

INTERPRETING DNA SEQUENCE

DNA Sequencing Overview

DNA sequencing involves the determination of the sequence of nucleotides in a sample of DNA. It is presently conducted using a modified PCR reaction where both normal and labeled dideoxy-nucleotides are included in the reaction mix. The dideoxy-nucleotides lack the OH at their 3' carbon atom and are labeled with fluorescent dyes (each nucleotide has a different color). During PCR, mostly normal nucleotides are added to the DNA template. However, when a labeled nucleotide happens to add to the growing DNA strand, chain elongation stops because there is no 3' OH to which the next nucleotide can be attached. Hence, the dideoxy method is also called the chain termination method. At the end of the sequencing reaction, the fragments with fluorescent termini are separated by length from longest to shortest using a polyacrylamide gel (either a big thin slab gel or a narrow capillary tube filled with gel solution) that is scanned with a laser detection device. As each band moves past a viewer, the laser excites the dye, and the color of fluorescence is read by a photocell and recorded on a computer. A computer program evaluates the resulting histogram of fluorescence peaks and "calls" the nucleotides in the order they passed through the viewer.

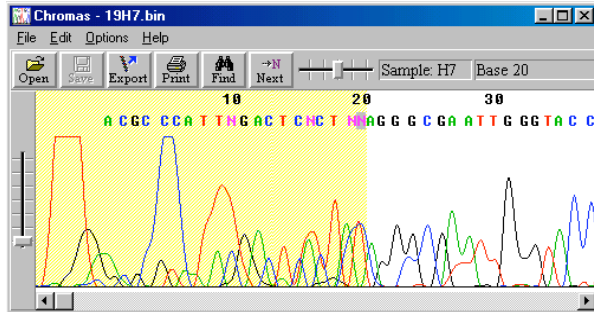
Interpreting Sequencing Results

The computer output for a sequencing run consists of chromatogram (*typically a sequencing facility will email you a file*) that can be opened in one of many free DNA sequence viewers that are available on the web. We will be using a program called FinchTV. When viewing chromatograms, there is some ambiguity at various sites along the DNA sequence as the sequencing machine that is used to "call" the bases can't always precisely determine what nucleotide is represented by either a broad peak or a set of overlaying peaks. In such cases, a letter other than A, C, G, or T is recorded, most commonly "N".

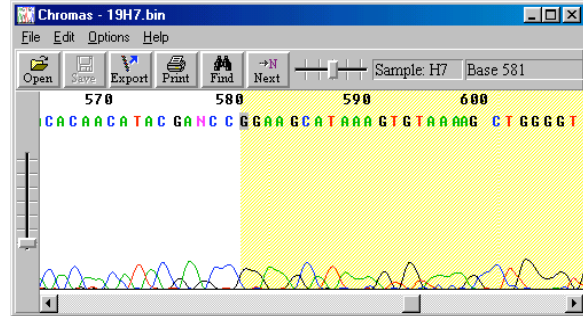
When you obtain a sequence you should proofread it to ensure that all ambiguous sites are correctly called and determine the overall quality of your data. At the very minimum you should proofread and evaluate the quality of your sequence prior to uploading it to NCBI. If your sequence is suboptimal, you want to redo the sequencing or have both a forward and reverse sequencing reaction done and compare your two results to alleviate any possible "miscalls".

Assignment: Viewing your Sequence and Analyses

Open the Finch TV program. (Finch TV folder is on desktop, open it and hit the icon that resembles a TV). You will be directed to “drag” a sequencing file into the “screen”. Drag in your sequencing file. **The first step** is to examine your sequence from beginning to end by using the horizontal and vertical scroll bars. Good sequence generally begins roughly around base 20, and is represented by tall distinct peaks that have little overlap. Poor reactions will result in low or multiple peaks as illustrated below (very typical of beginning or end of a reaction).



