

Hsa_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Ptr_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Ppy_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Mml_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Mfa_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Mne_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Ssc_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDV	103
Bta_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDV	103
Cfa_	TMEM66	ALTLHHDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTLDI	103
Mmu_	TMEM66	ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCEAYTPRVIQCQNKGWDGYDVQWECKTLDI	104
Rno_	TMEM66	ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTLDI	104
Ocu_	TMEM66	ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTLDV	97
Laf_	TMEM66	ALTLHYNRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVTOCQNKGWDGYDVQWECKTLDI	89
Mdo_	TMEM66	ALTLHYNRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVTOCQNKGWDGFDVQWECKAELDT	119
Gga_	TMEM66	VLTLHGRYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGYDVQWOCKADLEN	94
Xla_	TMEM66	TITLYADRYTNARRSAPVPQLQCVGGSAGCHAMVPQVVQCHNRGWDGLDVQWECKVMDN	93
Xtr_	TMEM66	AITLYADRYTNARRSSPVPQLQCVGGSAGCHAMVPQVVQCHNRGWDGFDVQWECKVMDN	93
Dre_	TMEM66	VLTLYRGRYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTMDN	93
Ssa_	TMEM66	VLTLYKGKYTTARRSSAVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGVDAQWECKTMDN	93
Tru_	TMEM66	VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTMDN	99
Tni_	TMEM66	TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVQWECKTMDN	89
Gac_	TMEM66	ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGMDIQWECRTMDN	92
Ppr_	TMEM66	VLTLYKGRYTTARRSSPVLQLQCAAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN	93
Cel_	TMEM66	AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVCANQGFDGSDVQWRCDADLPH	96
Cre_	TMEM66	AITLNKGKMTTGRRVAPTLQLKCVGG-SAKGAFTPKVVCNSQGFDGSDVQWRCDADLPH	96
Cbr_	TMEM66	AITLHKGKMTTGRRVAPALQLKCVGG-SAKQFSPKVVQCANCQGFDGSDVQWRCDADLPH	96

Sequence alignment and its applications

Dr Opeyemi Lawal

Canadian Research Institute for Food Safety
(CRIFS)

University of Guelph, Ontario
Canada

NCBI Database as a global resource

NCBI is a branch of NIH that created and curates an inexhaustible database and resource that advances science and health by providing access to biomedical and genomic information.

Public databases

General public databases

- NCBI – SRA - **RefSeq, GenBank** - <https://www.ncbi.nlm.nih.gov>
- ENA/EMBL - <https://www.ebi.ac.uk/>
- DDBJ - <https://www.ddbj.nig.ac.jp/index-e.html>
- UNIPROT - <https://www.uniprot.org>
- BIGSdb - <https://pubmlst.org/software/bigsdb>

Specialized databases

- SalFoS – *Salmonella* - <https://salfos.ibis.ulaval.ca>
- ENTEROBASE – Enteric bacteria - <https://enterobase.readthedocs.io/en/latest/>
- PhageDB – Actinophage - <https://phagesdb.org>
- GISAID – Viruses of public health importance - <https://gisaid.org>
- PHASTER – Phages - <https://phaster.ca>
- CARD – Antimicrobial Resistance genes - <http://card.mcmaster.ca>
- ResFinder - Antimicrobial Resistance genes - <https://cge.food.dtu.dk/services/ResFinder/>



Overview of NCBI Database

National Library of Medicine
National Center for Biotechnology Information

All Databases **Search**

NCBI Home

- [Resource List \(A-Z\)](#)
- [All Resources](#)
- [Chemicals & Bioassays](#)
- [Data & Software](#)
- [DNA & RNA](#)
- [Domains & Structures](#)
- [Genes & Expression](#)
- [Genetics & Medicine](#)
- [Genomes & Maps](#)
- [Homology](#)
- [Literature](#)
- [Proteins](#)
- [Sequence Analysis](#)
- [Taxonomy](#)
- [Training & Tutorials](#)
- [Variation](#)

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Popular Resources

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

NCBI News & Blog

New annotations in RefSeq! 14 Feb 2023
In December and January, the NCBI Eukaryotic Genome Annotation Pipeline released twenty-nine new

Announcing New Names for Eukaryotic Genome Annotations in RefSeq! 09 Feb 2023
The RefSeq eukaryotic genome annotation pipeline (EGAP) is moving to

NCBI at ACMG 2023 08 Feb 2023
Join us March 14-18 in Salt Lake City, Utah. We are excited to celebrate ClinVar's 10th anniversary and look

More...

