BLAST

Basic Local Alignment Search Tool

Lec 5 / Bioinformatics

By

Dr Delveen R. Ibrahim

Types of DNA sequencing based on the genome target :

1. **Whole genome sequencing** is used to initially determine the genome sequence of an organism.

2. **Whole exome sequencing (WES):** identifies all the protein-coding genes in the genome. Focusing on protein-coding exons (and excluding other regions of the genome) can lower the cost and time of sequencing, as exons make up about 1% of the human genome.

3. Targeted sequencing: This method generally produces a smaller amount of data than WGS, making analysis more manageable. There are several methods of targeted sequencing, each appropriate for specific applications. The most popular methods are **amplicon sequencing.**

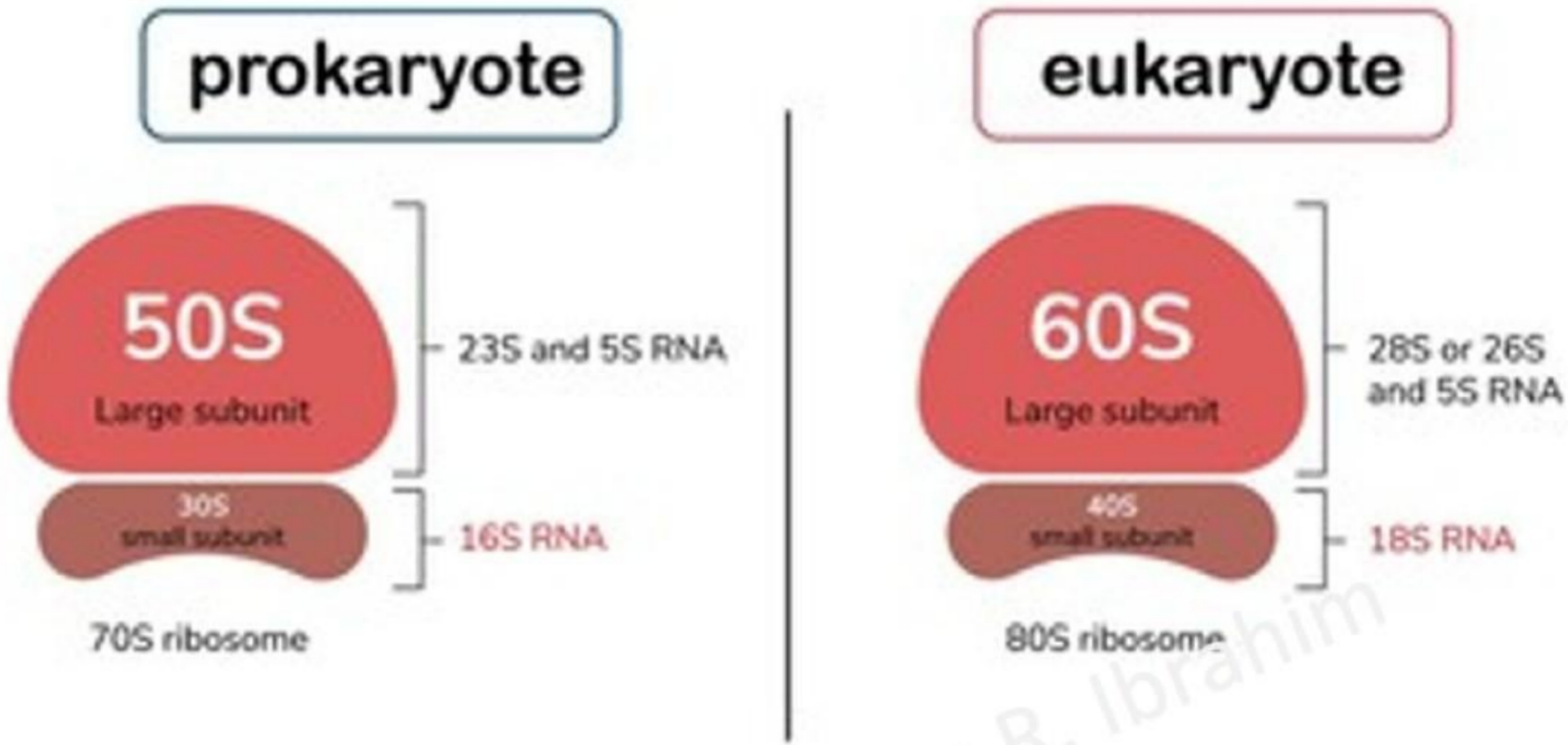
- **Amplicon sequencing** This method uses PCR to amplify DNA to make amplicons. The amplicons are indexed and sequenced, such as partial sequences of 16S rDNA , 18S rDNA and specific genes.

