Metagenomics and disease

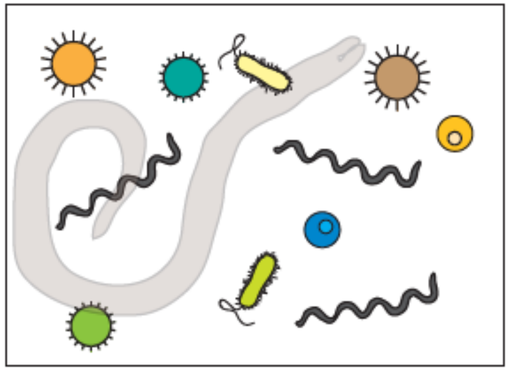
Mark Stenglein, GDW

Genomics of Disease in Wildlife

Metagenomics is the study of >1 genome

Many genomes

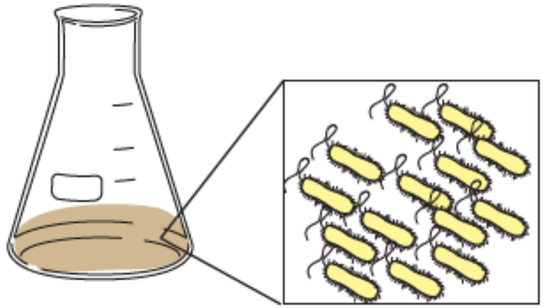
soil community

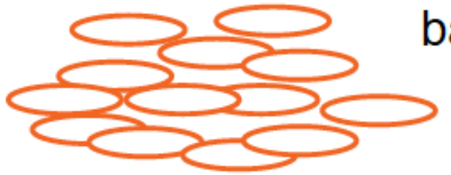


soil ‘metagenome’