Beginning of Sequence

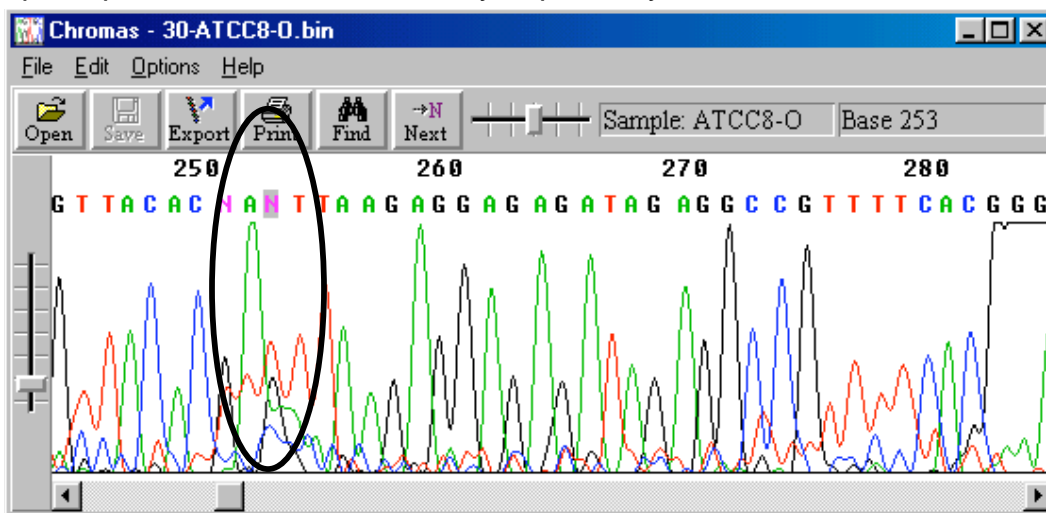


End of sequence

At what base (indicated by the numbers at the top of the sequence) does your sequence become of high quality? (Note: some will not be of high quality anywhere, if this is the case get another sequence or compare a labmate's)

How long is your sequence? Does the quality deteriorate towards the end? Do you find a stretch of AAA's at the end? (These AAA's are added by the Taq polymerase and are non-template derived.)

The second step is to check your sequence for ambiguity by searching for places where “N” is the base call instead of an A, C, G, or T. For example, in the case shown below, the correct nucleotide was likely a T. Do you find any instances where the N designation might not be appropriate? Also examine areas where two bases are superimposed on each other. Can you pick only one?



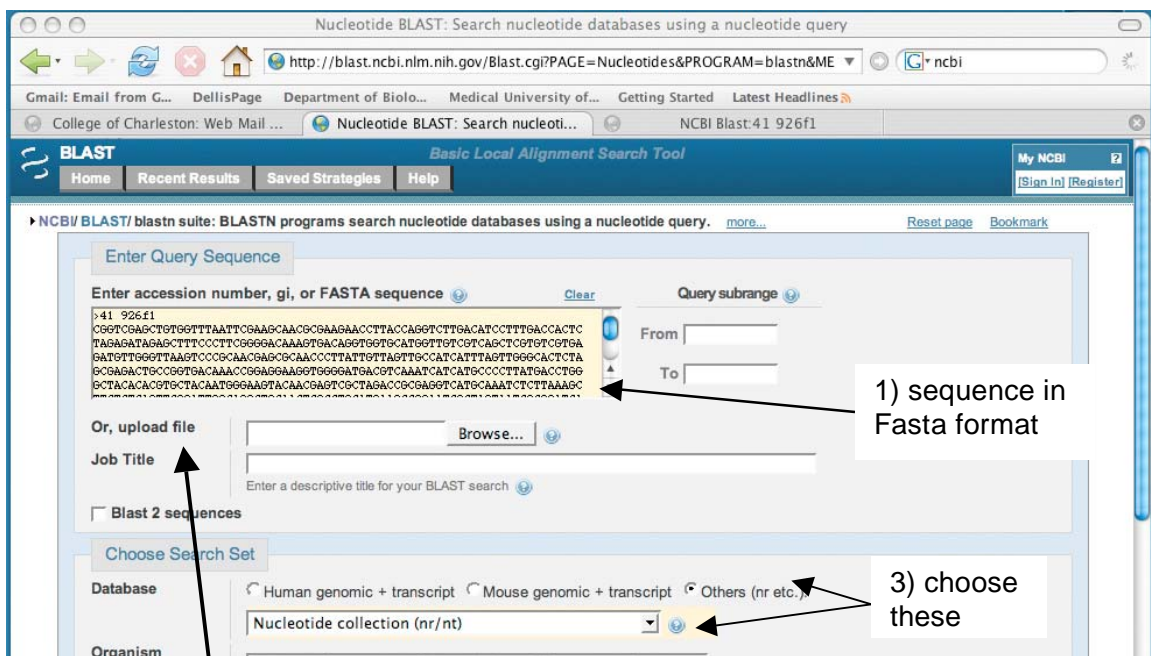
Look at the overall quality of the sequence, do you find that you have nice peaks that are easy to score? Does it appear that more than one product was sequenced?

