NCBI BLAST Services

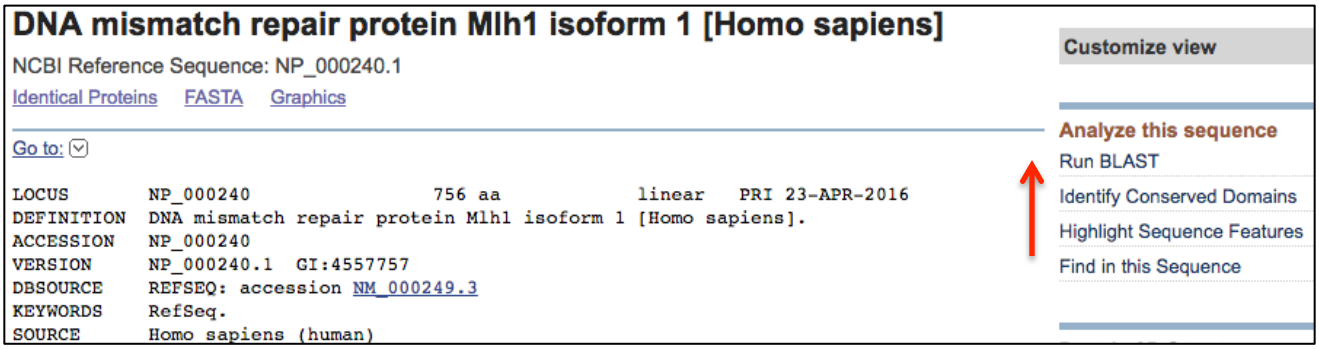
Protein BLAST

Query: human MLH1, NP_000240
Program: blastp
Database: refseq_protein
Goals:

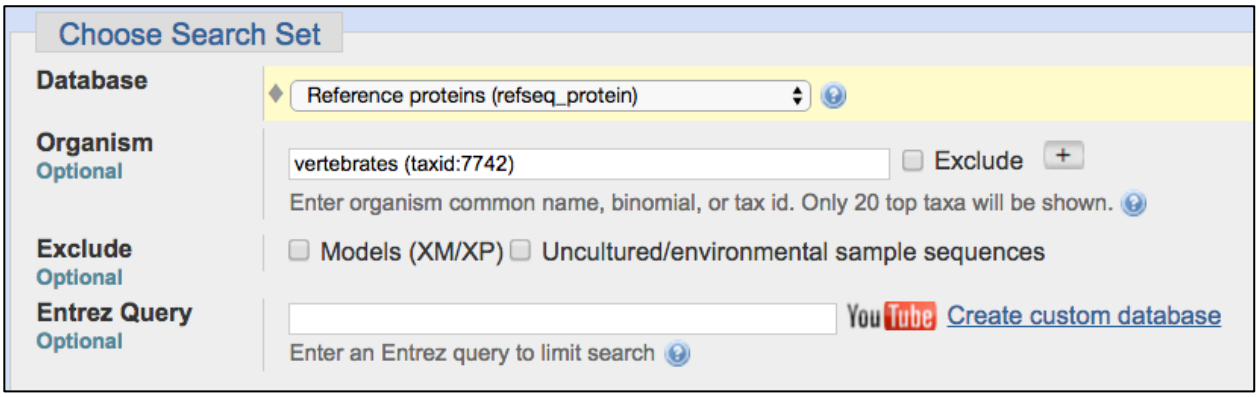
- Explore DNA mismatch repair proteins in vertebrates using Reference Sequences.
- Demonstrate the usefulness of organism limits, taxonomy report, and the link to multiple sequence alignment**.**

Procedure:

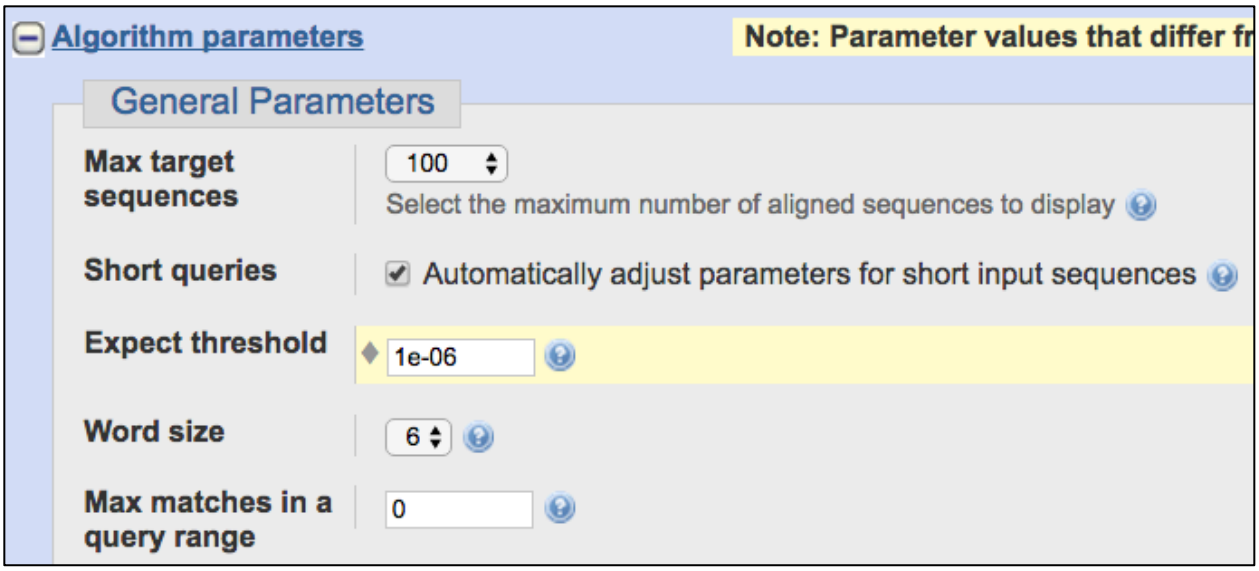
- Retrieve **NP_000240** from the Entrez protein service.
- Click “Run BLAST” under the Analyze this sequence portlet.



- Select **refseq_protein** as the database.
- Enter "vertebrates” in the Organism input box, select from the suggested list



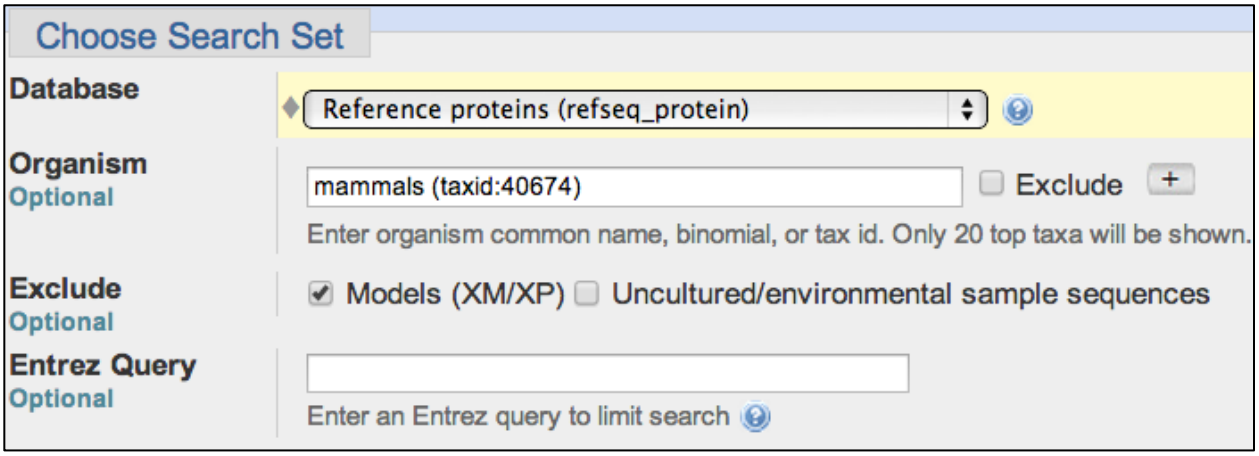
- Click **Algorithm Parameters** and increase the **Max target sequences** to 1000 and set the Expect threshold to a stringent value, 1e-6.



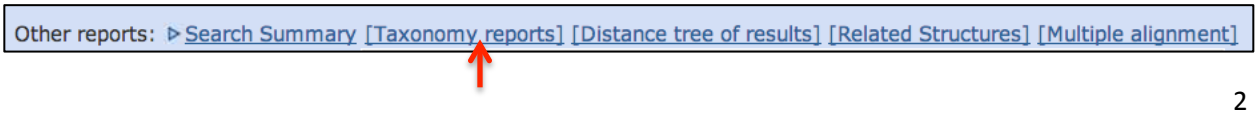
- Click **BLAST** to submit the search
- The matches have different types of RefSeq accessions. The XP_ entries represent proteins derived from gene models. NP_ accessions represent proteins derived from experimentally supported expressed sequences.

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.

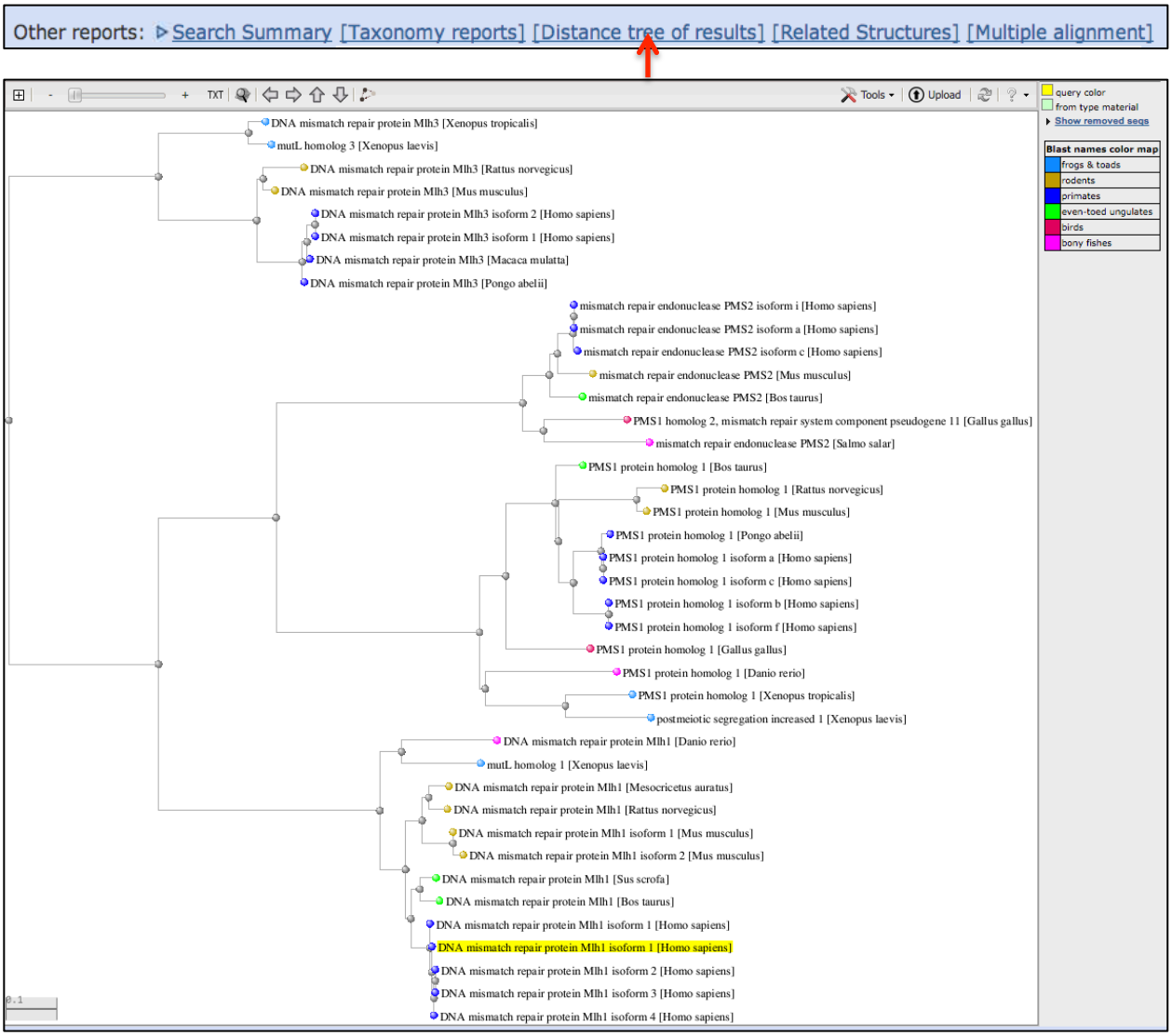
- From the results page, click **Edit and resubmit**
- Check the Exclude **Models (XM/XP)** checkbox



- 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.