1 genome

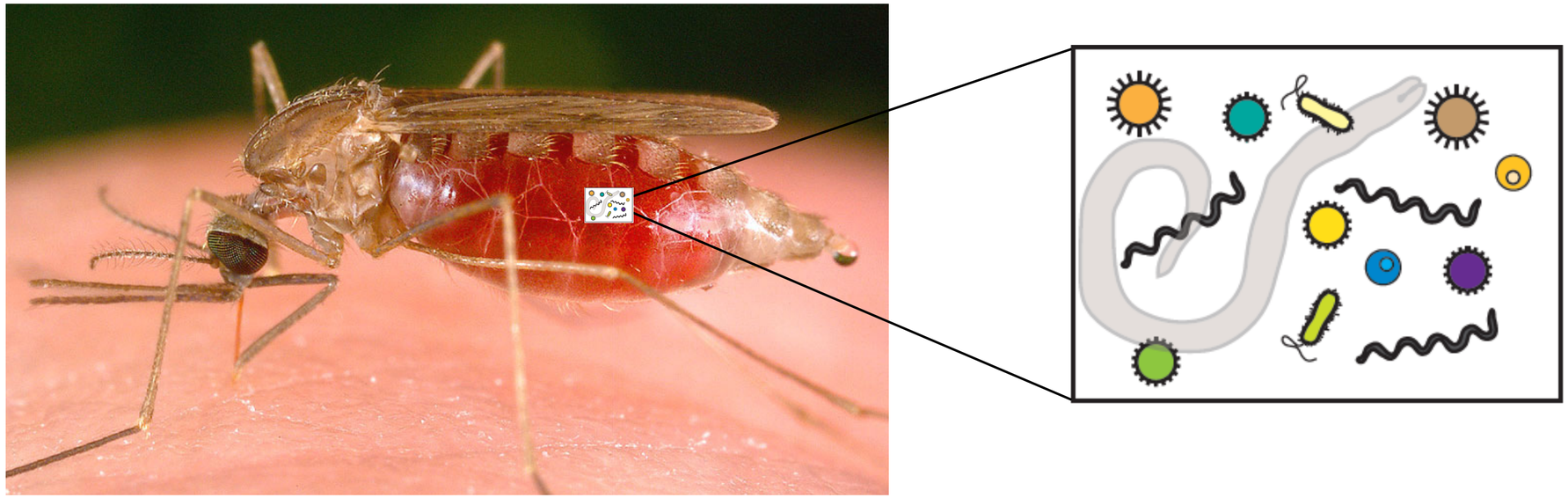
bacterial isolate



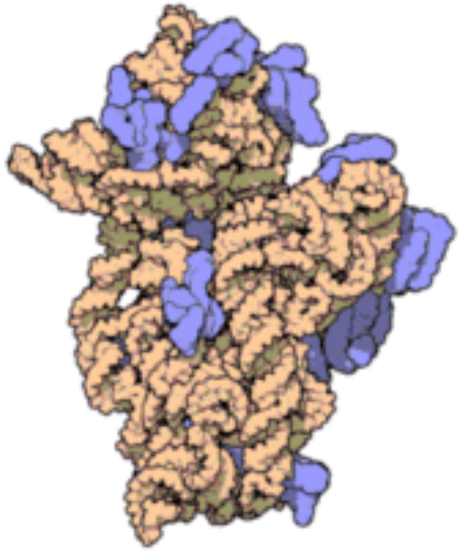


bacterial genome

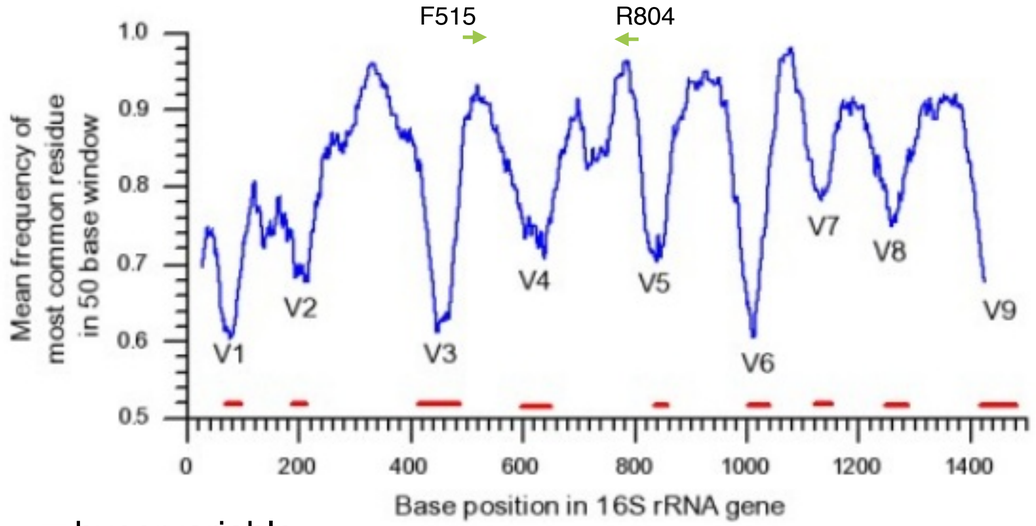
If you sequence total nucleic acid from an intact multicellular organism, you are doing metagenomics



16S sequencing (microbiome sequencing) could be considered a form of metagenomics

