

# BLAST

Basic Local Alignment Search Tool

So useful – it is now a verb in the literature



# Goals



What is BLAST and why is it important?

Principles of the algorithm

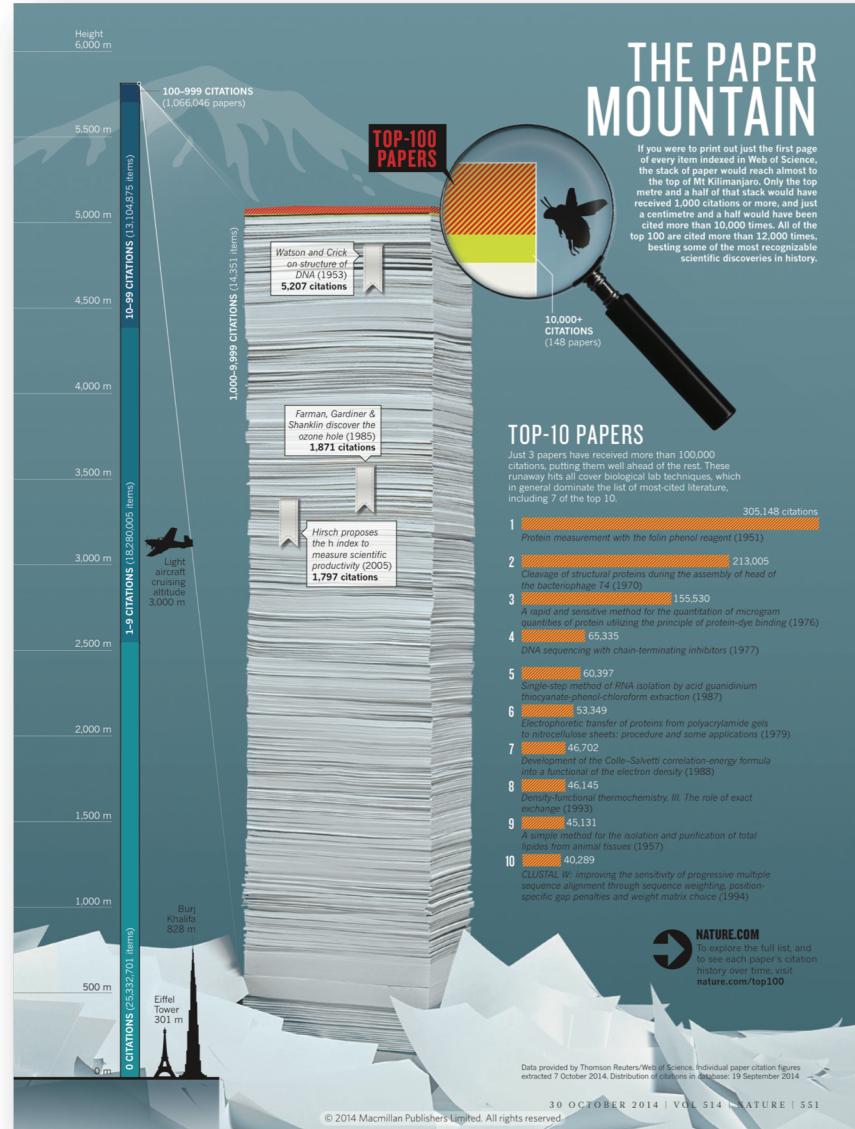
Online Examples

Command Line Implementation

# A Lot of BLASTing

- Where is BLAST on this list?
  - Altschul et al. 1990
    - #12 – 38,380 citations
      - 62,483 (Web of Science 8/1/2021) - #10
  - Altschul et al. 1997
    - #14 – 36,410 citations
      - 53,647 (Web of Science 8/1/2021) - #14
  - Combined: 4<sup>th</sup>!

