BLAST and Beyond!

Opener

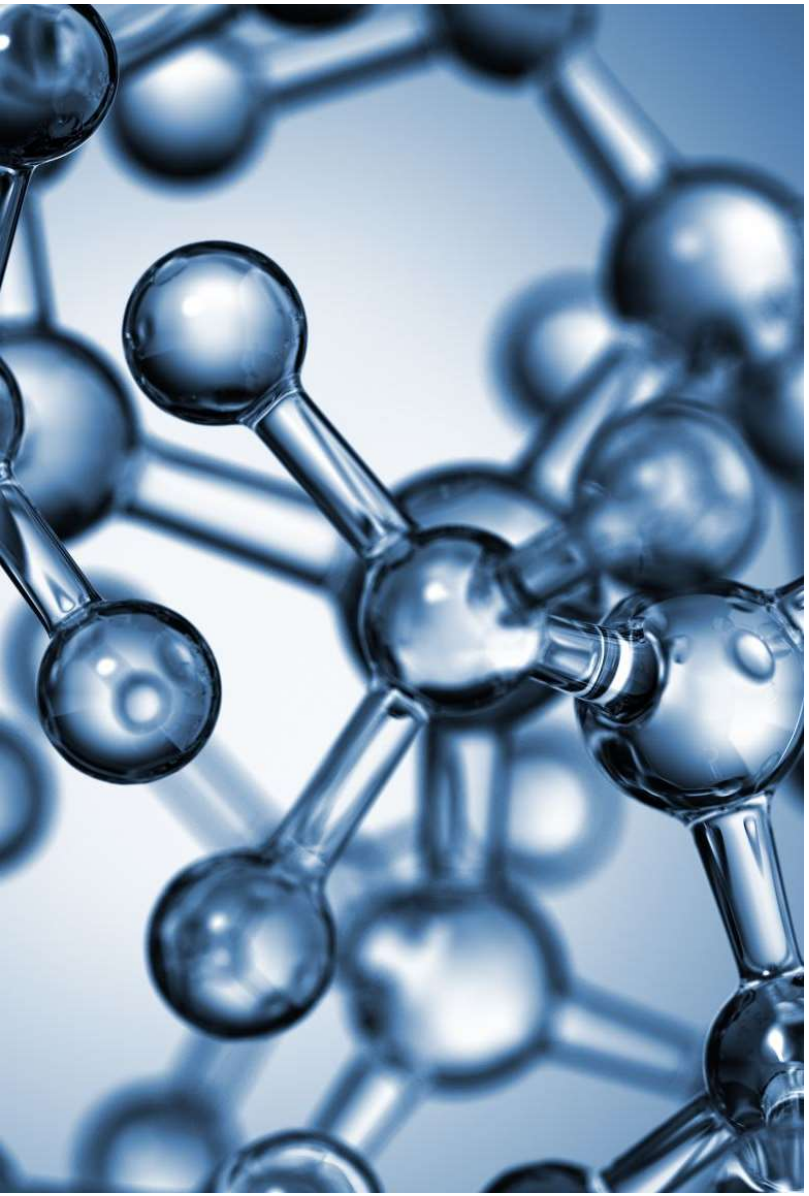
- Go to the following website and make an account:

https://portal.xsede.org

- Write your username on the front board

Resources:

