

# Using an existing metabolic model to (start to) build a new one!

Dr Sandy Macdonald, Bioscience Technology Facility Bioinformatician,  
Friday, 19th November, 2021

# Where to start?

- Let's imagine you have an annotated genome sequence for a bacterium and want to build a flux balance analysis (FBA) model for it. Where do you start?
- You could go through the genome, gene by gene, search for that gene in KEGG or MetaCyc to find what reaction it catalyses, and then add the reactions, metabolites, and genes to your model, typing out the SBML by hand. 😜 (*I did this in my postdoc in 2008*)
- Starting from an existing model/s and transferring as much knowledge from that model as possible to your new model would be much better!
- The result is not a complete model, but a very good starting point

# *E. coli* as a starting point

- As with many other things, *E. coli* is a good starting point, because:
  - Its genome is very well annotated
  - Its metabolism is (relatively) well understood
  - Many of its metabolic enzymes have been studied experimentally
  - There have been several versions now of *E. coli* FBA models, each time improved and with more genes included

**molecular systems biology** Setting standards in Systems Biology

*Mol Syst Biol.* 2007; 3: 121.  
Published online 2007 Jun 26. doi: [10.1038/msb4100155](https://doi.org/10.1038/msb4100155)  
PMCID: PMC1911197  
PMID: [17593909](https://pubmed.ncbi.nlm.nih.gov/17593909/)

**A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information**

Adam M Feist,<sup>1</sup> Christopher S Henry,<sup>2</sup> Jennifer L Reed,<sup>1</sup> Markus Krummenacker,<sup>3</sup> Andrew R Joyce,<sup>1</sup> Peter D Karp,<sup>3</sup> Linda J Broadbelt,<sup>2</sup> Vassily Hatzimanikatis,<sup>4</sup> and Bernhard Ø Palsson<sup>1,a</sup>

► Author information ► Article notes ► Copyright and License information ► Disclaimer

This article has been cited by other articles in PMC.

**Associated Data**

► Supplementary Materials

**Abstract** Go to: ▾

An updated genome-scale reconstruction of the metabolic network in *Escherichia coli* K-12 MG1655 is presented. This updated metabolic reconstruction includes: (1) an alignment with the latest genome annotation and the metabolic content of EcoCyc leading to the inclusion of the activities of 1260 ORFs, (2) characterization and quantification of the biomass components and maintenance requirements associated with growth of *E. coli* and (3) thermodynamic information for the included chemical reactions. The conversion of this metabolic network reconstruction into an *in silico* model is detailed. A new step in the metabolic reconstruction process, termed thermodynamic consistency analysis, is introduced, in which reactions were checked for consistency with thermodynamic reversibility estimates. Applications demonstrating the capabilities of the genome-scale metabolic model to predict high-throughput experimental growth and gene deletion phenotypic screens are presented. The increased scope and computational capability using this new reconstruction is expected to broaden the spectrum of both basic biology and applied systems biology studies of *E. coli* metabolism.

**Keywords:** computational biology, group contribution method, systems biology, thermodynamics

**Introduction** Go to: ▾

The process of extracting biochemical content from genome annotations and literature sources to

<https://pubmed.ncbi.nlm.nih.gov/17593909/>

# Assigning gene function

- How do we know the function of a gene?

1. Genome sequencing, finding open reading frames (start/stop codons), and hence coding sequences for genes

2. Assign gene functions based on sequence similarity (homology) to other annotated genome sequences

3. Expressing protein, characterising it experimentally, determining protein structure

<https://biocyc.org/gene?orgid=ECOLI&id=EG10024>

The screenshot shows the EcoCyc genome browser interface for the *Escherichia coli* K-12 substr. MG1655 reference genome. The top navigation bar includes links for "Change Current Database" (set to EcoCyc), "Search in Current Database" (with a search bar), and "show operations". The main content area displays a genomic map of the chromosome, specifically the aceE region. The map shows various genes (e.g., mutT, zapD, coaE, guaC, hofC, hoF, ppdD, nadC, ampD, ampE, aceE, pdhR, lpd, aceF, acnB, acnB, yacL, spaD) with their start and end coordinates (bp) and orientations. A legend indicates protein genes (orange arrows), RNA genes (blue arrows), transcription start sites (green arrows), and terminators (red arrows). Gene color indicates operon membership. Below the map is a detailed description of the pyruvate dehydrogenase complex from *E. coli*, including its substrate specificity and reaction mechanism. The bottom section of the screenshot shows a BLAST search results page for the pyruvate dehydrogenase E1 component (NP\_414656.1) against the *Staphylococcus aureus* database. The results table lists top hits, including various homologs of the E1 component from *S. aureus* with their percent identity, E-value, and accession numbers.