- 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 clustere



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

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.

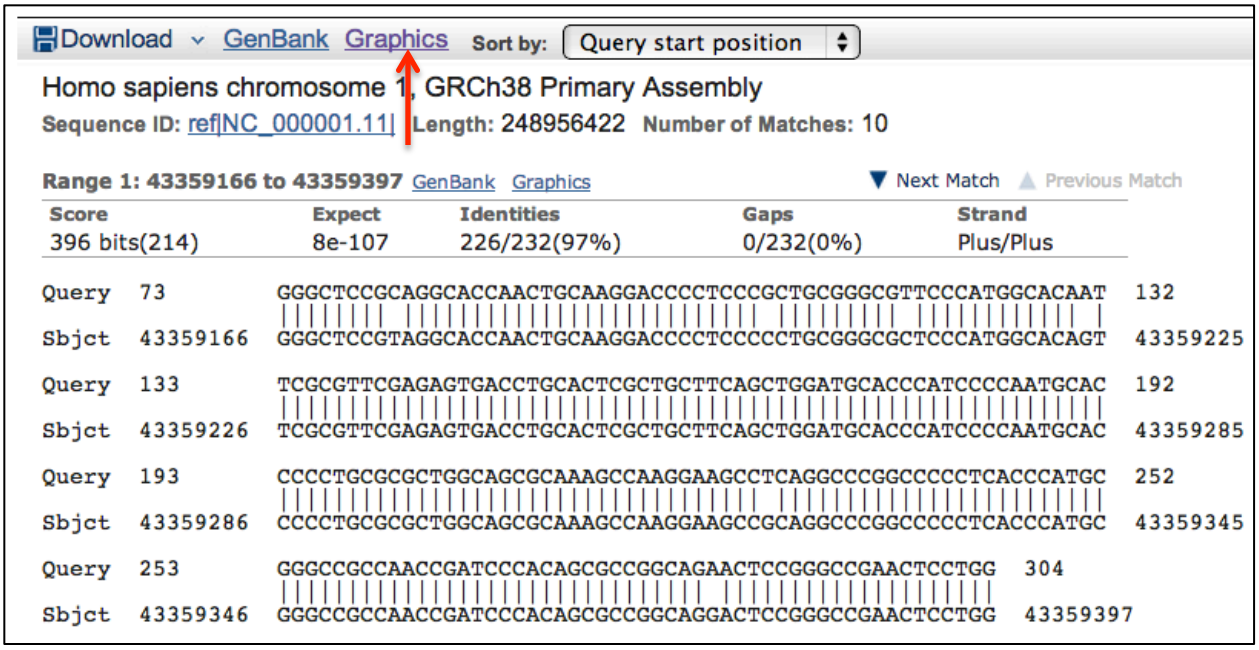
- Click the linked hit for Homo sapiens chromosome 1, GRCh38 Primary Assembly to go to the alignment
- Click **Query Start position** to arrange the matches according to exon order

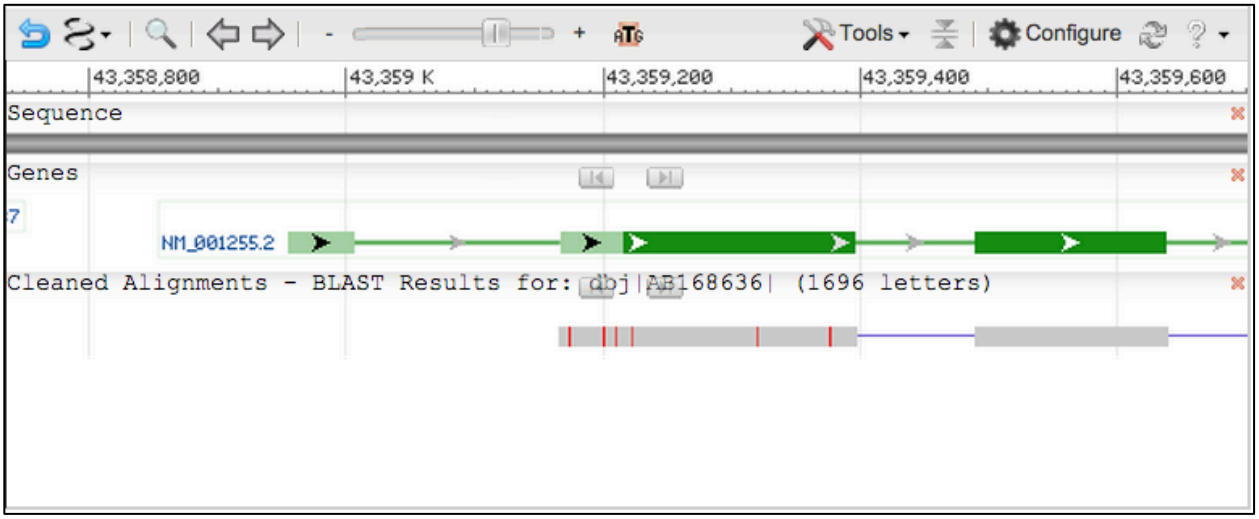
The first aligned segment starts at position 73 of the mRNA. Megablast misses the first exon hit as well as a match to some related transcripts. Re-running the search with blastn finds this hit. You will need to set the Expect threshold to 1e-6 to avoid additional non-significant matches.

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.





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 discontiguous 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).

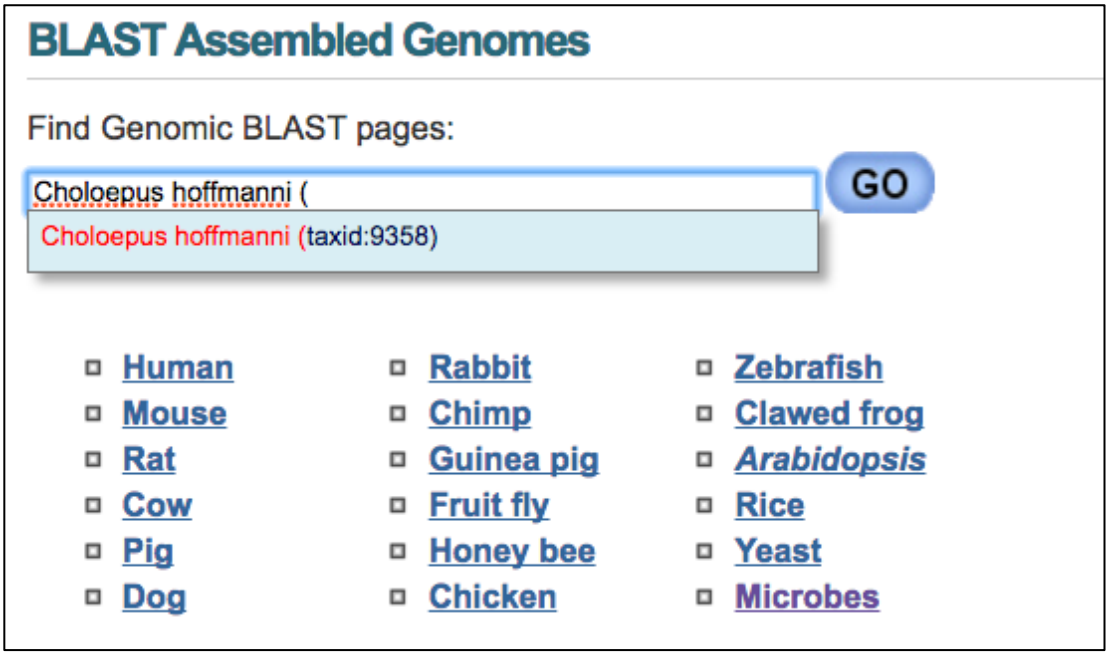
- Examine the alignment to the human transcript NM_001255.

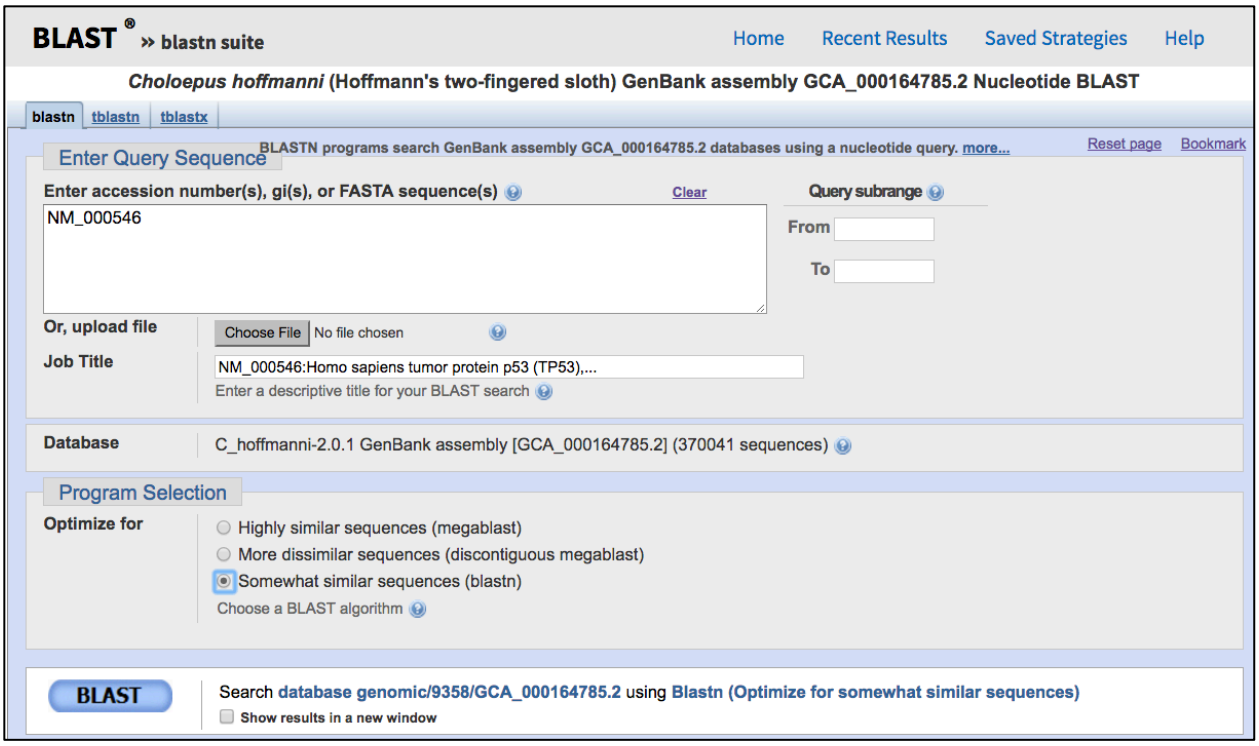
The macaque mRNA sequence has a single base deletion relative to the human transcript. This results in a frame shift making the protein translation diverge at the C- terminus. This is most likely a sequencing error as the other mammalian CDC20 proteins agree with the human sequence. You can use a blastx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

Independent exercise: Finding TP53 in the sloth (Choelepus 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 Choelepus hoffmani genome BLAST page.