https://github.com/dansh351/Murdock-PIS-Computational-Biology

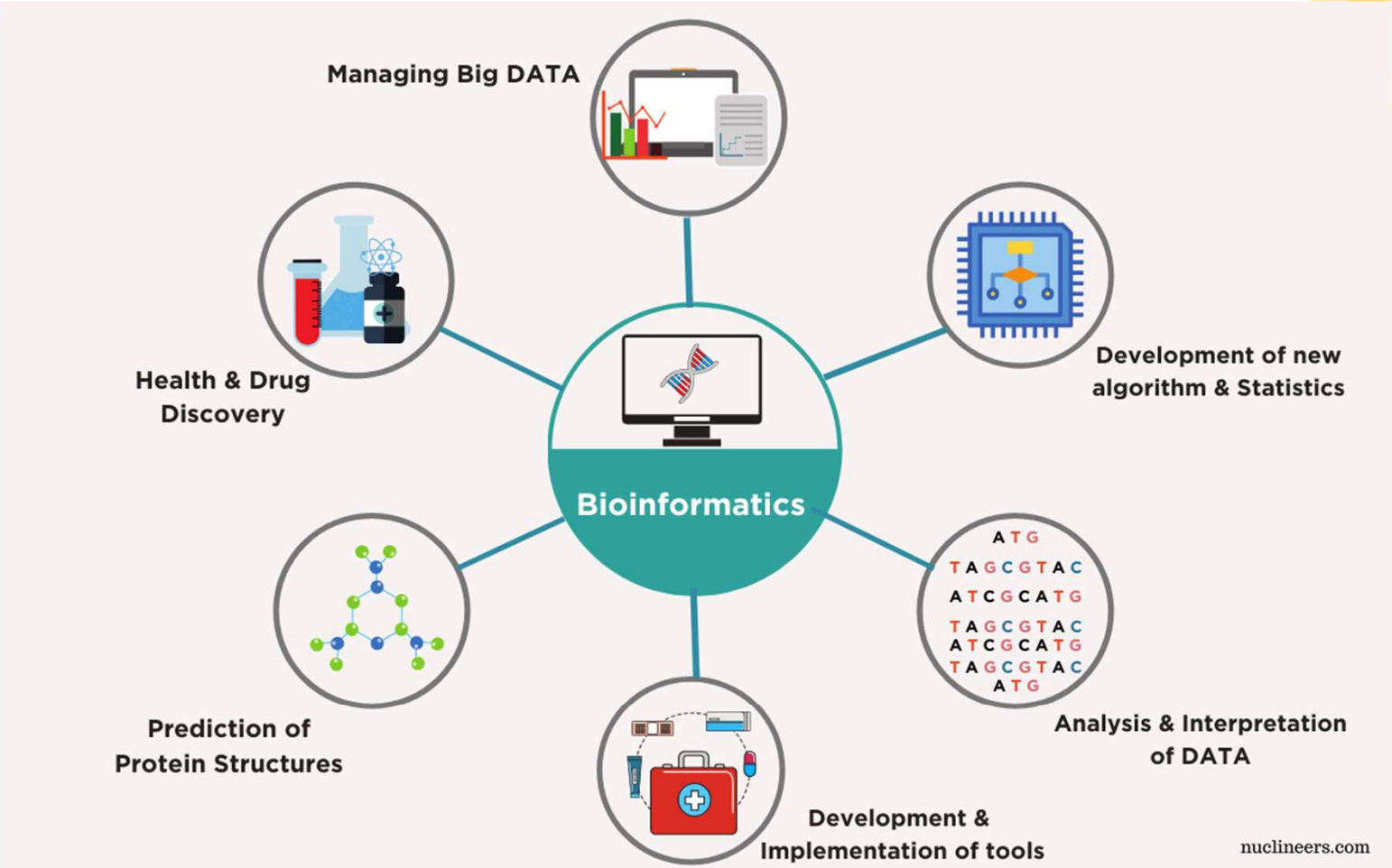


Bioinformatics

Bioinformatics is a field of computational science that has to do with the analysis of sequences of biological molecules. [It] usually refers to genes, DNA, RNA, or protein, and is particularly useful in comparing genes and other sequences in proteins and other sequences within an organism or between organisms, looking at evolutionary relationships between organisms, and using the patterns that exist across DNA and protein sequences to figure out what their function is.

Christopher P. Austin, M.D.

genome.gov


Managing Big DATA
Health & Drug Discovery
Development of new algorithm & Statistics
Bioinformatics
Prediction of Protein Structures
Analysis & Interpretation of DATA
Development & Implementation of tools
nuclineers.com

Why is bioinformatics important?

Why is bioinformatics important?

Quick Facts: Computer and Information Research Scientists	
2020 Median Pay	$126,830 per year $60.97 per hour
Typical Entry-Level Education	Master's degree
Work Experience in a Related Occupation	None
On-the-job Training	None
Number of Jobs, 2019	32,700
Job Outlook, 2019-29	15% (Much faster than average)
Employment Change, 2019-29	5,000

Basic Local Alignment Search Tool (BLAST)