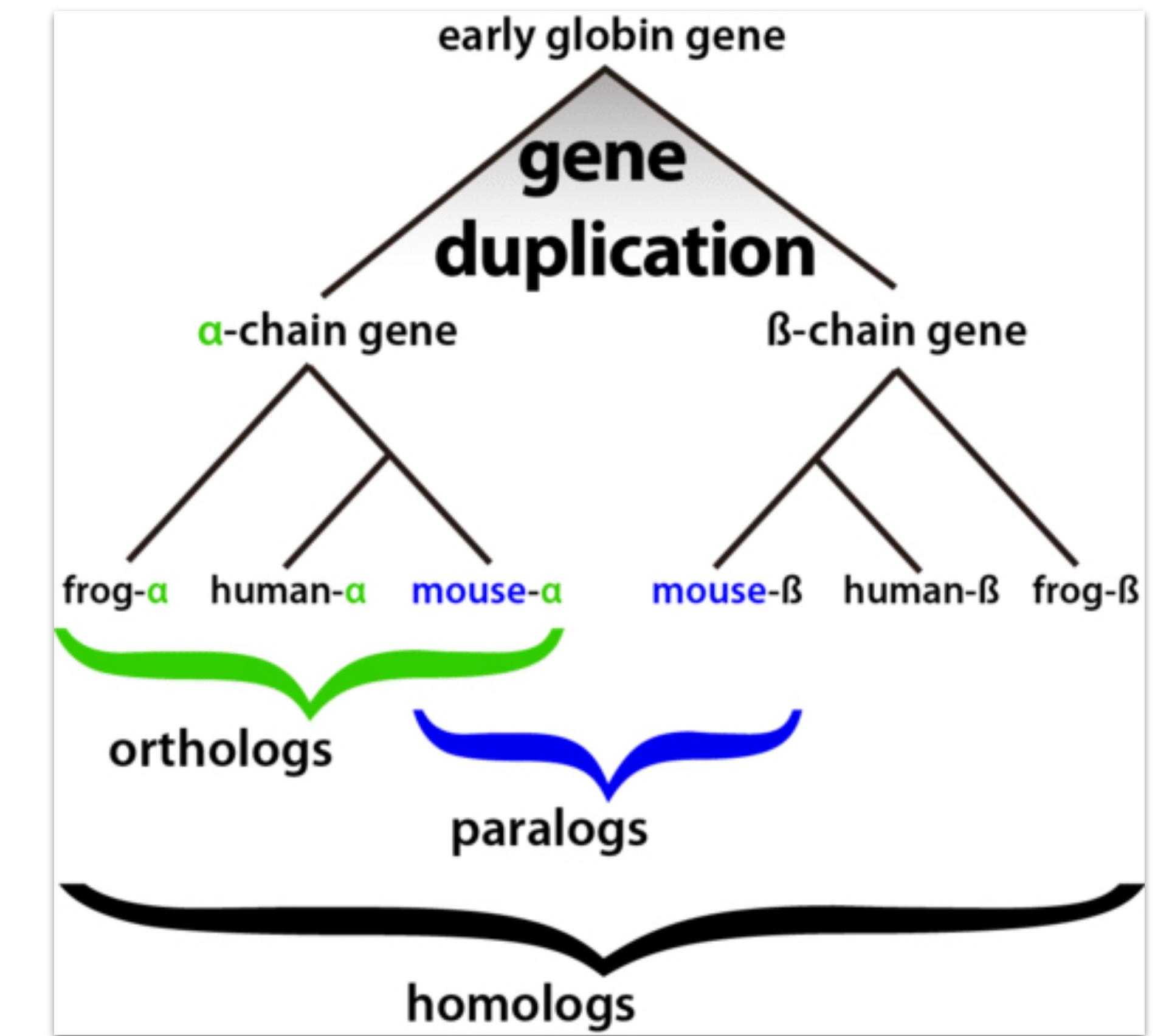
<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

[https://www.jbc.org/article/S0021-9258\(19\)70050-7/pdf](https://www.jbc.org/article/S0021-9258(19)70050-7/pdf)

# Homologs, orthologs, and paralogs

What's the difference?

- "Paralogs" arise from a gene duplication event in an organism or a distant ancestor and often acquire a different function
- "Orthologs" occur from a speciation event, so orthologs in two species, e.g. human and mouse  $\alpha$ -globins, will share a function
- "Homologs" can describe both paralogs and orthologs, but it's a vague term and I wouldn't encourage its use!
- **We're looking for orthologs**
- There are also analogs, xenologs, and more!
- Find out more here: [https://en.wikipedia.org/wiki/Sequence\\_homology](https://en.wikipedia.org/wiki/Sequence_homology)



# What are the pitfalls?

- Is your genome similar enough to *E. coli* to transfer gene functions?
  - **Use a more closely-related, well-characterised genome**
- Could two proteins share similar sequence but have different functions?
  - **Yes! These could be paralogs or promiscuous enzymes**
- What if we match against a protein that shares a domain (part of the sequence) but has a different overall function?
  - **We can set thresholds on our sequence searches to reduce the likelihood of this: % identity and % coverage**
- Finding the true ortholog is not easy!
  - **Building a phylogenetic tree is the best way to work this out**

# Picking the right starting point

So, we need to do the following:

1. Pick a closely related (to our "unknown" organism), well-characterised genome, with an FBA model, as our starting point.
2. Carefully match gene/protein sequences from our "known" organism against our "unknown" organism's gene/protein sequences (using BLAST).
3. Filter down the set of reactions in the "known" FBA model to only those with matches/hits in our "unknown" organism.

And that's just the start!

# The BiGG database

- Biochemical, Genetic, and Genomic database
- 108 FBA models, 48 single strain models, 30 species
- Includes human, mouse, yeast, parasites, quite a few species of bacteria
- 58 *E. coli* models!! (more than half of the models in the database)
- There are lots more FBA models elsewhere (e.g. with publications), but...
- ...the models in BiGG follow sensible rules for naming reactions and metabolites
- They have an API for grabbing data!

**BiGG Models**

Home Advanced Search Data Access Memote Validator [?](#)

