

## NCBI BLAST Services

### Protein BLAST

**Query:** human MLH1, NP\_000240

**Program:** blastp

**Database:** refseq\_protein

**Goals:**

- Explore DNA mismatch repair proteins in vertebrates using Reference Sequences.
- Demonstrate the usefulness of organism limits, taxonomy report, and the link to multiple sequence alignment.

**Procedure:**

- Retrieve **NP\_000240** from the Entrez protein service.
- Click "Run BLAST" under the Analyze this sequence portlet.

### DNA mismatch repair protein Mlh1 isoform 1 [Homo sapiens]

NCBI Reference Sequence: NP\_000240.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

---

Go to:

LOCUS	NP_000240	756 aa	linear	PRI 23-APR-2016
DEFINITION	DNA mismatch repair protein Mlh1 isoform 1 [Homo sapiens].			
ACCESSION	NP_000240			
VERSION	NP_000240.1 GI:4557757			
DBSOURCE	REFSEQ: accession <a href="#">NM_000249.3</a>			
KEYWORDS	RefSeq.			
SOURCE	Homo sapiens (human)			

Customize view

---

**Analyze this sequence**

[Run BLAST](#)

[Identify Conserved Domains](#)

[Highlight Sequence Features](#)

[Find in this Sequence](#)

- Select **refseq\_protein** as the database.
- Enter "vertebrates" in the Organism input box, select from the suggested list

### Choose Search Set

<b>Database</b>	<input type="text" value="Reference proteins (refseq_protein)"/>
<b>Organism</b> <small>Optional</small>	<input type="text" value="vertebrates (taxid:7742)"/> <input type="checkbox"/> Exclude <input type="button" value="+"/> <p style="font-size: small; margin-top: 2px;">Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ⓘ</p>
<b>Exclude</b> <small>Optional</small>	<input type="checkbox"/> Models (XM/XP) <input type="checkbox"/> Uncultured/environmental sample sequences
<b>Entrez Query</b> <small>Optional</small>	<input type="text"/> <input type="button" value="YouTube"/> <a href="#">Create custom database</a> <p style="font-size: small; margin-top: 2px;">Enter an Entrez query to limit search ⓘ</p>

- Click **Algorithm Parameters** and increase the **Max target sequences** to 1000 and set the Expect threshold to a stringent value, 1e-6.

**Algorithm parameters** Note: Parameter values that differ fr

**General Parameters**

**Max target sequences**    
 Select the maximum number of aligned sequences to display

**Short queries**  Automatically adjust parameters for short input sequences

**Expect threshold**    
 (This field is highlighted in yellow in the original image)

**Word size**

**Max matches in a query range**

- Click **BLAST** to submit the search
- The matches have different types of RefSeq accessions. The XP\_ entries represent proteins derived from gene models. NP\_ accessions represent proteins derived from experimentally supported expressed sequences.

Gene models may be incomplete due to missing data in the genome or may represent potential but unsupported splice variants. You can filter these from the database by using the "Exclude" option.

- From the results page, click **Edit and resubmit**
- Check the Exclude **Models (XM/XP)** checkbox

**Choose Search Set**

**Database**

**Organism**   Exclude    
 Optional  
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

**Exclude**  Models (XM/XP)  Uncultured/environmental sample sequences   
 Optional

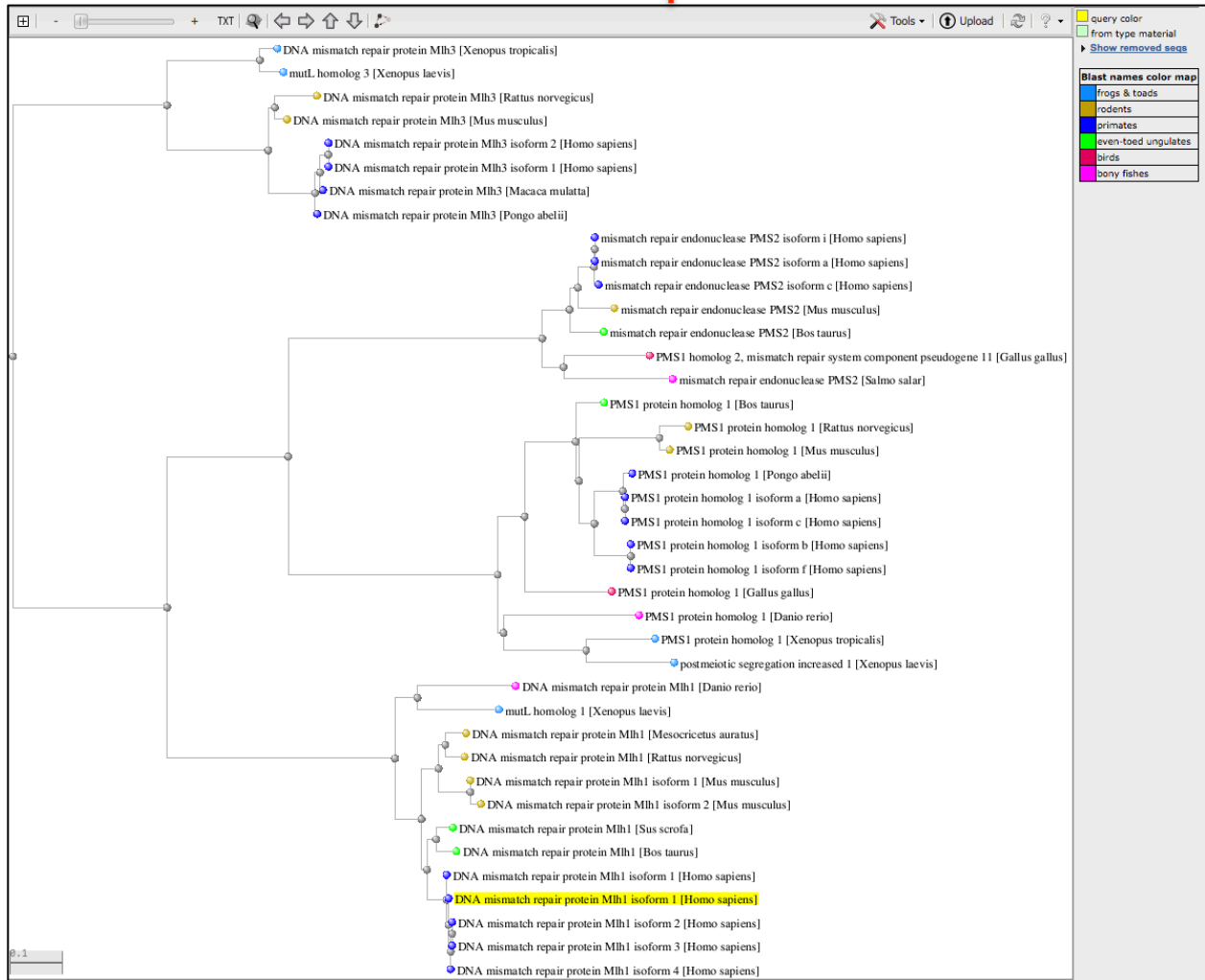
**Entrez Query**    
 Optional  
 Enter an Entrez query to limit search

- Click **BLAST** to submit the search.
- The results now contain only NP\_ style accessions, experimentally supported gene products. Click **Taxonomy report**, to see the DNA mismatch repair proteins products found in humans and other vertebrare.

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#) [Related Structures](#) [Multiple alignment](#)

- Return to the BLAST results, click on the **Distance tree of results** to see a graphic presentation of the relative between the different proteins. The distinct gene products form different clusters

Other reports: [▶ Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Related Structures\]](#) [\[Multiple alignment\]](#)



- You can extend this search to a Multiple Alignment with COBALT to obtain a more accurate tree.

Other reports: [▶ Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#) [\[Multiple alignment\]](#)

### Independent exercise: aromatic amino acid hydroxylases

**Query:** human tyrosine hydroxylase, NP\_954986

**Program:** blastp

**Database:** refseq\_protein

**Goals:**

- Identify members of the aromatic amino acid hydroxylase family in mammals and other groups

- Use taxonomy report, formatting options, TreeView, and links to explore results.
- Use the results to make a multiple alignment with COBALT and a phylogenetic tree from the COBALT alignments

**Procedure:**

You may need to "**Reset page**" first before starting this exercise. Follow the procedure above for creatine kinases.

