Environmental Microbiology: Bioinformatic exercises II

Identification, classification and phylogenetic tree construction of 16S rRNA sequences

26.01.2015

II Molecular (Culture-Independent) Analyses of Microbial Communities

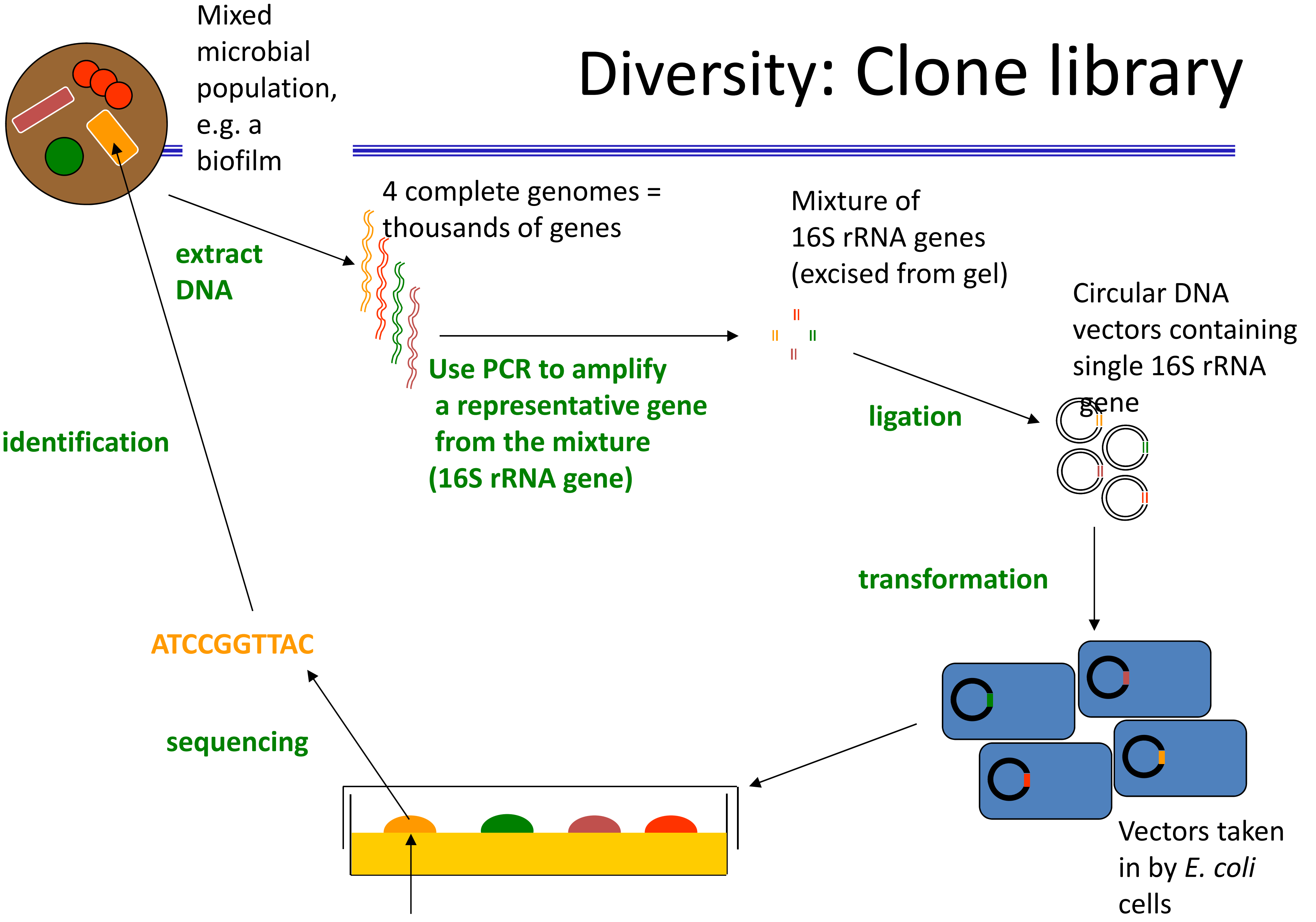
“Molecular Microbial Ecology“

IIC Linking specific genes to specific organisms using PCR

Molecular analysis of diversity based on 16S rDNA

- **Cloning** – generation of a clone library.
  - Enables the study of 16S genes in isolation.
- **Denaturing gradient gel electrophoresis** (DGGE).
  - Separates fragments of the 16S DNA, which have a different sequence.
- **Terminal restriction fragment length polymorphism** (T-RFLP)
  - Digestion of PCR products (16S DNA) resulting in different Fragments
- **All** of the above methods depend on **PCR amplification of the target genes** from the environment.

Diversity: Clone library



Mixed microbial population, e.g. a biofilm

extract DNA

4 complete genomes = thousands of genes

Use PCR to amplify a representative gene from the mixture (16S rRNA gene)

Mixture of 16S rRNA genes (excised from gel)

ligation

Circular DNA vectors containing single 16S rRNA gene

transformation

Vectors taken in by *E. coli* cells

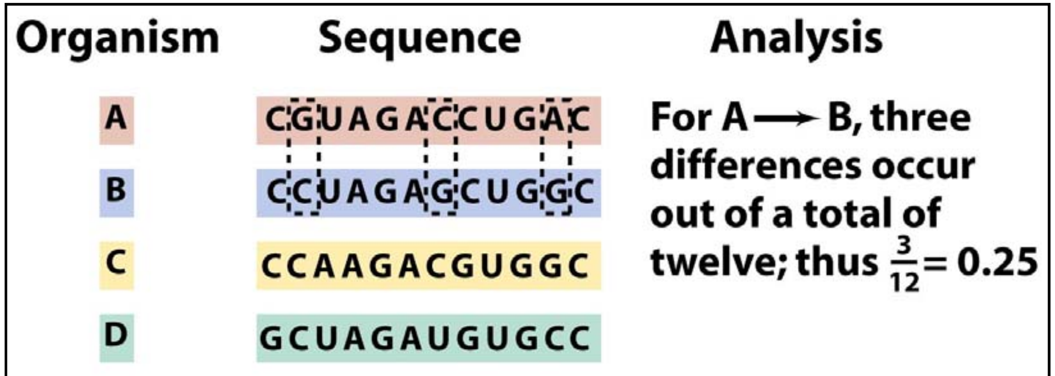
sequencing

ATCCGGTTAC

identification

Each clone (colony) contains 1000s of cells all with the same 16S rRNA gene = 1000s copies 16S rDNA

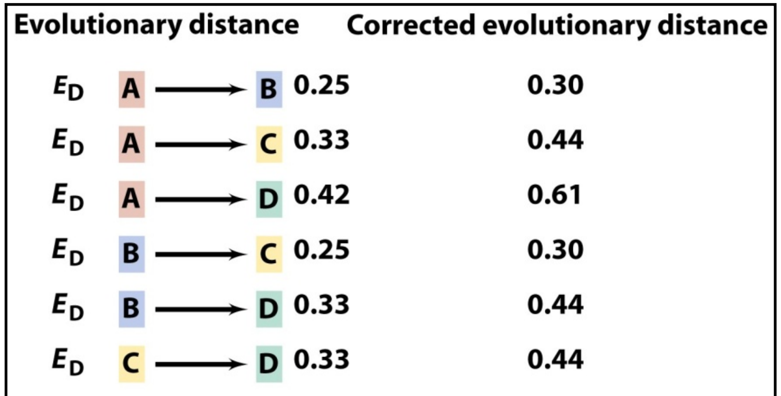
Sequence Analyses & Phylogenetic Tree Construction



Organism	Sequence	Analysis
A	CGUAGACCUGAC	For A ⟶ B, three differences occur out of a total of twelve; thus $\frac{3}{12} = 0.25$
B	CCUAGAGCUGGC	
C	CCAAGACGUGGC	
D	GCUAGAUGUGCC	

Sequence alignment and analysis

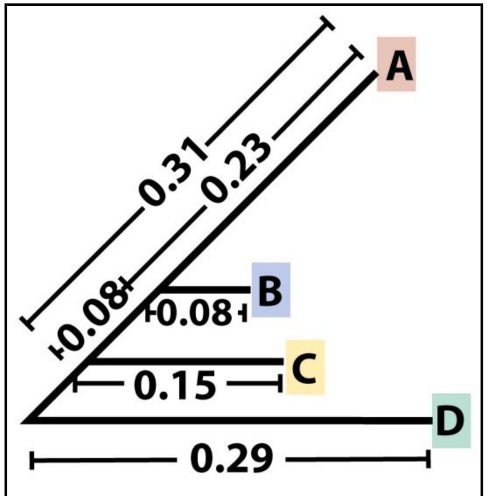
Figure 11-13a Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.



	Evolutionary distance	Corrected evolutionary distance
E_D A ⟶ B	0.25	0.30
E_D A ⟶ C	0.33	0.44
E_D A ⟶ D	0.42	0.61
E_D B ⟶ C	0.25	0.30
E_D B ⟶ D	0.33	0.44
E_D C ⟶ D	0.33	0.44

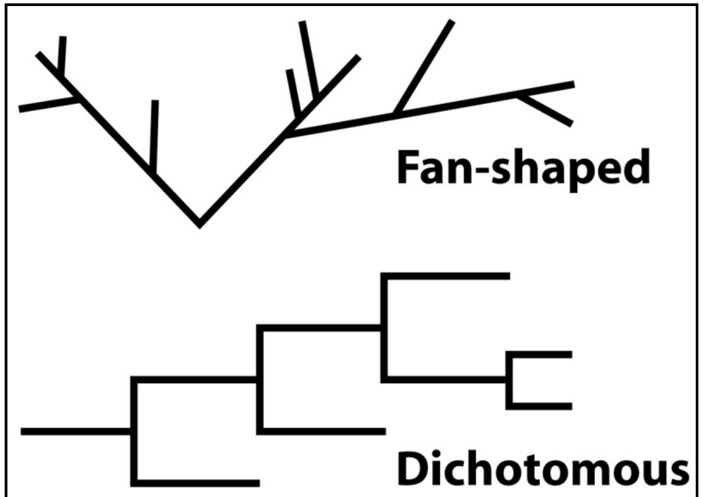
Calculation of evolutionary distance

Figure 11-13b Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.



Phylogenetic tree

Figure 11-13c Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.



Tree topologies

Figure 11-13d Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.

Sequencing results

You find the sequence file (MyExerziseSeq) on your desktop

```
>myExercizeSeq
AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAG
TGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCGCAAGACCAAA
GAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAGCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTG
GTCTGAGAGGATGACCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCT
GATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGA
CGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCG
CACGCAGGCGGTTTGTTAAGTCAGATGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGT
AGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAA
AGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGC
TAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAAACTCAAATGAATTGACGGGGGCCCGCACAAGCGGTGGAGCATGTGGTTT
AATTCGATGCAACGCGAAGAACCTTACCTGGTCTTGACATCCACAGAACTTTCCAGAGATGGATTGGTGCCTTCGGGAACTGTGAGACAGGTGCT
GCATGGCTGTCGTCAGCTCGTGTTGTGAAATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTATCTTTTGTTGCCAGCGGTCCGGCCGGGAAC
TCAAAGGAGACTGCCAGTGATAAACTGGAGGAAGGTGGGGATGACGTCAAGTCATCATGGCCCTTACGACCAGGGCTACACACGTGCTACAATGG
CGCATACAAAGAGAAGCGACCTCGCGAGAGCAAGCGGACCTCATAAAGTGCGTCGTAGTCCGGATTGGAGTCTGCAACTCGACTCCATGAAGTCG
GAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGGAGTGGGTTGCAAAAG
AAGTAGGTAGCTTAACCTTCGGGAGGGCGCTTACCACTTTGTGATTCATGACTGGGGTGAAGTCGTAACAAGGTAACCGTAGGGGAACC
```

What does this sequence encode? Really an rRNA sequence?

Is it the entire coding region?

Topics

- Identification of the sequence using BLAST against GenBank (Nucleotide BLAST at NCBI) and against the 16S rRNA sequence database at NCBI
- Classification of the sequence, identification of closest relatives and phylogenetic tree construction using the the Ribosomal Database Project (RDP) (a specialized database rRNA analyses)
- Aligning sequences using Clustal Omega it the EMBL-EBI server

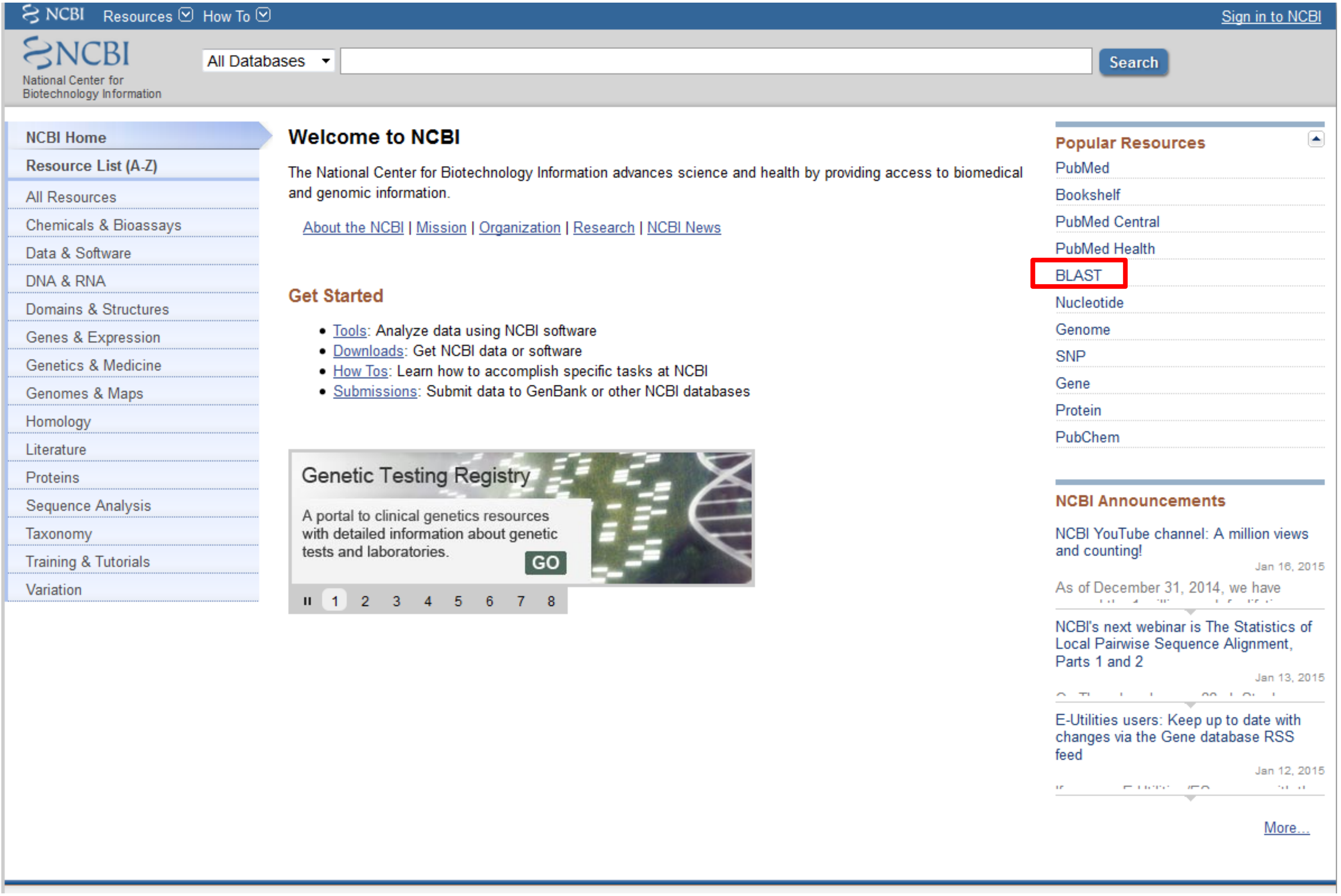
BLAST - **B**asic **L**ocal **A**lignment **S**earch **T**ool

The program

- compares nucleotide or protein sequences to sequence databases
- finds regions of local similarity between sequences
- calculates the statistical significance of matches
- BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Sequence identification

http://www.ncbi.nlm.nih.gov/ → BLAST



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

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genetic Testing Registry

A portal to clinical genetics resources with detailed information about genetic tests and laboratories.

Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements

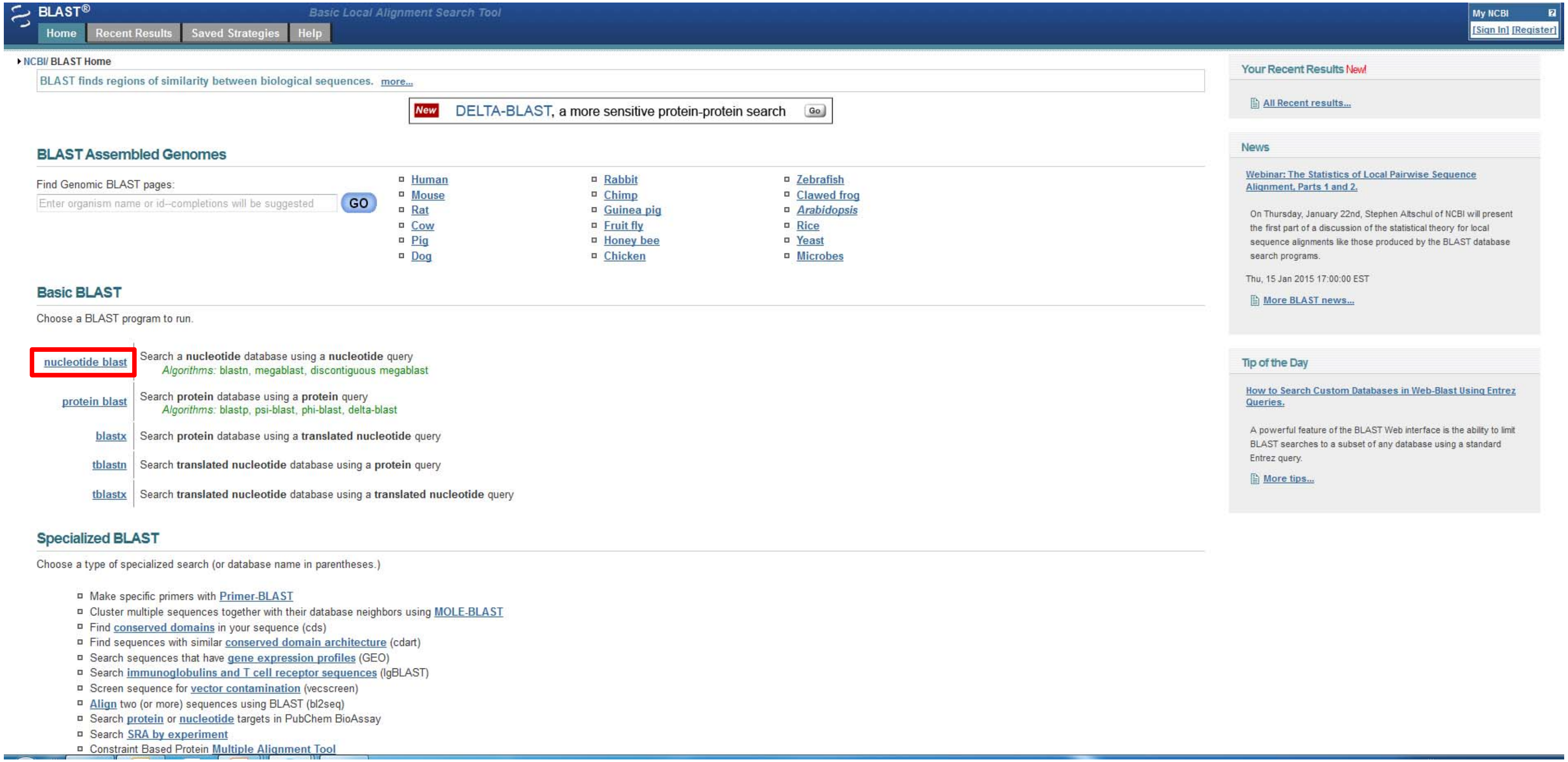
NCBI YouTube channel: A million views and counting!
Jan 16, 2015
As of December 31, 2014, we have

NCBI's next webinar is The Statistics of Local Pairwise Sequence Alignment, Parts 1 and 2
Jan 13, 2015

E-Utilities users: Keep up to date with changes via the Gene database RSS feed
Jan 12, 2015

More...

Nucleotide BLAST
(searching the GenBank database)



NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled Genomes

Find Genomic BLAST pages:

Enter organism name or id--completions will be suggested GO

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

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

Your Recent Results New!

All Recent results...

News

Webinar: The Statistics of Local Pairwise Sequence Alignment, Parts 1 and 2.

On Thursday, January 22nd, Stephen Altschul of NCBI will present the first part of a discussion of the statistical theory for local sequence alignments like those produced by the BLAST database search programs.

Thu, 15 Jan 2015 17:00:00 EST

More BLAST news...

Tip of the Day

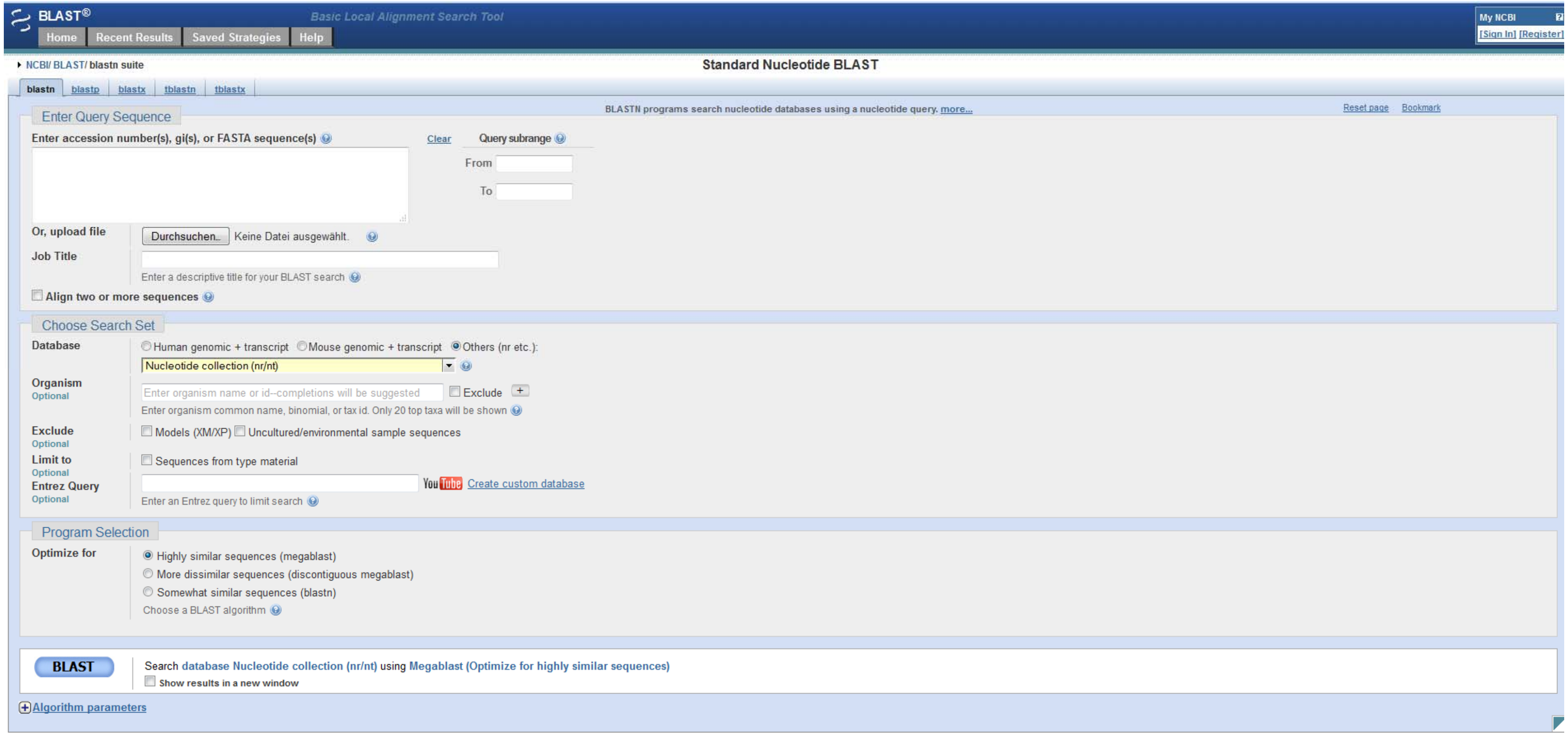
How to Search Custom Databases in Web-Blast Using Entrez Queries.

A powerful feature of the BLAST Web interface is the ability to limit BLAST searches to a subset of any database using a standard Entrez query.

More tips...

Nucleotide BLAST

Copy/paste query sequence (>mySeq) into the query sequence entry field → BLAST



NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file: Keine Datei ausgewählt.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

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

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

Limit to Optional: □ Sequences from type material

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

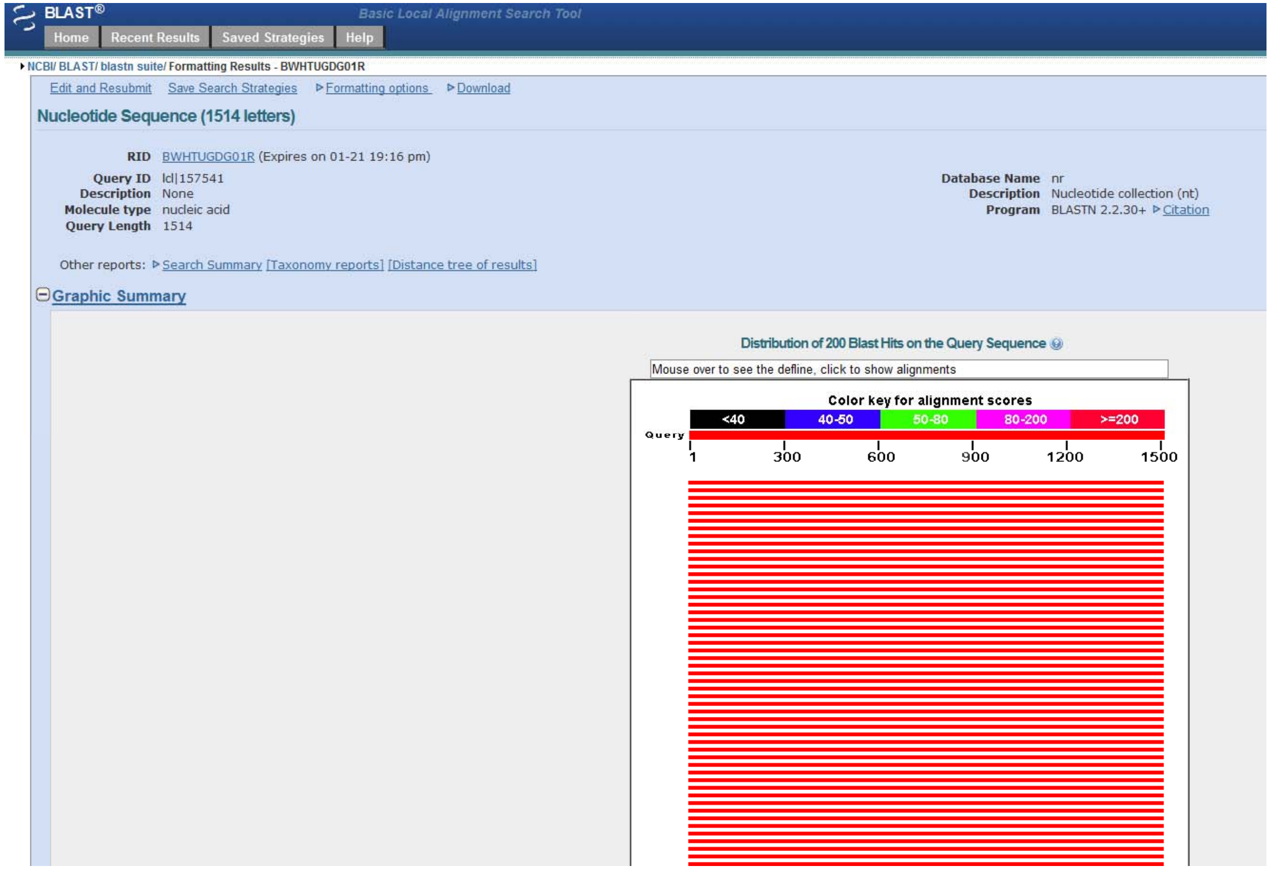
Optimize for:
- ● Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Nucleotide BLAST results



Nucleotide Sequence (1514 letters)

RID BWHTUGDG01R (Expires on 01-21 19:16 pm)

Query ID lcl|157541
Description None
Molecule type nucleic acid
Query Length 1514

Database Name nr
Description Nucleotide collection (nt)
Program BLASTN 2.2.30+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results]

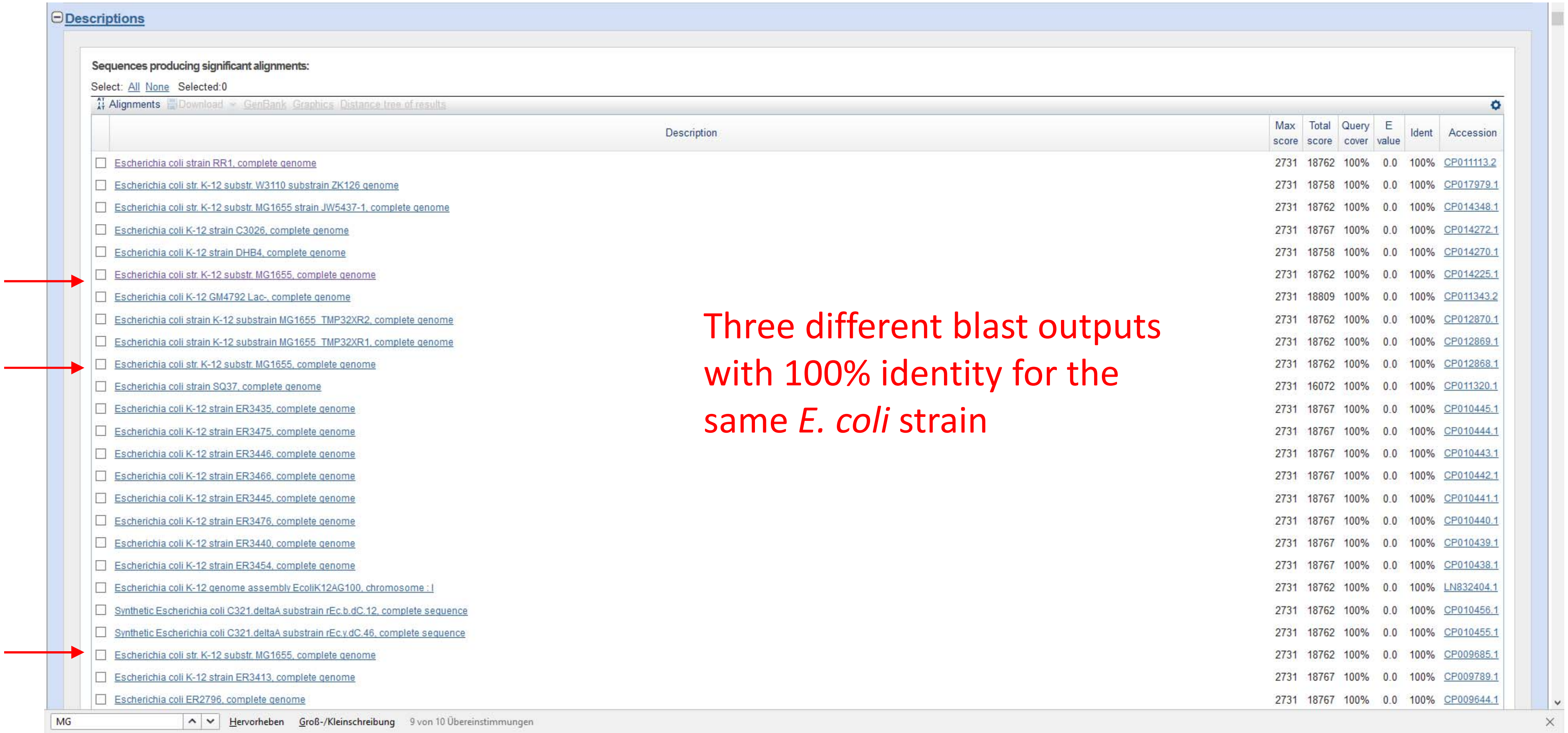
Graphic Summary

Distribution of 200 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

- The graphical display shows the sequences that BLAST was able to align
- Alignment scores are represented on the color bar at the top of the figure, with scores going from low (black) to high (red).
- The numbered line below the color bar represents the amino acid sequence of query sequence
- Below it are various sequences from several databases that were found to align to the query. The precise position of each sequence relative to the query sequence indicates the areas of sequence similarity.