Third step: determining homology. In other words, is your sequence similar to any other published sequences and if so, to what degree? This can be accomplished using BLAST, a program supported by the National Center for Biotechnology Information (NCBI).

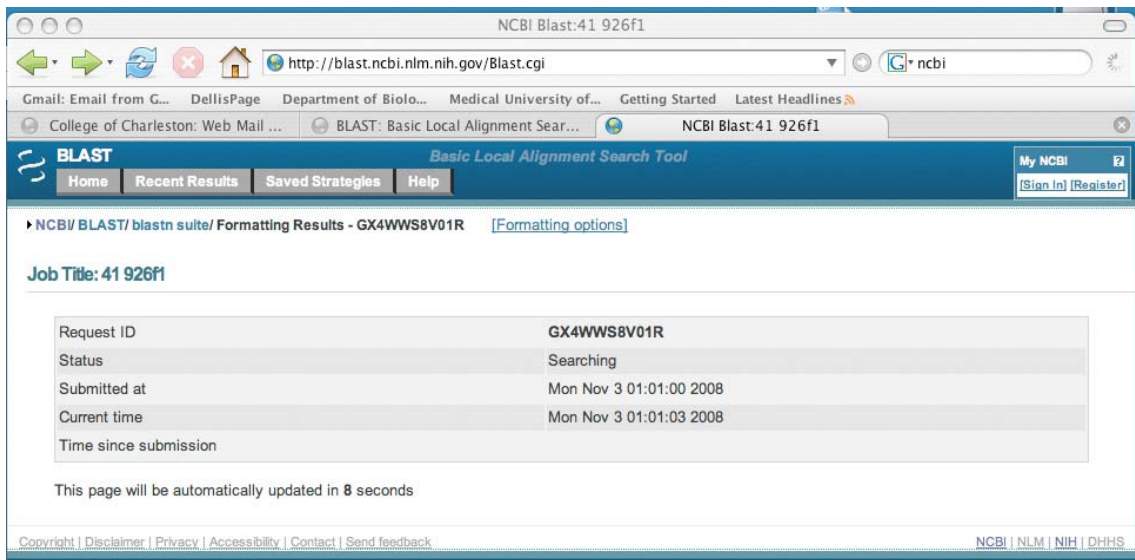
In Finch TV, select from the File dropdown menu the command “export FASTA sequence” and save the export onto the desktop. Open the exported file (you may need to open this up in the notepad program). You should see your sequence as a string of letters, copy the sequence and do a BLAST search at:

<http://www.ncbi.nlm.nih.gov/BLAST/cgi>

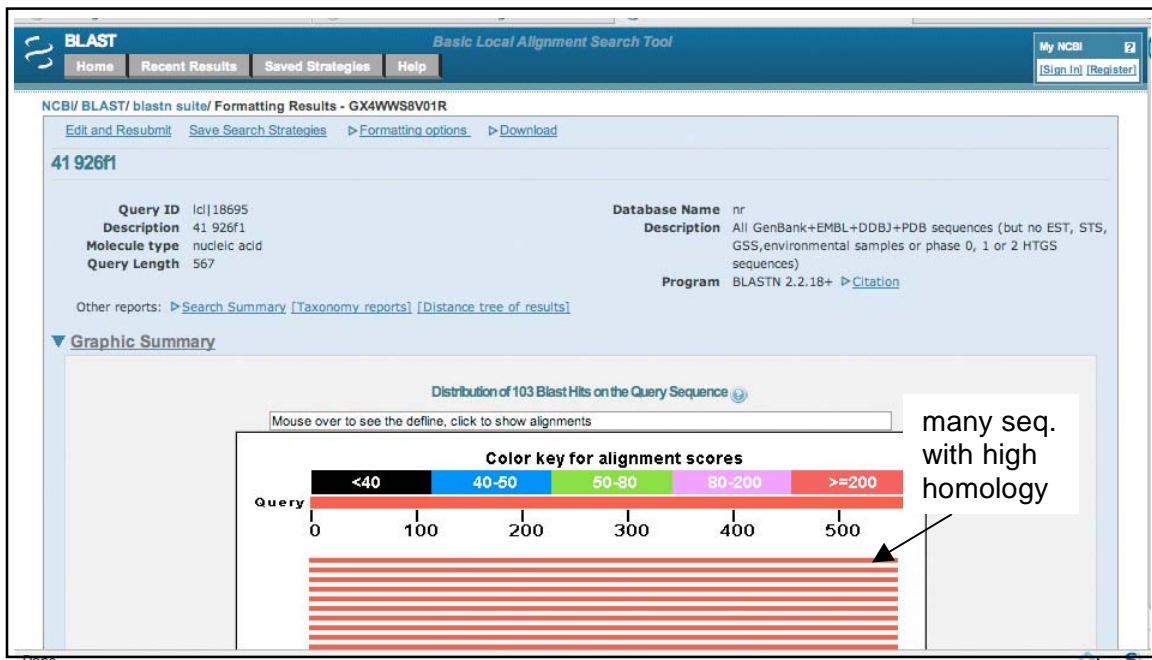
In the Basic Blast (left) section of this page, select “nucleotide BLAST.” Paste your sequence in the “Enter Query Sequence” box. In the “Choose Search Set” database, select “other” which includes all bacterial sequences. The drop-down menu should now say “Nucleotide collection (nr/nt).” You do not have to change any other parameters.



