#)

Home Search & Browse Submit & Update Software About ENA Support

Search results for **SRP015911**

Show more data from EMBL-EBI

Read  
Run (9)

Study  
Study (1)

Run (9 results found)

Download:  -  of 9 results in [XML](#)

Showing results 1 - 9 of 9 results

<a href="#">SRR653521</a>	Illumina Genome Analyzer IIx sequencing; GSM1072327: σ70 IP... <a href="#">more</a>
<a href="#">SRR653522</a>	Illumina Genome Analyzer IIx sequencing; GSM1072328: INPUT DNA B;... <a href="#">more</a>
<a href="#">SRR576935</a>	Illumina Genome Analyzer IIx sequencing; GSM1010221: σ70 IP... <a href="#">more</a>
<a href="#">SRR576936</a>	Illumina Genome Analyzer IIx sequencing; GSM1010222: σ70 IP... <a href="#">more</a>
<a href="#">SRR576937</a>	Illumina Genome Analyzer IIx sequencing; GSM1010223: aerobic... <a href="#">more</a>
<a href="#">SRR576934</a>	Illumina Genome Analyzer IIx sequencing; GSM1010220: FNR IP... <a href="#">more</a>
<a href="#">SRR653520</a>	Illumina Genome Analyzer IIx sequencing; GSM1072326: σ70 IP... <a href="#">more</a>
<a href="#">SRR576938</a>	Illumina Genome Analyzer IIx sequencing; GSM1010224: anaerobic... <a href="#">more</a>
<a href="#">SRR576933</a>	Illumina Genome Analyzer IIx sequencing; GSM1010219: FNR IP... <a href="#">more</a>

# ENA: European Nucleotide Archive

EMBL-EBI

ENA  
European Nucleotide Archive

Services Research Training About us

Home Search & Browse Submit & Update Software About ENA Support

Contact Helpdesk

Run: SRR653521

Illumina Genome Analyzer Iix sequencing; GSM1072327: o70 IP ChIP-seq Anaerobic B; Escherichia coli str. K-12 substr. MG1655star; ChIP-Seq

View: XML Download: XML

Submitting Centre GEO	Platform ILLUMINA	Model Illumina Genome Analyzer Iix	Read Count 8,534,578	Base Count 298,710,230
Library Layout SINGLE	Library Strategy ChIP-Seq	Library Source GENOMIC	Library Selection ChIP	Library Name
Broker Name NCBI				

Navigation Read Files

This table contains the files for run SRR653521

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 1 of 1 results in TEXT

Select columns

Showing results 1 - 1 of 1 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJNA176149	SAMN01907000	SRS387990	SRX220452	SRR653521	879462	Escherichia coli str. K-12 substr. MG1655star	Illumina Genome Analyzer Iix	SINGLE	File 1	File 1			File 1	File 1		

# FASTQ : Boites à outils

- FASTX ToolKit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/))
- Trim\_Galore ([https://www.bioinformatics.babraham.ac.uk/projects/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!