Exclude multistrain models from search

**Search Results [?](#)**

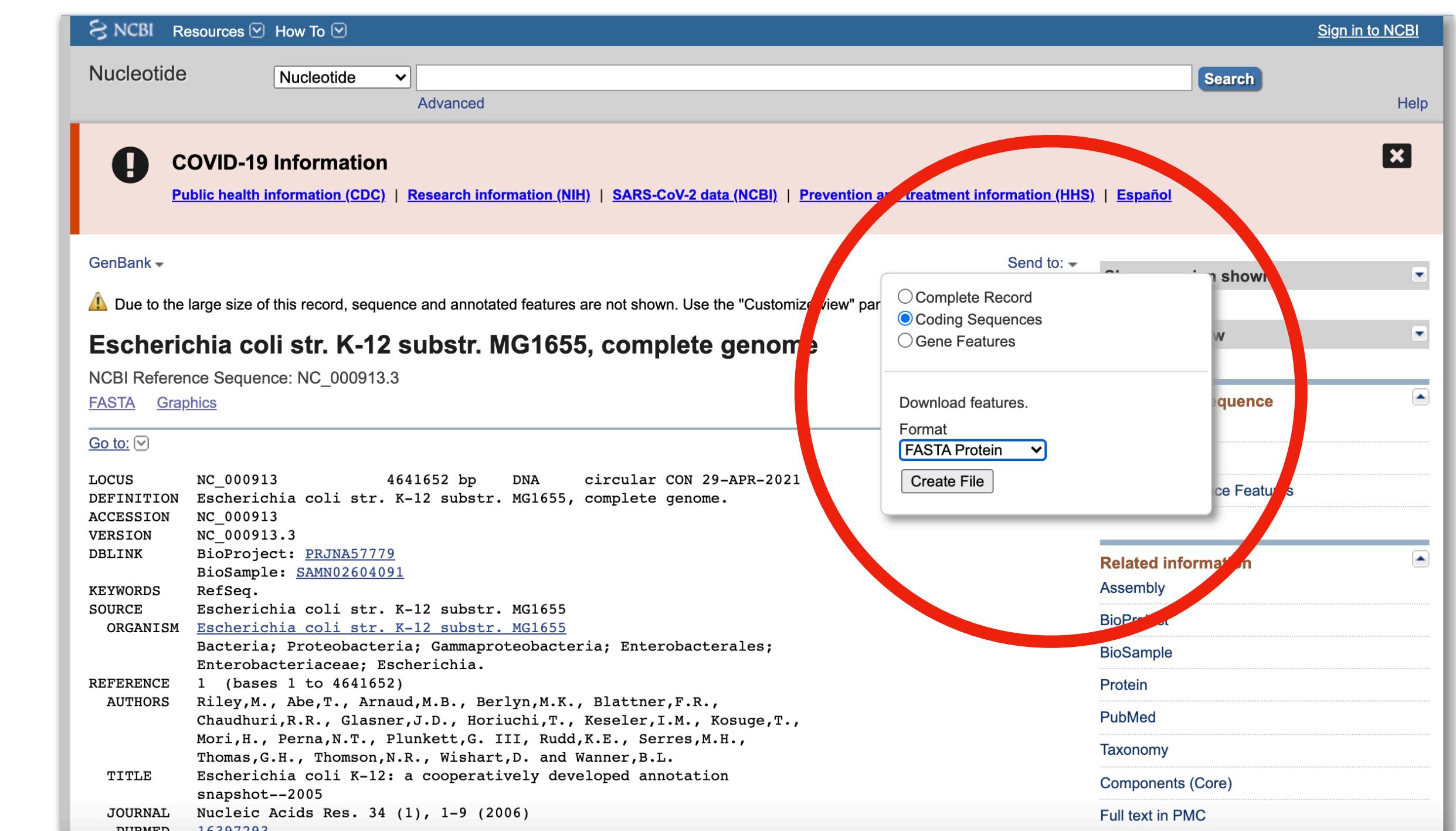
Models

BiGG ID	Organism	Metabolites	Reactions	Genes
iCN718	Acinetobacter baumannii AYE	888	1015	709
iYO844	Bacillus subtilis subsp. subtilis str. 168	990	1250	844
iRC1080	Chlamydomonas reinhardtii	1706	2191	1086
iCN900	Clostridioides difficile 630	885	1229	900
iHN637	Clostridium ljungdahlii DSM 13528	698	785	637
iCHOv1_DG44	Cricetulus griseus	2751	3942	1184
iCHOv1	Cricetulus griseus	4456	6663	1766
iEC042_1314	Escherichia coli 042	1926	2714	1314
iECP_1309	Escherichia coli 536	1941	2739	1309
iEC55989_1330	Escherichia coli 55989	1953	2756	1330
iECABU_c1320	Escherichia coli ABU 83972	1942	2731	1320
iAPEC01_1312	Escherichia coli APEC O1	1942	2735	1313
iEcoIC_1368	Escherichia coli ATCC 8739	1969	2768	1368

<http://bigg.ucsd.edu/models>

# Finding the "known" protein sequences

- The BiGG model page should have a link to the genome sequence for the "known" species on NCBI
- Download the **coding sequences** in **FASTA Protein** format, making sure to give the file a sensible name (the default is "sequence")
- Some of you will be using the *E. coli* K-12 MG1655 FBA model (called **iAF1260**), and you should use the link on the right to get the protein sequences
- The others are using the *S. aureus* N315 FBA model (called **iSB619**), and you should use this link to get your protein sequences: [https://www.ncbi.nlm.nih.gov/nuccore/NC\\_002745.2](https://www.ncbi.nlm.nih.gov/nuccore/NC_002745.2)



[https://www.ncbi.nlm.nih.gov/nuccore/NC\\_000913.3](https://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3)

# Finding the "unknown" protein sequences

- You'll need to pick a closely related species or another strain of the same species as your "known" genome/model
- The best place to find these in the NCBI nucleotide database using the filtering options or advanced search to limit it to the right species and RefSeq genomes that are "high quality".
- <https://www.ncbi.nlm.nih.gov/nuccore>
- Make sure to ignore plasmids. You can do this with the advanced search by giving a sequence length that is in a similar range to your "known" genome, e.g. 2,500,000 – 10,000,000 for *E. coli*
- Also watch out for the records that say "This entry is the master record... and contains no sequence data."
- I used the advanced search: (srcdb\_refseq[PROP]) AND ("Escherichia coli"[porgn]) AND ("2500000"[SLEN] : "10000000"[SLEN])

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (srcdb\_refseq[PROP]) AND ("Escherichia coli"[porgn]) AND ("2500000"[SLEN] : "10000000"[SLEN]) Search Create alert Advanced

COVID-19 Information Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Species Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

Bacteria (26,766) Customize ...

Molecule types genomic DNA/RNA (26,766) Customize ...

Source databases RefSeq (26,766) Customize ...

Sequence Type Nucleotide (26,766)

Genetic compartments Plasmid (3)

Sequence length Custom range...

Release date Custom range...

Revision date Custom range...

Items: 1 to 20 of 26766 << First < Prev Page 1 of 1339 Next > Last >>

1. Escherichia coli strain 1566m1, whole genome shotgun sequencing project  
5,632,019 bp other DNA  
This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.  
Accession: NZ\_JAJHPW000000000.1 GI: 2136361147  
GenBank

2. Escherichia coli strain 1566m1\_00000F\_arrow, whole genome shotgun sequence  
3,303,149 bp linear DNA  
Accession: NZ\_JAJHPW010000007.1 GI: 2136361049  
Assembly  
GenBank FASTA Graphics

3. Escherichia coli strain EF268, whole genome shotgun sequencing project  
5,115,263 bp other DNA  
This entry is the master record for a whole genome shotgun sequencing project and contains no sequence data.  
Accession: NZ\_JAJHJU000000000.1 GI: 2136360291  
GenBank

Results by taxon Top Organisms [Tree]  
Escherichia coli (26766)  
More...