- **Examples of amplicon sequencing:**

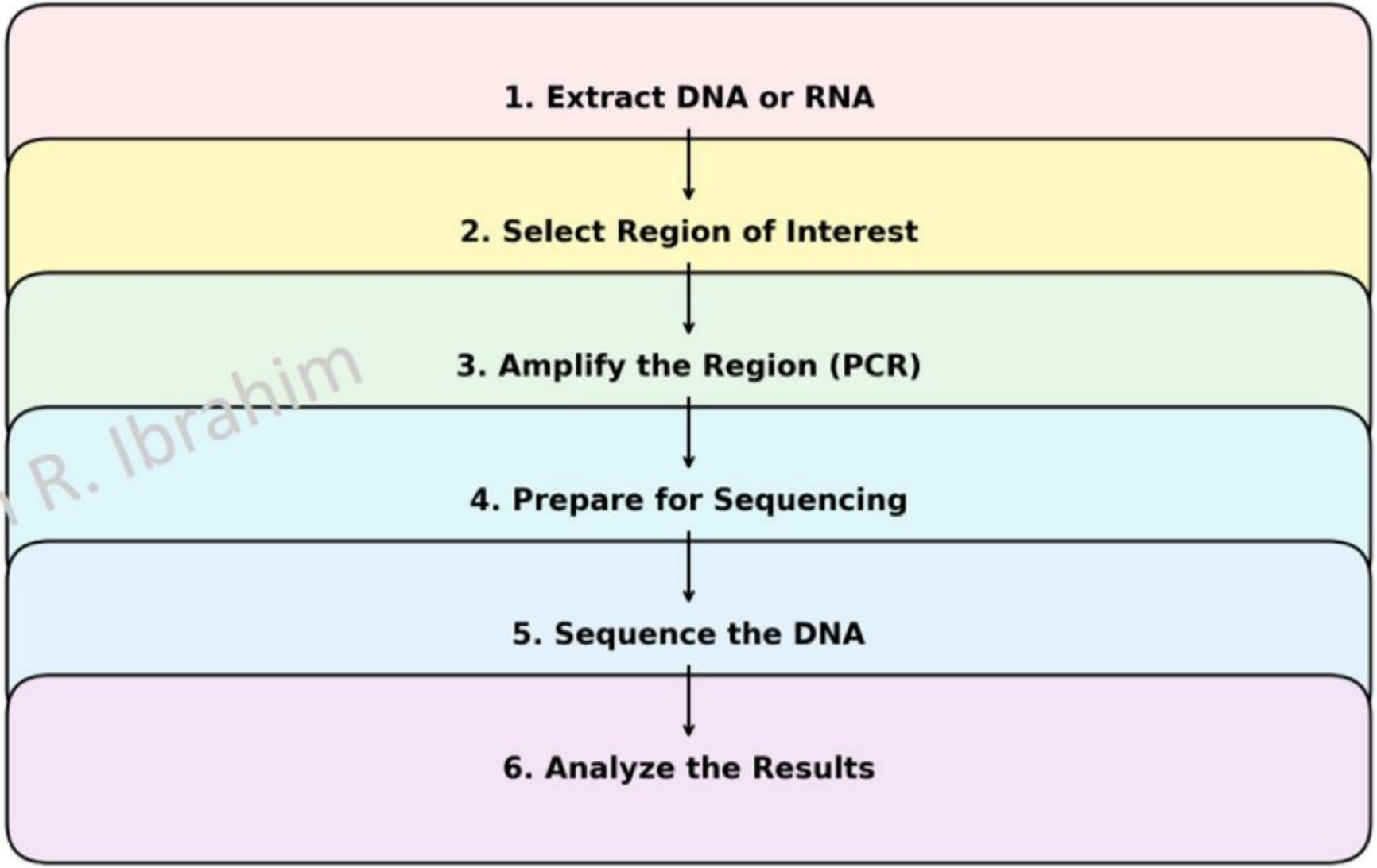
- **The 16S rRNA gene is usually for identification and classification as it is highly conserved between different species of bacteria and archaea**

- **18S rRNA is the structural RNA for the small component of eukaryotic cytoplasmic ribosomes, and thus one of the basic components of all eukaryotic cells**

ribosomal RNA (rRNA) subunit

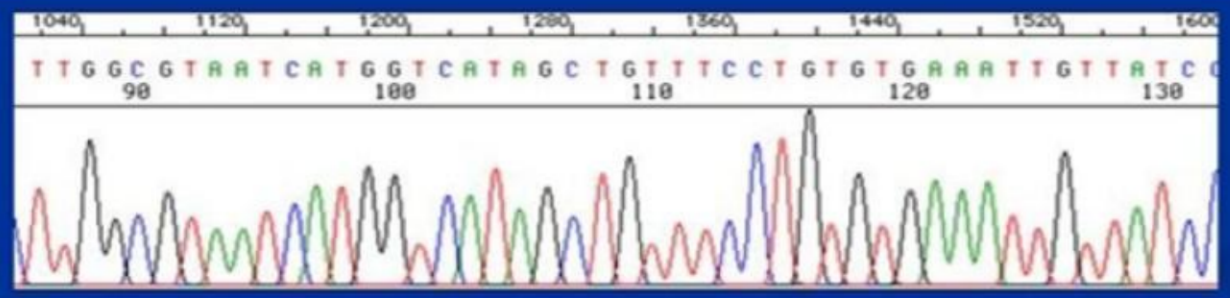


Steps of Amplicon sequencing



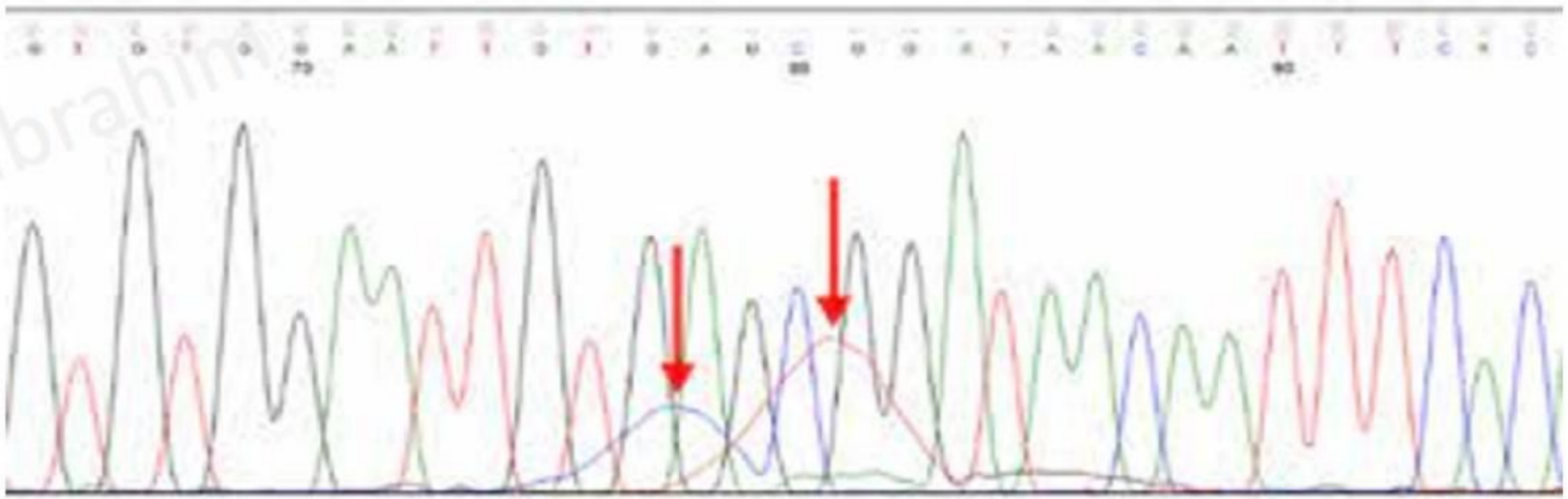
Now what after sequencing??