# FASTQ : Boites à outils....

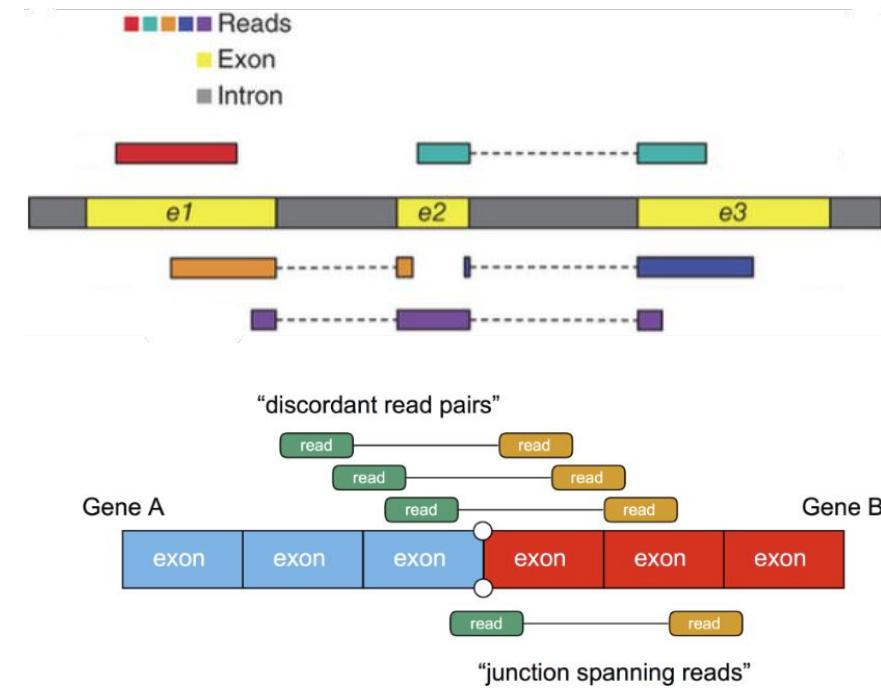
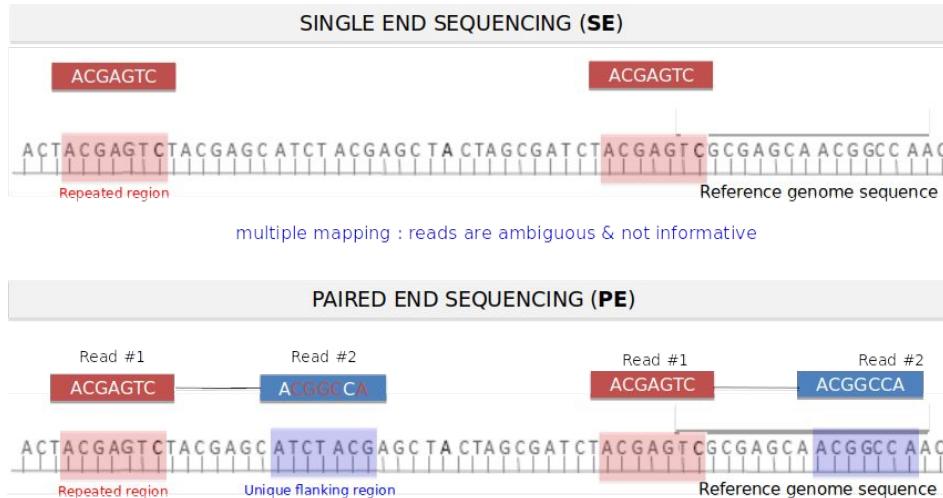
- Cutadapt (<https://cutadapt.readthedocs.io/en/latest/guide.html>)
- prinseq (<http://prinseq.sourceforge.net/manual.html>)
- tally (<https://www.ebi.ac.uk/~stijn/reaper/tally.html>)
- reaper (<http://www.ebi.ac.uk/~stijn/reaper/reaper.html>)
- seqtk (<https://github.com/lh3/seqtk>)
- deconseq (<http://deconseq.sourceforge.net/>)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

