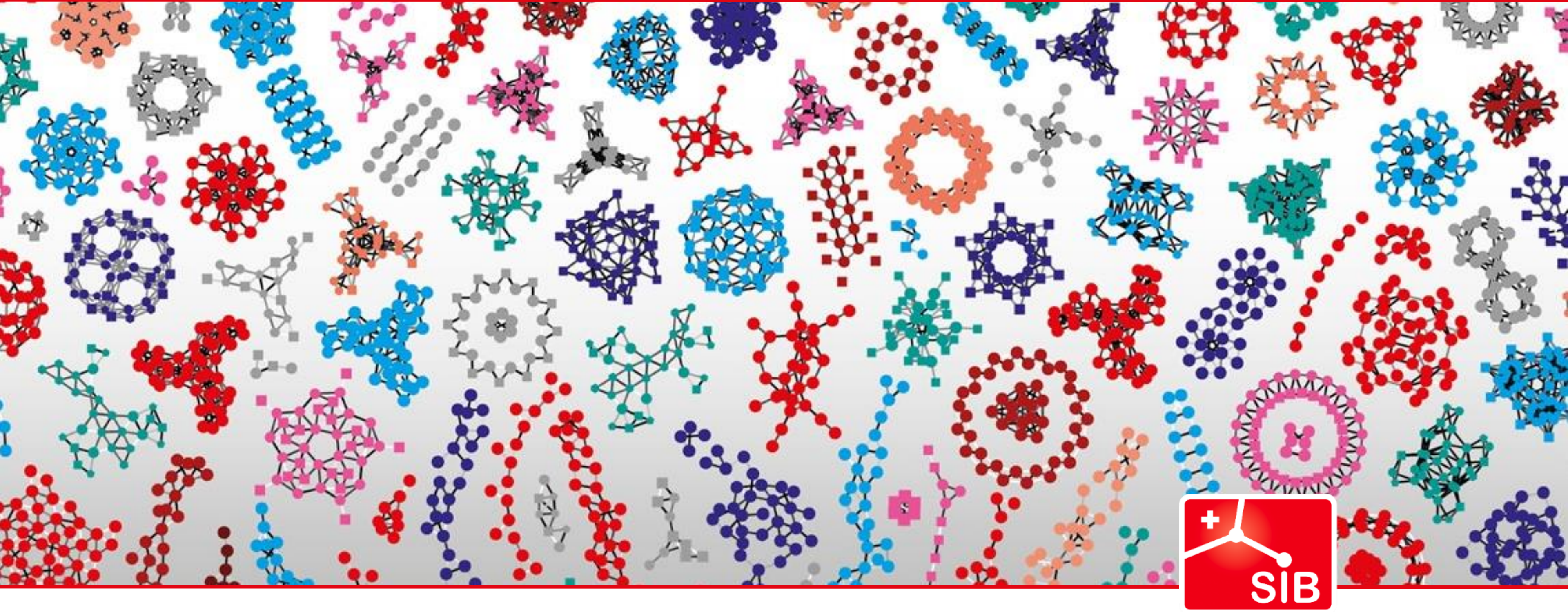


Swiss Institute of Bioinformatics

Proteomes in UniProtKB

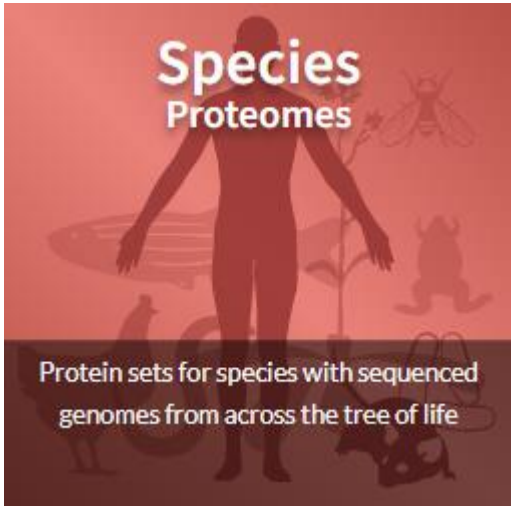
Protein sequence databases and sequence annotation

Streamed from Geneva, 14 October 2022

www.sib.swiss

Proteomes

- A proteome is the set of proteins thought to be expressed by an organism (completely sequenced genomes).
- A proteome is formed by all UniProtKB entries mapping to a genome assembly from EMBL-ENA or Ensembl or RefSeq.
- Also from VectorBase, WormBase, Parasite
- Viral proteomes are manually checked and verified and periodically added to the database (SIB)


Species
Proteomes
Protein sets for species with sequenced genomes from across the tree of life

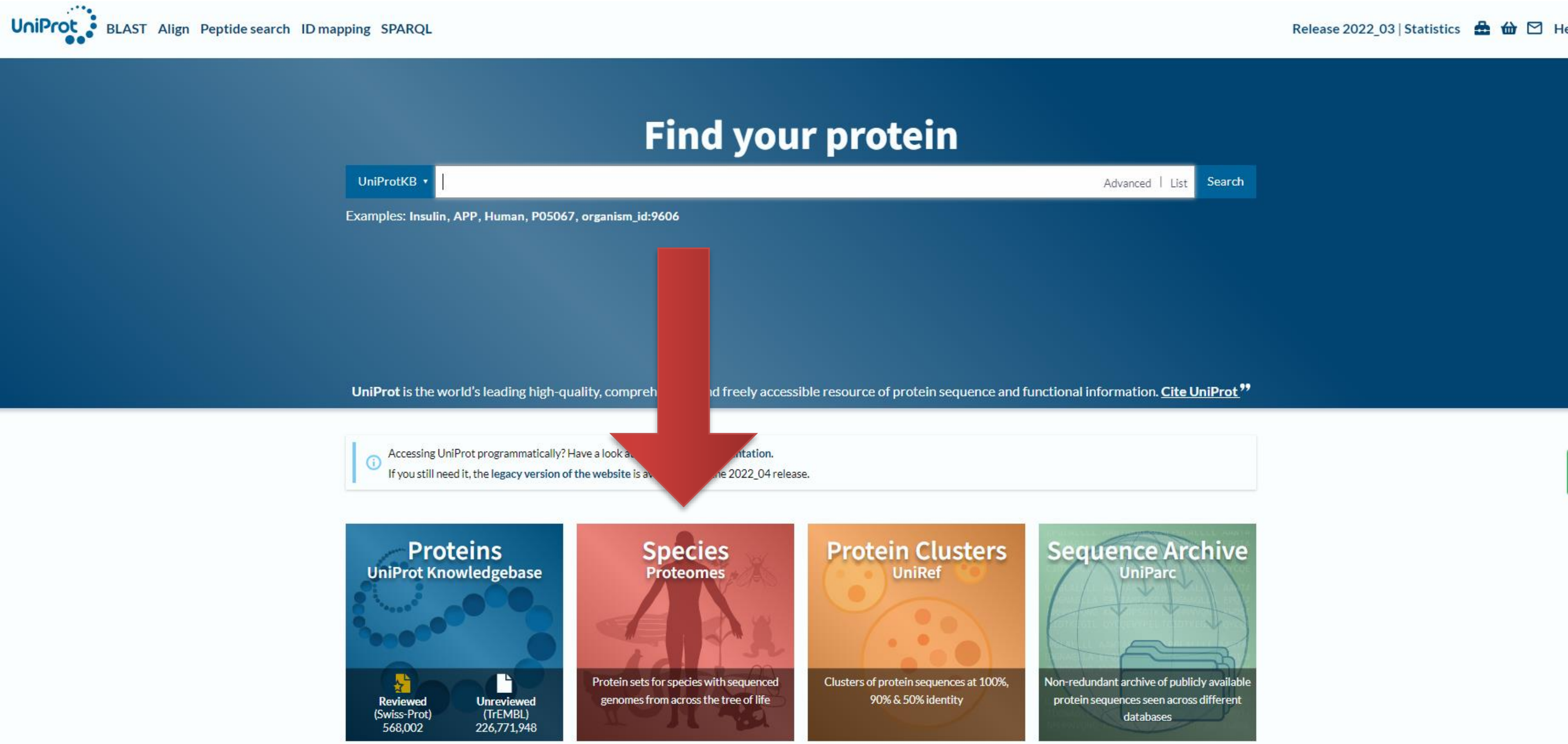
VectorBase Release 49 *beta* 5 Nov 2020 *Bioinformatics Resources for Invertebrate Vectors of Human Pathogens*

WormBase

WormBase ParaSite

ViralZone

UniProtKB Proteomes



Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, compreh d freely accessible resource of protein sequence and functional information. Cite UniProt

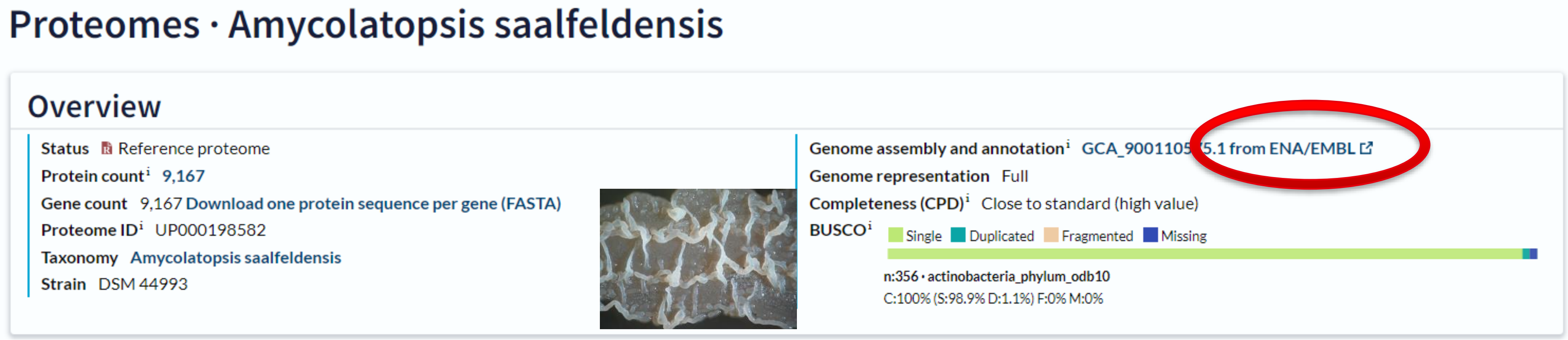
Accessing UniProt programmatically? Have a look a ntation.
If you still need it, the legacy version of the website is av he 2022_04 release.

Protein sets for species with sequenced genomes from across the tree of life

Clusters of protein sequences at 100%, 90% & 50% identity

Non-redundant archive of publicly available protein sequences seen across different databases

Proteomes



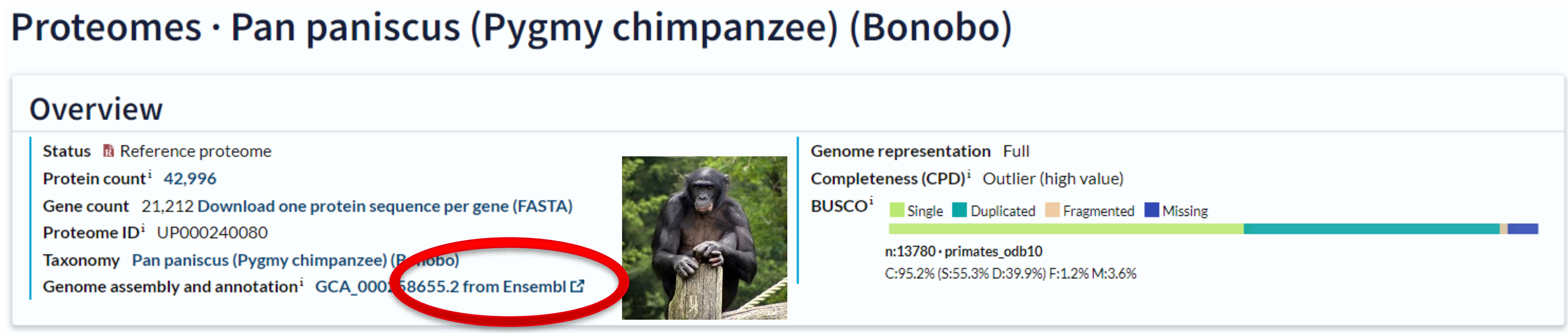
Proteomes · Amycolatopsis saalfeldensis

Overview

Status R Reference proteome
Protein count ⓘ 9,167
Gene count 9,167 Download one protein sequence per gene (FASTA)
Proteome ID ⓘ UP000198582
Taxonomy Amycolatopsis saalfeldensis
Strain DSM 44993

Genome assembly and annotation ⓘ GCA_900110575.1 from ENA/EMBL
Genome representation Full
Completeness (CPD) ⓘ Close to standard (high value)
BUSCO ⓘ 🟩 Single 🟦 Duplicated 🟧 Fragmented 🟦 Missing

n:356 · actinobacteria_phylum_odb10
C:100% (S:98.9% D:1.1%) F:0% M:0%



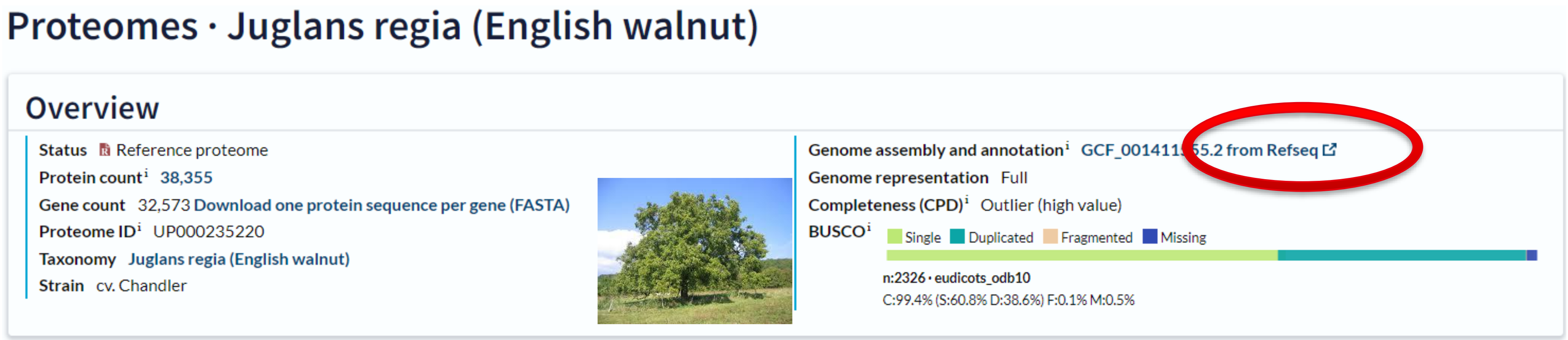
Proteomes · Pan paniscus (Pygmy chimpanzee) (Bonobo)

