# Proteome Comparison Service

ASM Microbe Workshop
June 1, 2017
New Orleans, LA

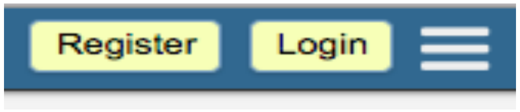
PATRIC
PATHOSYSTEMS RESOURCE INTEGRATION CENTER

# Background

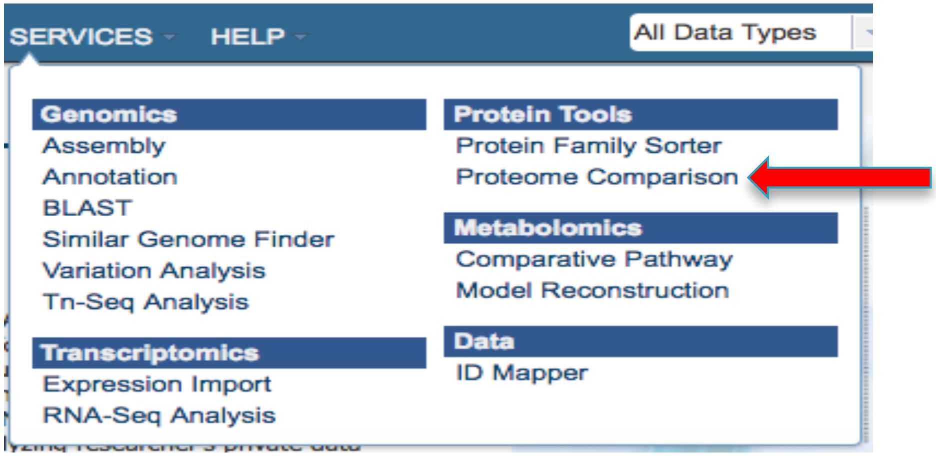
- Proteome Comparison tool can be used to identify insertions, deletions and protein homologs
- Use bi-directional blastp best hits to define homology
- The user selects a reference genome
- The user can add up to 9 genomes to compare to the reference genome
- Support both public and private genomes, a set of proteins saved in PATRIC as a feature group, and user fasta file

# The Proteome Comparison Service

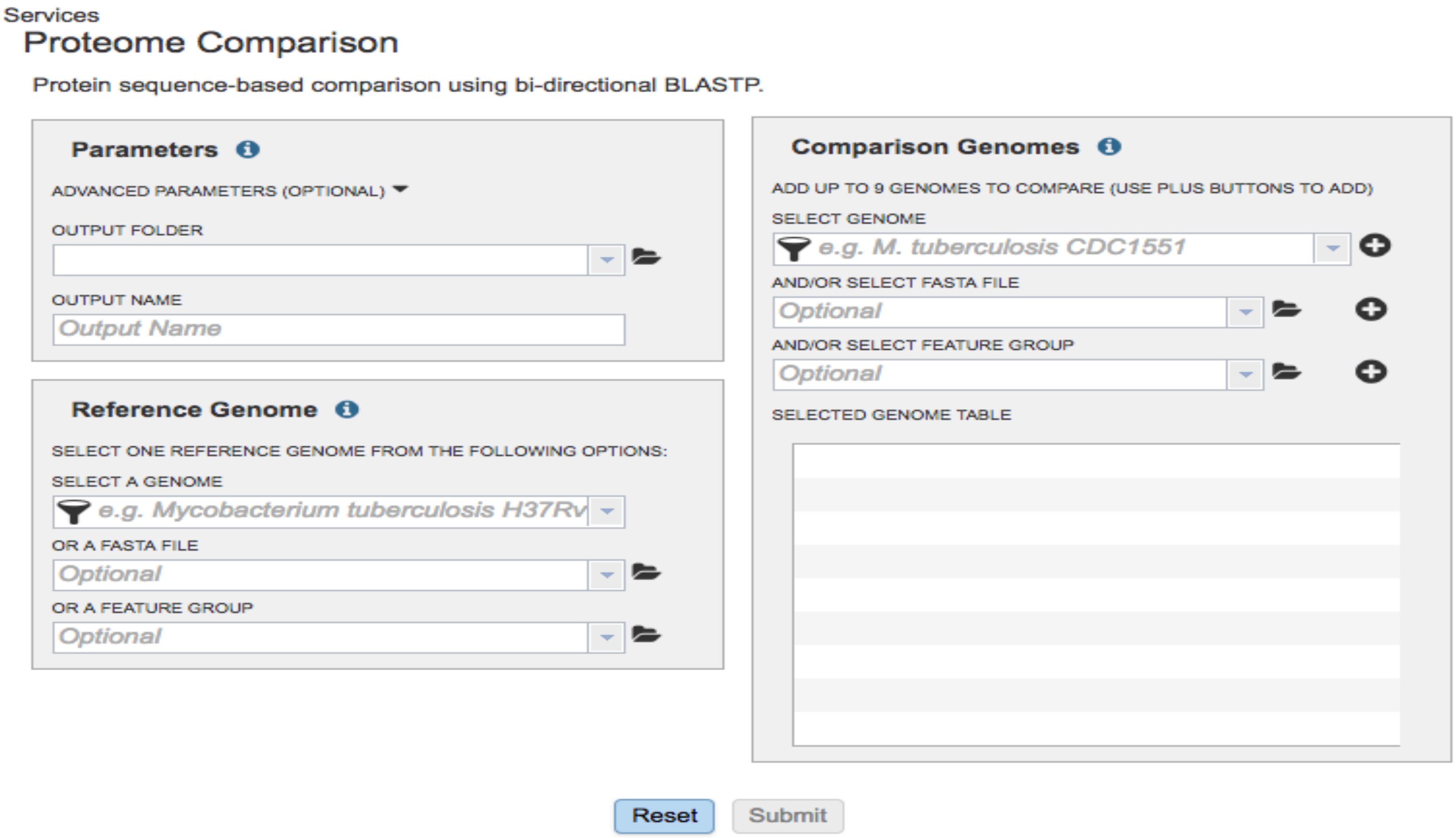
- Login to the PATRIC website at www.patricbrc.org



- On the PATRIC home page open the Services tab at the top of the page and select the Proteome Comparison service



# Proteome Comparison Submission Form



Services

Proteome Comparison

Protein sequence-based comparison using bi-directional BLASTP.

**Parameters**

ADVANCED PARAMETERS (OPTIONAL) ▾

OUTPUT FOLDER

OUTPUT NAME

Output Name

**Reference Genome**

SELECT ONE REFERENCE GENOME FROM THE FOLLOWING OPTIONS:

SELECT A GENOME

e.g. Mycobacterium tuberculosis H37Rv

OR A FASTA FILE

Optional

OR A FEATURE GROUP

Optional

**Comparison Genomes**

ADD UP TO 9 GENOMES TO COMPARE (USE PLUS BUTTONS TO ADD)