Find related data Database: Select Find items

Search details srcdb\_refseq[PROP] AND "Escherichia coli"[Primary Organism] AND ("2500000"[SLEN] : "10000000"[SLEN])

Recent activity Turn Off Clear

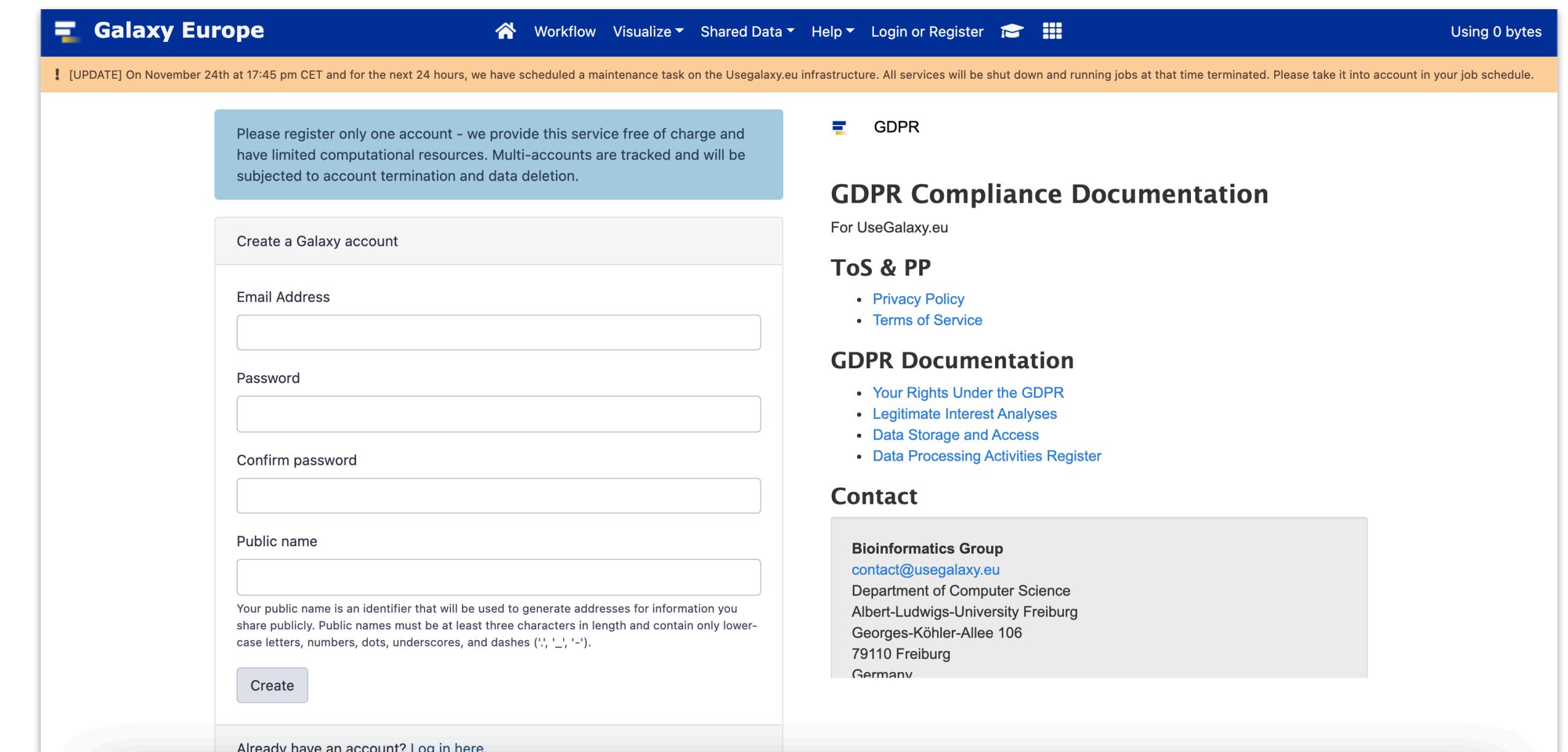
[https://www.ncbi.nlm.nih.gov/nuccore?term=%28srcdb\\_refseq%5BPROP%5D%29%20AND%20%28%22Escherichia%20coli%22%5Bporgn%5D%29%20AND%20%28%222500000%22%5BSLEN%5D%29&cmd=DetailsSearch](https://www.ncbi.nlm.nih.gov/nuccore?term=%28srcdb_refseq%5BPROP%5D%29%20AND%20%28%22Escherichia%20coli%22%5Bporgn%5D%29%20AND%20%28%222500000%22%5BSLEN%5D%29&cmd=DetailsSearch)

# Reciprocal BLAST hits (RBH)

- **BLAST** is a sequence similarity search method that's widely used
- Usually you'll search one or more sequences against a database of sequences to find the most similar matches (hits)
- You measure similarity by:
  - **% identity** (prop. of matching letters in the alignment)
  - **% coverage** (prop. of your query sequence covered by the alignment)
  - E value (probability that the match in the database occurred at random)
- A reciprocal BLAST hit BLASTs the hit back against all of the original sequences to ensure the original sequence is found again

# Galaxy (usegalaxy.eu)

- Galaxy is a web tool for running bits of bioinformatics software, like BLAST, that would normally be on the command line
- You can sign up for a free account here:  
<https://usegalaxy.eu/login>
- **Galaxy has a reciprocal BLAST function that we can use with our two sets of protein sequences to find matches**



# Create a new history

The screenshot shows the Galaxy Europe web interface. At the top, there is a navigation bar with links for Workflow, Visualize, Shared Data, Help, User, and a grid icon. A message at the top indicates a scheduled maintenance task from November 24th at 17:45 pm CET for 24 hours. The main content area includes a sidebar with tool categories like Tools, Get Data, Send Data, Collection Operations, and various file manipulation tools. A central box contains an update about the maintenance and a COVID-19 Research section. On the right, there is a History panel which is currently empty. News and Events sections are also present.

! [UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

**Tools**

- search tools
- Upload Data**

**Get Data**

**Send Data**

**Collection Operations**

**GENERAL TEXT TOOLS**

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group

**GENOMIC FILE MANIPULATION**

- Convert Formats
- FASTA/FASTQ
- Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore

Using 3%

**-- UPDATE -- 24 hours downtime starting from November 24th, 2021 at 17:45 pm CET**

On November 24th, 2021 at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated.

During this time we will be upgrading Galaxy to the latest version (21.09), migrating/upgrading the DB server and performing other maintenance.

Please take it into account in your job schedule.

**COVID-19 Research!**

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the [Galaxy SARS-CoV-2 portal](#). We mirror all public SARS-CoV-2 data from ENA in a [Galaxy data library](#) for your convenience. The Galaxy community has created [COVID-19 dedicated training materials](#). Please check our [recent activities](#) for more details.

If you need help submitting your data to public archives, like ENA, please [get in touch](#). We will support you in sharing your data.