Overview

Status R Reference proteome
Protein count ⓘ 42,996
Gene count 21,212 Download one protein sequence per gene (FASTA)
Proteome ID ⓘ UP000240080
Taxonomy Pan paniscus (Pygmy chimpanzee) (Bonobo)
Genome assembly and annotation ⓘ GCA_000258655.2 from Ensembl

Genome representation Full
Completeness (CPD) ⓘ Outlier (high value)
BUSCO ⓘ 🟩 Single 🟦 Duplicated 🟧 Fragmented 🟦 Missing

n:13780 · primates_odb10
C:95.2% (S:55.3% D:39.9%) F:1.2% M:3.6%



Proteomes · Juglans regia (English walnut)

Overview

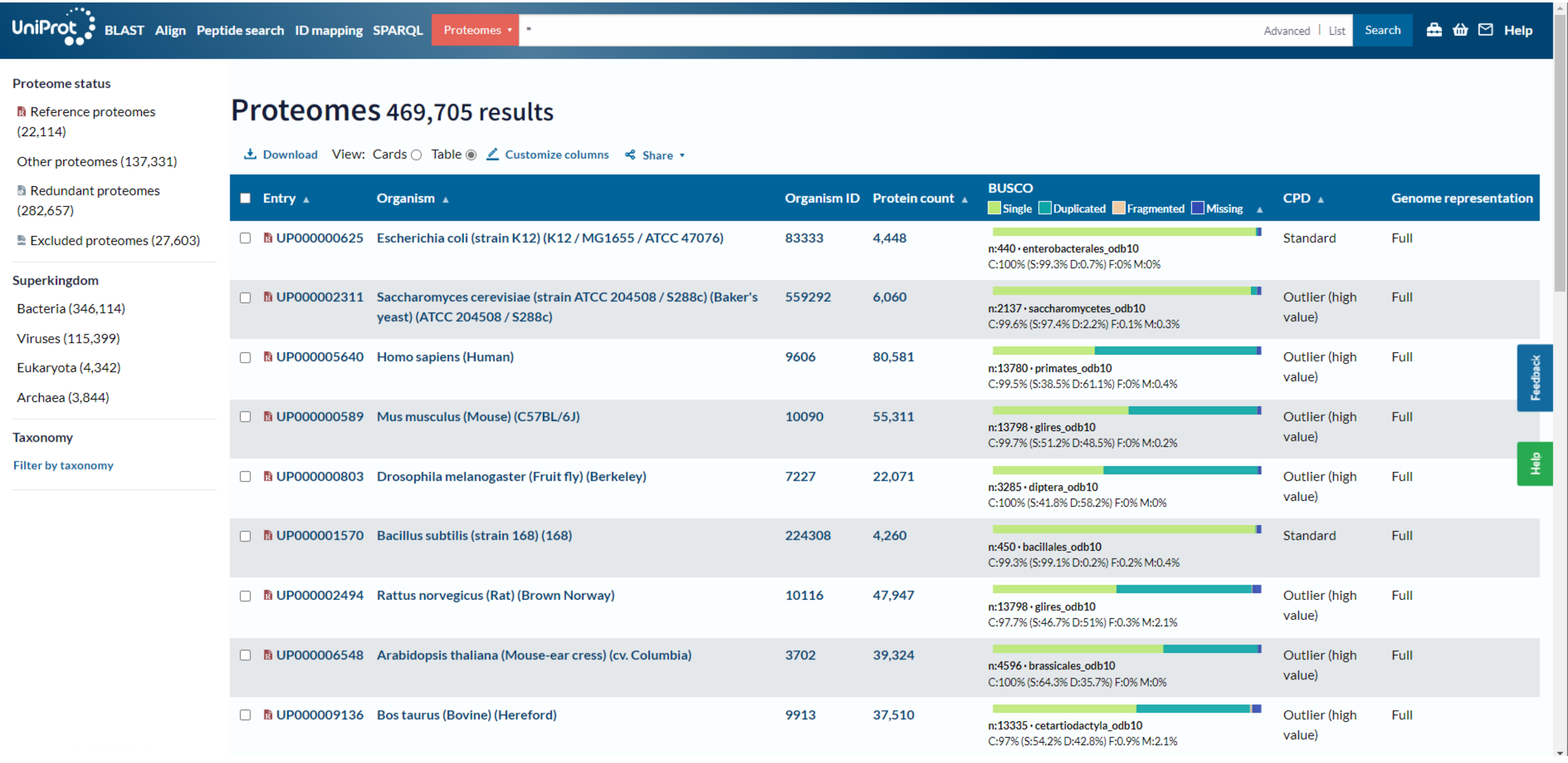
Status R Reference proteome
Protein count ⓘ 38,355
Gene count 32,573 Download one protein sequence per gene (FASTA)
Proteome ID ⓘ UP000235220
Taxonomy Juglans regia (English walnut)
Strain cv. Chandler

Genome assembly and annotation ⓘ GCF_001411555.2 from Refseq
Genome representation Full
Completeness (CPD) ⓘ Outlier (high value)
BUSCO ⓘ 🟩 Single 🟦 Duplicated 🟧 Fragmented 🟦 Missing

n:2326 · eudicots_odb10
C:99.4% (S:60.8% D:38.6%) F:0.1% M:0.5%

Proteomes

- The Proteomes webpage has been redesigned to enable users to view full details of their proteome(s) of interest in a single table view.



Proteome status

- Reference proteomes (22,114)
- Other proteomes (137,331)
- Redundant proteomes (282,657)
- Excluded proteomes (27,603)

Superkingdom

- Bacteria (346,114)
- Viruses (115,399)
- Eukaryota (4,342)
- Archaea (3,844)

Taxonomy

Filter by taxonomy

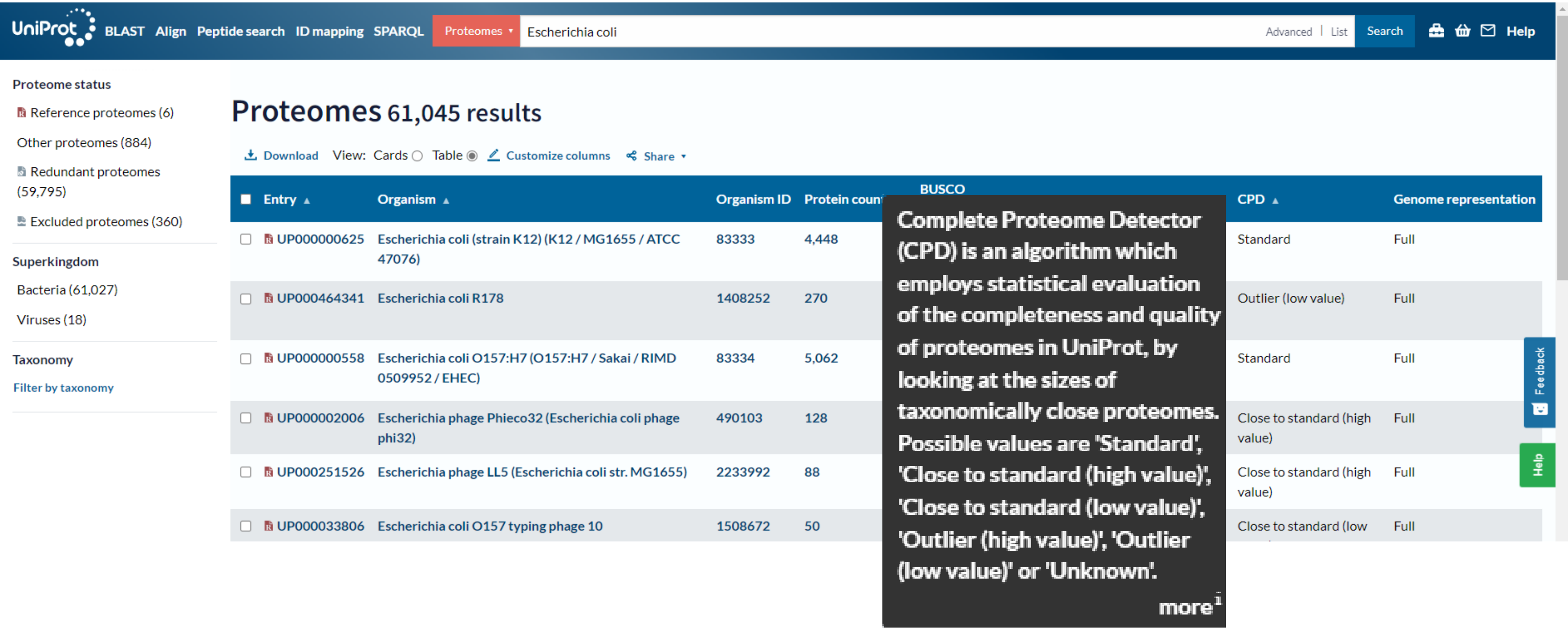
Proteomes 469,705 results

Download View: Cards ○ Table ● Customize columns Share ▾

□	Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing ▲	CPD ▲	Genome representation
□	UP000000625	Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448	n:440 · enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:0%	Standard	Full
□	UP000002311	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) (ATCC 204508 / S288c)	559292	6,060	n:2137 · saccharomycetes_odb10 C:99.6% (S:97.4% D:2.2%) F:0.1% M:0.3%	Outlier (high value)	Full
□	UP000005640	Homo sapiens (Human)	9606	80,581	n:13780 · primates_odb10 C:99.5% (S:38.5% D:61.1%) F:0% M:0.4%	Outlier (high value)	Full
□	UP000000589	Mus musculus (Mouse) (C57BL/6J)	10090	55,311	n:13798 · glires_odb10 C:99.7% (S:51.2% D:48.5%) F:0% M:0.2%	Outlier (high value)	Full
□	UP000000803	Drosophila melanogaster (Fruit fly) (Berkeley)	7227	22,071	n:3285 · diptera_odb10 C:100% (S:41.8% D:58.2%) F:0% M:0%	Outlier (high value)	Full
□	UP000001570	Bacillus subtilis (strain 168) (168)	224308	4,260	n:450 · bacillales_odb10 C:99.3% (S:99.1% D:0.2%) F:0.2% M:0.4%	Standard	Full
□	UP000002494	Rattus norvegicus (Rat) (Brown Norway)	10116	47,947	n:13798 · glires_odb10 C:97.7% (S:46.7% D:51%) F:0.3% M:2.1%	Outlier (high value)	Full
□	UP000006548	Arabidopsis thaliana (Mouse-ear cress) (cv. Columbia)	3702	39,324	n:4596 · brassicales_odb10 C:100% (S:64.3% D:35.7%) F:0% M:0%	Outlier (high value)	Full
□	UP000009136	Bos taurus (Bovine) (Hereford)	9913	37,510	n:13335 · cetartiodactyla_odb10 C:97% (S:54.2% D:42.8%) F:0.9% M:2.1%	Outlier (high value)	Full

CPD: Completeness in terms of number

We display the results of the ‘Complete Proteome Detector’ (CPD), an in-house algorithm which statistically evaluates the completeness and quality of each proteome by directly comparing it to those of a group of at least 3 closely taxonomically-related species. The CPD classifies each proteome as either ‘standard’, ‘close to standard’ or an ‘outlier’, according to protein count.



Escherichia coli

Proteome status

- Reference proteomes (6)
- Other proteomes (884)
- Redundant proteomes (59,795)
- Excluded proteomes (360)

Superkingdom

- Bacteria (61,027)
- Viruses (18)