SELECT GENOME

e.g. M. tuberculosis CDC1551

AND/OR SELECT FASTA FILE

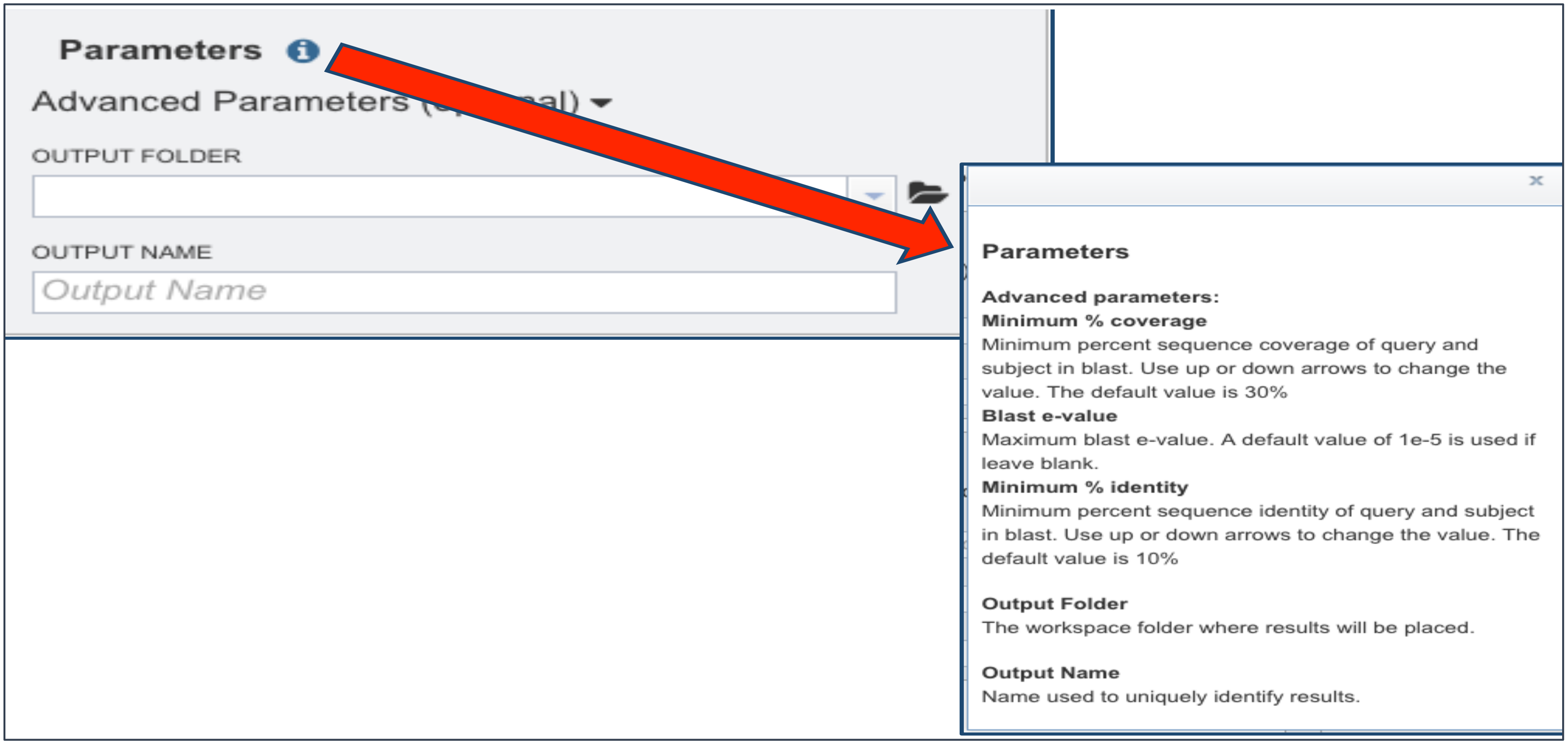
Optional

AND/OR SELECT FEATURE GROUP

Optional

SELECTED GENOME TABLE

# Setting parameters and selecting an output folder



**Parameters**

Advanced Parameters (optional)

OUTPUT FOLDER

OUTPUT NAME

*Output Name*

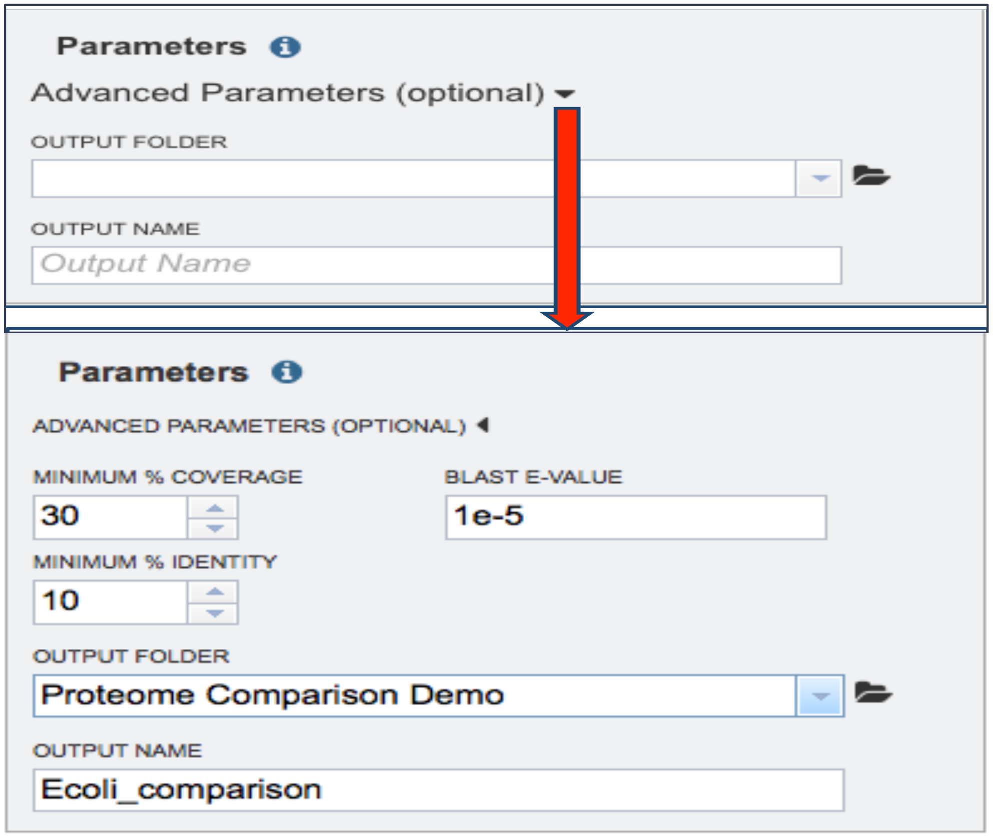
**Parameters**

**Advanced parameters:**
**Minimum % coverage**
Minimum percent sequence coverage of query and subject in blast. Use up or down arrows to change the value. The default value is 30%
**Blast e-value**
Maximum blast e-value. A default value of 1e-5 is used if leave blank.
**Minimum % identity**
Minimum percent sequence identity of query and subject in blast. Use up or down arrows to change the value. The default value is 10%

**Output Folder**
The workspace folder where results will be placed.

**Output Name**
Name used to uniquely identify results.

# Setting parameters and selecting an output folder



**Parameters**

Advanced Parameters (optional)

OUTPUT FOLDER

OUTPUT NAME

Output Name

**Parameters**

ADVANCED PARAMETERS (OPTIONAL)