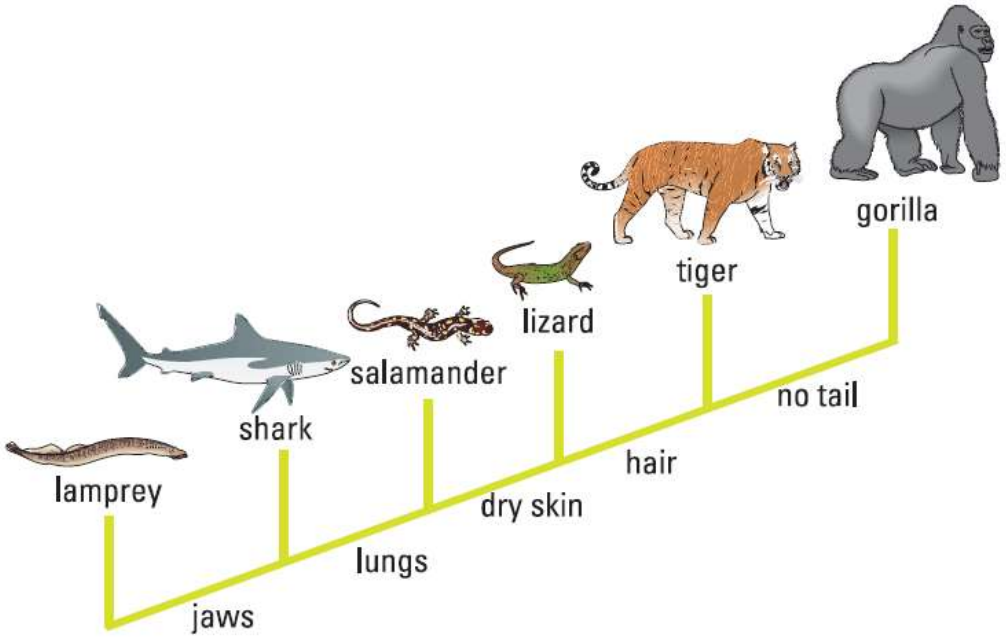


Figure 2. Cladogram of Several Animal Species

BLAST Lab (Briefly…)

Procedure

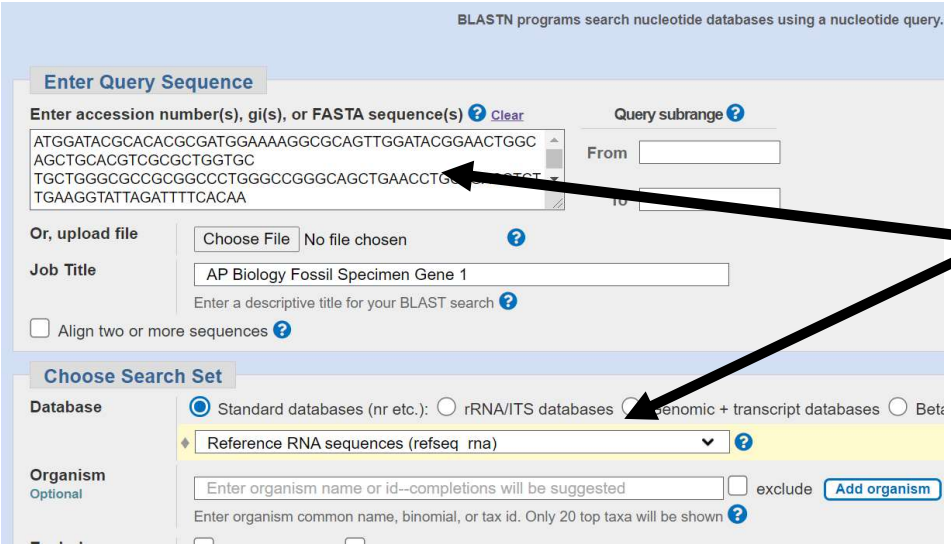
A team of scientists has uncovered the fossil specimen in Figure 3 near Liaoning Province, China. Make some general observations about the morphology (physical structure) of the fossil, and then record your observations in your notebook.

Little is known about the fossil. It appears to be a new species. Upon careful examination of the fossil, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes. Your task is to use BLAST to analyze these genes and determine the most likely placement of the fossil species on Figure 4.

https://blogging4biology.edublogs.org/2010/08/28/college-board-lab-files/

This is a pre-programmed set of parameters for BLAST to use. The downloaded files contain these parameters.

Upload the downloaded search strategy files here and click view.



BLAST will automatically input everything where it belongs.