Taxonomy

Filter by taxonomy

Proteomes 61,045 results

Download View: Cards ○ Table ● Customize columns Share ▾

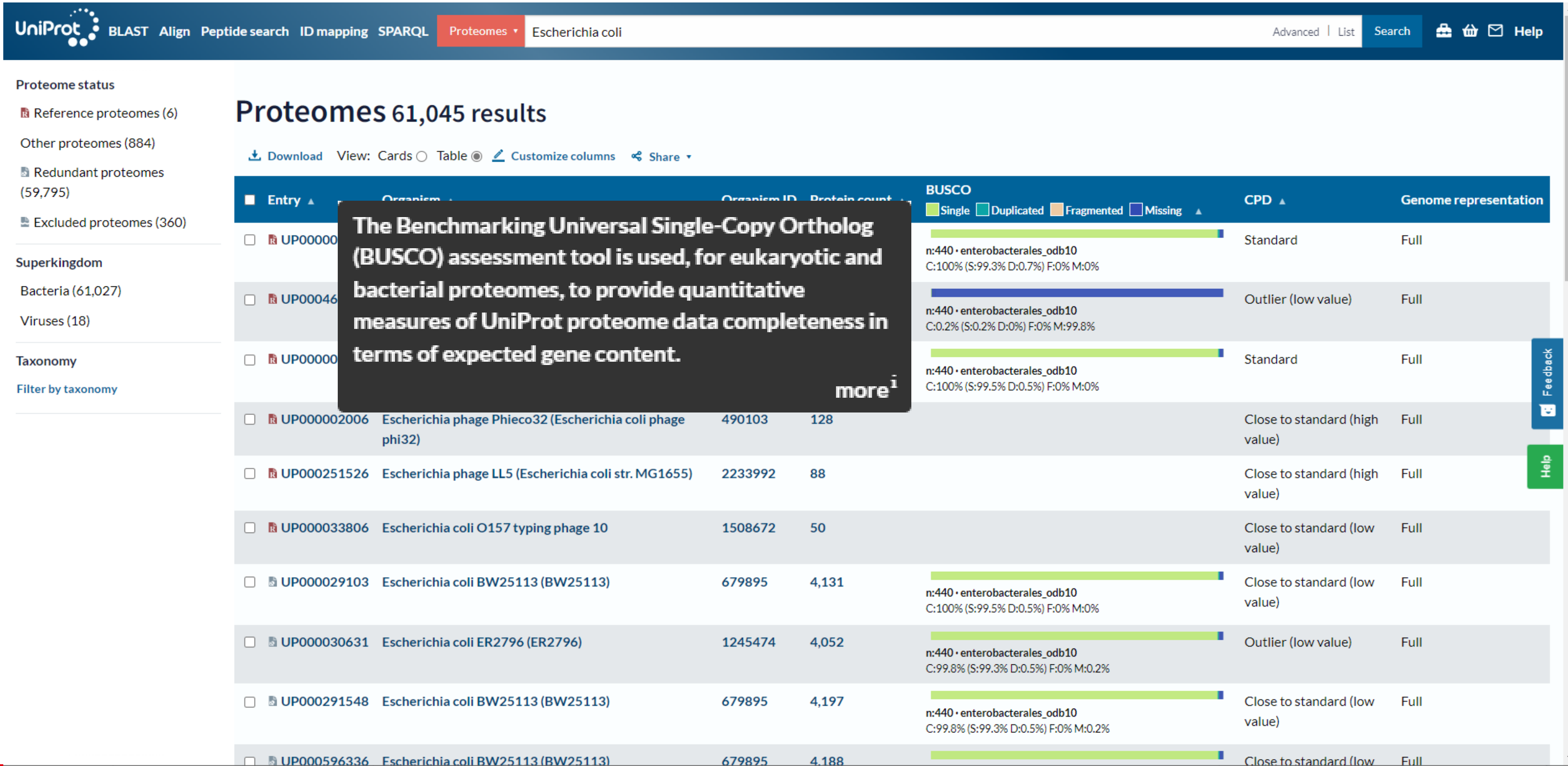
■	Entry ▲	Organism ▲	Organism ID	Protein count	BUSCO	CPD ▲	Genome representation
□	UP000000625	Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448		Standard	Full
□	UP000464341	Escherichia coli R178	1408252	270		Outlier (low value)	Full
□	UP000000558	Escherichia coli O157:H7 (O157:H7 / Sakai / RIMD 0509952 / EHEC)	83334	5,062		Standard	Full
□	UP000002006	Escherichia phage Phieco32 (Escherichia coli phage phi32)	490103	128		Close to standard (high value)	Full
□	UP000251526	Escherichia phage LL5 (Escherichia coli str. MG1655)	2233992	88		Close to standard (high value)	Full
□	UP000033806	Escherichia coli O157 typing phage 10	1508672	50		Close to standard (low	Full

Complete Proteome Detector (CPD) is an algorithm which employs statistical evaluation of the completeness and quality of proteomes in UniProt, by looking at the sizes of taxonomically close proteomes. Possible values are 'Standard', 'Close to standard (high value)', 'Close to standard (low value)', 'Outlier (high value)', 'Outlier (low value)' or 'Unknown'.

more

BUSCO: Completeness in term of gene content

The BUSCO score includes percentages of complete single-copy genes, complete duplicated genes, fragmented and missing genes, as well as the total number of orthologous clusters used in the BUSCO assessment.



Proteome status

- Reference proteomes (6)
- Other proteomes (884)
- Redundant proteomes (59,795)
- Excluded proteomes (360)

Superkingdom

- Bacteria (61,027)
- Viruses (18)

Taxonomy

Filter by taxonomy

Proteomes 61,045 results

Download View: Cards ○ Table ● Customize columns Share ▾

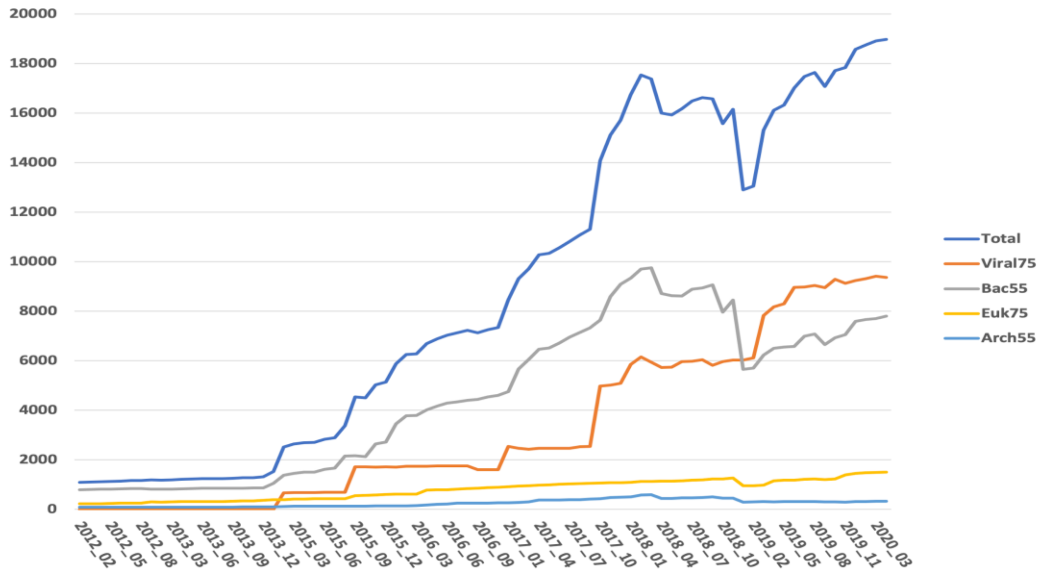
The Benchmarking Universal Single-Copy Ortholog (BUSCO) assessment tool is used, for eukaryotic and bacterial proteomes, to provide quantitative measures of UniProt proteome data completeness in terms of expected gene content.

more[i]

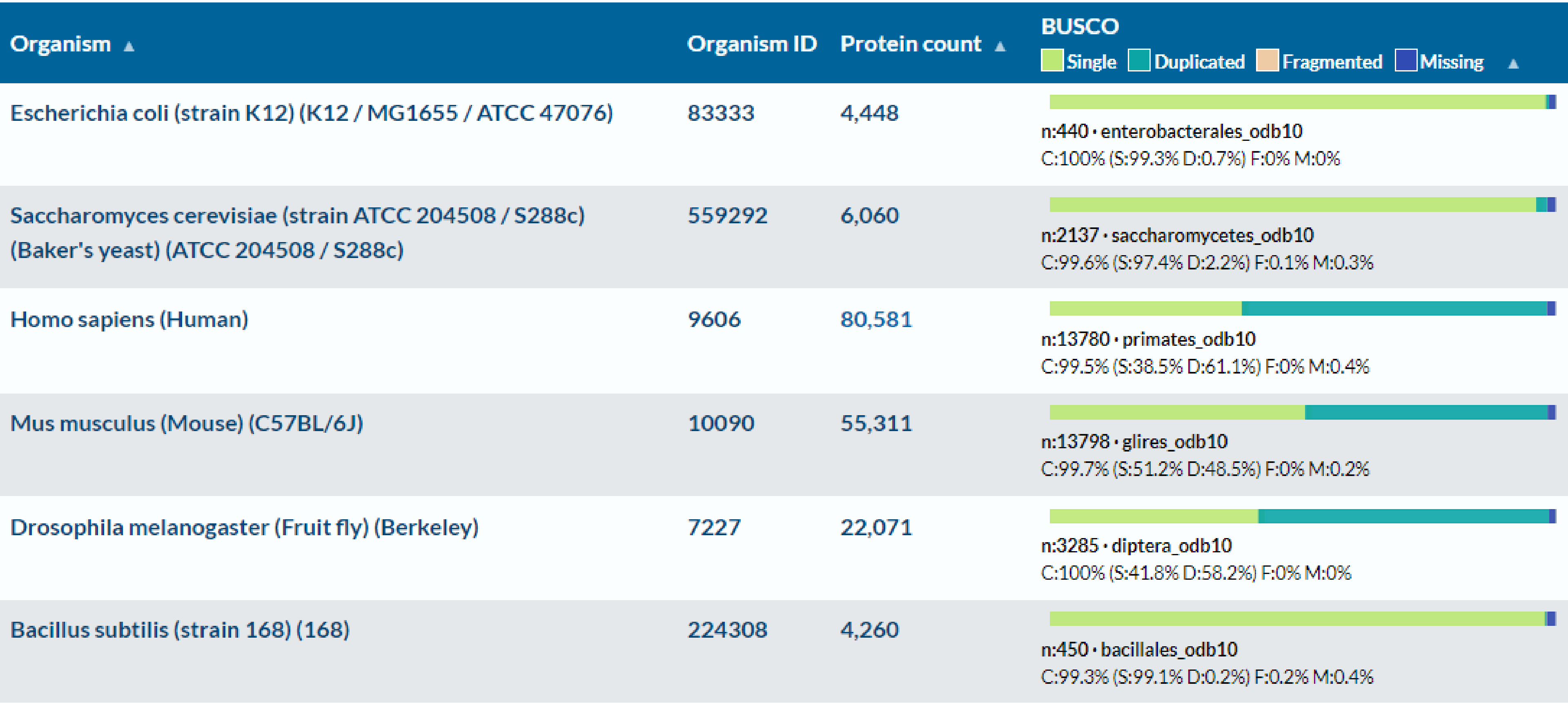
☐	Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing ▲	CPD ▲	Genome representation
☐	UP00000				n:440 • enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:0%	Standard	Full
☐	UP00046				n:440 • enterobacterales_odb10 C:0.2% (S:0.2% D:0%) F:0% M:99.8%	Outlier (low value)	Full
☐	UP00000				n:440 • enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Standard	Full
☐	UP000002006	Escherichia phage Phieco32 (Escherichia coli phage phi32)	490103	128		Close to standard (high value)	Full
☐	UP000251526	Escherichia phage LL5 (Escherichia coli str. MG1655)	2233992	88		Close to standard (high value)	Full
☐	UP000033806	Escherichia coli O157 typing phage 10	1508672	50		Close to standard (low value)	Full
☐	UP000029103	Escherichia coli BW25113 (BW25113)	679895	4,131	n:440 • enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Close to standard (low value)	Full
☐	UP000030631	Escherichia coli ER2796 (ER2796)	1245474	4,052	n:440 • enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Outlier (low value)	Full
☐	UP000291548	Escherichia coli BW25113 (BW25113)	679895	4,197	n:440 • enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Close to standard (low value)	Full
☐	UP000596336	Escherichia coli BW25113 (BW25113)	679895	4,188		Close to standard (low	Full

BUSCO: Completeness in term of gene content

BUSCO v3 identifies complete, duplicated, fragmented, and potentially missing genes by comparison to a defined set of near-universal single copy orthologs. As a result of our adoption of the BUSCO algorithm, a number of incomplete and poor quality proteomes were identified and removed from UniProtKB. This, coupled with a concomitant clean-up of metagenomic data, resulted in a drop in the number of bacterial sequences in 2018.



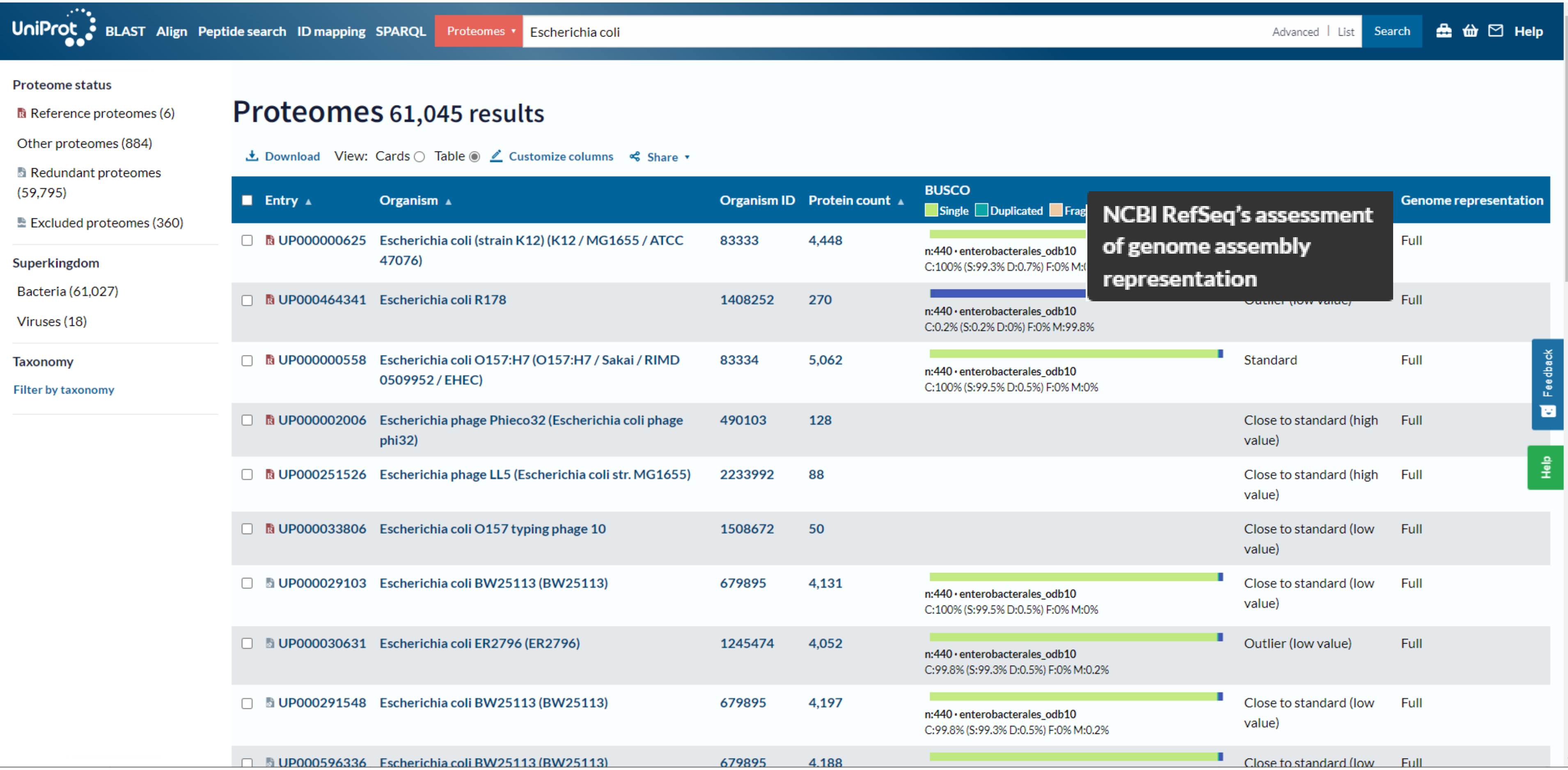
BUSCO: Completeness in term of gene content



Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟦 Duplicated 🟧 Fragmented 🟦 Missing ▲
Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448	n:440 · enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:0%
Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) (ATCC 204508 / S288c)	559292	6,060	n:2137 · saccharomycetes_odb10 C:99.6% (S:97.4% D:2.2%) F:0.1% M:0.3%
Homo sapiens (Human)	9606	80,581	n:13780 · primates_odb10 C:99.5% (S:38.5% D:61.1%) F:0% M:0.4%
Mus musculus (Mouse) (C57BL/6J)	10090	55,311	n:13798 · glires_odb10 C:99.7% (S:51.2% D:48.5%) F:0% M:0.2%
Drosophila melanogaster (Fruit fly) (Berkeley)	7227	22,071	n:3285 · diptera_odb10 C:100% (S:41.8% D:58.2%) F:0% M:0%
Bacillus subtilis (strain 168) (168)	224308	4,260	n:450 · bacillales_odb10 C:99.3% (S:99.1% D:0.2%) F:0.2% M:0.4%

UniProtKB – scope

We additionally provide the assessment of the genome assembly status imported from the source of the genome assembly and annotation (e.g. Ensembl or RefSeq).