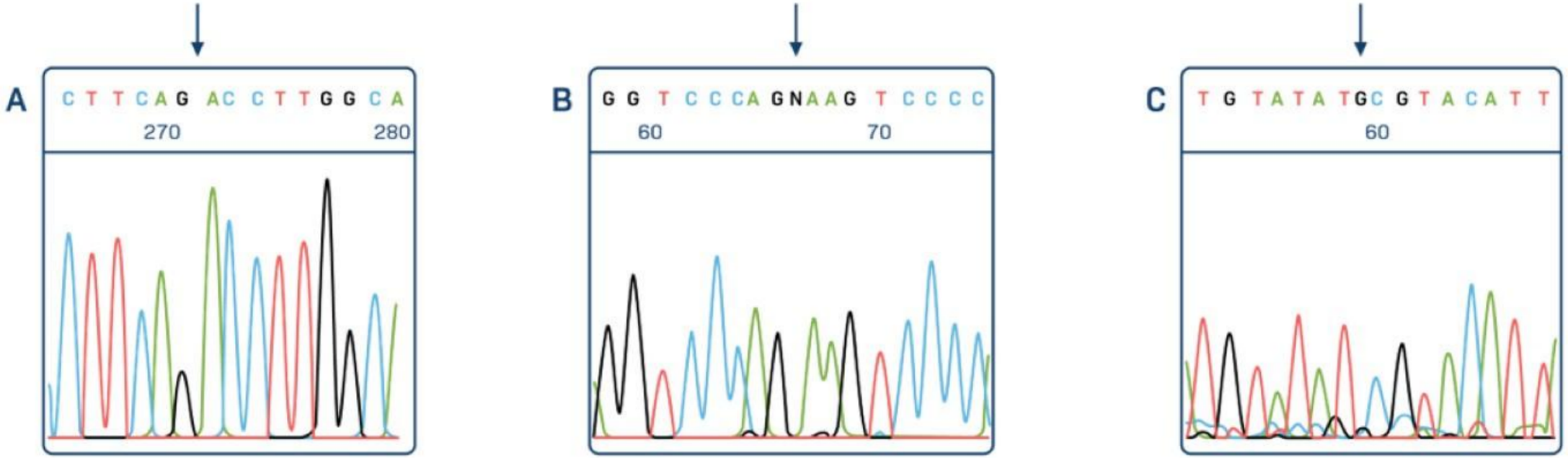
- You will get the Sequence results back
- **The final Data generated is complete chromatogram and the text version of DNA Sequence,** a **Raw data**.



ACAACTGGCGTGAATCATGGCCCTTCGGGGCCATTGTTTC
TCTGTGGAGGAGTGCCATGACGAAAGATGAACTGATTGC
CCGTCTCCGCTCGCTGGGTGAACAACTGAACCGTGATGT
CAGCCTGACGGGGACGAAAGAAGAACTGGCGCTCCGTGT
GGCAGAGCTGAAAGAGGAGCTTGATGACACGGATGAAAC
TGCCGGTCAGGACACCCCTCTCAGCCGGGAAAATGT

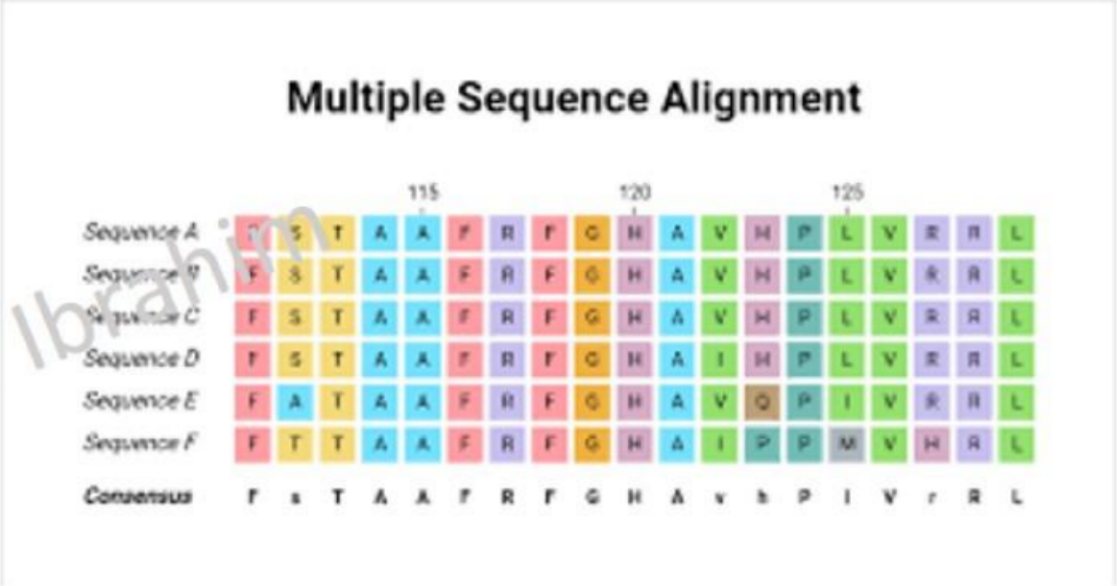
- **Sequence cleaning and trimming:** You must use specific software to analyze your sequence and to check the quality of the sequence, if there is any noise or bad quality sequence part it should be removed by using for example BioEdit software.




A
C T T C A G AC C T T G G C A
270
280
B
G G T C C C A G N A A G T C C C C
60
70
C
T G T A T A T G C G T A C A T T
60

Sequence alignment

- is the process of comparing two or more DNA, RNA, or protein sequences to find similarities and differences. It helps identify matching regions, mutations, and evolutionary relationships

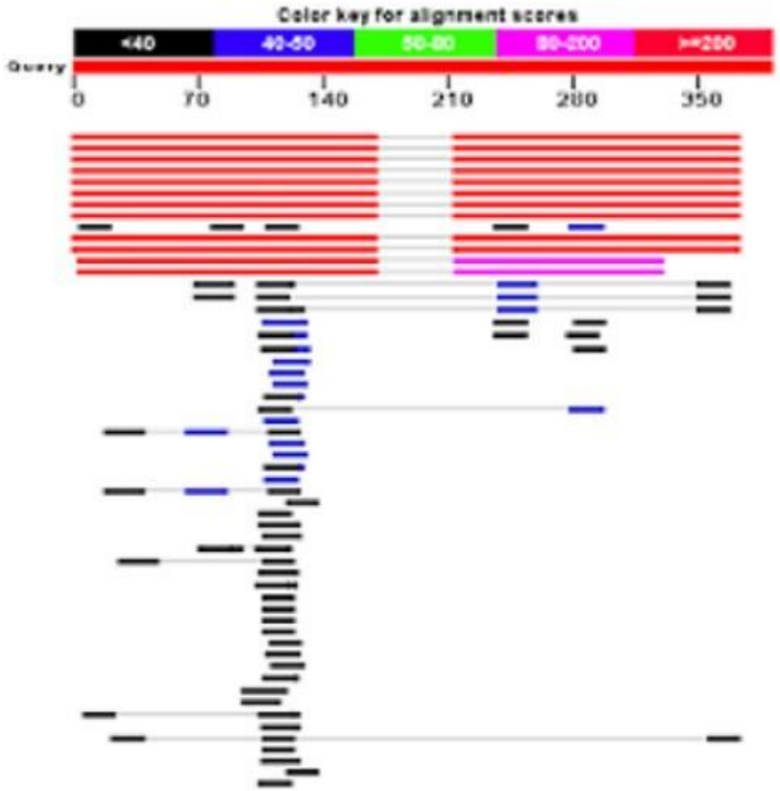


Sequence alignment:

- Many tools exist to analyze **DNA, RNA, and protein** sequences.
- Some tools are free to download or use online, while others require payment.
- **NCBI** offers various free tools for sequence analysis.
- **BLAST** is one of those tools.
- **BLAST: Blast is a bioinformatics tool used to determine the origin of unknown DNA or protein sequences**

Why Use BLAST?

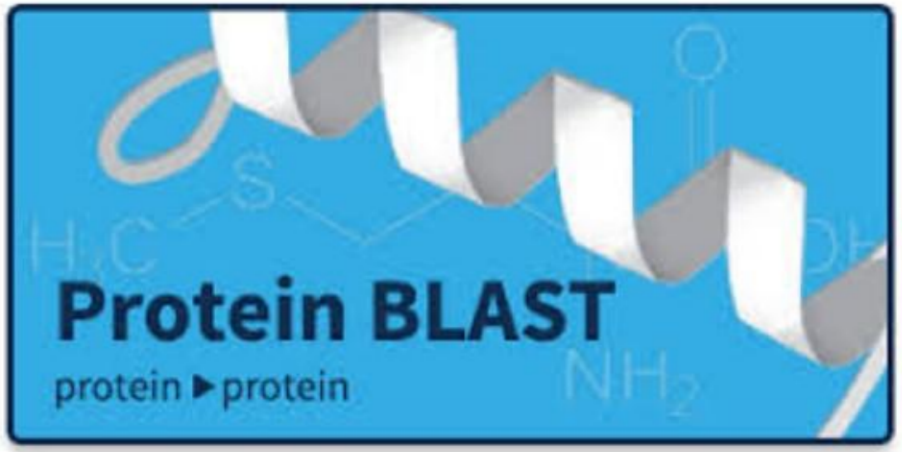
- BLAST helps find similar regions in DNA or protein sequences.
- It compares sequences to a database and checks how closely they match.
- Scientists use BLAST to study gene functions, evolutionary relationships, and identify related genes.



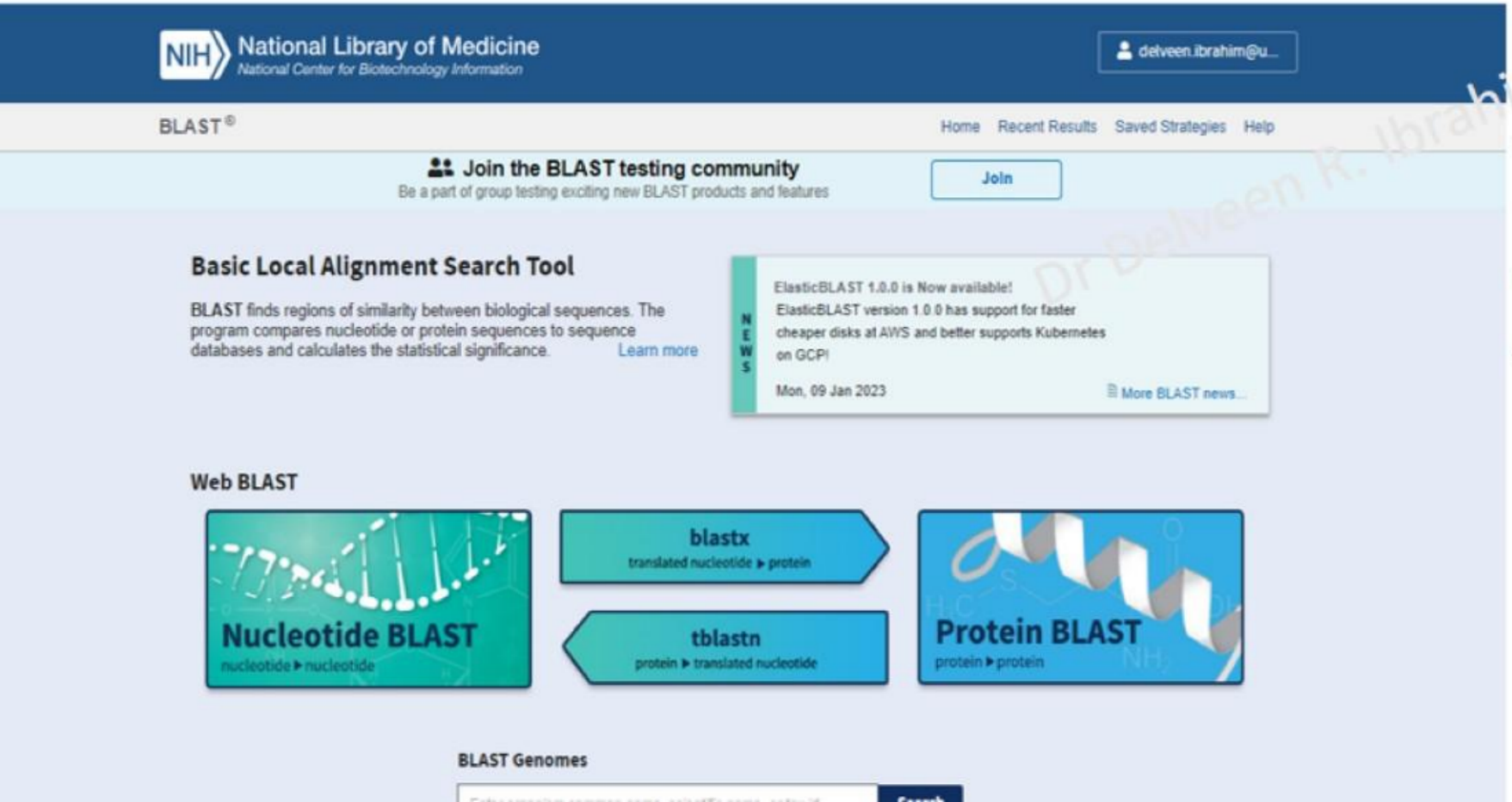
BLAST Type	Query	Database	Use Case
BLASTN	DNA	DNA	Find similar DNA sequences
BLASTP	Protein	Protein	Compare protein sequences
BLASTX	DNA (translated)	Protein	Find proteins from a DNA sequence
TBLASTN	Protein	DNA (translated)	Find DNA sequences that could code for a protein
TBLASTX	DNA (translated)	DNA (translated)	Compare translated DNA sequences

Why is BLAST popular?