Gene 1

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Gallus gallus collagen type V alpha 1 chain (COL5A1), mRNA	Gallus gallus	10091	10091	100%	0.0	99.32%	8250	NM_204790.3

Gene 2

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Drosophila melanogaster FI02063 full insert cDNA	Drosophila mel…	4420	4420	92%	0.0	99.88%	2458	BT050432.1

Gene 3

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	PREDICTED: Taeniopygia guttata ubiquitin conjugating enzyme E2 Q1 (UBE2Q1), mRNA	Taeniopygia gut…	2193	2193	95%	0.0	99.59%	1774	XM_030291666.3

Gene 4

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Alligator sinensis mitochondrion, complete genome	Alligator sinensis	1768	1768	100%	0.0	100.00%	16746	AF511507.1

Based on “Total score”, the fossil specimen can most likely be placed here.

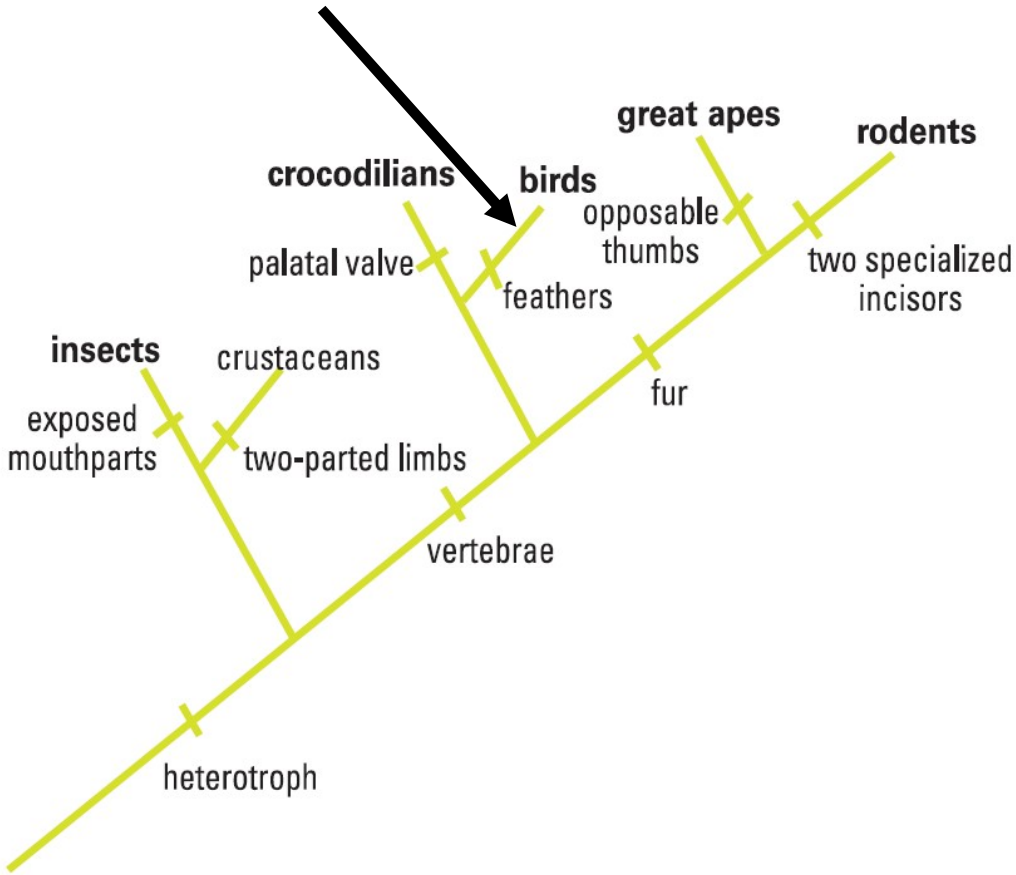


Figure 4. Fossil Cladogram

BLAST® » blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

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

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

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 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): YouTube 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