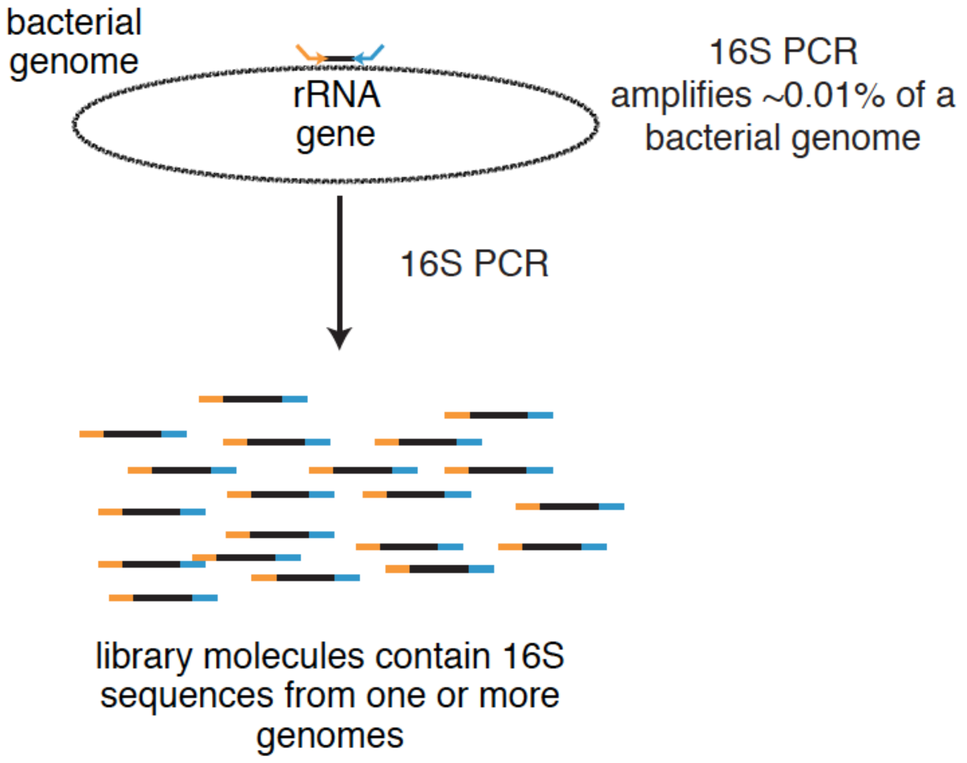


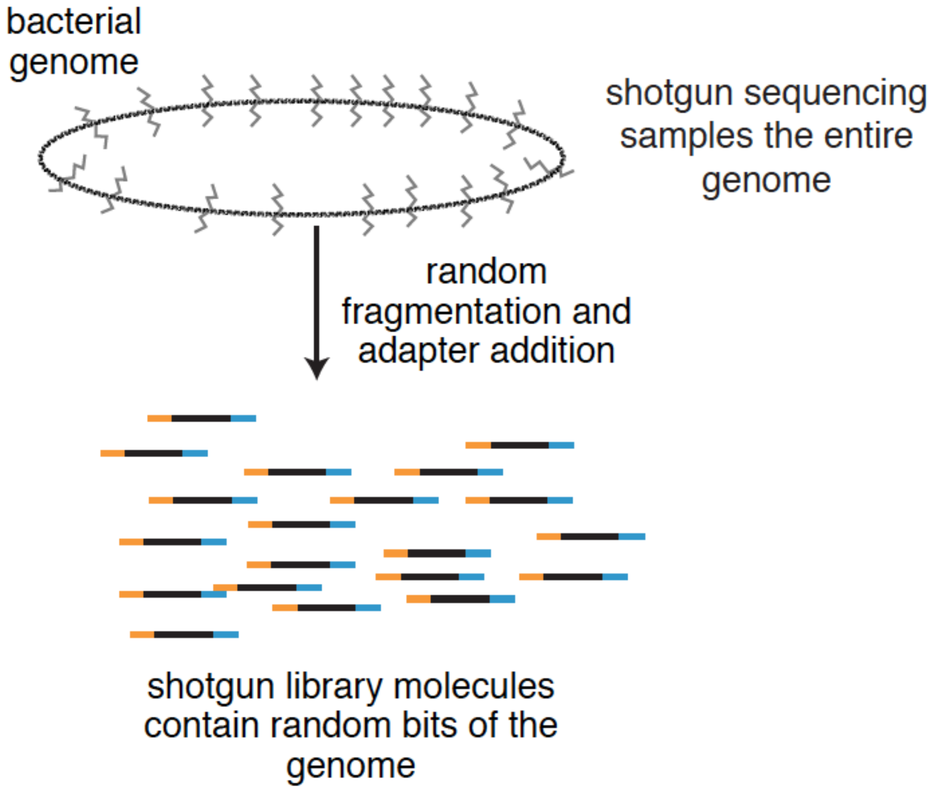
bacterial 30S ribosomal subunit
16S rRNA is in orange
(purple: ribosomal proteins)
image: wikipedia