MINIMUM % COVERAGE: 30

BLAST E-VALUE: 1e-5

MINIMUM % IDENTITY: 10

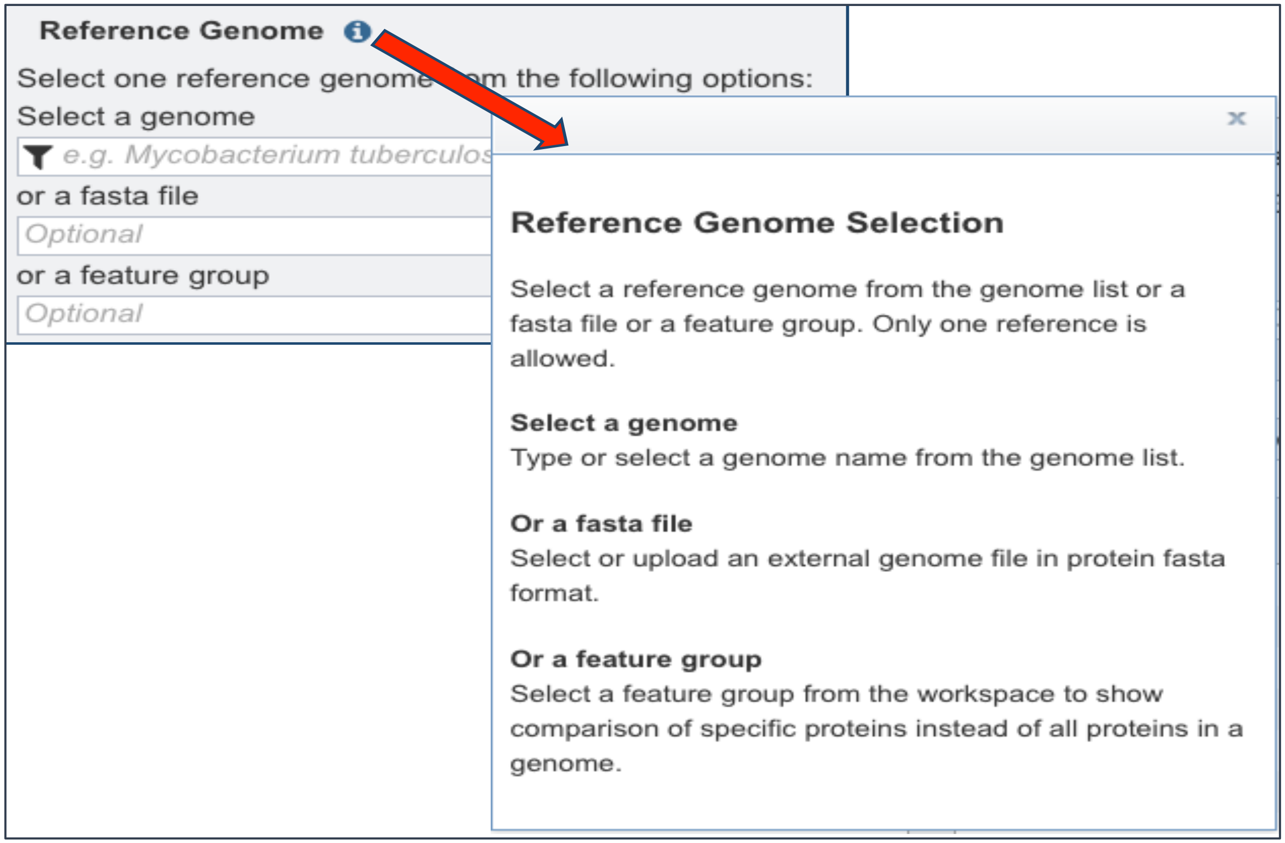
OUTPUT FOLDER: Proteome Comparison Demo

OUTPUT NAME: Ecoli_comparison

# Example

- Reference Genome
  - Escherichia coli str. K-12 substr. MG1655 (511145.12)
- Comparison Genomes
  - Escherichia coli strain swine65 (562.9957)
  - Escherichia coli strain MRSN388634 (562.10576)
  - Escherichia coli O104:H4 str. TY-2482 (1038844.18)
  - Escherichia coli O104:H4 str. GOS1 (1038927.4)

# Selecting the Reference Genome



**Reference Genome**

Select one reference genome from the following options:

Select a genome

or a fasta file

or a feature group

**Reference Genome Selection**

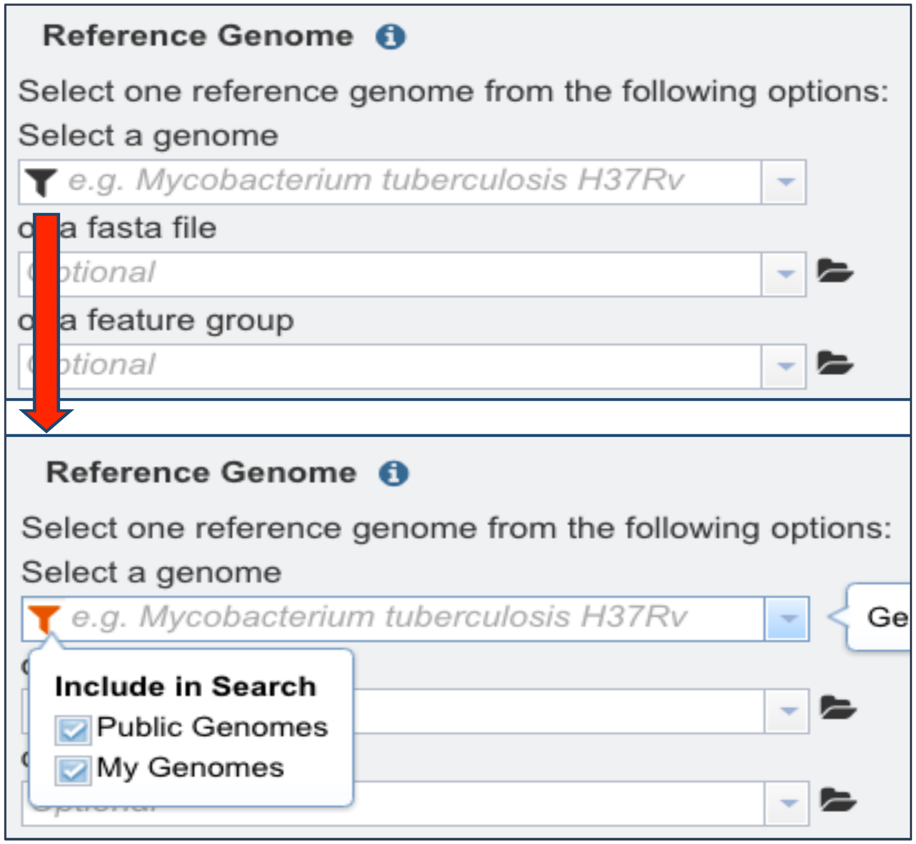
Select a reference genome from the genome list or a fasta file or a feature group. Only one reference is allowed.

**Select a genome**
Type or select a genome name from the genome list.

**Or a fasta file**
Select or upload an external genome file in protein fasta format.

**Or a feature group**
Select a feature group from the workspace to show comparison of specific proteins instead of all proteins in a genome.

# Selecting the Reference Genome



**Reference Genome**

Select one reference genome from the following options:

Select a genome

e.g. Mycobacterium tuberculosis H37Rv

or a fasta file

Optional

or a feature group

Optional

**Reference Genome**

Select one reference genome from the following options:

Select a genome

e.g. Mycobacterium tuberculosis H37Rv

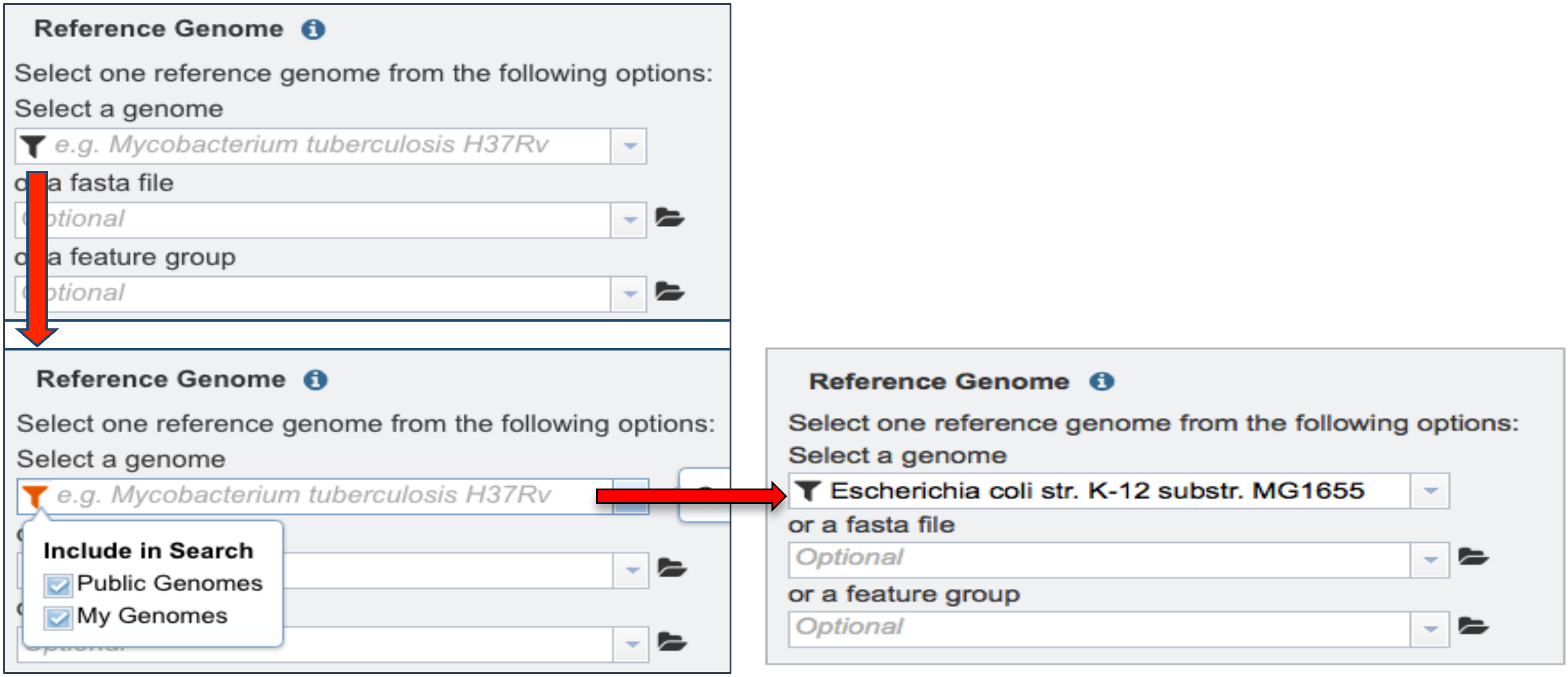
Ge

**Include in Search**

☑ Public Genomes

☑ My Genomes

# Selecting the Reference Genome



**Reference Genome**

Select one reference genome from the following options:

Select a genome

e.g. Mycobacterium tuberculosis H37Rv

or a fasta file

Optional

or a feature group

Optional

**Reference Genome**

Select one reference genome from the following options:

Select a genome

e.g. Mycobacterium tuberculosis H37Rv

**Include in Search**

☑ Public Genomes

☑ My Genomes

**Reference Genome**

Select one reference genome from the following options:

Select a genome

Escherichia coli str. K-12 substr. MG1655

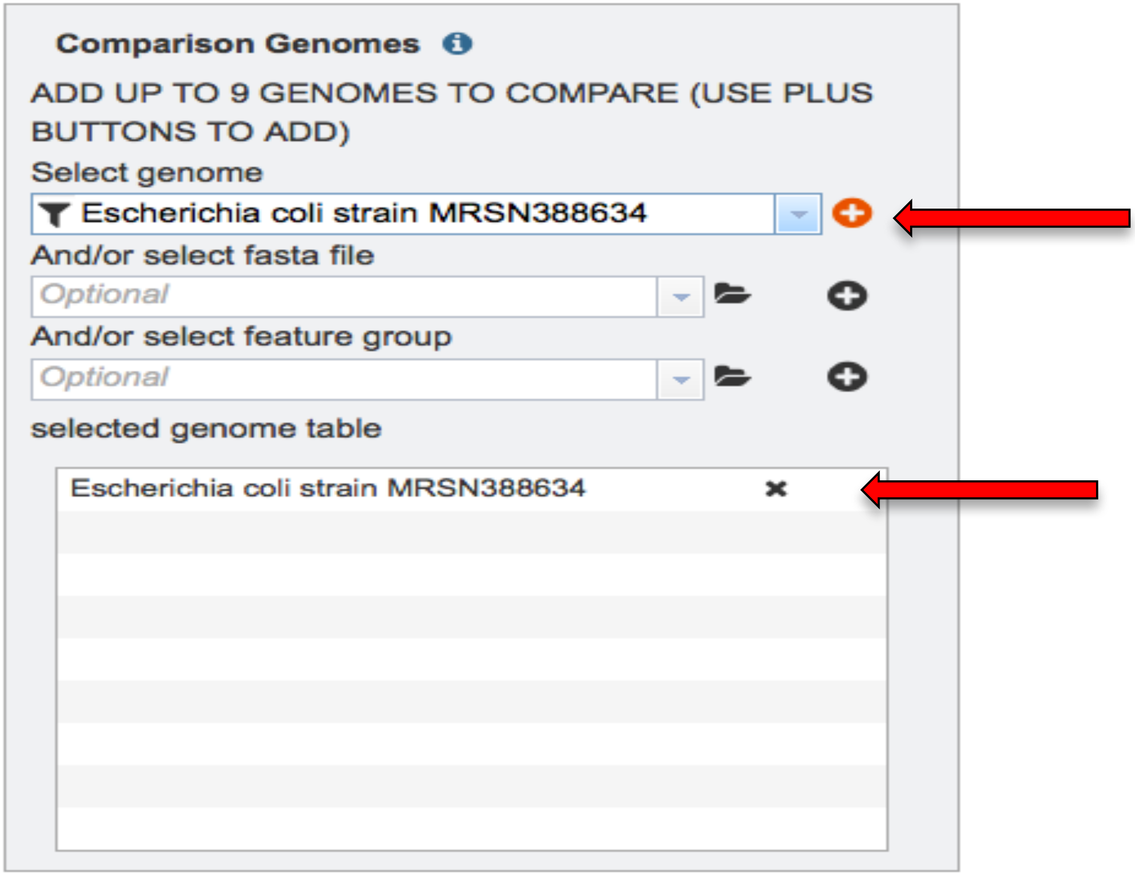
or a fasta file

Optional

or a feature group

Optional

# Selecting the Comparison Genomes



**Comparison Genomes**

ADD UP TO 9 GENOMES TO COMPARE (USE PLUS BUTTONS TO ADD)

Select genome

Escherichia coli strain MRSN388634

And/or select fasta file

Optional

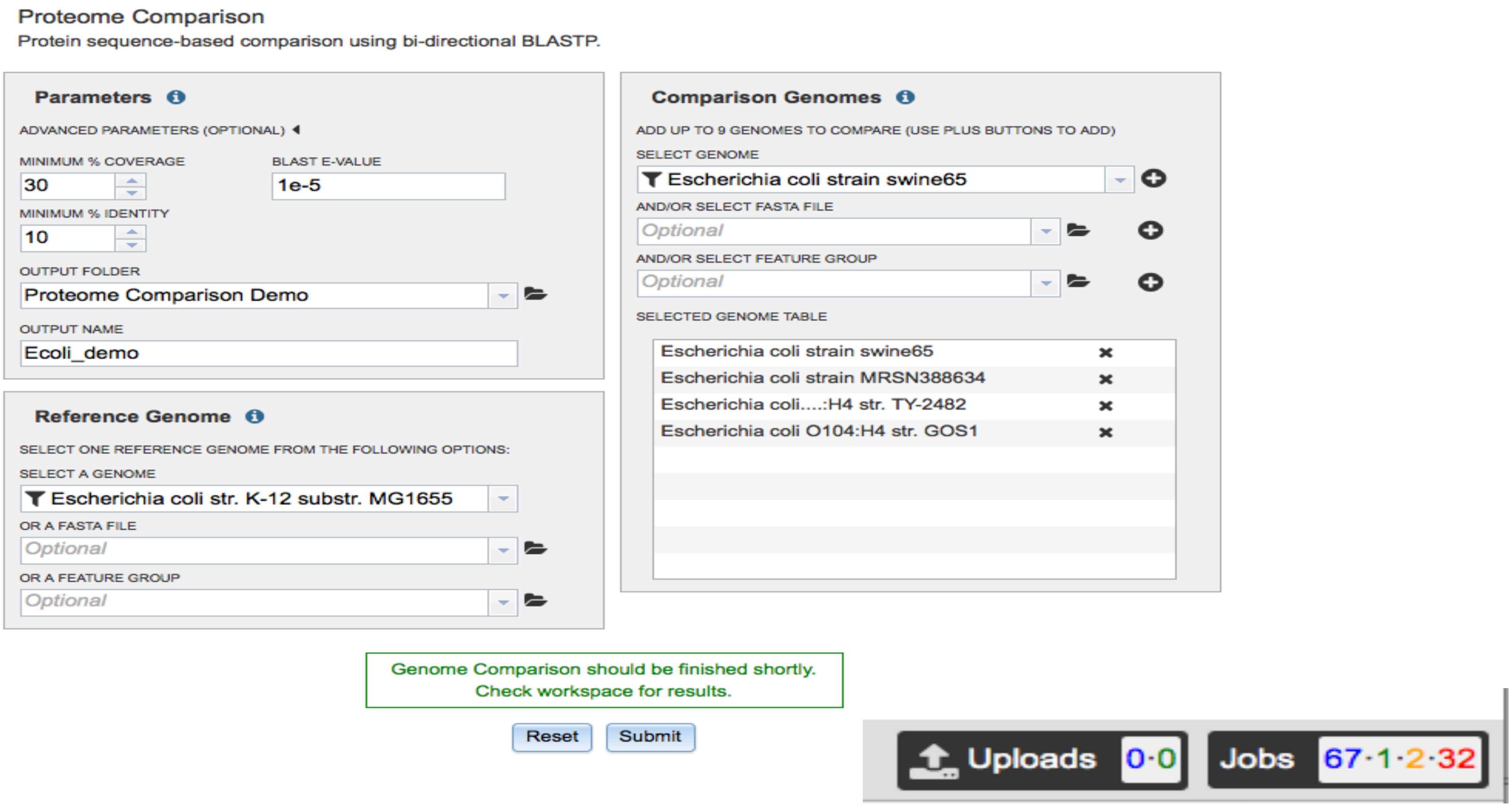
And/or select feature group

Optional

selected genome table

Escherichia coli strain MRSN388634 ✕

# Submitting the Job



Proteome Comparison

Protein sequence-based comparison using bi-directional BLASTP.