- Fast and sensitive
- Reliable
- Flexible
- Finds short similar regions anywhere in a sequence

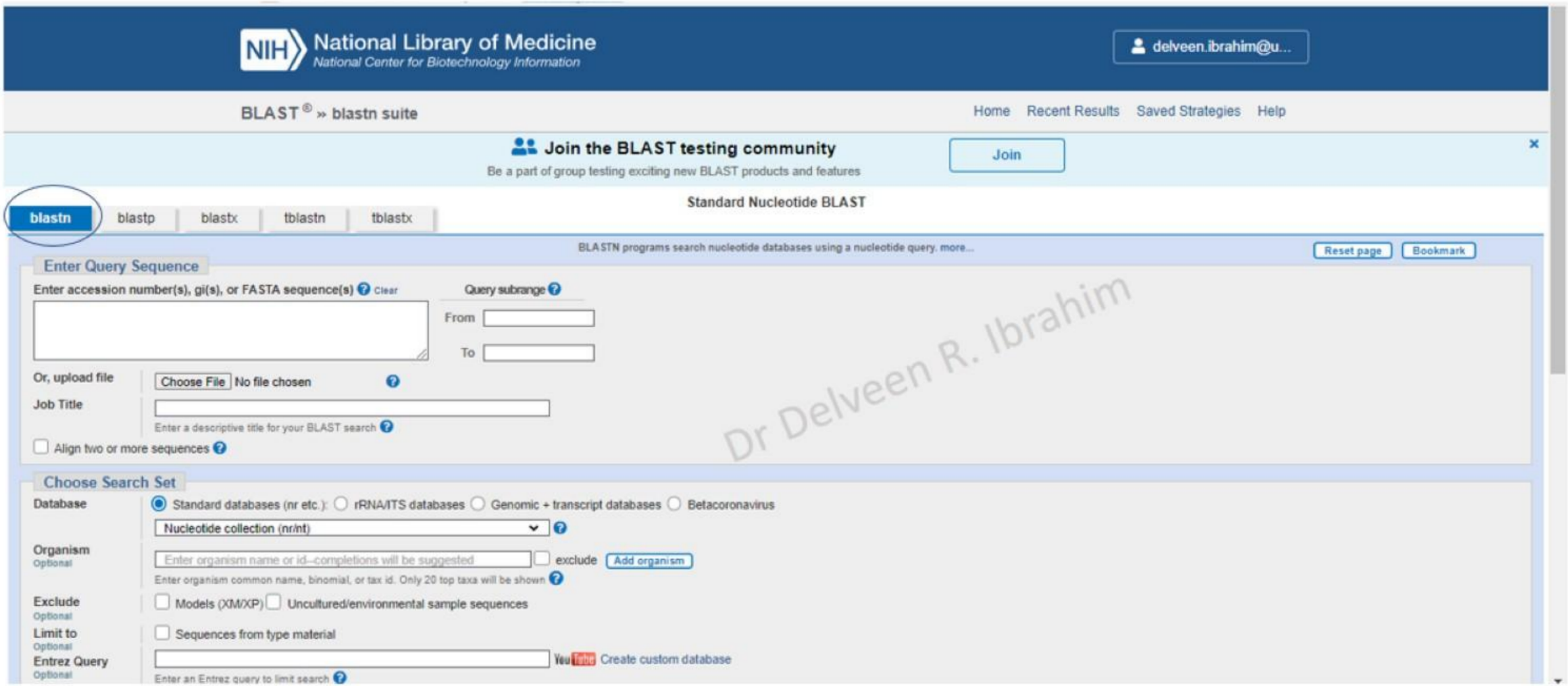


So , to find BLAST



- Simply type in your searching browser BLAST
- Or
- Use this link https://blast.ncbi.nlm.nih.gov/Blast.cgi

We are going to use blastn



National Library of Medicine
National Center for Biotechnology Information

delveen.ibrahim@u...

BLAST® » blastn suite

Home Recent Results Saved Strategies Help

Join the BLAST testing community
Be a part of group testing exciting new BLAST products and features
Join

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

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

Reset page Bookmark

Enter Query Sequence

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

Query subrange
From
To

Or, upload file Choose File No file chosen

Job Title KX668668.Nannizzia fulva strain GZUIFR-EB6401M...
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

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

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

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

Limit to Optional Sequences from type material

Entrez Query Optional Create custom database
Enter an Entrez query to limit search