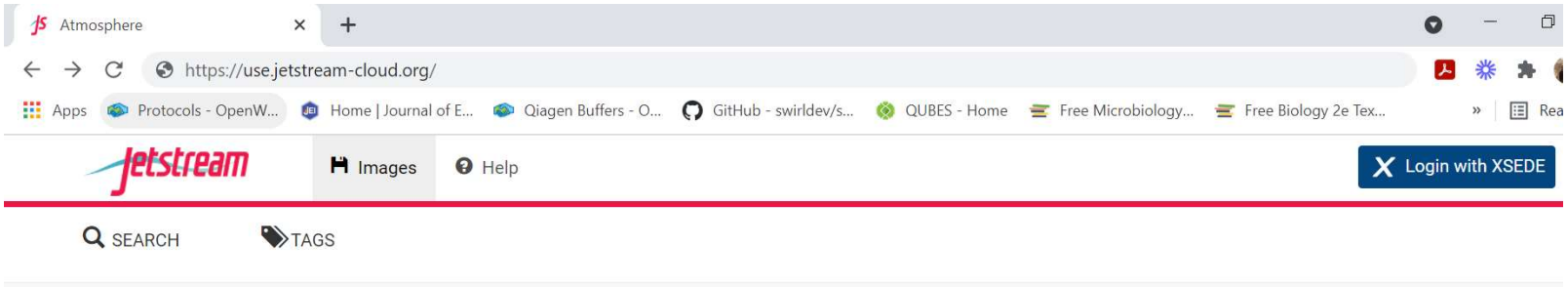
BLAST

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

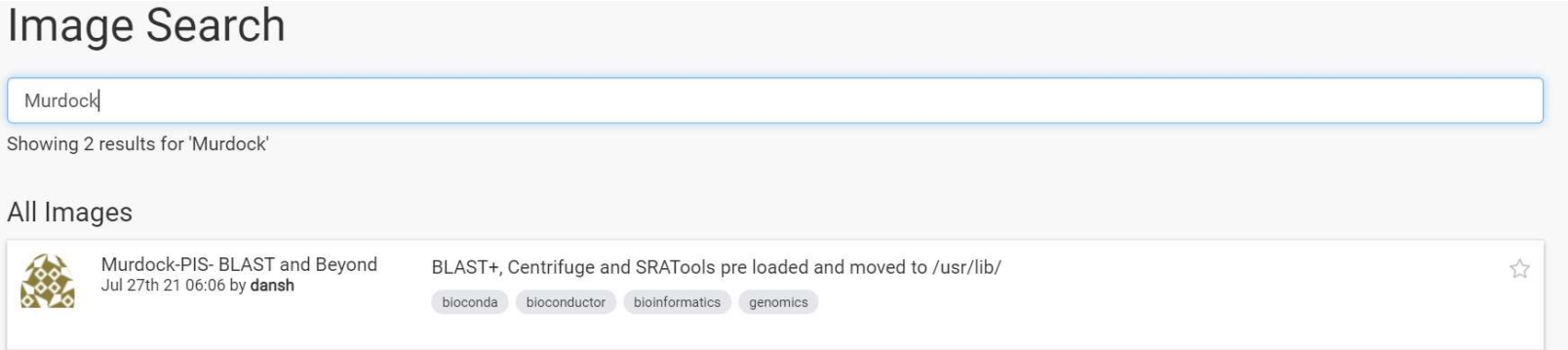
Note: Parameter values that differ from the default are highlighted in yellow and marked with + sign

+ Algorithm parameters



Log in using your XSEDE username and password

Once logged in, click the Images tab.



Search for the Murdock –PIS- BLAST and Beyond Image

← Murdock-PIS- BLAST and Beyond

+ ADD TO PROJECT | Launch

Created: 7/27/2021 06:06 am PDT

Created by: dansh

Description: BLAST+, Centrifuge and SRATools pre loaded and moved to /usr/lib/

Visibility: Public

Tags: bioconda | bioconductor | bioinformatics | genomics

Launch the Image

Basic Info

Instance Name

Murdock-PIS- BLAST and Beyond

Base Image Version

2.0

Project

dantest

Instance Count

1

Resources

Allocation Source

TG-SEE210005

Provider

Jetstream - Indiana University

Instance Size

m1.medium (CPU: 6, Mem: 16 GB, Disk: 60 GB)