Click the “Blast!” button at the bottom to submit your sequence data.



This screen will come up next. Finally (sometimes after a lengthy wait), a new window will appear showing any “hits” your sequence made. The results will be color coded and annotated so that you can either click on the horizontal homology bar(s) or you can simply scroll down the page.



The bars show what places along your sequence are similar to other published sequences; the colors indicate how many bases were involved in homology determination. The scores indicate how much homology occurred for particular hits. Clicking on a score or bar in the graphic representation will take you to the section of the results page that describes a sequence that is similar to yours.

Clicking on a “gi” link at the beginning of any line will take you to the GenBank accession page for a sequence showing similarity to yours. There you can find a wealth of information about the published sequence to which yours showed some homology.

Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
EU557008.1	Uncultured bacterium clone C56 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU557006.1	Uncultured bacterium clone C59 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU557004.1	Uncultured bacterium clone C62 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU557001.1	Uncultured bacterium clone C66 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU557000.1	Uncultured bacterium clone C72 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU556999.1	Uncultured bacterium clone C75 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU556998.1	Uncultured bacterium clone C80 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU556996.1	Uncultured bacterium clone C99 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EU285587.1	Enterococcus faecalis strain C19315led5A 16S ribosomal RNA gene, partial s	946	946	98%	0.0	97%	
EU547775.1	Enterococcus faecalis strain IJ-07 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
AB362599.1	Enterococcus faecalis gene for 16S rRNA, partial sequence, strain: NRIC 011	946	946	98%	0.0	97%	
EF653454.1	Enterococcus faecalis strain 47/3 16S ribosomal RNA gene, partial sequence	946	946	98%	0.0	97%	
EF608536.1	Uncultured bacterium clone PCD-8 16S ribosomal RNA gene, partial sequenc	946	946	98%	0.0	97%	
AM607463.1	Uncultured bacterium partial 16S rRNA gene, isolate BF0001D078	946	946	98%	0.0	97%	

Fourth step: examine the proposed identities of your sequence. Ignore any “uncultured bacterium” because it does not designate an identity. Look for high query coverage, low E value (chance of random match) and high max identity scores. In this case, many instances of *Enterococcus faecalis* came up.