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:

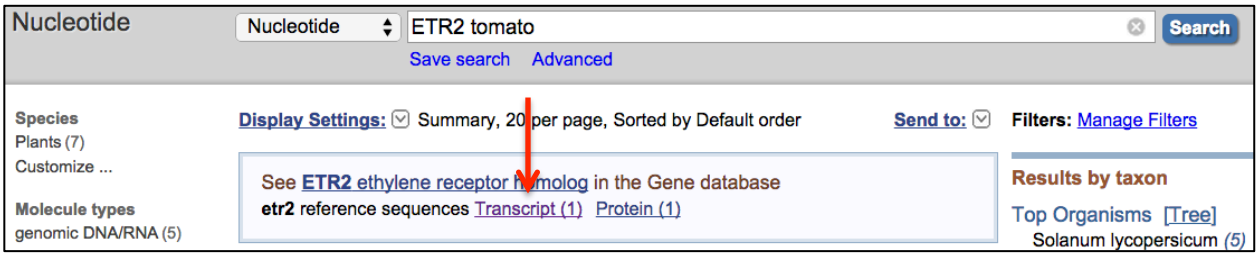
- Retrieve NM_001247224.1 from Entrez nucleotide and copy to the clipboard.
- Use the genomic database finder on the BLAST homepage to find the potato genome BLAST page**.**
- Run BLAST
- Find corresponding potato gene.

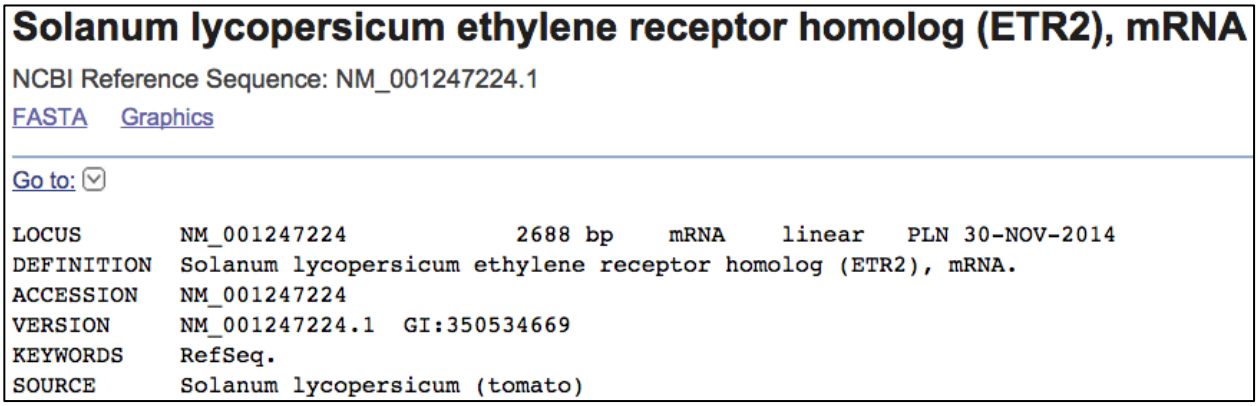
Notes

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.





Solanum lycopersicum ethylene receptor homolog (ETR2), mRNA

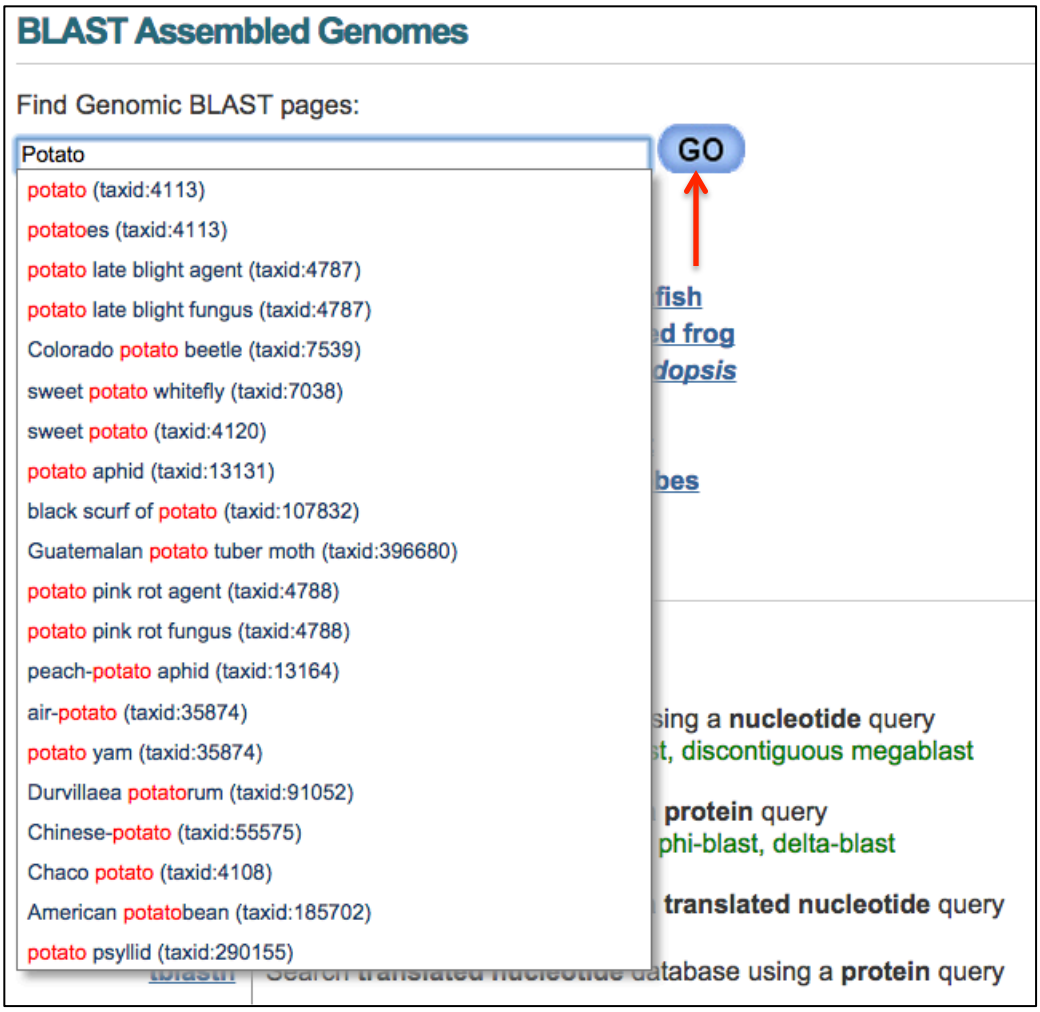
NCBI Reference Sequence: NM_001247224.1

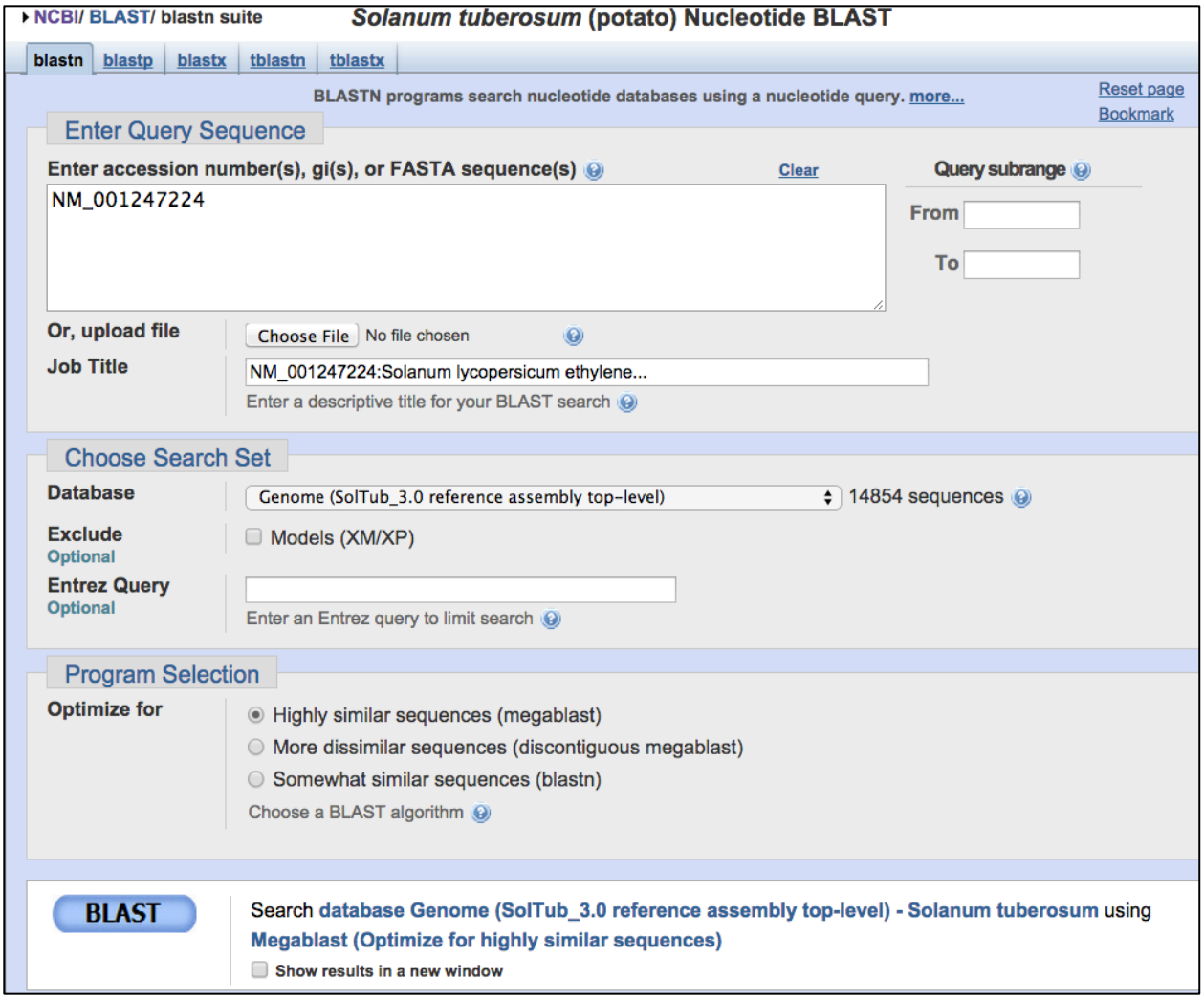
FASTA Graphics

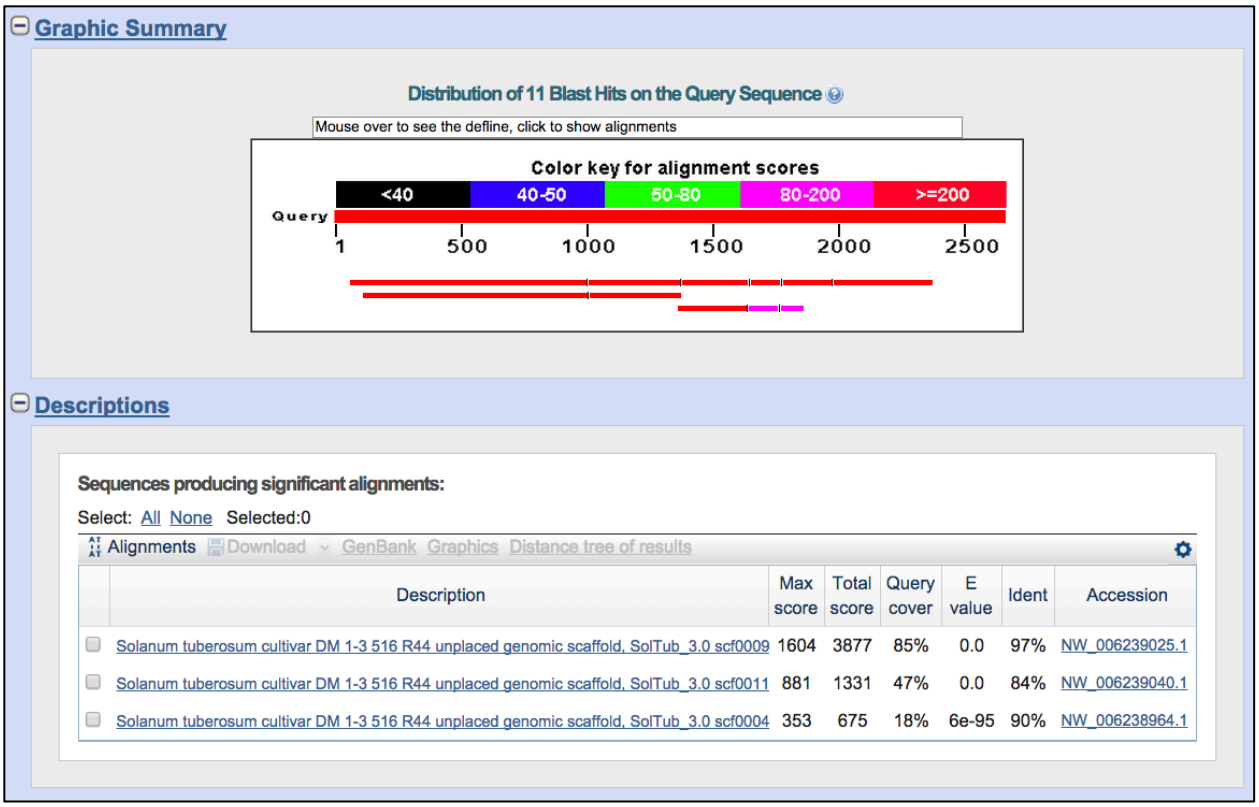
Go to:

```
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)
```