Nucleotide BLAST results



Descriptions

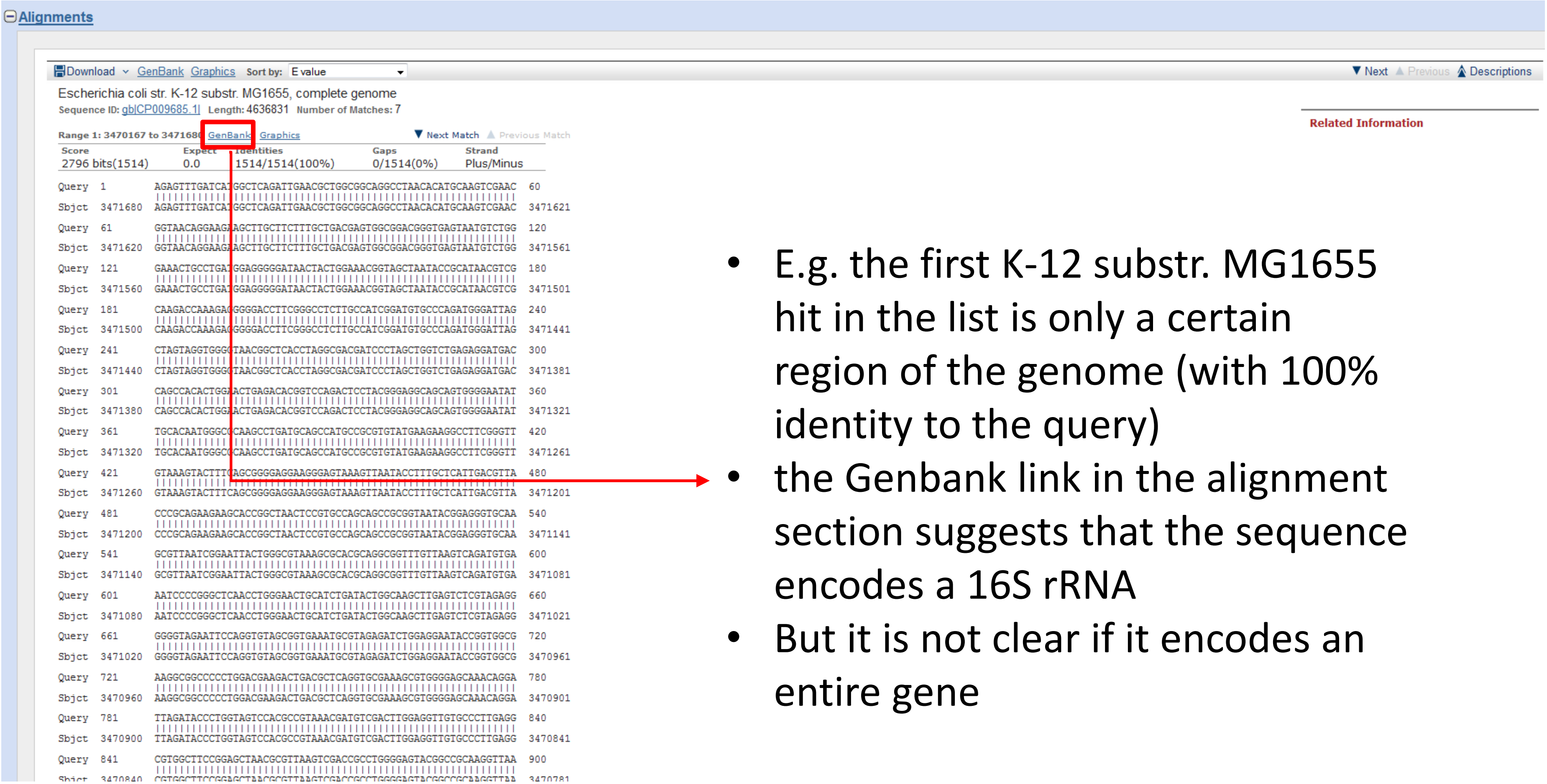
Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	Escherichia coli strain RR1, complete genome	2731	18762	100%	0.0	100%	CP011113.2
☐	Escherichia coli str. K-12 substr. W3110 substrain ZK126 genome	2731	18758	100%	0.0	100%	CP017979.1
☐	Escherichia coli str. K-12 substr. MG1655 strain JW5437-1, complete genome	2731	18762	100%	0.0	100%	CP014348.1
☐	Escherichia coli K-12 strain C3026, complete genome	2731	18767	100%	0.0	100%	CP014272.1
☐	Escherichia coli K-12 strain DHB4, complete genome	2731	18758	100%	0.0	100%	CP014270.1
☐	Escherichia coli str. K-12 substr. MG1655, complete genome	2731	18762	100%	0.0	100%	CP014225.1
☐	Escherichia coli K-12 GM4792 Lac-, complete genome	2731	18809	100%	0.0	100%	CP011343.2
☐	Escherichia coli strain K-12 substrain MG1655_TMP32XR2, complete genome	2731	18762	100%	0.0	100%	CP012870.1
☐	Escherichia coli strain K-12 substrain MG1655_TMP32XR1, complete genome	2731	18762	100%	0.0	100%	CP012869.1
☐	Escherichia coli str. K-12 substr. MG1655, complete genome	2731	18762	100%	0.0	100%	CP012868.1
☐	Escherichia coli strain SQ37, complete genome	2731	16072	100%	0.0	100%	CP011320.1
☐	Escherichia coli K-12 strain ER3435, complete genome	2731	18767	100%	0.0	100%	CP010445.1
☐	Escherichia coli K-12 strain ER3475, complete genome	2731	18767	100%	0.0	100%	CP010444.1
☐	Escherichia coli K-12 strain ER3446, complete genome	2731	18767	100%	0.0	100%	CP010443.1
☐	Escherichia coli K-12 strain ER3466, complete genome	2731	18767	100%	0.0	100%	CP010442.1
☐	Escherichia coli K-12 strain ER3445, complete genome	2731	18767	100%	0.0	100%	CP010441.1
☐	Escherichia coli K-12 strain ER3476, complete genome	2731	18767	100%	0.0	100%	CP010440.1
☐	Escherichia coli K-12 strain ER3440, complete genome	2731	18767	100%	0.0	100%	CP010439.1
☐	Escherichia coli K-12 strain ER3454, complete genome	2731	18767	100%	0.0	100%	CP010438.1
☐	Escherichia coli K-12 genome assembly EcoliK12AG100, chromosome : I	2731	18762	100%	0.0	100%	LN832404.1
☐	Synthetic Escherichia coli C321.deltaA substrain rEc.b.dC.12, complete sequence	2731	18762	100%	0.0	100%	CP010456.1
☐	Synthetic Escherichia coli C321.deltaA substrain rEc.y.dC.46, complete sequence	2731	18762	100%	0.0	100%	CP010455.1
☐	Escherichia coli str. K-12 substr. MG1655, complete genome	2731	18762	100%	0.0	100%	CP009685.1
☐	Escherichia coli K-12 strain ER3413, complete genome	2731	18767	100%	0.0	100%	CP009789.1
☐	Escherichia coli ER2796, complete genome	2731	18767	100%	0.0	100%	CP009644.1

Three different blast outputs with 100% identity for the same *E. coli* strain

Nucleotide BLAST results



Escherichia coli str. K-12 substr. MG1655, complete genome

Sequence ID: gb|CP009685.1| Length: 4636831 Number of Matches: 7

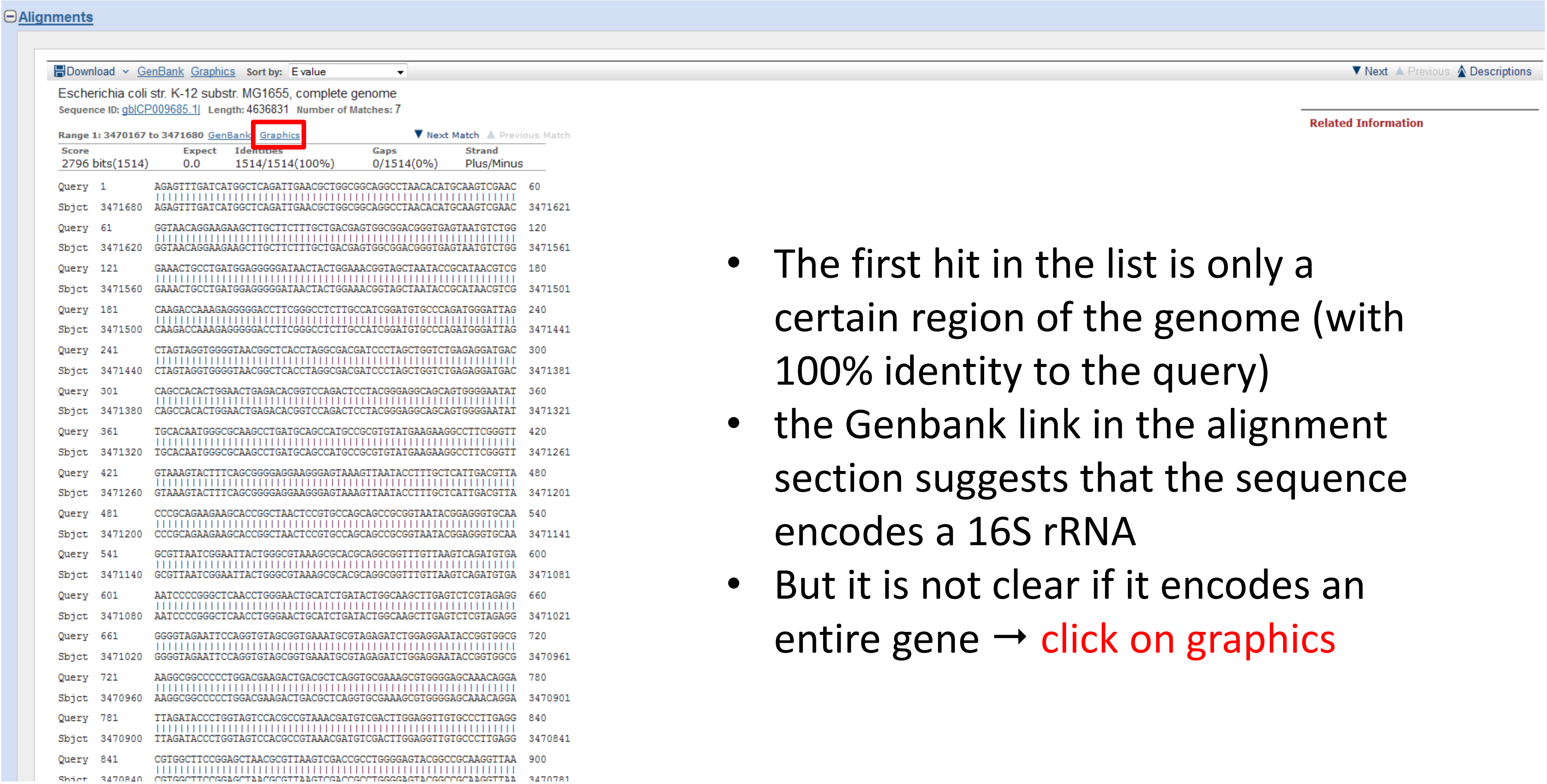
Range 1: 3470167 to 3471680 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
2796 bits(1514)	0.0	1514/1514(100%)	0/1514(0%)	Plus/Minus

```
Query  1        AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471680  AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  3471621

Query  61       GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471620  GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  3471561

Query  121      GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471560  GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  3471501

Query  181      CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471500  CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  3471441

Query  241      CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471440  CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  3471381

Query  301      CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471380  CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  3471321

Query  361      TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471320  TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  3471261

Query  421      GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  480
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471260  GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  3471201

Query  481      CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  540
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471200  CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  3471141

Query  541      GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471140  GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  3471081

Query  601      AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  660
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471080  AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  3471021

Query  661      GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  720
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471020  GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  3470961

Query  721      AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  780
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470960  AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  3470901

Query  781      TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  840
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470900  TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  3470841

Query  841      CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  900
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470840  CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  3470781
```

- E.g. the first K-12 substr. MG1655 hit in the list is only a certain region of the genome (with 100% identity to the query)
- the Genbank link in the alignment section suggests that the sequence encodes a 16S rRNA
- But it is not clear if it encodes an entire gene

Nucleotide BLAST results



⊟ Alignments

Download ⌄ GenBank Graphics Sort by: E value ▾

▼ Next ▲ Previous ▲ Descriptions

Escherichia coli str. K-12 substr. MG1655, complete genome

Sequence ID: gb|CP009685.1| Length: 4636831 Number of Matches: 7

Related Information

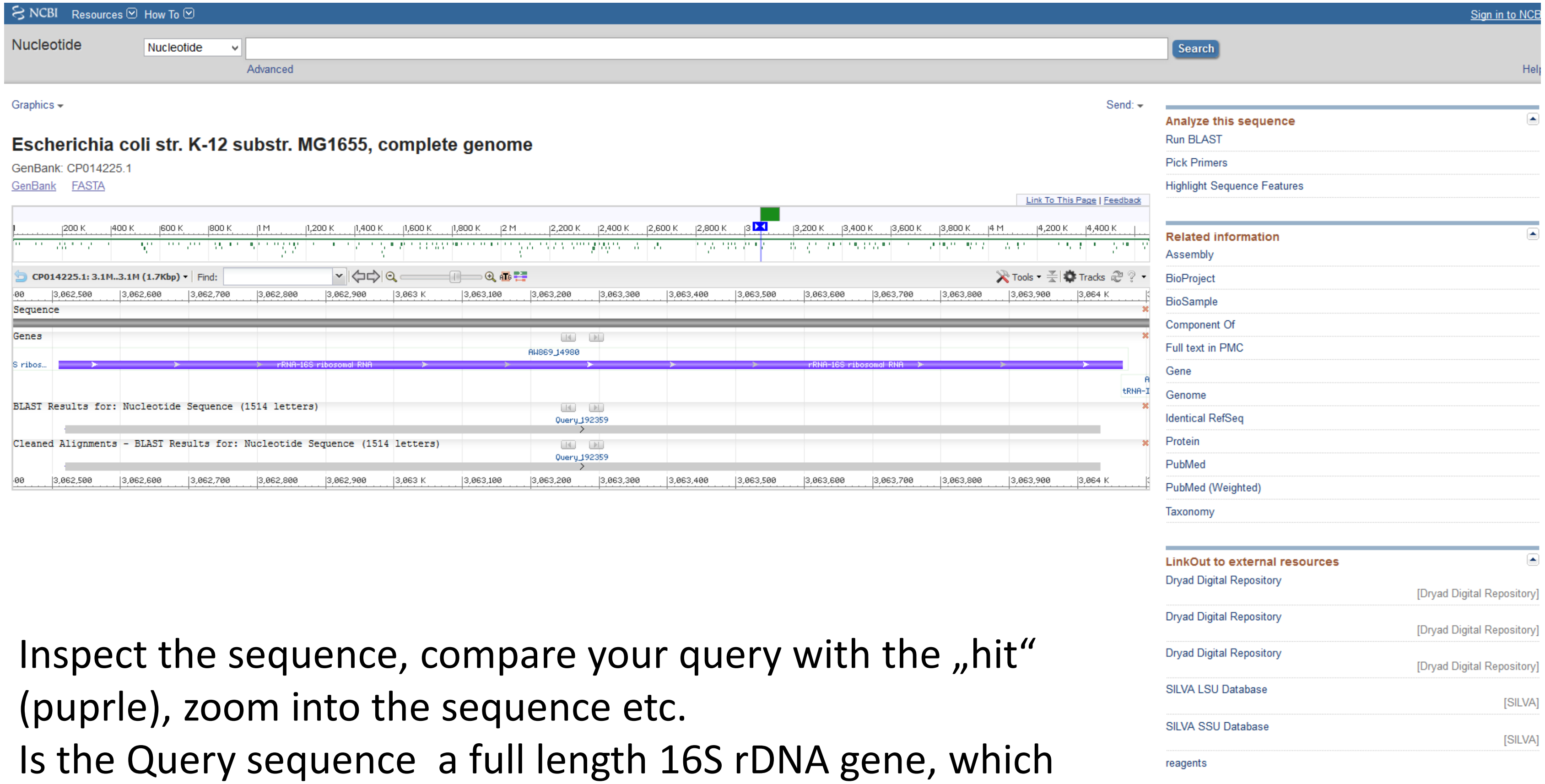
Range 1: 3470167 to 3471680 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
2796 bits(1514)	0.0	1514/1514(100%)	0/1514(0%)	Plus/Minus

```
Query  1        AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471680  AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  3471621

Query  61       GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471620  GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  3471561

Query  121      GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471560  GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  3471501

Query  181      CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471500  CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  3471441

Query  241      CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471440  CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  3471381

Query  301      CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471380  CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  3471321

Query  361      TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471320  TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  3471261

Query  421      GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  480
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471260  GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  3471201

Query  481      CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  540
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471200  CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  3471141

Query  541      GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471140  GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  3471081

Query  601      AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  660
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471080  AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  3471021

Query  661      GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  720
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3471020  GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  3470961

Query  721      AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  780
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470960  AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  3470901

Query  781      TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  840
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470900  TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  3470841

Query  841      CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  900
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3470840  CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  3470781
```

- The first hit in the list is only a certain region of the genome (with 100% identity to the query)
- the Genbank link in the alignment section suggests that the sequence encodes a 16S rRNA
- But it is not clear if it encodes an entire gene → click on graphics

Nucleotide BLAST results



Escherichia coli str. K-12 substr. MG1655, complete genome

GenBank: CP014225.1

GenBank FASTA

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Related information

Assembly

BioProject

BioSample

Component Of

Full text in PMC

Gene

Genome

Identical RefSeq

Protein

PubMed

PubMed (Weighted)

Taxonomy

LinkOut to external resources

Dryad Digital Repository [Dryad Digital Repository]

Dryad Digital Repository [Dryad Digital Repository]

Dryad Digital Repository [Dryad Digital Repository]

SILVA LSU Database [SILVA]

SILVA SSU Database [SILVA]

reagents

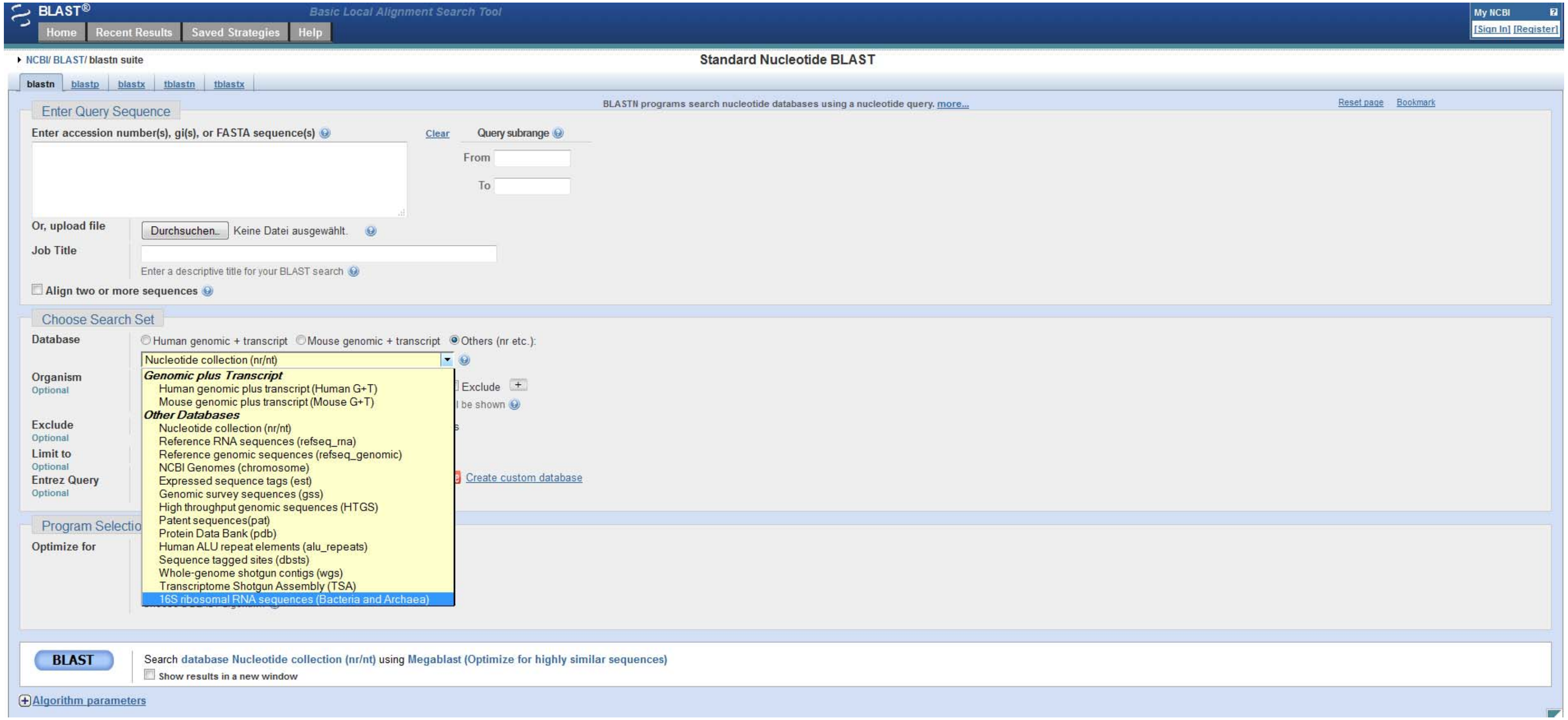
Inspect the sequence, compare your query with the „hit“ (puprle), zoom into the sequence etc.
Is the Query sequence a full length 16S rDNA gene, which parts are missing?

Limitations of GenBank

- GenBank entries can contain
  - Entire genes
  - Portions of genes
  - Many genes
- GenBank entries can be of uneven quality
  - Can be duplicates and/or inaccurate
  - The database is not a selection center
  - All data is treated equally
- GenBank entries are not the final word on particular genes
  - They have no authoritative biological meaning
  - They merely keep track of what was done
- Gene-centric databases are needed to compile everything that is known on a given gene and to correct potential errors

Nucleotide BLAST

Choose the 16S rRNA sequence database in the nucleotide blast databases menu to ensure that the query sequence encodes a 16S rRNA and to find the entire gene



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file: Keine Datei ausgewählt.