hypervariable

16S sequencing vs. shotgun metagenomics



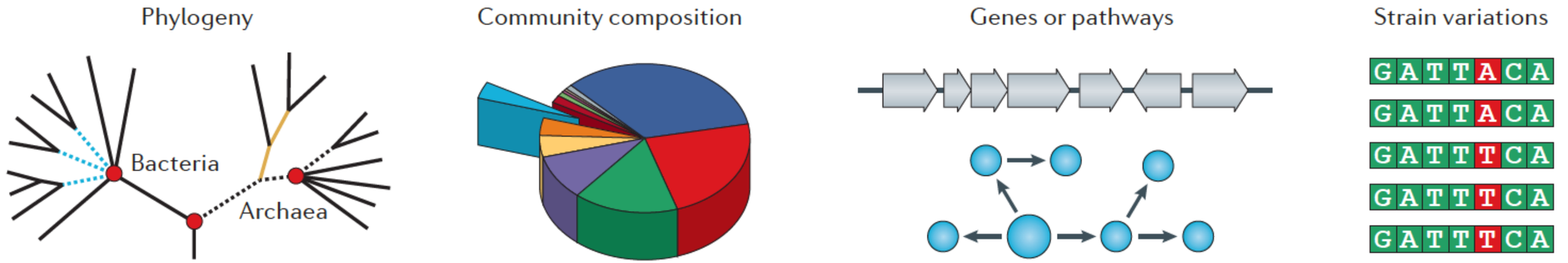


- Only bacteria and archaea surveyed
- Deeper sampling of bacterial diversity per $
- Relatively easy to make libraries and interpret results
- Simple way to sample microbial community diversity

- All organisms studied*
- Decreased sampling depth per $
- Enables analysis of other genomic features of organisms, e.g. antimicrobial resistant genes
- Analysis is significantly more difficult

Applications of metagenomics

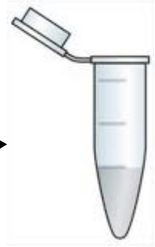
- Community composition analysis (environmental samples, microbiomes, …)
- Characterization of diet, bloodmeal composition
- Identification of genes of interest: AMR genes, industrially useful enzymes
- Pathogen detection and discovery



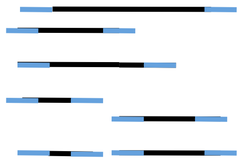
Lasken and McLean (2014) Nat Rev Gen

Pathogen detection and discovery using metagenomic sequencing

case and control tissues

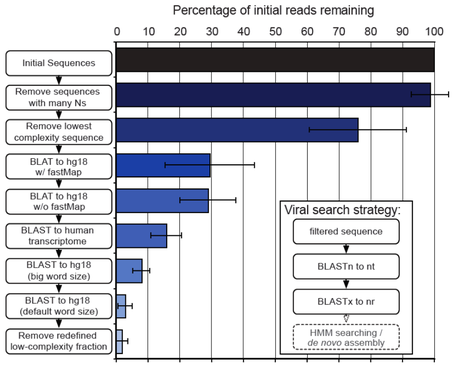


Nucleic acid



Library prep / barcode

Illumina sequencing

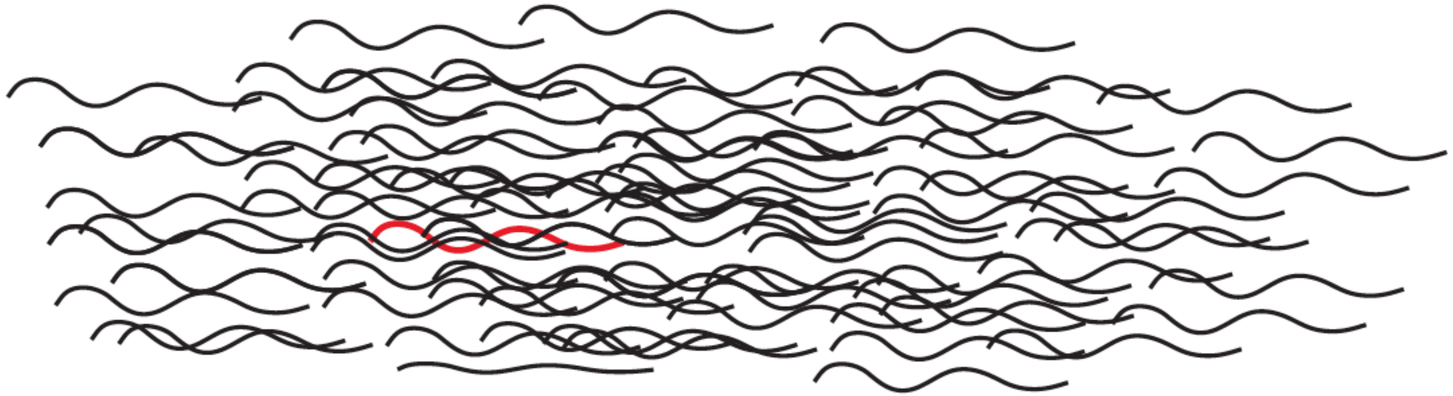


Computational Analysis

Follow-up

Some challenges for metagenomics pathogen discovery

1) Pathogen nucleic acid is typically present in a sea of host nucleic acid



~1 viral nucleic acid per 10^4-10^7 host nucleic acids

2) New pathogens have unknown sequences

TTTCAG?TTT?ACC????TG??AAA?ACATCC??TATAC???T?