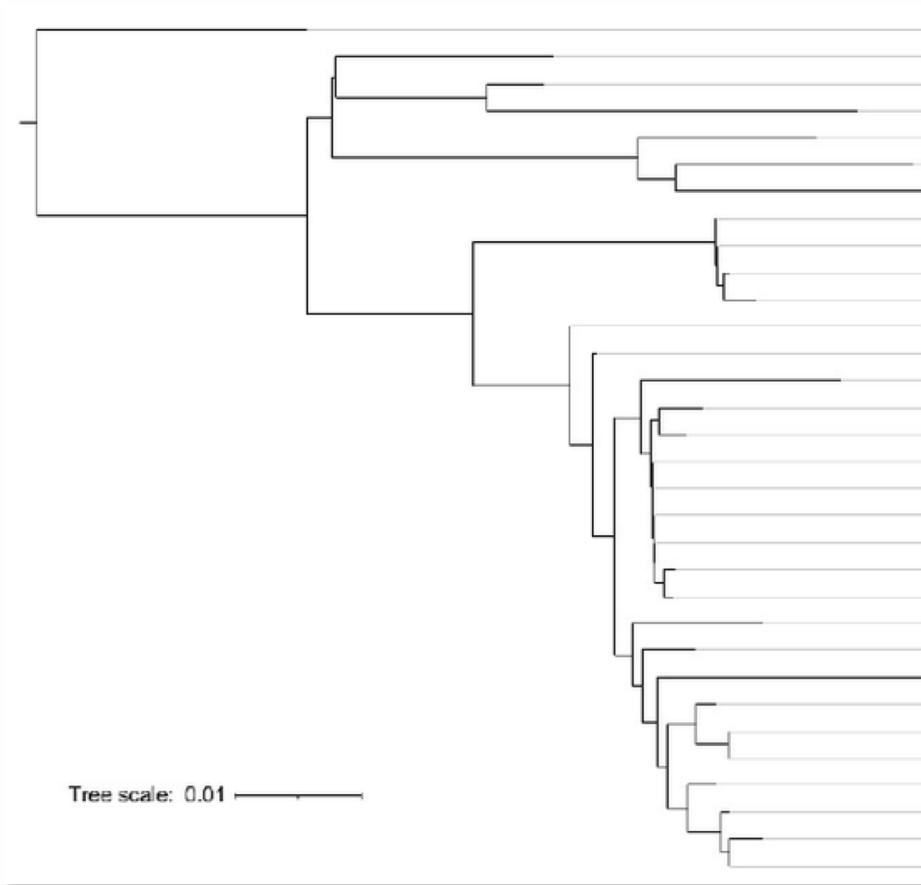
Van Noorden et al. 2014, *Nature*



# BLAST

- Search a query against a database

- Identify species
- Locate domains
- Assess function
- Establish phylogeny
- Mapping



# BLAST

- Sequence searching algorithm
- Finds the best local alignments
- Calculates statistical significance
- Similarity suggests homology
- Less sensitive than Smith-Waterman, but FASTER!

- Global vs Local Alignment

- Global alignment: entire sequences



- Local alignment: segments of sequences



- Local alignment often the most relevant
    - Depends on biological assumptions

# BLAST Flavors

Name	Query	Database
<b>blastn</b>	nucleotide	nucleotide
<b>blastp</b>	protein	protein
<b>blastx</b>	nucleotide	protein
<b>tblastx</b>	nucleotide	nucleotide
<b>tblastn</b>	protein	nucleotide
<b>PSI-blast</b>	protein	protein



# BLAST Databases: Protein

Name	Host	Description
nr	NCBI	Non-redundant, general
Refseq_protein	NCBI	Annotated and curated protein collection
SwissProt	SIB	Manually curated and reviewed proteins from UniProt
Trembl	EBI	Automatically annotated, non-reviewed proteins
PDB	Rutgers/UCSD/UCSC	Proteins with 3D structural information



# BLAST Databases: Nucleotide

Name	Host	Description
nt	NCBI	Non-redundant, general
Refseq_RNA	NCBI	Annotated and curated RNA sequence collection
Refseq_Genomics	NCBI	Sequenced and curated genomes
EST	NCBI	Expressed sequence tags
UNIVEC	NCBI	Vector contaminant database
WGS	NCBI	Draft, whole genome shotgun sequence assemblies
SRA	NCBI	Raw NGS datasets
Many more databases, e.g. barcoding, viral, tRNA, etc, custom-built databases		