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt)

Genomic plus Transcript
- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases
- Nucleotide collection (nr/nt)
- Reference RNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun contigs (wgs)
- Transcriptome Shotgun Assembly (TSA)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

Organism Optional: □ Exclude +

l be shown

Exclude Optional

Limit to Optional

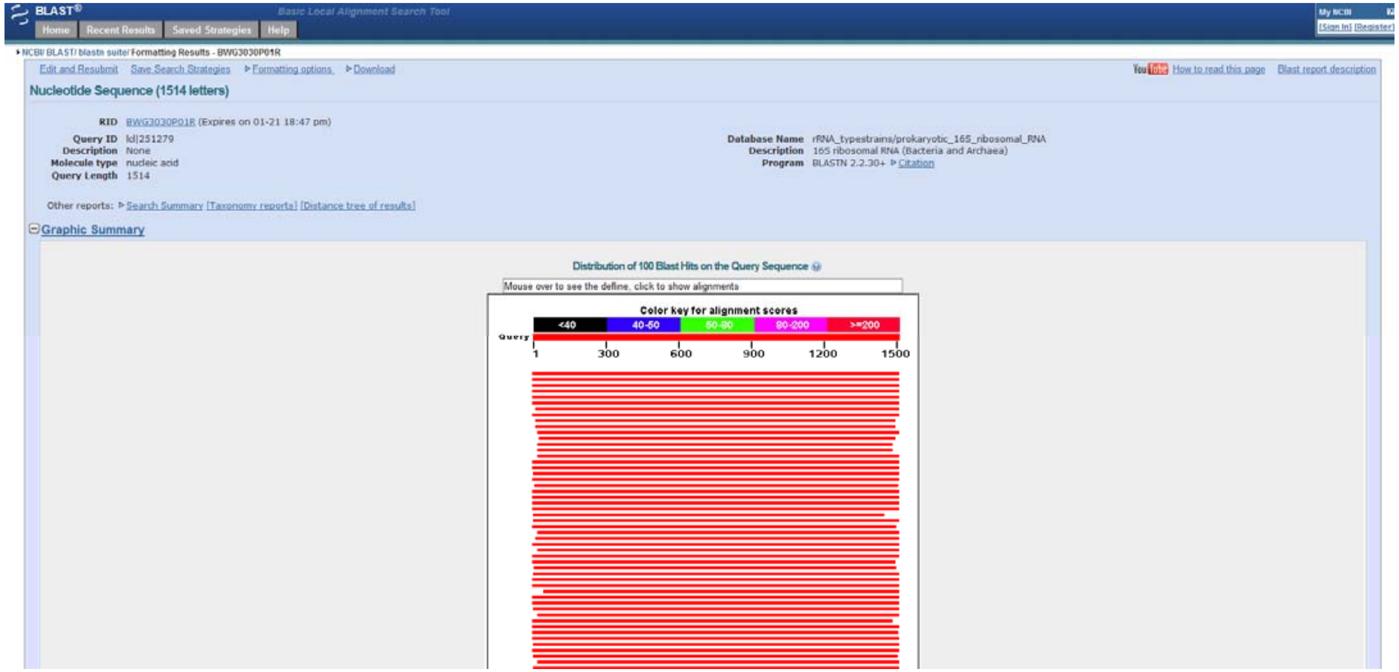
Entrez Query Optional

Optimize for

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

BLAST results



Nucleotide Sequence (1514 letters)

RID BWG3030P01R (Expires on 01-21 18:47 pm)

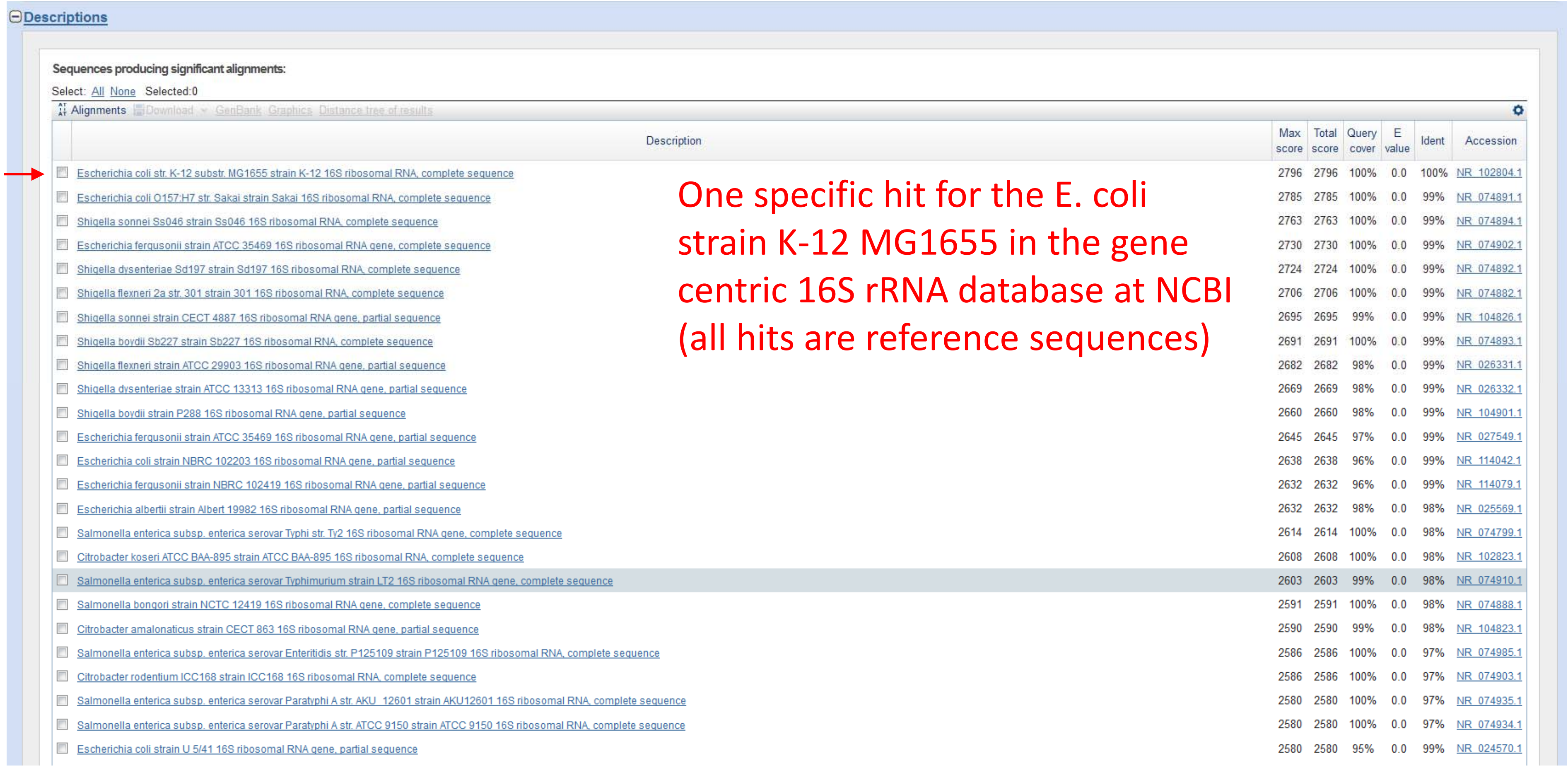
Query ID lcl|251279
Description None
Molecule type nucleic acid
Query Length 1514

Database Name rRNA_typestrains/prokaryotic_16S_ribosomal_RNA
Description 16S ribosomal RNA (Bacteria and Archaea)
Program BLASTN 2.2.30+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results]

Mouse over to see the define, click to show alignments

BLAST results



Sequences producing significant alignments:

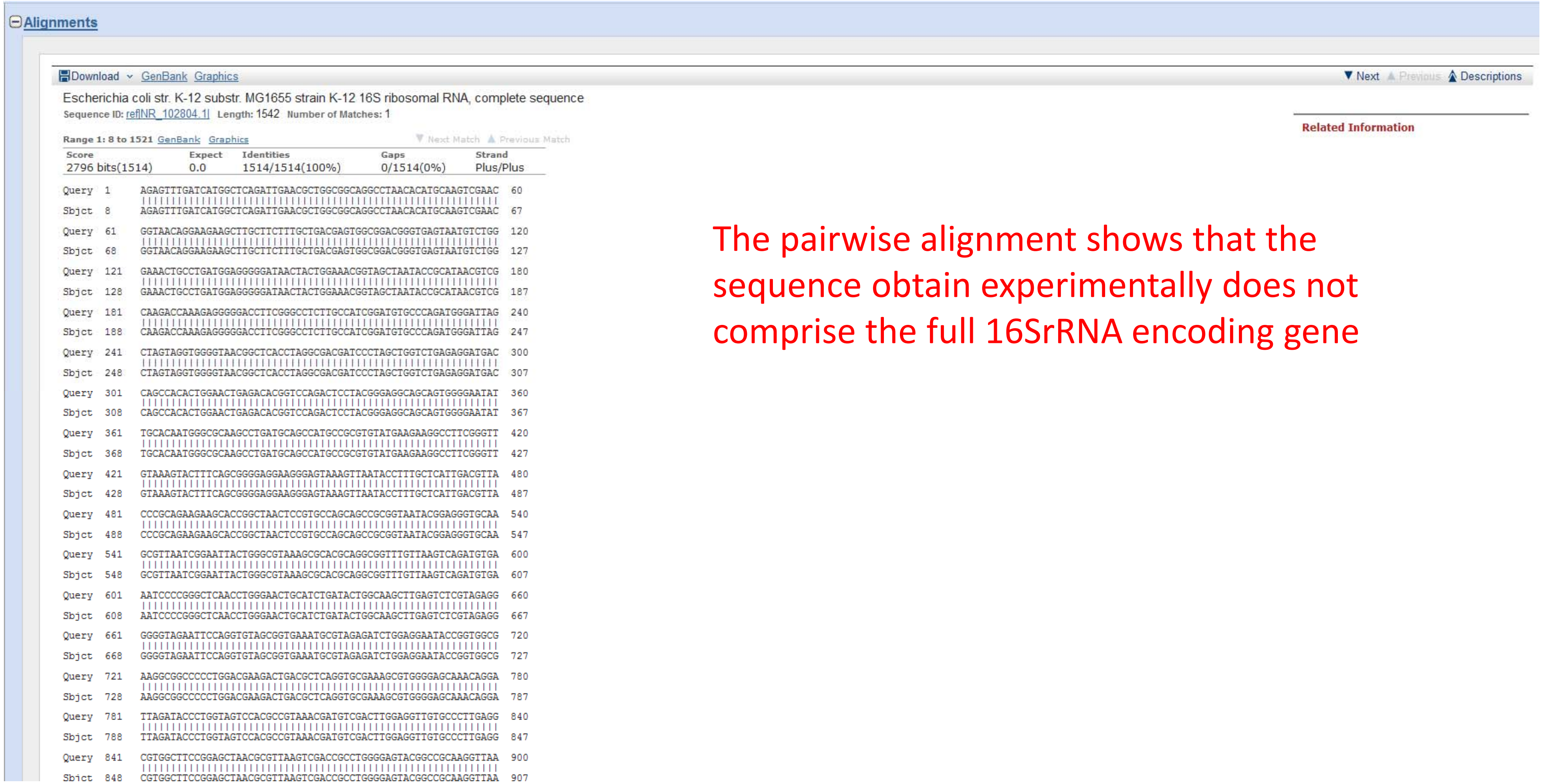
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Escherichia coli str. K-12 substr. MG1655 strain K-12 16S ribosomal RNA, complete sequence	2796	2796	100%	0.0	100%	NR_102804.1
□	Escherichia coli O157:H7 str. Sakai strain Sakai 16S ribosomal RNA, complete sequence	2785	2785	100%	0.0	99%	NR_074891.1
□	Shigella sonnei Ss046 strain Ss046 16S ribosomal RNA, complete sequence	2763	2763	100%	0.0	99%	NR_074894.1
□	Escherichia fergusonii strain ATCC 35469 16S ribosomal RNA gene, complete sequence	2730	2730	100%	0.0	99%	NR_074902.1
□	Shigella dysenteriae Sd197 strain Sd197 16S ribosomal RNA, complete sequence	2724	2724	100%	0.0	99%	NR_074892.1
□	Shigella flexneri 2a str. 301 strain 301 16S ribosomal RNA, complete sequence	2706	2706	100%	0.0	99%	NR_074882.1
□	Shigella sonnei strain CECT 4887 16S ribosomal RNA gene, partial sequence	2695	2695	99%	0.0	99%	NR_104826.1
□	Shigella boydii Sb227 strain Sb227 16S ribosomal RNA, complete sequence	2691	2691	100%	0.0	99%	NR_074893.1
□	Shigella flexneri strain ATCC 29903 16S ribosomal RNA gene, partial sequence	2682	2682	98%	0.0	99%	NR_026331.1
□	Shigella dysenteriae strain ATCC 13313 16S ribosomal RNA gene, partial sequence	2669	2669	98%	0.0	99%	NR_026332.1
□	Shigella boydii strain P288 16S ribosomal RNA gene, partial sequence	2660	2660	98%	0.0	99%	NR_104901.1
□	Escherichia fergusonii strain ATCC 35469 16S ribosomal RNA gene, partial sequence	2645	2645	97%	0.0	99%	NR_027549.1
□	Escherichia coli strain NBRC 102203 16S ribosomal RNA gene, partial sequence	2638	2638	96%	0.0	99%	NR_114042.1
□	Escherichia fergusonii strain NBRC 102419 16S ribosomal RNA gene, partial sequence	2632	2632	96%	0.0	99%	NR_114079.1
□	Escherichia albertii strain Albert 19982 16S ribosomal RNA gene, partial sequence	2632	2632	98%	0.0	98%	NR_025569.1
□	Salmonella enterica subsp. enterica serovar Typhi str. Ty2 16S ribosomal RNA gene, complete sequence	2614	2614	100%	0.0	98%	NR_074799.1
□	Citrobacter koseri ATCC BAA-895 strain ATCC BAA-895 16S ribosomal RNA, complete sequence	2608	2608	100%	0.0	98%	NR_102823.1
□	Salmonella enterica subsp. enterica serovar Typhimurium strain LT2 16S ribosomal RNA gene, complete sequence	2603	2603	99%	0.0	98%	NR_074910.1
□	Salmonella bongori strain NCTC 12419 16S ribosomal RNA gene, complete sequence	2591	2591	100%	0.0	98%	NR_074888.1
□	Citrobacter amalonaticus strain CECT 863 16S ribosomal RNA gene, partial sequence	2590	2590	99%	0.0	98%	NR_104823.1
□	Salmonella enterica subsp. enterica serovar Enteritidis str. P125109 strain P125109 16S ribosomal RNA, complete sequence	2586	2586	100%	0.0	97%	NR_074985.1
□	Citrobacter rodentium ICC168 strain ICC168 16S ribosomal RNA, complete sequence	2586	2586	100%	0.0	97%	NR_074903.1
□	Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601 strain AKU12601 16S ribosomal RNA, complete sequence	2580	2580	100%	0.0	97%	NR_074935.1
□	Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 strain ATCC 9150 16S ribosomal RNA, complete sequence	2580	2580	100%	0.0	97%	NR_074934.1
□	Escherichia coli strain U 5/41 16S ribosomal RNA gene, partial sequence	2580	2580	95%	0.0	99%	NR_024570.1

One specific hit for the E. coli strain K-12 MG1655 in the gene centric 16S rRNA database at NCBI (all hits are reference sequences)

BLAST results



Escherichia coli str. K-12 substr. MG1655 strain K-12 16S ribosomal RNA, complete sequence

Sequence ID: ref|NR_102804.1| Length: 1542 Number of Matches: 1

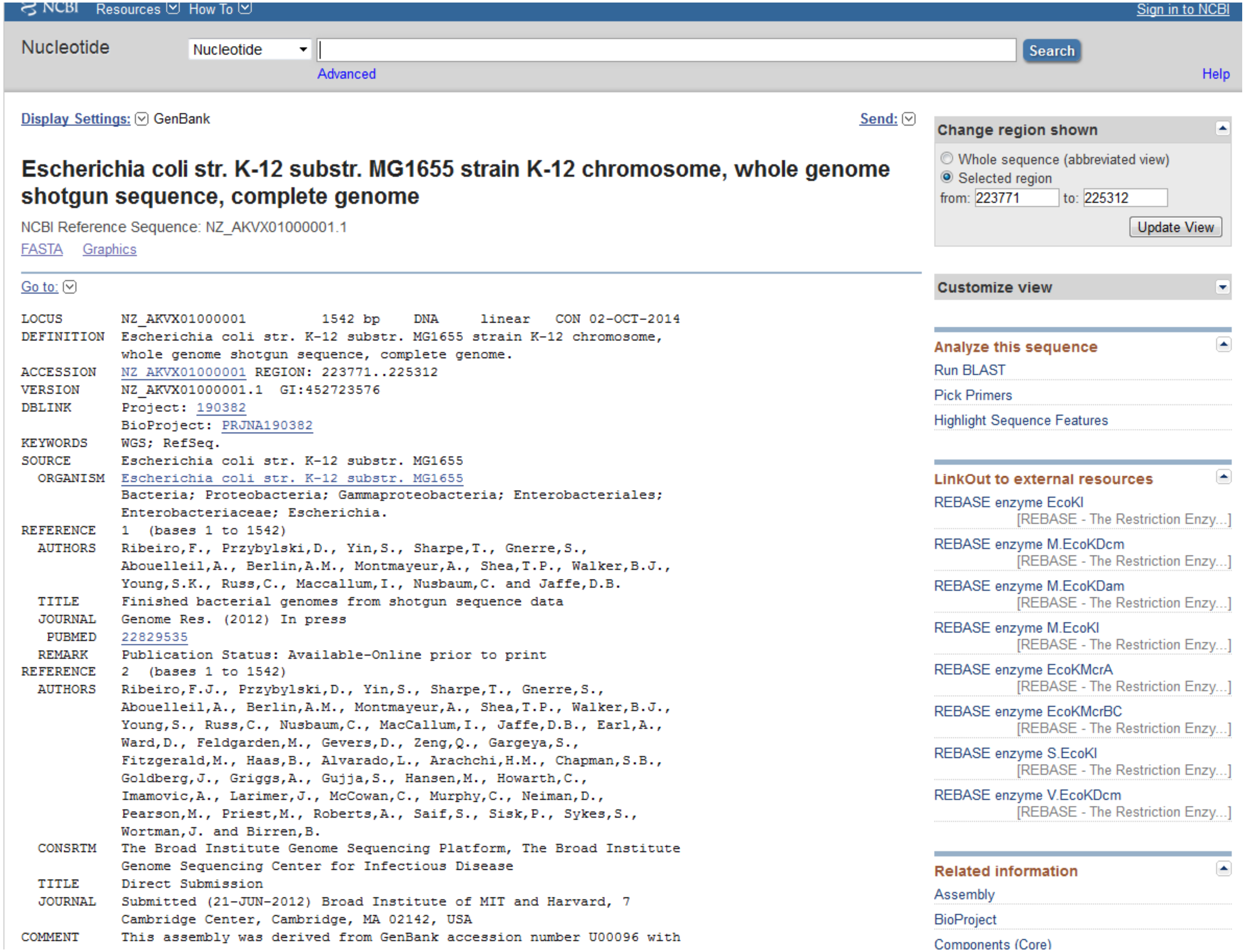
Range 1: 8 to 1521 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
2796 bits(1514)	0.0	1514/1514(100%)	0/1514(0%)	Plus/Plus

```
Query  1    AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  8    AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAAC  67

Query  61   GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  68   GGTAACAGGAAGAAGCTTGCTTCTTTGCTGACGAGTGGCGGACGGGTGAGTAATGTCTGG  127

Query  121  GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  128  GAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAATACCGCATAACGTCG  187

Query  181  CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  188  CAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTAG  247

Query  241  CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  248  CTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAGGATGAC  307

Query  301  CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  308  CAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATAT  367

Query  361  TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  368  TGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTT  427

Query  421  GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  428  GTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTAATACCTTTGCTCATTGACGTTA  487

Query  481  CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  488  CCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCAA  547

Query  541  GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  600
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  548  GCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCAGGCGGTTTGTTAAGTCAGATGTGA  607

Query  601  AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  660
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  608  AATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGG  667

Query  661  GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  720
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  668  GGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGTGGCG  727

Query  721  AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  780
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  728  AAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAAGCGTGGGGAGCAAACAGGA  787

Query  781  TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  840
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  788  TTAGATACCCTGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGG  847

Query  841  CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  900
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  848  CGTGGCTTCCGGAGCTAACGCGTTAAGTCGACCGCCTGGGGAGTACGGCCGCAAGGTTAA  907
```

The pairwise alignment shows that the sequence obtain experimentally does not comprise the full 16SrRNA encoding gene

BLAST results



Display Settings: GenBank

Escherichia coli str. K-12 substr. MG1655 strain K-12 chromosome, whole genome shotgun sequence, complete genome

NCBI Reference Sequence: NZ_AKVX01000001.1

```
LOCUS       NZ_AKVX01000001         1542 bp    DNA     linear   CON 02-OCT-2014
DEFINITION  Escherichia coli str. K-12 substr. MG1655 strain K-12 chromosome,
            whole genome shotgun sequence, complete genome.
ACCESSION   NZ_AKVX01000001 REGION: 223771..225312
VERSION     NZ_AKVX01000001.1  GI:452723576
DBLINK      Project: 190382
            BioProject: PRJNA190382
KEYWORDS    WGS; RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1542)
  AUTHORS   Ribeiro,F., Przybylski,D., Yin,S., Sharpe,T., Gnerre,S.,
            Abouelleil,A., Berlin,A.M., Montmayeur,A., Shea,T.P., Walker,B.J.,
            Young,S.K., Russ,C., Maccallum,I., Nusbaum,C. and Jaffe,D.B.
  TITLE     Finished bacterial genomes from shotgun sequence data
  JOURNAL   Genome Res. (2012) In press
   PUBMED   22829535
  REMARK    Publication Status: Available-Online prior to print
REFERENCE   2  (bases 1 to 1542)
  AUTHORS   Ribeiro,F.J., Przybylski,D., Yin,S., Sharpe,T., Gnerre,S.,
            Abouelleil,A., Berlin,A.M., Montmayeur,A., Shea,T.P., Walker,B.J.,
            Young,S., Russ,C., Nusbaum,C., MacCallum,I., Jaffe,D.B., Earl,A.,
            Ward,D., Feldgarden,M., Gevers,D., Zeng,Q., Gargeya,S.,
            Fitzgerald,M., Haas,B., Alvarado,L., Arachchi,H.M., Chapman,S.B.,
            Goldberg,J., Griggs,A., Gujja,S., Hansen,M., Howarth,C.,
            Imamovic,A., Larimer,J., McCowan,C., Murphy,C., Neiman,D.,
            Pearson,M., Priest,M., Roberts,A., Saif,S., Sisk,P., Sykes,S.,
            Wortman,J. and Birren,B.
  CONSRTM   The Broad Institute Genome Sequencing Platform, The Broad Institute
            Genome Sequencing Center for Infectious Disease
  TITLE     Direct Submission
  JOURNAL   Submitted (21-JUN-2012) Broad Institute of MIT and Harvard, 7
            Cambridge Center, Cambridge, MA 02142, USA
COMMENT     This assembly was derived from GenBank accession number U00096 with
```

Change region shown

○ Whole sequence (abbreviated view)
● Selected region
from: 223771 to: 225312
Update View

Customize view

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features

LinkOut to external resources

- REBASE enzyme EcoKI [REBASE - The Restriction Enzy...]
- REBASE enzyme M.EcoKDcm [REBASE - The Restriction Enzy...]
- REBASE enzyme M.EcoKDam [REBASE - The Restriction Enzy...]
- REBASE enzyme M.EcoKI [REBASE - The Restriction Enzy...]
- REBASE enzyme EcoKMcrA [REBASE - The Restriction Enzy...]
- REBASE enzyme EcoKMcrBC [REBASE - The Restriction Enzy...]
- REBASE enzyme S.EcoKI [REBASE - The Restriction Enzy...]
- REBASE enzyme V.EcoKDcm [REBASE - The Restriction Enzy...]

Related information

- Assembly
- BioProject
- Components (Core)

Nucleotide BLAST results

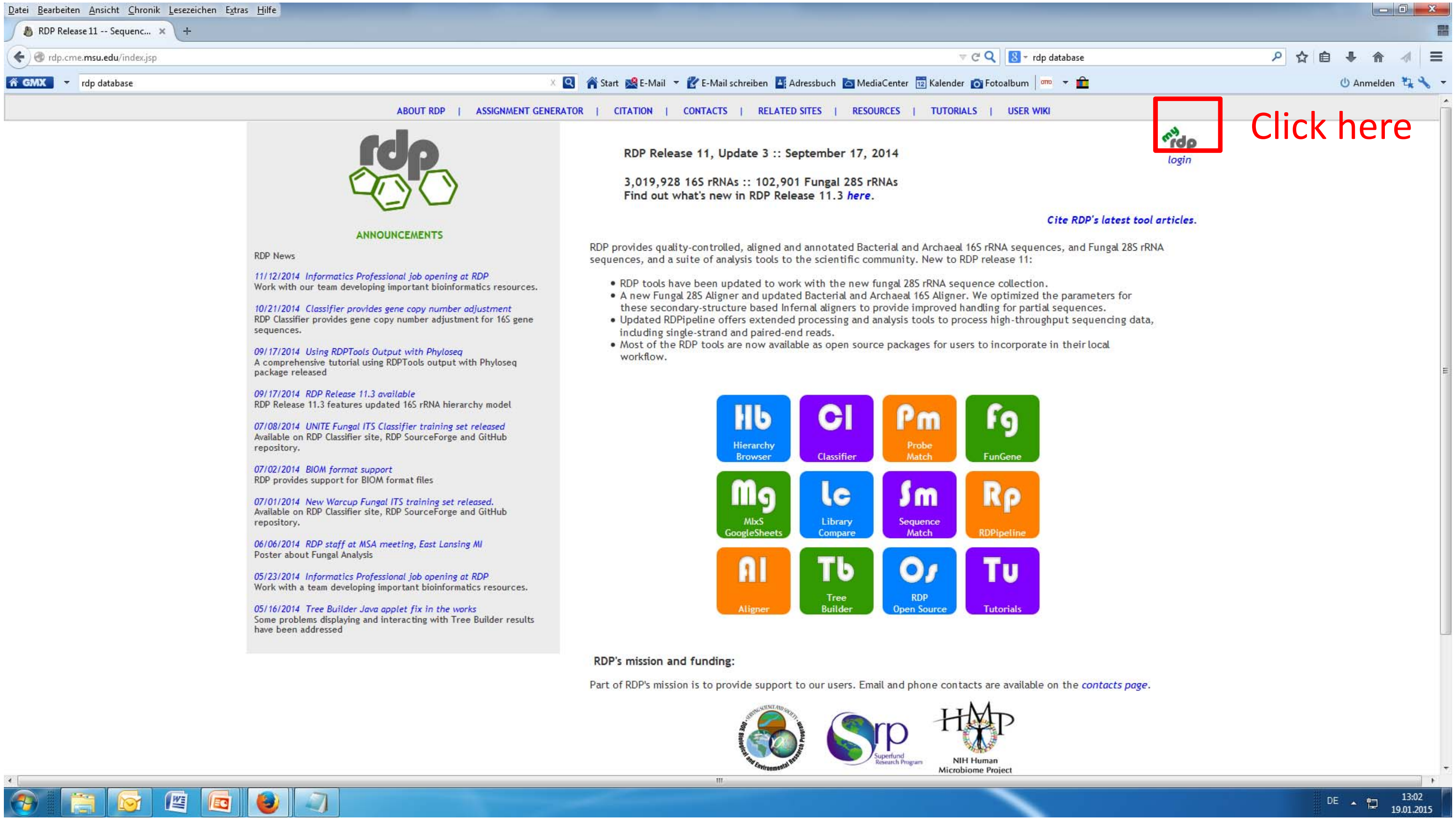
- The obtained sequence indeed encodes a 16S rRNA
- Most likely from *E. coli* K-12 MG1655
  - Full sequence coverage
  - 100% identical

rRNA sequence databases

For more specialized and accurate classification and phylogenies of rRNA (rDNA) sequences

- http://rdp.cme.msu.edu
- http://greengenes.lbl.gov/
- http://www.arb-silva.de/

Ribosomal database project (rdp) database



Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
RDP Release 11 -- Sequenc...
rdp.cme.msu.edu/index.jsp
rdp database
GMX
rdp database
Start
E-Mail
E-Mail schreiben
Adressbuch
MediaCenter
Kalender
Fotoalbum
Anmelden
ABOUT RDP | ASSIGNMENT GENERATOR | CITATION | CONTACTS | RELATED SITES | RESOURCES | TUTORIALS | USER WIKI
my rdp
login

Click here

ANNOUNCEMENTS

RDP News

11/12/2014 Informatics Professional job opening at RDP
Work with our team developing important bioinformatics resources.

10/21/2014 Classifier provides gene copy number adjustment
RDP Classifier provides gene copy number adjustment for 16S gene sequences.

09/17/2014 Using RDPTools Output with Phyloseq
A comprehensive tutorial using RDPTools output with Phyloseq package released

09/17/2014 RDP Release 11.3 available
RDP Release 11.3 features updated 16S rRNA hierarchy model

07/08/2014 UNITE Fungal ITS Classifier training set released
Available on RDP Classifier site, RDP SourceForge and GitHub repository.

07/02/2014 BIOM format support
RDP provides support for BIOM format files

07/01/2014 New Warcup Fungal ITS training set released.
Available on RDP Classifier site, RDP SourceForge and GitHub repository.

06/06/2014 RDP staff at MSA meeting, East Lansing MI
Poster about Fungal Analysis

05/23/2014 Informatics Professional job opening at RDP
Work with a team developing important bioinformatics resources.

05/16/2014 Tree Builder Java applet fix in the works
Some problems displaying and interacting with Tree Builder results have been addressed

RDP Release 11, Update 3 :: September 17, 2014

3,019,928 16S rRNAs :: 102,901 Fungal 28S rRNAs
Find out what's new in RDP Release 11.3 *here*.

Cite RDP's latest tool articles.

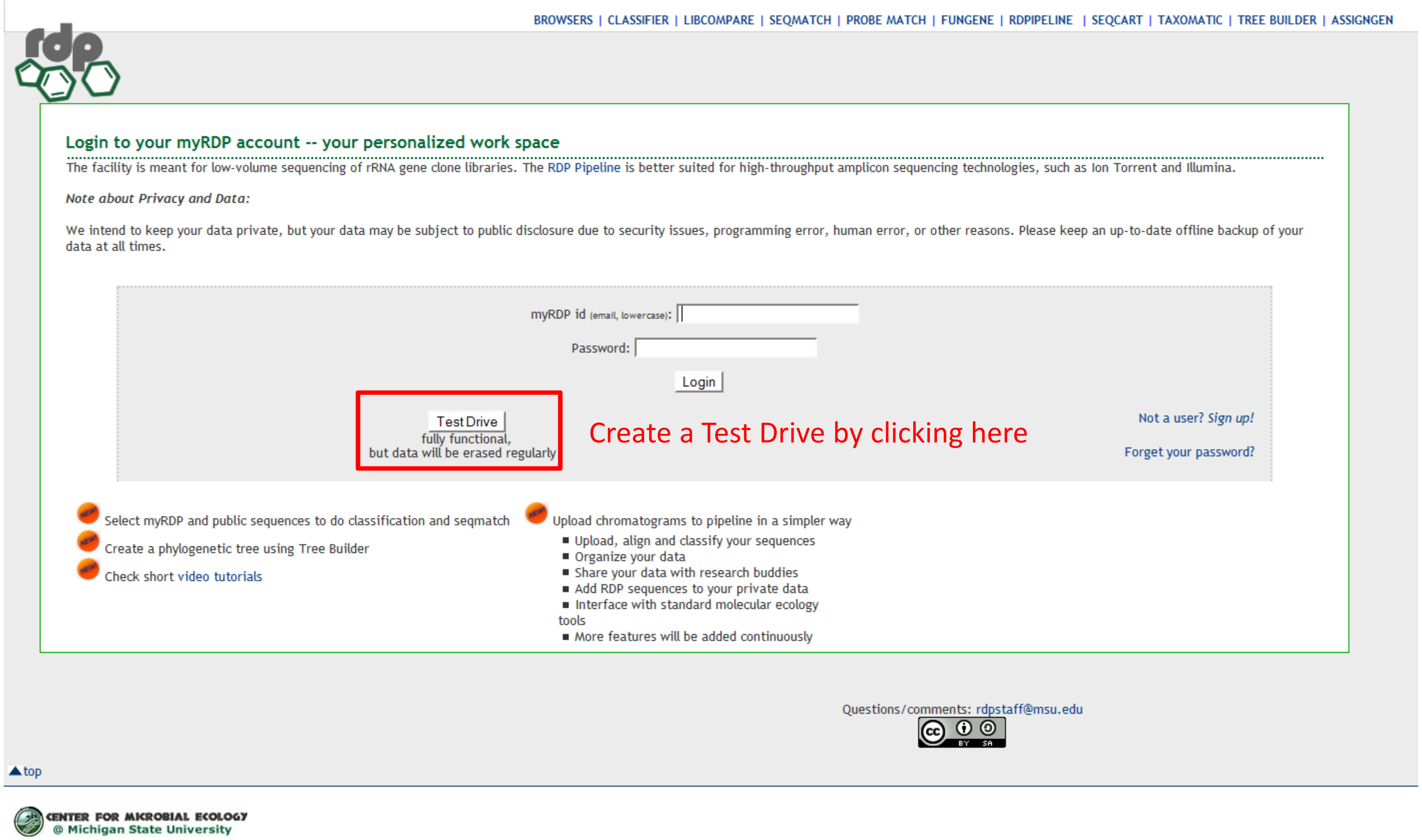
RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.