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

**News**

Nov 13, 2021 **Training Infrastructure feedback: FORCeS eScience course**

Nov 13, 2021

**Events**

Nov 2, 2021 - Nov 23, 2021 **Forces 2021**

Nov 25, 2021 **Galaxy Metabolomics Community** OPEN CHAT

History

search datasets **Create new history**

**E. coli reciprocal BLAST**

(empty)

This history is empty. You can load your own data or get data from an external source

# Click on "load your own data", select your two fasta files, and set type to fasta, click "Start"

The screenshot shows the Galaxy Europe web interface. On the left, a sidebar lists various tools and data types under categories like Tools, Get Data, Send Data, Collection Operations, and several genomic file manipulation sections. A central modal window titled "Download from web or upload from disk" displays a table of two uploaded files:

Name	Size	Type	Genome	Settings	Status
ec_k12_prots.fasta	2.2 MB	fasta	----- Additional ...	⚙️	0%
ec_um146_prots.fasta	2.3 MB	fasta	----- Additional ...	⚙️	0%

Below the table, there are filters for "Type (set all)" (set to "fasta") and "Genome (set all)" (set to "----- Additional ..."). At the bottom of the modal are buttons for "Choose local files", "Choose remote files", "Paste/Fetch data", "Start" (which is highlighted in blue), "Pause", "Reset", and "Close".

The main workspace on the right shows a history panel with a message: "This history is empty. You can load your own data or get data from an external source". Below the history is a news section with a card about "Training Infrastructure feedback: FORCeS eScience course" and an events section with cards for "Forces 2021" and "Galaxy Metabolomics Community".

# Search the Tools for "blast" and select "BLAST Reciprocal Best Hits (RBH)"

The screenshot shows the Galaxy Europe web interface. In the top left, the logo 'Galaxy Europe' is displayed. The top navigation bar includes links for Home, Workflow, Visualize, Shared Data, Help, User, and a grid icon. A progress bar at the top right indicates 'Using 3%'. A message at the top states: '[UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.'

In the left sidebar under 'Tools', the search term 'blast' is entered in a search bar, and a green button labeled 'Upload Data' is visible. Below the search bar, there are several tool entries:

- NCBI BLAST+ makeprofiledb Make profile database
- NCBI BLAST+ dustmasker masks low complexity regions
- NCBI BLAST+ segmasker low-complexity regions in protein sequences
- MAF-convert read MAF-format alignments and write them in another format.
- BLAST Reciprocal Best Hits (RBH)** from two FASTA files (selected)
- Trinotate
- SPRING Map with BLAST
- LUMPY preprocessing extracts discordant read pairs and split-read alignments from a BAM dataset
- MiGMAP mapper for full-length T- and B-cell repertoire sequencing

A central box contains an 'UPDATE' message: '-- UPDATE -- 24 hours downtime starting from November 24th, 2021 at 17:45 pm CET'. It details the maintenance schedule and upgrade process. Below this is a 'COVID-19 Research!' section with information about SARS-CoV-2 analysis and training materials.

The right sidebar shows the 'History' section with two datasets listed:

- E. coli reciprocal BLAST (2 shown, 4.5 MB)
- 2: ec\_um146\_prot.fasta
- 1: ec\_k12\_prot.fasta

At the bottom, a quote by Prof. Stephen Hawking is displayed: "Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

Two news items are listed:

- Nov 13, 2021: Training Infrastructure feedback: FORCeS eScience course
- Nov 2, 2021 - Nov 23, 2021: Forces 2021

An event is also listed: Nov 25, 2021: Metabolomics Community Meeting.

At the bottom right, there is a green button labeled 'OPEN CHAT'.

# Select your two fasta files, change the "Molecule type" to "protein", and "blastp", then click "Execute"

The screenshot shows the Galaxy Europe web interface with the following details:

- Tools Panel:** Shows a search bar with "blast" and a "blast" entry. Below it are buttons for "Upload Data" and "Show Sections".
- Job Overview:** A message indicates a maintenance task scheduled on November 24th at 17:45 pm CET.
- Tool Configuration:**
  - BLAST Reciprocal Best Hits (RBH):** Galaxy Version 0.1.11. This is the active tool.
  - Inputs:** Species A: "ec\_k12\_prot.fasta" (uploaded). Species B: "ec\_um146\_prot.fasta" (uploaded).
  - Molecule type of FASTA inputs:** Set to "protein".
  - Type of BLAST:** Set to "blastp" (Traditional BLASTP to compare a protein query to a protein database).
  - Minimum percentage identity for BLAST matches:** Set to 70%.
  - Minimum percentage query coverage for BLAST matches:** Set to 50%.
- History:** Shows previous runs:
  - E. coli reciprocal BLAST (2 shown, 4.5 MB)
  - 2: ec\_um146\_prot.fasta
  - 1: ec\_k12\_prot.fasta

# Wait for the reciprocal BLAST task to turn from orange to green when it's complete

The screenshot shows the Galaxy Europe web interface. On the left, a sidebar lists various tools: blast, NCBI BLAST+ makeprofiledb, NCBI BLAST+ dustmasker, NCBI BLAST+ segmasker, MAF-convert, BLAST Reciprocal Best Hits (RBH), Trinotate, SPRING Map with BLAST, LUMPY preprocessing, and MiGMAP. The main area displays a successful execution of the 'blast' tool, which used two inputs (ec\_k12\_prots.fasta and ec\_um146\_prots.fasta) and produced one output (BLAST RBH: ec\_k12\_prots.fasta vs ec\_um146\_prots.fasta). A message encourages users to cite the Galaxy platform. The right side shows a history panel with three entries: E. coli reciprocal BLAST, 3: BLAST RBH: ec\_k12\_prots.fasta vs ec\_um146\_prots.fasta, and 2: ec\_um146\_prots.fasta. The first entry is highlighted in orange, while the others are green.

Galaxy Europe

Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾

Using 3%

[UPDATE] On November 24th at 17:45 pm CET and for the next 24 hours, we have scheduled a maintenance task on the Usegalaxy.eu infrastructure. All services will be shut down and running jobs at that time terminated. Please take it into account in your job schedule.

**Tools**

blast

Upload Data

Show Sections

NCBI BLAST+ makeprofiledb Make profile database

NCBI BLAST+ dustmasker masks low complexity regions

NCBI BLAST+ segmasker low-complexity regions in protein sequences

MAF-convert read MAF-format alignments and write them in another format.

**BLAST Reciprocal Best Hits (RBH)** from two FASTA files

Trinotate

SPRING Map with BLAST

LUMPY preprocessing extracts discordant read pairs and split-read alignments from a BAM dataset

MiGMAP mapper for full-length T- and B-cell repertoire sequencing

Executed **BLAST Reciprocal Best Hits (RBH)** and successfully added 1 job to the queue.