**Parameters**

ADVANCED PARAMETERS (OPTIONAL)

MINIMUM % COVERAGE: 30

BLAST E-VALUE: 1e-5

MINIMUM % IDENTITY: 10

OUTPUT FOLDER: Proteome Comparison Demo

OUTPUT NAME: Ecoli_demo

**Reference Genome**

SELECT ONE REFERENCE GENOME FROM THE FOLLOWING OPTIONS:

SELECT A GENOME: Escherichia coli str. K-12 substr. MG1655

OR A FASTA FILE: Optional

OR A FEATURE GROUP: Optional

**Comparison Genomes**

ADD UP TO 9 GENOMES TO COMPARE (USE PLUS BUTTONS TO ADD)

SELECT GENOME: Escherichia coli strain swine65

AND/OR SELECT FASTA FILE: Optional

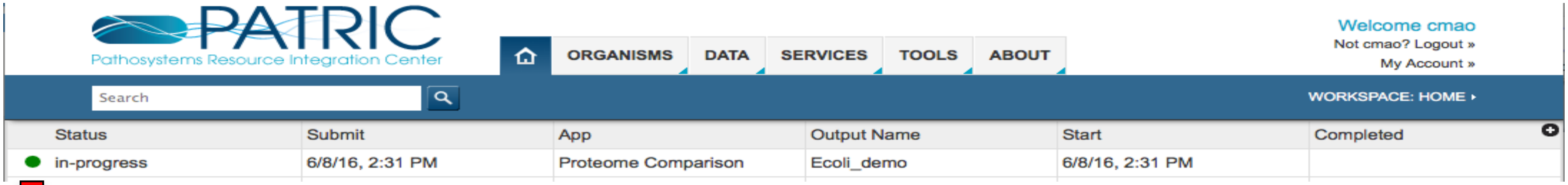
AND/OR SELECT FEATURE GROUP: Optional

SELECTED GENOME TABLE

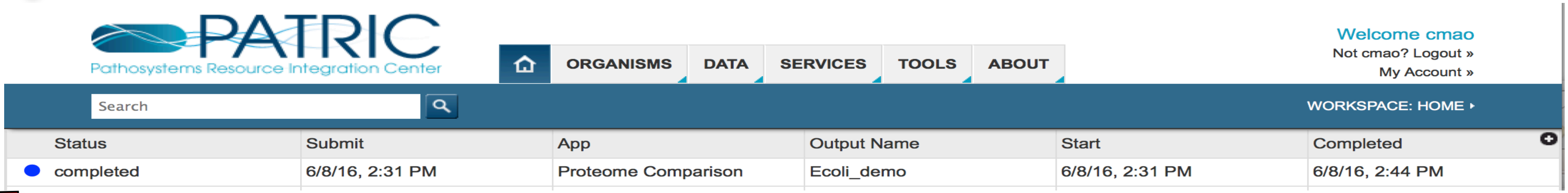
| | |
|---|---|
| Escherichia coli strain swine65 | ✖ |
| Escherichia coli strain MRSN388634 | ✖ |
| Escherichia coli....:H4 str. TY-2482 | ✖ |
| Escherichia coli O104:H4 str. GOS1 | ✖ |

Genome Comparison should be finished shortly.
Check workspace for results.

# Monitor Running Services on the Job Page



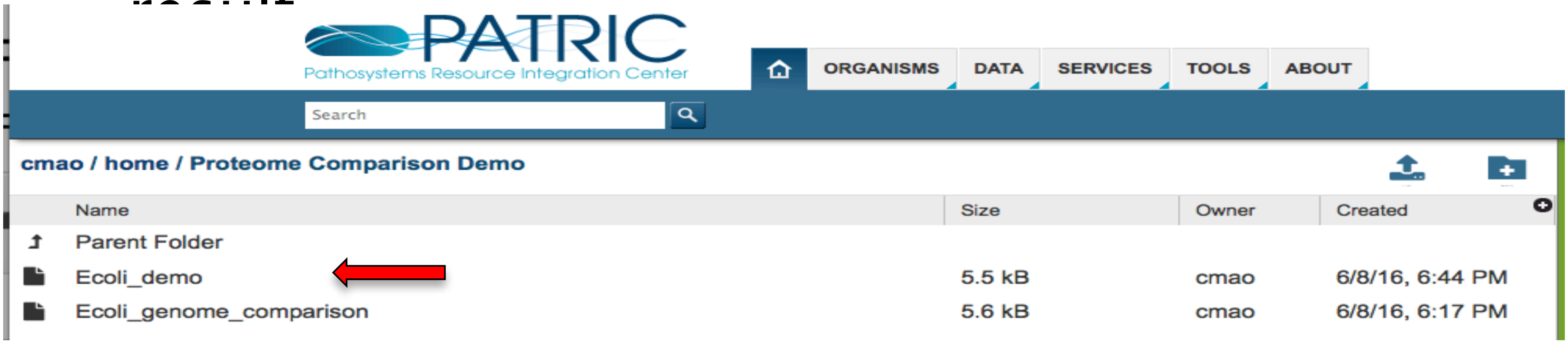
| Status | Submit | App | Output Name | Start | Completed |
|---|---|---|---|---|---|
| ● in-progress | 6/8/16, 2:31 PM | Proteome Comparison | Ecoli_demo | 6/8/16, 2:31 PM | |



| Status | Submit | App | Output Name | Start | Completed |
|---|---|---|---|---|---|
| ● completed | 6/8/16, 2:31 PM | Proteome Comparison | Ecoli_demo | 6/8/16, 2:31 PM | 6/8/16, 2:44 PM |

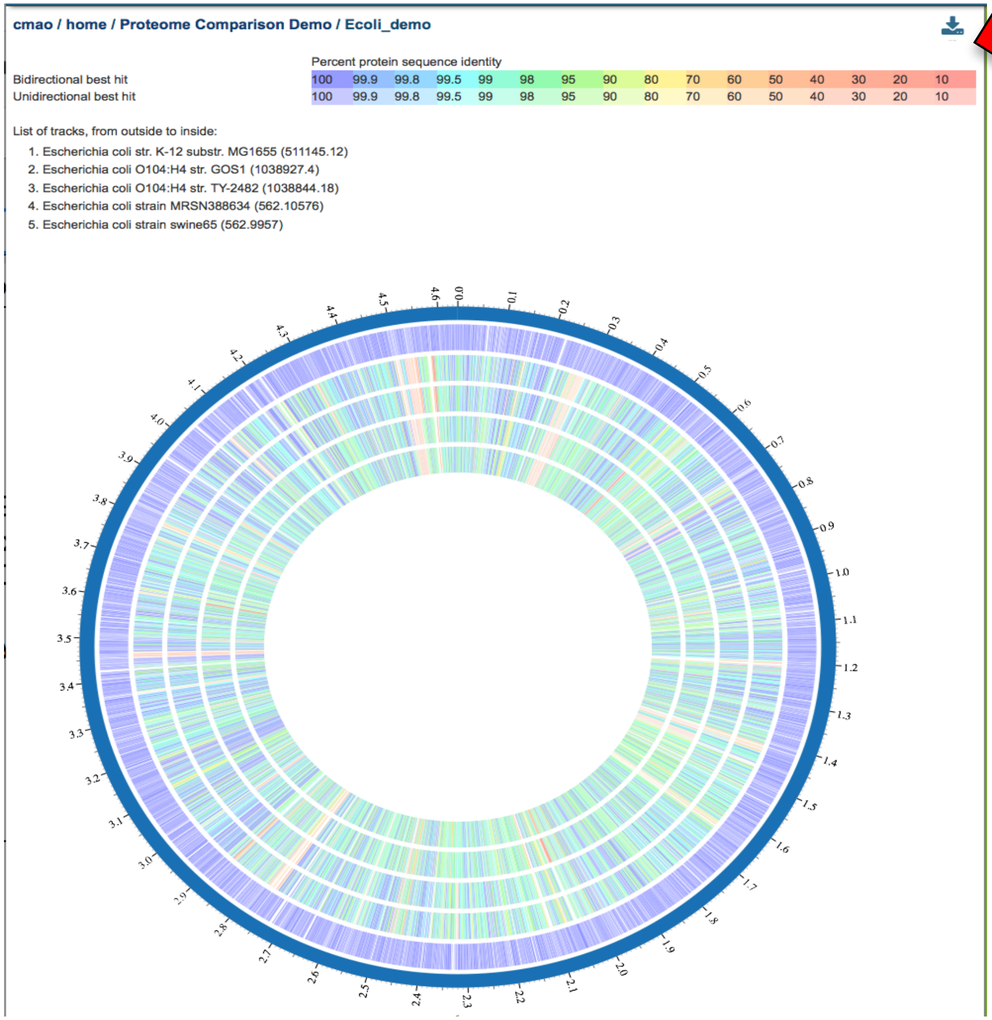
# Proteome Comparison Service Results

- The result is written to your workspace
- The result can be displayed by clicking on the result