RDP's mission and funding:

Part of RDP's mission is to provide support to our users. Email and phone contacts are available on the *contacts page*.

rdp database project



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp

Login to your myRDP account -- your personalized work space

The facility is meant for low-volume sequencing of rRNA gene clone libraries. The RDP Pipeline is better suited for high-throughput amplicon sequencing technologies, such as Ion Torrent and Illumina.

Note about Privacy and Data:

We intend to keep your data private, but your data may be subject to public disclosure due to security issues, programming error, human error, or other reasons. Please keep an up-to-date offline backup of your data at all times.

myRDP id (email, lowercase):

Password:

fully functional,
but data will be erased regularly

Create a Test Drive by clicking here

Not a user? *Sign up!*

Forget your password?

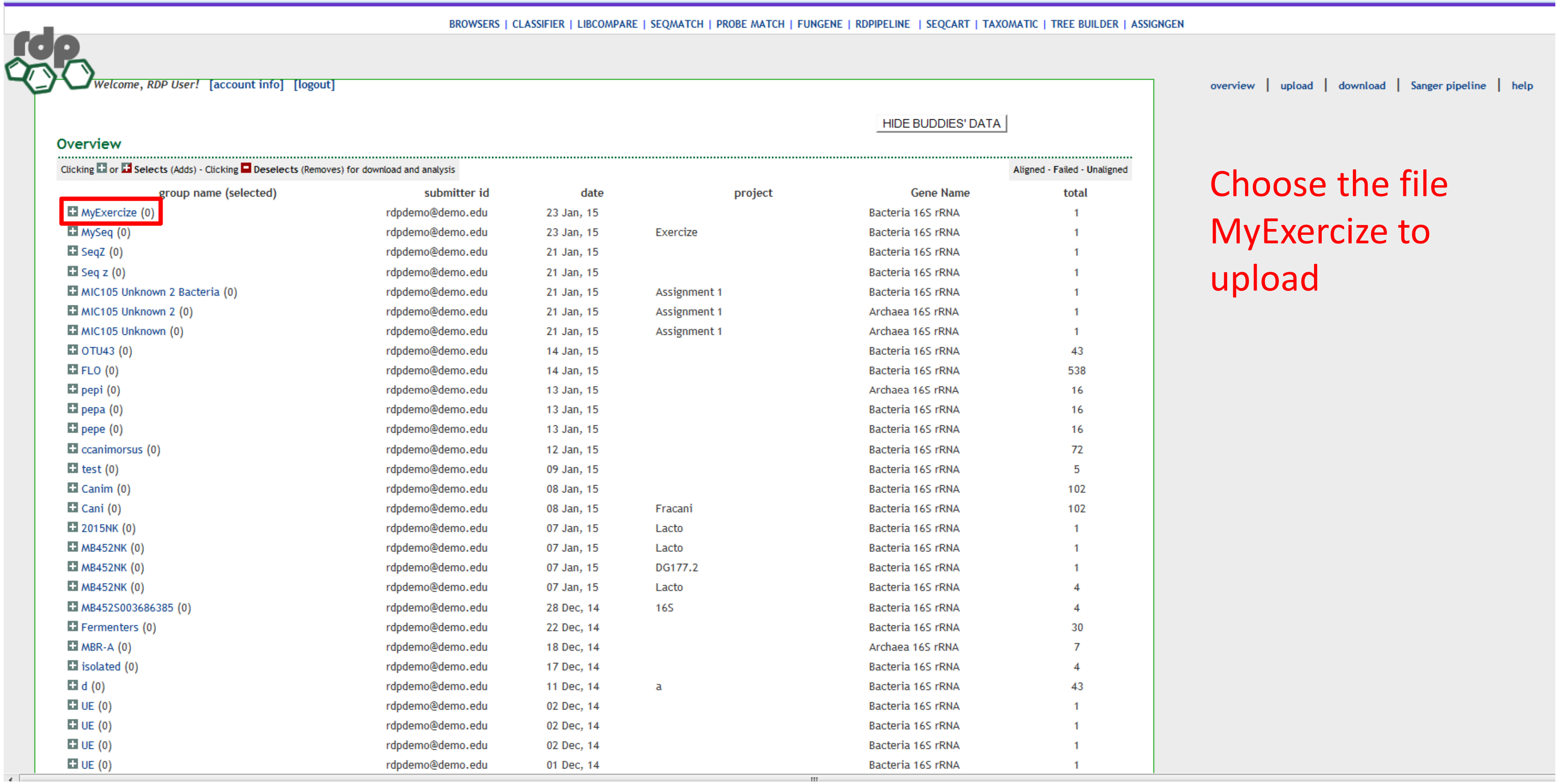
- Select myRDP and public sequences to do classification and seqmatch
- Create a phylogenetic tree using Tree Builder
- Check short video tutorials

Upload chromatograms to pipeline in a simpler way

- Upload, align and classify your sequences
- Organize your data
- Share your data with research buddies
- Add RDP sequences to your private data
- Interface with standard molecular ecology tools
- More features will be added continuously

Questions/comments: rdpstaff@msu.edu

Upload sequence


BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
overview | upload | download | Sanger pipeline | help
HIDE BUDDIES' DATA

Welcome, RDP User! [account info] [logout]

Overview

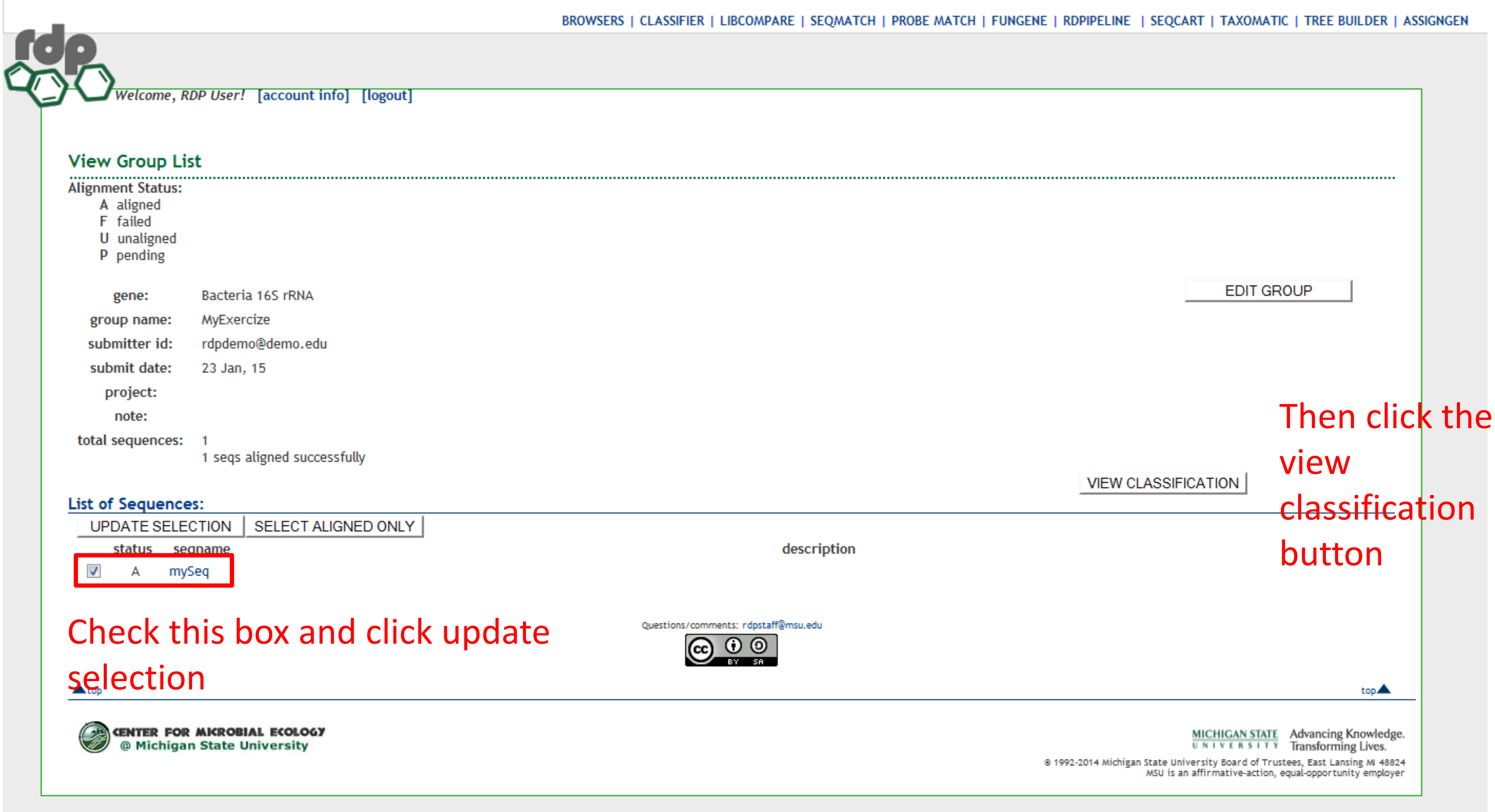
Clicking ⊞ or ⊞ **Selects** (Adds) - Clicking ⊟ **Deselects** (Removes) for download and analysis

Aligned - Failed - Unaligned

group name (selected)	submitter id	date	project	Gene Name	total
⊞ MyExercize (0)	rdpdemo@demo.edu	23 Jan, 15		Bacteria 16S rRNA	1
⊞ MySeq (0)	rdpdemo@demo.edu	23 Jan, 15	Exercize	Bacteria 16S rRNA	1
⊞ SeqZ (0)	rdpdemo@demo.edu	21 Jan, 15		Bacteria 16S rRNA	1
⊞ Seq z (0)	rdpdemo@demo.edu	21 Jan, 15		Bacteria 16S rRNA	1
⊞ MIC105 Unknown 2 Bacteria (0)	rdpdemo@demo.edu	21 Jan, 15	Assignment 1	Bacteria 16S rRNA	1
⊞ MIC105 Unknown 2 (0)	rdpdemo@demo.edu	21 Jan, 15	Assignment 1	Archaea 16S rRNA	1
⊞ MIC105 Unknown (0)	rdpdemo@demo.edu	21 Jan, 15	Assignment 1	Archaea 16S rRNA	1
⊞ OTU43 (0)	rdpdemo@demo.edu	14 Jan, 15		Bacteria 16S rRNA	43
⊞ FLO (0)	rdpdemo@demo.edu	14 Jan, 15		Bacteria 16S rRNA	538
⊞ pepi (0)	rdpdemo@demo.edu	13 Jan, 15		Archaea 16S rRNA	16
⊞ pepa (0)	rdpdemo@demo.edu	13 Jan, 15		Bacteria 16S rRNA	16
⊞ pepe (0)	rdpdemo@demo.edu	13 Jan, 15		Bacteria 16S rRNA	16
⊞ ccanimorsus (0)	rdpdemo@demo.edu	12 Jan, 15		Bacteria 16S rRNA	72
⊞ test (0)	rdpdemo@demo.edu	09 Jan, 15		Bacteria 16S rRNA	5
⊞ Canim (0)	rdpdemo@demo.edu	08 Jan, 15		Bacteria 16S rRNA	102
⊞ Cani (0)	rdpdemo@demo.edu	08 Jan, 15	Fracani	Bacteria 16S rRNA	102
⊞ 2015NK (0)	rdpdemo@demo.edu	07 Jan, 15	Lacto	Bacteria 16S rRNA	1
⊞ MB452NK (0)	rdpdemo@demo.edu	07 Jan, 15	Lacto	Bacteria 16S rRNA	1
⊞ MB452NK (0)	rdpdemo@demo.edu	07 Jan, 15	DG177.2	Bacteria 16S rRNA	1
⊞ MB452NK (0)	rdpdemo@demo.edu	07 Jan, 15	Lacto	Bacteria 16S rRNA	4
⊞ MB452S003686385 (0)	rdpdemo@demo.edu	28 Dec, 14	16S	Bacteria 16S rRNA	4
⊞ Fermenters (0)	rdpdemo@demo.edu	22 Dec, 14		Bacteria 16S rRNA	30
⊞ MBR-A (0)	rdpdemo@demo.edu	18 Dec, 14		Archaea 16S rRNA	7
⊞ isolated (0)	rdpdemo@demo.edu	17 Dec, 14		Bacteria 16S rRNA	4
⊞ d (0)	rdpdemo@demo.edu	11 Dec, 14	a	Bacteria 16S rRNA	43
⊞ UE (0)	rdpdemo@demo.edu	02 Dec, 14		Bacteria 16S rRNA	1
⊞ UE (0)	rdpdemo@demo.edu	02 Dec, 14		Bacteria 16S rRNA	1
⊞ UE (0)	rdpdemo@demo.edu	02 Dec, 14		Bacteria 16S rRNA	1
⊞ UE (0)	rdpdemo@demo.edu	01 Dec, 14		Bacteria 16S rRNA	1

Choose the file MyExercize to upload

Classifying the query sequence



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
EDIT GROUP
VIEW CLASSIFICATION
UPDATE SELECTION
SELECT ALIGNED ONLY
top
top
Questions/comments: rdpstaff@msu.edu
CENTER FOR MICROBIAL ECOLOGY @ Michigan State University
MICHIGAN STATE UNIVERSITY Advancing Knowledge. Transforming Lives.
© 1992-2014 Michigan State University Board of Trustees, East Lansing MI 48824
MSU is an affirmative-action, equal-opportunity employer

View Group List

Alignment Status:
A aligned
F failed
U unaligned
P pending

gene:	Bacteria 16S rRNA
group name:	MyExercize
submitter id:	rdpdemo@demo.edu
submit date:	23 Jan, 15
project:	
note:	
total sequences:	1 1 seqs aligned successfully

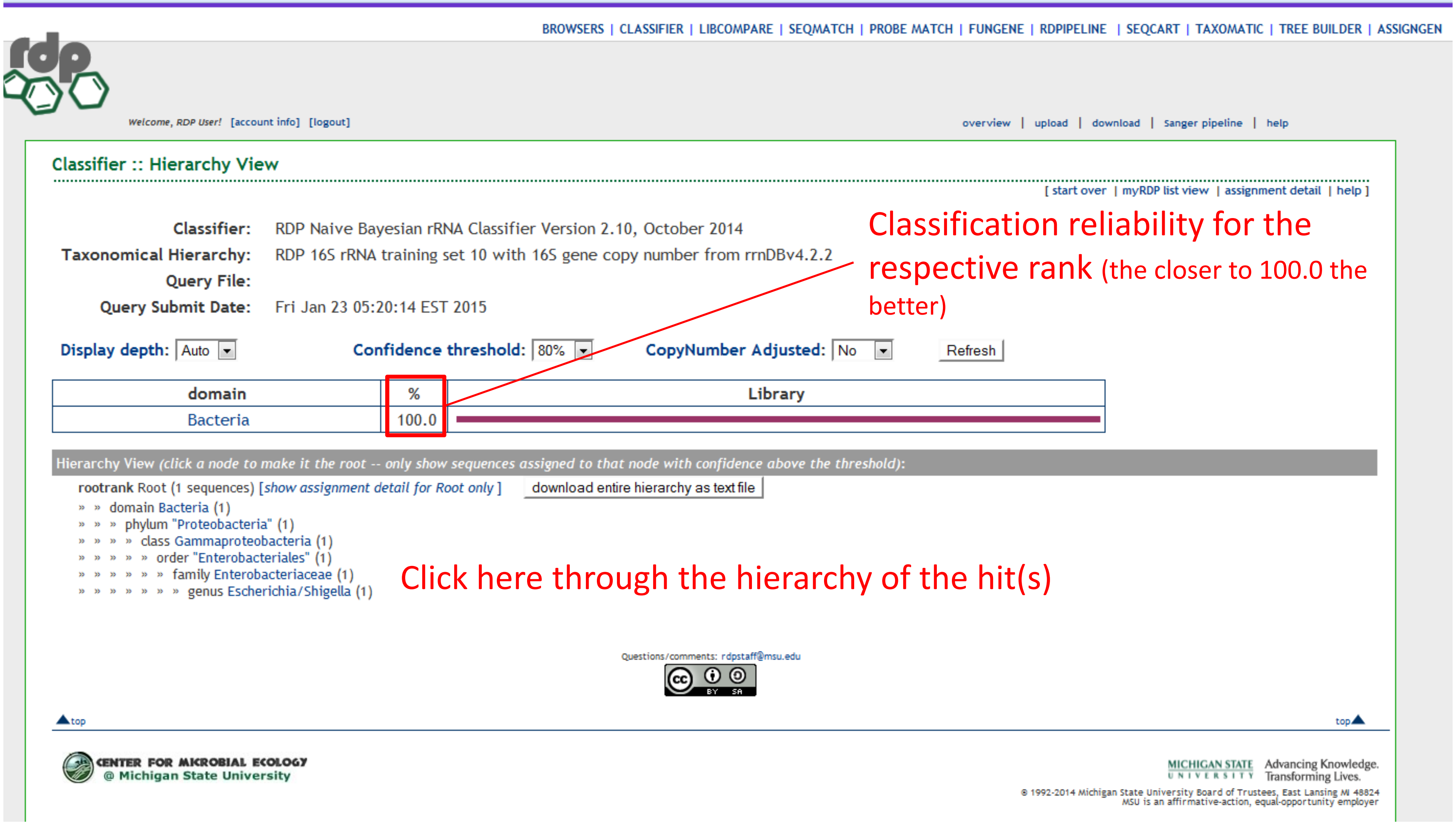
List of Sequences:

	status	seqname	description
☑	A	mySeq	

Then click the view classification button

Check this box and click update selection

Classification results


BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help
[start over | myRDP list view | assignment detail | help]
Refresh
download entire hierarchy as text file
top
top

Classifier :: Hierarchy View

Classifier: RDP Naive Bayesian rRNA Classifier Version 2.10, October 2014
Taxonomical Hierarchy: RDP 16S rRNA training set 10 with 16S gene copy number from rrnDBv4.2.2
Query File:
Query Submit Date: Fri Jan 23 05:20:14 EST 2015

Display depth: Auto **Confidence threshold:** 80% **CopyNumber Adjusted:** No

domain	%	Library
Bacteria	100.0	

Hierarchy View (*click a node to make it the root -- only show sequences assigned to that node with confidence above the threshold*):

rootrank Root (1 sequences) [*show assignment detail for Root only*]

» » domain Bacteria (1)
» » » phylum "Proteobacteria" (1)
» » » » class Gammaproteobacteria (1)
» » » » » order "Enterobacteriales" (1)
» » » » » » family Enterobacteriaceae (1)
» » » » » » » genus Escherichia/Shigella (1)

Questions/comments: rdpstaff@msu.edu

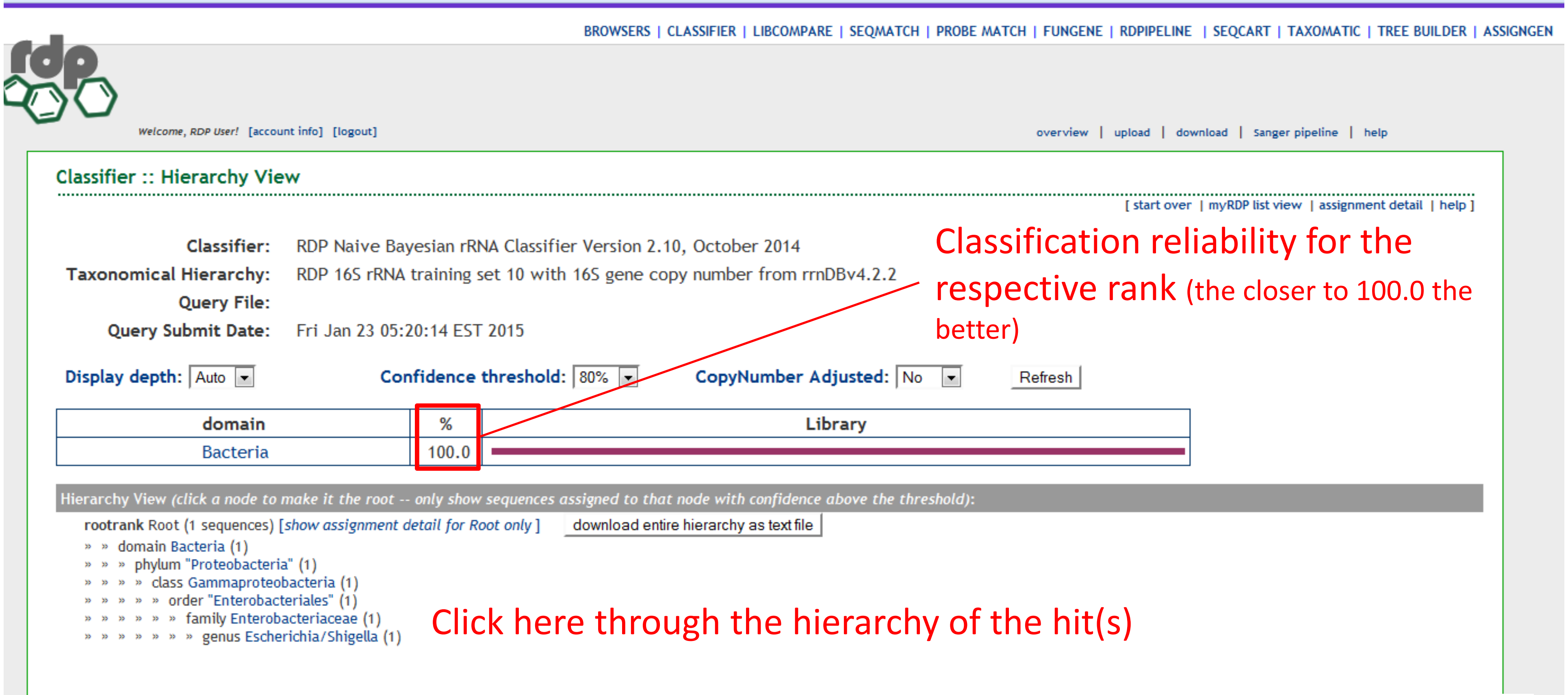
CENTER FOR MICROBIAL ECOLOGY
@ Michigan State University

MICHIGAN STATE UNIVERSITY Advancing Knowledge. Transforming Lives.

Classification reliability for the respective rank (the closer to 100.0 the better)

Click here through the hierarchy of the hit(s)

Classification results



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | Sanger pipeline | help
[start over | myRDP list view | assignment detail | help]
Refresh
download entire hierarchy as text file

Classifier :: Hierarchy View

Classifier: RDP Naive Bayesian rRNA Classifier Version 2.10, October 2014

Taxonomical Hierarchy: RDP 16S rRNA training set 10 with 16S gene copy number from rrnDBv4.2.2

Query File:

Query Submit Date: Fri Jan 23 05:20:14 EST 2015

Display depth: Auto Confidence threshold: 80% CopyNumber Adjusted: No

domain	%	Library
Bacteria	100.0	

Hierarchy View (*click a node to make it the root -- only show sequences assigned to that node with confidence above the threshold*):

rootrank Root (1 sequences) [*show assignment detail for Root only*]

```
» » domain Bacteria (1)
» » » phylum "Proteobacteria" (1)
» » » » class Gammaproteobacteria (1)
» » » » » order "Enterobacteriales" (1)
» » » » » » family Enterobacteriaceae (1)
» » » » » » » genus Escherichia/Shigella (1)
```

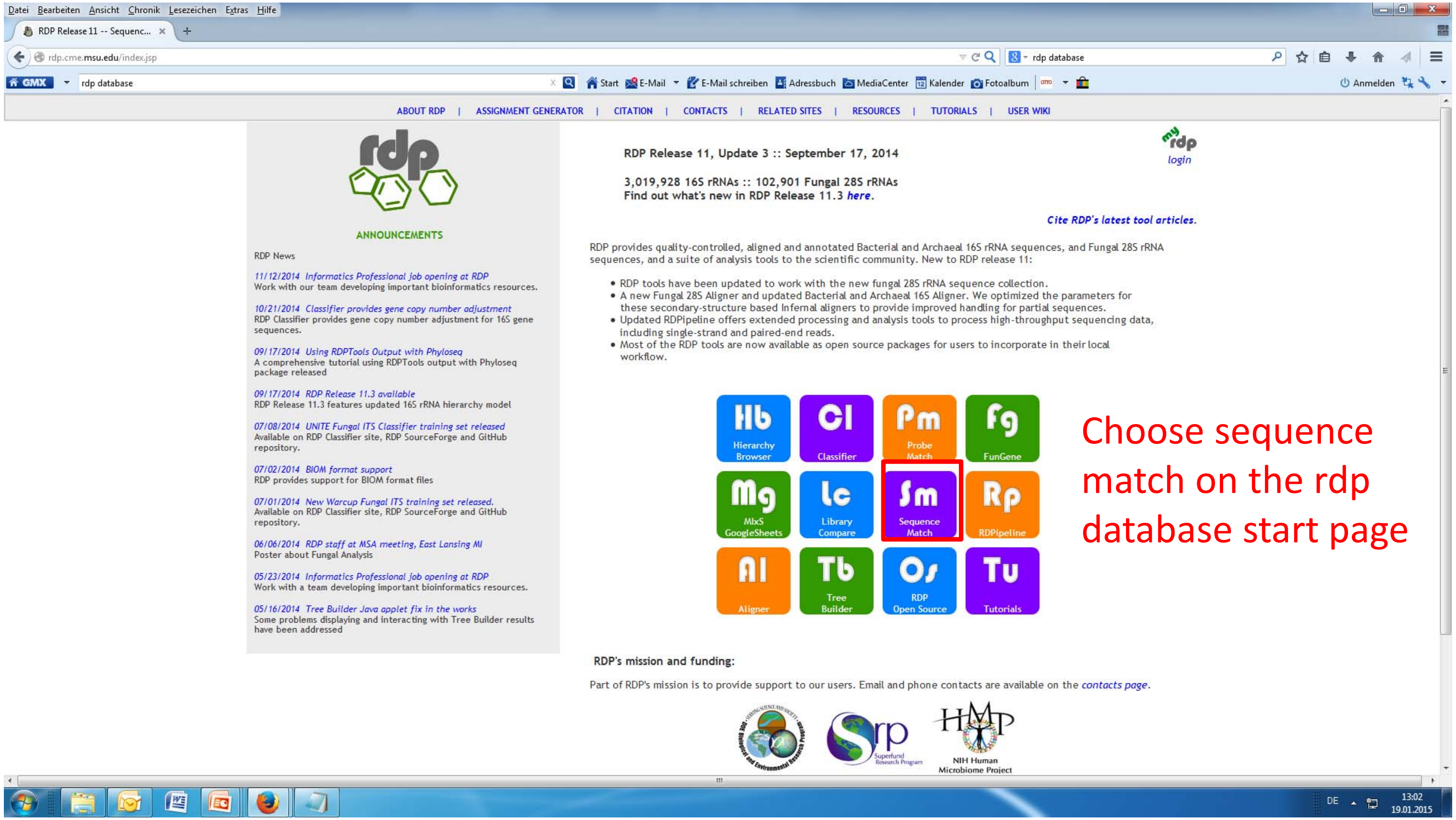
Classification reliability for the respective rank (the closer to 100.0 the better)

Click here through the hierarchy of the hit(s)

As already indicated by the BLAST searches, you can now be quite sure that:

- The obtained sequence indeed encodes a 16S rRNA
- Most likely from the genus *Escherichia/Shigella*

Finding the closest relatives to your sequence - Seqmatch



ANNOUNCEMENTS

RDP News

11/12/2014 Informatics Professional job opening at RDP
Work with our team developing important bioinformatics resources.

10/21/2014 Classifier provides gene copy number adjustment
RDP Classifier provides gene copy number adjustment for 16S gene sequences.