- Run the search first against mammals Reference Sequences. Set the e-value cutoff to 1e-6 to see only closely related proteins. Compare the results with and without the XP\_ filter.
- Use the TreeView display to examine the relationships in the group of NP\_ proteins.

What are the different members of this group in humans?

What is meant by "isoforms" in the case of tyrosine hydroxylase?

Which one of the mammalian aromatic amino acid hydroxylases is the product of two genes?

Note the e-value of the hit to phenylalanine-4-hydroxylase from *Pongo abelii*

- Click **Edit and resubmit** to get back to the search page.
- Expand **Algorithm parameters** and set the Max target sequences to 5000 and the Expect threshold to 1e-4.
- Remove the Organism limit.
- Remove the XP\_ exclusion.
- Click **BLAST** to resubmit the search.

You can use the formatting options to now filter your results for certain kinds of hits.

- Click **Formatting options**
- Type "fishes" in the Organism box on the **Formatting options** page and click **Reformat**.

What two fishes are represented? Are there additional genes represented in the fish?

- Type "bacteria" in the Organism box on the **Formatting options** page and click **Reformat**.

What domain is missing in these hits that was present in the eukaryotic proteins?

- Type "hypothetical protein" in the Entrez query box on the **Formatting options** page and click **Reformat**.

What organisms are represented in these results?

- Now type *Pongo abelii* in the Organism box on the **Formatting options** page and click **Reformat**.

Why is the e-value different than the one you noted previously for the search against only mammalian NP\_ RefSeqs?

**Additional practice:** Explore BLAST results for prolactin (NP\_000939) for mammals especially compare mouse and rat to human. There is an additional gene family in fishes.

## Nucleotide BLAST and Genomic BLAST

### CDC20 and human genome

**Query:** Macaque CDC20 mRNA, AB168636

**Program:** nucleotide BLAST page with megablast and blastn

**Database:** human genomic + transcript, mouse genomic+transcript

- Map a sequence onto various genomes
- Compare the speed and sensitivity of various algorithms
- Use the different sorting options in BLAST results
- Use formatting options, CDS feature.

#### Procedure:

- Retrieve AB168636 from Entrez nucleotide and follow the link to **Run BLAST**.
- Select the **Human genomic + transcripts** database, click **BLAST**.
- Examine the **Graphic summary** and **Descriptions** sections.

Notice that there are separate sections for the transcripts and genomic regions. There are two genome assemblies represented: the reference genome, GRCh38 and an alternate assembly, CHM1\_1.1, a hydatidiform mole assembly. The latter is useful because it has a single haplotype. There are hits to chromosome 9 and chromosome 1 in the three assemblies. The retro-transposed pseudogene on Chromosome 9 actually ranks higher than the functional gene because the single uninterrupted single hit outscores the individual exon hits for the functional gene. Re-sorting the output by **Total score** and/or **Max Ident** bring match to the functional gene to the top of the list.

- Click on the linked hit to Homo sapiens chromosome 9, GRCh38 Primary Assembly and examine the alignment to the pseudogene.

Notice the single nearly complete alignment with no introns. The poly-A tail from the mRNA is even present in the genome. This is an example of an apparent retro-transcribed mRNA that has been inserted into the genome.

- Click the linked hit for Homo sapiens chromosome 1, GRCh38 Primary Assembly to go to the alignment
- Click **Query Start position** to arrange the matches according to exon order

The first aligned segment starts at position 73 of the mRNA. Megablast misses the first exon hit as well as a match to some related transcripts. Re-running the search with blastn finds this hit. You will need to set the Expect threshold to 1e-6 to avoid additional non-significant matches.

### Linking to the Graphical Sequence Viewer

Displaying the BLAST hits on the annotated chromosome in the Graphical Sequence Viewer provides important genomic context for the aligned regions.

- Follow the main 'Graphics' link at the top of the alignments the hits on chromosome 9 and chromosome 1 to display the hits in the Graphical Sequence Viewer. Make a note of the surrounding genes.

Download ▾ GenBank Graphics Sort by: Query start position

Homo sapiens chromosome 1, GRCh38 Primary Assembly  
 Sequence ID: [ref|NC\\_000001.11|](#) Length: 248956422 Number of Matches: 10

Range 1: 43359166 to 43359397 [GenBank](#) [Graphics](#) ▾ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
396 bits(214)	8e-107	226/232(97%)	0/232(0%)	Plus/Plus
Query 73	GGGCTCCGCAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGTTCCCATGGCACAAT	132		
Sbjct 43359166	GGGCTCCGTAGGCACCAACTGCAAGGACCCCTCCCGCTGCGGGCGCTCCCATGGCACAGT	43359225		
Query 133	TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCAATGCAC	192		
Sbjct 43359226	TCGCGTTCGAGAGTGACCTGCACTCGCTGCTTCAGCTGGATGCACCCATCCCAATGCAC	43359285		
Query 193	CCCCGCGCGCTGGCAGCGCAAAGCCAAGGAAGCCTCAGGCCCGGCCCTCACCCATGC	252		
Sbjct 43359286	CCCCGCGCGCTGGCAGCGCAAAGCCAAGGAAGCCTCAGGCCCGGCCCTCACCCATGC	43359345		
Query 253	GGGCCGCCAACCGATCCACAGCGCCGGCAGAACTCCGGGCCGAACTCCTGG	304		
Sbjct 43359346	GGGCCGCCAACCGATCCACAGCGCCGGCAGAACTCCGGGCCGAACTCCTGG	43359397		

Notice also that the search did not find a match to the first exon of the human gene. This is a consequence of the algorithm choice (megablast). You can compare these results to what you get with discontinuous megablast and blastn. You can also compare these results to what you obtain in a search of the mouse mRNA and transcript database. You will need to use blastn rather than megablast to find the functional gene, which on chromosome 4.

### Formatting Options CDS Feature

- Open **Format options** link, check **CDS Features**, click **Reformat**

This adds the translation to the nucleotide alignment if coding regions are annotated on the query or subject (database sequence).

- Examine the alignment to the human transcript NM\_001255.

The macaque mRNA sequence has a single base deletion relative to the human transcript. This results in a frame shift making the protein translation diverge at the C- terminus. This is most likely a sequencing error as the other mammalian CDC20 proteins agree with the human sequence. You can use a blastx search with AB168636 to demonstrate this frame shift as well.

### Independent exercise: Finding TP53 in the sloth (*Choloepus hoffmani*) assembly

**Query:** Human TP53 transcript variant 1, mRNA, NM\_000546

**Program and Database:** Use the Genome BLAST finder on the BLAST homepage to get the *Choloepus hoffmani* genome BLAST page.

## BLAST Assembled Genomes

Find Genomic BLAST pages:

<input type="checkbox"/> <a href="#">Human</a>	<input type="checkbox"/> <a href="#">Rabbit</a>	<input type="checkbox"/> <a href="#">Zebrafish</a>
<input type="checkbox"/> <a href="#">Mouse</a>	<input type="checkbox"/> <a href="#">Chimp</a>	<input type="checkbox"/> <a href="#">Clawed frog</a>
<input type="checkbox"/> <a href="#">Rat</a>	<input type="checkbox"/> <a href="#">Guinea pig</a>	<input type="checkbox"/> <a href="#">Arabidopsis</a>
<input type="checkbox"/> <a href="#">Cow</a>	<input type="checkbox"/> <a href="#">Fruit fly</a>	<input type="checkbox"/> <a href="#">Rice</a>
<input type="checkbox"/> <a href="#">Pig</a>	<input type="checkbox"/> <a href="#">Honey bee</a>	<input type="checkbox"/> <a href="#">Yeast</a>
<input type="checkbox"/> <a href="#">Dog</a>	<input type="checkbox"/> <a href="#">Chicken</a>	<input type="checkbox"/> <a href="#">Microbes</a>

**BLAST** ® » blastn suite [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

*Choloepus hoffmanni* (Hoffmann's two-fingered sloth) GenBank assembly GCA\_000164785.2 Nucleotide BLAST

blastn tblastn tblastx

Enter Query Sequence BLASTN programs search GenBank assembly GCA\_000164785.2 databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#)  [To](#)

NM\_000546

Or, upload file  No file chosen [Choose a BLAST algorithm](#)

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

Database  [Choose a BLAST algorithm](#)

Program Selection

Optimize for

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

[Choose a BLAST algorithm](#)

Search database genomic/9358/GCA\_000164785.2 using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

**Goals:**

Use megablast and blastn to identify the scaffold that contains the TP53 gene.