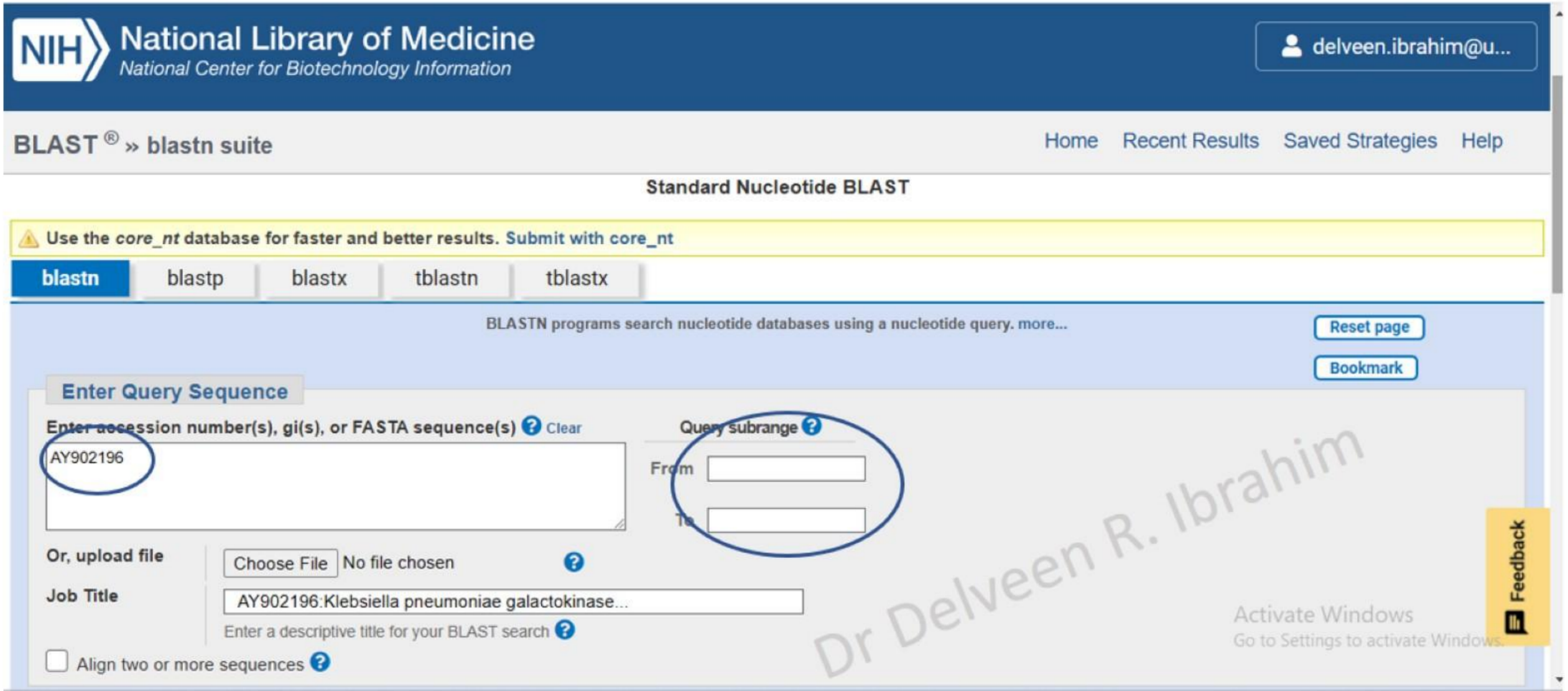
Program Selection

Optimize for Highly similar sequences (megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window

+ Algorithm parameters

You can use FASTA file or accession no. in the blast searching box , please write your note about each part when I explain it



Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

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

Core nucleotide database (core_nt)

Organism
Optional

exclude Add organism

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

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Limit to
Optional

Sequences from type material

Entrez Query
Optional

Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search **database core_nt** using **Megablast (Optimize for highly similar sequences)**

Show results in a new window

Now , you will get this window after you clicked on BLAST at the end of the previous window, you can check the similarity and many other features for your query and the other data that you got in your search table down

National Center for Biotechnology Information

BLAST® » blastn suite » results for RID-X9C4DUUD013

Home Recent Results Saved Strategies Help

Join the BLAST testing community
Be a part of group testing exciting new BLAST products and features

Join

< Edit Search | Save Search | Search Summary | How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	KX668868:Nannizzia fulva strain GZUIFR-EB6401M...
RID	X9C4DUUD013 *Search expires on 01-29 12:20 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	KX668868.1
Description	Nannizzia fulva strain GZUIFR-EB6401M internal transcrit ...
Molecule type	nucleic acid
Query Length	616
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity [ ] to [ ] E value [ ] to [ ] Query Coverage [ ] to [ ]

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Nannizzia fulva strain GZUIFR-EB6401M internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA ge...	Nannizzia fulva	1138	1138	100%	0.0	100.00%	616	KX668868.1
Nannizzia fulva strain Ahv-29 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and inter...	Nannizzia fulva	1112	1112	99%	0.0	99.51%	622	KT192466.1
Arthroderma fulvum strain CS785 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Nannizzia fulva	1112	1112	99%	0.0	99.51%	652	JX413541.1
Nannizzia sp. strain GZUIFR 21.489 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1112	1112	99%	0.0	99.51%	627	MZ019843.1

Feedback

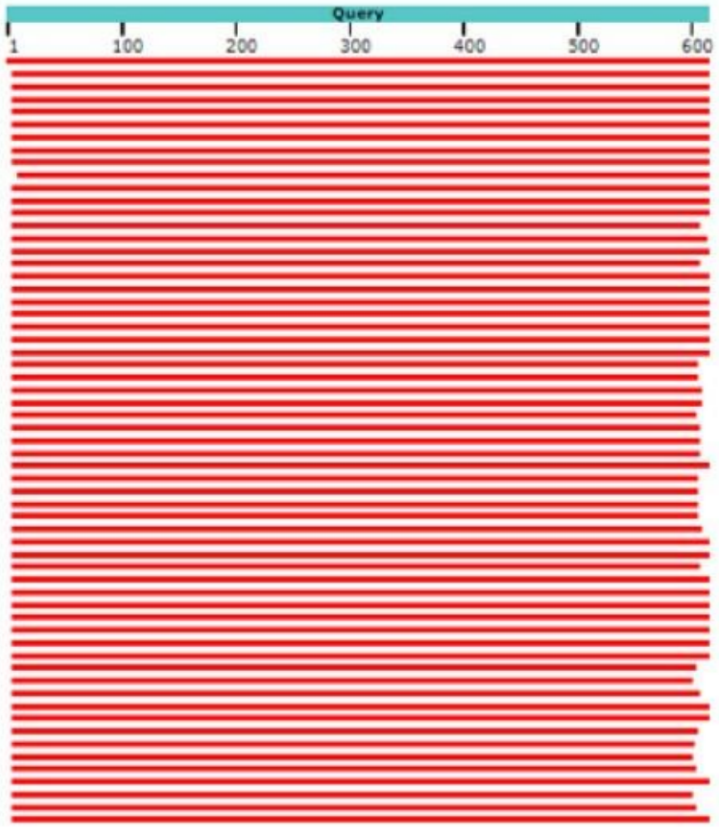
Descriptions | **Graphic Summary** | Alignments | Taxonomy