Allocation Used

72% of 100500 SUs from TG-SEE210005

Resources Instance will Use

A total 6 of 132 alloted CPUs

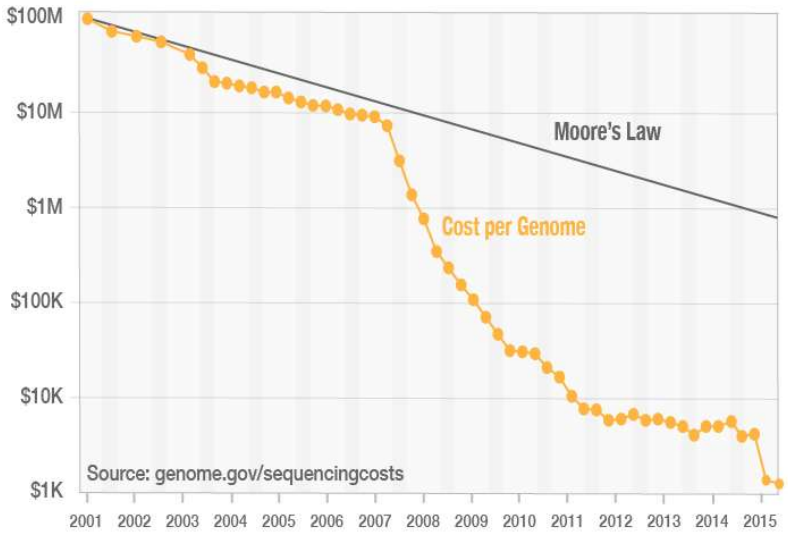
A total 16 of 360 alloted GBs of Memory

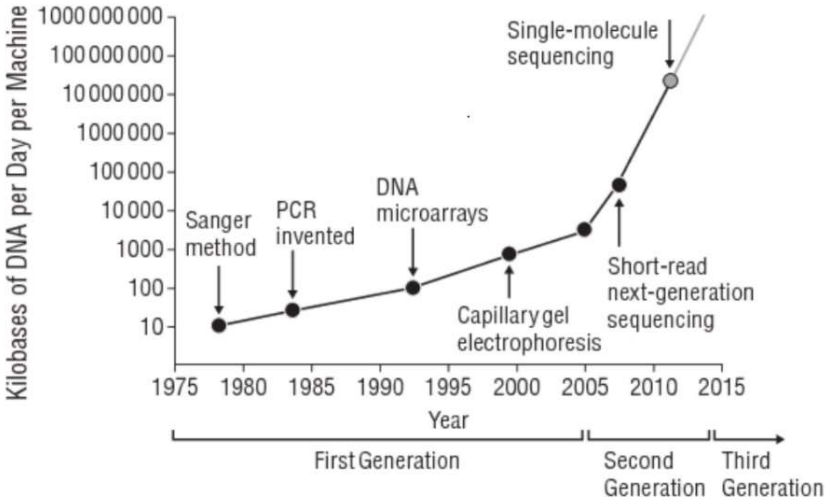
Make sure the image is “medium” sized and the Provider is set to Jetstream-Indiana University