09/17/2014 Using RDPTools Output with Phyloseq
A comprehensive tutorial using RDPTools output with Phyloseq package released

09/17/2014 RDP Release 11.3 available
RDP Release 11.3 features updated 16S rRNA hierarchy model

07/08/2014 UNITE Fungal ITS Classifier training set released
Available on RDP Classifier site, RDP SourceForge and GitHub repository.

07/02/2014 BIOM format support
RDP provides support for BIOM format files

07/01/2014 New Warcup Fungal ITS training set released.
Available on RDP Classifier site, RDP SourceForge and GitHub repository.

06/06/2014 RDP staff at MSA meeting, East Lansing MI
Poster about Fungal Analysis

05/23/2014 Informatics Professional job opening at RDP
Work with a team developing important bioinformatics resources.

05/16/2014 Tree Builder Java applet fix in the works
Some problems displaying and interacting with Tree Builder results have been addressed

RDP Release 11, Update 3 :: September 17, 2014

3,019,928 16S rRNAs :: 102,901 Fungal 28S rRNAs
Find out what's new in RDP Release 11.3 *here*.

Cite RDP's latest tool articles.

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

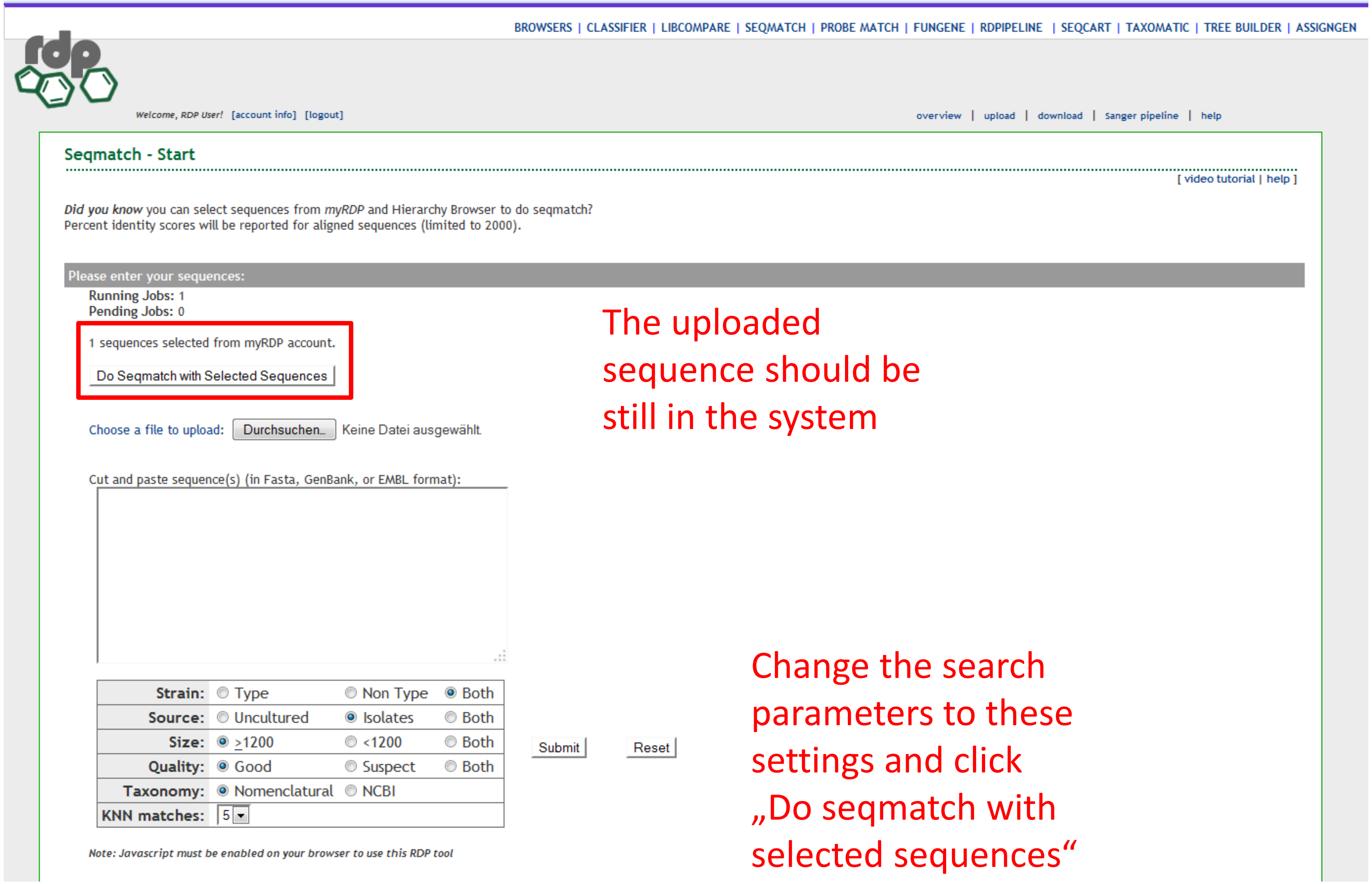
- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.

Choose sequence match on the rdp database start page

RDP's mission and funding:

Part of RDP's mission is to provide support to our users. Email and phone contacts are available on the *contacts page*.

Seqmatch



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help
[video tutorial | help]

Seqmatch - Start

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:

Running Jobs: 1
Pending Jobs: 0

1 sequences selected from myRDP account.

Do Seqmatch with Selected Sequences

Choose a file to upload: Durchsuchen... Keine Datei ausgewählt.

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

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	● Isolates	○ Both
Size:	● >1200	○ <1200	○ Both
Quality:	● Good	○ Suspect	○ Both
Taxonomy:	● Nomenclatural	○ NCBI	
KNN matches:	5		

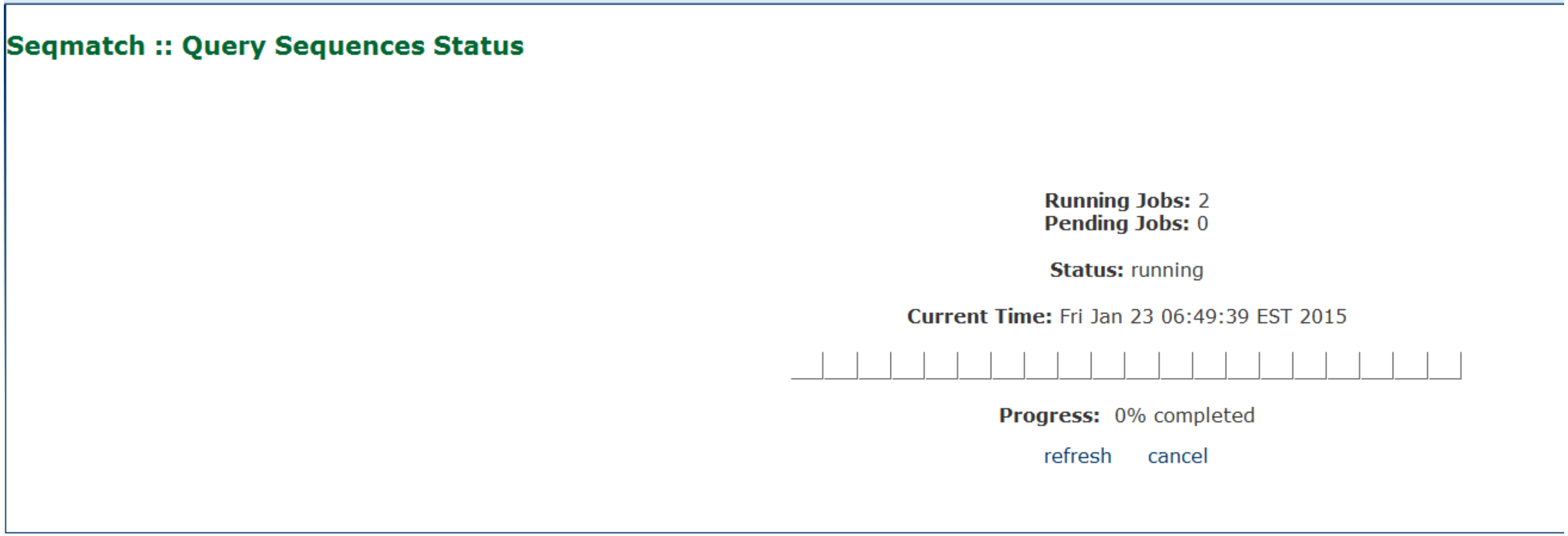
Submit Reset

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

The uploaded sequence should be still in the system

Change the search parameters to these settings and click „Do seqmatch with selected sequences“

Seqmatch



Seqmatch :: Query Sequences Status

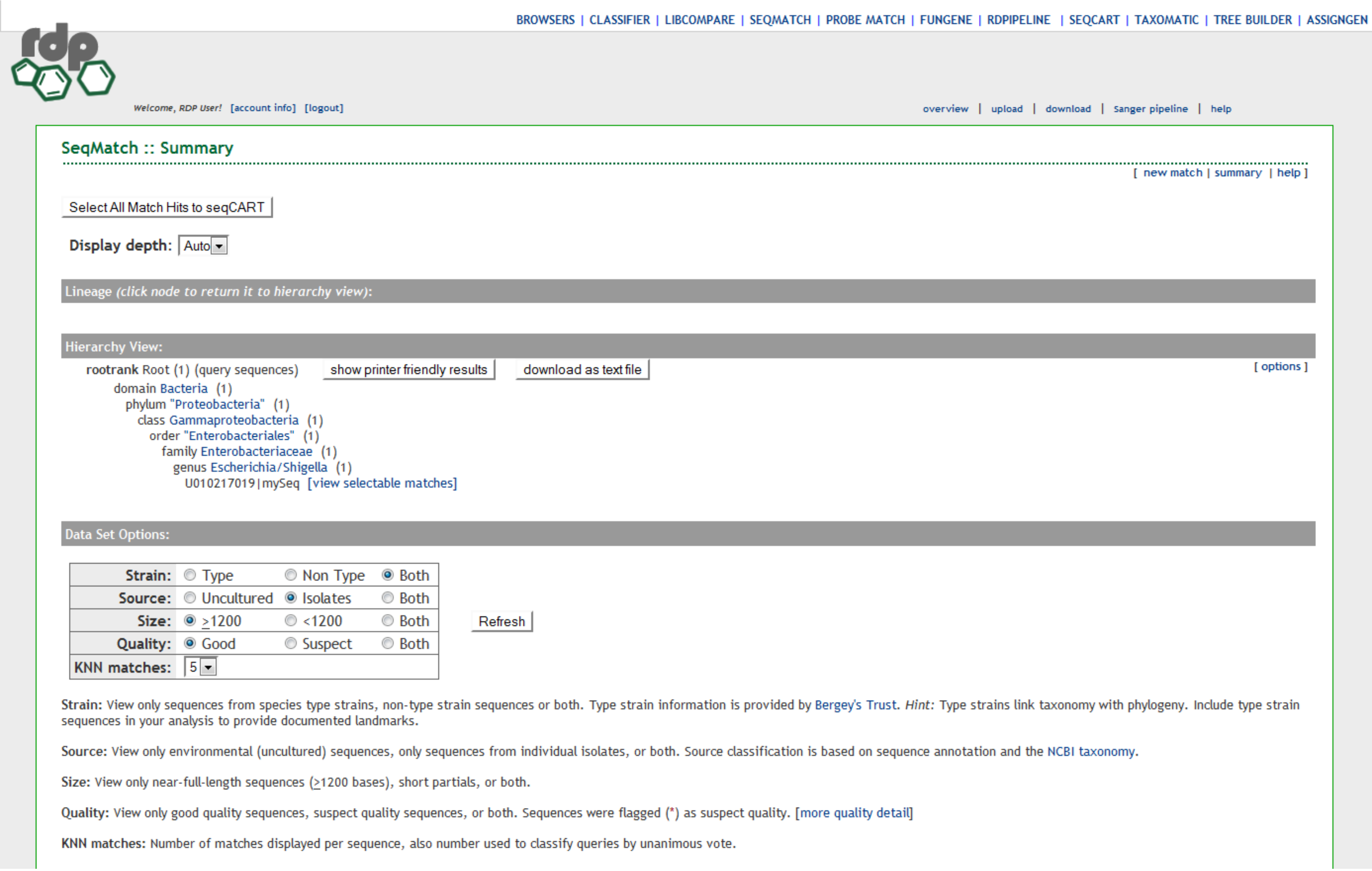
Running Jobs: 2
Pending Jobs: 0

Status: running

Current Time: Fri Jan 23 06:49:39 EST 2015

Progress: 0% completed

Seqmatch - Results



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

SeqMatch :: Summary

[new match | summary | help]
Select All Match Hits to seqCART

Display depth: Auto

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

Hierarchy View:

show printer friendly results
download as text file
[options]

```
rootrank Root (1) (query sequences)
   domain Bacteria (1)
     phylum "Proteobacteria" (1)
       class Gammaproteobacteria (1)
         order "Enterobacteriales" (1)
           family Enterobacteriaceae (1)
             genus Escherichia/Shigella (1)
               U010217019|mySeq [view selectable matches]
```

Data Set Options:

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	● Isolates	○ Both
Size:	● ≥1200	○ <1200	○ Both
Quality:	● Good	○ Suspect	○ Both
KNN matches:	5		

Strain: View only sequences from species type strains, non-type strain sequences or both. Type strain information is provided by Bergey's Trust. *Hint:* Type strains link taxonomy with phylogeny. Include type strain sequences in your analysis to provide documented landmarks.

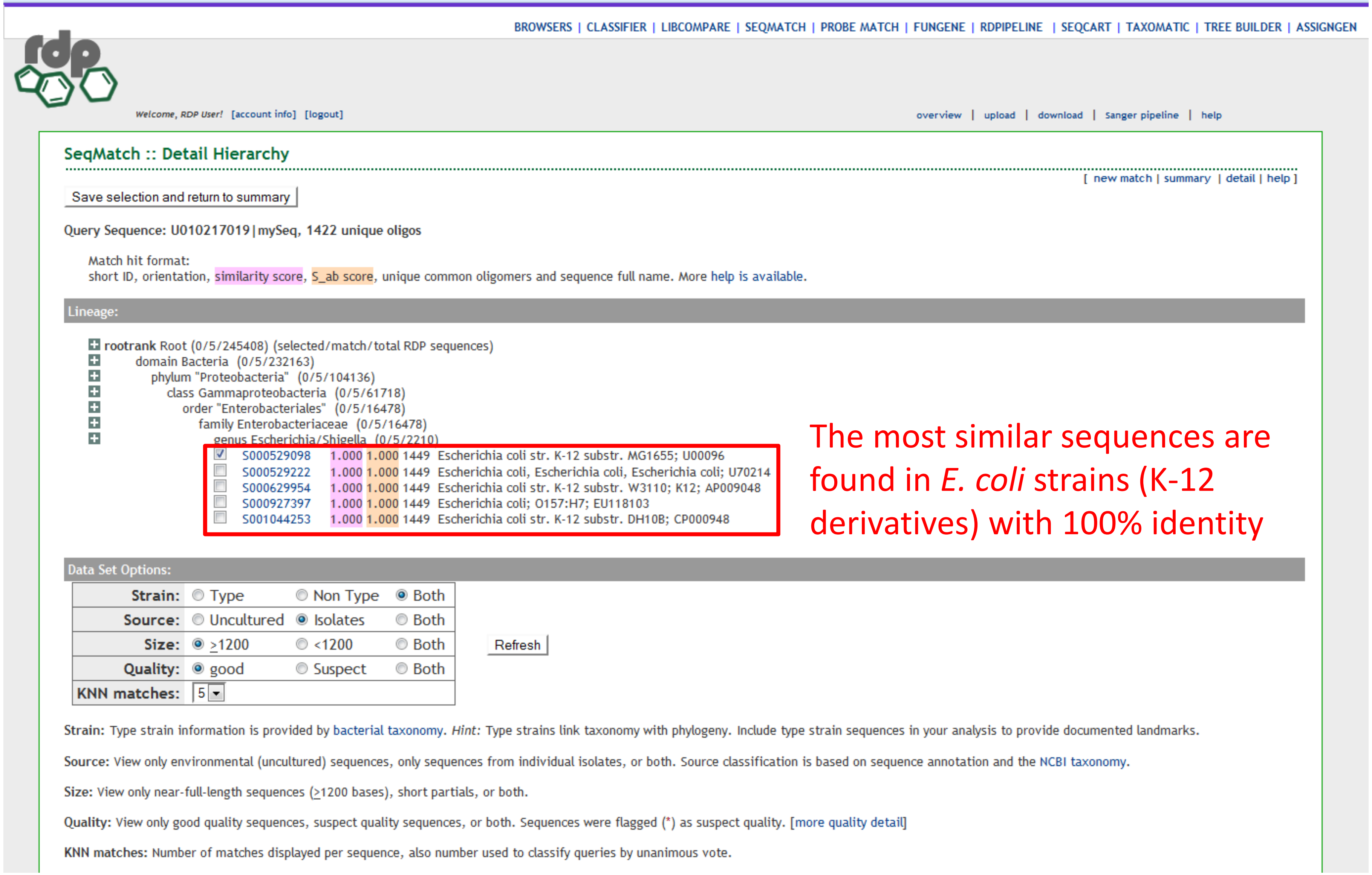
Source: View only environmental (uncultured) sequences, only sequences from individual isolates, or both. Source classification is based on sequence annotation and the NCBI taxonomy.

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. [more quality detail]

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

Seqmatch - Results



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

SeqMatch :: Detail Hierarchy

[new match | summary | detail | help]

Save selection and return to summary

Query Sequence: U010217019|mySeq, 1422 unique oligos

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

Lineage:

```
⊞ rootrank Root (0/5/245408) (selected/match/total RDP sequences)
⊞     domain Bacteria  (0/5/232163)
⊞       phylum "Proteobacteria"  (0/5/104136)
⊞         class Gammaproteobacteria  (0/5/61718)
⊞           order "Enterobacteriales"  (0/5/16478)
⊞             family Enterobacteriaceae  (0/5/16478)
⊞               genus Escherichia/Shigella  (0/5/2210)
                 ☑ S000529098  1.000 1.000 1449 Escherichia coli str. K-12 substr. MG1655; U00096
                 ☐ S000529222  1.000 1.000 1449 Escherichia coli, Escherichia coli, Escherichia coli; U70214
                 ☐ S000629954  1.000 1.000 1449 Escherichia coli str. K-12 substr. W3110; K12; AP009048
                 ☐ S000927397  1.000 1.000 1449 Escherichia coli; O157:H7; EU118103
                 ☐ S001044253  1.000 1.000 1449 Escherichia coli str. K-12 substr. DH10B; CP000948
```

The most similar sequences are found in *E. coli* strains (K-12 derivatives) with 100% identity

Data Set Options:

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	● Isolates	○ Both
Size:	● ≥1200	○ <1200	○ Both
Quality:	● good	○ Suspect	○ Both
KNN matches:	5		

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.

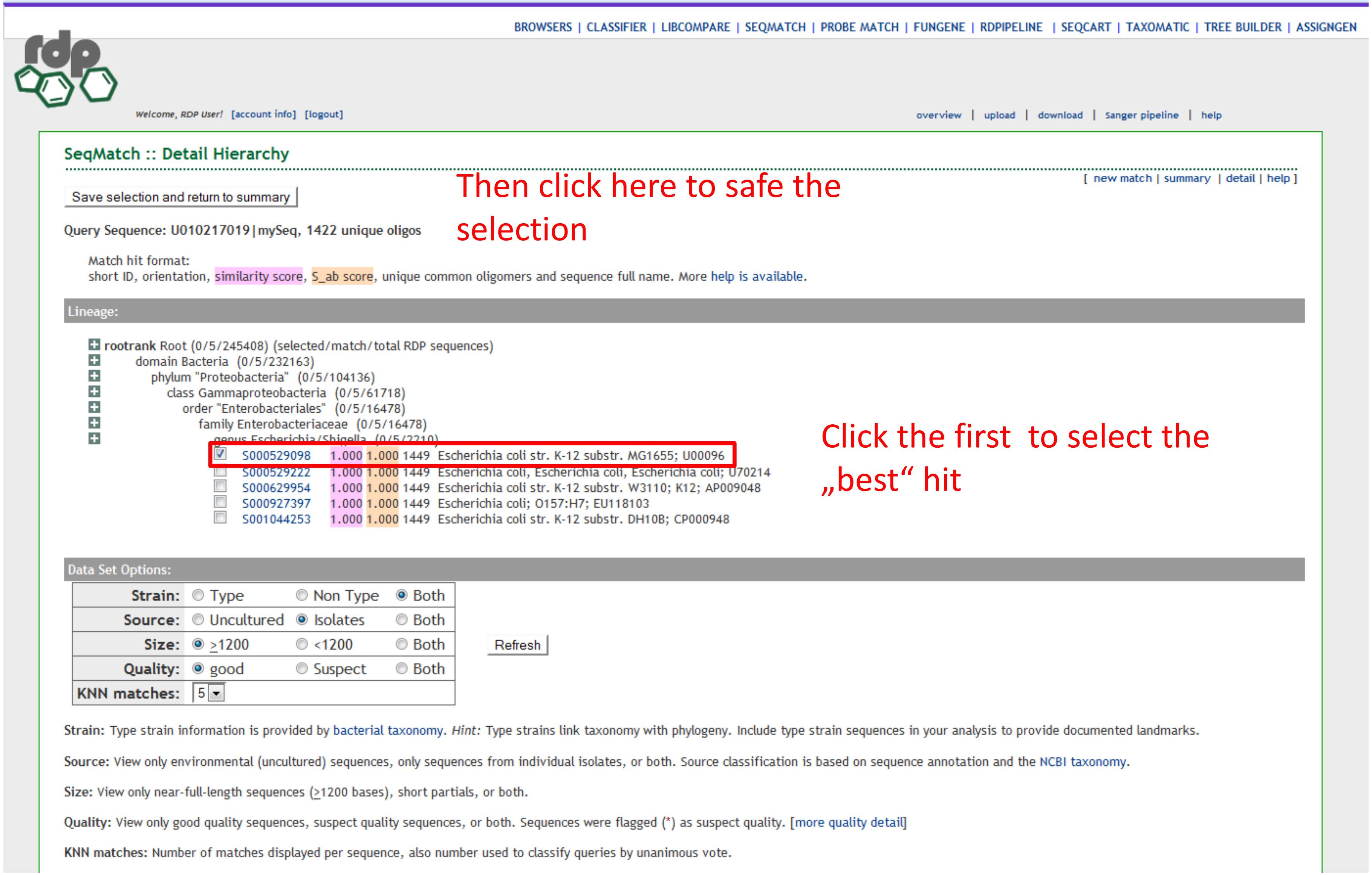
Source: View only environmental (uncultured) sequences, only sequences from individual isolates, or both. Source classification is based on sequence annotation and the NCBI taxonomy.

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. [more quality detail]

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

Seqmatch - Results



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

SeqMatch :: Detail Hierarchy

[new match | summary | detail | help]

Save selection and return to summary

Then click here to safe the selection

Query Sequence: U010217019|mySeq, 1422 unique oligos

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

Lineage:

```
rootrank Root (0/5/245408) (selected/match/total RDP sequences)
  domain Bacteria  (0/5/232163)
    phylum "Proteobacteria"  (0/5/104136)
      class Gammaproteobacteria  (0/5/61718)
        order "Enterobacteriales"  (0/5/16478)
          family Enterobacteriaceae  (0/5/16478)
            genus Escherichia/Shigella  (0/5/2210)
              ☑ S000529098  1.000 1.000 1449  Escherichia coli str. K-12 substr. MG1655; U00096
              □ S000529222  1.000 1.000 1449  Escherichia coli, Escherichia coli, Escherichia coli; U70214
              □ S000629954  1.000 1.000 1449  Escherichia coli str. K-12 substr. W3110; K12; AP009048
              □ S000927397  1.000 1.000 1449  Escherichia coli; O157:H7; EU118103
              □ S001044253  1.000 1.000 1449  Escherichia coli str. K-12 substr. DH10B; CP000948
```

Click the first to select the „best“ hit

Data Set Options:

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	● Isolates	○ Both
Size:	● >1200	○ <1200	○ Both
Quality:	● good	○ Suspect	○ Both
KNN matches:	5		

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.

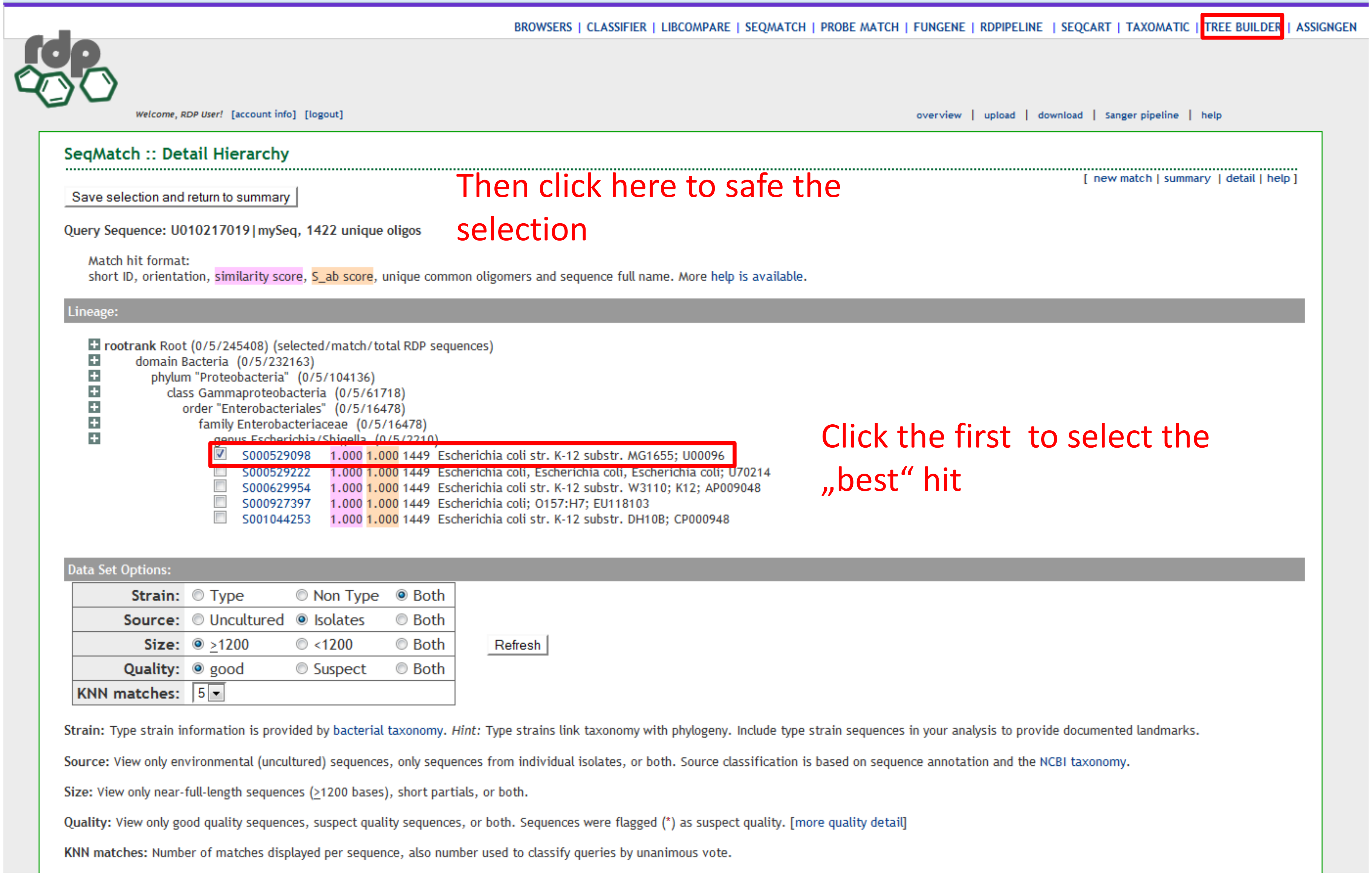
Quality: View only good quality sequences, suspect quality sequences, or both. Sequences were flagged (*) as suspect quality. [more quality detail]

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

Four steps in building a phylogenetic tree

1. Choosing the sequence type and set
2. Alignment of sequence data (in rdp this is done automatically)
3. Search for the best tree (in rdp a distance based method i.e. special form of the neighbor joining method is used)
4. Evaluation of tree reproducibility (bootstrapping)

Building a phylogenetic tree



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help
[new match | summary | detail | help]
Save selection and return to summary

SeqMatch :: Detail Hierarchy

Then click here to safe the selection

Query Sequence: U010217019|mySeq, 1422 unique oligos

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

Lineage:

rootrank Root (0/5/245408) (selected/match/total RDP sequences)
domain Bacteria (0/5/232163)
phylum "Proteobacteria" (0/5/104136)
class Gammaproteobacteria (0/5/61718)
order "Enterobacteriales" (0/5/16478)
family Enterobacteriaceae (0/5/16478)
genus Escherichia/Shigella (0/5/2210)

☑ S000529098 1.000 1.000 1449 Escherichia coli str. K-12 substr. MG1655; U00096
□ S000529222 1.000 1.000 1449 Escherichia coli, Escherichia coli, Escherichia coli; U70214
□ S000629954 1.000 1.000 1449 Escherichia coli str. K-12 substr. W3110; K12; AP009048
□ S000927397 1.000 1.000 1449 Escherichia coli; O157:H7; EU118103
□ S001044253 1.000 1.000 1449 Escherichia coli str. K-12 substr. DH10B; CP000948

Click the first to select the „best“ hit

Data Set Options:

Strain:	○ Type	○ Non Type	● Both
Source:	○ Uncultured	● Isolates	○ Both
Size:	● >1200	○ <1200	○ Both
Quality:	● good	○ Suspect	○ Both
KNN matches:	5		

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.