Selecting the first *Enterococcus* sequence outlined in the box in the above figure (genbank# EU285587.1) will take you to the comparison between your query and the match. The majority of discrepancies are sequence gaps, with a few misalignments.

```
>|_ab|EU285587.1| Enterococcus faecalis strain C19315led5A 16S ribosomal RNA gene,
partial sequence
Length=1456

Score = 946 bits (512), Expect = 0.0
Identities = 550/566 (97%), Gaps = 12/566 (2%)
Strand=Plus/Plus

Query 1      CCGTTCGAGC-TGTGGTTAATTGCAAGCAACGCGAAGAACCTTACCAGGTTTGGACATCC 59
Sbjct 893     CCGTGGAGCATGTGGTTAATTGCAAGCAACGCGAAGAACCTTACCAGGTTTGGACATCC 952

Query 60     TTTGACCACTCTAGAGATAGAGCTTTCCTTCGGGGACAAAGTGACAGTGGTGCATGGT 119
Sbjct 953     TTTGACCACTCTAGAGATAGAGCTTTCCTTCGGGGACAAAGTGACAGTGGTGCATGGT 1012

Query 120    TGTCTGACGCTCGTGTGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTATT 179
Sbjct 1013    TGTCTGACGCTCGTGTGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTATT 1072

Query 180    GTTAGTTGCCATCATTTAGTTGGGCACTTAGCGAGACTGCCGGTGACAAACCGGAGGAA 239
Sbjct 1073    GTTAGTTGCCATCATTTAGTTGGGCACTTAGCGAGACTGCCGGTGACAAACCGGAGGAA 1132

Query 240    GGTGGGGATGACCTCAAATCATCATGCCCTTATGACCTGGGCTACACAGTGTACAAT 299
Sbjct 1133    GGTGGGGATGACCTCAAATCATCATGCCCTTATGACCTGGGCTACACAGTGTACAAT 1192

Query 300    GGGAAGTACAACGAGTCGCTAGACCCGAGGTCATGCAAACTCTTAAAGCTTCTCTCAG 359
Sbjct 1193    GGGAAGTACAACGAGTCGCTAGACCCGAGGTCATGCAAACTCTTAAAGCTTCTCTCAG 1252

Query 360    TTCGGATTGCGAGGCTGCAACTCGCCTGCATGAAGCCGGAATCGTAGTAATCGCGGATC 419
Sbjct 1253    TTCGGATTG-CAGGTCGCAACTCGCCTGCATGAAGCCGGAATCGTAGTAATCGCGGATC 1311

Query 420    AGCACGCCCGGTTGAATACGTTGCCCGGGCCCTTGTACACCCCGCTCACACCACGAGA 479
Sbjct 1312    AGCACGCCCGGTTGAATACGTTCCCGGG-CCTTGTACACACC-CCCGCTCACACCACGAGA 1370

Query 480    GTTTGTAACACCAGTCCG-GAGGTACCTTTT-GGAGC-A-CGCCTTAGTGG-AT 534
Sbjct 1371    GTTTGTAACACCAGTCCGTTGAGGTAACCTTTTGGAGCCAGCCGCTAAGGTGGGAT 1430

Query 535    AGATGAT-GGGTGA-GTTC-TAACA 557
Sbjct 1431    AGATGATGGGGTGAAGT-CCTAACA 1455
```

one misalignment

several sequence gaps

Fifth step: now go back to YOUR sequence, check these areas, and decide if you would change anything. If you do change anything on your chromatogram, then go back and re-BLAST your sequence. Did any of the identity values change because of your corrections? What are the old and new values?

How does your sequence compare to its best match(es)? Fill out the BLAST results in the Table, then access the Ribosomal Database.