Proteome status

- Reference proteomes (6)
- Other proteomes (884)
- Redundant proteomes (59,795)
- Excluded proteomes (360)

Superkingdom

- Bacteria (61,027)
- Viruses (18)

Taxonomy

Filter by taxonomy

Proteomes 61,045 results

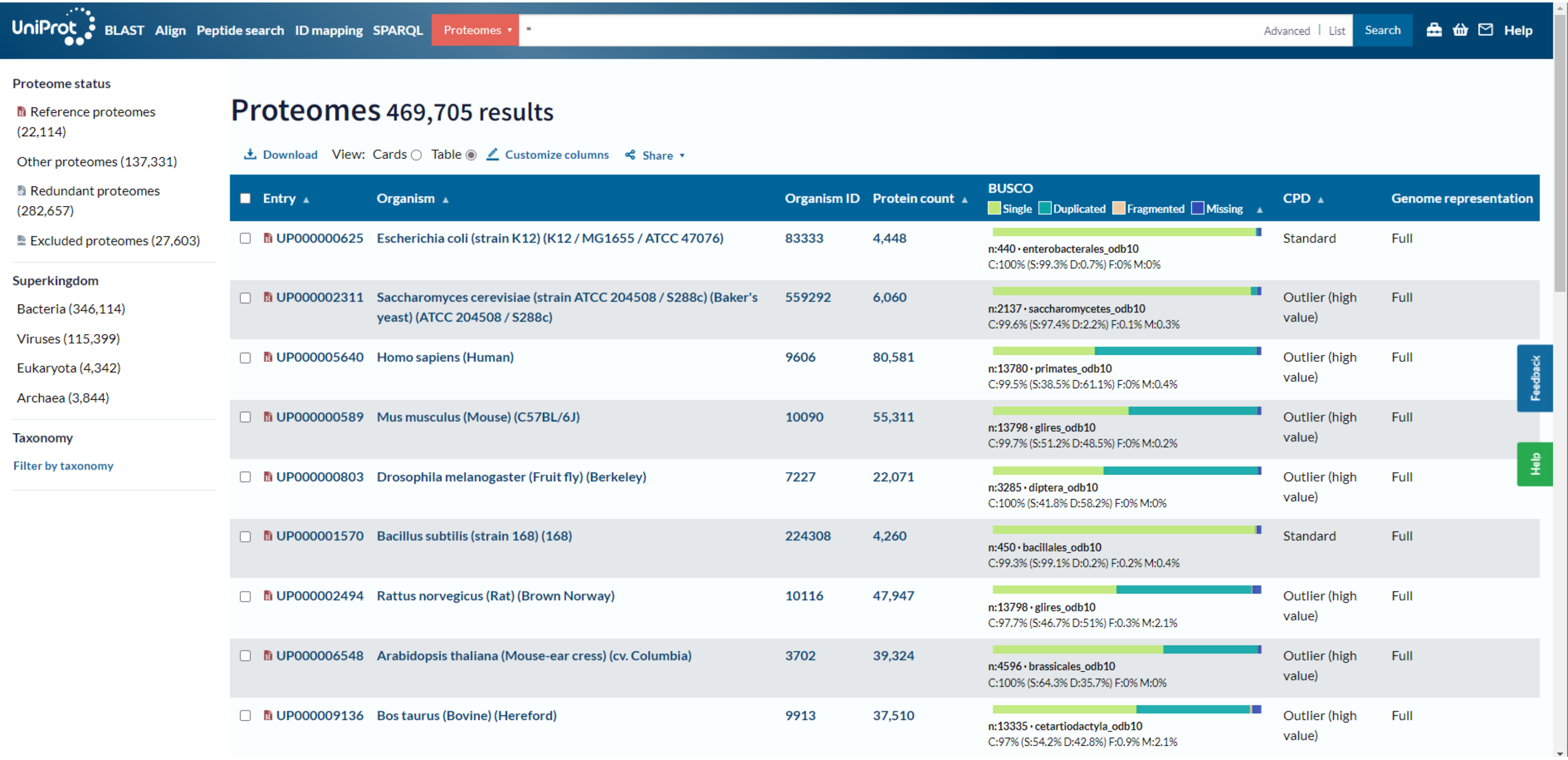
Download View: Cards ○ Table ● Customize columns Share ▾

□ Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟦 Duplicated 🟧 Frag		Genome representation
□ UP000000625	Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448	n:440 · enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:(		Full
□ UP000464341	Escherichia coli R178	1408252	270	n:440 · enterobacterales_odb10 C:0.2% (S:0.2% D:0%) F:0% M:99.8%	Outlier (low value)	Full
□ UP000000558	Escherichia coli O157:H7 (O157:H7 / Sakai / RIMD 0509952 / EHEC)	83334	5,062	n:440 · enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Standard	Full
□ UP000002006	Escherichia phage Phieco32 (Escherichia coli phage phi32)	490103	128		Close to standard (high value)	Full
□ UP000251526	Escherichia phage LL5 (Escherichia coli str. MG1655)	2233992	88		Close to standard (high value)	Full
□ UP000033806	Escherichia coli O157 typing phage 10	1508672	50		Close to standard (low value)	Full
□ UP000029103	Escherichia coli BW25113 (BW25113)	679895	4,131	n:440 · enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Close to standard (low value)	Full
□ UP000030631	Escherichia coli ER2796 (ER2796)	1245474	4,052	n:440 · enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Outlier (low value)	Full
□ UP000291548	Escherichia coli BW25113 (BW25113)	679895	4,197	n:440 · enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Close to standard (low value)	Full
□ UP000596336	Escherichia coli BW25113 (BW25113)	679895	4,188		Close to standard (low	Full

NCBI RefSeq's assessment of genome assembly representation

Proteomes

- UniProt release 2022_04 contains over 230 million sequence records, with almost 470,000 proteomes originating from completely sequenced viral, bacterial, archaeal and eukaryotic genomes available through the UniProtKB Proteomes



Proteome status

- Reference proteomes (22,114)
- Other proteomes (137,331)
- Redundant proteomes (282,657)
- Excluded proteomes (27,603)

Superkingdom

- Bacteria (346,114)
- Viruses (115,399)
- Eukaryota (4,342)
- Archaea (3,844)

Taxonomy

Filter by taxonomy

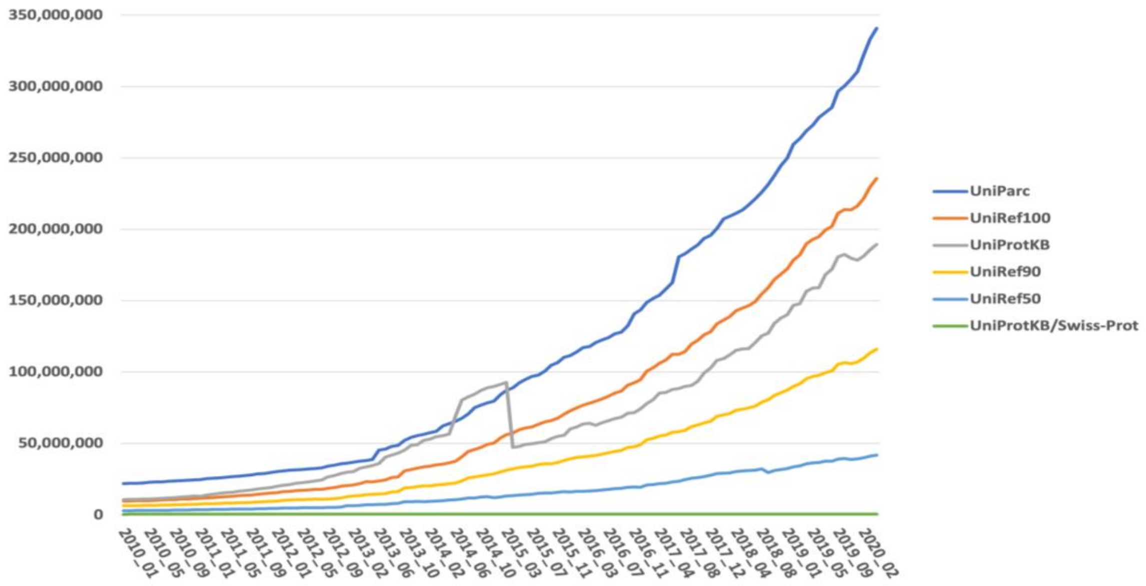
Proteomes 469,705 results

Download View: Cards ○ Table ● Customize columns Share ▾

□	Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing ▲	CPD ▲	Genome representation
□	UP000000625	Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448	n:440 · enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:0%	Standard	Full
□	UP000002311	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) (ATCC 204508 / S288c)	559292	6,060	n:2137 · saccharomycetes_odb10 C:99.6% (S:97.4% D:2.2%) F:0.1% M:0.3%	Outlier (high value)	Full
□	UP000005640	Homo sapiens (Human)	9606	80,581	n:13780 · primates_odb10 C:99.5% (S:38.5% D:61.1%) F:0% M:0.4%	Outlier (high value)	Full
□	UP000000589	Mus musculus (Mouse) (C57BL/6J)	10090	55,311	n:13798 · glires_odb10 C:99.7% (S:51.2% D:48.5%) F:0% M:0.2%	Outlier (high value)	Full
□	UP000000803	Drosophila melanogaster (Fruit fly) (Berkeley)	7227	22,071	n:3285 · diptera_odb10 C:100% (S:41.8% D:58.2%) F:0% M:0%	Outlier (high value)	Full
□	UP000001570	Bacillus subtilis (strain 168) (168)	224308	4,260	n:450 · bacillales_odb10 C:99.3% (S:99.1% D:0.2%) F:0.2% M:0.4%	Standard	Full
□	UP000002494	Rattus norvegicus (Rat) (Brown Norway)	10116	47,947	n:13798 · glires_odb10 C:97.7% (S:46.7% D:51%) F:0.3% M:2.1%	Outlier (high value)	Full
□	UP000006548	Arabidopsis thaliana (Mouse-ear cress) (cv. Columbia)	3702	39,324	n:4596 · brassicales_odb10 C:100% (S:64.3% D:35.7%) F:0% M:0%	Outlier (high value)	Full
□	UP000009136	Bos taurus (Bovine) (Hereford)	9913	37,510	n:13335 · cetartiodactyla_odb10 C:97% (S:54.2% D:42.8%) F:0.9% M:2.1%	Outlier (high value)	Full

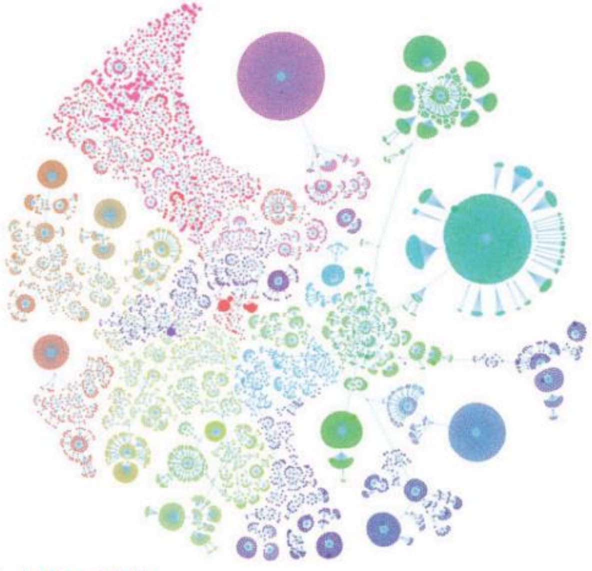
Too many…

The UniProt Knowledgebase (UniProtKB) has witnessed exponential growth in the last few years with a two-fold increase in the number of entries in 2014. This follows the vastly increased submission of multiple genomes for the same or closely related organisms. This increase was accompanied by a high level of redundancy in unreviewed UniProtKB (TrEMBL), and many sequences were over-represented in the database.