# How it Works: Making Words

## Nucleotide

11-letter words (seeds)

ACTACGTGCTATGC

ACTACGTGCTA

CTACGTGCTAT

TACGTGCTATG

ACGTGCTATGC

## Protein

3-letter words (seeds)

PQGDEF

PQG

QGD

GDE

DEF



# How it Works

## Nucleotide

CATG**CTTCGCGGGAT**GCCA

11-mer word size

**CTTCGCGGGAT**

**CTTCGCGGGAT**  
CTTG**CTTCGCGGGAT**GGTA

← **CTTCGCGGGAT** →  
CTTG**CTTCGCGGGAT**GGTA

## Protein

HWR**AHYTCSYAI**

3-mer word size

**AHY**

**AHY**  
RSH**AHYCCSYAI**

← **AHY** →  
RSH**AHYCCSYAI**



# BLAST Scoring and E-values

- Nucleotide sequences search for 11-letter matches
  - $4^{11} = 4,194,304$  combinations
  - Match = +5, mismatch = -4
  - Only scores above a threshold ( $T$ ) are kept

ACTACGTGCTA  
ACTACGTGCTA  
 $5+5+5+5+5+5+5+5+5+5 = 55$

ACTACG**T**GCTA  
ACAAGA**T**GGTA  
 $5+5-4+5-4-4+5+5-4+5+5 = 19$

# BLAST Scoring and E-values

- Proteins use a BLOSUM62 scoring matrix
  - $20 \times 20 \times 20 = 8,000$  possible 3-letter words
  - All possible amino acid pairs are given a score
  - All combinations above a threshold (T) are kept
    - Minimizes search space

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																		C		
S	-1	4																	S		
T	-1	1	5																T		
P	-3	-1	-1	7															P		
A	0	1	0	-1	4														A		
G	-3	0	-2	-2	0	6													G		
N	-3	1	0	-2	-2	6													N		
D	-3	0	-1	-1	-2	-1	1	1	6										D		
E	-4	0	-1	-1	-1	-2	0	2	5										E		
Q	-3	0	-1	-1	-1	-2	0	0	2	5									Q		
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8								H		
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5							R		
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5						K		
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5					M		
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4				I		
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4			L		
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		V		
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		F		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	3	7		Y	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