The tool uses 2 inputs:

- 1: ec\_k12\_prots.fasta
- 2: ec\_um146\_prots.fasta

It produces this output:

- 3: BLAST RBH: ec\_k12\_prots.fasta vs ec\_um146\_prots.fasta

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

We need your support ...

If Galaxy helped with the analysis of your data, please do not forget to **cite**:

Afgan E et al. 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 44, W3–W10 doi:10.1093/nar/gkw343

And please **acknowledge** the Freiburg Galaxy server:

The Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre

**History**

search datasets

E. coli reciprocal BLAST

3 shown

4.5 MB

3: BLAST RBH: ec\_k12\_prots.fasta vs ec\_um146\_prots.fasta

2: ec\_um146\_prots.fasta

1: ec\_k12\_prots.fasta

# Have a look at your results, then click the little disk icon to download and save the file

Screenshot of the Galaxy Europe interface showing a BLAST search results table and a history panel.

**Tools:**

- blast
- Upload Data
- Show Sections
- NCBI BLAST+ makeprofiledb
- NCBI BLAST+ dustmasker
- NCBI BLAST+ segmasker
- MAF-convert
- BLAST Reciprocal Best Hits (RBH) from two FASTA files
- Trinotate
- SPRING Map with BLAST
- LUMPY preprocessing
- MiGMAP mapper

**Table Headers:** #A\_id, B\_id, A\_length, B\_length, A\_qcovhsp

**Table Data:**

#A_id	B_id	A_length	B_length	A_qcovhsp
Icl NC_000913.3_prot_NP_414542.1_1	Icl NC_017632.1_prot_WP_001386572.1_4449	21	21	100
Icl NC_000913.3_prot_NP_414543.1_2	Icl NC_017632.1_prot_WP_001264707.1_4450	820	820	100
Icl NC_000913.3_prot_NP_414544.1_3	Icl NC_017632.1_prot_WP_000241660.1_4451	310	310	100
Icl NC_000913.3_prot_NP_414545.1_4	Icl NC_017632.1_prot_WP_000781042.1_4452	428	428	100
Icl NC_000913.3_prot_NP_414546.1_5	Icl NC_017632.1_prot_WP_000738733.1_4453	98	98	100
Icl NC_000913.3_prot_NP_414547.1_6	Icl NC_017632.1_prot_WP_000906188.1_4454	258	258	100
Icl NC_000913.3_prot_NP_414548.1_7	Icl NC_017632.1_prot_WP_001112615.1_4455	476	476	100
Icl NC_000913.3_prot_NP_414549.1_8	Icl NC_017632.1_prot_WP_000130187.1_4456	317	317	100
Icl NC_000913.3_prot_NP_414550.1_9	Icl NC_017632.1_prot_WP_001094685.1_4457	195	195	100
Icl NC_000913.3_prot_NP_414551.1_10	Icl NC_017632.1_prot_WP_000528533.1_4458	188	188	100
Icl NC_000913.3_prot_NP_414552.1_11	Icl NC_017632.1_prot_WP_001102393.1_4459	237	237	100
Icl NC_000913.3_prot_NP_414554.1_13	Icl NC_017632.1_prot_WP_000843687.1_4460	134	134	100
Icl NC_000913.3_prot_NP_414555.1_14	Icl NC_017632.1_prot_WP_000516135.1_4461	638	638	100
Icl NC_000913.3_prot_NP_414556.1_15	Icl NC_017632.1_prot_WP_001118464.1_4462	376	376	100
Icl NC_000913.3_prot_NP_414559.1_17	Icl NC_017632.1_prot_WP_000809168.1_4463	69	50	72
Icl NC_000913.3_prot_NP_414560.1_19	Icl NC_017632.1_prot_WP_000681386.1_4467	388	388	100
Icl NC_000913.3_prot_NP_414561.1_20	Icl NC_017632.1_prot_WP_000062878.1_4468	301	301	100
Icl NC_000913.3_prot_NP_414564.1_23	Icl NC_017632.1_prot_WP_001274021.1_4469	87	87	100
Icl NC_000913.3_prot_NP_414565.1_24	Icl NC_017632.1_prot_WP_001337277.1_4470	72	72	100
Icl NC_000913.3_prot_NP_414566.1_25	Icl NC_017632.1_prot_WP_000767329.1_4471	313	313	100
Icl NC_000913.3_prot_NP_414567.1_26	Icl NC_017632.1_prot_WP_001286813.1_4472	938	938	100
Icl NC_000913.3_prot_NP_414568.1_27	Icl NC_017632.1_prot_WP_000083369.1_4473	164	164	100
Icl NC_000913.3_prot_NP_414569.1_28	Icl NC_017632.1_prot_WP_000004655.1_4474	149	149	100
Icl NC_000913.3_prot_NP_414570.1_29	Icl NC_017632.1_prot_WP_001166403.1_4475	316	316	100
Icl NC_000913.3_prot_NP_414571.1_30	Icl NC_017632.1_prot_WP_001239167.1_4476	304	304	99

**History:**

- E. coli reciprocal BLAST (3 shown, 4.88 MB)
- 3: BLAST RBH: ec\_k12\_rts.fasta vs ec\_um146\_prots.fasta (3,601 lines, 1 comments, format: tabular, database: ?)

**Download Icon:** A green circular icon with a white disk symbol, located at the bottom right of the history panel.

# Things to think about

- Check the **number of sequences** in both your "known" fasta file, and the "unknown" fasta file. You can find this by clicking the fasta filename in the history bar at the side to expand it.
- Check the number of lines in your BLAST results file the same way. The number of lines is the number of hits the "known" sequences had against the "unknown" ones.
- What proportion have hits?
- Do the hits have high percentage identity and coverage?
- What would have happened if we'd changed the default settings for the identity and coverage before running the reciprocal BLAST?