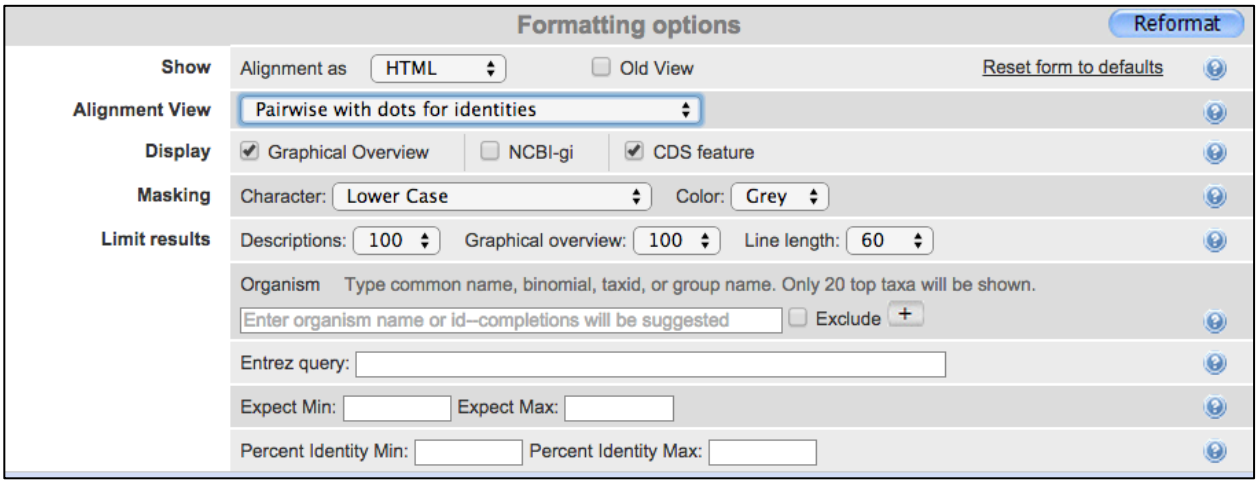
Use the Genome database selector to find the potato genome BLAST page.

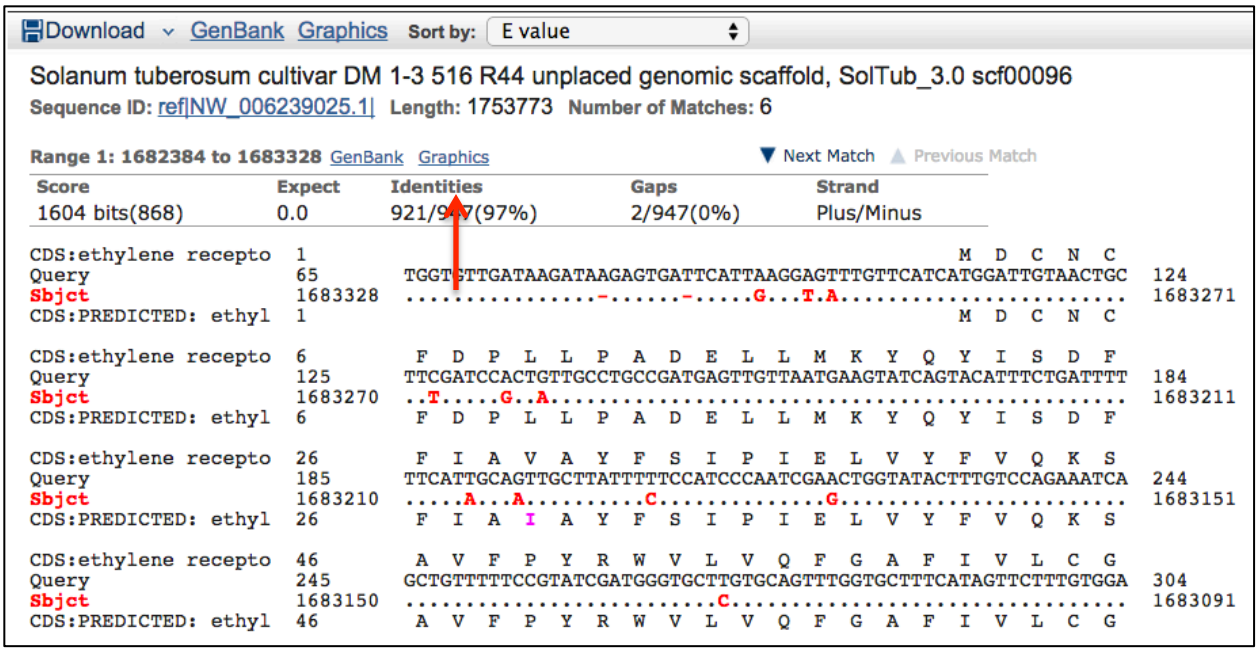

BLAST Assembled Genomes
Find Genomic BLAST pages:
Potato
GO
potato (taxid:4113)
potatoes (taxid:4113)
potato late blight agent (taxid:4787)
potato late blight fungus (taxid:4787)
Colorado potato beetle (taxid:7539)
sweet potato whitefly (taxid:7038)
sweet potato (taxid:4120)
potato aphid (taxid:13131)
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 potatobean (taxid:185702)
potato psyllid (taxid:290155)
fish
d frog
dopsis
bes
sing a nucleotide query
st, discontiguous megablast
protein query
phi-blast, delta-blast
translated nucleotide query
tblastn
tabase 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
NM_001247224
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
NM_001247224:Solanum lycopersicum ethylene...
Enter a descriptive title for your BLAST search
Choose Search Set
Database
Genome (SolTub_3.0 reference assembly top-level)
14854 sequences
Exclude
Optional
Models (XM/XP)
Entrez Query
Optional
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
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

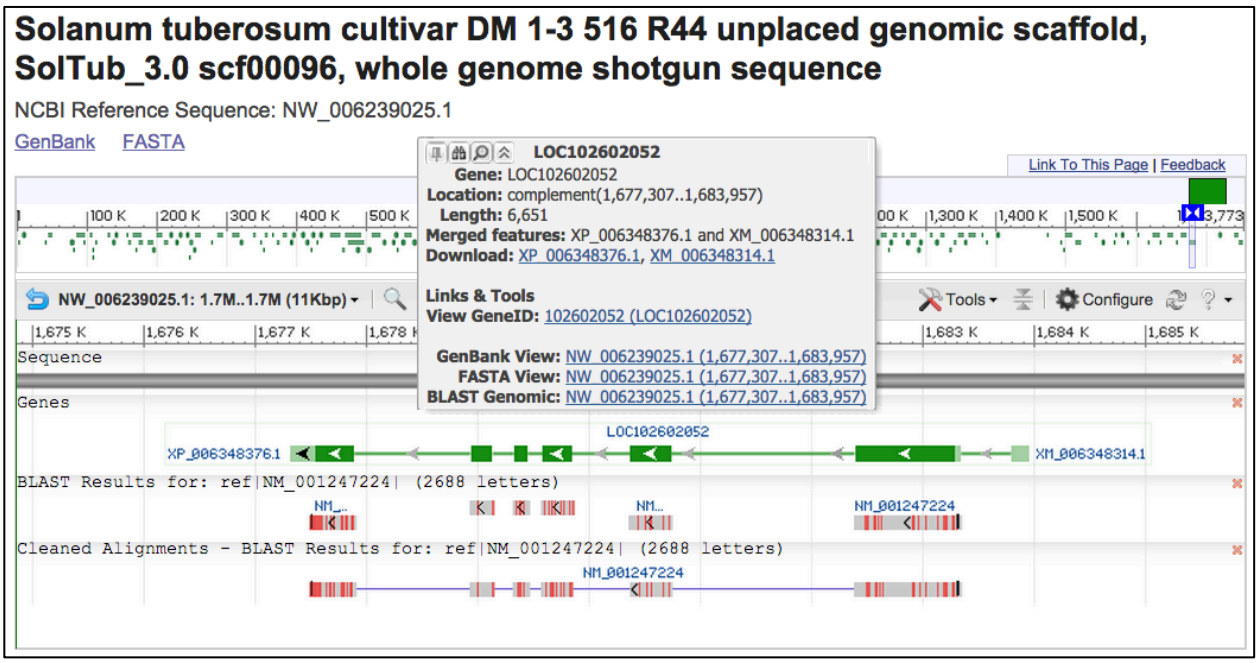


Use the Formatting options to highlight differences (dots for identities) and show coding regions (CDS).





Use the Graphic link on the first aligned segment (exon) to see the potato genome and the corresponding gene.



You can use edit and resubmit to show the increased sensitivity of blastn compared to megablast. You should pick up additional genes and match the upstream, untranslated exon and the 3' UTR.

Microbial Genomes BLAST

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

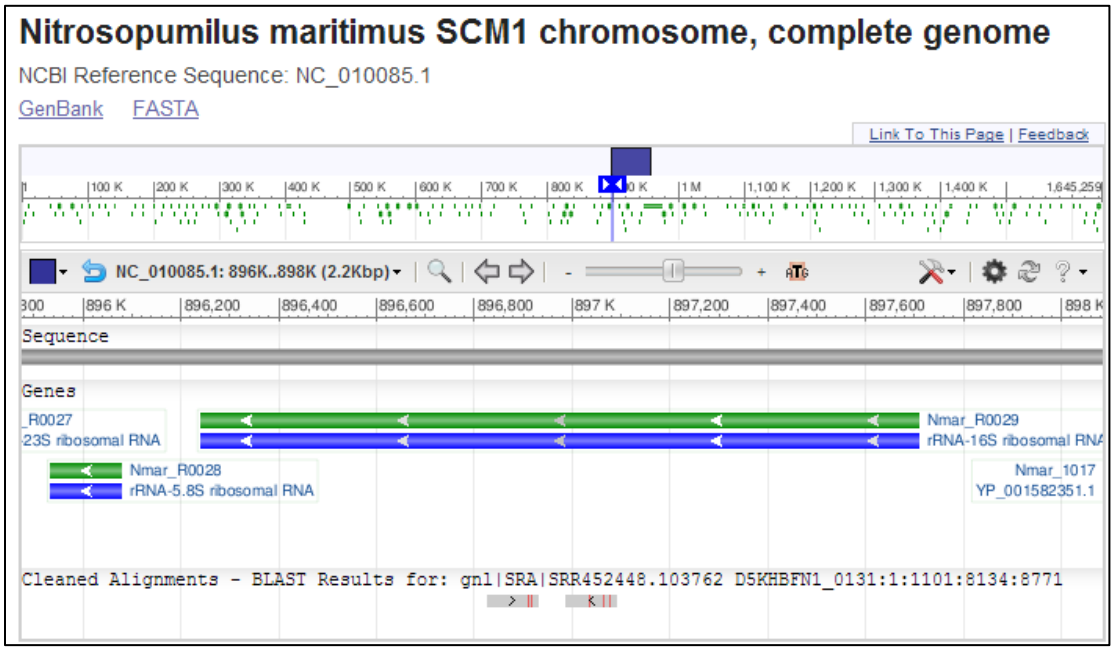
>gnl|SRA|SRR452448.103762 D5KHBFN1_0131:1:1101:8134:8771
ACCCCTTGTGGTGCTCCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC
AGGCGGCAAACTTAACGGCTTTCCTGCCGCACTGCATTTGGTGGTAAAATGCTTTGATCT
ATCGATGACCACCTGTGGCGAAGGCGGTCTACTAGAACACGTCCGACGGTGAGGGATGAA
AGCTGGGGGAGCAAACCGGA

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.