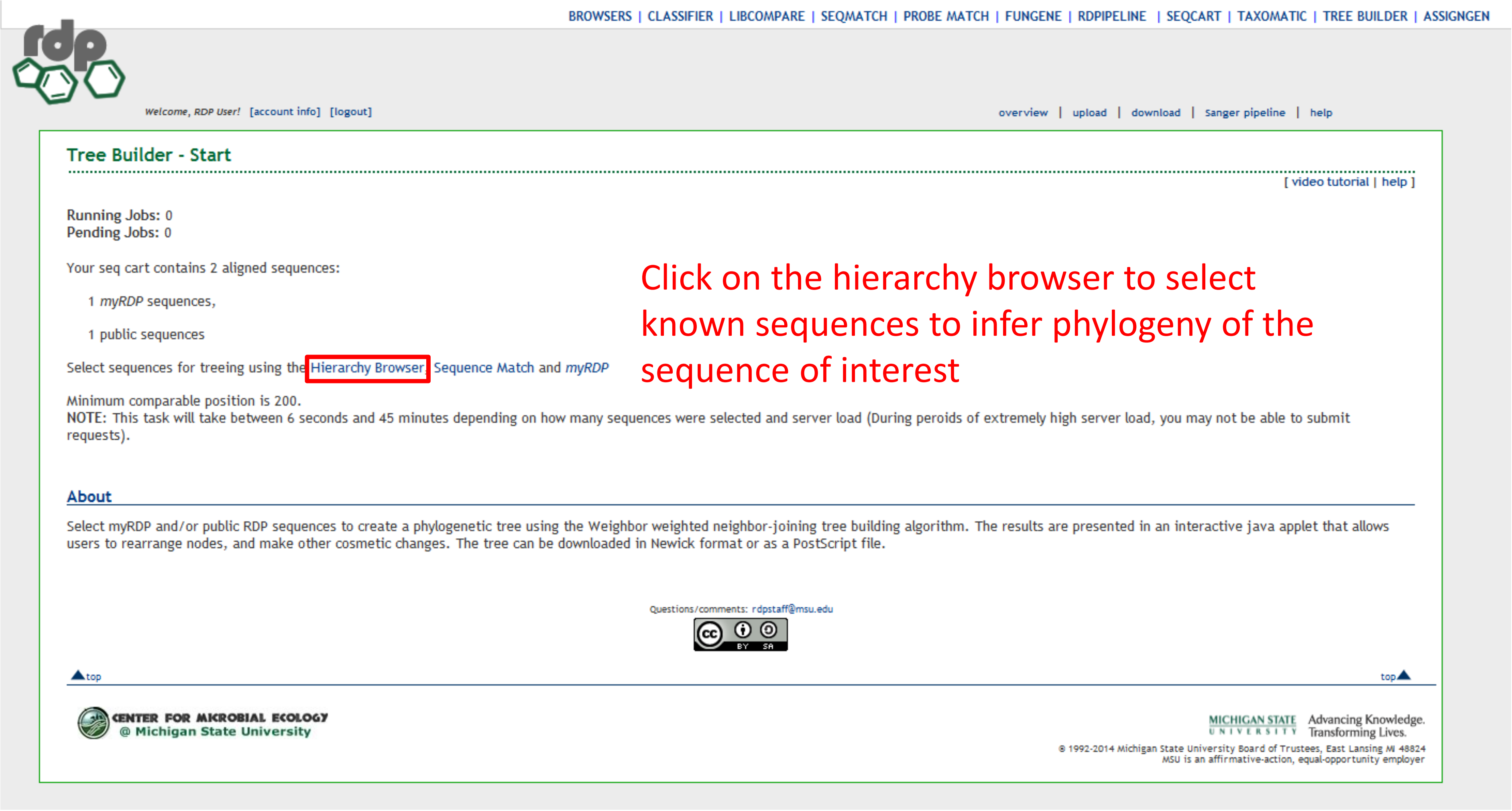
Source: View only environmental (uncultured) sequences, only sequences from individual isolates, or both. Source classification is based on sequence annotation and the NCBI taxonomy.

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. [more quality detail]

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

Building a phylogenetic tree



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help
[video tutorial | help]
top
top

Tree Builder - Start

Running Jobs: 0
Pending Jobs: 0

Your seq cart contains 2 aligned sequences:

1 *myRDP* sequences,

1 public sequences

Select sequences for treeing using the Hierarchy Browser, Sequence Match and *myRDP*

Minimum comparable position is 200.
NOTE: This task will take between 6 seconds and 45 minutes depending on how many sequences were selected and server load (During peroids of extremely high server load, you may not be able to submit requests).

About

Select myRDP and/or public RDP sequences to create a phylogenetic tree using the Weighbor weighted neighbor-joining tree building algorithm. The results are presented in an interactive java applet that allows users to rearrange nodes, and make other cosmetic changes. The tree can be downloaded in Newick format or as a PostScript file.

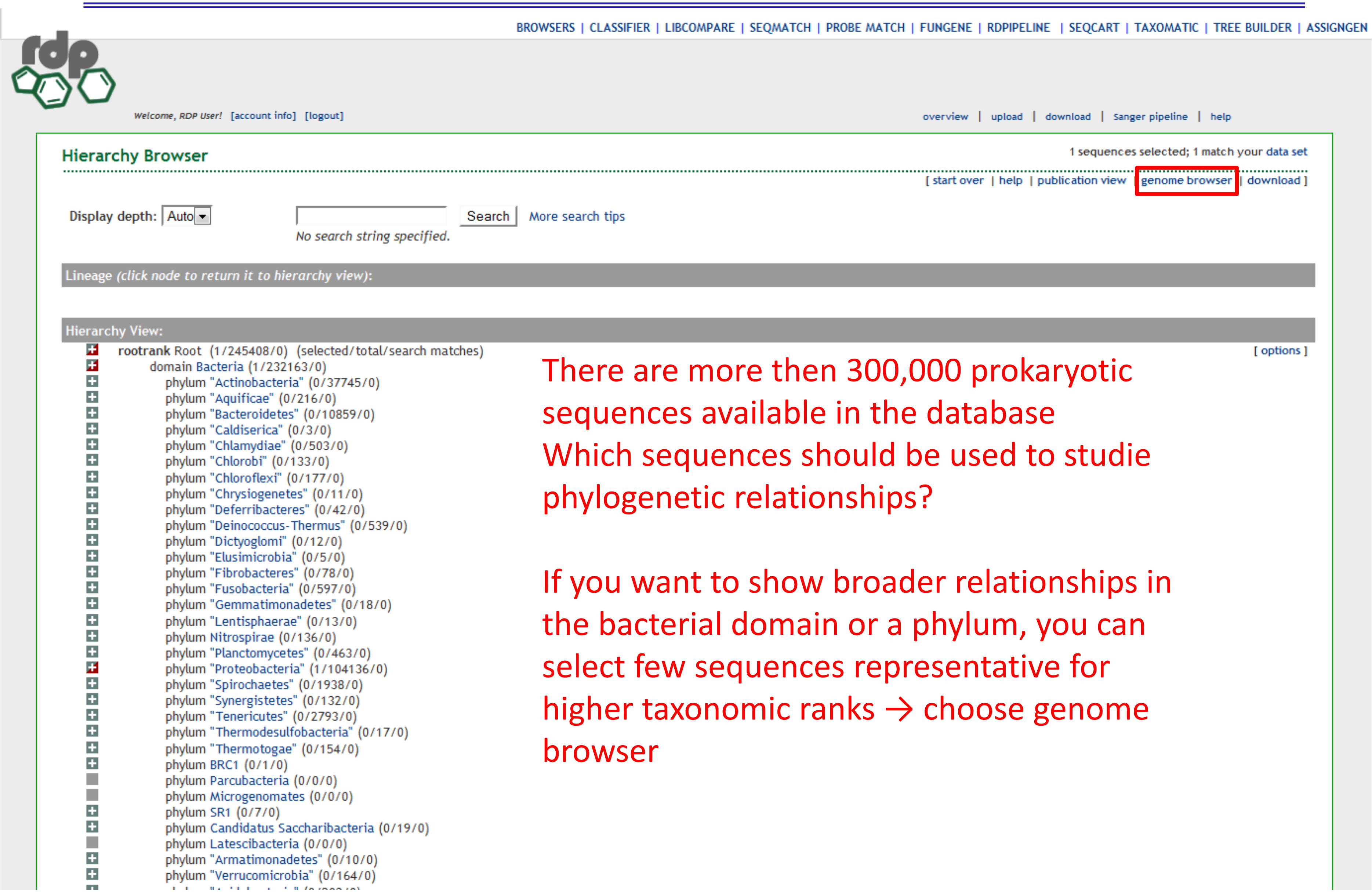
Questions/comments: rdpstaff@msu.edu

CENTER FOR MICROBIAL ECOLOGY
@ Michigan State University

MICHIGAN STATE UNIVERSITY Advancing Knowledge. Transforming Lives.

Click on the hierarchy browser to select known sequences to infer phylogeny of the sequence of interest

Building a phylogenetic tree



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Hierarchy Browser

1 sequences selected; 1 match your data set

[start over | help | publication view | genome browser | download]

Display depth: Auto ▾ Search More search tips

No search string specified.

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

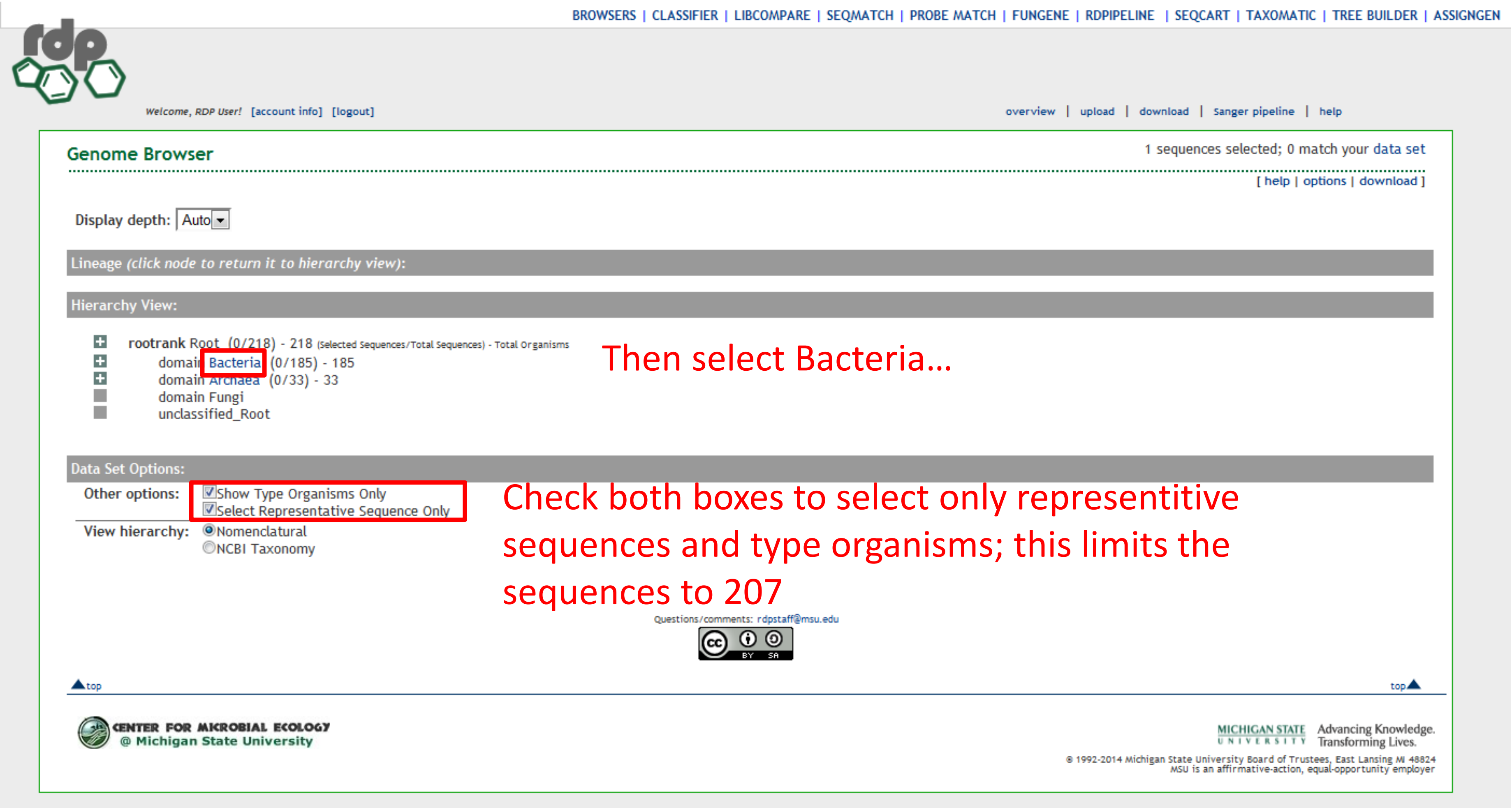
Hierarchy View: [options]

rootrank Root (1/245408/0) (selected/total/search matches)
domain Bacteria (1/232163/0)
phylum "Actinobacteria" (0/37745/0)
phylum "Aquificae" (0/216/0)
phylum "Bacteroidetes" (0/10859/0)
phylum "Caldiserica" (0/3/0)
phylum "Chlamydiae" (0/503/0)
phylum "Chlorobi" (0/133/0)
phylum "Chloroflexi" (0/177/0)
phylum "Chrysiogenetes" (0/11/0)
phylum "Deferribacteres" (0/42/0)
phylum "Deinococcus-Thermus" (0/539/0)
phylum "Dictyoglomi" (0/12/0)
phylum "Elusimicrobia" (0/5/0)
phylum "Fibrobacteres" (0/78/0)
phylum "Fusobacteria" (0/597/0)
phylum "Gemmatimonadetes" (0/18/0)
phylum "Lentisphaerae" (0/13/0)
phylum Nitrospirae (0/136/0)
phylum "Planctomycetes" (0/463/0)
phylum "Proteobacteria" (1/104136/0)
phylum "Spirochaetes" (0/1938/0)
phylum "Synergistetes" (0/132/0)
phylum "Tenericutes" (0/2793/0)
phylum "Thermodesulfobacteria" (0/17/0)
phylum "Thermotogae" (0/154/0)
phylum BRC1 (0/1/0)
phylum Parcubacteria (0/0/0)
phylum Microgenomates (0/0/0)
phylum SR1 (0/7/0)
phylum Candidatus Saccharibacteria (0/19/0)
phylum Latescibacteria (0/0/0)
phylum "Armatimonadetes" (0/10/0)
phylum "Verrucomicrobia" (0/164/0)

There are more then 300,000 prokaryotic sequences available in the database
Which sequences should be used to studie phylogenetic relationships?

If you want to show broader relationships in the bacterial domain or a phylum, you can select few sequences representative for higher taxonomic ranks → choose genome browser

Sequence selection



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

1 sequences selected; 0 match your data set

[help | options | download]

Display depth: Auto ▾

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

Hierarchy View:

- rootrank Root (0/218) - 218 (Selected Sequences/Total Sequences) - Total Organisms
  - domain Bacteria (0/185) - 185
  - domain Archaea (0/33) - 33
  - domain Fungi
  - unclassified_Root

Then select Bacteria…

Data Set Options:

Other options: ☑ Show Type Organisms Only
☑ Select Representative Sequence Only

View hierarchy: ● Nomenclatural
○ NCBI Taxonomy

Check both boxes to select only representitive sequences and type organisms; this limits the sequences to 207

Questions/comments: rdpstaff@msu.edu

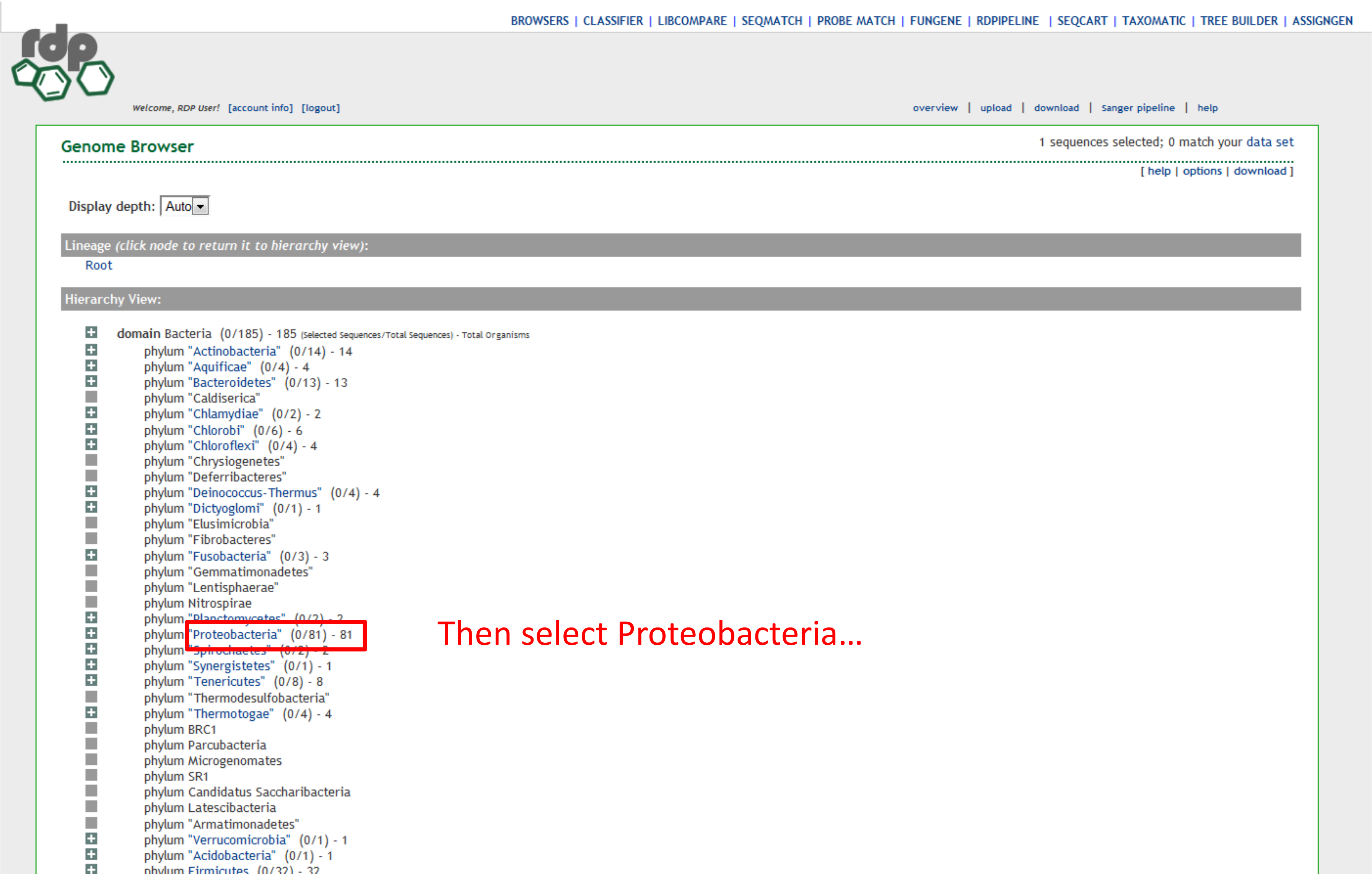
top

top

CENTER FOR MICROBIAL ECOLOGY
@ Michigan State University

MICHIGAN STATE UNIVERSITY Advancing Knowledge. Transforming Lives.

Sequence selection



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

1 sequences selected; 0 match your data set

[help | options | download]

Display depth: Auto ▾

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

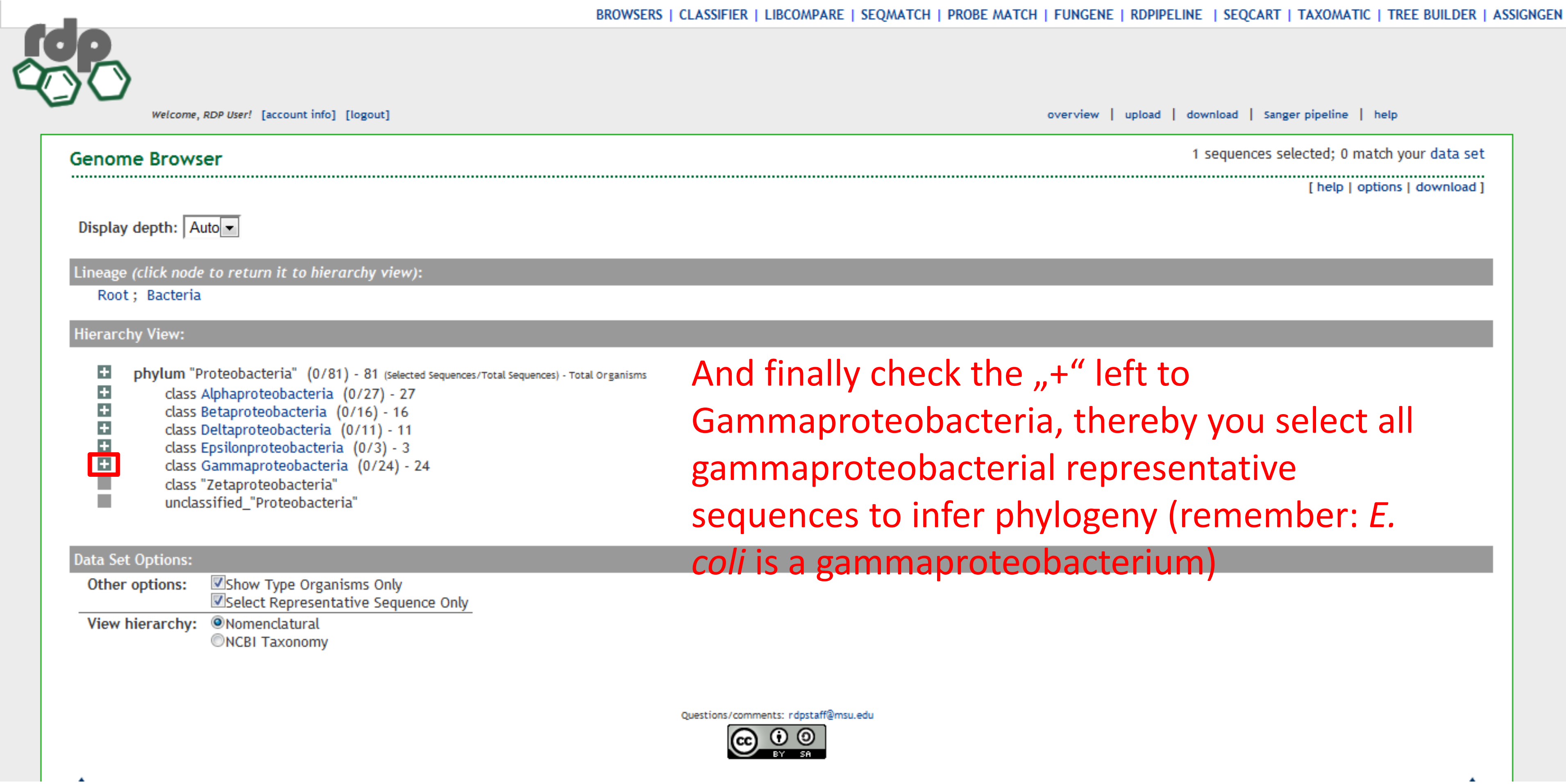
Root

Hierarchy View:

⊞ domain Bacteria (0/185) - 185 (Selected Sequences/Total Sequences) - Total Organisms
- ⊞ phylum "Actinobacteria" (0/14) - 14
- ⊞ phylum "Aquificae" (0/4) - 4
- ⊞ phylum "Bacteroidetes" (0/13) - 13
- ■ phylum "Caldiserica"
- ⊞ phylum "Chlamydiae" (0/2) - 2
- ⊞ phylum "Chlorobi" (0/6) - 6
- ⊞ phylum "Chloroflexi" (0/4) - 4
- ■ phylum "Chrysiogenetes"
- ■ phylum "Deferribacteres"
- ⊞ phylum "Deinococcus-Thermus" (0/4) - 4
- ⊞ phylum "Dictyoglomi" (0/1) - 1
- ■ phylum "Elusimicrobia"
- ■ phylum "Fibrobacteres"
- ⊞ phylum "Fusobacteria" (0/3) - 3
- ■ phylum "Gemmatimonadetes"
- ■ phylum "Lentisphaerae"
- ■ phylum Nitrospirae
- ⊞ phylum "Planctomycetes" (0/2) - 2
- ⊞ phylum "Proteobacteria" (0/81) - 81
- ⊞ phylum "Spirochaetes" (0/2) - 2
- ⊞ phylum "Synergistetes" (0/1) - 1
- ⊞ phylum "Tenericutes" (0/8) - 8
- ■ phylum "Thermodesulfobacteria"
- ⊞ phylum "Thermotogae" (0/4) - 4
- ■ phylum BRC1
- ■ phylum Parcubacteria
- ■ phylum Microgenomates
- ■ phylum SR1
- ■ phylum Candidatus Saccharibacteria
- ■ phylum Latescibacteria
- ■ phylum "Armatimonadetes"
- ⊞ phylum "Verrucomicrobia" (0/1) - 1
- ⊞ phylum "Acidobacteria" (0/1) - 1
- ⊞ phylum Firmicutes (0/32) - 32

Then select Proteobacteria…

Sequence selection



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Genome Browser

1 sequences selected; 0 match your data set

[help | options | download]

Display depth: Auto

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

Root ; Bacteria

Hierarchy View:

⊞ phylum "Proteobacteria" (0/81) - 81 (Selected Sequences/Total Sequences) - Total Organisms
⊞ class Alphaproteobacteria (0/27) - 27
⊞ class Betaproteobacteria (0/16) - 16
⊞ class Deltaproteobacteria (0/11) - 11
⊞ class Epsilonproteobacteria (0/3) - 3
⊞ class Gammaproteobacteria (0/24) - 24
class "Zetaproteobacteria"
unclassified_"Proteobacteria"

And finally check the „+“ left to Gammaproteobacteria, thereby you select all gammaproteobacterial representative sequences to infer phylogeny (remember: *E. coli* is a gammaproteobacterium)

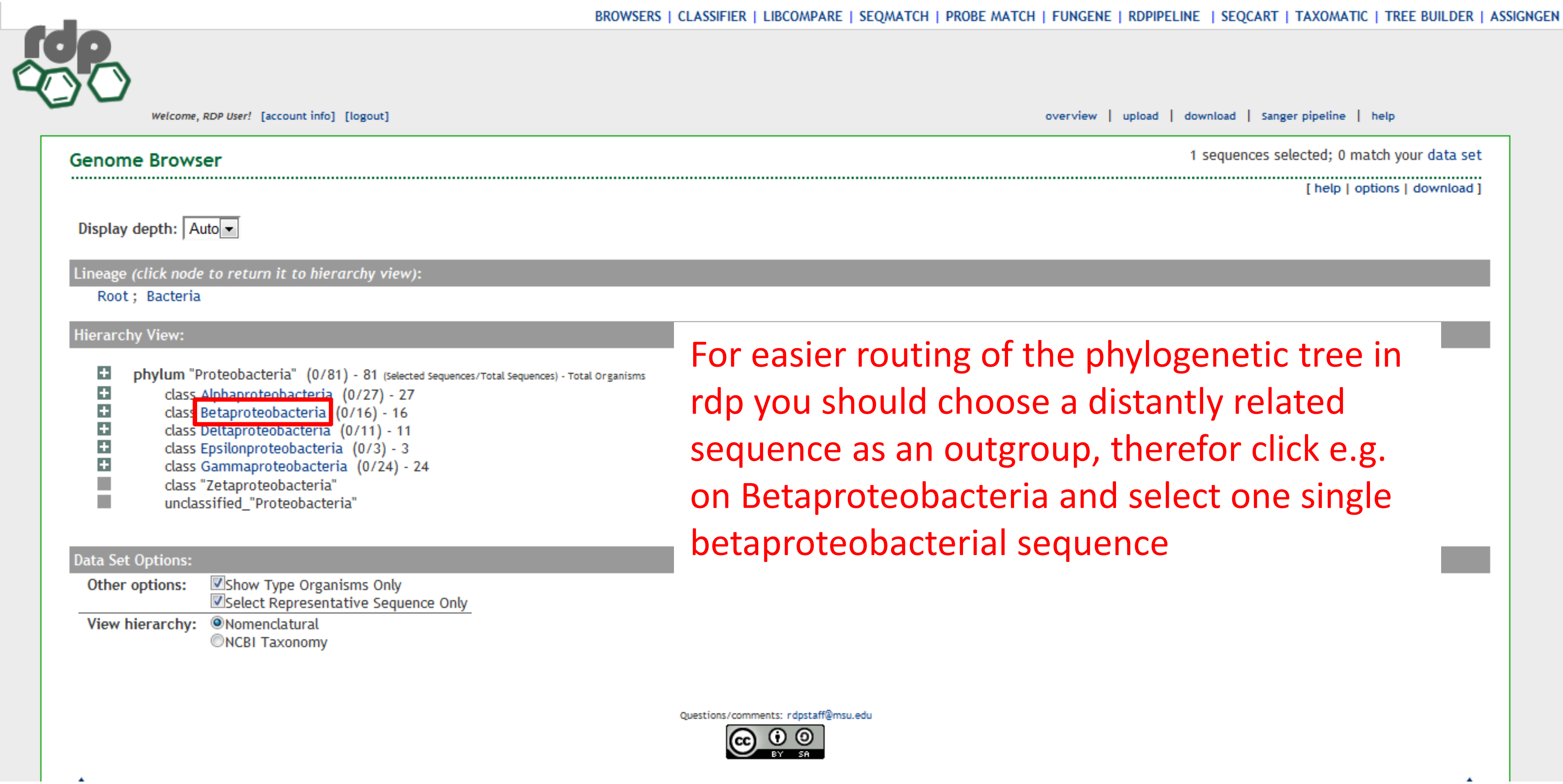
Data Set Options:

Other options: ☑ Show Type Organisms Only
☑ Select Representative Sequence Only

View hierarchy: ● Nomenclatural
○ NCBI Taxonomy

Questions/comments: rdpstaff@msu.edu

Sequence selection



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

1 sequences selected; 0 match your data set

[help | options | download]

Display depth: Auto

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

Root ; Bacteria

Hierarchy View:

- phylum "Proteobacteria" (0/81) - 81 (Selected Sequences/Total Sequences) - Total Organisms
  - class Alphaproteobacteria (0/27) - 27
  - class Betaproteobacteria (0/16) - 16
  - class Deltaproteobacteria (0/11) - 11
  - class Epsilonproteobacteria (0/3) - 3
  - class Gammaproteobacteria (0/24) - 24
  - class "Zetaproteobacteria"
  - unclassified_"Proteobacteria"

Data Set Options:

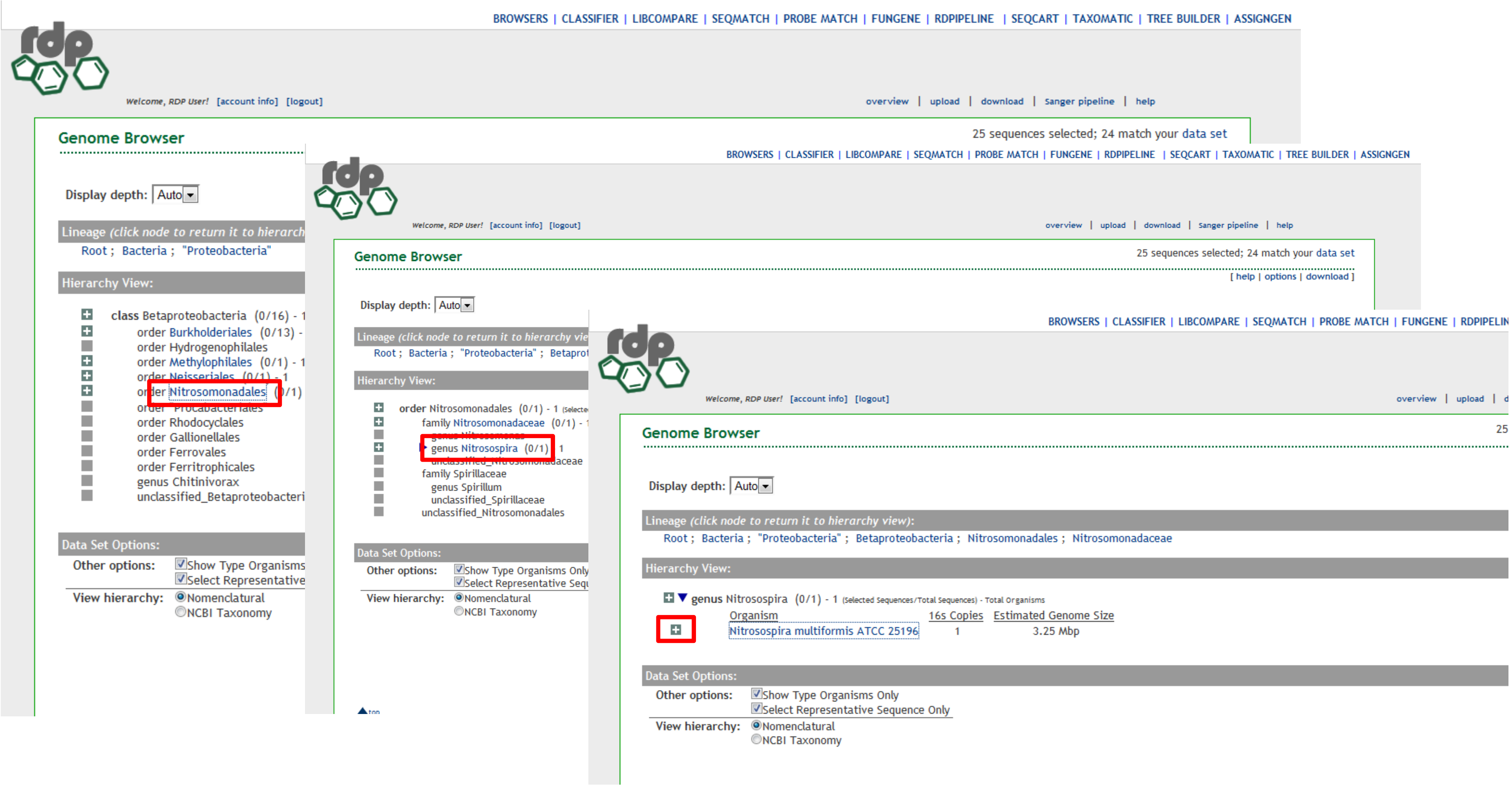
Other options: ☑ Show Type Organisms Only ☑ Select Representative Sequence Only

View hierarchy: ● Nomenclatural ○ NCBI Taxonomy

Questions/comments: rdpstaff@msu.edu

For easier routing of the phylogenetic tree in rdp you should choose a distantly related sequence as an outgroup, therefor click e.g. on Betaproteobacteria and select one single betaproteobacterial sequence

Sequence selection



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

25 sequences selected; 24 match your data set

Display depth: Auto

Lineage (click node to return it to hierarch

Root ; Bacteria ; "Proteobacteria"

Hierarchy View:

class Betaproteobacteria (0/16) - 1
order Burkholderiales (0/13) -
order Hydrogenophilales
order Methylophilales (0/1) - 1
order Neisseriales (0/1) - 1
order Nitrosomonadales (0/1)
order Procabacteriales
order Rhodocyclales
order Gallionellales
order Ferrovales
order Ferritrophicales
genus Chitinivorax
unclassified_Betaproteobacteri

Data Set Options:

Other options: ☑Show Type Organisms
☑Select Representative

View hierarchy: ●Nomenclatural
○NCBI Taxonomy

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

25 sequences selected; 24 match your data set

[help | options | download]

Display depth: Auto

Lineage (click node to return it to hierarchy vie

Root ; Bacteria ; "Proteobacteria" ; Betaprot

Hierarchy View:

order Nitrosomonadales (0/1) - 1 (Selecte
family Nitrosomonadaceae (0/1) - 1
genus Nitrosomonas
genus Nitrosospira (0/1) - 1
unclassified_Nitrosomonadaceae
family Spirillaceae
genus Spirillum
unclassified_Spirillaceae
unclassified_Nitrosomonadales

Data Set Options:

Other options: ☑Show Type Organisms Only
☑Select Representative Sequ

View hierarchy: ●Nomenclatural
○NCBI Taxonomy

▲top

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELIN

Welcome, RDP User! [account info] [logout]

overview | upload | d

Genome Browser

25

Display depth: Auto

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

Root ; Bacteria ; "Proteobacteria" ; Betaproteobacteria ; Nitrosomonadales ; Nitrosomonadaceae

Hierarchy View:

genus Nitrosospira (0/1) - 1 (Selected Sequences/Total Sequences) - Total Organisms

	Organism	16s Copies	Estimated Genome Size
⊞	Nitrosospira multiformis ATCC 25196	1	3.25 Mbp

Data Set Options:

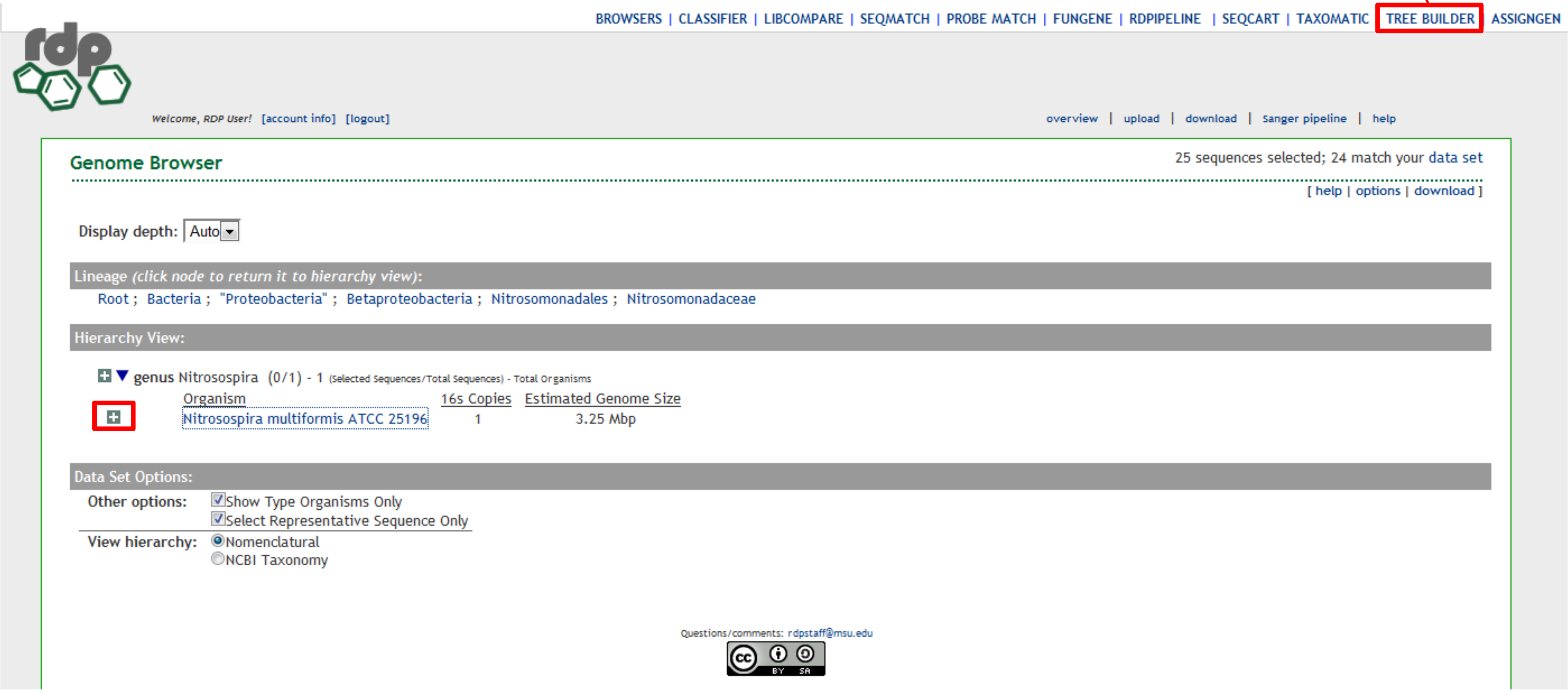
Other options: ☑Show Type Organisms Only
☑Select Representative Sequence Only

View hierarchy: ●Nomenclatural
○NCBI Taxonomy

Nitrosomonadales → Nitrosospira → klick the „+“ left to Nitrosospira multiformis

Sequence selection

Then click on tree builder



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN

Welcome, RDP User! [account info] [logout]

overview | upload | download | sanger pipeline | help

Genome Browser

25 sequences selected; 24 match your data set

[help | options | download]

Display depth: Auto

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

Root ; Bacteria ; "Proteobacteria" ; Betaproteobacteria ; Nitrosomonadales ; Nitrosomonadaceae

Hierarchy View:

⊞ ▼ genus Nitrosospira (0/1) - 1 (Selected Sequences/Total Sequences) - Total Organisms

	Organism	16s Copies	Estimated Genome Size
⊞	Nitrosospira multiformis ATCC 25196	1	3.25 Mbp

Data Set Options:

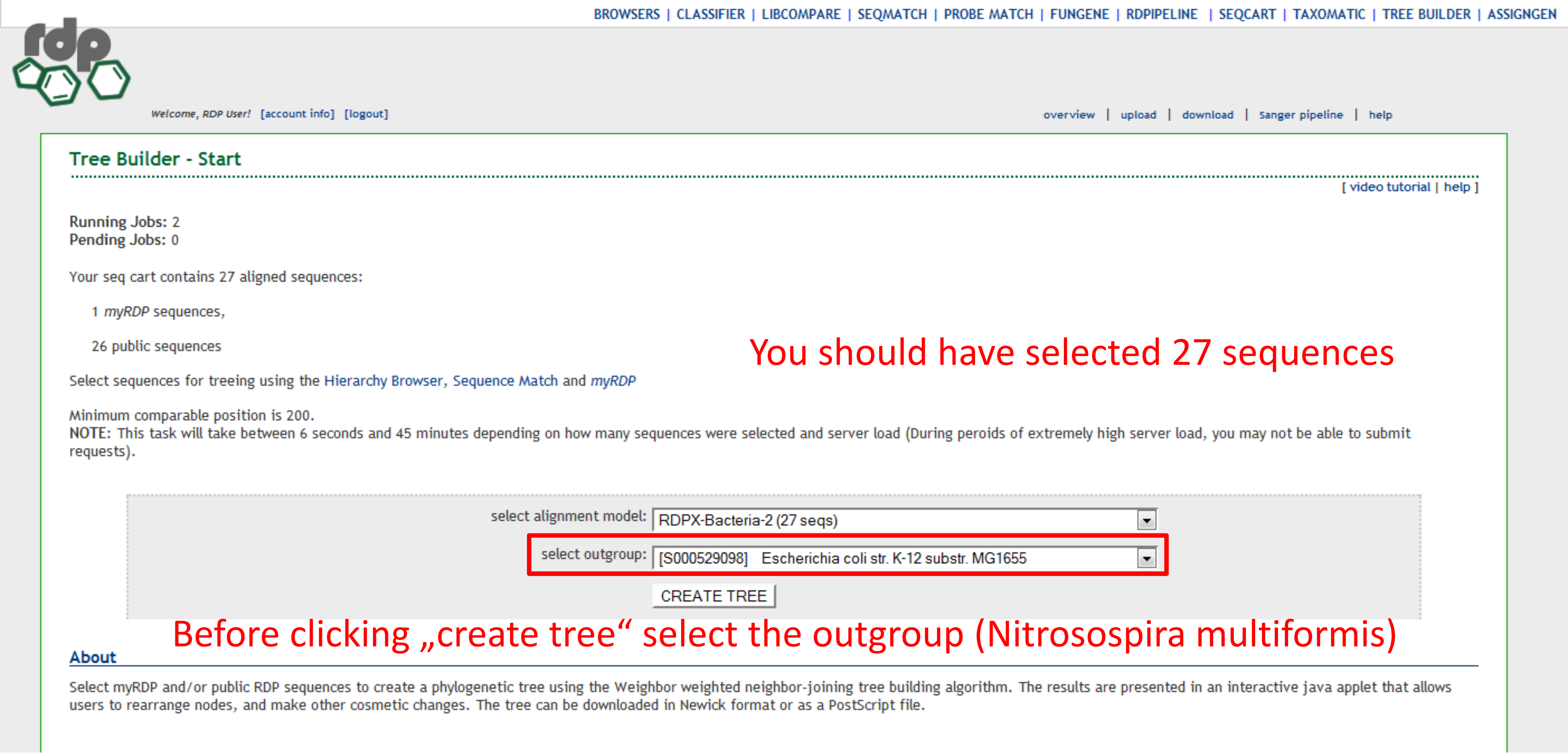
Other options: ☑ Show Type Organisms Only
☑ Select Representative Sequence Only

View hierarchy: ● Nomenclatural ○ NCBI Taxonomy

Questions/comments: rdpstaff@msu.edu

Nitrosomonadales → Nitrosospira → klick the „+“ left to Nitrosospira multiformis

Building the phylogenetic tree – Tree builder



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Tree Builder - Start

[video tutorial | help]

Running Jobs: 2
Pending Jobs: 0

Your seq cart contains 27 aligned sequences:

1 *myRDP* sequences,

26 public sequences

Select sequences for treeing using the Hierarchy Browser, Sequence Match and *myRDP*

Minimum comparable position is 200.
NOTE: This task will take between 6 seconds and 45 minutes depending on how many sequences were selected and server load (During peroids of extremely high server load, you may not be able to submit requests).

select alignment model: RDPX-Bacteria-2 (27 seqs)

select outgroup: [S000529098] Escherichia coli str. K-12 substr. MG1655

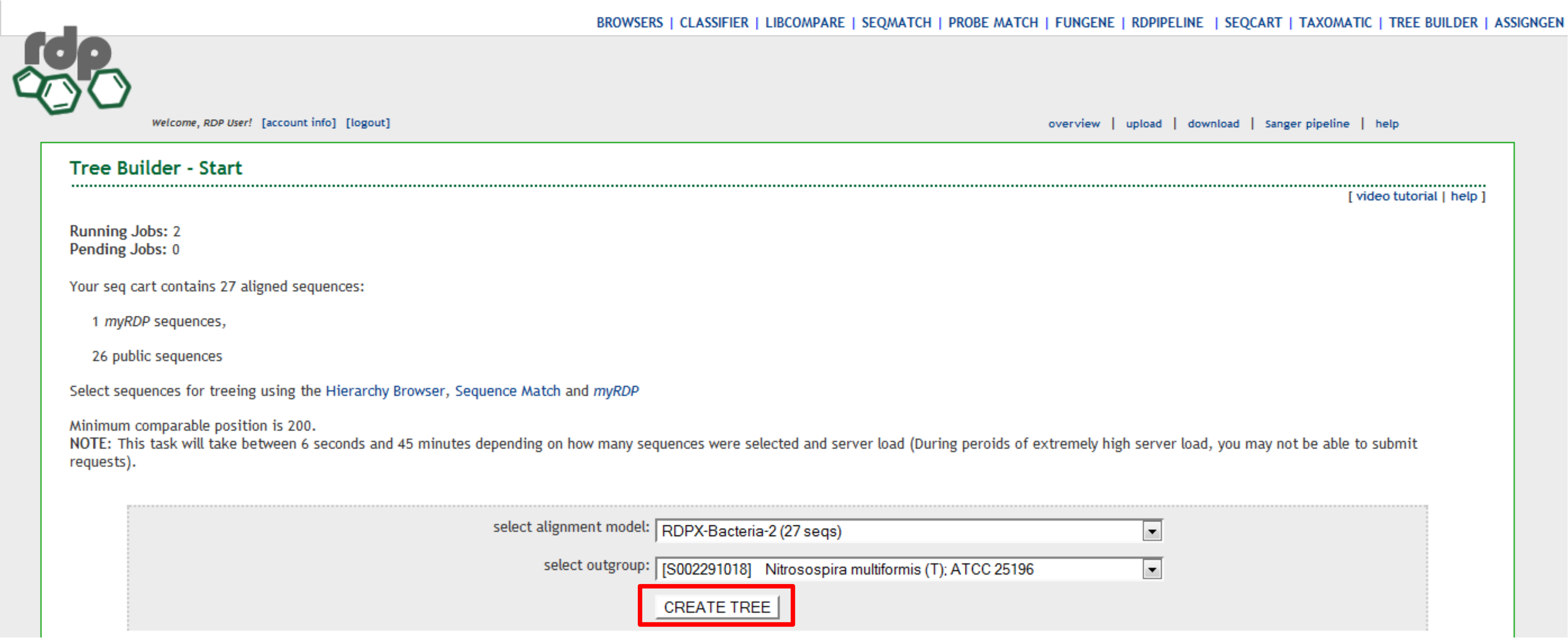
About

Select myRDP and/or public RDP sequences to create a phylogenetic tree using the Weighbor weighted neighbor-joining tree building algorithm. The results are presented in an interactive java applet that allows users to rearrange nodes, and make other cosmetic changes. The tree can be downloaded in Newick format or as a PostScript file.

You should have selected 27 sequences

Before clicking „create tree“ select the outgroup (Nitrosospira multiformis)

Building the phylogenetic tree – Tree builder



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIGNGEN
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Tree Builder - Start

[video tutorial | help]

Running Jobs: 2
Pending Jobs: 0

Your seq cart contains 27 aligned sequences:

1 *myRDP* sequences,

26 public sequences

Select sequences for treeing using the Hierarchy Browser, Sequence Match and *myRDP*

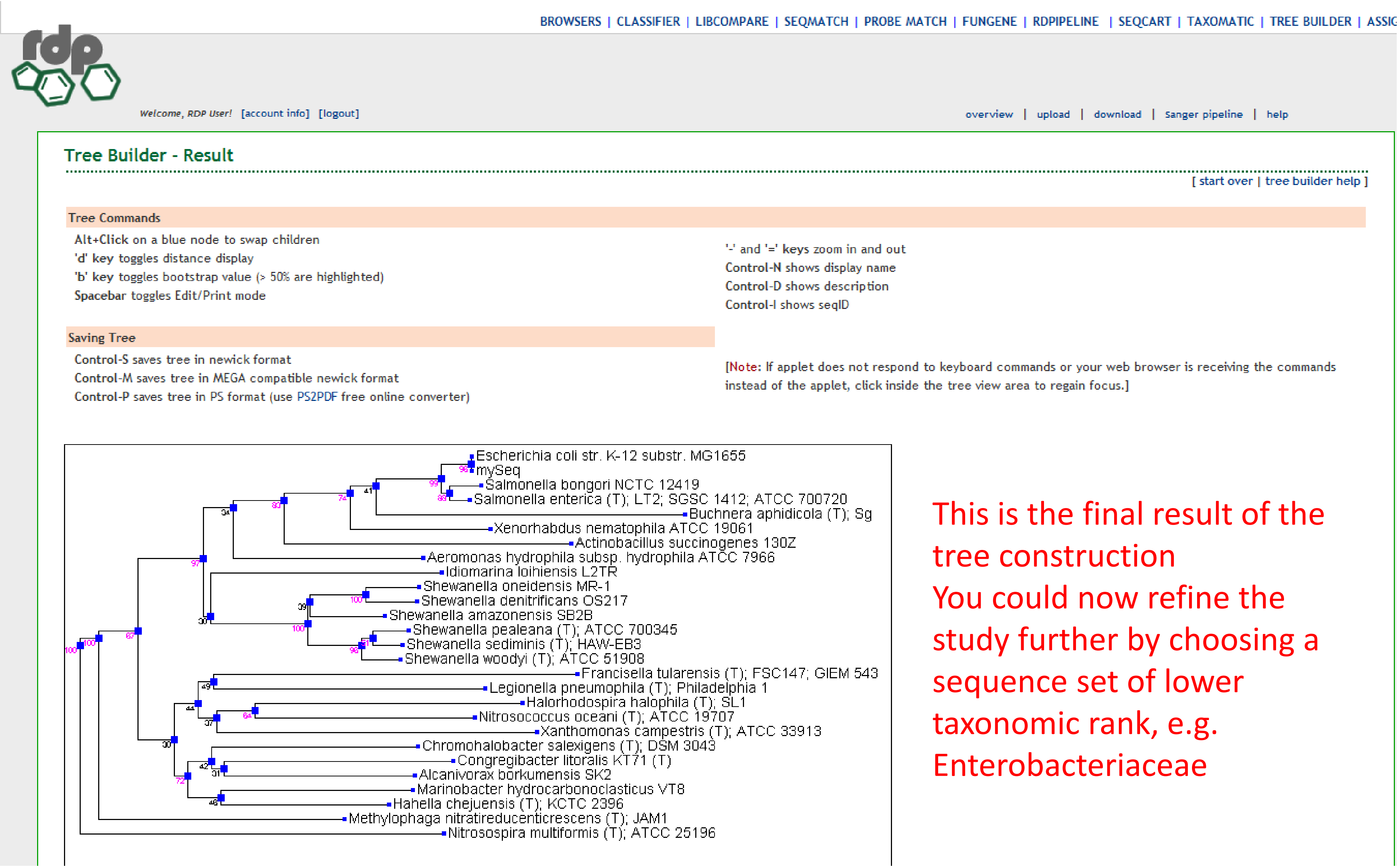
Minimum comparable position is 200.
NOTE: This task will take between 6 seconds and 45 minutes depending on how many sequences were selected and server load (During peroids of extremely high server load, you may not be able to submit requests).

select alignment model: RDPX-Bacteria-2 (27 seqs)

select outgroup: [S002291018] Nitrosospira multiformis (T); ATCC 25196

Then create the tree (confirm all requests from Java etc.), this might take a few minutes

The final tree



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIG
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Tree Builder - Result

[start over | tree builder help]

Tree Commands

Alt+Click on a blue node to swap children
'd' key toggles distance display
'b' key toggles bootstrap value (> 50% are highlighted)
Spacebar toggles Edit/Print mode

'-' and '=' keys zoom in and out
Control-N shows display name
Control-D shows description
Control-I shows seqID

Saving Tree

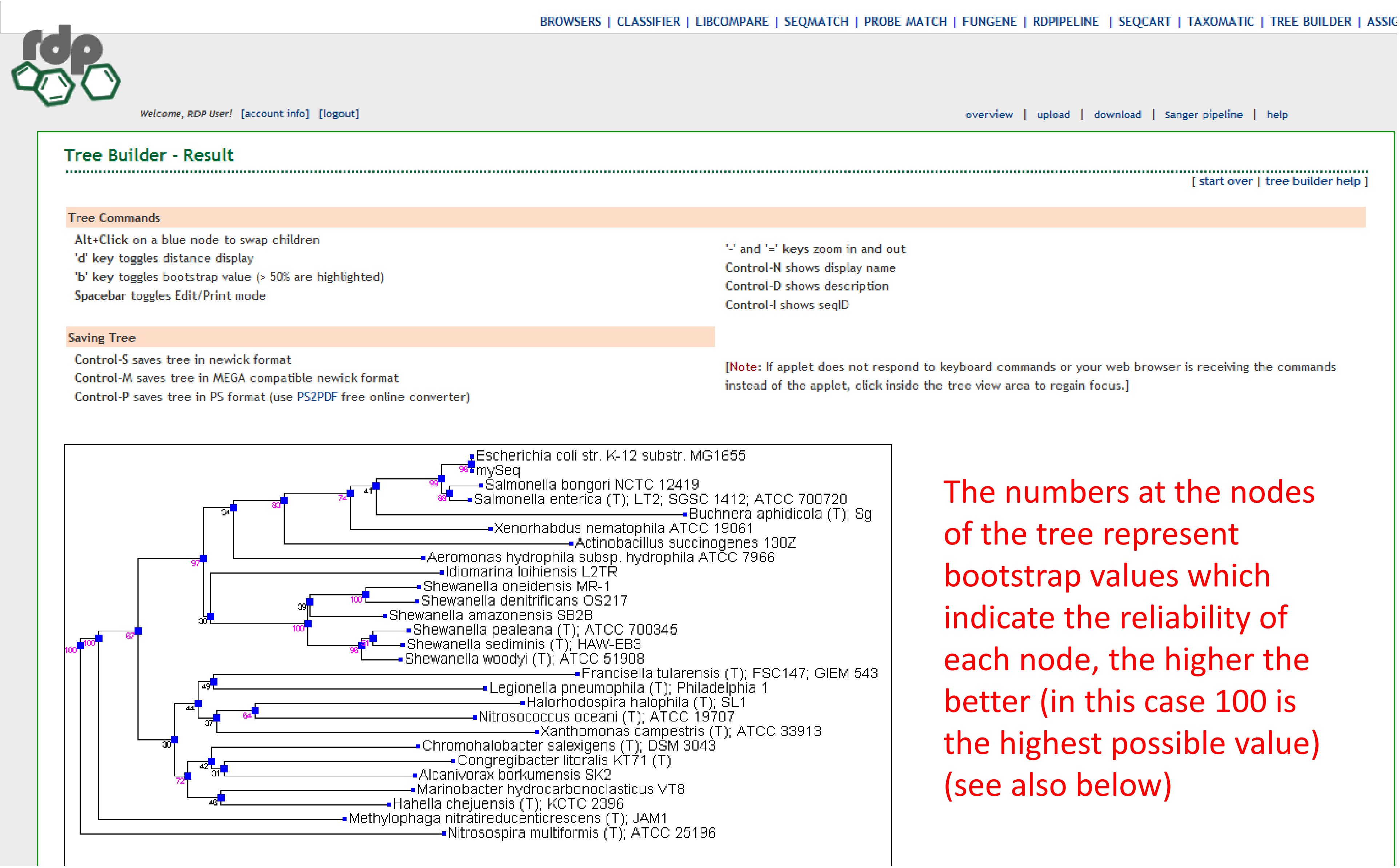
Control-S saves tree in newick format
Control-M saves tree in MEGA compatible newick format
Control-P saves tree in PS format (use PS2PDF free online converter)

[Note: If applet does not respond to keyboard commands or your web browser is receiving the commands instead of the applet, click inside the tree view area to regain focus.]