Suite et boites à outils généralistes

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

# 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"

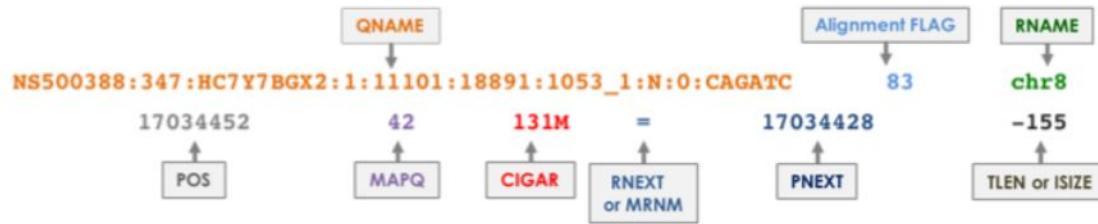
```

## 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...

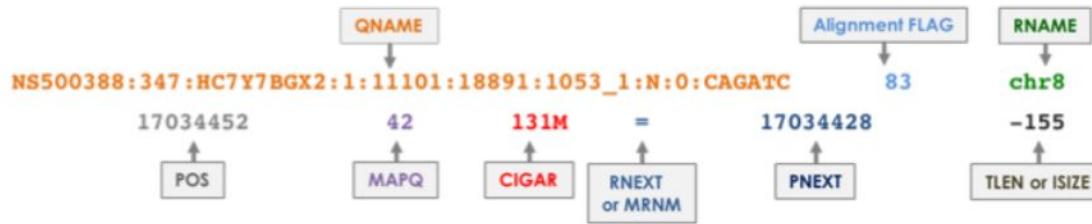
COLUMN	NAME	CONTENT	
1	QNAME	NS500388:347:HC7Y7BGX2:1:1110...	
2	FLAG	83	
3	RNAME	chr8	
4	POS	17034452	
5	MAPQ	42	
6	CIGAR	131M	
7	RNEXT	=	
8	PNEXT	17034428	
9	TLEN	-155	
10	SEQ	TTAATTACATCCATATACT ...	
11	QUAL	EEA<EEEEEEEAEAA<E	
12	NM-tag	NM:i:23	
13	MD-tag	MD:Z:2G0G13G2G1G6 ...	
14	XM-tag	XM:Z...hh.....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

# SAM/BAM : Sequence Alignment/Map



1	<b>QNAME</b>	ID of the reads
2	<b>FLAG</b>	Alignment flag (Combination of bitwise FLAGS)
3	<b>RNAME</b>	Reference name (Chromosome name)
4	<b>POS</b>	Position in reference (5' position)
5	<b>MAPQ</b>	Mapping quality
6	<b>CIGAR</b>	Alignment description in CIGAR format (M:match ; I:insertion ; D:deletion ; S:substitution)
7	<b>RNEXT or MRNM</b>	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	<b>PNEXT</b>	Left most position of where the next alignment in this gap maps to the reference
9	<b>TLEN or ISIZE</b>	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>

**Picard**  
build: 1.135.0

A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

**Decoding SAM flags**

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag:  [Explain](#)

[Switch to mate](#) Toggle first in pair / second in pair

**Summary:**

- read paired (0x1)
- read mapped in proper pair (0x2)
- read reverse strand (0x10)
- first in pair (0x40)
- second in pair (0x80)
- not primary alignment (0x100)
- read fails platform/vendor quality checks (0x200)
- read is PCR or optical duplicate (0x400)
- supplementary alignment (0x800)

# 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:

Reference sequence with aligned reads	CIGAR string	Explanation
C T G C A T G T T A G A T A A * * G A T A G C T G T G C T A A A G G A T A * C T G	<b>1M2I4M1D3M</b>	Insertion & Deletion
G A T A A * G G A T A	<b>5M1P1I4M</b>	Padding & Insertion
T G T T A [redacted] T G C T A	<b>5M15N5M</b>	Spliced read
a a a C A T G T T A G	<b>3S8M</b>	Soft clipping
A A A C A T G T T A G	<b>3H8M</b>	Hard clipping

# Outils de manipulation des fichiers SAM/BAM

Samtools

Home

Download ▾

Workflows ▾

Documentation ▾

Support ▾

## Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

**Samtools** Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format

**BCFtools** Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants

**HTSlip** A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlip internally, but these source packages contain their own copies of htllib so they can be built independently.

### Download

Source code releases can be downloaded from [GitHub](#) or [Sourceforge](#):

 [Source release details](#)

### Workflows

We have described some standard workflows using Samtools:

- WGS/WES Mapping to Variant Calls
- Using CRAM within Samtools

### Documentation

- Manuals
- HowTos
- Specifications
- Duplicate Marking
- Zlib Benchmarks
- CRAM Benchmarks
- Publications

### Support

- Mailing Lists
- HTSlip issues
- BCFtools issues
- Samtools issues

# 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)
addrplacerg adds or replaces RG tags
markup   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]
```

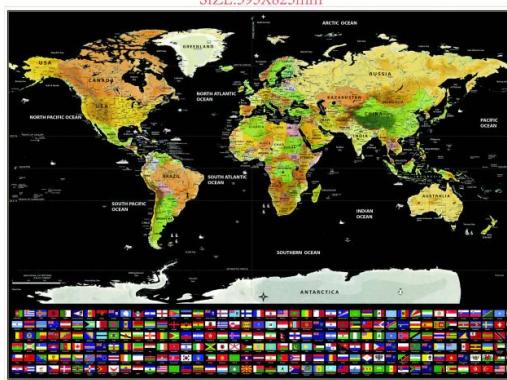
```
(ngs) olivier@dellolinux:~$ samtools sort
Usage: samtools sort [options...] [in.bam]

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)
- GRCh37 - hg19 (Feb 2009)
- GRCh38 - 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>)

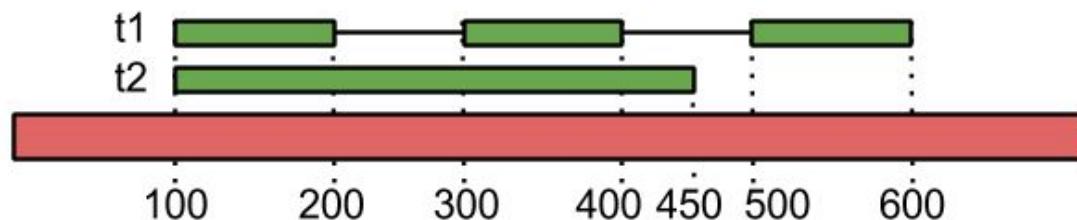
The GFF format is a flat tab-delimited file, each line of which corresponds to an annotation, or feature.

```
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>