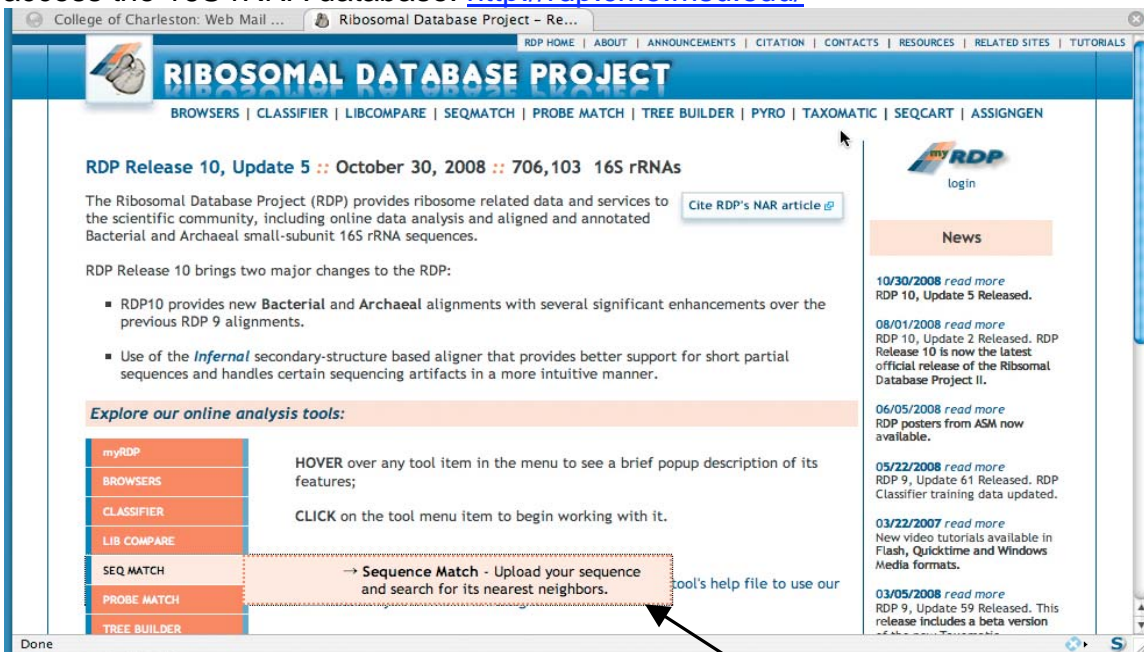
BLAST RESULTS

	NCBI Best match (<i>Genus species</i>)	% Identity	Organization where sequence was uploaded
Your Sequence			
Alternate #1			
Alternate #2			

RDB RESULTS

	RDB Best match (<i>Genus species</i>)	% Identity	Organization where sequence was uploaded
Your Sequence			
Alternate #1			
Alternate #2			

To access the 16S rRNA database: <http://rdp.cme.msu.edu/>



Choose "Sequence Match"

Here is the start page:

Seqmatch - Start [\[help\]](#)

Did you know you can select sequences from *myRDP* and Hierarchy Browser to do seqmatch? Percent identity scores will be reported for aligned sequences (limited to 2000).

Please enter your sequences:

Choose a file to upload: ← you can upload your ".seq" file from the desktop

Cut and paste sequence(s) (in Fasta, GenBank, or EMBL format):

```
>41_926f1
CGGTCGAGCTGTGGTTAATTGGAAGCAACGGAAGAACCCTACCAGGCTTT
TAGAGATAGAGCTTTCCTTCGGGGACAAAGTGACAGGTGGTGCATGGTTGT
GATGTTGGGTTAAGTCCCGCAACGAGCGCAACCTTATTGTTAGTTGCCATC
CGGAGACTGCCGGTGACAAACCGGAGGAGGTGGGGATGACGTCAAATCATC
GCTACACACCTGCTACAAATGGGAAGTACAACGAGTCGCTAGACCGCGAGGTC
TTCTCTCAGTTCCGATGGCAGGCTGCAACTCGCCTGCATGAAGCCGGAATC
GCACCGCGGGTGAATACGTTGCCGGGGCCTTGTACACACCGCCGTCACAC
CCGAAGTCGGGAGGTACCCTTTGGAGCACCGCCTTAGGTGGATAGATGATG
CACAAAAN
```

← here is your sequence in Fasta format

Strain: <input type="radio"/> Type <input type="radio"/> Non Type <input checked="" type="radio"/> Both	<input type="button" value="Submit"/> <input type="button" value="Reset"/>
Source: <input type="radio"/> Uncultured <input type="radio"/> Isolates <input checked="" type="radio"/> Both	
Size: <input checked="" type="radio"/> ≥1200 <input type="radio"/> <1200 <input type="radio"/> Both	
Quality: <input checked="" type="radio"/> Good <input type="radio"/> Suspect <input type="radio"/> Both	
Taxonomy: <input checked="" type="radio"/> Nomenclatural <input type="radio"/> NCBI	
KNN matches: <input type="text" value="20"/>	

Note: Javascript must be enabled on your browser to use this RDP tool

← Paste in sequence. Click "Submit"

Options

Strain: Type strain information is provided by [bacterial taxonomy](#). *Hint:* Type strains link taxonomy with phylogeny. Include type strain sequences in your analysis to provide documented landmarks.

Source: View only environmental (uncultured) sequences, only sequences from individual isolates, or both. Source classification is based on sequence annotation and the [NCBI taxonomy](#).

Size: View only near-full-length sequences (≥1200 bases), short partials, or both.

Quality: View only good quality sequences, suspect quality sequences, or both. Sequences were flagged as suspect quality using Pintail. [\[more quality detail\]](#)

Taxonomy: View sequences placed into a new phylogenetically consistent higher-order [bacterial taxonomy](#) overlaid on the 16S rRNA classification. For the nomenclatural taxonomy, a set of well characterized (vetted) sequences was provided by these workers. Other sequences were placed into this scheme using the [RDP Naïve Bayesian classifier](#).