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)*
ACCCCTTGTGGTGCTCCCCCGCCAATTCCTTTAAGTTTCATACTTGCGTACGTACTTCCC
AGGCGGCAAACTTAACGGCTTTCCTGCCGCACTGCATTTG
>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 GBSQMEQ01CJRIM *Forward (Biological)*
ACACAACCTGTAGGTTTAGGAAGACCGGCCATCTTACAAGCTCCCTTTCCGGGTCCTGAC
GGAAAAAGCTCATAAATGTGTTTCAGTTTGAAACCGGTGACTTTTGAAAGTACCCGAACC
ATCGGTGCGATACCGTTTTTCTCATAATAATCACGAAGTACATCGATGACTTTCTGGTGT
TCGTCATTGAGCTCTTCAATGCCTTCTTCCTTCTTTACCCACTGAACCCATTCCTGGCTC
CAGTTGTCATATGAGTCGATGAAACCGTCTTCGTCGACCGTGAAAGTTTTTCCCATAAAT
TCTACACTTGGCATTTAATCGTCCTCCTTTGGTGTTATATTATTTTCTTTCGGCTCGCGG
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.

>gnl|SRA|SRR452448.103068966 D5KHBFN1_0131:1:2208:1519:200066
GGTGAGTAATGCTTGGGAACTTGCCTTTGCGAGGGGGATAACAGTTGGAAACGACTGCTA
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 dara

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

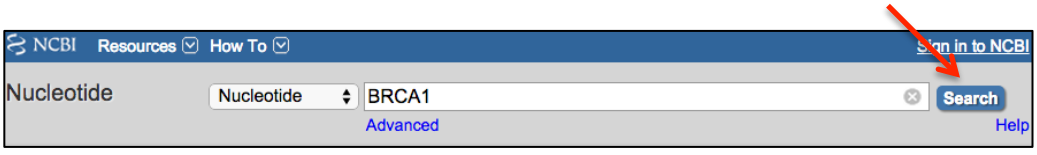
Query: Human BRCA1 exon 15 plus flanks (NG_005905.2 from 146746 to 147056).
Organism limit: human
Database: Reference Genome from selected organisms
Avoid known SNPs: On and off

- 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.

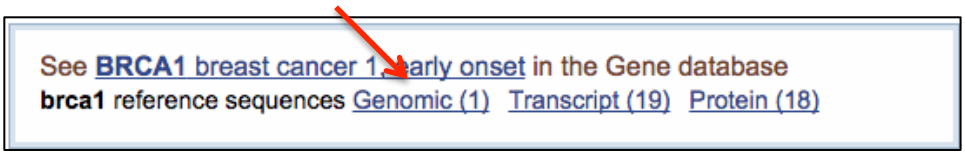
Notes

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

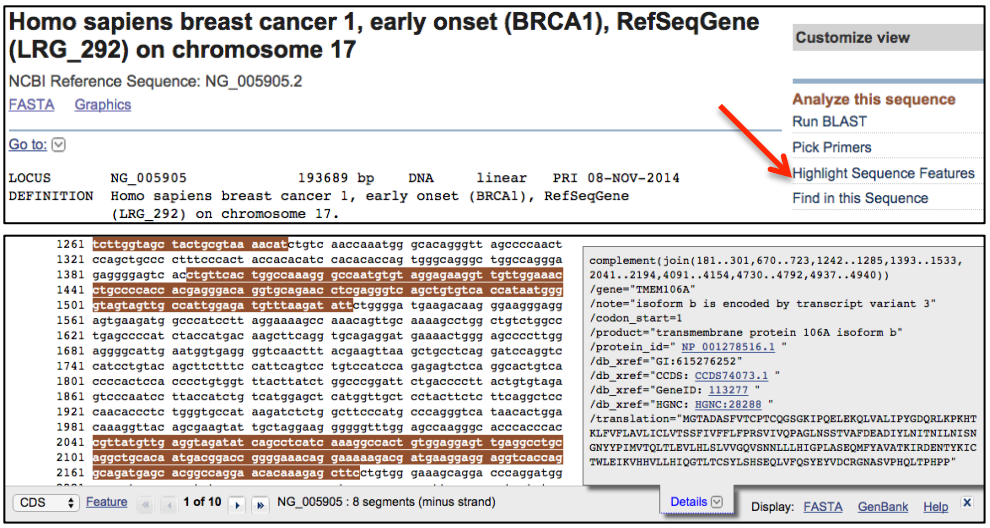
1. Search BRCA1 in the NCBI Nucleotide system.



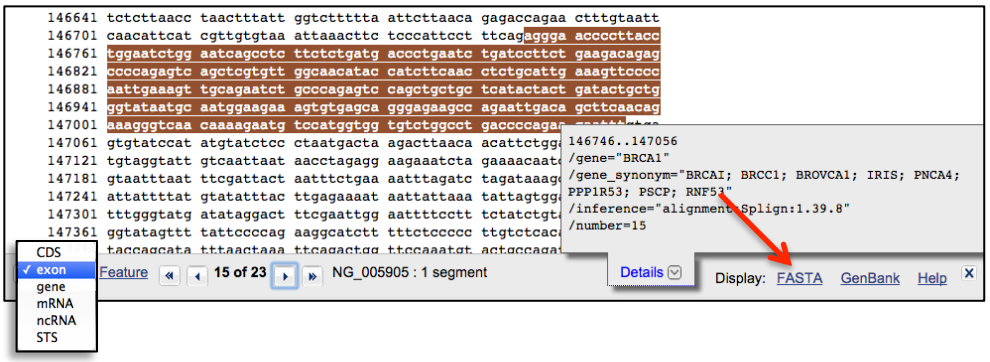
2. Follow the Genomic link in the Gene Sensor box at the top of the Nucleotide results to retrieve the RefSeqGene record (NG_005905) for the BRCA1 gene.



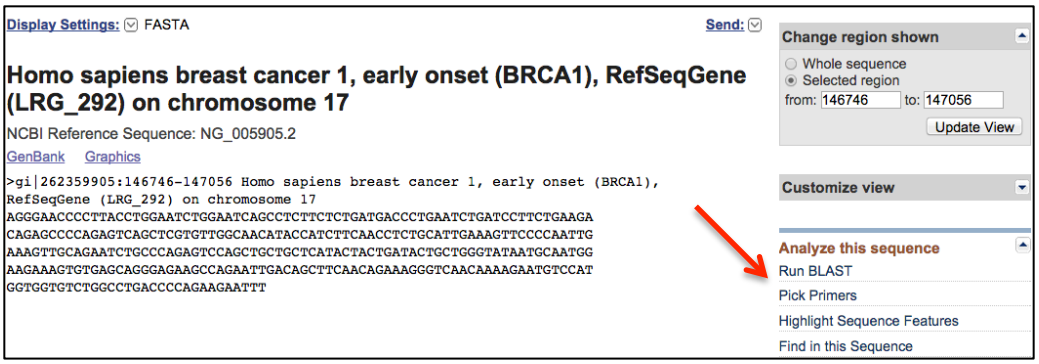
3. Click the "Highlight Sequence Features" in the right-hand column of the sequence record to activate feature highlighting. You will see the coding sequence (CDS) feature of the gene highlighted.



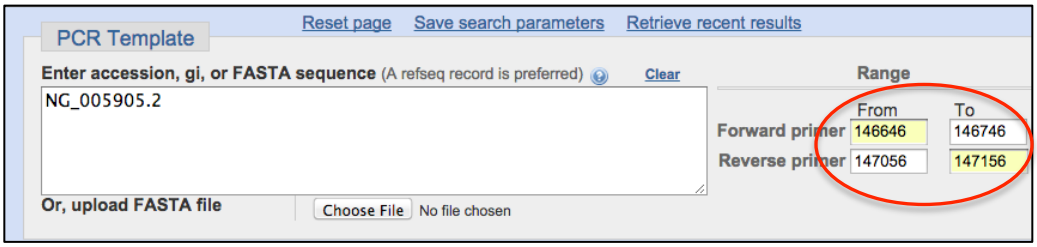
4. Change the "Feature" pull-down list at the bottom left of the sequence display from "CDS" to "exon" and then navigate to exon 15.



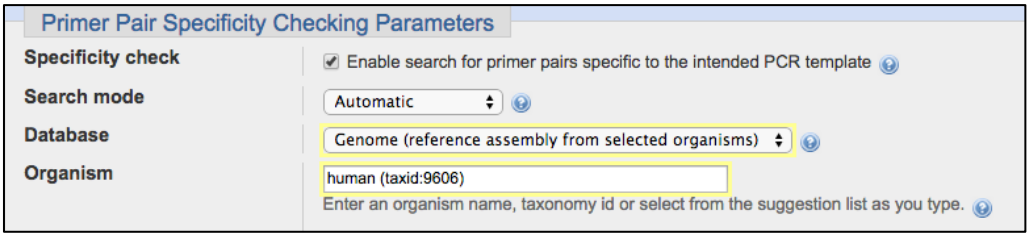
5. Follow the FASTA link to display the highlighted exon as a separate view. Then follow the link in the right-hand column of the sequence display to "Pick Primers."



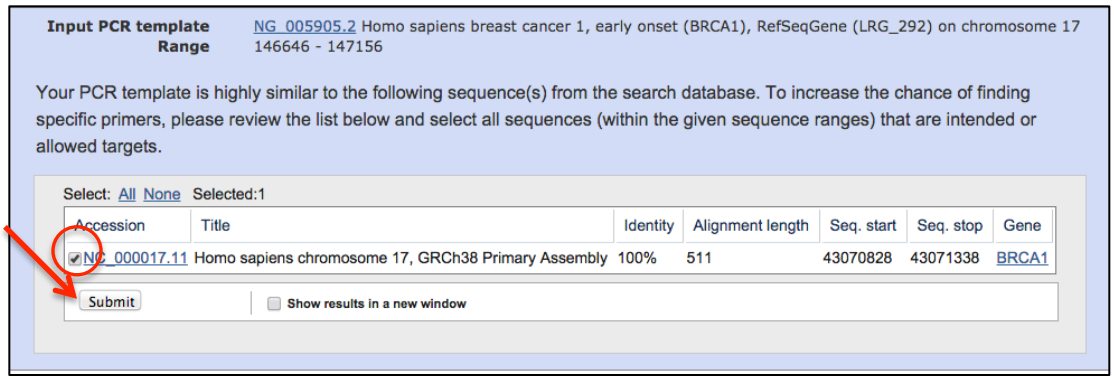
6. Edit the primer ranges in Primer BLAST so that the forward and reverse primers will bind upstream and downstream of the exon. For example, set the forward primer range from 146646 to 146746 and the reverse primer from 147056 to 147156. This will provide sufficient upstream and downstream sequence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