BLAST+ Demonstration

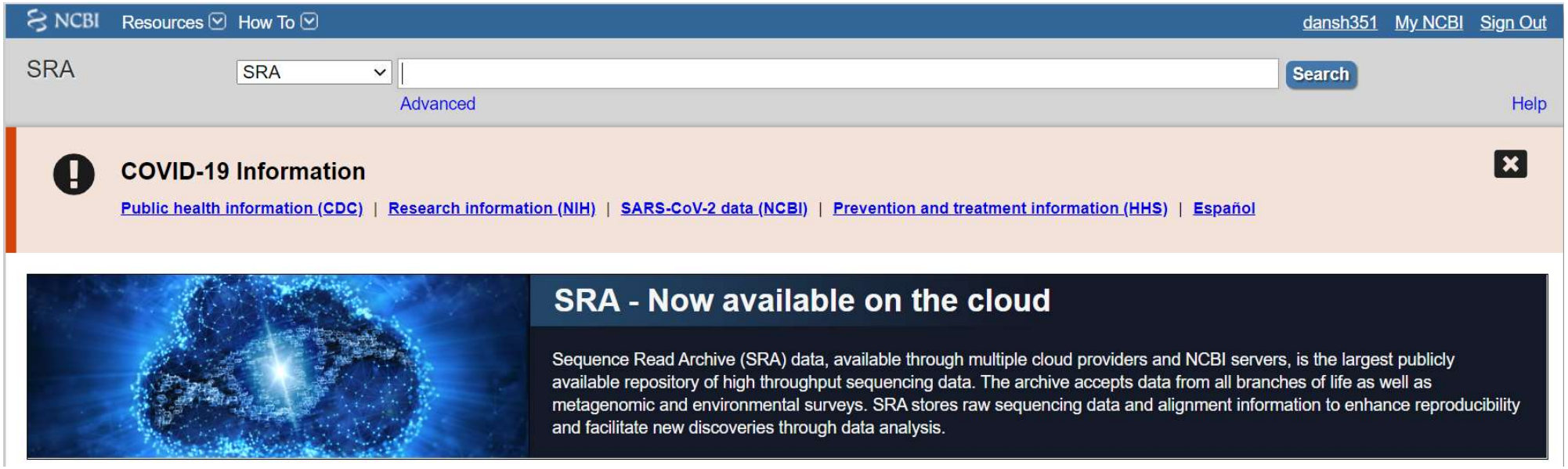
Beyond BLAST

- BLAST is an extremely powerful, and accurate sequence aligner
- It is the most used bioinformatics tool today, and there are over 200,000 BLAST searches conducted per week
- However…
  - Its 20+ years old
  - Limited to FASTA sequences
  - Very accurate, but slow.
  - **Its not the only tool out there!**


$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
Cost per Genome
Source: genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015


Kilobases of DNA per Day per Machine
1000 000 000
100 000 000
10 000 000
1000 000
100 000
10 000
1000
100
10
Single-molecule sequencing
Sanger method
PCR invented
DNA microarrays
Capillary gel electrophoresis
Short-read next-generation sequencing
1975 1980 1985 1990 1995 2000 2005 2010 2015
Year
First Generation
Second Generation
Third Generation

NCBI Sequence Read Archive (SRA)



“Messing About” with Metagenomics

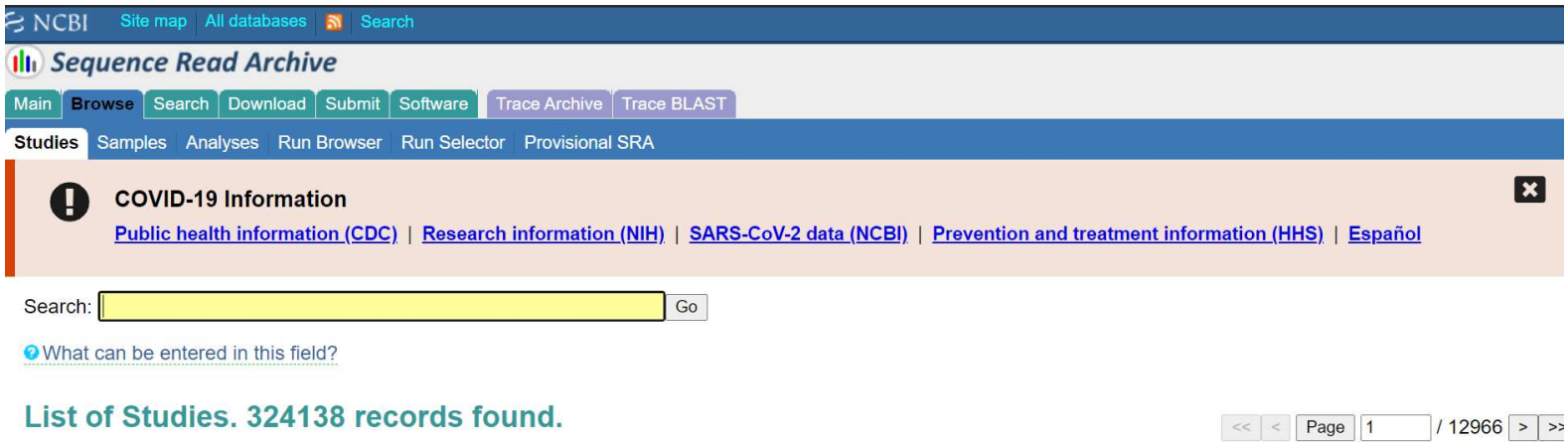
```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions

export PATH=/usr/lib/centrifuge-1.0.4-beta:$PATH
export CENTRIFUGE_INDEXES=/usr/lib/centrifuge-1.0.4-beta/Index
export PATH=/usr/lib/sratoolkit.2.11.0-centos_linux64/bin:$PATH
```



Once you get to the Studies tab of SRA, type metagenome into the search bar

#	Accession	Title	Project	Center
1.	DRP003219 download	Effects of a subchronic and mild social defeat stress on murine cecal microbiota	336524	NILGS
2.	DRP007245 download	wetland eukaryotic community	728004	PUSAN
3.	DRP007251 download	wetland bacterial community	728002	PUSAN
4.	DRP007252 download	wetland fungal community	728003	PUSAN
5.	DRP007363 download	Mangrove metagenome	735343	KOCHI
6.	DRP007452 download	Blood microbiome in febrile patients	743243	NUGSM

Browse around until you get to the study you like, then click the accession number on the left.