P Q G

P E G

$$7+2+6 = 15$$

P Q G

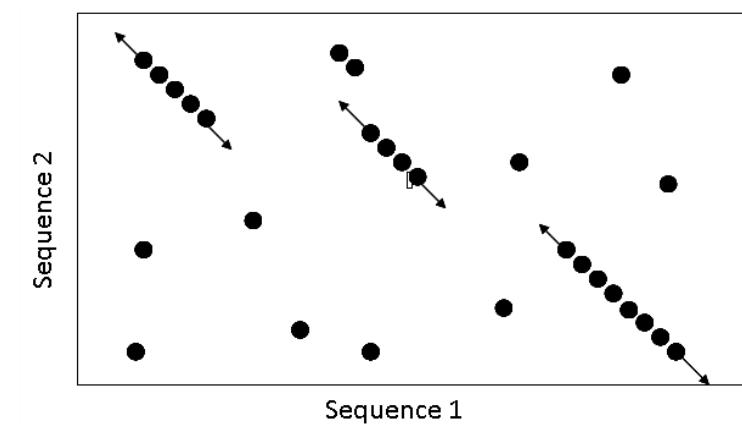
E Q R

$$-1+5+-2 = 2$$

# Extending Matches

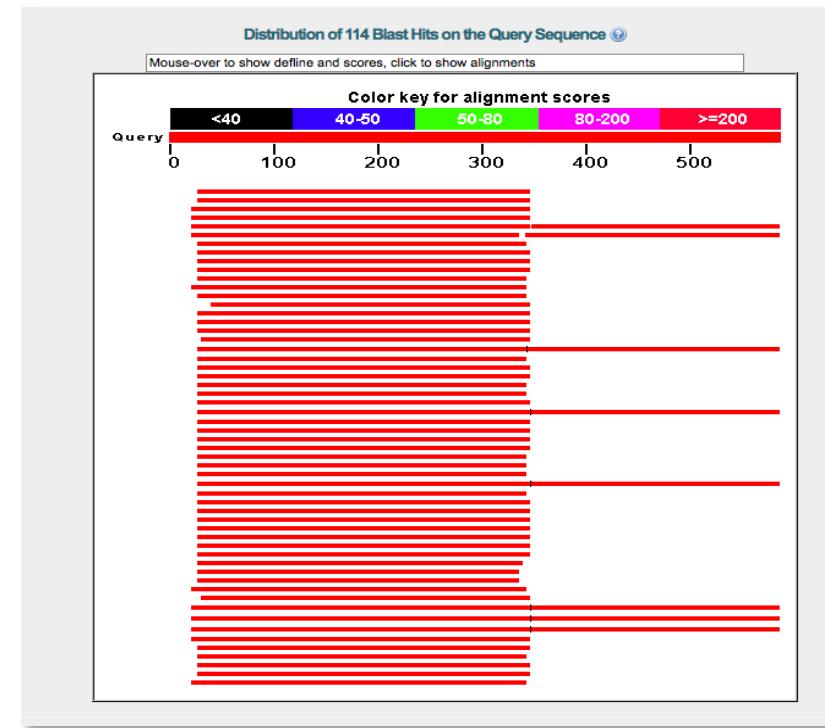
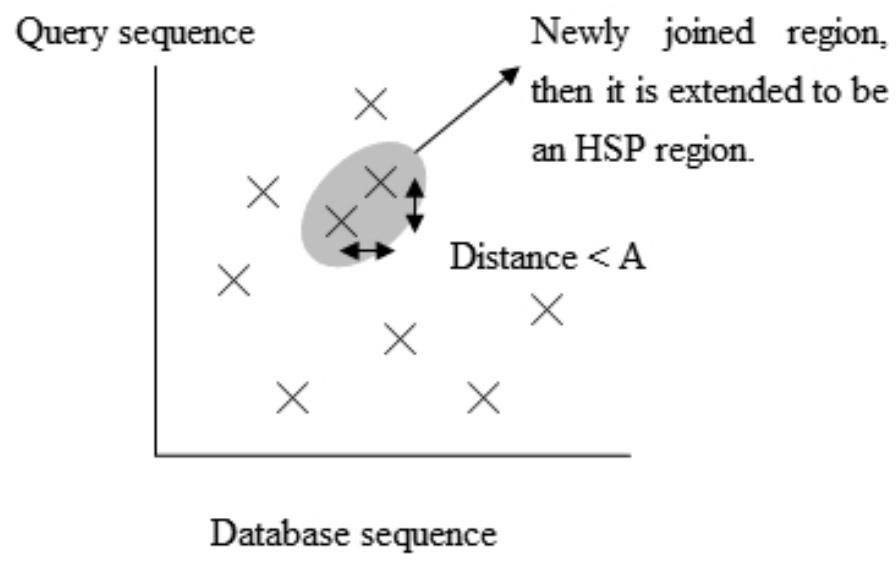
- Match = HSP (High-scoring Sequence Pair)
  - Match is found and extended as long as score stays above a threshold value
    - After finished extending, the HSP is kept if above the cutoff score (S)

Query sequence: R P P Q G L F  
Database sequence: D P P E G V V  
Score: -2 7 7 2 6 1 -1  
Optimal accumulated score =  $7+7+2+6+1 = 23$