*(the defaults (>70% identity, >50% coverage) are a good rule of thumb)*

# The perennial bioinformatics problem: IDs!

- Our fasta sequences look like this:

```
>lcl|NC_000913.3_prot_NP_414579.4_37 [gene=caIC] [locus_tag=b0037] ...
MDIIGGQHLRQMWDLADVYGHKTALICESSGGVNRYSYLELNQEINRTANLFYTLGIRKGDKVALHLD
NCPEFIFCWFGGLAKIGAIMVPINARLLCEESAWILQNSQACLLVTSAQFYPMYQQIQQQEDATQLRHICLT
DVALPADDGVSSFTQLKNQQPATLCYAPPLSTDDEAEIFLFTSGTTSRPKGVVITHYNLRFAGYYSAWQCA
LRDDDVYLTVMMPAFHIDCQCTAACMAAFSAGATFVLVEKYSARAFWGQVQKYRATVTECIPMMIRTLMVQP
PSANDQQHRLREVMFYLNLSSEQEKDAFCERFGVRLLTSGMTETIVGIIGDRPGDKRRWPSIGRVGFCYE
AEIRDDHNRPLPAGEIGEICIKGIPGKTIFKEYFLNPQATAKVLEADGWLHTGDTGYRDEEDFFYFVDRR
CNMIKRGGENVSCVELENIIAAHPKIQDIVVVGKDSIRDEAIKAFVVLNEGETLSEEEFFRFCEQNMAK
FKVPSYLEIRKDLPRNCSGKIIRKNLK
```

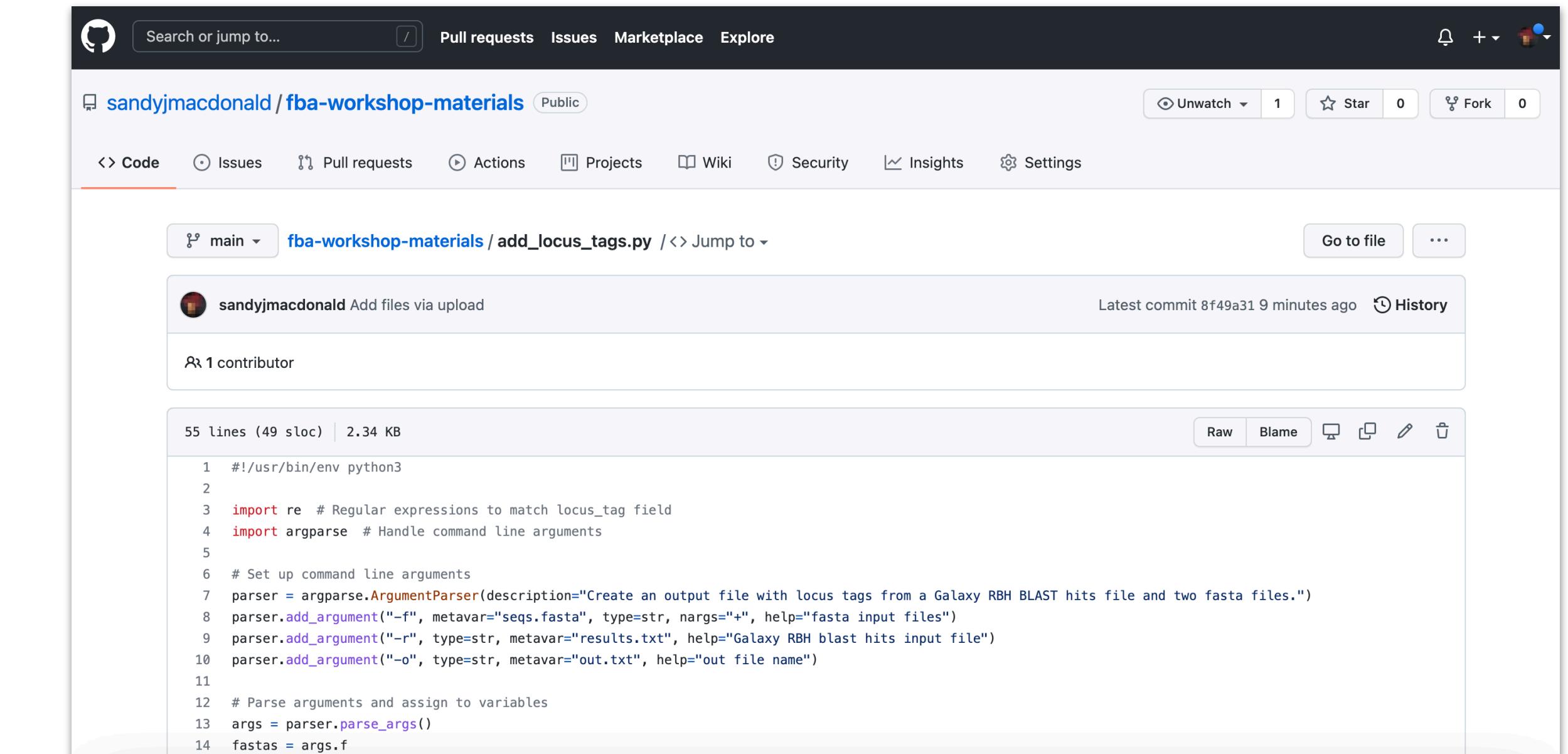
- But the gene reaction rules in the model look like this:

```
{"id": "CRNDCL2",
"name": "D-Carnitine-CoA Ligase",
"metabolites":
{
    "adp_c": 1.0,
    "atp_c": -1.0,
    "coa_c": -1.0,
    "crnDcoa_c": 1.0,
    "crn_D_c": -1.0,
    "pi_c": 1.0
},
"lower_bound": 0.0,
"upper_bound": 999999.0,
"gene_reaction_rule": "b0037",
"subsystem": "Oxidative Phosphorylation"},
```

# Python to the rescue!

## My add\_locus\_tags.py script

- I've written a Python script that takes the two fasta files, the blast results file, and the name you want to call your output file, and it creates an output file with the matching locus tags on each line
- You can find it here: [https://github.com/sandyjmacdonald/fba-workshop-materials/blob/main/add\\_locus\\_tags.py](https://github.com/sandyjmacdonald/fba-workshop-materials/blob/main/add_locus_tags.py)
- Put the Python script in the same directory as your two fasta files and blast results file and then run it by typing this in the terminal:



The screenshot shows a GitHub repository page for 'fba-workshop-materials'. The 'Code' tab is selected, displaying the 'main' branch. The file 'add\_locus\_tags.py' is shown with 55 lines of code. The code is a Python script that imports re and argparse, sets up command line arguments for input files and output file, and then parses the arguments to assign them to variables.

```
1  #!/usr/bin/env python3
2
3  import re # Regular expressions to match locus_tag field
4  import argparse # Handle command line arguments
5
6  # Set up command line arguments
7  parser = argparse.ArgumentParser(description="Create an output file with locus tags from a Galaxy RBH BLAST hits file and two fasta files.")
8  parser.add_argument("-f", metavar="seqs.fasta", type=str, nargs="+", help="fasta input files")
9  parser.add_argument("-r", metavar="results.txt", type=str, help="Galaxy RBH blast hits input file")
10 parser.add_argument("-o", metavar="out.txt", type=str, help="out file name")
11
12 # Parse arguments and assign to variables
13 args = parser.parse_args()
14 fastas = args.f
```

```
python3 add_locus_tags.py -f known.fasta unknown.fasta -r blast_results.tabular -o results_locus_tags.txt
```

