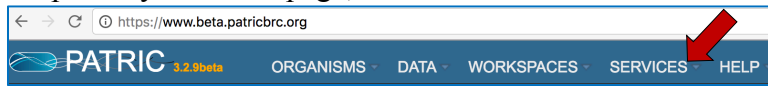


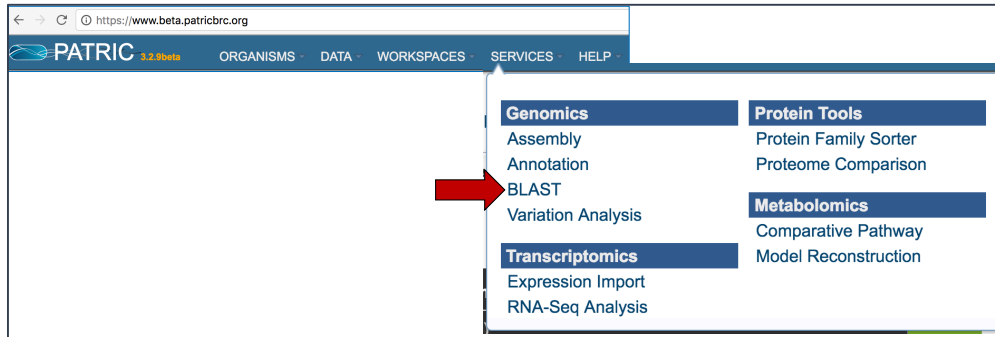
Submitting a BLAST (Basic Local Alignment Search Tool) job at PATRIC

I. Locating the BLAST Service

1. At the top of any PATRIC page, find the Services tab.



2. Click on BLAST (red arrow).



3. This will open up the BLAST landing page where researchers can do nucleotide or amino acid BLAST searches.

A screenshot of the PATRIC BLAST landing page. The page title is 'Services BLAST'. Below the title, a description states: 'The BLAST Search allows you to search against public or private genomes in PATRIC or other reference databases using a DNA or protein sequence and find matching genomes, genes, RNAs, or proteins.' The main form area is titled 'Sequence' and contains a large text input box with the placeholder text: 'Enter a query nucleotide or protein sequence to search. Multiple query sequences are currently not supported.' Below the text box are two dropdown menus labeled 'Program' and 'Database'. At the bottom of the form, there is a link for 'ADVANCED OPTIONS' and a 'Search' button.

II. Loading a sequence and choosing a type of BLAST

1. Cut and paste a sequence into the Sequence box. Depending upon the sequence, this will open the drop down box under Program, showing the types of BLAST available(1).

Definitions of the types of BLAST searches are as follows:

- BLAST: searches nucleotide databases using a nucleotide query
- BLASTP: searches protein databases using a protein query
- BLASTX: searches protein databases using a translated nucleotide query
- TBLASTX: searches translated nucleotide databases using a translated nucleotide query
- TBLASTN: searches translated nucleotide subjects using a protein query

Services

BLAST

The BLAST Search allows you to search against public or private genomes in PATRIC or other reference databases using a DNA or protein sequence and find matching genomes, genes, RNAs, or proteins.

Sequence

```
>fig|511145.12.peg.3032|b2938|VBIescColl29921_3032| Biosynthetic arginine decarboxylase (EC 4.1.1.19) [Escherichia coli str. K-12 substr. MG1655 | 511145.12]
MSSQEASKMLRTYNIWGNYYDVNELGHISVCPDPVPEARVDLAQLVKTREAGQRL
PALFCFPQILQHRLRSINAAFKRARESYGNGDYFLVYPKVNQHRRVIESLIHSGEPLG
LEAGSKAELMAVLAHAGMTRSVIVCNGYKDREYIRLALIGKMGHKVYLVEIKMSEIAIV
LDEAERLNVVPRLGVRARLASQSGKQSSGGEKSKFGLAATQVLQVETLREAGRLDSL
QLLHFHLSQMANIRDIATGVRESARFYVELHKLGVNIQCFDVGGLGVDYEGTRSQSDC
SVNYGLNEYANNI IWAIGDACEENGLPHPTVITESGRAVTAHHTVLVSNII GVERNEYTV
PTAPADAPRALQSMWETWQEMHEPGTRRSLEWLHDSQMDLHDIIHIGYSSGIFSLQERA
WAEQLYLSMCHEVQQLDPQNRHRPIIDELQERMADKMYVNFSLFQSPMDAWGIDQLFP
VLPLEGLDQVPERRAVLLDITCDSGDAIDHYIDGDIATTMPPEYDPENPPMLGFFFMVG
AYQEILGNMHNLFGDTEAVDVVFVFDGSEVELSDEGDTVADMLQYVQLDPKTLTQFRD
QVKKTDLDAELQQQFLEEFAGLYGYTYLEDE
```

Program

blastp - search protein database using a protein query

blastp - search protein database using a protein query

tblastn - search translated nucleotide database using a protein query

ADVANCED OPTIONS ▾

Search

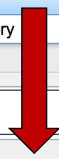
2. Clicking on the algorithm of choice will close the drop down box and display the choice in the Program text box.