Sequence Alignments

- Sequence alignments is a powerful approach to compare DNA or protein sequences to assess relatedness such as common evolutionary lineage or common structural function
- DNA sequences and the protein sequences they encode evolve by mutation followed by natural selection
- Assess and identify mutations, insertion and deletion of nucleotides or substitution of one amino acid for another or the insertion or deletion of one or multiple adjacent amino acids
- Infer important evolutionary information
- Design degenerate primers

Sequence Alignments

Global Sequence Alignments

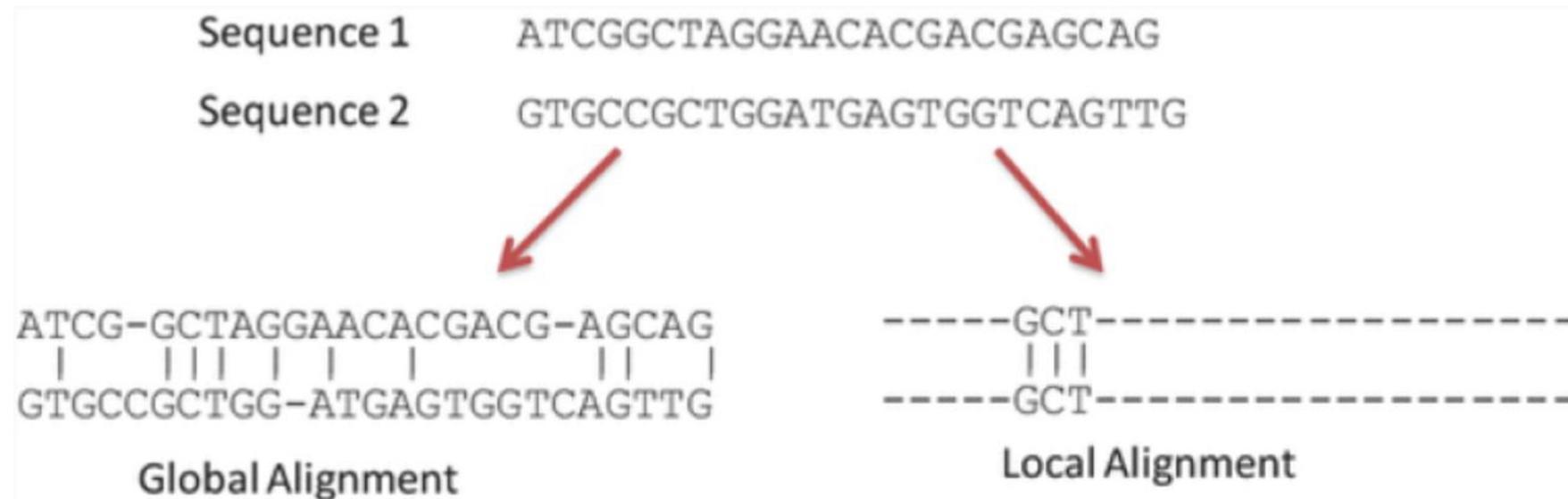
Compare two sequences in their entirety; the gap penalty is assessed regardless of whether gaps are located internally within a sequence, or at the end of one or both sequences

- The Needleman and Wunsch Algorithm

Local Sequence Alignments

Find best matching sequences within the two search sequences.