**Independent exercise: Identifying Potato ETR2 annotation**

**Query:** Tomato ethylene receptor homolog (ETR2), mRNA CDC20 mRNA, NM\_001247224.1

**Program:** Potato genome BLAST page with megablast and blastn

**Database:** SolTub 3.0 reference assembly top level

- Map a sequence onto the genomes
- Compare the speed and sensitivity of various algorithms
- Use the different sorting options in BLAST results
- Use formatting options, CDS feature.

**Procedure:**

- Retrieve NM\_001247224.1 from Entrez nucleotide and copy to the clipboard.
- Use the genomic database finder on the BLAST homepage to find the potato genome BLAST page.
- Run BLAST
- Find corresponding potato gene.

**Notes**

You can find the tomato transcript very quickly by searching Nucleotide with

ETR2 tomato



Then use the Gene Sensor to link to the transcript sequence.

Nucleotide

[Save search](#) [Advanced](#)

Species [Display Settings:](#)  Summary, 20 per page, Sorted by Default order [Send to:](#)  [Filters:](#) [Manage Filters](#)

Plants (7)

Customize ...

Molecule types

genomic DNA/RNA (5)

See [ETR2 ethylene receptor homolog](#) in the Gene database  
[etr2](#) reference sequences [Transcript \(1\)](#) [Protein \(1\)](#)

**Results by taxon**

Top Organisms [\[Tree\]](#)

Solanum lycopersicum (5)

## Solanum lycopersicum ethylene receptor homolog (ETR2), mRNA

NCBI Reference Sequence: NM\_001247224.1

[FASTA](#) [Graphics](#)

[Go to:](#)

```

LOCUS       NM_001247224                2688 bp    mRNA    linear    PLN 30-NOV-2014
DEFINITION Solanum lycopersicum ethylene receptor homolog (ETR2), mRNA.
ACCESSION  NM_001247224
VERSION    NM_001247224.1  GI:350534669
KEYWORDS   RefSeq.
SOURCE     Solanum lycopersicum (tomato)
  
```

Use the Genome database selector to find the potato genome BLAST page.

### BLAST Assembled Genomes

Find Genomic BLAST pages:

Potato

- potato (taxid:4113)
- potatoes (taxid:4113)
- potato late blight agent (taxid:4787)
- potato late blight fungus (taxid:4787)
- Colorado potato beetle (taxid:7539)
- sweet potato whitefly (taxid:7038)
- sweet potato (taxid:4120)
- potato aphid (taxid:13131)
- black scurf of potato (taxid:107832)
- Guatemalan potato tuber moth (taxid:396680)
- potato pink rot agent (taxid:4788)
- potato pink rot fungus (taxid:4788)
- peach-potato aphid (taxid:13164)
- air-potato (taxid:35874)
- potato yam (taxid:35874)
- Durvillaea potatorum (taxid:91052)
- Chinese-potato (taxid:55575)
- Chaco potato (taxid:4108)
- American potato bean (taxid:185702)
- potato psyllid (taxid:290155)

**GO**

[fish](#)  
[d frog](#)  
[dopsis](#)  
[bes](#)

using a **nucleotide** query  
st, **discontiguous megablast**

**protein** query  
**phi-blast, delta-blast**

**translated nucleotide** query

[blast](#) Search translated nucleotide database using a **protein** query

NCBI/ BLAST/ blastn suite **Solanum tuberosum (potato) Nucleotide BLAST**

blastn blastp blastx tblastn tblastx

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**

NM\_001247224 **From**  **To**

**Or, upload file**  No file chosen [Choose](#)

**Job Title**  [Choose](#)  
 Enter a descriptive title for your BLAST search [Choose](#)

### Choose Search Set

**Database**  14854 sequences [Choose](#)

**Exclude**  Models (XM/XP) [Optional](#)

**Entrez Query**  [Optional](#)  
 Enter an Entrez query to limit search [Choose](#)

### Program Selection

**Optimize for**

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

Choose a BLAST algorithm [Choose](#)

**BLAST** Search database **Genome (SolTub\_3.0 reference assembly top-level) - Solanum tuberosum** using **Megablast (Optimize for highly similar sequences)**

Show results in a new window

**Graphic Summary**

Distribution of 11 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

**Color key for alignment scores**

<40	40-50	50-80	80-200	>=200
-----	-------	-------	--------	-------

Query 1 500 1000 1500 2000 2500

---

**Descriptions**

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> <a href="#">Solanum tuberosum cultivar DM 1-3 516 R44 unplaced genomic scaffold, SolTub_3.0 scf0009</a>	1604	3877	85%	0.0	97%	<a href="#">NW_006239025.1</a>
<input type="checkbox"/> <a href="#">Solanum tuberosum cultivar DM 1-3 516 R44 unplaced genomic scaffold, SolTub_3.0 scf0011</a>	881	1331	47%	0.0	84%	<a href="#">NW_006239040.1</a>
<input type="checkbox"/> <a href="#">Solanum tuberosum cultivar DM 1-3 516 R44 unplaced genomic scaffold, SolTub_3.0 scf0004</a>	353	675	18%	6e-95	90%	<a href="#">NW_006238964.1</a>

Use the Formatting options to highlight differences (dots for identities) and show coding regions (CDS).

**Formatting options** Reformat

Show Alignment as:   Old View [Reset form to defaults](#)

Alignment View:

Display:  Graphical Overview  NCBI-gi  CDS feature

Masking: Character:  Color:

Limit results: Descriptions:  Graphical overview:  Line length:

Organism: Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.  
  Exclude

Entrez query:

Expect Min:  Expect Max:

Percent Identity Min:  Percent Identity Max:



## Microbial Genomes BLAST

**Query:** SRA read from SRR452448, a metagenome from a Gulf of California hydrothermal vent plume (1600 – 2000 m deep)

```
>gn1 | SRA | SRR452448.103762 D5KHBFN1_0131:1:1101:8134:8771
ACCCCTTGTTGGTGCTCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC
AGGCGGCAAACCTTAACGGCTTTCTGCCGCACTGCATTTGGTGGTAAAATGCTTTGATCT
ATCGATGACCACCTGTGGCGAAGGCGGTCTACTAGAACACGTCGGACGGTGAGGGATGAA
AGCTGGGGGAGCAAACCGGA
```

**Program:** nucleotide BLAST page with megablast and blastn

**Database:** microbial genomes, representative, all, complete

**Purpose:** Identify and map unknown microbial sequence

### Procedure:

- Copy/paste the above sequences into the microbial genomes BLAST page.
- Select the **Representative genomes database**, click **BLAST**.
- Examine the **Graphic summary** and **Descriptions** sections.
- Investigate how changing the database and the BLAST program affects the results.

Examine the Descriptions section to find the best matching bacterial or archaeal genome. In the alignments section you can see that nearby genes are identified. For the hydrothermal vent plume these 16s regions. You can see your BLAST hits in genomic context by clicking the graphics link for all matches.