How to make useful proteomes

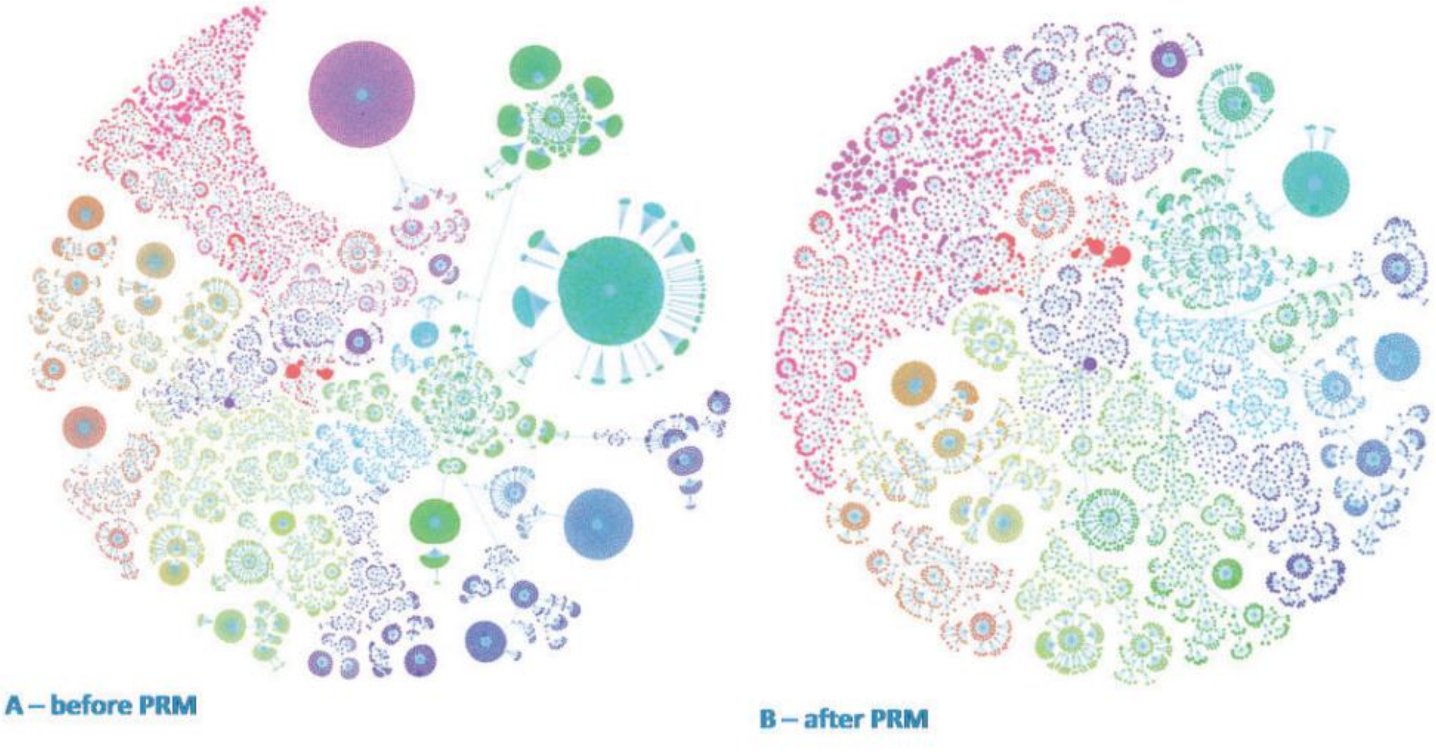
This was especially true for bacterial species where different strains of the same species have been sequenced and submitted (e.g. 1,692 strains of Mycobacterium tuberculosis, corresponding to 5.97 million entries). High redundancy led to an increase in the size of UniProtKB, and thus to the amount of data to be processed internally and by our users, but also to repetitive results in BLAST searches for over-represented sequences.



A – before PRM

Redundant proteomes

A redundancy removal process was first introduced in 2015. This process identifies and removes almost identical proteomes of the same species before their inclusion in UniProtKB and places their sequences in UniParc. Currently this process has removed ~38% of all complete proteomes (~241 million proteins) from UniProtKB.



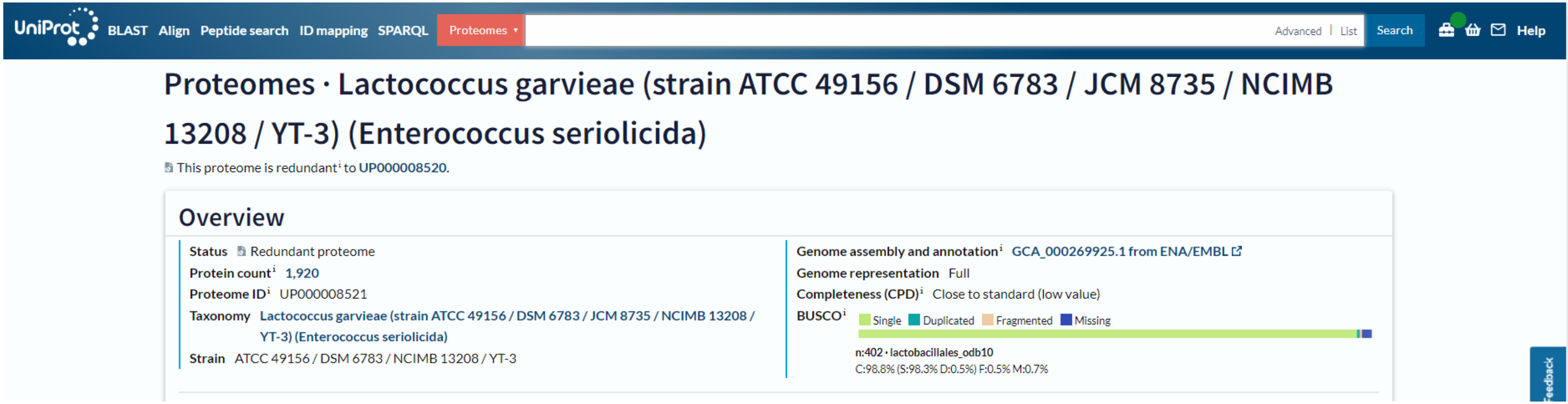
A – before PRM

B – after PRM

Redundant proteomes

The redundant proteome sequences are available through UniParc to researchers and stable proteome identifiers (of the form UPXXXXXXXXX, where Xs are integers) are maintained for each redundant proteome to ensure findability.

labelled with a specific icon



Proteomes · Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) (Enterococcus seriolicida)

This proteome is redundant to UP000008520.

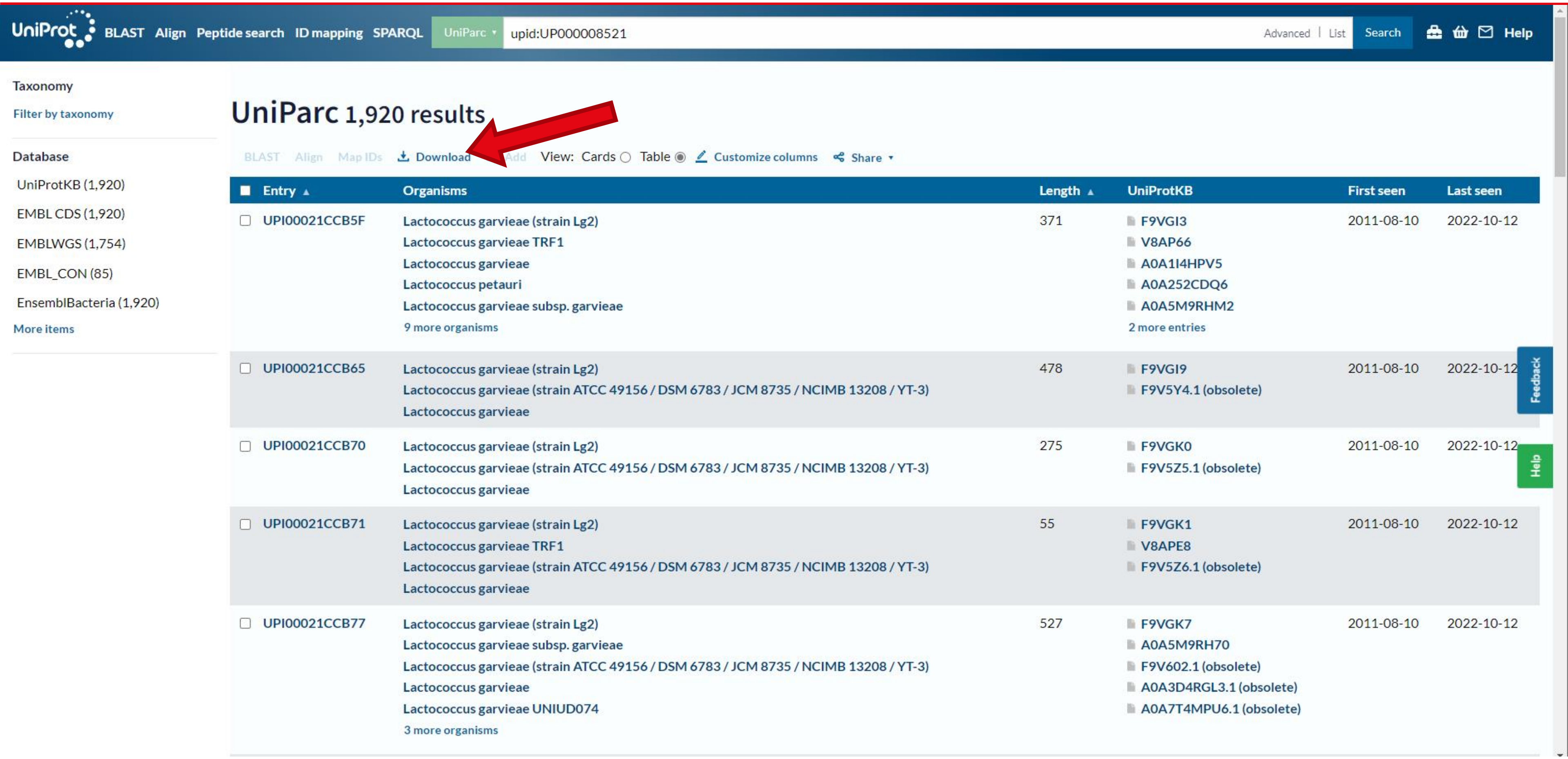
Overview

Status Redundant proteome
Protein count 1,920
Proteome ID UP000008521
Taxonomy Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) (Enterococcus seriolicida)
Strain ATCC 49156 / DSM 6783 / NCIMB 13208 / YT-3

Genome assembly and annotation GCA_000269925.1 from ENA/EMBL
Genome representation Full
Completeness (CPD) Close to standard (low value)
BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing
n:402 · lactobacillales_odb10
C:98.8% (S:98.3% D:0.5%) F:0.5% M:0.7%

https://www.uniprot.org/proteomes/UP000008521

Redundant proteomes



upid:UP000008521

Taxonomy

Filter by taxonomy

Database

UniProtKB (1,920)

EMBL CDS (1,920)

EMBLWGS (1,754)

EMBL_CON (85)

EnsemblBacteria (1,920)

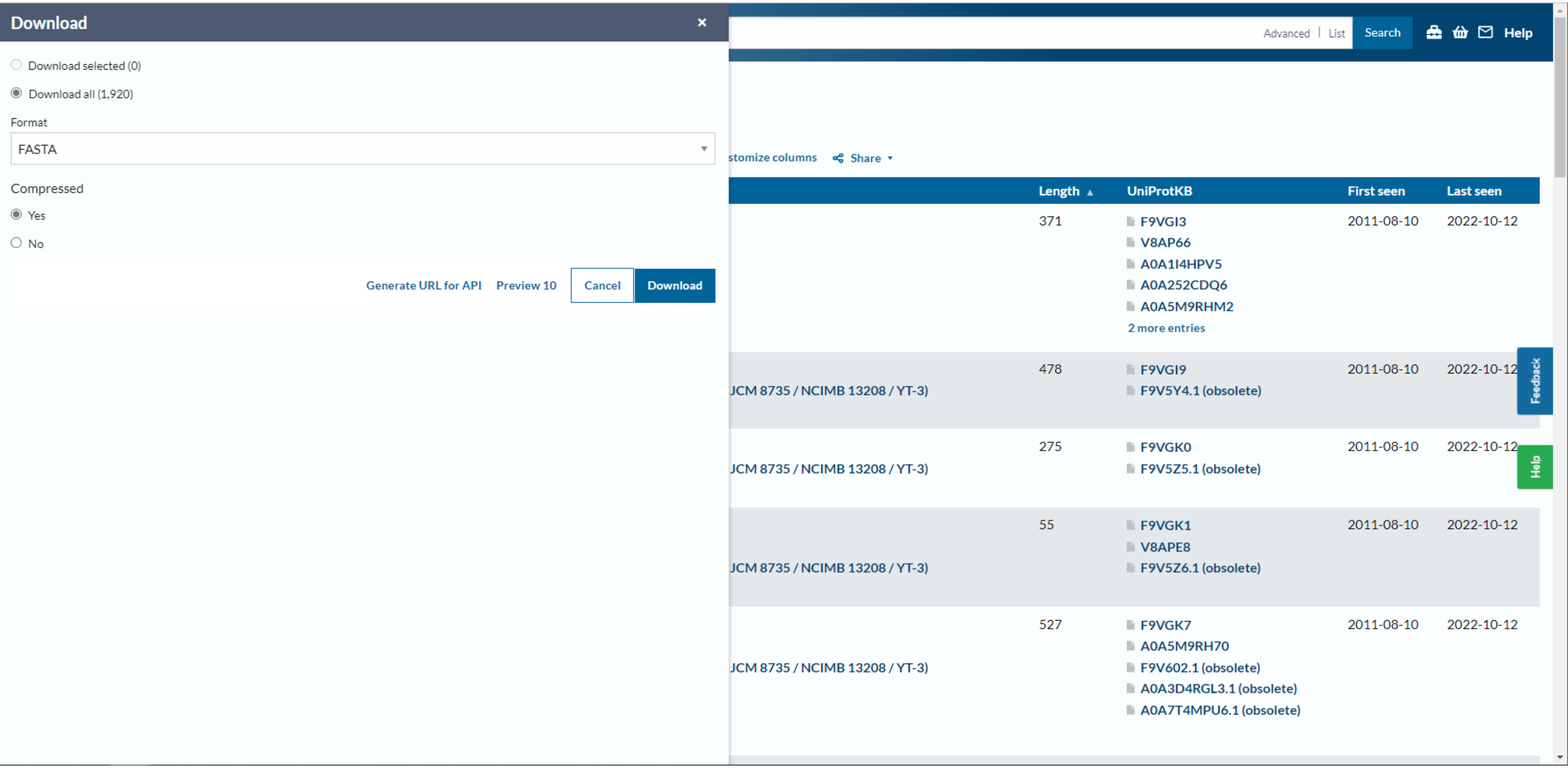
More items

UniParc 1,920 results

View: Cards ○ Table ●

Entry	Organisms	Length	UniProtKB	First seen	Last seen
UPI00021CCB5F	Lactococcus garvieae (strain Lg2) Lactococcus garvieae TRF1 Lactococcus garvieae Lactococcus petauri Lactococcus garvieae subsp. garvieae 9 more organisms	371	F9VGI3 V8AP66 A0A1I4HPV5 A0A252CDQ6 A0A5M9RHM2 2 more entries	2011-08-10	2022-10-12
UPI00021CCB65	Lactococcus garvieae (strain Lg2) Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) Lactococcus garvieae	478	F9VGI9 F9V5Y4.1 (obsolete)	2011-08-10	2022-10-12
UPI00021CCB70	Lactococcus garvieae (strain Lg2) Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) Lactococcus garvieae	275	F9VGK0 F9V5Z5.1 (obsolete)	2011-08-10	2022-10-12
UPI00021CCB71	Lactococcus garvieae (strain Lg2) Lactococcus garvieae TRF1 Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) Lactococcus garvieae	55	F9VGK1 V8APE8 F9V5Z6.1 (obsolete)	2011-08-10	2022-10-12
UPI00021CCB77	Lactococcus garvieae (strain Lg2) Lactococcus garvieae subsp. garvieae Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) Lactococcus garvieae Lactococcus garvieae UNIUD074 3 more organisms	527	F9VGK7 A0A5M9RH70 F9V602.1 (obsolete) A0A3D4RGL3.1 (obsolete) A0A7T4MPU6.1 (obsolete)	2011-08-10	2022-10-12