KNN matches: Number of matches displayed per sequence, also number used to classify queries by unanimous vote.

Next, the status page comes up:

Seqmatch :: Query Sequences Status

Status: running

Current Time: Mon Nov 03 02:13:40 EST 2008

Progress: 33% completed

refresh cancel

Finally, the results appear:

Select All Match Hits to seqCART

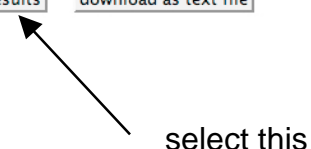
Display depth: Autc ▾

Lineage (click node to return it to hierarchy view):

Hierarchy View:

no rank Root (1) (query sequences) [show printer friendly results](#) [download as text file](#) [options]

- domain Bacteria (1)
 - phylum Firmicutes (1)
 - class "Bacilli" (1)
 - order "Lactobacillales" (1)
 - family "Enterococcaceae" (1)
 - genus Enterococcus (1)
 - 41 [\[view selectable matches\]](#)

 select this

Data Set Options:

This is the bottom half of the screen. You can try different parameters if you want.

Data Set Options:

Strain:	<input type="radio"/> Type	<input type="radio"/> Non Type	<input checked="" type="radio"/> Both
Source:	<input type="radio"/> Uncultured	<input type="radio"/> Isolates	<input checked="" type="radio"/> Both
Size:	<input checked="" type="radio"/> ≥1200	<input type="radio"/> <1200	<input type="radio"/> Both
Quality:	<input checked="" type="radio"/> good	<input type="radio"/> Suspect	<input type="radio"/> Both
KNN matches:	20 ▾		

[Refresh](#)


Strain: Type strain information is provided by bacterial taxonomy. *Hint:* Type strains link taxonomy with phylogeny. Include type strain sequences in your analysis to provide documented landmarks.

Source: View only environmental (uncultured) sequences, only sequences from individual isolates, or both. Source classification is based on sequence annotation and the NCBI taxonomy.

Size: View only near-full-length sequences (≥1200 bases), short partials, or both.

Quality: View only good quality sequences, suspect quality sequences, or both. Sequences were flagged (*) as suspect quality using Pintail. [\[more quality detail\]](#)

KNN matches: Number of matches displayed per sequence, also number used to classify queries by unanimous vote.



Seqmatch: version 3
RDP Data: release 10.5
Data Set: both type and non-type strains, both environmental (uncultured) sequences and isolates, near-full-length sequences (≥ 1200 bases), good quality sequences
Comments: 286680 sequences were included in the search
The screening was based on 7-base oligomers
Query Submit Date: Mon Nov 03 02:13:40 EST 2008

Match hit format: short ID, orientation, similarity score, S_{ab} score, unique common oligomers and sequence full name. More help is available.

Lineage:
Results for Query Sequence: 41, 553 unique oligos

```

no rank Root (20) (match sequences)
  domain Bacteria (20)
    phylum Firmicutes (20)
      class "Bacilli" (20)
        order "Lactobacillales" (20)
          family "Enterococaceae" (20)
            genus Enterococcus (20)
              5000001976 not_calculated 0.863 1466 Enterococcus faecalis; LMG 7937; AJ301831
              5000012997 not_calculated 0.863 1417 Enterococcus faecalis; AF039902
              5000261697 not_calculated 0.863 1439 Enterococcus faecalis; EC-12; AB154827
              5000375664 not_calculated 0.863 1486 Enterococcus faecalis; SLS; AY692453
              5000394239 not_calculated 0.863 1430 Enterococcus faecalis; DJ1; AF447490
              5000409542 not_calculated 0.863 1426 Enterococcus faecalis; NAB11; AY395018
              5000428402 not_calculated 0.863 1434 Enterococcus faecalis; JH2-2; AF070224
              5000430944 not_calculated 0.863 1439 Enterococcus faecalis; RO90; AF515223
              5000468458 not_calculated 0.863 1470 Enterococcus faecalis; SFL; AY850358
              5000485949 not_calculated 0.863 1402 Enterococcus faecalis; C13115; AY550919
              5000495530 not_calculated 0.863 1435 Enterococcus faecalis; K-4; AB036835
              5000514055 not_calculated 0.863 1437 Enterococcus faecalis; ChDC YE2; AY942558
              5000536332 not_calculated 0.863 1460 uncultured bacterium; rRNA255; AY959028
              5000617875 not_calculated 0.863 1473 Enterococcus faecalis; mother C4; 7C4; AM157433
              5000619967 not_calculated 0.863 1445 Enterococcus faecalis; D3; DQ239694
              5000628286 not_calculated 0.863 1440 Enterococcus faecalis V583; AEO16830
              5000628288 not_calculated 0.863 1441 Enterococcus faecalis V583; AEO16830
              5000628291 not_calculated 0.863 1441 Enterococcus faecalis V583; AEO16830
              5000710240 not_calculated 0.866 1401 uncultured bacterium; aaa47a08; DQB18629
              5000837260 not_calculated 0.864 1409 uncultured bacterium; P7D82-658; EF509508