Download [GenBank](#) [Graphics](#) Sort by: E value

Nitrosopumilus maritimus SCM1 chromosome, complete genome  
Sequence ID: [ref|NC\\_010085.1](#) Length: 1645259 Number of Matches: 2

Range 1: 896830 to 896927 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
168 bits(186)	1e-39	96/98(98%)	0/98(0%)	Plus/Plus

Features: [rRNA-16S ribosomal RNA](#)

```
Query 1 ACCCCTTGTTGGTGCTCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC 60
      |||
Sbjct 896830 ACCCCTTGTTGGTGCTCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC 896889

Query 61 AGGCGGCAAACCTTAACGGCTTTCTGCCGCACTGCATT 98
      |||
Sbjct 896890 AGGCGGCAAACCTTAACGGCTTTCTGCCGCACTGCATT 896927
```

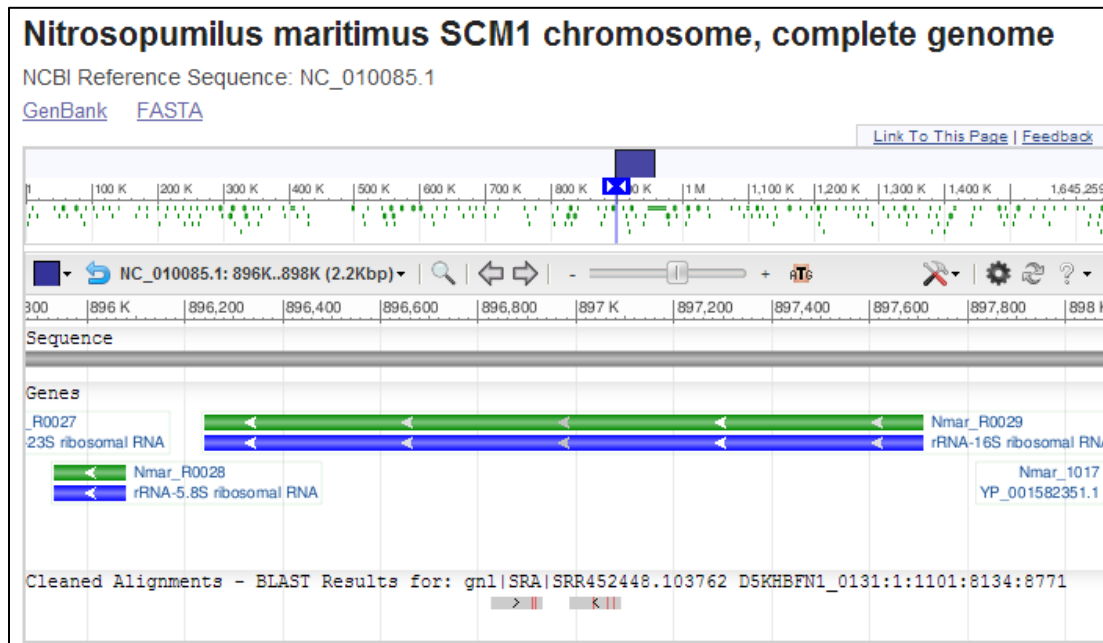
Range 2: 896991 to 897088 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Identities	Gaps	Strand
163 bits(180)	4e-38	95/98(97%)	0/98(0%)	Plus/Minus

Features: [rRNA-16S ribosomal RNA](#)

```
Query 103 GGTAAAATGCTTTGATCTATCGATGACCACCTGTGGCGAAGGCGGTCTACTAGAACACGT 162
      |||
Sbjct 897088 GGTAAAATCCTTTGATCTATTGATGACCACCTGTGGCGAAGGCGGTCTACCAGAACACGT 897029

Query 163 CCGACGGTGAGGGATGAAAGCTGGGGGAGCAAACCGGA 200
      |||
Sbjct 897028 CCGACGGTGAGGGATGAAAGCTGGGGGAGCAAACCGGA 896991
```



Your query will split because the SRA runs in this case are paired reads. For instance the first sequence above is the following two reads.

```
>gnl|SRA|SRR452448.103762.1 D5KHBFN1_0131:1:1101:8134:8771 (Biological)
ACCCCTTGTGGTGCTCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC
AGGCGGCAAACCTAACGGCTTTCCCTGCCGCACTGCATTTG
>gnl|SRA|SRR452448.103762.2 D5KHBFN1_0131:1:1101:8134:8771 (Biological)
GTGGTAAAATGCTTTGATCTATCGATGACCACCTGTGGCGAAGGCGGTCTACTAGAACAC
GTCCGACGGTGAGGGATGAAAGCTGGGGGAGCAAACCGGA
```

These hit the Nitrosopumilus 16S gene as two alignments as shown in the graphic above.

For bacteria there are sometimes multiple widely separated hits per genome because the 16S genes are in multiple copies.

### Independent exercise: Identifying Function and Organism for SRA reads

Use Microbial Genomes BLAST to identify the best match for this read from Arsenic contaminated marine sediment. You will need to adjust the algorithm to find meaningful hits

```
>gnl|SRA|ERR149036.350421.2 GBSQMEQ01CJRM Forward (Biological)
ACACAACCTGTAGGTTTAGGAAGACCGGCCATCTTACAAGCTCCCTTTCCGGGTCCCTGAC
GGAAAAAGCTCATAAATGTGTTTCAGTTTGAACCGGTGACTTTTAAAAGTACCCGAACC
ATCGGTGCGATACCGTTTTTCTCATAATAATCACGAAGTACATCGATGACTTTCTGGTGT
TCGTCATTGAGCTTTCAATGCCTTCTTCCCTTTTACCCACTGAACCCATTCCCTGGCTC
CAGTTGTCAATGAGTCGATGAAACCGTCTTCGTCGACCGTGAAGTTTTTCCCATAAAT
TCTACACTTGGCATTAAATCGTCCCTCCTTTGGTGTATATATTTTCTTTCCGGCTCGCGG
CCTGTAATCTACATTTTTAAAAGACTAGTTCTATCGCAATGAAAAAATTAATGTCAACAT
CAAAAATCAACTGAGACTGCCAAGGCACACAGGGGATAGGN
```

The read from the contaminated sediment corresponds to a gene involved in anaerobic respiration. What is this gene? In what process is it involved?

You can also try the following 16S paired read from the hydrothermal vent plume to identify the source organism.

```
>gnIIISRAISRR452448.103068966 D5KHBFN1_0131:1:2208:1519:200066
GGTGAGTAATGCTTGGGAACCTTGCTTTGCGAGGGGGATAACAGTTGGAAACGACTGCTA
ATACCGCATAACGTCTACGGACCAAACGGGGCTTAGGCTCATATTCCCCACTGCTGCCTC
CCGTAGGAGTCTGGACCGTGTCTCAGTTCCAGTGTGGCTGATCTTCCTCTCAGTACAGCT
AGAGATCGTTGCCTTGGTAA
```

Notice that BLAST finds multiple copies of the rRNA gene cassette in the genome. Would you expect this bacterial species to be in the deep ocean based on the information on the sequence record and linked publication?

Investigate how changing the BLAST program and the nature of the database changes your results.

## SRA BLAST

We can look at gene expression in the melanoma cell lines reported in BioProject ([PRJNA152041](#)). The paper associated with the GEO series shows differential expression of a number of genes including ([CXCL8](#)) in the more metastatic melanoma cells.

### Query:

Human RefSeq RNA for CXCL8, NM\_000584



**Program:** nucleotide BLAST page with megablast and blastn

**Database:** melanoma cell line RNA Seq data

SRX119449 GSM873648: HEMn human normal melanocyte RNA-Seq

SRX119450 GSM873649: A375 human primary melanoma RNA-Seq

SRX119451 GSM873650: A2058 human metastatic melanoma RNA-Seq

**Settings:** Max target seqs 20000, Expect threshold 1e-16

**Purpose:** Find reads for target sequences in the SRA data. Notice the increased expression of CXCL8 in the metastatic cell line.

**Procedure:**

- Copy/paste the above sequences into the SRA BLAST page.
- Select the appropriate SRA experiment, click **BLAST**.
- Examine the **Graphic summary** and **Descriptions** sections.

Investigate how changing the database and the BLAST program affects the results

### Independent exercise: Depth profile of ammonia oxidation

You can repeat the above exercise using other RefSeq mRNAs for genes of interest: [HIF1A](#), [THSB1](#).

## Align two or more Sequences, Global alignment, and Multiple-alignment

Align 2 sequences

**Query 1:** Human Albumin, NP\_000468

**Query 2:** Human GC, NP\_000574

**Program:** blastp

**Procedure:**

- Retrieve NP\_000468 from the Entrez protein system.
- Follow the link to **Run BLAST** from the **Analyze this sequence** portlet on the protein record.
- Check the box that reads **Align 2 or more sequences**.
- Enter NP\_000574 in the subject sequence box.
- Click BLAST
- Expand and examine the **Dot Matrix View**

Off-diagonal elements show that more than one local alignment is found between these two sequences with a repeated domain structure.