Program

blastp - search protein database using a protein query

blastp - search protein database using a protein query

tblastn - search translated nucleotide database using a protein query



Sequence

```
>fig|511145.12.peg.3032|b2938|VBIescColl29921_3032| Biosynthetic arginine decarboxylase (EC 4.1.1.19) [Escherichia coli str. K-12 substr. MG1655 | 511145.12]
MSSQEASKMLRTYNIWGNYYDVNELGHISVCPDPVPEARVDLAQLVKTREAGQRL
PALFCFPQILQHRLRSINAAFKRARESYGNGDYFLVYPKVNQHRRVIESLIHSGEPLG
LEAGSKAELMAVLAHAGMTRSVIVCNGYKDREYIRLALIGKMGHKVYLVEIKMSEIAIV
LDEAERLNVVPRLGVRARLASQSGKQSSGGEKSKFGLAATQVLQVETLREAGRLDSL
QLLHFHLSQMANIRDIATGVRESARFYVELHKLGVNIQCFDVGGLGVDYEGTRSQSDC
SVNYGLNEYANNI IWAIGDACEENGLPHPTVITESGRAVTAHHTVLVSNII GVERNEYTV
PTAPADAPRALQSMWETWQEMHEPGTRRSLEWLHDSQMDLHDIIHIGYSSGIFSLQERA
WAEQLYLSMCHEVQQLDPQNRHRPIIDELQERMADKMYVNFSLFQSPMDAWGIDQLFP
VLPLEGLDQVPERRAVLLDITCDSGDAIDHYIDGDIATTMPPEYDPENPPMLGFFFMVG
AYQEILGNMHNLFGDTEAVDVVFVFDGSEVELSDEGDTVADMLQYVQLDPKTLTQFRD
QVKKTDLDAELQQQFLEEFAGLYGYTYLEDE
```

Program

blastp - search protein database using a protein query

Database

Reference or Representative Genome proteins (faa)

III. Selecting a database

1. PATRIC provides a variety of databases that selected sequences can be compared to. If the sequence selected is a protein, the available databases are as follows:

1a. The Reference or Representative Genome proteins (faa) includes those genomes that RefSeq has given special status(2). The reference genomes represent the highest quality dataset that is supported by curation by NCBI scientific staff, and the representative genomes are another high-quality selection that were identified at RefSeq by clustering genomes and applying weighting metrics that include consideration of species-level taxonomic classification (e.g., a preference for type strain) and assembly quality (e.g. a preference for complete genomes but WGS is allowed).

1b. Plasmid contig proteins (faa) include all the proteins that are found on contigs that are identified as being from plasmids.

1c. Specialty gene reference proteins (faa) contain all the genes used by PATRIC to tag genes that are of special interest. These include genes that have been identified as being virulence factors, as important in antibiotic resistance or susceptibility, are homologs with human genes, or have been investigated as being a drug target.

1d. Search within a selected genome(s) allows researchers to choose specific genomes that they wish to BLAST against.

1e. Search within selected genome group allows researcher to BLAST against any of the genome groups that they have created and are stored in their workspace.

1f. Search with selected taxon allows researchers to BLAST their sequence against any taxon level available in PATRIC.

Database
Reference or Representative Genome proteins (faa)
Reference or Representative Genome proteins (faa)
plasmid contigs proteins (faa)
Specialty gene reference proteins (faa)
Search within selected genomes
Search within selected genome group
Search within selected taxon

2. If the sequence selected for BLAST analysis is nucleotide, the available databases are as follows:

2a. The Reference or Representative Genome genes (fna), or fasta nucleic acid, includes those genomes that RefSeq has given special status(2). .fna is used generically to specify nucleic acids. The reference genomes represent the highest quality dataset that is supported by curation by NCBI scientific staff, and the representative genomes are another high-quality selection that were identified at RefSeq by clustering genomes and applying weighting metrics that include consideration of species-level taxonomic classification (e.g., a preference for type strain) and assembly quality (e.g. a preference for complete genomes but WGS is allowed). This will include non-coding sequences, like intergenic regions.