- The Smith-Waterman Algorithm



Sequence Alignments

Global vs Local Sequence Alignments

Global Sequence Alignment	Local Sequence Alignment
Align end-to-end of the two sequences	Finds local region with the highest level of similarity between the two sequences
Contains all letters from both the query and target sequence	Aligns a fraction of the query sequence to a fraction of the target sequence
Two sequences with approximately the same length are suitable for global alignment	Any two sequences with high level of matches without considering the rest of the sequences regions
Aligns closely related sequences	Suitable for aligning more divergent sequences or distantly related sequences
Usually for comparing homologous genes – like two genes with same function	Used for finding conserved regions or domain in alignment

Sequence Alignment methods

Pairwise Sequence Alignments

Identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid)

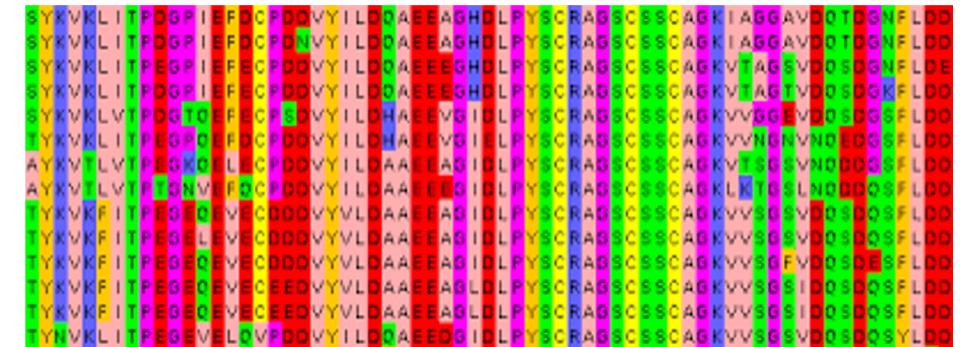
Multiple Sequence Alignments

Alignment of three or more nucleotide or protein sequences of similar length to infer homology and the evolutionary relationships between the sequences.

Pairwise sequence alignment



Multiple Sequence alignment



Sequence Alignments

Pairwise vs Multiple Sequence Alignments

Pairwise Sequence Alignment	Multiple Sequence Alignment
Compares two nucleotides or protein sequences	Compares three or more nucleotides or protein sequences
Can be global or local alignments	Generally global multiple sequence alignment
Simple algorithm is used relative to MSA	Complex algorithm is used
Needleman-Wunsch algorithm and Smith-Waterman algorithm are used	Iterative pairwise alignment method called progressive alignment method is used.
Used for identifying conserved regions between two sequences	Detect regions of variability or conservation region(s) in biological sequences
Used for similarity searches in a database	Phylogenetic analysis, Detection of homology etc
BLAST, LALIGN, EMBOSS (Needle and Water)	MUSCLE, MAFFT, CLUSTALW, T-Coffee

BLAST Algorithm

- **Basic Local Alignment Search Tool** is an algorithm developed by Steven Altschul and Samuel Karlin in 1990
- Designed to find regions of local similarity between biological sequences
- Heuristic and fast, but approximate method of alignment
- The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches
- BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families or homologs



Overview of NCBI BLAST Programs

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

ElasticBLAST 1.0.0 is Now available!
ElasticBLAST version 1.0.0 has support for faster cheaper disks at AWS and better supports Kubernetes on GCP!
Mon, 09 Jan 2023 [More BLAST news...](#)

Protein BLAST

Web BLAST

Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human
Microbes

Mouse
Rat

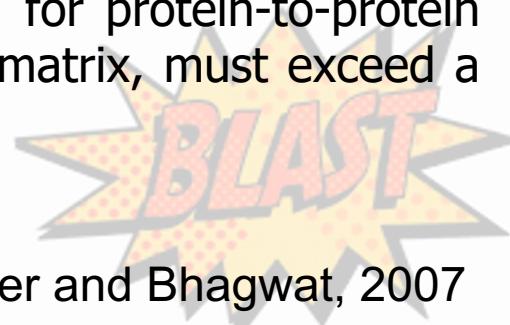


BLAST Programs

Blast programs	*Query sequence	Database
blastn	Nucleotide	Nucleotide
blastp	Protein	Protein
blastx	Nucleotide => Protein	Protein
tblastn	Protein	Nucleotide => Protein
tblastx	Nucleotide => Protein	Nucleotide => Protein

How BLAST programs work

- Blast searches begin with a query sequence that will be matched against sequence databases specified by the user
- Begins by breaking down the query sequence into a series of short overlapping “words”
- Default word size for blastn is 11 nucleotides, while 3 amino acids for blastp
- BLAST scans the database looking for matches between the “words” indexed in the “query” and strings found within the database sequences
- For nucleotide-to-nucleotide searches, these matches must be exact; for protein-to-protein searches, the score of the match as determined using a substitution matrix, must exceed a specified threshold