Needleman-Wunsch Global Sequence Alignment

**Query 1:** Human Albumin, NP\_000468

**Query 2:** Human GC, NP\_000574

**Program:** Protein

**Procedure:**

- Click on the **Global Sequence Alignment Tool** link in the **Specialized BLAST** section of the BLAST homepage.
- Click the **Protein** tab over the Query sequence text area.
- Click the **Align** button

The tool finds a single global alignment between the two sequences.

Align more than two sequences (BLAST) and extend to a multiple-alignment

**Query 1:** Human Albumin, NP\_000468

**Query 2:** Human AFP, Human AFM, Human GC proteins

NP\_001125

NP\_001124

NP\_000574

Enter these one per line.

**Procedure:**

- Retrieve NP\_000468 from the Entrez protein system.
- Follow the link to **Run BLAST** from the **Analyze this sequence** portlet on the protein record.
- Check the box that reads **Align 2 or more sequences**.
- Enter NP\_000574, NP\_001125, NP\_001124, one accession per line, in the subject sequence box.
- Click BLAST
- From the results click the Multiple Alignment link
- Generate the Phylogenetic Tree from the COBALT results.

**Explanatory Notes:**

The "Align 2 (or more) sequences" service is now combined with Basic BLAST. Checking the "Align two or more sequences" on the BLAST form will transform the BLAST form to allow direct comparison of two input sequences. This service produces only local alignments since this is BLAST. In cases such as the albumin family used here -- where there is a set of repeated domains, more than one alignment is found. This is easily seen in the dot matrix graphic of the alignments found between albumin and the vitamin D binding protein. The new Needleman-Wunsch alignment tool allows a global comparison of albumin and the vitamin D binding protein and produces the single best alignment that includes all residues.

Entering more than two sequences in the search boxes allows a search against a small custom database. In this case comparing the albumin sequence to the other three members of the family produces pairwise local alignments equivalent to a small database search. As before there are more than one local alignment reported for some sequences. The new COBALT extension to BLAST linked through "Other reports" produces a true global multiple alignment of the four proteins. The Download link at the top of the COBALT output allows export of the alignment for local editing. The Phylogenetic Tree link produces a more accurate distance tree of the albumin protein family than could be obtained from the BLAST

alignments. COBALT is available as an extension on all protein BLAST results. A direct interface to COBALT is linked to the “Specialized BLAST” section of the BLAST homepage.

### Independent practice: align two or more Sequences, Global alignment, and Multiple alignment

Perform Align 2 Sequences and a global alignment with Human spectrin alpha chain, brain isoform 3, NP\_001182461 and *Drosophila* beta spectrin, NP\_523388.

Perform a multiple alignment directly from a set of protein results.

- Retrieve 2353 from the HomoloGene database.
- Click on the **Links** menu and follow the link to Protein
- Click on **Align sequences with COBALT** in the **Analyze these sequences** portlet.
- Click the Align button in COBALT
- Remove (uncheck) any aberrant XP\_ sequences and Re-align them.
- Generate the Phylogenetic Tree from the final alignment.

## Primer BLAST

### Designing primers specific to an exon of a gene

**Query:** Human BRCA1 exon 15 plus flanks (NG\_005905.2 from 146746 to 147056).

**Organism limit:** human

**Database:** Reference Genome from selected organisms

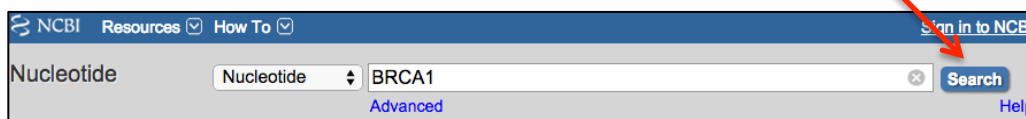
**Avoid known SNPs:** On and off

- Use the gene sensor to retrieve the RefSeq Gene the Entrez nucleotide system.
- Find exon 15 using the Highlight Sequence Features tool on the nucleotide record/
- Display exon 15
- Follow the link to **Pick Primers** from the **Analyze this sequence** portlet on the subsequence.
- Select **Use new graphic view** at the bottom of the form to see results in the graphical sequence viewer.
- Run the search with the default settings.

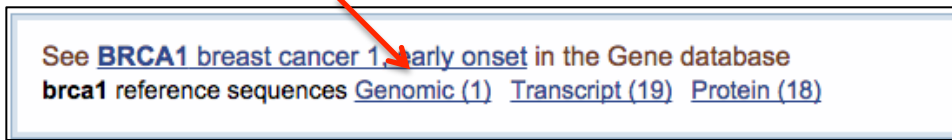
### Notes

You can use the Gene Sensor to quickly find primers to amplify an exon from BRCA1

1. Search BRCA1 in the [NCBI Nucleotide system](#).



- Follow the Genomic link in the Gene Sensor box at the top of the Nucleotide results to retrieve the RefSeqGene record ([NG\\_005905](#)) for the BRCA1 gene.



- Click the “Highlight Sequence Features” in the right-hand column of the sequence record to activate feature highlighting. You will see the coding sequence (CDS) feature of the gene highlighted.

**Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG\_292) on chromosome 17** Customize view

NCBI Reference Sequence: NG\_005905.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NG\_005905 193689 bp DNA linear PRI 08-NOV-2014

DEFINITION Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG\_292) on chromosome 17.

**Analyze this sequence**

Run BLAST

Pick Primers

**Highlight Sequence Features**

Find in this Sequence

---

```

1261 ccttggtagc taactcgttaa aacatctgtc aaccaaatgg gcaacagggtt agccccaact
1321 ccagctgccc ctttccaact acaacaacac caacaaccag tgggcaaggc tggccaggga
1381 gaggggagtc acctgtttcac tggccaaggc gccaatgtgt aggagaaggt tgttggaaac
1441 ttgcccacc acgagggaca ggtgcagaac ctogagggtc agctgtgtca ccataatggg
1501 ttagtagttg ccattggaga tgtttaagat attctgggga tgaagacaag ggaagggagg
1561 agtgaagatg gccatccctt aggaaaagcc aaacagttgc aaaaagctgg ctgtctggcc
1621 tgaagcccat ctaccatgac aagottcagg tgcagaggat gaaaactggg agccctctgg
1681 aggggaattg aatggtgag gttcaacttc acgaaatcaa gctgctcag gatccaggtc
1741 catctgtgac agctctctc catctagtc tgcacatcca gagagttcca gcaactgtca
1801 ccccaactca cccctgtggt taactatct ggcocggatt ctgacocctt actgtgtaga
1861 gcccaactcc taacatctg tcaatgagct catgtttgct cactcttctc ttaaggtccc
1921 caacaacctc tgggtgcat aagatctctg gottcccatg cccagggcca taacaactga
1981 caaaggttac agcgaagtat tgcataagg gggggtttg agccaaggcc acccaaccac
2041 ggtatggtt agtagatat caagctcact aaaggccact gtaggaggtg tgaagcctgg
2101 aggtcgaca atgacggacc ggggaacag gaaaagacc atgaaggagg aggtcaccag
2161 gcaagtaggc acgcccagga acacaagag cttcctgtgg gaagcagga cccagatggt
    
```

complement(join(181..301,670..723,1242..1285,1393..1533,2041..2194,4091..4154,4730..4792,4937..4940))

/gene="TMEM106A"

/note="isoform b is encoded by transcript variant 3"

/codon\_start=1

/product="transmembrane protein 106A isoform b"

/protein\_id="NP\_001278516.1"

/db\_xref="GI:615276252"

/db\_xref="CCDS:CCDS74073.1"

/db\_xref="GeneID:113277"

/db\_xref="HGNC:HGNC:28288"

/translation="MGTADASFVCTCGSGKIPQELKQLVALIPYDQRLPKPHTKLPVFLAVLICLVTSSTFVFFLFRSIVVOPAGLNSSTVAFDEADYININILNISMGNYPIMVQLTLEVLHLSLVVQVSNLLHIGFLASEQMFYAVATKIRIDENTYIICTWLEIKVHVVLLHIQTLTCSYLSRSQVLFQSYEYVDCRGNASVPHQLTFHPF"

CDS Feature 1 of 10 NG\_005905 : 8 segments (minus strand) Details Display: FASTA GenBank Help