2b. The Reference or Representative Genome features (ffn) is the FASTA nucleotide of gene regions, and this database contains all the coding regions across this special selection of genomes.

2c. The Reference and Representative Genome features (frn) is the FASTA non-coding RNA, and includes all the non-coding RNA regions for a genome (tRNA, rRNA).

2d. PATRIC 16sRNA genes (frn) includes all the 16s rRNA genes across all the genomes available in PATRIC.

2e. Transcriptomic genomes (fna) will BLAST against all the genome sequences that have expression data associated with them that are publically available in PATRIC. This will include non-coding sequences, like intergenic regions.

2f. Transcriptomics Genomes feature (ffn) will BLAST against all the coding sequences from the genomes that have expression data associated with them that are publically available in PATRIC.

2g. Plasmid contigs (fna) will BLAST against all the sequences identified as coming from plasmids that are available in PATRIC. This will include non-coding sequences, like intergenic regions.

2h. Plasmid contig feature (ffn) will BLAST against all the coding sequences identified as coming from plasmids that are available in PATRIC.

2i. Search within selected genomes allows researchers to choose specific genomes that they wish to BLAST against.

2j. Search within selected genome group allows researcher to BLAST against any of the genome groups that they have created and are stored in their workspace.

2k. Search with selected taxon allows researchers to BLAST their sequence against any taxon level available in PATRIC.

Database
Reference or Representative Genomes (fna)
Reference or Representative Genomes (fna)
Reference or Representative Genome features (ffn)
Reference or Representative Genome RNAs features (frn)
PATRIC 16s RNA Genes (frn)
Transcriptomics Genomes (fna)
Transcriptomics Genomes features (ffn)
plasmid contigs (fna)
plasmid contigs features (ffn)
Search within selected genomes
Search within selected genome group
Search within selected taxon

IV. BLASTing against gene features or contigs

- Depending upon query type, researchers will be able to choose to search entire genomes or limit the search to only features. When BLASTN, TBLASTN, or TBLASTX are selected, researchers can choose to search against either contigs or features. When BLASTP or BLASTX are selected, the search is limited to features.

ADD GENOMES TO SEARCH: <input type="text" value="e.g. M. tuberculosis CDC1551"/>		
SEARCH FOR: <div> Genomic sequences (contigs) <div> Genomic sequences (contigs) </div> Genomic features (genes, proteins or RNAs) </div>		
MAX HITS: <input type="text" value="50"/>	E VALUE THRESHOLD: <input type="text" value="10"/>	

V. Adjusting the BLAST parameters

1. Once a database to BLAST against is selected, researchers have the option of further refining the BLAST job by using the Advanced Options.

The diagram illustrates the first step of adjusting BLAST parameters. It shows two screenshots of the NCBI BLAST web interface. The top screenshot shows the 'Database' dropdown menu set to 'Reference or Representative Genomes (fna)' and a 'Search' button. A red arrow points to the 'ADVANCED OPTIONS' link. A large black arrow points down to the second screenshot, which shows the 'ADVANCED OPTIONS' section expanded, revealing the 'BLAST Parameters' section. This section includes a 'MAX HITS' dropdown menu (set to 50) and an 'E VALUE THRESHOLD' input field (set to 10). A 'Search' button is also visible at the bottom of the expanded section.

2. Researchers can adjust both the number of hits returned, and the E value threshold. There are limits to the number of hits returned. To see the available number, click on the arrow at the end of the text box under Max Hits. This will open a drop down box that allows researchers to choose 1, 10, 50, 100 or 500 hits.

The diagram illustrates the second step of adjusting BLAST parameters. It shows two screenshots of the NCBI BLAST web interface. The top screenshot shows the 'BLAST Parameters' section with the 'MAX HITS' dropdown menu (set to 50) and the 'E VALUE THRESHOLD' input field (set to 10). A red arrow points to the dropdown arrow of the 'MAX HITS' menu. A large black arrow points down to the second screenshot, which shows the 'MAX HITS' dropdown menu open, displaying a list of options: 1, 10, 50, 100, and 500. The '50' option is highlighted. The 'Search' button is visible at the bottom of the interface.

VI. Submitting the BLAST job

1. Once the sequence has been uploaded, the program and database selected, the BLAST parameters adjusted, the job can be started by clicking the Search button at the bottom of the page.