Redundant proteomes



Download

○ Download selected (0)

● Download all (1,920)

Format

FASTA ▾

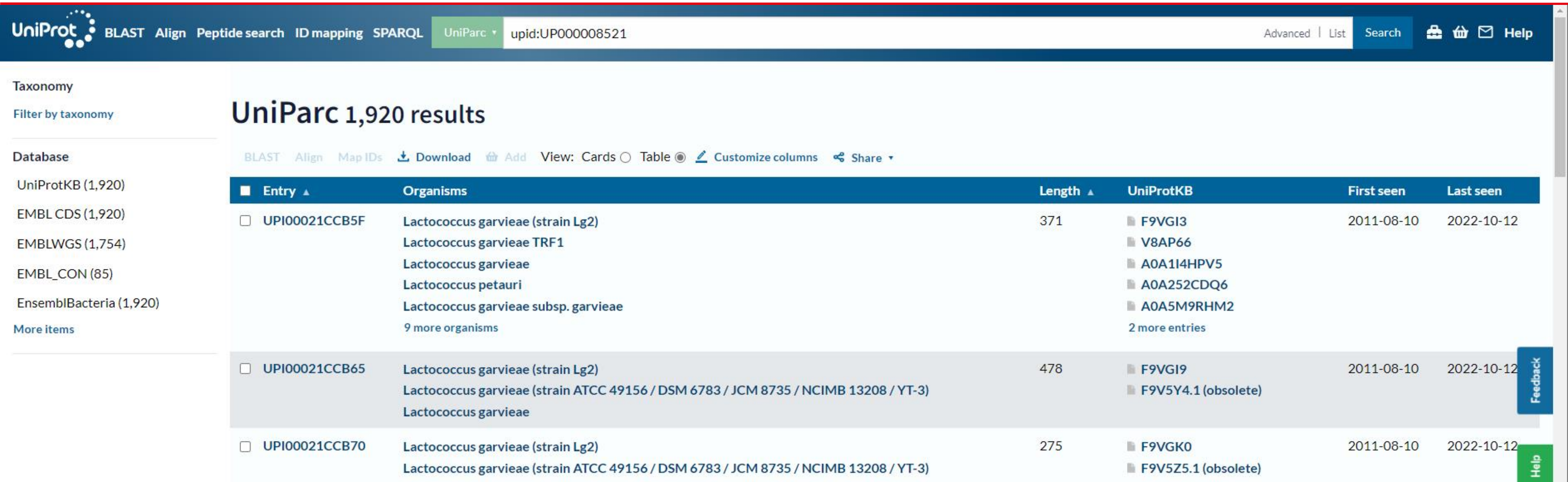
Compressed

● Yes

○ No

	Length ▲	UniProtKB	First seen	Last seen
	371	F9VGI3 V8AP66 A0A1I4HPV5 A0A252CDQ6 A0A5M9RHM2 2 more entries	2011-08-10	2022-10-12
JCM 8735 / NCIMB 13208 / YT-3)	478	F9VGI9 F9V5Y4.1 (obsolete)	2011-08-10	2022-10-12
JCM 8735 / NCIMB 13208 / YT-3)	275	F9VGK0 F9V5Z5.1 (obsolete)	2011-08-10	2022-10-12
JCM 8735 / NCIMB 13208 / YT-3)	55	F9VGK1 V8APE8 F9V5Z6.1 (obsolete)	2011-08-10	2022-10-12
JCM 8735 / NCIMB 13208 / YT-3)	527	F9VGK7 A0A5M9RH70 F9V602.1 (obsolete) A0A3D4RGL3.1 (obsolete) A0A7T4MPU6.1 (obsolete)	2011-08-10	2022-10-12

Redundant proteomes



UniProt BLAST Align Peptide search ID mapping SPARQL | UniParc ▾ | upid:UP000008521 | Advanced | List | Search | Help

Taxonomy

Filter by taxonomy

Database

UniProtKB (1,920)

EMBL CDS (1,920)

EMBLWGS (1,754)

EMBL_CON (85)

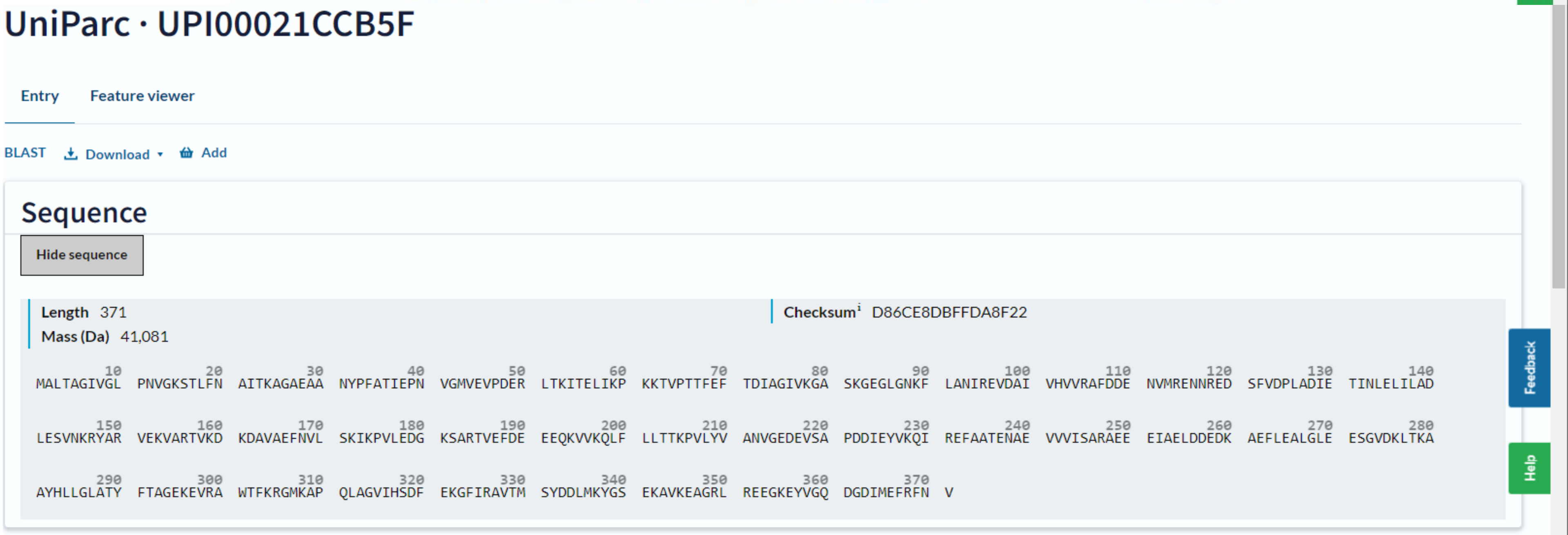
EnsemblBacteria (1,920)

More items

UniParc 1,920 results

BLAST Align Map IDs Download Add View: Cards ○ Table ● Customize columns Share ▾

□ Entry ▲	Organisms	Length ▲	UniProtKB	First seen	Last seen
□ UPI00021CCB5F	Lactococcus garvieae (strain Lg2) Lactococcus garvieae TRF1 Lactococcus garvieae Lactococcus petauri Lactococcus garvieae subsp. garvieae 9 more organisms	371	F9VGI3 V8AP66 A0A1I4HPV5 A0A252CDQ6 A0A5M9RHM2 2 more entries	2011-08-10	2022-10-12
□ UPI00021CCB65	Lactococcus garvieae (strain Lg2) Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3) Lactococcus garvieae	478	F9VGI9 F9V5Y4.1 (obsolete)	2011-08-10	2022-10-12
□ UPI00021CCB70	Lactococcus garvieae (strain Lg2) Lactococcus garvieae (strain ATCC 49156 / DSM 6783 / JCM 8735 / NCIMB 13208 / YT-3)	275	F9VGK0 F9V5Z5.1 (obsolete)	2011-08-10	2022-10-12



UniParc · UPI00021CCB5F

Entry | Feature viewer

BLAST | Download ▾ | Add

Sequence

Hide sequence

Length 371 **Checksum** D86CE8DBFFDA8F22

Mass (Da) 41,081

```
        10          20          30          40          50          60          70          80          90         100         110         120         130         140
MALTAGIVGL  PNVGKSTLFN  AITKAGAEAA  NYPFATIEPN  VGMVEVPDER  LTKITELIKP  KKTVPTTFEF  TDIAGIVKGA  SKGEGLGNKF  LANIREVDAI  VHVVRAFDDE  NVMRENNRED  SFVDPLADIE  TINLELILAD

       150         160         170         180         190         200         210         220         230         240         250         260         270         280
LESVNKRYAR  VEKVARTVKD  KDAVAEFNVL  SKIKPVLEDG  KSARTVEFDE  EEQKVVKQLF  LLTTKPVLYV  ANVGEDEVSA  PDDIEYVKQI  REFAATENAE  VVVISARAEE  EIAELDDEDK  AEFLEALGLE  ESGVDKLTKA

       290         300         310         320         330         340         350         360         370
AYHLLGLATY  FTAGEKEVRA  WTFKRGMKAP  QLAGVIHSDF  EKGFIRAVTM  SYDDLMKYGS  EKAVKEAGRL  REEGKEYVGQ  DGDIMEFRFN  V
```

Redundant proteomes

- Proteome's redundancy is only applied to bacteria and fungi for the moment.

- See FAQ:

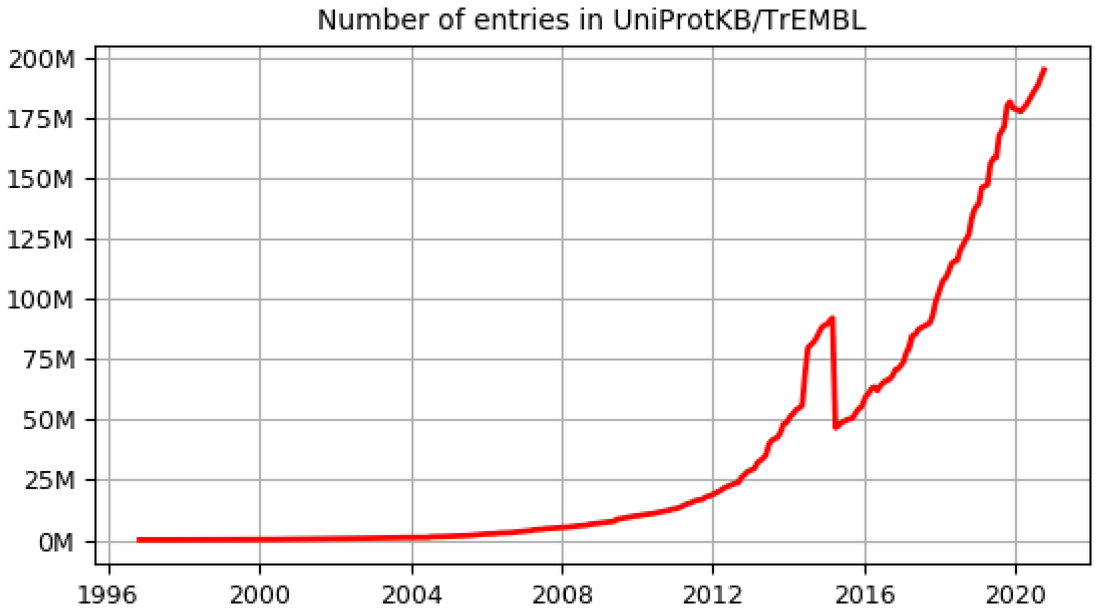
  https://www.uniprot.org/help/redundancy
- https://www.uniprot.org/help/proteome_redundancy
- http://insideuniprot.blogspot.ch/2015_05_01_archive.html

- If you need ‘to protect’ a proteome for a good reason, please contact us !

- help@uniprot.org

Still too many to be useful…

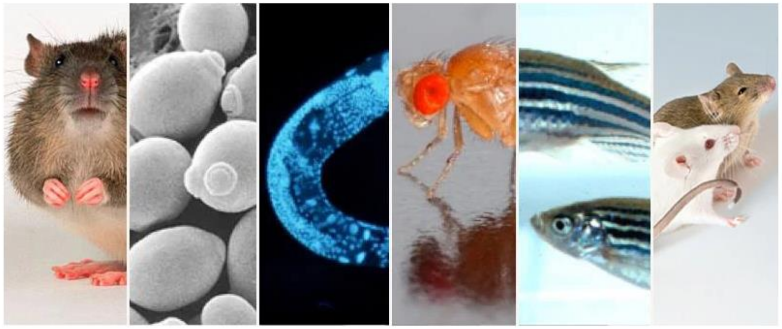
- ‘Redundant’ proteomes are excluded from UniProtKB (since 2015) (bacteria & fungi).
- Corresponding protein sequences are only available in UniParc …



https://www.uniprot.org/help/proteome_redundancy

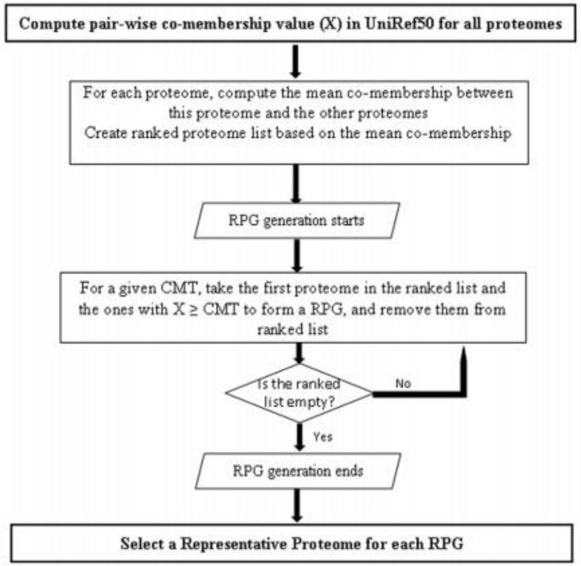
Reference proteomes

For the remaining proteomes we provide a Reference Proteome set (~9% of total proteomes) selected by the research community and supplemented with selected proteomes from a computational clustering to provide the best annotated proteome in their cluster.