3) Misannotated sequences in databases confound results

4) Case/control studies can give you information about correlation but not causation

How sensitive is NGS for pathogen detection?
In theory, a single read is sufficient to identify a pathogen
(but that's cutting it a little close)

A single read pair aligning to **caprine herpesvirus-2**
amongst ~0.5M mule deer reads

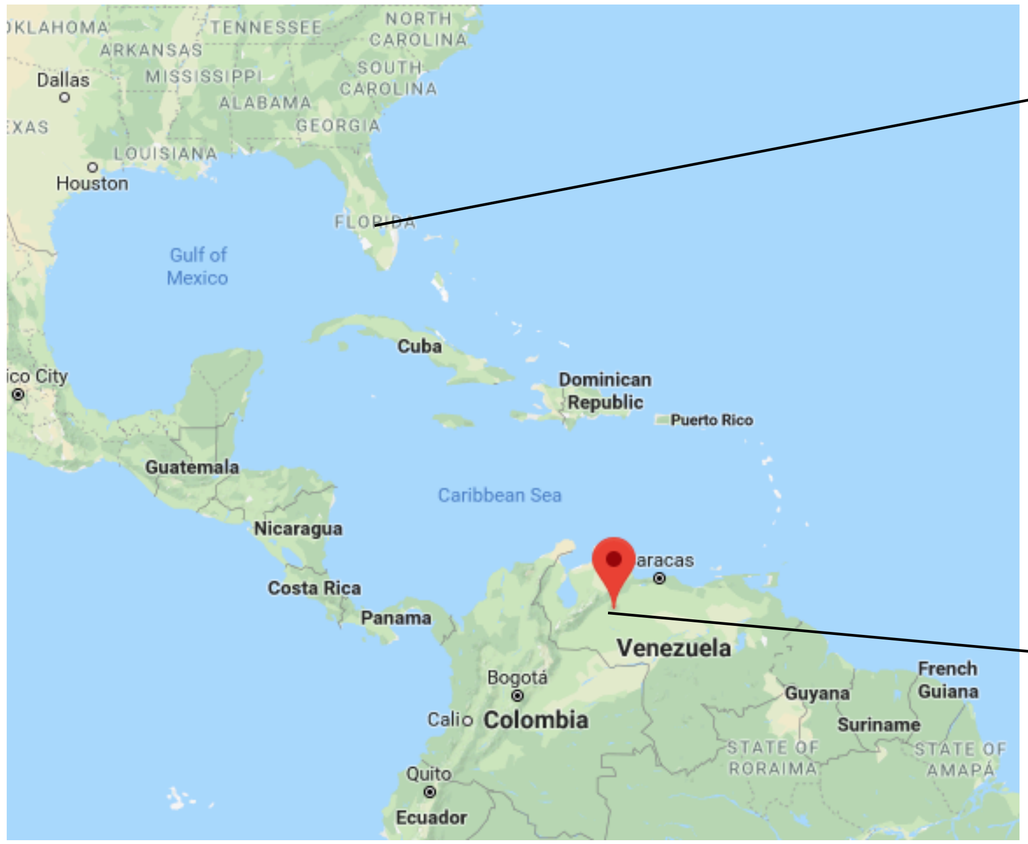
Identification of this pathogen completely consistent with histopathology

case had been tested for *ovine* herpesvirus 2

Laura Hoon-Hanks, DVM
Samples from: Karen Fox DVM, CO Parks & Wildlife

PCR is generally more sensitive than NGS for targeted pathogen detection

Example of metagenomic pitfall: Guanarito virus sequence supposedly in a pool of *Culex cedecei* mosquitoes collected in the Florida Everglades