# Assembling HSPs

- HSPs, after extension, are assembled into a longer alignment



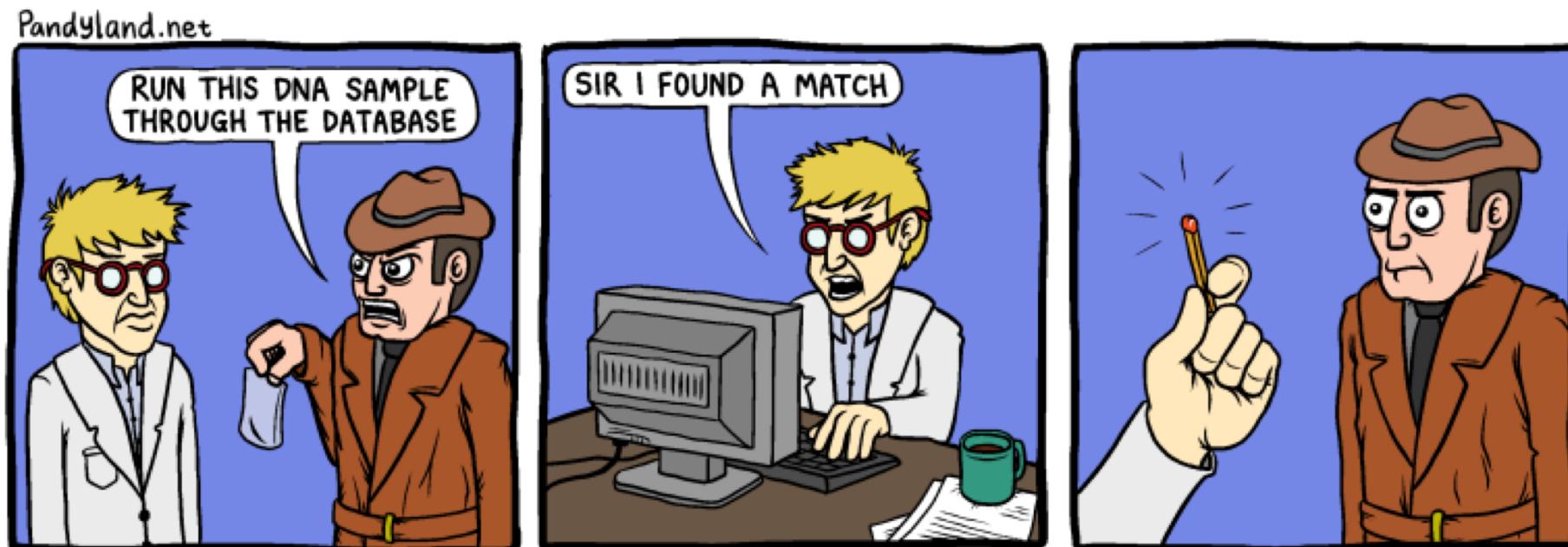
# Output

- Max/Total Score: quality of the alignment
  - Higher the score the better the match
- Query Coverage: what proportion of the query the particular HSP covers
- E-value: probability that a match  $\geq$  Max Score occurs by random chance (based on database size)
- Max Identity: For that HSP, the % of bases that match

Accession	Total Score	Query Coverage	E-value	Max Ident
X56286.1	579	54%	7e-162	99%
AF091629.1	573	54%	3e-160	99%
L48348.1	481	55%	2e-132	93%

# Interpretation

- The matches you get are only acceptable matches, not necessarily the optimal match
- Your search is only as good as your database
  - If the optimal match is not in the database, you will not find it.
  - If you have sequences not in the database, SUBMIT THEM!



# Take Away Points

- BLAST is a powerful tool for database searching
- Very fast, but at the expense of sensitivity
- Flexible (types, databases)
- Interpret results carefully
- Help make it grow!



# DATABASES

Where are the genomic data?

# SRA

- Sequence Read Archive
  - <https://www.ncbi.nlm.nih.gov/sra/>

The screenshot shows the NCBI SRA homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, and a user account (rfitak@email.arizona.edu). The main search bar has 'SRA' selected. Below the search bar is a 'COVID-19 Information' banner with links to CDC, NIH, and HHS resources. The main content area features a large image of a glowing blue molecular structure and a text box stating 'SRA - Now available on the cloud'. Below this, there are three columns: 'Getting Started' (with links to How to Submit, How to search and download, How to use SRA in the cloud, and Submit to SRA), 'Tools and Software' (with links to Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser, and SRA Run Selector), and 'Related Resources' (with links to Submission Portal, Trace Archive, dbGaP Home, BioProject, and BioSample).