Other NCBI BLAST programs

Standalone and API BLAST



Download BLAST

Get BLAST databases and executables



Use BLAST API

Call BLAST from your application



Use BLAST in the cloud

Start an instance at a cloud provider

Specialized searches

SmartBLAST



Find proteins highly similar
to your query

Primer-BLAST



Design primers specific to
your PCR template

Global Align



Compare two sequences
across their entire span
(Needleman-Wunsch)

CD-search



Find conserved domains in
your sequence

IgBLAST



Search immunoglobulins
and T cell receptor
sequences

VecScreen



Search sequences for
vector contamination

CDART



Find sequences with
similar conserved domain
architecture

Multiple Alignment



Align sequences using
domain and protein
constraints

MOLE-BLAST



Establish taxonomy for
uncultured or
environmental sequences



Application of BLAST programs

Species identification – Identification of species or finding gene homologs using the nucleotide sequences of known genes or whole-genome sequences. For example identification of unknown species using the nucleotide sequences of the 16S rRNA gene

Detection of domains or regions of interests – Domains and /or regions of interests could be detected or localized in nucleotide and/or protein sequences using blast programs. Eg locating promoter region or motif in a gene

Phylogenetic analyses – Sequence alignments that are the starting point for constructing phylogenies can be generated using blast program

Comparison - Blast-based database comparison to identify genes of interests in WGS – AMR genes, Virulence genes, Plasmids, Phage etc



Practical session

Retrieving data from NCBI database

General public databases

- NCBI – SRA - **RefSeq, GenBank** - <https://www.ncbi.nlm.nih.gov>
- ENA/EMBL - <https://www.ebi.ac.uk/>
- DDBJ - <https://www.ddbj.nig.ac.jp/index-e.html>
- UNIPROT - <https://www.uniprot.org>
- BIGSdb - <https://pubmlst.org/software/bigsdb>

Specialized databases

- SalFoS – *Salmonella* - <https://salfos.ibis.ulaval.ca>
- ENTEROBASE – Enteric bacteria - <https://enterobase.readthedocs.io/en/latest/>
- PhageDB – Actinophage - <https://phagesdb.org>
- GISAID – Viruses of public health importance - <https://gisaid.org>
- PHASTER – Phages - <https://phaster.ca>
- CARD – Antimicrobial Resistance genes - <http://card.mcmaster.ca>
- ResFinder - Antimicrobial Resistance genes - <https://cge.food.dtu.dk/services/ResFinder/>



Retrieving data from NCBI database

<https://www.ncbi.nlm.nih.gov>

National Library of Medicine
National Center for Biotechnology Information

All Databases **Search**

Log in

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

New annotations in RefSeq! 14 Feb 2023
In December and January, the NCBI Eukaryotic Genome Annotation Pipeline released twenty-nine new

Announcing New Names for Eukaryotic Genome Annotations in RefSeq! 09 Feb 2023
The RefSeq eukaryotic genome annotation pipeline (EGAP) is moving to

NCBI at ACMG 2023 08 Feb 2023
Join us March 14-18 in Salt Lake City, Utah. We are excited to celebrate ClinVar's 10th anniversary and look

[More...](#)



Retrieving data from NCBI database

Specific gene

Gene name - *mph(A)*

Gene accession no. - NG_047985.1

Specific bacterial species

Species name - *Escherichia coli* K-12

Accession no. - GCA_000005845.2

SRA accession no. - SRR13921543

How about retrieving sequences of a specific gene from WGS annotation file

-> *mph(A)* from *E. coli* pEC9952-1 - acc no - CP104596.1

<https://www.ncbi.nlm.nih.gov>



Identify gene homologs using NCBI BLAST program

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

BLAST® » blastn suite

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)

From
To

Or, upload file no file selected [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database Standard databases (nr etc.) rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt) [?](#)

Organism Optional

Enter organism name or id--completions will be suggested exclude [Add organism](#)

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude Optional

Models (XM/XP) Uncultured/environmental sample sequences

Sequences from type material

[YouTube](#) Create custom database

Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for

Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

Paste Sequence or Provide Accession number



Overview of BLAST results

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

Job Title NG_047985.1 Escherichia coli mph(A) gene for...