Escherichia coli str. K-12 substr. MG1655
mySeq
Salmonella bongori NCTC 12419
Salmonella enterica (T); LT2; SGSC 1412; ATCC 700720
Buchnera aphidicola (T); Sg
Xenorhabdus nematophila ATCC 19061
Actinobacillus succinogenes 130Z
Aeromonas hydrophila subsp. hydrophila ATCC 7966
Idiomarina loihiensis L2TR
Shewanella oneidensis MR-1
Shewanella denitrificans OS217
Shewanella amazonensis SB2B
Shewanella pealeana (T); ATCC 700345
Shewanella sediminis (T); HAW-EB3
Shewanella woodyi (T); ATCC 51908
Francisella tularensis (T); FSC147; GIEM 543
Legionella pneumophila (T); Philadelphia 1
Halorhodospira halophila (T); SL1
Nitrosococcus oceani (T); ATCC 19707
Xanthomonas campestris (T); ATCC 33913
Chromohalobacter salexigens (T); DSM 3043
Congregibacter litoralis KT71 (T)
Alcanivorax borkumensis SK2
Marinobacter hydrocarbonoclasticus VT8
Hahella chejuensis (T); KCTC 2396
Methylophaga nitratireducenticrescens (T); JAM1
Nitrosospira multiformis (T); ATCC 25196
98
99
88
41
74
80
34
97
100
39
30
100
51
98
100
100
67
49
44
64
37
30
42
31
72
46

This is the final result of the tree construction
You could now refine the study further by choosing a sequence set of lower taxonomic rank, e.g. Enterobacteriaceae

The final tree



BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | FUNGENE | RDPIPELINE | SEQCART | TAXOMATIC | TREE BUILDER | ASSIG
rdp
Welcome, RDP User! [account info] [logout]
overview | upload | download | sanger pipeline | help

Tree Builder - Result

[start over | tree builder help]

Tree Commands

Alt+Click on a blue node to swap children
'd' key toggles distance display
'b' key toggles bootstrap value (> 50% are highlighted)
Spacebar toggles Edit/Print mode

'-' and '=' keys zoom in and out
Control-N shows display name
Control-D shows description
Control-I shows seqID

Saving Tree

Control-S saves tree in newick format
Control-M saves tree in MEGA compatible newick format
Control-P saves tree in PS format (use PS2PDF free online converter)

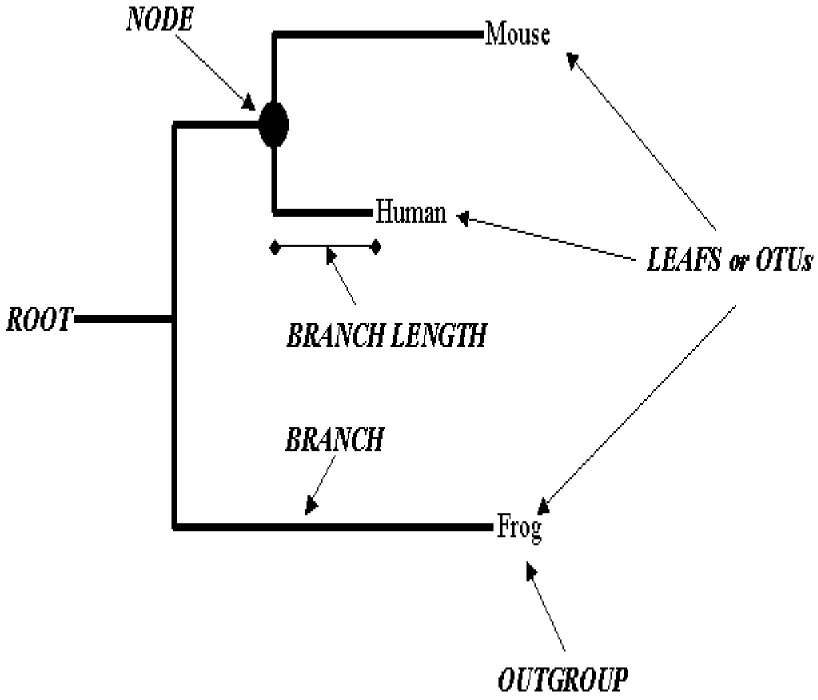
[Note: If applet does not respond to keyboard commands or your web browser is receiving the commands instead of the applet, click inside the tree view area to regain focus.]

Escherichia coli str. K-12 substr. MG1655
mySeq
Salmonella bongori NCTC 12419
Salmonella enterica (T); LT2; SGSC 1412; ATCC 700720
Buchnera aphidicola (T); Sg
Xenorhabdus nematophila ATCC 19061
Actinobacillus succinogenes 130Z
Aeromonas hydrophila subsp. hydrophila ATCC 7966
Idiomarina loihiensis L2TR
Shewanella oneidensis MR-1
Shewanella denitrificans OS217
Shewanella amazonensis SB2B
Shewanella pealeana (T); ATCC 700345
Shewanella sediminis (T); HAW-EB3
Shewanella woodyi (T); ATCC 51908
Francisella tularensis (T); FSC147; GIEM 543
Legionella pneumophila (T); Philadelphia 1
Halorhodospira halophila (T); SL1
Nitrosococcus oceani (T); ATCC 19707
Xanthomonas campestris (T); ATCC 33913
Chromohalobacter salexigens (T); DSM 3043
Congregibacter litoralis KT71 (T)
Alcanivorax borkumensis SK2
Marinobacter hydrocarbonoclasticus VT8
Hahella chejuensis (T); KCTC 2396
Methylophaga nitratireducenticrescens (T); JAM1
Nitrosospira multiformis (T); ATCC 25196
98
99
89
41
74
80
34
97
100
39
30
100
51
98
67
100
100
49
44
64
37
30
42
31
72
46

The numbers at the nodes of the tree represent bootstrap values which indicate the reliability of each node, the higher the better (in this case 100 is the highest possible value) (see also below)

Some further informations: Reading Your Tree

- There’s a lot of vocabulary in a tree
- **Nodes** correspond to common ancestors
- The **root** is the oldest ancestor
  - Often artificial
  - Only meaningful with a good outgroup
- Trees can be un-rooted
- Branch lengths are only meaningful when the tree is scaled and refer to the degree of differences



Building a phylogenetic Tree

- There are two types of tree-reconstruction methods
  - Distance-based methods
  - Statistical methods

- Statistical methods are the most accurate
  - Maximum likelihood of success
  - Parsimony

- Statistical methods take more time
  - Limited to small datasets

Distance-based Methods for Tree Reconstruction

- Distance-based methods are the most popular
  - Neighbor Joining (NJ)
  - UPGMA

- Distance-based methods involve 2 steps:
  - Measure the distances between pairs of sequences in the MSA
  - Transform the distance matrix into a tree

Bootstrapping

- Use bootstrapping to verify the solidity of each node
- ClustalW and Phylip do bootstrap operations automatically
- Bootstrapping involves these steps:
  - Select a subset of your MSA
  - Redo the tree
  - Repeat this operation N times (100 or 1000 times if you can)
  - Compute a consensus tree of the N trees
  - Measure how many of the N trees agree with the consensus tree on each node

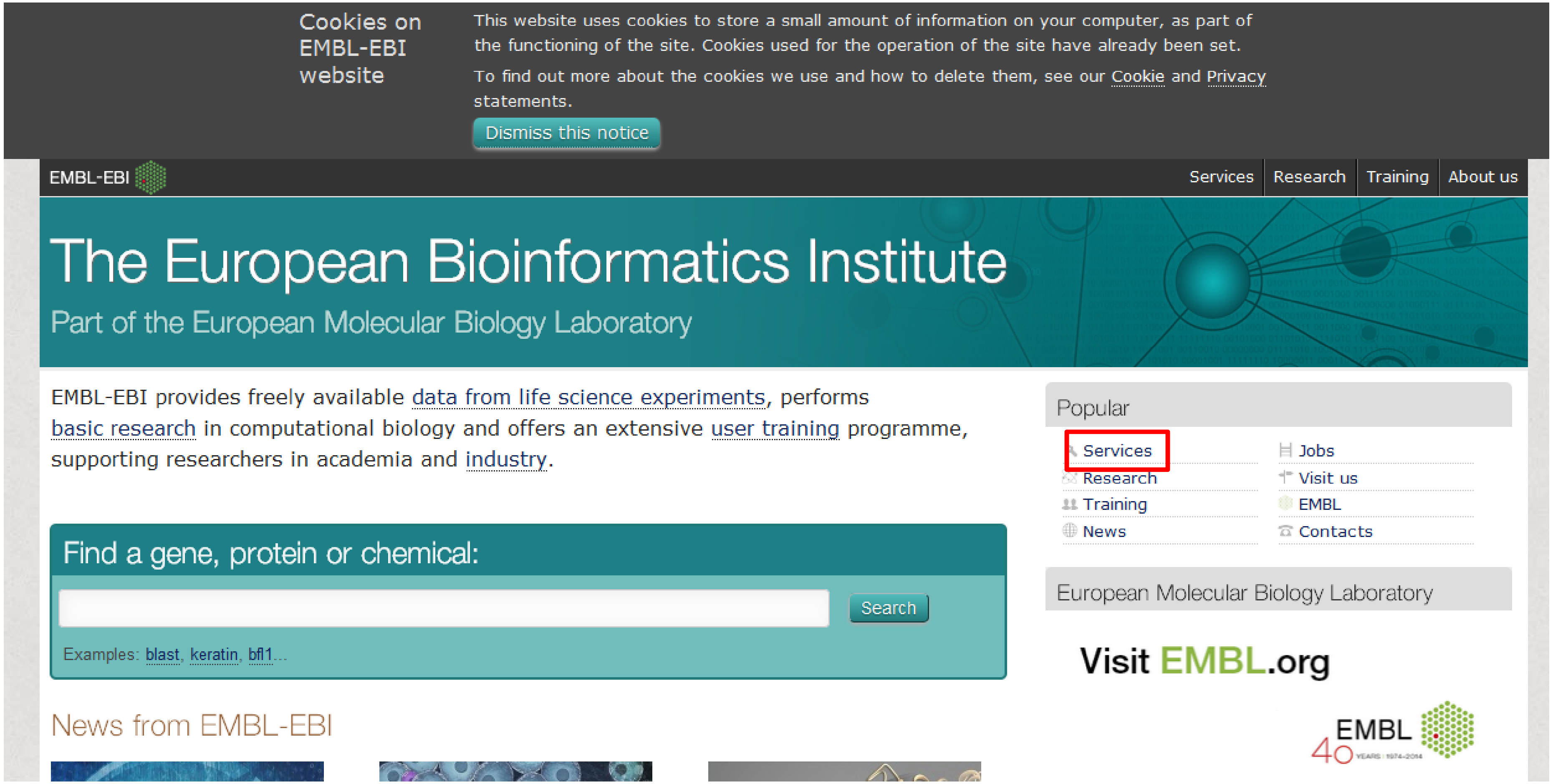
- Each node gets a bootstrap figure between 0 and N
- High bootstrap ⇔ good node

Doing a sequence alignment

- Different alignment programms available, e.g. Clustal, Muscel, Kalign etc.
- Can be downloaded as stand alone software (expasy.org)
- Or run on servers, e.g. ebi-embl (http://www.ebi.ac.uk/)

Doing a sequence alignment

- Go to http://www.ebi.ac.uk/



Cookies on EMBL-EBI website

This website uses cookies to store a small amount of information on your computer, as part of the functioning of the site. Cookies used for the operation of the site have already been set.

To find out more about the cookies we use and how to delete them, see our Cookie and Privacy statements.

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

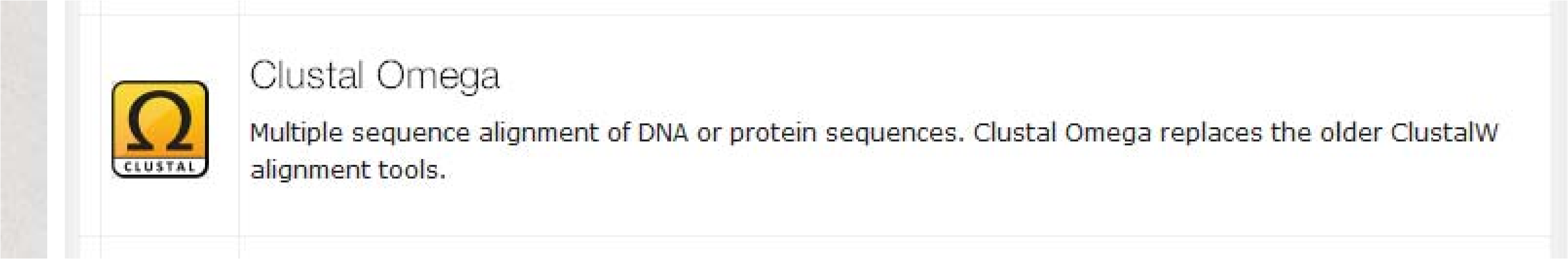
Examples: blast, keratin, bfl1...

News from EMBL-EBI

Choose services and then DNA & RNA

Doing a sequence alignment

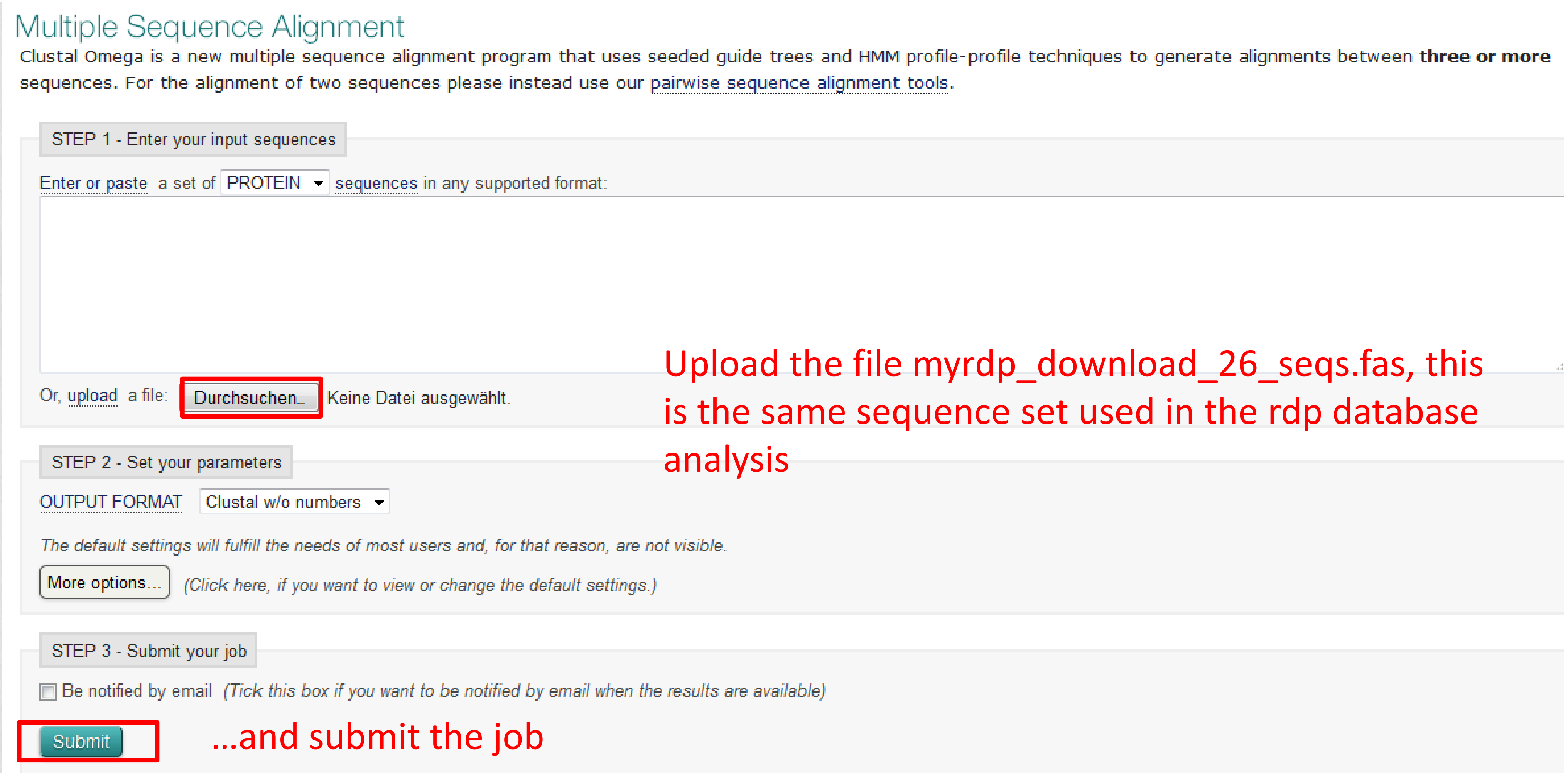
- Scroll down the page and select Clustal Omega



Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Doing a sequence alignment



Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN sequences in any supported format:

Or, upload a file: Durchsuchen... Keine Datei ausgewählt.

STEP 2 - Set your parameters

OUTPUT FORMAT Clustal w/o numbers

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

STEP 3 - Submit your job

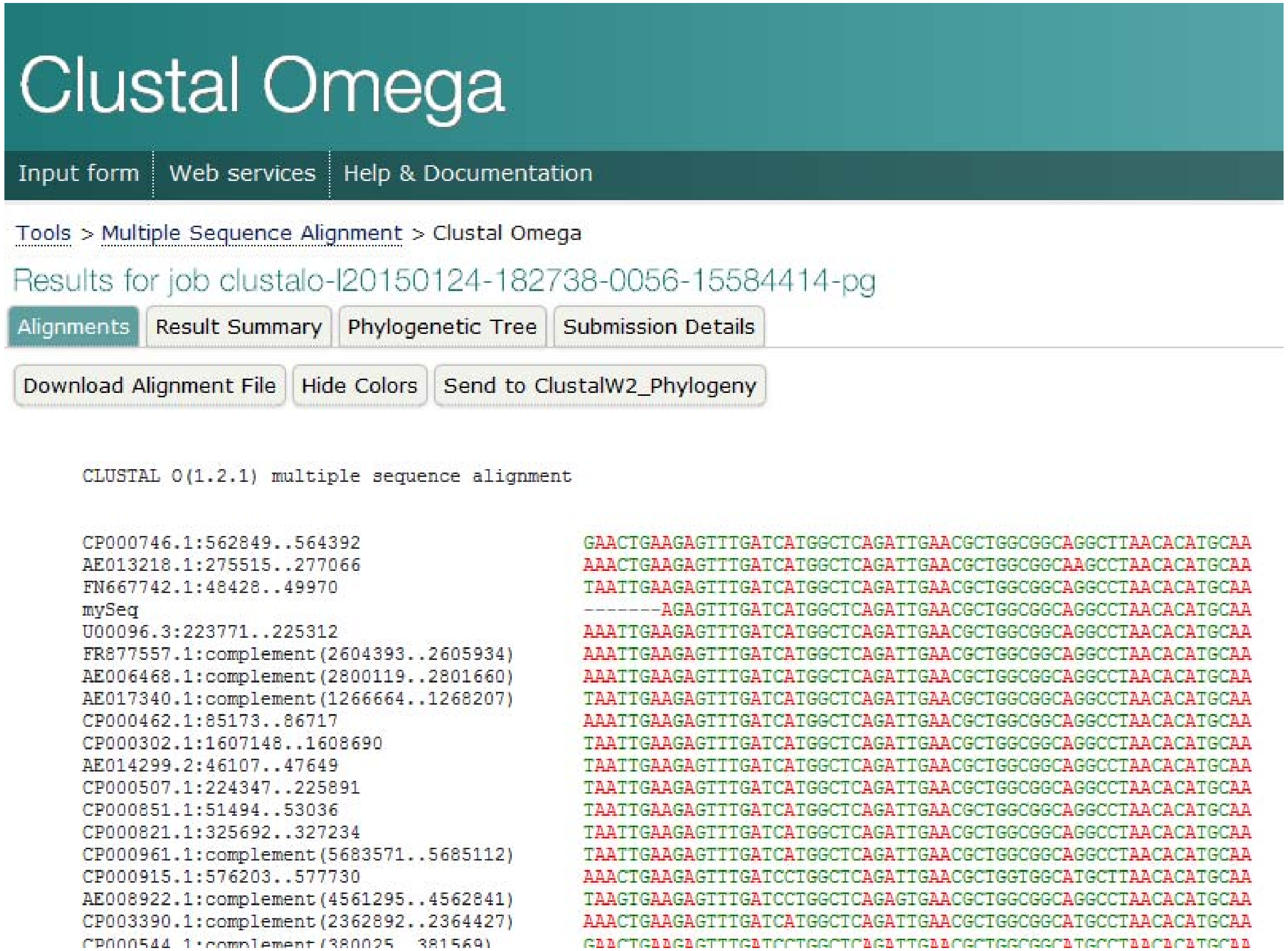
□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Upload the file myrdp_download_26_seqs.fas, this is the same sequence set used in the rdp database analysis

...and submit the job

Doing a sequence alignment



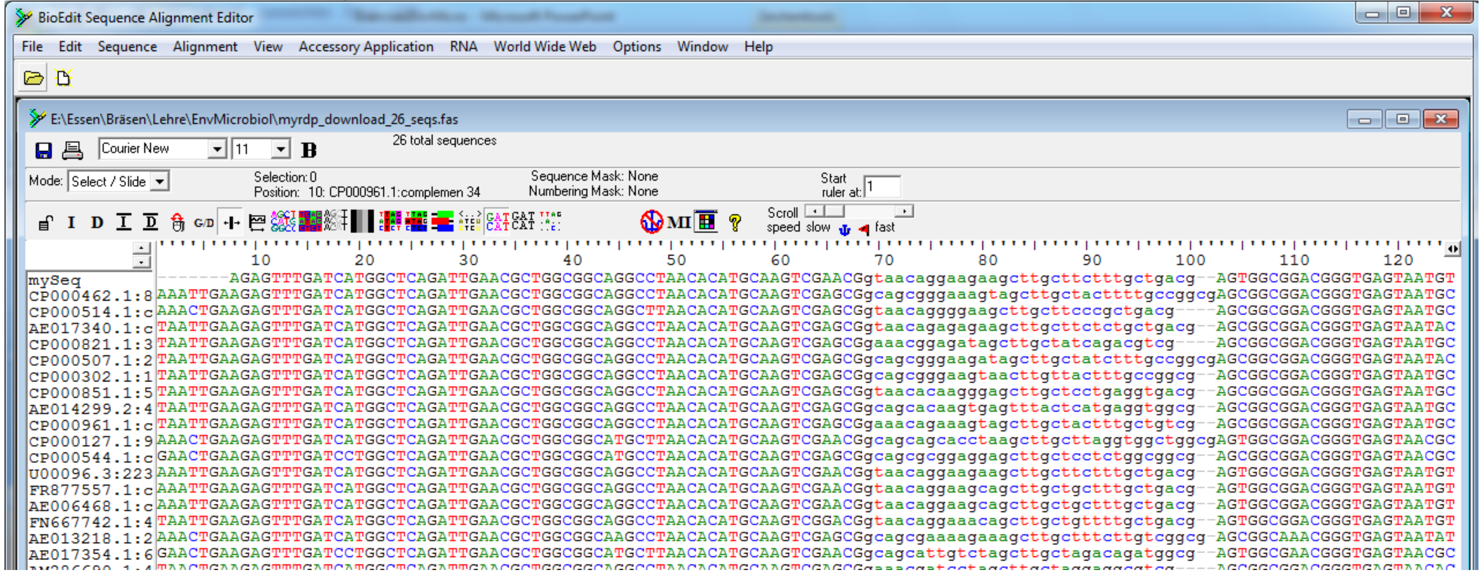
Results for job clustalo-I20150124-182738-0056-15584414-pg

```
CLUSTAL O(1.2.1) multiple sequence alignment


CP000746.1:562849..564392                    GAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTTAACACATGCAA
AE013218.1:275515..277066                    AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAAGCCTAACACATGCAA
FN667742.1:48428..49970                      TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
mySeq                                        -------AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
U00096.3:223771..225312                      AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
FR877557.1:complement(2604393..2605934)      AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
AE006468.1:complement(2800119..2801660)      AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
AE017340.1:complement(1266664..1268207)      TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000462.1:85173..86717                      AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000302.1:1607148..1608690                  TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
AE014299.2:46107..47649                      TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000507.1:224347..225891                    TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000851.1:51494..53036                      TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000821.1:325692..327234                    TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000961.1:complement(5683571..5685112)      TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP000915.1:576203..577730                    AAACTGAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGTGGCATGCTTAACACATGCAA
AE008922.1:complement(4561295..4562841)      TAAGTGAAGAGTTTGATCCTGGCTCAGAGTGAACGCTGGCGGCAGGCCTAACACATGCAA
CP003390.1:complement(2362892..2364427)      AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCATGCCTAACACATGCAA
CP000544.1:complement(380025..381569)        GAACTGAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCCTAACACATGCAA
```

- Scroll down to inspect the alignment
- Try out the options (e.g. coloring)
- A tree building option is also implemented try it out and compare the tree to that obtained from rdp

Doing a sequence alignment

- You can also download the alignment file which can than be loaded to alignment viewer and editing software available at e.g. www.expasy.org/
- For example bioedit



26 total sequences

Selection: 0
Position: 10: CP000961.1:complemen 34

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
                 10        20        30        40        50        60        70        80        90       100       110       120
mySeq        -------AGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGgtaacaggaagaagcttgcttctttgctgacg--AGTGGCGGACGGGTGAGTAATGT
CP000462.1:8 AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgcagcgggaaagtagcttgctacttttgccggcgAGCGGCGGACGGGTGAGTAATGC
CP000514.1:c AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCTTAACACATGCAAGTCGAGCGgtaacaggggaagcttgcttcccgctgacg----AGCGGCGGACGGGTGAGTAATGC
AE017340.1:c TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgtaacagagagaagcttgcttctctgctgacg--AGCGGCGGACGGGTGAGTAATAC
CP000821.1:3 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgaaacggagatagcttgctatcagacgtcg----AGCGGCGGACGGGTGAGTAATGC
CP000507.1:2 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgcagcgggaagatagcttgctatctttgccggcgAGCGGCGGACGGGTGAGTAATAC
CP000302.1:1 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgcagcgggaagtaacttgttactttgccggcg--AGCGGCGGACGGGTGAGTAATGC
CP000851.1:5 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgtaacacaagggagcttgctcctgaggtgacg--AGCGGCGGACGGGTGAGTAATGC
AE014299.2:4 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgcagcacaagtgagtttactcatgaggtggcg--AGCGGCGGACGGGTGAGTAATGC
CP000961.1:c TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgaaacagaaagtagcttgctactttgctgtcg--AGCGGCGGACGGGTGAGTAATGC
CP000127.1:9 AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCATGCTTAACACATGCAAGTCGAACGgcagcagcacctaagcttgcttaggtggctggcgAGTGGCGGACGGGTGAGTAACGC
CP000544.1:c GAACTGAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCCTAACACATGCAAGTCGAGCGgcagcgcggaggagcttgctcctctggcggcg--AGCGGCGGACGGGTGAGTAACGC
U00096.3:223 AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGgtaacaggaagaagcttgcttctttgctgacg--AGTGGCGGACGGGTGAGTAATGT
FR877557.1:c AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGgtaacaggaagcagcttgctgctttgctgacg--AGTGGCGGACGGGTGAGTAATGT
AE006468.1:c AAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGgtaacaggaagcagcttgctgctttgctgacg--AGTGGCGGACGGGTGAGTAATGT
FN667742.1:4 TAATTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGGACGgtaacaggaaacagcttgctgttttgctgacg--AGTGGCGGACGGGTGAGTAATGT
AE013218.1:2 AAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAAGCCTAACACATGCAAGTCGAGCGgcagcgaaaagaaagcttgctttcttgtcggcg-AGCGGCAAACGGGTGAGTAATAT
AE017354.1:6 GAACTGAAGAGTTTGATCCTGGCTCAGATTGAACGCTGGCGGCATGCTTAACACATGCAAGTCGAACGgcagcattgtctagcttgctagacagatggcg--AGTGGCGAACGGGTGAGTAACGC
AM286690.1:4 TAACTGAAGAGTTTGATCATGGCTCAGATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAGCGgaaacgatcctagcttgctaggaggcgtcg----AGCGGCGGACGGGTGAGTAACAC
```

Most important databases

- www.ncbi.nlm.nih.gov (The US site of the joint international DNA sequence repository (GenBank))
- www.ddbj.nig.ac.jp (Its counterpart in Japan)
- www.ebi.ac.uk/embl/ (Its counterpart in Europe (EMBL)); with links to
- www.expasy.org/sprot/ this is a very good starting point when analyzing proteins