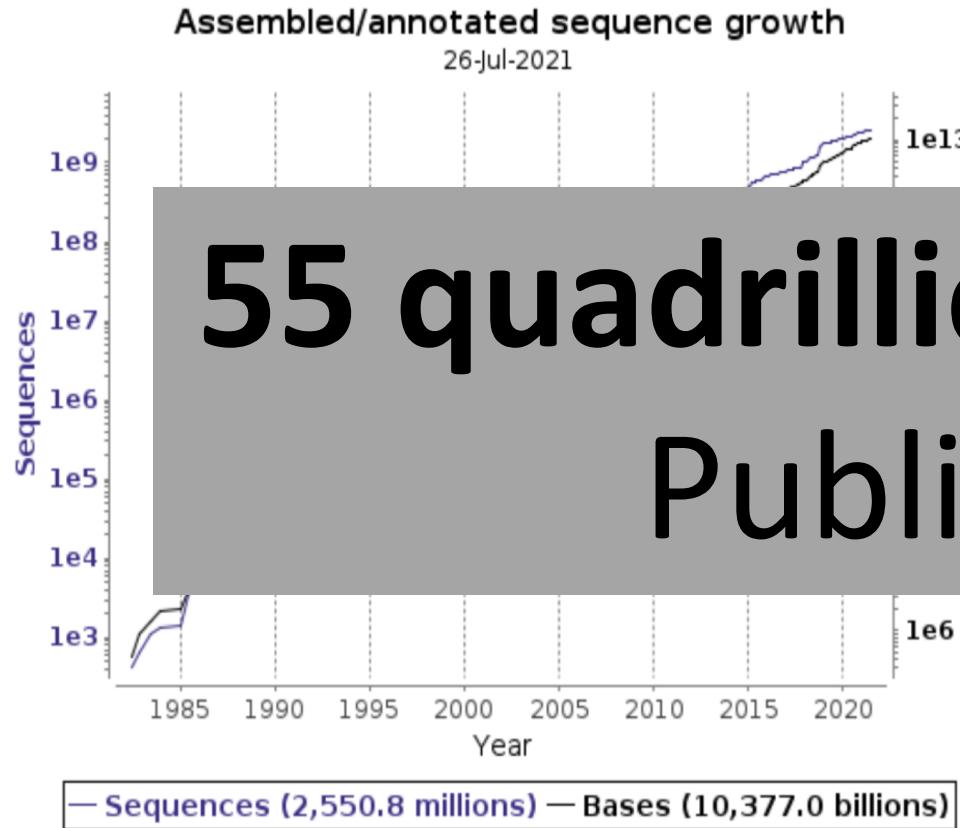
# So how big is it?

Guesses?



# DNA Sequence Databases (GenBank, SRA, ENA)

Assembled/annotated sequence growth



Reads growth



<https://www.ebi.ac.uk/ena/browser/about/statistics>

# SRA Demo...

# How to search all that SRA data?

NCBI Resources ▾ How To ▾ rfitak@email.arizona.edu My NCBI Sign Out

SRA  Advanced

Getting Started Submission Quick Start Search and Download SRA in the Cloud

### Search in BigQuery

#### Overview

SRA has deposited its metadata into BigQuery to provide the bioinformatics community with programmatic access to this data. You can now search across the entire SRA by sequencing methodologies and sample attributes. NCBI is piloting this in BigQuery to help users leverage the benefits of elastic scaling and parallel execution of queries. BigQuery has a large collection of client libraries that can be used within your workflow. You can also interact with it on a web browser.

The Big Query resource contains a tables for SRA metadata and computed metadata on SRA runs.

#### Tables

The list of tables can be found here: [SRA cloud-based tables](#).

Please read about the [SRA Taxonomy Analysis Tool](#) to learn how the analysis is carried out.