# Filtering your model

- With our list of locus tags that match from our "known" genome against our "unknown" genome, we can now go through the reactions in the "known" model and check whether the genes in the gene reaction rule have matches
- What's the best way to do this filtering? Python again!
- **COBRApy** is a Python library for working with FBA models:  
<https://github.com/opencobra/cobrapy>
- We can use it to check our list of locus tags against each reaction, keep only the matching ones, and then swap the locus tags, replacing the "known" locus tags with the "unknown" (now also known!) ones

# Python to the rescue again!

## My filter\_fba\_model.py script

- To make the process a bit easier, I've made a Python script that uses COBRApy to filter an FBA model in json format if you give it a text file. like the one we created, that maps the known locus tags to the unknown ones
- You can find it in the same GitHub repository: [https://github.com/sandyjmacdonald/fba-workshop-materials/blob/main/filter\\_fba\\_model.py](https://github.com/sandyjmacdonald/fba-workshop-materials/blob/main/filter_fba_model.py)
- Run it by typing, for example:

```
python3 filter_fba_model.py -l locus_tags_map.txt -m model.json
```

Model summary:

1138 genes matched from 1261 genes in original model (90%)

Missing genes:

b0423,b3629,b1800,b0731,b2538,b0341,b2092,b2542,b0616,b1907,b2484,b1296,b2485,b1385,b1363,b0351,b3627,b1488,b3628,b0349,b2027,b4177,b4474,b3579,b0615,b4291,b2487,b3213,b3875,b0353,b2530,b1297,b1302,b3374,b2036,b2541,b2920,b4287,b2203,b0333,b3990,b0260,b1483,b2486,b3991,b2482,b3577,b3578,b0732,b2481,b1484,b0614,b3502,b2038,b4031,b0584,b0150,b2037,b0340,b3625,b4301,b0339,b3622,b3648,b2539,b2492,b2205,b0261,b2490,b3626,b4321,b2204,b4288,b0650,b3927,b3212,b0352,b0617,b3623,b3624,b2094,b2540,b2489,b2393,b2034,b2202,b1102,b3992,s0001,b1486,b1398,b2488,b1298,b4407,b4289,b2035,b2917,b1801,b1485,b2093,b4120,b3370,b1386,b0348,b0273,b1301,b0070,b4356,b1300,b0347,b1778,b1006,b2483,b3371,b3966,b1487,b2919,b4290,b2032,b2010,b0350,b2033,b2206

2266 reactions added from 2382 reactions in original model (95%)

Missing reactions:

42A12B00Xpp,ACALDtpp,ADSS,ACONIs,ALAALAD,AD0CBLtonex,ALDD19xr,ALDD2x,AOBUTDs,AS03t8pp,ARBTNexs,ATPHs,CBItonex,CBL1tonex,CINNDO,CITL,C02tpp,CPGNexs,CYNTAH,CYNTt2pp,CPGNTonex,DATPHs,3HCINNMH,3HPPPNH,DGK1,DHCIND,DHCINDO,DHPPD,DHPTDCs,ECA40ALpp,FE3DCITabcpp,FE3DCITtonex,FALDtpp,FALGTHLs,FE3HOXexs,FE3HOXtonex,FECRMexs,FECRMtonex,FRULYSDG,FRULYSE,FRULYSK,FRULYSt2pp,FEENTERexs,FEENTERtonex,FRUURt2rpp,FEOXAMexs,FEOXAMtonex,GALCTNLt2pp,GALT1,GALTptsp,GGGABADs,GGGABADr,GGGABAH,GGPTRCO,GGPTRCS,GK1,GLUSy,GLCTR2,GLCTR3,GLYALDtpp,GLYBt2pp,GLYCtpp,HCINNMt2rpp,GTPHs,HCYSMT,HCYSMT2,HEPK2,HEPT4,H2St1pp,H2tpp,HKNDDH,HKNTDH,HOPNTAL,HPPPND0,HPPPNT2rpp,MALDDH,MANGLYCptspp,MANPGH,MELIBt2pp,MCITS,MMCD,METOX1s,METOX2s,MMETt2pp,METOXR2,MMM2,N03R1bpp,N03R2bpp,N0tpp,016A4Lpp,016AP1pp,016AP2pp,016AP3pp,016AT,016AUNDtpp,016GALFT,016GLCT1,PACCOAL,MOAT3C,O2tpp,OP4ENH,PEAMNOpp,PPCSCT,N20tpp,PPPNDO,PSCLYSt2pp,RHAT1,S02tpp,THZPSN,TDPDRE,URAt2pp\_copy2,UREAtpp,TYROXDApp,XYLUT2pp,XYLut2pp,UDP GALM

Filtered model written to example\_files/filtered\_iAF1260.json

# You're only just at the start!

- There's still a *lot* left to do!
- Your model will likely have a lot of gaps in it (and errors)
- Gap-filling, which COBRAPy can do, will give you a working model but ideally you need to find candidate genes for those gap-filled reactions
  - **Try re-running the BLAST with lower cutoffs to see if any of these missing genes turn up. Promiscuous enzymes? Side reactions?**
- Are all the assignments from our BLAST correct?
  - **Check if the genes in the "unknown" genome have been characterised.**
- Do any exchange reactions need to be added to make pathways work?
- What should the biomass reaction be?

# Give it a go...

- Run through this process and pick your own "unknown" strain or species
- I'd recommend starting with another strain of the same species and if your new model is too good then try a different species
- Does your model work at all?
- If not, then can you try gap-filling?

# Key things

- What "unknown" organism did you pick, and why?
- What proportion of the "known" genes had hits to "unknown" genes?
- Pick another set of BLAST cutoffs and see how this changes
- What proportion of genes and reactions are retained in your filtered model compared to the original model?
- Can you think of reasons why genes could be missing?

# Questions? Queries?

Pop me an email with any questions or queries:

[s.macdonald@york.ac.uk](mailto:s.macdonald@york.ac.uk)