```

:: Result Format

Each match result line contains six elements, from left to right,

- A short ID used to uniquely identify the RDP sequence. A click will return the simple entry, including the sequence.
- The orientation of the query sequence when the match is performed. "-" means the query sequence has been reverse-complemented. A top match hit with "-" orientation usually indicates the query sequence is a minus strand.
- A similarity score. SeqMatch reports the percent sequence identity over all pairwise comparable positions when run with aligned myRDP sequences. (Comparable positions are aligned positions containing a base in both sequences). Note that the rank order may differ between S_{ab} and pairwise identity scores, but the top 20 S_{ab} scores will contain the closest sequence by pairwise identity about 95% of the time (Cole et al). If two sequences do not overlap, the similarity between these two sequences will be displayed as "?".
- A seqmatch score (S_{ab}). These are the number of (unique) 7-base oligomers shared between your sequence and a given RDP sequence divided by the lowest number of unique oligos in either of the two sequences.
- The number of uniquely occurring oligomers within a given sequence (Olis). If the same oligomer occurs more than once then they are counted only once; thus this number only approximately reflects the sequence length. Counting only unique oligos compensates somewhat for composition bias (for example, inserts tend to be GC-rich and it becomes very likely that the same GC-rich oligos occur several times; by counting these only once, this artifact becomes less severe).
- Full name. The definition line from the RDP distribution, often the same as Genus/species/string name and accno.

Did results from NCBI and Ribosomal Database at the Genus species level come out the same? Fill out the Tables on page 6.

If there was a discrepancy between NCBI and the Ribosomal Database project which would you suspect is generally more accurate? Explain why.

NOTE: ANSWER THE QUESTIONS ON THE HANDOUT AND ON THE LAST PAGES (PAGES 11-12). KEEP THE HANDOUT AND TURN IN THE LAST PAGE.

REFERENCES

FinchTV; <http://www.geospiza.com/finchtv.html>

National Center for Biotechnology Information; <http://www.ncbi.nlm.nih.gov/>

Ribosomal Database Project; <http://rdp.cme.msu.edu/>

Your Name: _____

Date: _____

Sequence: _____ Bacteria Identity: _____ % Identity: _____

Answer the questions in your handout and on this sheet.

1a. At what base (indicated by the numbers at the top of the sequence) does your sequence become of high quality?

1b. How long is your sequence? Does the quality deteriorate towards the end? Do you find a stretch of AAA's at the end?

2a. Do you find any instances where the N designation might not be appropriate? If so, give the nucleotide number of the "N" base and the base that you designated to replace N.

2b. Look at the overall quality of the sequence, do you find that you have nice peaks that are easy to score? Does it appear that more than one product was sequenced?

5a. check the areas that contain mismatches or deletions, and decide if you would change anything. If you do change anything on your chromatogram, then go back and re-BLAST your sequence. Did any of the identity values change because of your corrections? What are the old and new values?

5b. How does your sequence compare to its best match(es)? Fill out the BLAST results in the Table, then access the Ribosomal Database.

BLAST RESULTS

	NCBI Best match (<i>Genus species</i>)	% Identity	Organization where sequence was uploaded
Your Sequence			
Alternate #1			
Alternate #2			

RDB RESULTS

	RDB Best match (<i>Genus species</i>)	% Identity	Organization where sequence was uploaded
Your Sequence			
Alternate #1			
Alternate #2			

5c. Did results from NCBI and Ribosomal Database at the Genus species level come out the same? Fill out the Tables on page 7.

5d. If there was a discrepancy between NCBI and the Ribosomal Database project which would you suspect is generally more accurate? Explain why.