hover to see the title *click to show alignments*

Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

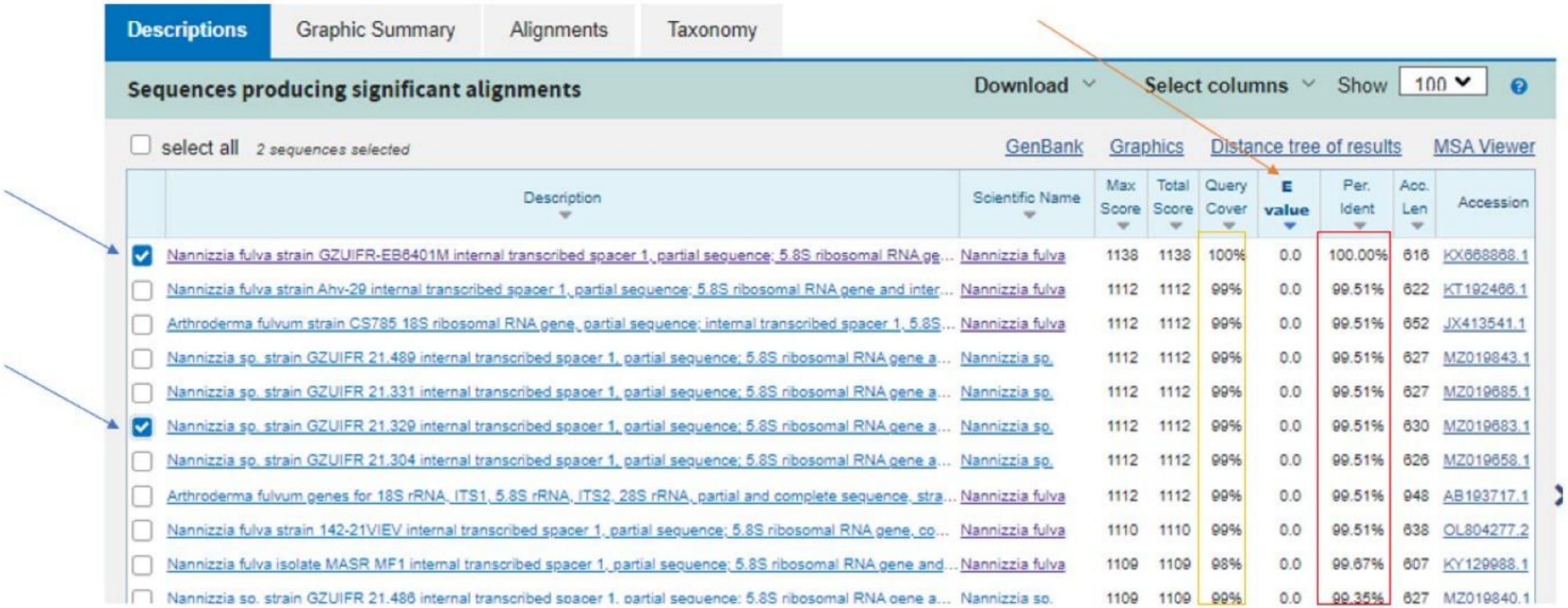
100 sequences selected

Distribution of the top 100 Blast Hits on 100 subject sequences



Feedback

Select only the first and 6th line, options , now choose alignment



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Nannizzia fulva strain GZUIFR-EB6401M internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA ge...	Nannizzia fulva	1138	1138	100%	0.0	100.00%	616	KX668868.1
Nannizzia fulva strain Ahv-29 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and inter...	Nannizzia fulva	1112	1112	99%	0.0	99.51%	622	KT192466.1
Arthroderma fulvum strain CS785 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S...	Nannizzia fulva	1112	1112	99%	0.0	99.51%	652	JX413541.1
Nannizzia sp. strain GZUIFR 21.489 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1112	1112	99%	0.0	99.51%	627	MZ019843.1
Nannizzia sp. strain GZUIFR 21.331 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1112	1112	99%	0.0	99.51%	627	MZ019685.1
Nannizzia sp. strain GZUIFR 21.329 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1112	1112	99%	0.0	99.51%	630	MZ019683.1
Nannizzia sp. strain GZUIFR 21.304 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1112	1112	99%	0.0	99.51%	626	MZ019658.1
Arthroderma fulvum genes for 18S rRNA, ITS1, 5.8S rRNA, ITS2, 28S rRNA, partial and complete sequence, stra...	Nannizzia fulva	1112	1112	99%	0.0	99.51%	948	AB193717.1
Nannizzia fulva strain 142-21VIEV internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, co...	Nannizzia fulva	1110	1110	99%	0.0	99.51%	638	OL804277.2
Nannizzia fulva isolate MASR MF1 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and...	Nannizzia fulva	1109	1109	98%	0.0	99.67%	607	KY129988.1
Nannizzia sp. strain GZUIFR 21.486 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene a...	Nannizzia sp.	1109	1109	99%	0.0	99.35%	627	MZ019840.1

Download GenBank Graphics ▼Next ▲Previous ◄Descriptions

Nannizzia fulva strain GZUIFR-EB6401M internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

Sequence ID: KX668868.1 Length: 616 Number of Matches: 1

Range 1: 1 to 616 GenBank Graphics ▼Next Match ▲Previous Match

Score	Expect	Identities	Gaps	Strand
1138 bits(616)	0.0	616/616(100%)	0/616(0%)	Plus/Plus

```
Query  1    ACTCGGCATACGGCCCGCCCCCGGATGCGTCCGAGGGCGGTGCCGCCGGCCACACGCCCA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ACTCGGCATACGGCCCGCCCCCGGATGCGTCCGAGGGCGGTGCCGCCGGCCACACGCCCA  60

Query  61   TTCTTGTCTATCTACCTGGTTGCCTCGGCGGGCCGCGCACTCGTGCCGCCCGGGGATGTT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   TTCTTGTCTATCTACCTGGTTGCCTCGGCGGGCCGCGCACTCGTGCCGCCCGGGGATGTT  120

Query  121  TATCTCCCTGGACCGCGCCCGCCGGAGGAGCGATTAAAACCCATGAATACTGTTCAGTCT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TATCTCCCTGGACCGCGCCCGCCGGAGGAGCGATTAAAACCCATGAATACTGTTCAGTCT  180

Query  181  GAGCGTAAGCAAGTGAAATCAGTTAAAACTTTCAACAACGGATCTCTTGGTTCCGGCATC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  GAGCGTAAGCAAGTGAAATCAGTTAAAACTTTCAACAACGGATCTCTTGGTTCCGGCATC  240

Query  241  GATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCCGTGAATCATCG  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GATGAAGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCCGTGAATCATCG  300

Query  301  AATCTTTGAACGCACATTGCGCCCTCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTCA  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  AATCTTTGAACGCACATTGCGCCCTCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTCA  360

Query  361  TTTCAACCCCTCAAGCCCGGCTTGTGTGATGGACGACCGTCCCGCCCTCCCCGATCCCGG  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361  TTTCAACCCCTCAAGCCCGGCTTGTGTGATGGACGACCGTCCCGCCCTCCCCGATCCCGG  420
```