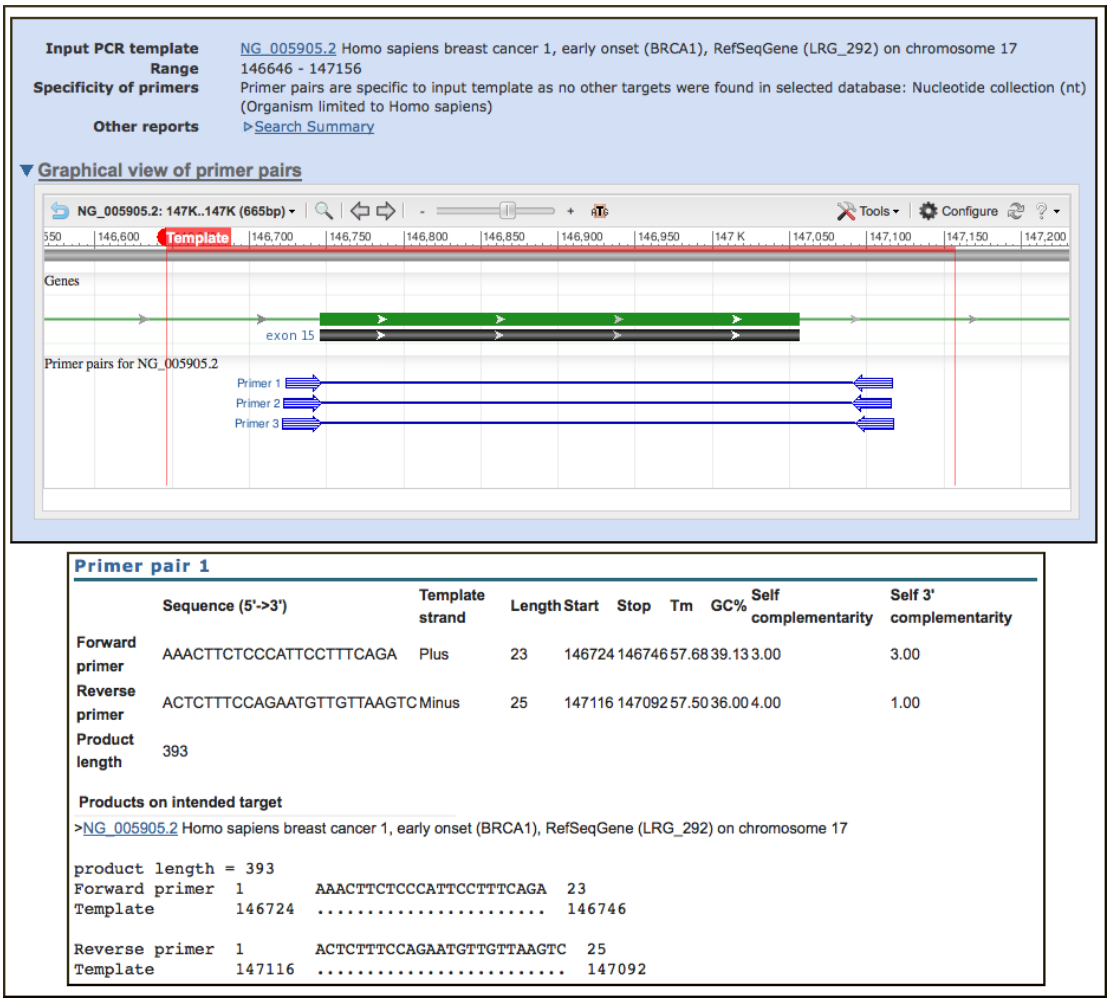


7. We want these primers to amplify only the target region from the human genome sequence. Set the database for Primer-BLAST to perform a specificity check to “Genome (reference assembly from selected organisms)” and leave the Organism limit set to human.



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.





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?

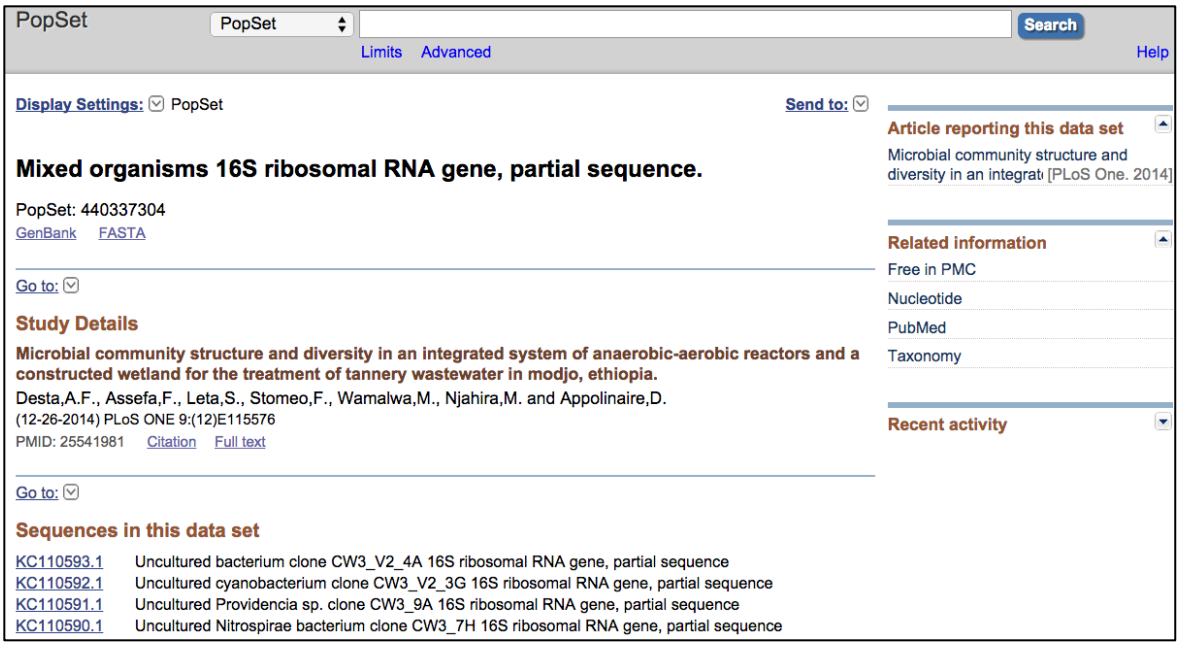
Using MOLE-BLAST to cluster targeted sequences

Query: 16S sequences from wastewater metagenome
Database: 16S reference sequences, nr

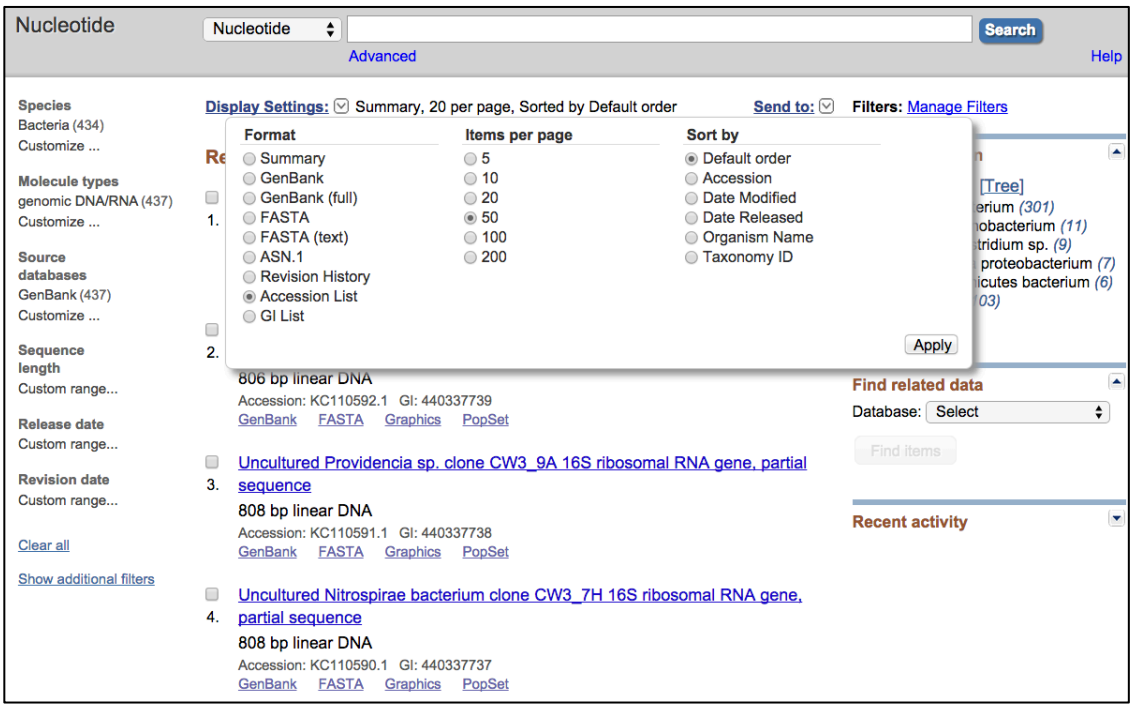
- Retrieve PopSet: 440337304 (www.ncbi.nlm.nih.gov/popset/440337304)
- Follow the link to nucleotide
- Copy the first 30 accessions to cluster in MOLE-BLAST
- Cluster with the 16S reference sequences, and nr

Notes

Follow the link to nucleotide from the PopSet record.

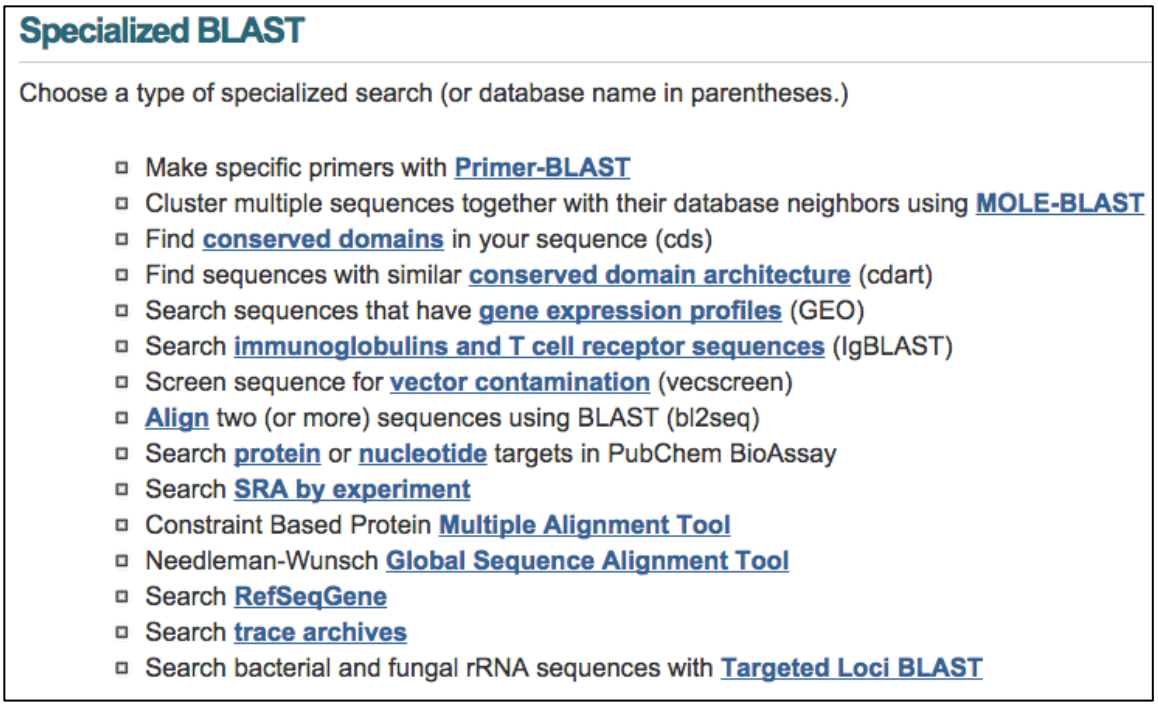


Use the Display settings to get and accession list of the first 50 records and copy these to your clipboard.