cmao / home / Proteome Comparison Demo

| Name | Size | Owner | Created |
|---|---|---|---|
| Parent Folder | | | |
| Ecoli_demo | 5.5 kB | cmao | 6/8/16, 6:44 PM |
| Ecoli_genome_comparison | 5.6 kB | cmao | 6/8/16, 6:17 PM |



cmao / home / Proteome Comparison Demo / Ecoli_demo

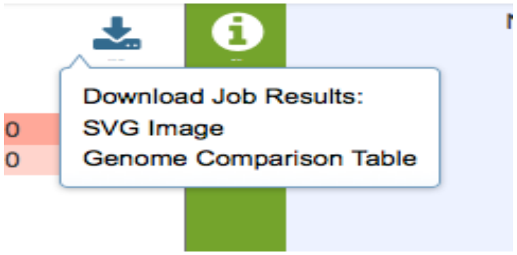
Percent protein sequence identity

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
| Unidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |

List of tracks, from outside to inside:

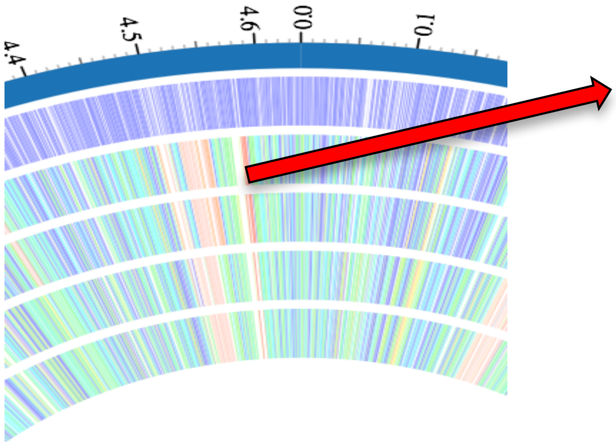
1. Escherichia coli str. K-12 substr. MG1655 (511145.12)
2. Escherichia coli O104:H4 str. GOS1 (1038927.4)
3. Escherichia coli O104:H4 str. TY-2482 (1038844.18)
4. Escherichia coli strain MRSN388634 (562.10576)
5. Escherichia coli strain swine65 (562.9957)

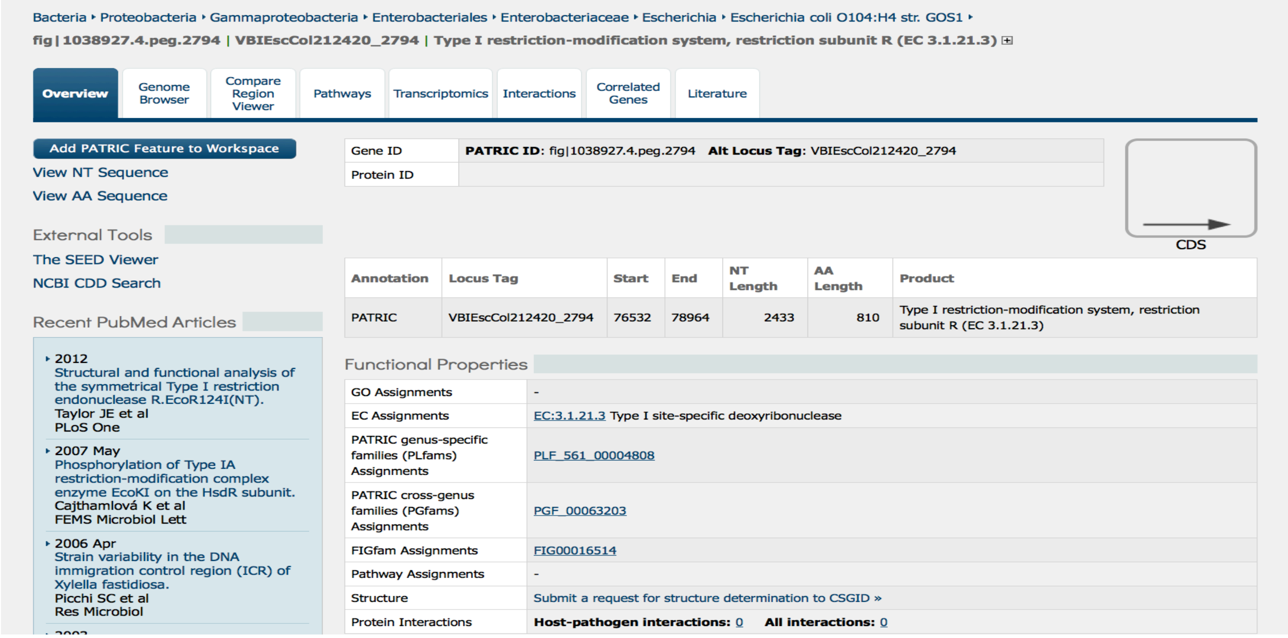
download



Download Job Results:
SVG Image
Genome Comparison Table

# Features on Graph Link to Gene Page





Bacteria ‣ Proteobacteria ‣ Gammaproteobacteria ‣ Enterobacteriales ‣ Enterobacteriaceae ‣ Escherichia ‣ Escherichia coli O104:H4 str. GOS1 ‣

**fig|1038927.4.peg.2794 | VBIEscCol212420_2794 | Type I restriction-modification system, restriction subunit R (EC 3.1.21.3)** ⊞

External Tools

The SEED Viewer

NCBI CDD Search

Recent PubMed Articles

- 2012
  Structural and functional analysis of the symmetrical Type I restriction endonuclease R.EcoR124I(NT).
  Taylor JE et al
  PLoS One
- 2007 May
  Phosphorylation of Type IA restriction-modification complex enzyme EcoKI on the HsdR subunit.
  Cajthamlová K et al
  FEMS Microbiol Lett
- 2006 Apr
  Strain variability in the DNA immigration control region (ICR) of Xylella fastidiosa.
  Picchi SC et al
  Res Microbiol
- 2003

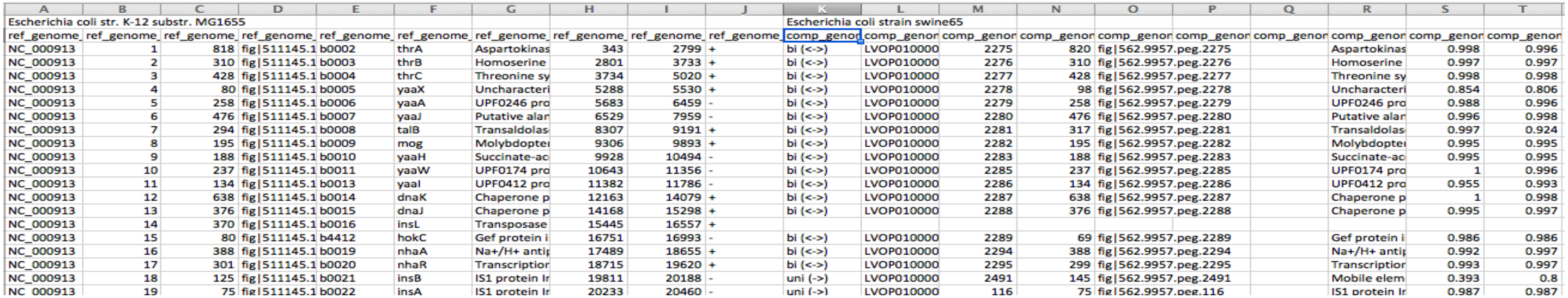
| | |
|---|---|
| Gene ID | **PATRIC ID**: fig\|1038927.4.peg.2794 **Alt Locus Tag**: VBIEscCol212420_2794 |
| Protein ID | |

| Annotation | Locus Tag | Start | End | NT Length | AA Length | Product |
|---|---|---|---|---|---|---|
| PATRIC | VBIEscCol212420_2794 | 76532 | 78964 | 2433 | 810 | Type I restriction-modification system, restriction subunit R (EC 3.1.21.3) |