#### The Basics of SQL

The basic SQL query has three parts or statements:

- SELECT: Identifies which columns from the selected table(s) to show. The \* indicates "all columns"
- FROM: Identifies table(s) to query
- WHERE: Joins tables using the identical columns in both tables and sets filters on the query

BigQuery

+ COMPOSE NEW QUERY

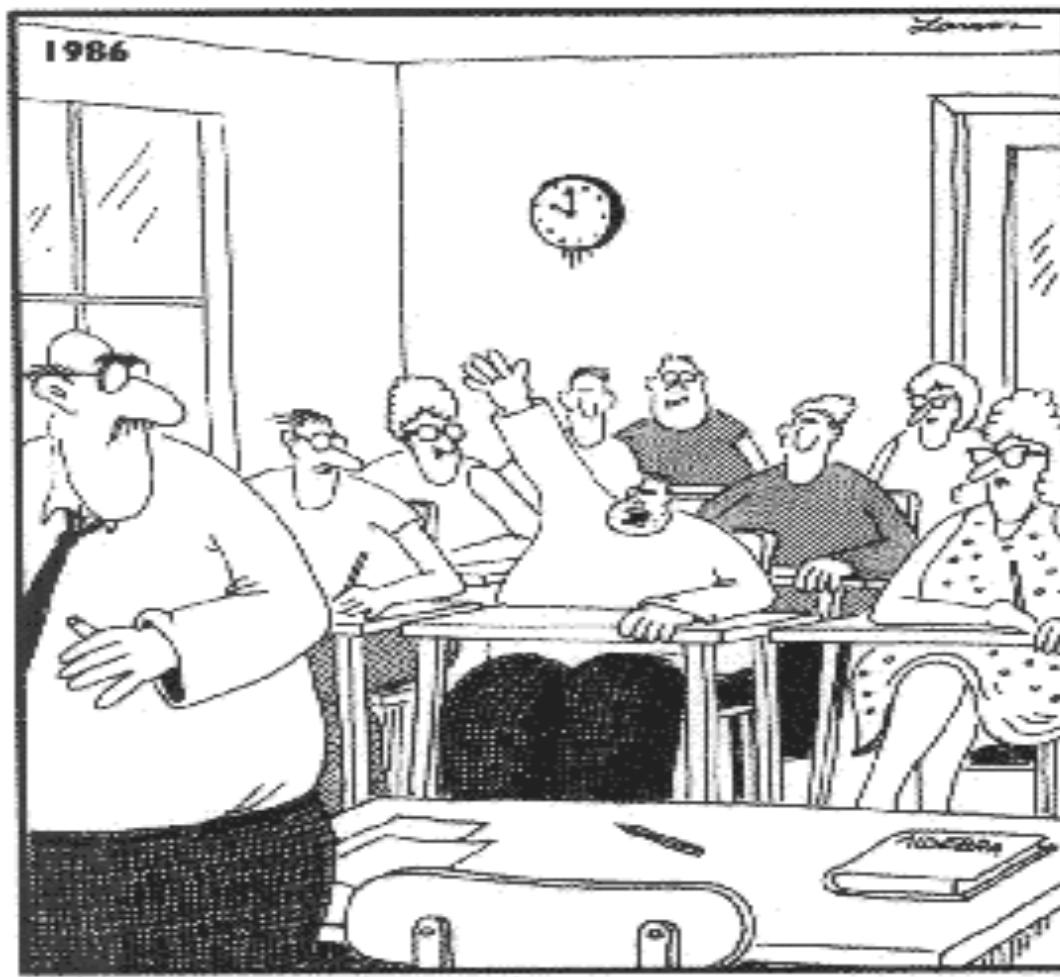
Unsaved query Edited

```
1 SELECT *
2 FROM `nih-sra-datastore.sra.metadata`
3 WHERE organism = 'Homo sapiens'
```



# BigQuery





"Mr. Osborne, may I be excused? My brain is full."

# Practice Examples

- Example 1: SRA Blast (<https://www.ncbi.nlm.nih.gov/sra>)
  - Click “SRA-BLAST” link
  - Query: M55627.1
    - o *Coccidioides immitis* (Valley fever fungus) ssuRNA
  - Project: SRX633288
    - o Puma 454 transcriptome reads
- Example 2: Blast an assembly (<https://blast.ncbi.nlm.nih.gov/>)
  - Select “Nucleotide BLAST”
  - Query:
    - o TruSeq Universal Adapter
      - o AATGATA CGGCG ACCACCGAGATCTACACTCTTCCCTACACGACGCTCTCCGATCT
  - Database: nt
  - Organism: *Cyprinus carpio* (*taxid:7962*)