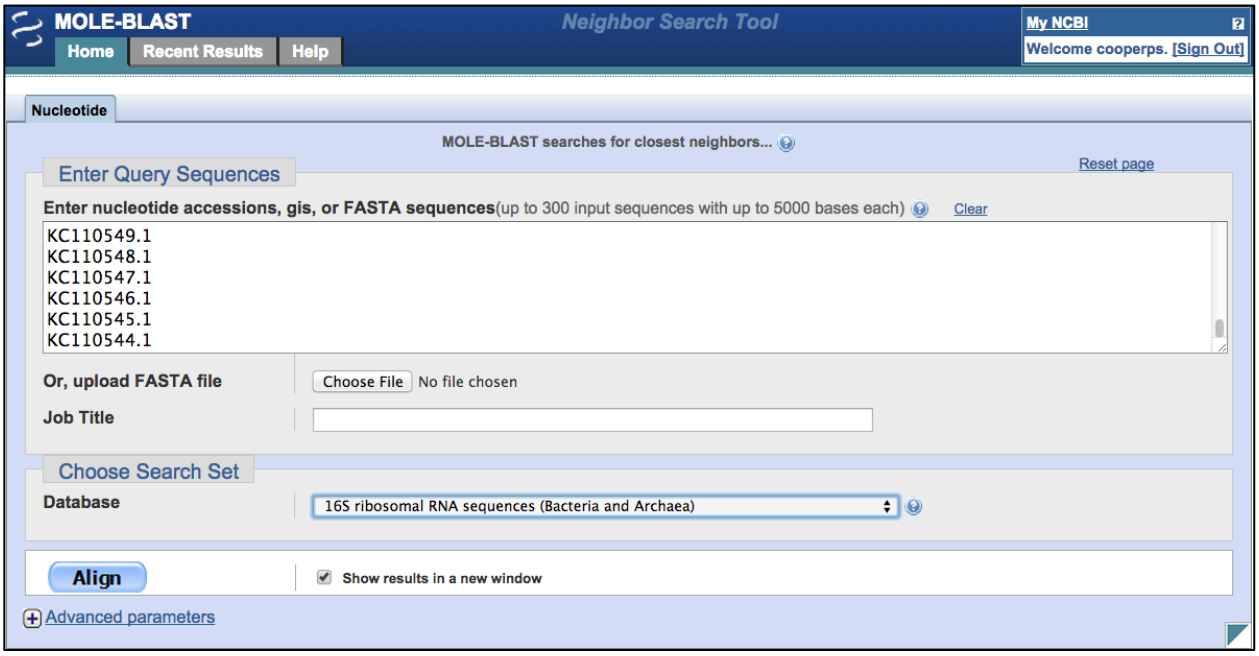


```
KC110593.1
KC110592.1
KC110591.1
KC110590.1
KC110589.1
KC110588.1
KC110587.1
KC110586.1
KC110585.1
KC110584.1
KC110583.1
KC110582.1
KC110581.1
KC110580.1
KC110579.1
KC110578.1
KC110577.1
KC110576.1
KC110575.1
KC110574.1
KC110573.1
KC110572.1
KC110571.1
KC110570.1
KC110569.1
KC110568.1
KC110567.1
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
```

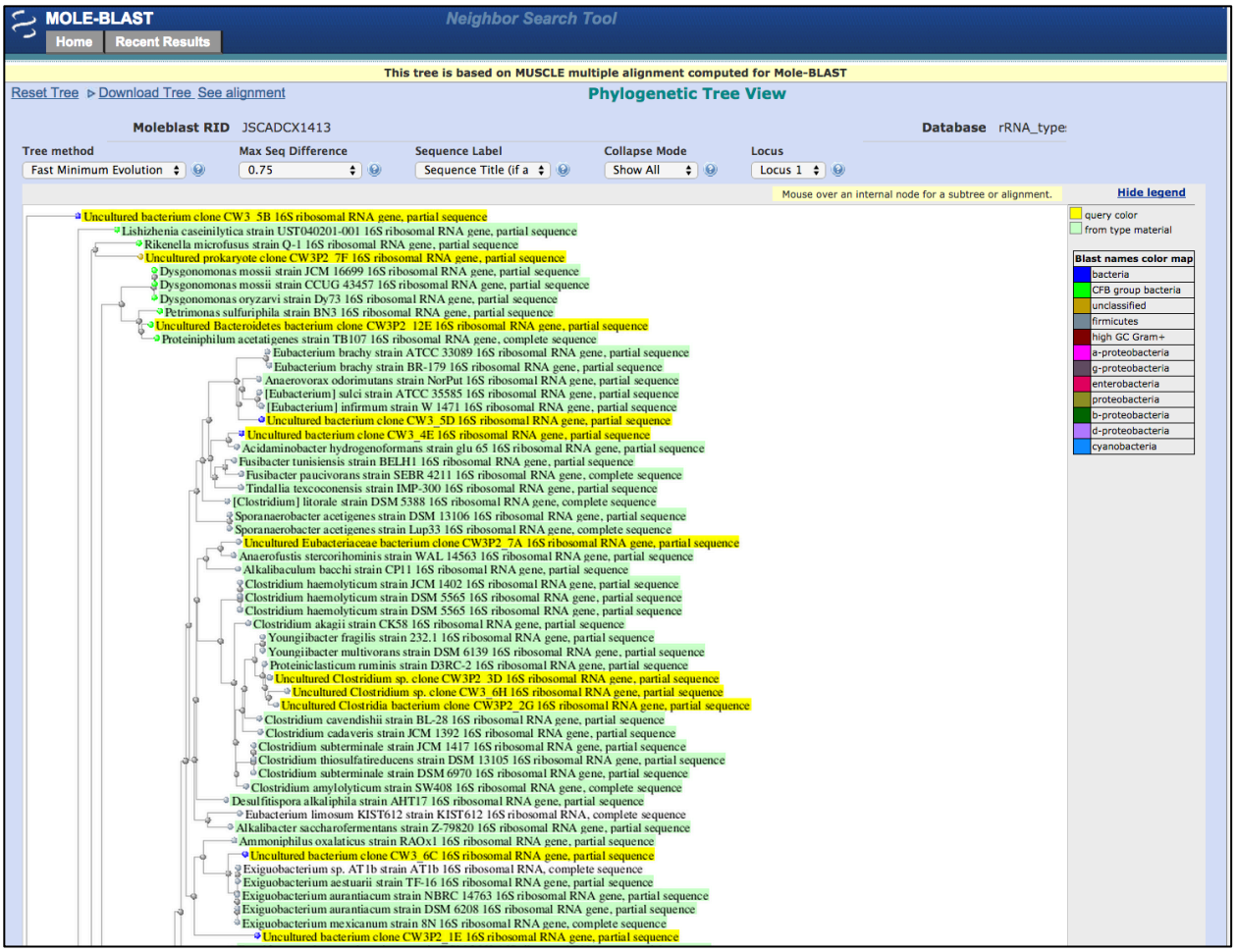
Link to MOLE-BLAST from the “Specialized BLAST” section of the BLAST homepage.

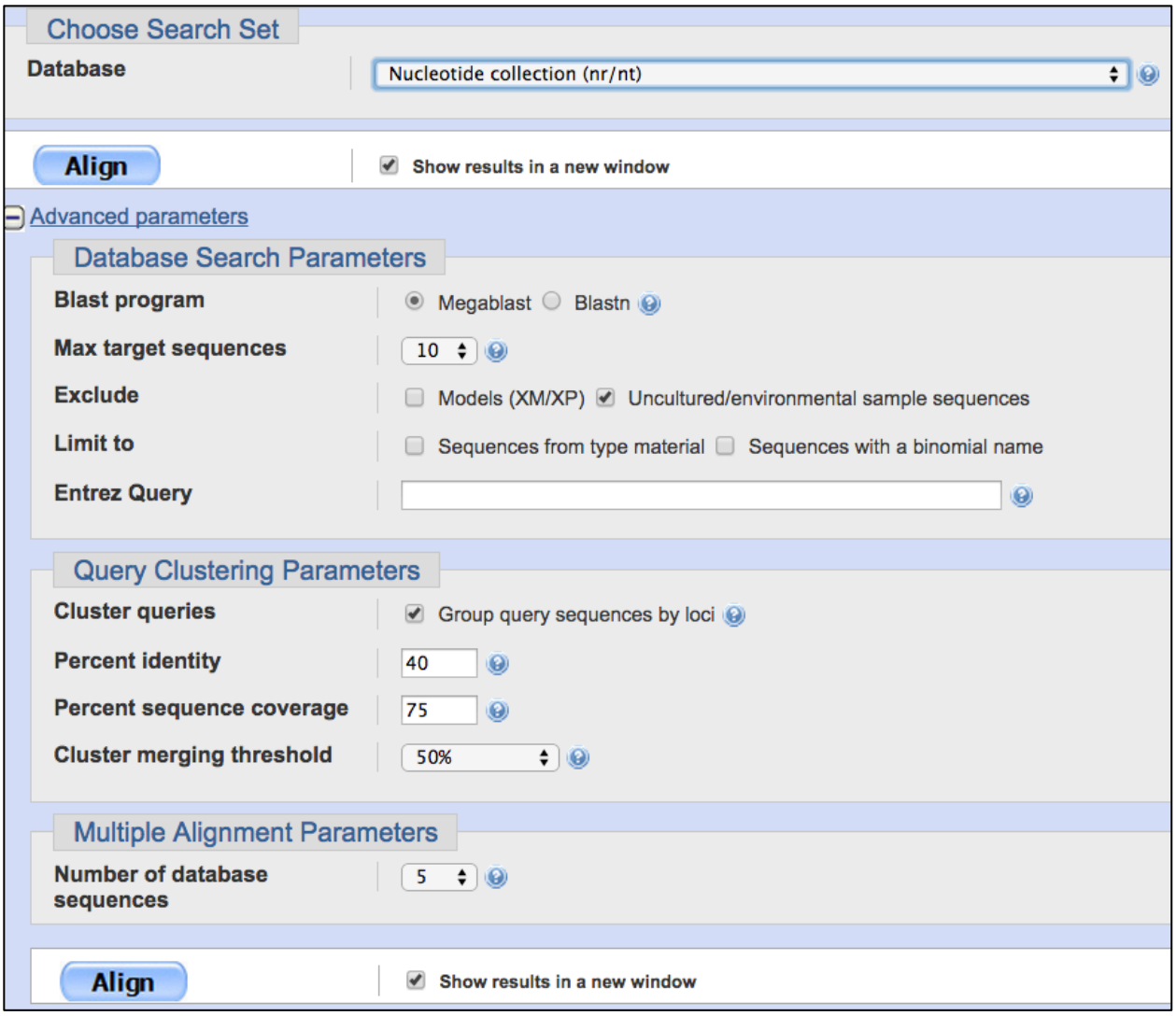


Paste the query accessions in the form and set the database to 16S rRNA. It's helpful to click the "Show results in new window" box so you can adjust settings and resubmit your search.



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.