Functional Properties

| | |
|---|---|
| GO Assignments | - |
| EC Assignments | EC:3.1.21.3 Type I site-specific deoxyribonuclease |
| PATRIC genus-specific families (PLfams) Assignments | PLF_561_00004808 |
| PATRIC cross-genus families (PGfams) Assignments | PGF_00063203 |
| FIGfam Assignments | FIG00016514 |
| Pathway Assignments | - |
| Structure | Submit a request for structure determination to CSGID » |
| Protein Interactions | **Host-pathogen interactions:** 0 **All interactions:** 0 |

# Genome Comparison Table

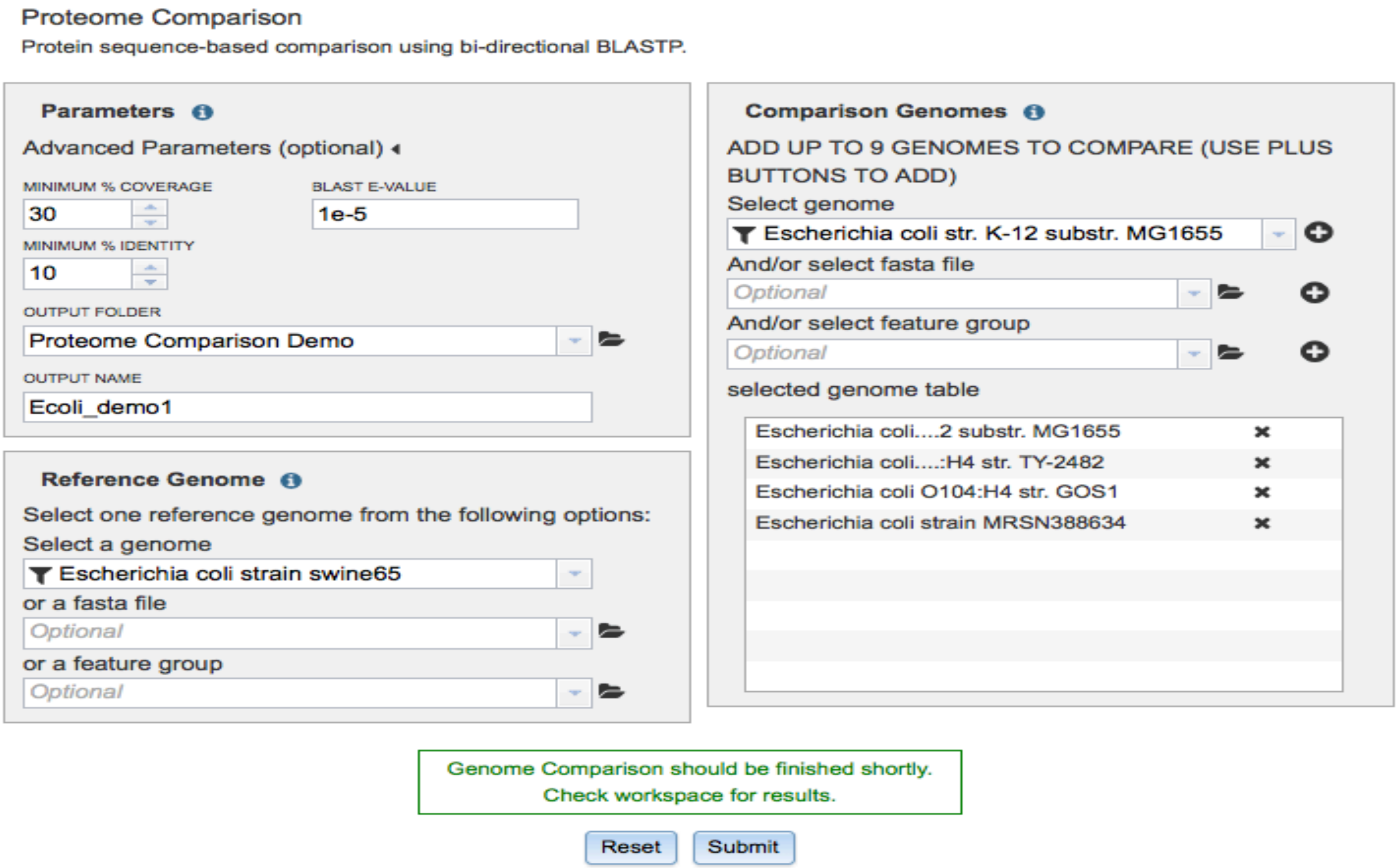


| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Escherichia coli str. K-12 substr. MG1655 | | | | | | | | | | Escherichia coli strain swine65 | | | | | | | | | |
| ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | ref_genome_ | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor | comp_genor |
| NC_000913 | 1 | 818 | fig\|511145.1 | b0002 | thrA | Aspartokinas | 343 | 2799 | + | bi (<->) | LVOP010000 | 2275 | 820 | fig\|562.9957.peg.2275 | | | Aspartokinas | 0.998 | 0.996 |
| NC_000913 | 2 | 310 | fig\|511145.1 | b0003 | thrB | Homoserine | 2801 | 3733 | + | bi (<->) | LVOP010000 | 2276 | 310 | fig\|562.9957.peg.2276 | | | Homoserine | 0.997 | 0.997 |
| NC_000913 | 3 | 428 | fig\|511145.1 | b0004 | thrC | Threonine sy | 3734 | 5020 | + | bi (<->) | LVOP010000 | 2277 | 428 | fig\|562.9957.peg.2277 | | | Threonine sy | 0.998 | 0.998 |
| NC_000913 | 4 | 80 | fig\|511145.1 | b0005 | yaaX | Uncharacteri | 5288 | 5530 | + | bi (<->) | LVOP010000 | 2278 | 98 | fig\|562.9957.peg.2278 | | | Uncharacteri | 0.854 | 0.806 |
| NC_000913 | 5 | 258 | fig\|511145.1 | b0006 | yaaA | UPF0246 pro | 5683 | 6459 | - | bi (<->) | LVOP010000 | 2279 | 258 | fig\|562.9957.peg.2279 | | | UPF0246 pro | 0.988 | 0.996 |
| NC_000913 | 6 | 476 | fig\|511145.1 | b0007 | yaaJ | Putative alar | 6529 | 7959 | - | bi (<->) | LVOP010000 | 2280 | 476 | fig\|562.9957.peg.2280 | | | Putative alar | 0.996 | 0.998 |
| NC_000913 | 7 | 294 | fig\|511145.1 | b0008 | talB | Transaldolas | 8307 | 9191 | + | bi (<->) | LVOP010000 | 2281 | 317 | fig\|562.9957.peg.2281 | | | Transaldolas | 0.997 | 0.924 |
| NC_000913 | 8 | 195 | fig\|511145.1 | b0009 | mog | Molybdopte | 9306 | 9893 | + | bi (<->) | LVOP010000 | 2282 | 195 | fig\|562.9957.peg.2282 | | | Molybdopte | 0.995 | 0.995 |
| NC_000913 | 9 | 188 | fig\|511145.1 | b0010 | yaaH | Succinate-ac | 9928 | 10494 | - | bi (<->) | LVOP010000 | 2283 | 188 | fig\|562.9957.peg.2283 | | | Succinate-ac | 0.995 | 0.995 |
| NC_000913 | 10 | 237 | fig\|511145.1 | b0011 | yaaW | UPF0174 pro | 10643 | 11356 | - | bi (<->) | LVOP010000 | 2285 | 237 | fig\|562.9957.peg.2285 | | | UPF0174 pro | 1 | 0.996 |
| NC_000913 | 11 | 134 | fig\|511145.1 | b0013 | yaal | UPF0412 pro | 11382 | 11786 | - | bi (<->) | LVOP010000 | 2286 | 134 | fig\|562.9957.peg.2286 | | | UPF0412 pro | 0.955 | 0.993 |
| NC_000913 | 12 | 638 | fig\|511145.1 | b0014 | dnaK | Chaperone p | 12163 | 14079 | + | bi (<->) | LVOP010000 | 2287 | 638 | fig\|562.9957.peg.2287 | | | Chaperone p | 1 | 0.998 |
| NC_000913 | 13 | 376 | fig\|511145.1 | b0015 | dnaJ | Chaperone p | 14168 | 15298 | + | bi (<->) | LVOP010000 | 2288 | 376 | fig\|562.9957.peg.2288 | | | Chaperone p | 0.995 | 0.997 |
| NC_000913 | 14 | 370 | fig\|511145.1 | b0016 | insL | Transposase | 15445 | 16557 | + | | | | | | | | | | |
| NC_000913 | 15 | 80 | fig\|511145.1 | b4412 | hokC | Gef protein i | 16751 | 16993 | - | bi (<->) | LVOP010000 | 2289 | 69 | fig\|562.9957.peg.2289 | | | Gef protein i | 0.986 | 0.986 |
| NC_000913 | 16 | 388 | fig\|511145.1 | b0019 | nhaA | Na+/H+ antip | 17489 | 18655 | + | bi (<->) | LVOP010000 | 2294 | 388 | fig\|562.9957.peg.2294 | | | Na+/H+ antip | 0.992 | 0.997 |
| NC_000913 | 17 | 301 | fig\|511145.1 | b0020 | nhaR | Transcriptior | 18715 | 19620 | + | bi (<->) | LVOP010000 | 2295 | 299 | fig\|562.9957.peg.2295 | | | Transcriptior | 0.993 | 0.997 |
| NC_000913 | 18 | 125 | fig\|511145.1 | b0021 | insB | IS1 protein Ir | 19811 | 20188 | - | uni (->) | LVOP010000 | 2491 | 145 | fig\|562.9957.peg.2491 | | | Mobile elem | 0.393 | 0.8 |
| NC_000913 | 19 | 75 | fig\|511145.1 | b0022 | insA | IS1 protein Ir | 20233 | 20460 | - | uni (->) | LVOP010000 | 116 | 75 | fig\|562.9957.peg.116 | | | IS1 protein Ir | 0.987 | 0.987 |