- Change the “Feature” pull-down list at the bottom left of the sequence display from “CDS” to “exon” and then navigate to exon 15.

```

146641 tctcttaacc taactttatt ggtcttttta attcttaaca gagaccagaa ctttgaatt
146701 caacattcat cgttgtgtaa attaaacttc tccattctct ttcagggga accccttaac
146761 tggaaatctg aatcagcctc ttctctgatg accctgaato tgatcctctt gaagacagag
146821 cccagagtc agctgtgttt ggcaacatac catctcaac ctotgcatg aagttccccc
146881 aattgaaagt tgcagaatct gccacagtc cagctctgctc toataactact gatactgctg
146941 ggtataatgc aatggaagaa agtgtgagca gggagaagcc aqaattgaca gottcaacag
147001 aaagggtcaa caaaagaatg tcaatggtg tctctgctct gaccocagaa
147061 gttgatccat atgtatctcc ctaatgaact agacttaaca acattctggt
147121 tgtaggattt gtoaatatc aacctagagg aagaatccta gaaaacaat
147181 gtaatttaat ttogatctact aatttctgaa aatttagatc tagataaag
147241 attattttat gtatatttac ttgagaaaat aattattaaa tattagtgg
147301 tttgggtatg atataggact ttcgaattgg aattttcctt tctatctgt
147361 ggtatagttt tattcccag aaggcaatctt ttctccccc ttgtctcaat
taccacata ttttaactaa ttcaactca ttccaaatct actccaact
    
```

146746..147056

/gene="BRCA1"

/gene\_synonym="BRCA1; BRCC1; BROVCA1; IRIS; PNCA4; PPP1R53; PSCP; RNF53"

/inference="alignment\_Splign:1.39.8"

/number=15

Feature 15 of 23 NG\_005905 : 1 segment Details Display: FASTA GenBank Help

CDS

**exon**

gene

mRNA

ncRNA

STS

- Follow the FASTA link to display the highlighted exon as a separate view. Then follow the link in the right-hand column of the sequence display to “Pick Primers.”

6. Edit the primer ranges in Primer BLAST so that the forward and reverse primers will bind upstream and downstream of the exon. For example, set the forward primer range from 146646 to 146746 and the reverse primer from 147056 to 147156. This will provide sufficient upstream and downstream sequence for Primer-BLAST to find acceptable binding sites.

7. We want these primers to amplify only the target region from the human genome sequence. Set the database for Primer-BLAST to perform a specificity check to “Genome (reference assembly from selected organisms)” and leave the Organism limit set to human.

8. Run the search with these settings by clicking the “Get Primers” button. An intermediate page appears that identifies a match to the chromosome 17 sequence (NC\_000017.11). Check the box next to the accession to confirm that this is an allowed target and click the “Submit” button.

**Input PCR template** [NG\\_005905.2](#) Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG\_292) on chromosome 17  
**Range** 146646 - 147156

Your PCR template is highly similar to the following sequence(s) from the search database. To increase the chance of finding specific primers, please review the list below and select all sequences (within the given sequence ranges) that are intended or allowed targets.

Select: [All](#) [None](#) Selected: 1

Accession	Title	Identity	Alignment length	Seq. start	Seq. stop	Gene
<input checked="" type="checkbox"/> <a href="#">NG_000017.11</a>	Homo sapiens chromosome 17, GRCh38 Primary Assembly	100%	511	43070828	43071338	<a href="#">BRCA1</a>

Show results in a new window

**Input PCR template** [NG\\_005905.2](#) Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG\_292) on chromosome 17  
**Range** 146646 - 147156  
**Specificity of primers** Primer pairs are specific to input template as no other targets were found in selected database: Nucleotide collection (nt) (Organism limited to Homo sapiens)  
[Other reports](#) [> Search Summary](#)

**Graphical view of primer pairs**

**Primer pair 1**

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
<b>Forward primer</b>	AAACTTCTCCATTCTTCAGA	Plus	23	146724	146746	57.68	39.13	3.00	3.00
<b>Reverse primer</b>	ACTCTTCCAGAATGTTGTTAAGTC	Minus	25	147116	147092	57.50	36.00	4.00	1.00
<b>Product length</b>	393								

**Products on intended target**  
[>NG\\_005905.2](#) Homo sapiens breast cancer 1, early onset (BRCA1), RefSeqGene (LRG\_292) on chromosome 17

```

product length = 393
Forward primer 1   AAACTTCTCCATTCTTCAGA 23
Template          146724  ..... 146746

Reverse primer 1   ACTCTTCCAGAATGTTGTTAAGTC 25
Template          147116  ..... 147092
    
```

**Independent practice: amplifying exon 2 of MLH1**

Repeat the above procedure to find primers to amplify the second exon of MLH1.

## Independent practice: mapping primers onto a genome and a gene using primer BLAST

### Primers:

Forward 5'-AATGGATGATTTGATGCTGTCCC-3'

Reverse 5'-CGTGCAAGTCACAGACTTGGC-3'.

**Database:** Reference Genome for human, chimp. Other mammalian representative genomes

- Use Primer-BLAST to map the primer onto the human reference genome and identify the gene they amplify.
- How large is the amplified product?
- Are there any expected non-specific products?
- Use the graphical display of the amplified region to determine
  - If this a coding or a non-coding region
  - If there are any pathogenic variants in the amplified region
- Explore other reference and representative genomes to see if the primers will work in other species For instance, will they work in mouse? Chimp? Other primates?

## Using MOLE-BLAST to cluster targeted sequences

**Query:** 16S sequences from wastewater metagenome

**Database:** 16S reference sequences, nr

- Retrieve PopSet: 440337304 ([www.ncbi.nlm.nih.gov/popset/440337304](http://www.ncbi.nlm.nih.gov/popset/440337304))
- Follow the link to nucleotide
- Copy the first 30 accessions to cluster in MOLE-BLAST
- Cluster with the 16S reference sequences, and nr

### Notes

Follow the link to nucleotide from the PopSet record.

PopSet PopSet Search

[Limits](#) [Advanced](#) [Help](#)

---

**Display Settings:**  PopSet **Send to:**

**Mixed organisms 16S ribosomal RNA gene, partial sequence.**

PopSet: 440337304  
[GenBank](#) [FASTA](#)

---

**Go to:**

**Study Details**

**Microbial community structure and diversity in an integrated system of anaerobic-aerobic reactors and a constructed wetland for the treatment of tannery wastewater in modjo, ethiopia.**  
 Desta,A.F., Assefa,F., Leta,S., Stomeo,F., Wamalwa,M., Njahira,M. and Appolinaire,D.  
 (12-26-2014) PLoS ONE 9:(12)E115576  
 PMID: 25541981 [Citation](#) [Full text](#)

---

**Go to:**

**Sequences in this data set**

[KC110593.1](#) Uncultured bacterium clone CW3\_V2\_4A 16S ribosomal RNA gene, partial sequence  
[KC110592.1](#) Uncultured cyanobacterium clone CW3\_V2\_3G 16S ribosomal RNA gene, partial sequence  
[KC110591.1](#) Uncultured Providencia sp. clone CW3\_9A 16S ribosomal RNA gene, partial sequence  
[KC110590.1](#) Uncultured Nitrospirae bacterium clone CW3\_7H 16S ribosomal RNA gene, partial sequence

**Article reporting this data set**

Microbial community structure and diversity in an integrati [PLoS One. 2014]

---

**Related information**

Free in PMC

Nucleotide

PubMed

Taxonomy

---

**Recent activity**

Use the Display settings to get and accession list of the first 50 records and copy these to your clipboard.