RID ZA91V84V013 Search expires on 02-23 03:06 am [Download All](#)

Program BLASTN [?](#) [Citation](#)

Database nt [See details](#)

Query ID lcl|Query_39855

Description NG_047985.1 Escherichia coli mph(A) gene for Mph(A) f...

Molecule type dna

Query Length 1106

Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name
[+ Add organism](#)

Percent Identity E value Query Coverage

to to to

[Filter](#) [Reset](#)

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download Select columns Show 100 [?](#)

select all 100 sequences selected

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Escherichia coli strain EC9952 plasmid pEC9952-1, complete sequence	Escherichia coli	2043	2303	100%	0.0	100.00%	157264	CP104506.1
Klebsiella pneumoniae TA8711 plasmid pTA8711-2 DNA, complete sequence	Klebsiella pneumoniae	2043	2160	100%	0.0	100.00%	39088	AP027265.1
Klebsiella pneumoniae strain KLP00221 plasmid pKLP00221_2, complete sequence	Klebsiella pneumoniae	2043	2339	100%	0.0	100.00%	242282	OP378664.1
Klebsiella pneumoniae strain KLP00172 plasmid pKLP00172_3, complete sequence	Klebsiella pneumoniae	2043	2218	100%	0.0	100.00%	110030	OP378655.1
Escherichia coli strain EC00668 plasmid pEC00668_2, complete sequence	Escherichia coli	2043	2278	100%	0.0	100.00%	145392	OP378621.1
Escherichia coli strain EC00606 plasmid pEC00606_7, complete sequence	Escherichia coli	2043	2222	100%	0.0	100.00%	18954	OP378612.1
Escherichia coli strain EC00592 plasmid pEC00592_3, complete sequence	Escherichia coli	2043	2455	100%	0.0	100.00%	156672	OP378607.1
Aeromonas salmonicida strain ZAS chromosome, complete genome	Aeromonas salmonic...	2043	2101	100%	0.0	100.00%	4885367	CP110647.1
Vector pEGY EC_13655, complete sequence	Vector pEGY EC_13...	2043	2517	100%	0.0	100.00%	98347	ON707124.1



Overview of BLAST results

Sequences producing significant alignments							Download	Select columns	Show	100	?
							GenBank	Graphics	Distance tree of results	MSA Viewer	
			Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> select all	100 sequences selected										

- **Max Score:** the highest alignment score calculated from the sum of the rewards for matched nucleotides or amino acids and penalties for mismatches and gaps.
- **Tot Score:** the sum of alignment scores of all segments from the same subject sequence.
- **Query Cover:** the percent of the query length that is included in the aligned segments.
- **E Value:** the number of alignments expected by chance with the calculated score or better. The expect value is the default sorting metric; for significant alignments the E value should be very close to zero.
- **Perc Ident:** the highest percent identity for a set of aligned segments to the same subject sequence.



Overview of BLAST results

◀ Edit Search Save Search Search Summary ▾

❓ How to read this report? 🎥 BLAST Help Videos ⏪ Back to Traditional Results Page

Job Title NG_047985.1 Escherichia coli mph(A) gene for...

RID ZA91V84V013 Search expires on 02-23 03:06 am Download All ▾

Program BLASTN ⓘ Citation ▾

Database nt See details ▾

Query ID lcl|Query_39855

Description NG_047985.1 Escherichia coli mph(A) gene for Mph(A) f...

Molecule type dna

Query Length 1106

Other reports Distance tree of results MSA viewer ⓘ

Filter Results

Organism only top 20 will appear exclude
Type common name, binomial, taxid or group name