Data begins with the reference (2A-J) and includes the following:
accession number for the contig in the reference genome (Column A); the order number of this gene in the genome (B); size in amino acids (C); PATRIC locus tag (D); RefSeq locus tag (E); gene name (F); functional annotation (G); start location for the gene on the contig (H); end of the gene on the contig (I); and strand that the gene is located on (J). This is followed by information on the comparison genomes.

This data in columns K-T for row 2 (for the first comparison genome) include: data on the type of BLAST hit (Column K, bi- or uni-directional, or missing); contig that the gene is located on (L); the order number of this gene in the genome (M); size in amino acids (N); PATRIC locus tag (O); RefSeq locus tag (P); gene name (Q); functional description (R); percent identity of the BLAST hit (S); and sequence coverage compared to the reference (T).

# Using a Different Reference Genome



Proteome Comparison

Protein sequence-based comparison using bi-directional BLASTP.

**Parameters**

Advanced Parameters (optional)

MINIMUM % COVERAGE: 30

BLAST E-VALUE: 1e-5

MINIMUM % IDENTITY: 10

OUTPUT FOLDER: Proteome Comparison Demo

OUTPUT NAME: Ecoli_demo1

**Reference Genome**

Select one reference genome from the following options:

Select a genome: Escherichia coli strain swine65

or a fasta file: Optional

or a feature group: Optional

**Comparison Genomes**

ADD UP TO 9 GENOMES TO COMPARE (USE PLUS BUTTONS TO ADD)

Select genome: Escherichia coli str. K-12 substr. MG1655

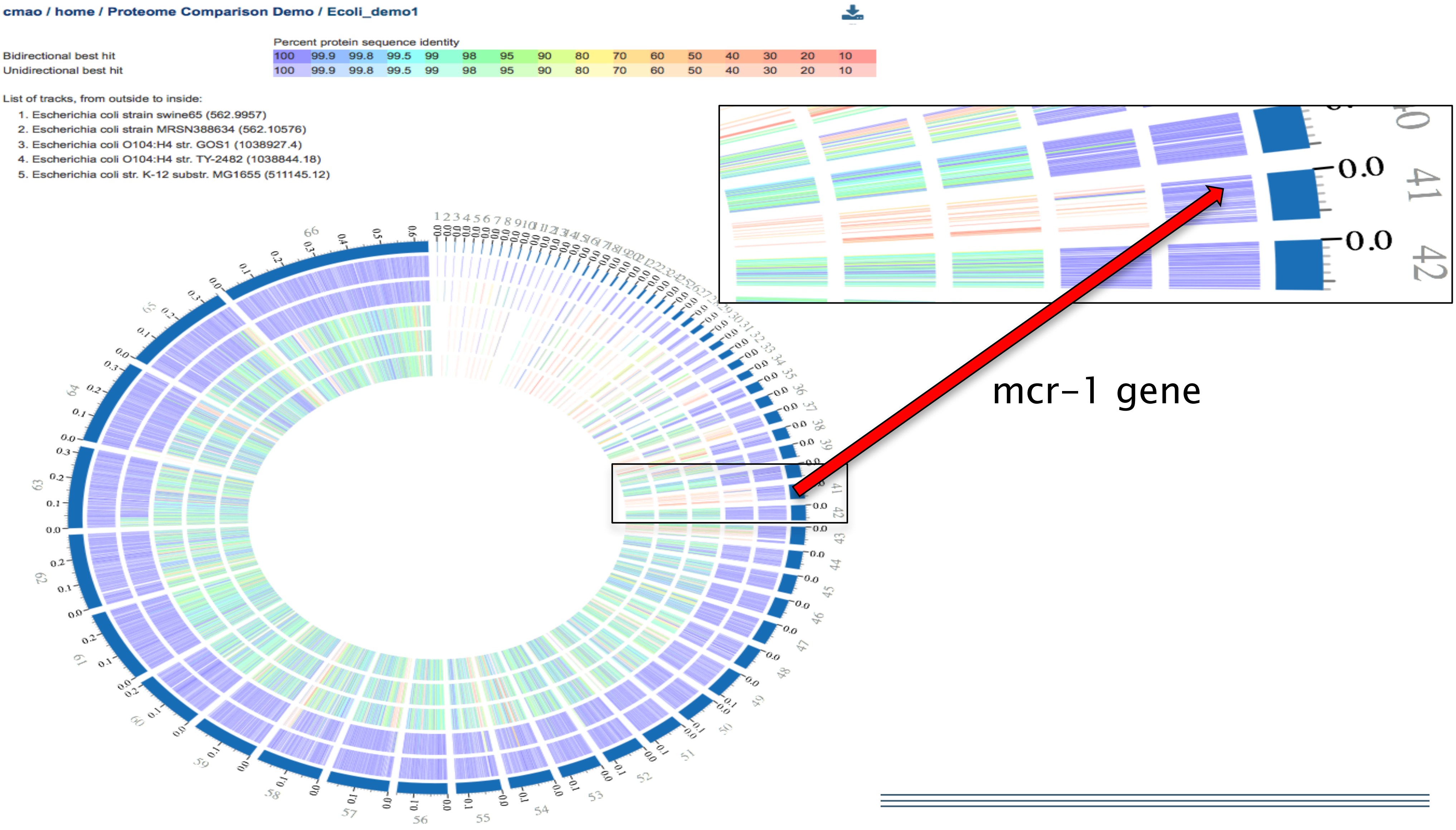
And/or select fasta file: Optional

And/or select feature group: Optional

selected genome table

| |
|---|
| Escherichia coli....2 substr. MG1655 |
| Escherichia coli....:H4 str. TY-2482 |
| Escherichia coli O104:H4 str. GOS1 |
| Escherichia coli strain MRSN388634 |

Genome Comparison should be finished shortly.
Check workspace for results.



cmao / home / Proteome Comparison Demo / Ecoli_demo1

Percent protein sequence identity

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
| Unidirectional best hit | 100 | 99.9 | 99.8 | 99.5 | 99 | 98 | 95 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |

List of tracks, from outside to inside:

1. Escherichia coli strain swine65 (562.9957)
2. Escherichia coli strain MRSN388634 (562.10576)
3. Escherichia coli O104:H4 str. GOS1 (1038927.4)
4. Escherichia coli O104:H4 str. TY-2482 (1038844.18)
5. Escherichia coli str. K-12 substr. MG1655 (511145.12)

mcr-1 gene

# mcr-1 homologs

| E. coli strain swine65 | | E. coli strain MRSN388634 | | | | E.coli O104:H4 str. TY-2482 | | | |
|---|---|---|---|---|---|---|---|---|---|
| patric_id | gene | patric_id | direction | identity | coverage | patric_id | direction | identity | coverage |
| fig\|562.9957.peg.578 | mcr-1 | fig\|562.10576.peg.4702 | bi | 1 | 0.998 | fig\|1038927.4.peg.4771 | uni | 0.346 | 0.922 |