Nucleotide Nucleotide Search

[Advanced](#) [Help](#)

---

**Species**  
 Bacteria (434)  
 Customize ...

**Molecule types**  
 genomic DNA/RNA (437)  
 Customize ...

**Source databases**  
 GenBank (437)  
 Customize ...

**Sequence length**  
 Custom range...

**Release date**  
 Custom range...

**Revision date**  
 Custom range...

[Clear all](#)  
[Show additional filters](#)

**Find related data**

Database:

---

**Recent activity**

**Filters: [Manage Filters](#)**

**Display Settings:**  Summary, 20 per page, Sorted by Default order **Send to:**

Format	Items per page	Sort by
<input type="radio"/> Summary	<input type="radio"/> 5	<input checked="" type="radio"/> Default order
<input type="radio"/> GenBank	<input type="radio"/> 10	<input type="radio"/> Accession
<input type="radio"/> GenBank (full)	<input type="radio"/> 20	<input type="radio"/> Date Modified
<input type="radio"/> FASTA	<input checked="" type="radio"/> 50	<input type="radio"/> Date Released
<input type="radio"/> FASTA (text)	<input type="radio"/> 100	<input type="radio"/> Organism Name
<input type="radio"/> ASN.1	<input type="radio"/> 200	<input type="radio"/> Taxonomy ID
<input type="radio"/> Revision History		
<input checked="" type="radio"/> Accession List		
<input type="radio"/> GI List		

1. **806 bp linear DNA**  
 Accession: KC110592.1 GI: 440337739  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

2.  [Uncultured Providencia sp. clone CW3\\_9A 16S ribosomal RNA gene, partial sequence](#)

3. **808 bp linear DNA**  
 Accession: KC110591.1 GI: 440337738  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

4.  [Uncultured Nitrospirae bacterium clone CW3\\_7H 16S ribosomal RNA gene, partial sequence](#)

5. **808 bp linear DNA**  
 Accession: KC110590.1 GI: 440337737  
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)



KC110593.1  
KC110592.1  
KC110591.1  
KC110590.1  
KC110589.1  
KC110588.1  
KC110587.1  
KC110586.1  
KC110585.1  
KC110584.1  
KC110583.1  
KC110582.1  
KC110581.1  
KC110580.1  
KC110579.1  
KC110578.1  
KC110577.1  
KC110576.1  
KC110575.1  
KC110574.1  
KC110573.1  
KC110572.1  
KC110571.1  
KC110570.1  
KC110569.1  
KC110568.1  
KC110567.1  
KC110566.1  
KC110565.1  
KC110564.1  
KC110563.1  
KC110562.1  
KC110561.1  
KC110560.1  
KC110559.1  
KC110558.1  
KC110557.1  
KC110556.1  
KC110555.1  
KC110554.1  
KC110553.1  
KC110552.1  
KC110551.1  
KC110550.1  
KC110549.1  
KC110548.1  
KC110547.1  
KC110546.1  
KC110545.1  
KC110544.1

[Link to MOLE-BLAST from the "Specialized BLAST" section of the BLAST homepage.](#)

## Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Cluster multiple sequences together with their database neighbors using [MOLE-BLAST](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins and T cell receptor sequences](#) (IgBLAST)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two (or more) sequences using BLAST (bl2seq)
- Search [protein](#) or [nucleotide](#) targets in PubChem BioAssay
- Search [SRA by experiment](#)
- Constraint Based Protein [Multiple Alignment Tool](#)
- Needleman-Wunsch [Global Sequence Alignment Tool](#)
- Search [RefSeqGene](#)
- Search [trace archives](#)
- Search bacterial and fungal rRNA sequences with [Targeted Loci BLAST](#)

Paste the query accessions in the form and set the database to 16S rRNA. It's helpful to click the "Show results in new window" box so you can adjust settings and resubmit your search.

**MOLE-BLAST** Neighbor Search Tool

Home Recent Results Help My NCBI Welcome cooperps. [Sign Out]

Nucleotide

MOLE-BLAST searches for closest neighbors...

Enter Query Sequences Reset page

Enter nucleotide accessions, gis, or FASTA sequences (up to 300 input sequences with up to 5000 bases each) Clear

KC110549.1  
KC110548.1  
KC110547.1  
KC110546.1  
KC110545.1  
KC110544.1

Or, upload FASTA file Choose File No file chosen

Job Title

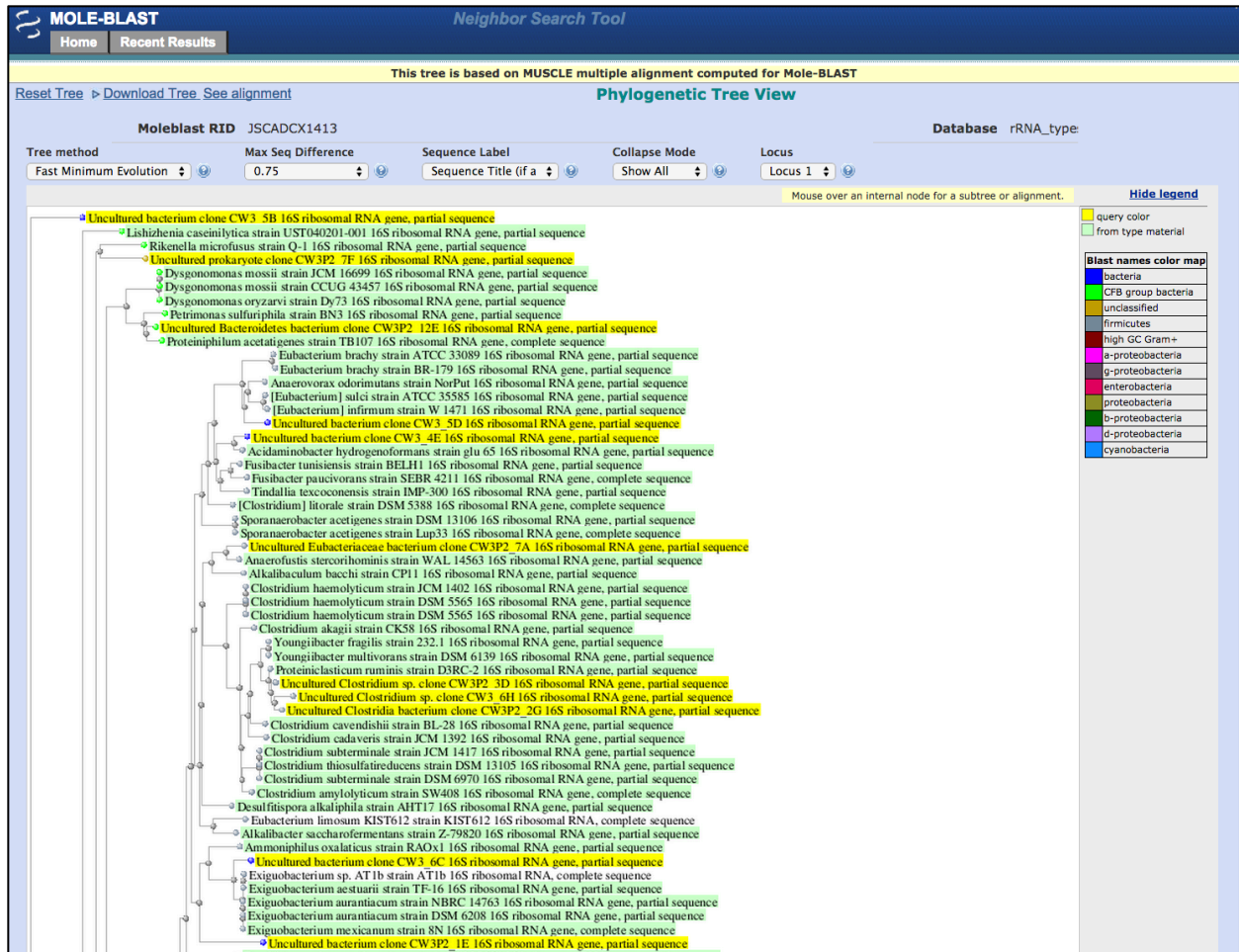
Choose Search Set

Database 16S ribosomal RNA sequences (Bacteria and Archaea)

**Align**  Show results in a new window

[Advanced parameters](#)

The MOLE-BLAST output provides a distribution of the bacterial classes in the metagenome. In some cases you can assign a likely genus. Notice that MOLE BLAST separates the clusters into different "loci" in this case this is based on the quality of the alignments. In certain cases this would separate different genes if present in the data.



### Choose Search Set

Database: Nucleotide collection (nr/nt)

**Align** |  Show results in a new window

**Advanced parameters**

#### Database Search Parameters

Blast program:  Megablast  Blastn

Max target sequences: 10

Exclude:  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to:  Sequences from type material  Sequences with a binomial name

Entrez Query:

#### Query Clustering Parameters

Cluster queries:  Group query sequences by loci

Percent identity: 40

Percent sequence coverage: 75

Cluster merging threshold: 50%

#### Multiple Alignment Parameters

Number of database sequences: 5

**Align** |  Show results in a new window

### Independent practice: MOLE-BLAST

You can try the above search again against nr. Set the exclude uncultured box for cleaner results.