Sequence

>fig|511145.12.peg.3032|b2938|VBIescCol129921_3032| Biosynthetic arginine decarboxylase (EC 4.1.1.19) [Escherichia coli str. K-12 substr. MG1655 | 511145.12]
MSSQEAQKMLRTYNIAGWNNYYDVNELGHISVCPDDPVPARVDLAQLVKTRAQQRLL
PALFCFPQILQHLRLSINAAPKRARESYGYNGDYFLVYPIKVNQHRVIESLIHSGEPLG
LEAGSKAELMAVLAHAGMTRSVIVCNGYKDREYIRLALIGKMGHKVYLIEKMSAIAIV
LDEAERLGNVVRRLASQSGCKWQSSGGEKSKFGLAATQVLQVETLREAGRLDSL
QLLHFLHLSQMANIRDIATGVRESAREYVELHKLGVNIQCFDVGGLGVDYEGTRSQSDC
SVNYGLNEYANNIWAIGDACEENGLPHPTVITESTGRAVTAHHTVLVSNIGVERNEYTV
PTAPADAPRALQSMWETWQEMHEPGTRRLREWLHDSQMDLHDHIGYSSGIFSLQERA
WAEQLYLSMCHEVQKQLDPQNRHAPRIDELQERMAKMYNVSFLQSPMDAWGIDQLFP
VLPLEGLDQVPERRAVLLDITCDSGAIDHYIDGDIATTMPPEYDPENPMLGFFVMVG
AYQEILGNMHNILFGDTEAVDVVFVFDGSEVEVLSDEGDTVADMLQYVQLDPKTLTLTQFRD
QVKKTLDLAELQQQFLEEFAGLYGYTYLEDE

Program

blastp - search protein database using a protein query

Database

Reference or Representative Genome proteins (faa)

ADVANCED OPTIONS 4

BLAST Parameters

MAX HITS:

50

E VALUE THRESHOLD:

10

Search

VII. Examining the BLAST results

- When the BLAST results are ready, the page will reload showing the name of the organism, the query and subject coverage, the score and the E value. Depending on the type of BLAST selected, researchers will also see locus tags, gene symbols and functional descriptions for the features, or information about the genomic contigs.

Services

BLAST

The BLAST Search allows you to search against public or private genomes in PATRIC or other reference databases using a DNA or protein sequence and find matching genomes, genes, RNAs, or proteins.

Edit form and resubmit

<input type="checkbox"/>	Genome	PATRIC ID	RefSeq Locus Tag	Gene	Product	Length (NT)	Length (AA)	ALN Length	Identit (%)	Query cover (%)	Subject cover (%)	Score	E value	
<input type="checkbox"/>	Shigella sonnei strain H140920393	fig 624.361.peg.2754	ERS432308_02697	speA	Biosynthetic arginine decarboxylase (EC 4	1977	658	658	98	88	96	1273	0	
<input type="checkbox"/>	Escherichia coli strain NCTC9001	fig 562.7382.peg.118	ERS451419_01175	speA	Biosynthetic arginine decarboxylase (EC 4	1977	658	658	98	88	96	1273	0	
<input type="checkbox"/>	Escherichia coli str. K-12 substr. MG1655	fig 511145.12.peg.30	b2938	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1272	0	
<input type="checkbox"/>	Escherichia coli O83:H1 str. NRG 857C	fig 685038.3.peg.294	NRG857_14440		Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Escherichia coli IA39	fig 585057.6.peg.348	ECIA39_3358	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Escherichia coli UMN028	fig 585056.7.peg.346	ECUMN_3290	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Shigella sp. D9	fig 556266.3.peg.165	ShiD9_0101000101C		Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Escherichia coli O157:H7 str. Sakai	fig 386585.9.peg.398	ECs3814		Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Shigella sonnei Sd046	fig 300269.12.peg.36	SSON_3092	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Shigella boydii Sb227	fig 300268.11.peg.37	SBO_3051	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Escherichia coli O104:H4 str. 2011C-3493	fig 1133852.3.peg.99	O3K_04755		Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1271	0	
<input type="checkbox"/>	Escherichia fergusonii ATCC 35469	fig 585054.5.peg.278	EFER_2878	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1269	0	
<input type="checkbox"/>	Shigella flexneri 2a str. 301	fig 198214.7.peg.348	SF2931	speA	Biosynthetic arginine decarboxylase (EC 4	1917	638	638	98	88	99	1267	0	
<input type="checkbox"/>	Escherichia albertii TW07627	fig 502347.3.peg.922	ESCA87627_0889	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	98	88	100	1260	0	
<input type="checkbox"/>	Shigella dysenteriae Sd197	fig 300267.13.peg.37	SDY_3134	speA	Biosynthetic arginine decarboxylase (EC 4	1875	624	624	98	87	100	1254	0	
<input type="checkbox"/>	Salmonella enterica subsp. enterica serov. fig 290319.15.peg.31	SPA2950	speA	Biosynthetic arginine decarboxylase (EC 4	1899	632	632	96	88	100	1246	0		
<input type="checkbox"/>	Citrobacter koseri ATCC BAA-895	fig 290338.8.peg.360	KO_04316		Biosynthetic arginine decarboxylase (EC 4	1899	632	632	96	88	100	1246	0	