ALLIANCE of GENOME RESOURCES

QfO Quest for Orthologs



Compute pair-wise co-membership value (X) in UniRef50 for all proteomes

For each proteome, compute the mean co-membership between this proteome and the other proteomes
Create ranked proteome list based on the mean co-membership

RPG generation starts

For a given CMT, take the first proteome in the ranked list and the ones with X ≥ CMT to form a RPG, and remove them from ranked list

Is the ranked list empty?

No

Yes

RPG generation ends

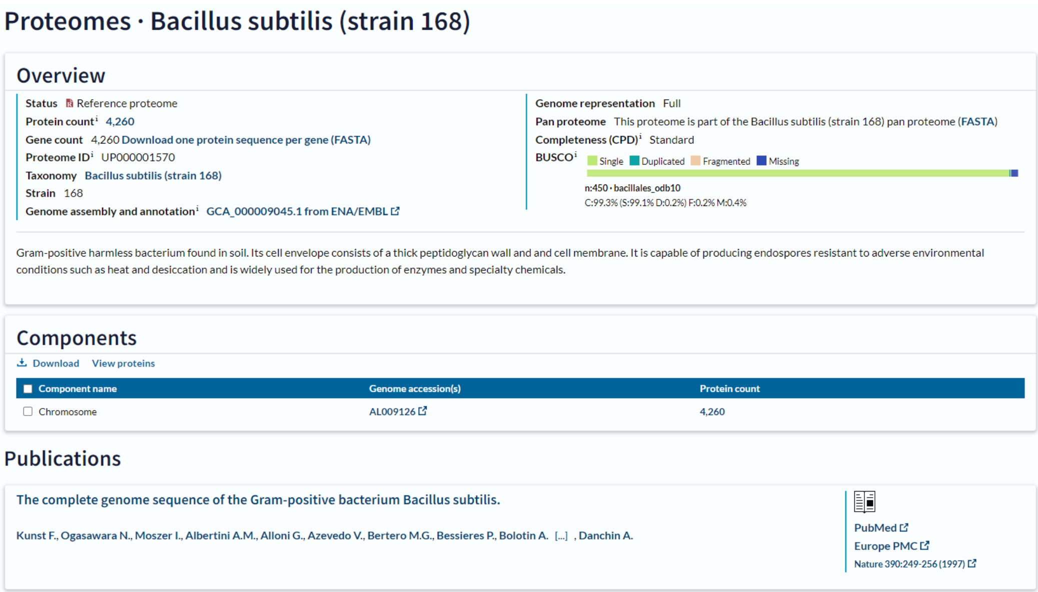
Select a Representative Proteome for each RPG

Figure 1. Flow chart of the method used to select Representative Proteomes. For details please see [materials and methods] section.
doi:10.1371/journal.pone.0018910.g001

Citation: Chen C, Natale DA, Finn RD, Huang H, Zhang J, et al. (2011) Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. PLoS ONE 6(4): e18910. doi:10.1371/journal.pone.0018910

Reference proteomes

labelled with a specific icon



Proteomes · Bacillus subtilis (strain 168)

Overview

Status Reference proteome
Protein count 4,260
Gene count 4,260 Download one protein sequence per gene (FASTA)
Proteome ID UP000001570
Taxonomy Bacillus subtilis (strain 168)
Strain 168
Genome assembly and annotation GCA_000009045.1 from ENA/EMBL

Genome representation Full
Pan proteome This proteome is part of the Bacillus subtilis (strain 168) pan proteome (FASTA)
Completeness (CPD) Standard
BUSCO Single Duplicated Fragmented Missing
n:450 · bacillales_odb10
C:99.3% (S:99.1% D:0.2%) F:0.2% M:0.4%

Gram-positive harmless bacterium found in soil. Its cell envelope consists of a thick peptidoglycan wall and and cell membrane. It is capable of producing endospores resistant to adverse environmental conditions such as heat and desiccation and is widely used for the production of enzymes and specialty chemicals.

Components

Download View proteins

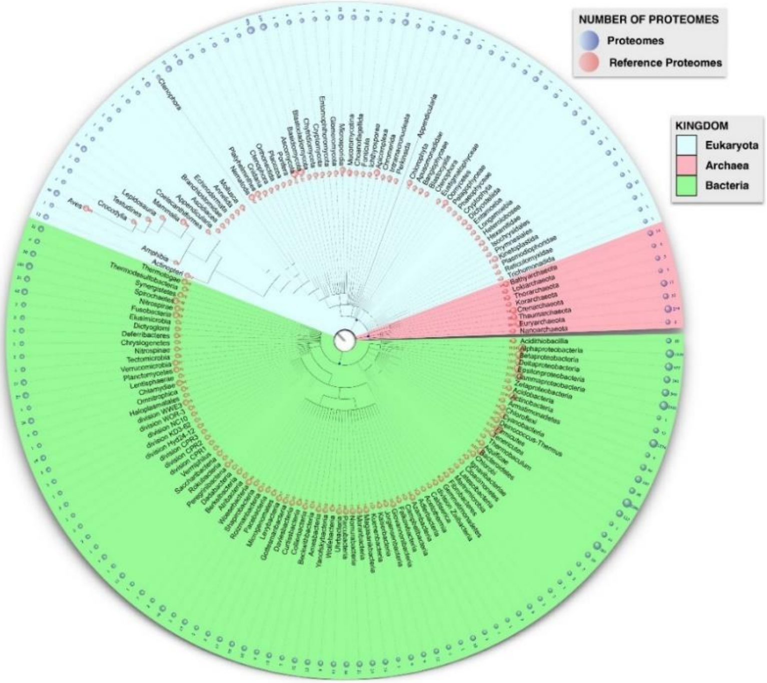
■ Component name	Genome accession(s)	Protein count
□ Chromosome	AL009126	4,260

Publications

The complete genome sequence of the Gram-positive bacterium Bacillus subtilis.

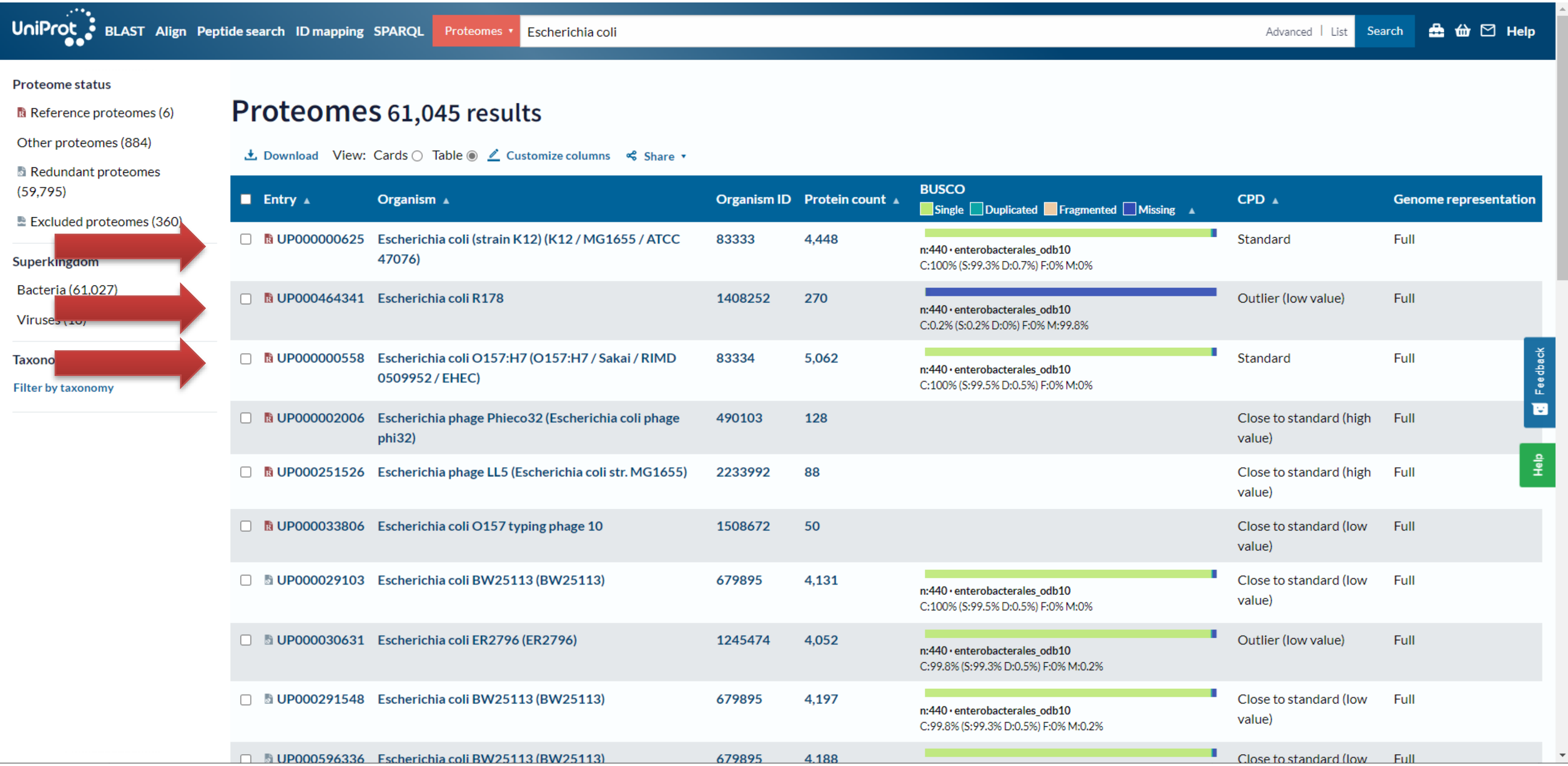
Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessieres P., Bolotin A. [...] , Danchin A.

PubMed
Europe PMC
Nature 390:249-256 (1997)



UniProtKB contains over 22,000 reference proteomes from all branches of the ToL at https://www.uniprot.org/proteomes/

UniProtKB – scope



Proteome status

- Reference proteomes (6)
- Other proteomes (884)
- Redundant proteomes (59,795)
- Excluded proteomes (360)

Superkingdom

- Bacteria (61,027)
- Viruses (18)

Taxonomy

Filter by taxonomy

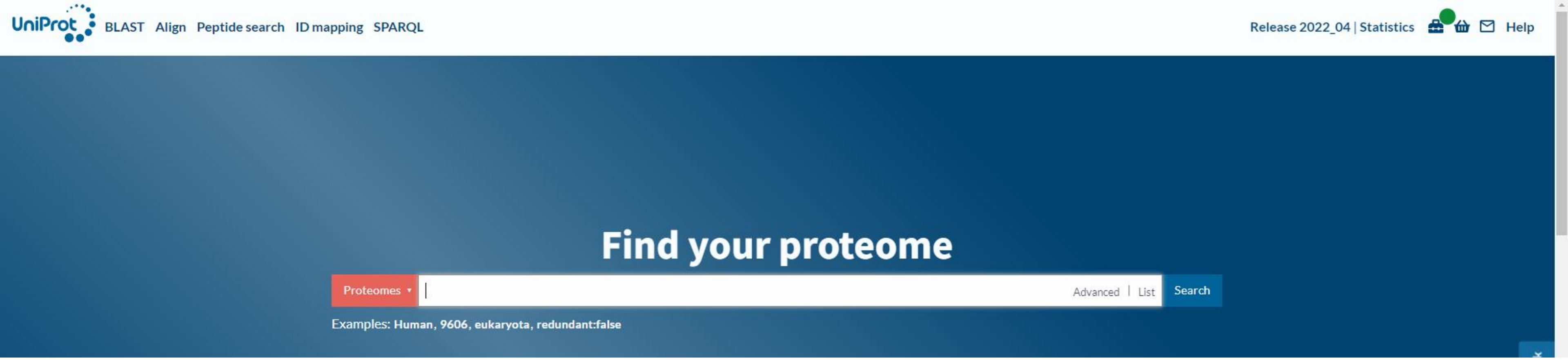
Proteomes 61,045 results

Download View: Cards ○ Table ● Customize columns Share ▾

□	Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing ▲	CPD ▲	Genome representation
□	UP000000625	Escherichia coli (strain K12) (K12 / MG1655 / ATCC 47076)	83333	4,448	n:440 • enterobacterales_odb10 C:100% (S:99.3% D:0.7%) F:0% M:0%	Standard	Full
□	UP000464341	Escherichia coli R178	1408252	270	n:440 • enterobacterales_odb10 C:0.2% (S:0.2% D:0%) F:0% M:99.8%	Outlier (low value)	Full
□	UP000000558	Escherichia coli O157:H7 (O157:H7 / Sakai / RIMD 0509952 / EHEC)	83334	5,062	n:440 • enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Standard	Full
□	UP000002006	Escherichia phage Phieco32 (Escherichia coli phage phi32)	490103	128		Close to standard (high value)	Full
□	UP000251526	Escherichia phage LL5 (Escherichia coli str. MG1655)	2233992	88		Close to standard (high value)	Full
□	UP000033806	Escherichia coli O157 typing phage 10	1508672	50		Close to standard (low value)	Full
□	UP000029103	Escherichia coli BW25113 (BW25113)	679895	4,131	n:440 • enterobacterales_odb10 C:100% (S:99.5% D:0.5%) F:0% M:0%	Close to standard (low value)	Full
□	UP000030631	Escherichia coli ER2796 (ER2796)	1245474	4,052	n:440 • enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Outlier (low value)	Full
□	UP000291548	Escherichia coli BW25113 (BW25113)	679895	4,197	n:440 • enterobacterales_odb10 C:99.8% (S:99.3% D:0.5%) F:0% M:0.2%	Close to standard (low value)	Full
□	UP000596336	Escherichia coli BW25113 (BW25113)	679895	4,188		Close to standard (low	Full