Download GenBank Graphics ▼Next ▲Previous ◄Descriptions

Nannizzia sp. strain GZUIFR 21.329 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and large subunit ribosomal RNA gene, partial sequence

Sequence ID: MZ019683.1 Length: 630 Number of Matches: 1

Range 1: 13 to 624 GenBank Graphics ▼Next Match ▲Previous Match

Score	Expect	Identities	Gaps	Strand
1112 bits(602)	0.0	609/612(99%)	1/612(0%)	Plus/Plus

```
Query  6    GCATACGGCCCGCCCCCGGATGCGTCCGAGGGCGGTGCCGCCGGCCACACGCCCATTCTT  65
            || ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  13   GCAGACGGCCCGCCCCCGGATGCGTCCGAGGGCGGTGCCGCCGGCCACACGCCCATTCTT  72

Query  66   GTCTATCTACCTGGTTGCCTCGGCGGGCCGCGCACTCGTGCCGCCCGGGGATGTTTATCT  125
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  73   GTCTATCTACCTGGTTGCCTCGGCGGGCCGCGCACTCGTGCCGCCCGGGGATGTTTATCT  132

Query  126  CCCTGGACCGCGCCCGCCGGAGGAGCGATTAAAACCCATGAATACTGTTCAGTCTGAGCG  185
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  133  CCCTGGACCGCGCCCGCCGGAGGAGCGATTAAAACCCATGAATACTGTTCAGTCTGAGCG  192

Query  186  TAAGCAAGTGAAATCAGTTAAAACTTTCAACAACGGATCTCTTGGTTCCGGCATCGATGA  245
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  193  TAAGCAAGTGAAATCAGTTAAAACTTTCAACAACGGATCTCTTGGTTCCGGCATCGATGA  252

Query  246  AGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCCGTGAATCATCGAATCT  305
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  253  AGAACGCAGCGAAATGCGATAAGTAATGTGAATTGCAGAATTCCGTGAATCATCGAATCT  312

Query  306  TTGAACGCACATTGCGCCCTCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTCATTTCA  365
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  313  TTGAACGCACATTGCGCCCTCTGGTATTCCGGGGGGCATGCCTGTTCGAGCGTCATTTCA  372

Query  366  ACCCCTCAAGCCCGGCTTGTGTGATGGACGACCGTCCCGCCCTCCCCGATCCCGGGGAGG  425
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  373  ACCCCTCAAGCCCGGCTTGTGTGATGGACGACCGTCCCGCCCTCCCCGATCCCGGGGAGG  432

Query  426  GGGACGCGCCCGAAAAGCAGTGGCCAGGCCGCGATTCCGGCTCCTGGGCGAATGGGCAAC  485
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  433  GGGACGCGCCCGAAAAGCAGTGGCCAGGCCGCGATTCCGGCTCCTGGGCGAATGGGCAAC  492

Query  486  AAACCTACGCCTCTAGGACCGGCCGGCTCTCTGGCCTAGTTTCCGTTAGGGATGAACTTC  545
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  493  AAACCTACGCCTCTAGGACCGGCCGGCTCTCTGGCCTAGTTTCCGTTAGGGATGAACTTC  552

Query  546  CCTACGATTCAGGTTGACCTCGGATCAGGTAGGGATACCCGCTGAACTTAAGCATATCAA  605
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  553  CCTACGATTCAGGTTGACCTCGGATCAGGTAGGGATACCCGCTGAACTTAAGCATATCAA  612

Query  606  -AAGCGGAGGA  616
            |||| |||||||
Sbjct  613  TAAGGCGGAGGA  624
```

Feature	**E-value**	**Identity Value**
Definition	The expected number of matches found by chance in a database.	The percentage of identical nucleotides or amino acids in the aligned region.
Lower value means...	A more statistically significant match.	More differences between the sequences.
Higher value means...	The match is likely random.	A higher degree of sequence similarity.
Depends on	Database size, sequence length, and similarity.	Only the alignment between two sequences.
Typical use	Determines if a match is statistically relevant.	Assesses how closely two sequences are related.

• A **lower E-value** (closer to 0) means a more significant match.
•A **higher E-value** suggests the match might have occurred by random chance.

For example:
•**E = 0.000001** → The match is likely real.
•**E = 1** → The match is probably due to chance.

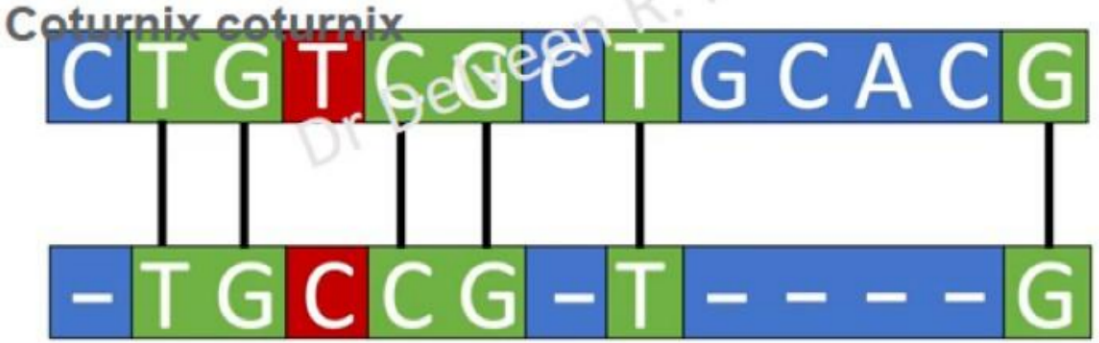
•So:

•**E-value** helps decide if a match is **meaningful or random**.
•**Identity value** shows **how similar** two sequences are.
•A **low E-value + high identity** suggests a strong match.
•**Query Cover[age]:** the percent of the query length that is included in the alignment (comparison).

https://youtu.be/UDdIzNibERQ

https://youtu.be/WRKQGwh Mw0

How to use BLAST for Your Genomic Research



SCIENCE BUDDIES

Task1:

- Use this accession (**AY902196**) number and answer the following questions by using blastn:
- To which gene does this accession belong?
- Complete the following table, choosing the top hit for each organism.
- Specify the organism before you blastn

Organism	Genbank accession	Hit description	% identity	E-value
Enterobacter spp.				
Klebsiella spp.				
Shigella spp.				
Pseudomonas spp.				

- Looking at both the **% identity and E-value** scores, which of the above hits (excluding *K. pneumoniae*), do you think is the best match to your **query sequence**?