Try MOLE-BLAST against fungal ITS reference sequences using a fungal leaf litter ITS PopSet: 298572961 ([www.ncbi.nlm.nih.gov/popset/298572961](http://www.ncbi.nlm.nih.gov/popset/298572961)).

## Using SmartBLAST to identify unknown sequences.

**Query:** ORF from striped bass TSA record GBAA01198466

```

.....
>lcl|Sequence 1 ORF:2178..6716 Frame -3
MQKSPVEDANFFSKYFFWWASPLLRKGFTKKLELSDVYKAPSFDLADNLSERLEREWDREVVSAKNQPRL
MRALARCFIGPFAFFGVLLYLGEASKTVQPQLLGRIGSFDPPHAPERSQGYFLALGLCLLFTARFLLLO
PAIFGLHHLGMQIRIALFSLIYKKTCLKSSRVLDKISTGQLVSLMSAHLNKLDESGLAHFVWITPLQCI
LCVGLIWEELIEVNGFCALAAALTLGIIQAWLSQKMGPHRVKRAKMINRRALALTSEIVENIHSVKAAGWED
VMETIIKNIRQDEMTLTKRIGSLRYFYASASYFFSAILVIVSAIVPHALSKGIILRRIFFTASYCMVLRMT
LTRQLPGSIQMWDYTLALVKKIEEFLMKEEYRVLEYNLTTTEVELVNVSAWDEGIGELFEKIKQENKAN
GHLNGDAGLFFTNLYITPVLKNIISLYLEKGMKMLAVAGSTGSGKSSLLMMILGELVPSEGKIRHSGRISFS
PQTSWIIPGTIRDNIIFGLTYDEYRYTSVIKACQLEEDFALLPEKDKTHLMEGGVTLGGQORARLGLARA
VYKDADLYLLDAPFTHLDIVTEKEIFEKCVCKLMASKTRIVVTSKLEHLKRADKILLHNGDCYFYGTFS
ELQAKRPFSSLLGLEAYDNINAERRSSILTETLRRVSIIDETAIFRGPDPHQFRQPPPPITVSGSQG
HPGGDGYPEKRKQSLILSPLAAARKFSFIGNSQQTANTTQSMTEEGVRELSEKFSVVPEDDQVEEVLV
RGNMYHHGLQHLNGORRQSVLAFITNSQOERREIQSFRKKSITPQCDLASELDIYARRLSKDSVYD
ISEEVDEEDMEQCFADERENIFETTSWSTYLRYITNRSVLYVLIIFIVFVFIIEVAGSVIGIFLITDTIW
RDSANPSSPNYIDEQHPNASSTPVHLAVIVTPTSAYYIIYIYVATSESVLALGFFRGLPLVHTLLTVSKR
LHEQMLSAVIRAPMAVLNMTMKTGRIMNRFTKDMATIDMLPLVVFDLIQTLIVTGAIFTVSIIMRPIFL
AAIPLAVIFVVLRYKFLRTGQQLKLEAEARSPIFSHLIISLKGWLTIRAFGRQTYFETLFHKALNHTHTA
TWFHYLATLRWFLFRCDMIFVLFSSAAAFIAVGTNQDKPGEVGIIVALAMLILGTFQWAVITSITVDGLM
RSVDRVFKFIDLPTTEPMPGKSGGKGGPDLVIDNPHAQDYWPNRGQMDVQGLTVKYTEAGRAVLNDISFS
VDGGQSIGLLGRTGSGKSTLLSALLRLASTDGEISIDGVSWSVSLHTWRKAFGVVPQRVFILTGTFRMN
LDPHGRYSDEELWRVAEEVGLKSVIEQFPDKLDFQLEDGGNVLSNGHKQLLCLARSILSKARILLLDEPS
AYLDPITLQVLRKTKQSFSGCTVILSEHRVEPLLECQSFLEIEGSAIKSYDSIQKLLNETSHLKQAMSP
ADRLHLFPTLHRLNSSKRAPQQTAKISSLPPEAAEDEVHDTL

```

**Database:** Smart-BLAST database

Copy and paste the above sequence into the SmartBLAST form (<http://blast.ncbi.nlm.nih.gov/smartblast/>) and click the BLAST button.

### Notes:

SmartBLAST quickly identifies the striped bass protein as a likely homolog of CFTR in zebrafish and human from the model organism (landmark) database. The matches from nr show close matches to a fish species (European seabass) in the same family (Moronidae) as the striped bass. The other matches from nr are from other perciform fish, which are more closely related to the striped bass than the zebrafish.

**Summary** Please, let us know what you think

A concise summary of the three best matches in the sequence database together with the two best matches from well-studied reference species, showing phylogenetic relationships based on multiple sequence alignment and conserved protein domains.

Conserved domains for the query:	ABC <sub>nr</sub>	P-loop <sub>nr</sub>	ABC <sub>me</sub>	P-loop <sub>me</sub>
cystic fibrosis transmembrane conductance regi ...	█	█	█	█
cystic fibrosis transmembrane conductance regi ...	█	█	█	█
CFTR	█	█	█	█
Cystic fibrosis transmembrane conductance regi ...	█	█	█	█
<b>Your query: 1 ORF:2178..6716 Frame -3</b>	█	█	█	█
cystic fibrosis transmembrane conductance regi ...	█	█	█	█

[See full multiple alignment](#)   [Legend](#)

Another useful feature of SmartBLAST is that it often allows you to see homologs in more distantly related model organisms than you could see with the default settings in ordinary protein BLAST on the web. In this case you can easily identify homologs in *Drosophila melanogaster*, *Arabidopsis thaliana* and *Caenorhabditis elegans* by looking at the Additional BLAST hits section of the output. These matches from the SmartBLAST landmark database are not visible in a protein BLAST search against nr unless you set the number of target sequences to a very high number. This is because the large number of matching vertebrate proteins in nr overwhelms the output.

**Independent practice: identifying a protein translation from the mango**

**Query:** ORF from Mangifera indica (Mango) TSA record GBCV01016775

```

.....
>lcl|Sequence 1 ORF:40..2235 Frame +1
MDDMETETAEVSLPEPKIQRLSESVVNRIAAGEVIQRPLSAVKELVENS LDANSTSINVVVKDGGGLKLIQ
VSDDGHGIRYEDLPILCERHTTSKLSKYEDLLSIKSMGFRGEALASMTYVGHVTVTTITKGQLHG YRVSY
RDGVMEHEPKPCA AVKGTQIMVENLFYNMIARRKTLQNSSDDYTKIVDLLSRLAIHHINVGFS CRKHGAA
RADVH SVTTSRRLDSIRT VYGVS VVRSLMNI EASD SDFSSSFKMDGFI SGNYSYVAKKTTMVL F INDRLV
ECSALKRAIEIVYTATLPKASKPFIYMSIVLPSEHV D VNVHPTKREVSLLNQEIIEKIQSVVELKLRHS
NEAISYQEQTVESSPSSMGTSKDLQLNNSLPGPKSQKVP MHKMVRTDSSDPAGRLHAYLQTKPHNHLAE
KSSLSAVRSSVRQRRNPSETADLTSIQELIDDI EGNCHSGLLEIVRHCTYIGMADDVFALLQHNT HLYLA
NVVNLSKELMYQQVLRRAFHNAIQLSEPA PLAELIVLALKEEDLDPESENDDLKEKIAEMNTE LLKQK
GEMLEEFYFCIKIDTHGNLSRLPVILDQYTPDMDRVPEFVLC LGNDVDWEEKNCFO SIAAALGNFYALHL
PLMPNPSGEGLVYYKKEKAFTNPEDGQPSKNTGDDVEMEVDIDHEL FSEAEAAWAQREWSIQHVLF PAMR
LFLKPPTSMATNGTFVQVATLEKLYKIFERC
    
```

**Database:** Smart-BLAST database

**Goals:** Find the closest match in the SmartBLAST database. Find homologs in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*.