- Clicking on a single check box in front of a specific return will do two things. It will populate the vertical green bar with all the possible downstream analysis tools or processes that can be deployed with that selection. With a single choice, those possibilities include;
 - The ability to download information on the sequence
 - The fasta file (protein or nucleotide)
 - The ability to look at other identifiers linked to the gene by the ID Mapping tool, what pathway incudes the selected gene

Download Selection

View FASTA Data

Multiple Sequence Alignment

ID Mapping

Pathway Summary

Copy selection to a new or existing group

Switch to Genome List View. Press and Hold for more options.

Switch to Feature List View. Press and Hold for more options.

HIDE

DOWNLOAD

FASTA

MSA

ID MAP

PTHWY

GROUP

GENOME

FEATURE

FASTA

PTHWY

GENOME

VIII. Submitting another BLAST job

1. At the top of the BLAST result page, researchers can click on the Edit from and resubmit button to initiate another BLAST job

Services

BLAST

The BLAST Search allows you to search against public or private genomes in PATRIC or other reference databases using a DNA or protein sequence and find matching genomes, genes, RNAs, or proteins.

Edit form and resubmit

2. This will reload the page, showing the original parameters used for the first BLAST job. These can be adjusted, with a second job submitted by clicking the Search button at the bottom of the page.

Services
BLAST

The BLAST Search allows you to search against public or private genomes in PATRIC or other reference databases using a DNA or protein sequence and find matching genomes, genes, RNAs, or proteins.

Sequence

```
>fig|511145.12.peg.3032|b2938|VBIEscCol129921_3032| Biosynthetic arginine decarboxylase (EC 4.1.1.19) [Escherichia coli str. K-12 substr. MG1655 | 511145.12]
MSSQEASKMLRTYNTANWGNYYDVNELGHISVCPDPDVPPEARVDLAQLVKTREAQGQRL
PALFCFPQILQHRRLRSINAAFKRARESYGYNGDYFLVVPPIKVNQHRRVIESLIHSGEPLG
LEAGSKAELMAVLAHAGMTRSVIVCNGYKDREYIRLALIGEKMGHKVVILIEKMSEIATV
LDEAERLNVVPRLGVRARLASQSGKQSGGKSKFGLAATQVLQVETLREAGRLDSL
QLLHFHLSQMANIRDITATGVRESARFYVELHKLGVNIQCFDVGGGLGVDYEGTRSQSDC
SVNYGLNEYANNIIWAIGDACEENGLPHPTVITESGRAVTAHHTVLVSNIIIGVERNEYTV
PTAPAEADAPRALQSMWETWQEMHEPOTRRSLREWLHDSQMDLHDIIHIGYSSGIFSLQERA
WAEQLYLSNCHVEVQKLDLPQNAHRPIIDELQERMAKMYVNFSLFQSMFPAWGIDQLFP
VLPLEGLDQVPERRAVLDDTCDSGDGIDHYIDGDIATTMPPEYDPENPPMLGFFPMVG
AYQEILGNMNLFGDTEAVDVVFVFDGSEVELSDEGDTVADMLQYVQLDPKTLTQFRD
QVKKTDLDAELQQQFLEEFAGLYGYTYLEDE
```

Program

blastp - search protein database using a protein query

Database

Search within selected genomes

ADVANCED OPTIONS

ADD GENOMES TO SEARCH:

Brucella melitensis biovar Abortus 2308

Brucella melitensis biovar Abortus 2308

SEARCH FOR:

Genomic features (genes, proteins or RNAs)

BLAST Parameters

MAX HITS:

50

E VALUE THRESHOLD:

10

Search

References

1. Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezuk, Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic acids research*, **41**, W29-33.
2. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciuffo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, **44**, D733-745.