E.coli (3 reference proteomes; 2 chromosomes, 1 plasmid) & E.coli phages (3 reference proteomes)

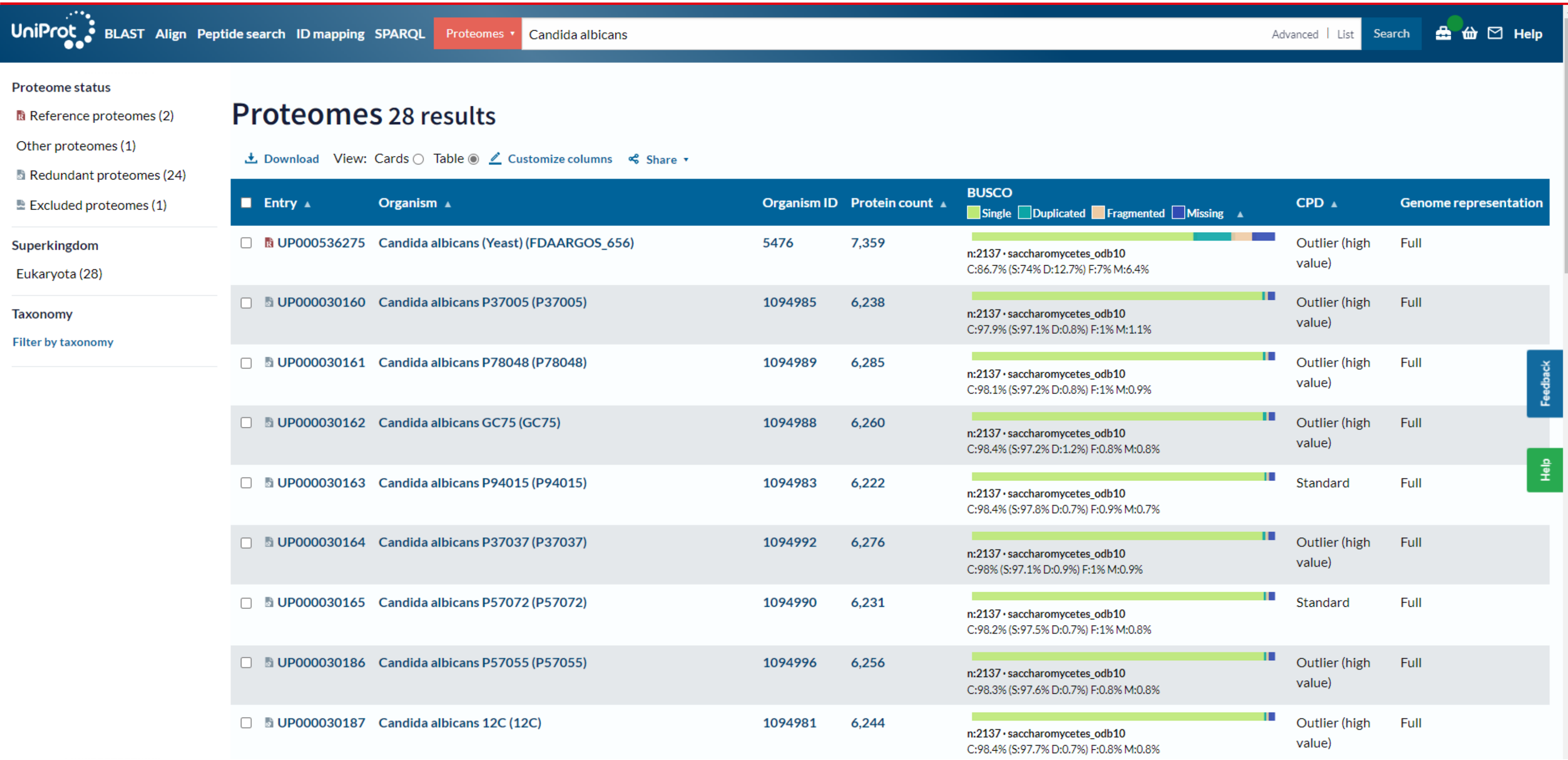
Exercise



- @UniProtKB: look for the Candida albicans proteomes
- How many proteomes ?
- What is the Proteome ID of the Candida albicans Reference Proteome ?

- To which strain does it correspond ?
- How many Swiss-Prot records ?
- How many TrEMBL records ?

Exercise



Candida albicans

Proteome status

- Reference proteomes (2)
- Other proteomes (1)
- Redundant proteomes (24)
- Excluded proteomes (1)

Superkingdom

- Eukaryota (28)

Taxonomy

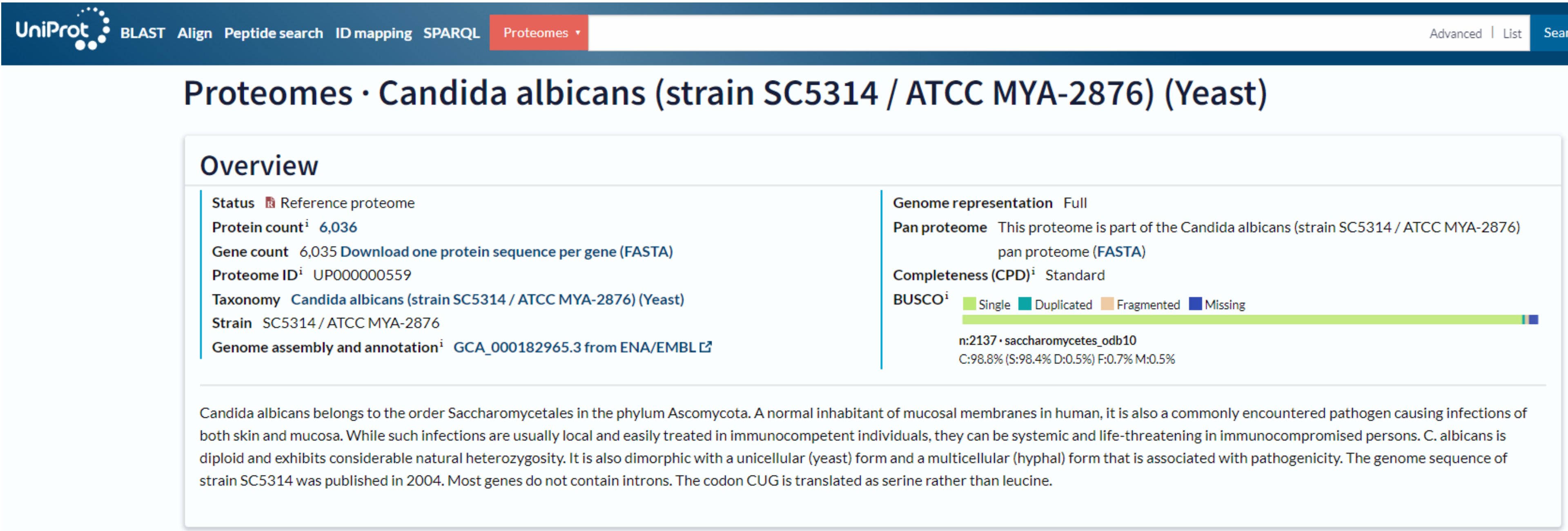
Filter by taxonomy

Proteomes 28 results

Download View: Cards ○ Table ● Customize columns Share ▾

□	Entry ▲	Organism ▲	Organism ID	Protein count ▲	BUSCO 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing ▲	CPD ▲	Genome representation
□	UP000536275	Candida albicans (Yeast) (FDAARGOS_656)	5476	7,359	n:2137 · saccharomycetes_odb10 C:86.7% (S:74% D:12.7%) F:7% M:6.4%	Outlier (high value)	Full
□	UP000030160	Candida albicans P37005 (P37005)	1094985	6,238	n:2137 · saccharomycetes_odb10 C:97.9% (S:97.1% D:0.8%) F:1% M:1.1%	Outlier (high value)	Full
□	UP000030161	Candida albicans P78048 (P78048)	1094989	6,285	n:2137 · saccharomycetes_odb10 C:98.1% (S:97.2% D:0.8%) F:1% M:0.9%	Outlier (high value)	Full
□	UP000030162	Candida albicans GC75 (GC75)	1094988	6,260	n:2137 · saccharomycetes_odb10 C:98.4% (S:97.2% D:1.2%) F:0.8% M:0.8%	Outlier (high value)	Full
□	UP000030163	Candida albicans P94015 (P94015)	1094983	6,222	n:2137 · saccharomycetes_odb10 C:98.4% (S:97.8% D:0.7%) F:0.9% M:0.7%	Standard	Full
□	UP000030164	Candida albicans P37037 (P37037)	1094992	6,276	n:2137 · saccharomycetes_odb10 C:98% (S:97.1% D:0.9%) F:1% M:0.9%	Outlier (high value)	Full
□	UP000030165	Candida albicans P57072 (P57072)	1094990	6,231	n:2137 · saccharomycetes_odb10 C:98.2% (S:97.5% D:0.7%) F:1% M:0.8%	Standard	Full
□	UP000030186	Candida albicans P57055 (P57055)	1094996	6,256	n:2137 · saccharomycetes_odb10 C:98.3% (S:97.6% D:0.7%) F:0.8% M:0.8%	Outlier (high value)	Full
□	UP000030187	Candida albicans 12C (12C)	1094981	6,244	n:2137 · saccharomycetes_odb10 C:98.4% (S:97.7% D:0.7%) F:0.8% M:0.8%	Outlier (high value)	Full

Exercise



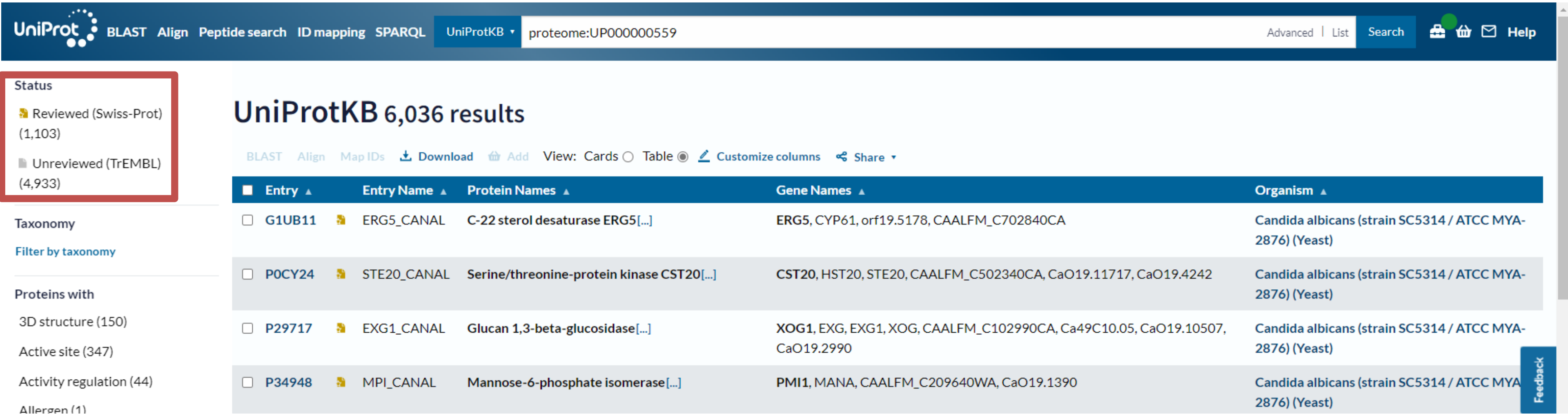
Proteomes · Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)

Overview

Status Reference proteome
Protein count[i] 6,036
Gene count 6,035 Download one protein sequence per gene (FASTA)
Proteome ID[i] UP000000559
Taxonomy Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)
Strain SC5314 / ATCC MYA-2876
Genome assembly and annotation[i] GCA_000182965.3 from ENA/EMBL

Genome representation Full
Pan proteome This proteome is part of the Candida albicans (strain SC5314 / ATCC MYA-2876) pan proteome (FASTA)
Completeness (CPD)[i] Standard
BUSCO[i] 🟩 Single 🟩 Duplicated 🟧 Fragmented 🟦 Missing
n:2137 · saccharomycetes_odb10
C:98.8% (S:98.4% D:0.5%) F:0.7% M:0.5%

Candida albicans belongs to the order Saccharomycetales in the phylum Ascomycota. A normal inhabitant of mucosal membranes in human, it is also a commonly encountered pathogen causing infections of both skin and mucosa. While such infections are usually local and easily treated in immunocompetent individuals, they can be systemic and life-threatening in immunocompromised persons. C. albicans is diploid and exhibits considerable natural heterozygosity. It is also dimorphic with a unicellular (yeast) form and a multicellular (hyphal) form that is associated with pathogenicity. The genome sequence of strain SC5314 was published in 2004. Most genes do not contain introns. The codon CUG is translated as serine rather than leucine.



proteome:UP000000559

Status
Reviewed (Swiss-Prot) (1,103)
Unreviewed (TrEMBL) (4,933)

Taxonomy
Filter by taxonomy

Proteins with
3D structure (150)
Active site (347)
Activity regulation (44)
Allergen (1)

UniProtKB 6,036 results

View: Cards ○ Table ●

□ Entry ▲		Entry Name ▲	Protein Names ▲	Gene Names ▲	Organism ▲
□ G1UB11		ERG5_CANAL	C-22 sterol desaturase ERG5[...]	ERG5, CYP61, orf19.5178, CAALFM_C702840CA	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)
□ P0CY24		STE20_CANAL	Serine/threonine-protein kinase CST20[...]	CST20, HST20, STE20, CAALFM_C502340CA, CaO19.11717, CaO19.4242	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)
□ P29717		EXG1_CANAL	Glucan 1,3-beta-glucosidase[...]	XOG1, EXG, EXG1, XOG, CAALFM_C102990CA, Ca49C10.05, CaO19.10507, CaO19.2990	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)
□ P34948		MPI_CANAL	Mannose-6-phosphate isomerase[...]	PMI1, MANA, CAALFM_C209640WA, CaO19.1390	Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)