+ Add organism

Percent Identity E value Query Coverage

Percent Identity E value Query Coverage

Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments

Download Select columns Show 100 ⓘ

select all 100 sequences selected

Description		Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Escherichia coli strain EC9952 plasmid pEC9952-1, complete sequence	Escherichia coli	2043	2393	100%	0.0	100.00%	157264	CP104596.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae TA8711 plasmid pTA8711-2 DNA, complete sequence	Klebsiella pneumoniae	2043	2160	100%	0.0	100.00%	39088	AP027265.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae strain KLP00221 plasmid pKLP00221_2, complete sequence	Klebsiella pneumoniae	2043	2339	100%	0.0	100.00%	242282	OP378664.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae strain KLP00172 plasmid pKLP00172_3, complete sequence	Klebsiella pneumoniae	2043	2218	100%	0.0	100.00%	110030	OP378655.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00668 plasmid pEC00668_2, complete sequence	Escherichia coli	2043	2278	100%	0.0	100.00%	145392	OP378621.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00606 plasmid pEC00606_7, complete sequence	Escherichia coli	2043	2222	100%	0.0	100.00%	18954	OP378612.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00592 plasmid pEC00592_3, complete sequence	Escherichia coli	2043	2455	100%	0.0	100.00%	156672	OP378607.1
<input checked="" type="checkbox"/>	Aeromonas salmonicida strain ZAS chromosome, complete genome	Aeromonas salmonicida	2043	2101	100%	0.0	100.00%	4885367	CP110647.1
<input checked="" type="checkbox"/>	Vector pEGY_EC_13, complete sequence	Vector pEGY_EC_13	2043	2517	100%	0.0	100.00%	98347	ON707124.1



Overview of BLAST results

[Edit Search](#)[Save Search](#)[Search Summary](#)[How to read this report?](#)[BLAST Help Videos](#)[Back to Traditional Results Page](#)

Job Title NG_047985.1 Escherichia coli mph(A) gene for...

RID ZA91V84V013 Search expires on 02-23 03:06 am [Download All](#)Program BLASTN [Citation](#)Database nt [See details](#)

Query ID lcl|Query_39855

Description NG_047985.1 Escherichia coli mph(A) gene for Mph(A) fa...

Molecule type dna

Query Length 1106

Other reports [Distance tree of results](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

 to

E value

 to

Query Coverage

 to [Filter](#)[Reset](#)[Descriptions](#)[Graphic Summary](#)[Alignments](#)[Taxonomy](#)

Sequences producing significant alignments

[Download](#)[Select columns](#)

Show 100

 select all 100 sequences selected[GenBank](#)[Graphics](#)[Distance tree of results](#)[MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Escherichia coli strain EC9952 plasmid pEC9952-1, complete sequence	Escherichia coli	2043	2393	100%	0.0	100.00%	157264	CP104596.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae TA8711 plasmid pTA8711-2 DNA, complete sequence	Klebsiella pneumoniae	2043	2160	100%	0.0	100.00%	39088	AP027265.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae strain KLP00221 plasmid pKLP00221_2, complete sequence	Klebsiella pneumoniae	2043	2339	100%	0.0	100.00%	242282	OP378664.1
<input checked="" type="checkbox"/>	Klebsiella pneumoniae strain KLP00172 plasmid pKLP00172_3, complete sequence	Klebsiella pneumoniae	2043	2218	100%	0.0	100.00%	110030	OP378655.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00668 plasmid pEC00668_2, complete sequence	Escherichia coli	2043	2278	100%	0.0	100.00%	145392	OP378621.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00606 plasmid pEC00606_7, complete sequence	Escherichia coli	2043	2222	100%	0.0	100.00%	18954	OP378612.1
<input checked="" type="checkbox"/>	Escherichia coli strain EC00592 plasmid pEC00592_3, complete sequence	Escherichia coli	2043	2455	100%	0.0	100.00%	156672	OP378607.1
<input checked="" type="checkbox"/>	Aeromonas salmonicida strain ZAS chromosome, complete genome	Aeromonas salmonicida	2043	2101	100%	0.0	100.00%	4885367	CP110647.1
<input checked="" type="checkbox"/>	Vector pEGY_EC_13655, complete sequence	Vector pEGY_EC_13655	2043	2517	100%	0.0	100.00%	98347	ON707124.1



Tasks

1. Extract the nucleotide sequences as well as the corresponding translated amino acid sequences of the following genes from the accession numbers listed below
mecC - CP028165.1
sul2 - MZ465529.1
tuf - CP031196.1
2. Identify and download the multiple sequence alignments of the top 50 homologs for each genes above using $\geq 90\%$ coverage and similarity

Advance level (Optional)

3. Construct the phylogenetic tree using the sequence alignments in no (2) and the gene families and/or clusters