A. gambie ovaries Metagenome

Identifiers: SRA: SRP041157
BioProject: PRJNA244534

Study Type: Metagenomics

Abstract: The study aims at characterizing the reproductive tract microbiome of natural populations of A. gambiae.

Related SRA data

Experiments: 1 (1 samples)
Runs: 2 (87.6Gbp; 54.1Gb)

Additional objects:

File type	count
fastq	4

This page will give some information on the study conducted

Click on “Runs”

Found 2 Items

	Run	Bases	Bytes
1	SRR1238105	45.97 G	28.77 Gb
2	SRR1238106	41.59 G	25.33 Gb

Scroll until you find the Accession number

```
[js-157-254] dantest ~-->fasterq-dump --split-files DRR128241
```

In the Terminal, type the following, using your accession number

Centrifuge!

Centrifuge
Classifier for metagenomic sequences

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Centrifuge is a very rapid and memory-efficient system for the classification of DNA sequences from microbial samples, with better sensitivity than and comparable accuracy to other leading systems. The system uses a novel indexing scheme based on the Burrows-Wheeler transform (BWT) and the Ferragina-Manzini (FM) index, optimized specifically for the metagenomic classification problem. Centrifuge requires a relatively small index (e.g., 4.3 GB for ~4,100 bacterial genomes) yet provides very fast classification speed, allowing it to process a typical DNA sequencing run within an hour. Together these advances enable timely and accurate analysis of large metagenomics data sets on conventional desktop computers.

OSI certified

```
[js-157-254] dantest ~-->centrifuge -x p+h+v -U DRR125241.fastq -S wetland_class
ification.tsv --report-file wetland_report.tsv
```

↑

Type this in your command line and watch the magic happen

```
 GNU nano 2.3.1                File: wetland_report.tsv

name    taxID   taxRank genomeSize      numReads        numUniqueReads  abundance
Azospirillum brasilense 192     species 13978806        2       0       0.0
Brucella        234     genus   4068975 1       1       0.0
Pseudomonas fluorescens 294     species 6526868 3       1       0       0.0
Pseudomonas stutzeri    316     species 4548207 1       0       8.95947e-09
Xanthomonas campestris  339     species 7516154 1       0       0.0
Xanthomonas campestris pv. campestris   340     leaf    14898506        1       0       0.0
Legionella pneumophila  446     species 6459328 1       0       6.30863e-09
Escherichia     561     genus   7192399 1       0       0.0
Escherichia coli        562     species 7253110 97      0       5.65493e-09
Serratia liquefaciens   614     species 5282719 1       1       7.71374e-09
Shigella        620     genus   4815334 4       0       0.0
Shigella sonnei 624     species 5137894 5       0       0
Rhodobacter     1060    genus   4407391 1       0       0.0
Rhodobacter sphaeroides 1063    species 4585882 2       0       0.0
Porphyrobacter neustonensis     1112    species 3090363 1       1       0.0
Staphylococcus epidermidis      1282    species 2601987 1       0       0.0
Streptococcus salivarius        1304    species 2213879 1       1       1.84064e-08
Propionibacterium acnes 1747    species 2515790 49      24      0
Mycobacterium bovis     1765    species 5455659 1       0       0.0
Mycobacterium tuberculosis      1773    species 4956632 1       0       0.0
Streptomyces albus      1888    species 6841649 2       0       0.0
Streptomyces clavuligerus       1901    species 7590758 1       1       0.0
Streptomyces glaucescens        1907    species 7623774 1       1       0.0
Aeromicrobium erythreum 2041    species 3629239 1       0       0.0
Homo sapiens    9606    species 3238442024      672     546     4.61799e-09
Escherichia virus Lambda        10710   species 48502   10      0       0
```

“Ask the Data” Project