MOLE-BLAST
Neighbor Search Tool
Home
Recent Results
This tree is based on MUSCLE multiple alignment computed for Mole-BLAST
Reset Tree
Download Tree
See alignment
Phylogenetic Tree View
Moleblast RID JSCADCX1413
Database rRNA_type
Tree method
Fast Minimum Evolution
Max Seq Difference
0.75
Sequence Label
Sequence Title (if a
Collapse Mode
Show All
Locus
Locus 1
Mouse over an internal node for a subtree or alignment.
Hide legend
query color
from type material
Blast names color map
bacteria
CFB group bacteria
unclassified
firmicutes
high GC Gram+
a-proteobacteria
g-proteobacteria
enterobacteria
proteobacteria
b-proteobacteria
d-proteobacteria
cyanobacteria
Uncultured bacterium clone CW3_5B 16S ribosomal RNA gene, partial sequence
Lishizhenia caseinilytica strain UST040201-001 16S ribosomal RNA gene, partial sequence
Rikenella microfusus strain Q-1 16S ribosomal RNA gene, partial sequence
Uncultured prokaryote clone CW3P2_7F 16S ribosomal RNA gene, partial sequence
Dysgonomonas mossii strain JCM 16699 16S ribosomal RNA gene, partial sequence
Dysgonomonas mossii strain CCUG 43457 16S ribosomal RNA gene, partial sequence
Dysgonomonas oryzarvi strain Dy73 16S ribosomal RNA gene, partial sequence
Petrimonas sulfuriphila strain BN3 16S ribosomal RNA gene, partial sequence
Uncultured Bacteroidetes bacterium clone CW3P2_12E 16S ribosomal RNA gene, partial sequence
Proteiniphilum acetatigenes strain TB107 16S ribosomal RNA gene, complete sequence
Eubacterium brachy strain ATCC 33089 16S ribosomal RNA gene, partial sequence
Eubacterium brachy strain BR-179 16S ribosomal RNA gene, partial sequence
Anaerovorax odorimutans strain NorPut 16S ribosomal RNA gene, partial sequence
[Eubacterium] sulci strain ATCC 35585 16S ribosomal RNA gene, partial sequence
[Eubacterium] infirmum strain W 1471 16S ribosomal RNA gene, partial sequence
Uncultured bacterium clone CW3_5D 16S ribosomal RNA gene, partial sequence
Uncultured bacterium clone CW3_4E 16S ribosomal RNA gene, partial sequence
Acidaminobacter hydrogenoformans strain glu 65 16S ribosomal RNA gene, partial sequence
Fusibacter tunisiensis strain BELH1 16S ribosomal RNA gene, partial sequence
Fusibacter paucivorans strain SEBR 4211 16S ribosomal RNA gene, complete sequence
Tindallia texcoconensis strain IMP-300 16S ribosomal RNA gene, partial sequence
[Clostridium] litorale strain DSM 5388 16S ribosomal RNA gene, complete sequence
Sporanaerobacter acetigenes strain DSM 13106 16S ribosomal RNA gene, partial sequence
Sporanaerobacter acetigenes strain Lup33 16S ribosomal RNA gene, complete sequence
Uncultured Eubacteriaceae bacterium clone CW3P2_7A 16S ribosomal RNA gene, partial sequence
Anaerofustis stercorihominis strain WAL 14563 16S ribosomal RNA gene, partial sequence
Alkalibaculum bacchi strain CP11 16S ribosomal RNA gene, partial sequence
Clostridium haemolyticum strain JCM 1402 16S ribosomal RNA gene, partial sequence
Clostridium haemolyticum strain DSM 5565 16S ribosomal RNA gene, partial sequence
Clostridium haemolyticum strain DSM 5565 16S ribosomal RNA gene, partial sequence
Clostridium akagii strain CK58 16S ribosomal RNA gene, partial sequence
Youngiibacter fragilis strain 232.1 16S ribosomal RNA gene, partial sequence
Youngiibacter multivorans strain DSM 6139 16S ribosomal RNA gene, partial sequence
Proteiniclasticum ruminis strain D3RC-2 16S ribosomal RNA gene, partial sequence
Uncultured Clostridium sp. clone CW3P2_3D 16S ribosomal RNA gene, partial sequence
Uncultured Clostridium sp. clone CW3_6H 16S ribosomal RNA gene, partial sequence
Uncultured Clostridia bacterium clone CW3P2_2G 16S ribosomal RNA gene, partial sequence
Clostridium cavendishii strain BL-28 16S ribosomal RNA gene, partial sequence
Clostridium cadaveris strain JCM 1392 16S ribosomal RNA gene, partial sequence
Clostridium subterminale strain JCM 1417 16S ribosomal RNA gene, partial sequence
Clostridium thiosulfatireducens strain DSM 13105 16S ribosomal RNA gene, partial sequence
Clostridium subterminale strain DSM 6970 16S ribosomal RNA gene, partial sequence
Clostridium amylolyticum strain SW408 16S ribosomal RNA gene, complete sequence
Desulfitispora alkaliphila strain AHT17 16S ribosomal RNA gene, partial sequence
Eubacterium limosum KIST612 strain KIST612 16S ribosomal RNA, complete sequence
Alkalibacter saccharofermentans strain Z-79820 16S ribosomal RNA gene, partial sequence
Ammoniphilus oxalaticus strain RAOx1 16S ribosomal RNA gene, partial sequence
Uncultured bacterium clone CW3_6C 16S ribosomal RNA gene, partial sequence
Exiguobacterium sp. AT1b strain AT1b 16S ribosomal RNA, complete sequence
Exiguobacterium aestuarii strain TF-16 16S ribosomal RNA gene, partial sequence
Exiguobacterium aurantiacum strain NBRC 14763 16S ribosomal RNA gene, partial sequence
Exiguobacterium aurantiacum strain DSM 6208 16S ribosomal RNA gene, partial sequence
Exiguobacterium mexicanum strain 8N 16S ribosomal RNA gene, complete sequence
Uncultured bacterium clone CW3P2_1E 16S ribosomal RNA gene, partial sequence



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).

Using SmartBLAST to identify unknown sequences.

Query: ORF from striped bass TSA record GBAA01198466

.....

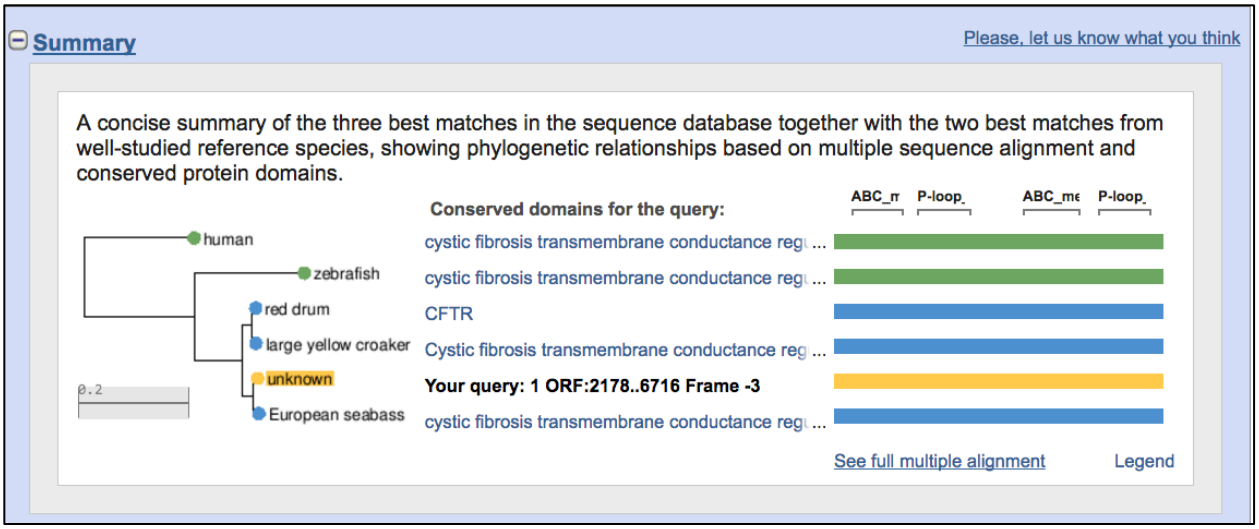
```
>lcl|Sequence 1 ORF:2178..6716 Frame -3
MQKSPVEDANFFSKYFFWWASPLLRKGFTKKLELSDVYKAPSFDLADNLSERLEREWDREVVSAKNQPRL
MRALARCFIGPFAFFGVLLYLGEASKTVQPQLLGRIIGSFDPFHAPERSQGYFLALGLCLLFTARFLLLQ
PAIFGLHHLGMQIRIALFSLIYKKTLKLSSRVLDKISTGQLVSLMSAHLNKLDESLGLAHFVWITPLQCI
LCVGLIWELIEVNGFCALAALTLLGIIQAWLSQKMGPHRVKRAGMINRRLALTSEIVENIHSVKAYGWED
VMETIIKNIRQDEMTLTRKIGSLRYFYSASYFFSAILVIVSAIVPHALSKGIILRRIFTTASYCMVLRMT
LTRQLPGSIQMWYDTLALVKKIEEFLMKEEYRVLEYNLTTTEVELVNVSASWDEGIGELFEKIKQENKAN
GHLNGDAGLFFTNLYITPVLKNISLYLEKGKMLAVAGSTGSGKSSLLMMILGELVPSEGKIRHSGRISFS
PQTSWIIPGTIRDNILFGLTYDEYRYTSVIKACQLEEDFALLPEKDKTHLMEGGVTLSGGQRARLGLARA
VYKDADLYLLDAPFTHLDIVTEKEIFEKCVCKLMASKTRIVVTSKLEHLKRADKILLLHNGDCYFYGTFS
ELQAKRPDFSSLLLGLEAYDNINAERRSSILTETLRRVSIDETAIFRGPDPIRQSFRQPPPPITVSGSQG
HPGGDGYPEKRKQSLILSPLAAARKFSFIGNSQQTANTTQSMTIEEGVRELSERKFSVVPEDDQVEEVLP
RGNMYHHGLQHLNGQRRQSVLAFITNSQGQERREQIQSSFRKKLSITPQCDLASELDIYARRLSKDSVYD
ISEEVDEEDMEQCFADERENIFETTSWSTYLRYITTNRSLVYVLIFIVFVFIIEVAGSVIGIFLITDTIW
RDSANPSSPNYIDEQHPNASSTPVHLAVIVTPTSAYYIIYIYVATSESVLALGFFRGLPLVHTLLTVSKR
LHEQMLSAVIRAPMAVLNTMKTGRIMNRFTKDMATIDDMLPLVVFDLIQLTLIVTGAIFTVSIMRPYIFL
AAIPLAVIFVVLRKYFLRTGQQLKLLEAEARSPIFSHLIISLKGLWTIRAFGRQTYFETLFHKALNTHTA
TWFHYLATLRWFLFRCDMIFVLFFSAAAFIAVGTNQDKPGEVGIIVALAMLILGTFQWAVITSITVDGLM
RSVDRVFKFIDLPTEEPMPGKSGGKGGPDLVIDNPHAQDYWPNRGQMDVQGLTVKYTEAGRAVLNDISFS
VDGGQSIGLLGRTGSGKSTLLSALLRLASTDGEISIDGVSWSSVSLHTWRKAFGVVPQRVFILTGTFRMN
LDPHGRYSDEELWRVAEEVGLKSVIEQFPDKLDFQLEDGGNVLSNGHKQLLCLARSILSKARILLLDEPS
AYLDPITLQVLRKTLKQSFSGCTVILSEHRVEPLLECQSFLMIEGSAIKSYDSIQKLLNETSHLKQAMSP
ADRLHLFPTLHRLNSSKRAPQQTAKISSLPEEAEDEVHDTRL
```

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.



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
MDDMETETAEVSLEPPKIQRLEESVVNRIAAGEVIQRPLSAVKELVENSLDANSTSINVVVKDGGLKLIQ
VSDDGHGIRYEDLPILCERHTTSKLSKYEDLLSIKSMGFRGEALASMTYVGHVTVTTITKGQLHGYRVSY
RDGVMEHEPKPCAAVKGTQIMVENLFYNMIARRKTLQNSSDDYTKIVDLLSRLAIHHINVGFSCRKHGAA
RADVHSVTTSSRLDSIRTVYGVSVVRSLMNIEASDSDFSSSSFKMDGFISGSNYVAKKTTMVLFINDRLV
ECSALKRAIEIVYTATLPKASKPFIYMSIVLPSEHVDVNVHPTKREVSLLNQEIIIEKIQSVVELKLRHS
NEAISYQEQTVESSPSSSMGTSKDLQLNNLSPGPKSQKVPMHKMVRTDSSDPAGRLHAYLQTKPHNHLAE
KSSLSAVRSSVRQRRNPSETADLTSIQELIDDIEGNCHSGLLEIVRHCTYIGMADDVFALLQHNTHLYLA
NVVNLSKELMYQQVLRRFAHFNAIQLSEPAPLAELIVLALKEEDLDPESSENDDLKEKIAEMNTELLKQK
GEMLEEYFCIKIDTHGNLSRLPVILDQYTPDMDRVPEFVLCLGNDVDWEEEKNCFQSIAAALGNFYALHL
PLMPNPSGEGLVYYKKEKAFTNPEDGQPSKNTGDDVEMEVDIDHELFSEAEAAWAQREWSIQHVLFPAMR
LFLKPPTSMATNGTFVQVATLEKLYKIFERC
```

Database: Smart-BLAST database

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