Arenavirus
Cause of Venezuelan hemorrhagic fever
Not known to be arthropod borne

image: American Society of Mammalogists

One of the putative Guanarito virus sequences

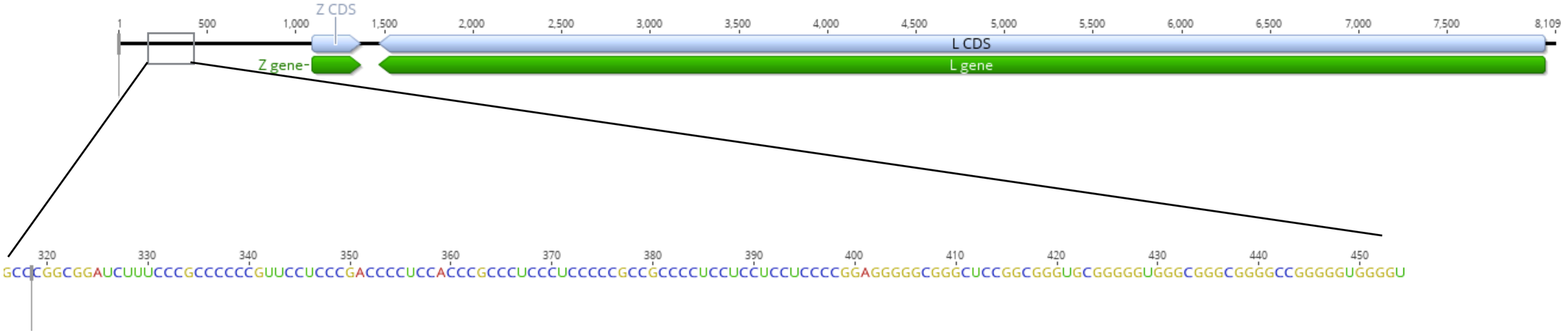
```
>NODE_274_length_640_cov_11681_ID_547
GCGGGGGTGGGCGGGCGGGGCCGGGGGTGGGGTCGGCGGGGGACCGTCCCCCGACCGGCGACCGGCCGCCGCCGGGC
GCATTTCCACCGCGGCGGTGCGCCGCGACCGGCTCCGGGACGGCTGGGAAGGCCCGGCGGGGAAGGTGGCTCGGGGG
GCCCCGTCCCGCCCCGTCTTCCCCCCGCCCGCGTCCTCCCCCGGGAGGGCGCGGGTCGGGGCGGCGGCGGCGGTGGC
GGCGGGACCACCCCCCGAGTGTTACAGCCCCCCGGCAGCAGCACTCGCCGAATCCCGGGGCCGAGGGAGCGAGACCC
GTCGCCGCGCTCTCCCCCCTCCCGGCGCCCACCCCCGCGGGGGCCCCCCGCGCGGGGGTCCCCCCCCGCGGGGGCGCG
CCGGCGGTCCTCGTGGGGGGCCGGGCCACCCCTCCCACGGCGCGACCGCTCTCCCACCCCCTCCCCGCACCCCCGGC
GACGGGGGCCCGCGCGGGTGGGGGCGGGGCGGACTGTCCCCAGTGCGCCCCGGGCGGGTCGCGCCGTCGGGCCCGGG
GGGGTTCTCTCGGGGCCACGCGCGCGTCCCTCGAAGAGGGGGACGGCGGAGCGAGCGCACGGGGTCGGCGGCGATGT
CGGCTACCCACCCGACCCGTCTTG
```

This sequence is linked on the GitHub agenda. What is this sequence?

BLAST the sequence vs. the NCBI nucleotide database

Description	Max score	Total score	Query cover	E value	Ident	Accession
Guanarito mammarenavirus isolate CVH-960201 segment L, complete sequence	1081	1081	94%	0.0	99%	KU746283.1
Guanarito mammarenavirus isolate CVH-950801 segment S, complete sequence	1064	1064	90%	0.0	99%	KU746280.1
Chimpanzee 28S ribosomal RNA gene fragment	826	826	100%	0.0	89%	M30950.1
Gorilla 28S ribosomal RNA gene fragment	817	817	99%	0.0	89%	M30951.1
Homo sapiens external transcribed spacer 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, internal transcribed spacer 2, 28S ribosomal RNA gene, and e	808	808	100%	0.0	89%	KY962518.1
Homo sapiens clone BAC JH1 genomic sequence	808	1612	100%	0.0	89%	MF164269.1
Homo sapiens RNA, 45S pre-ribosomal N2 (RNA45SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146144.1
Homo sapiens RNA, 28S ribosomal N2 (RNA28SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146148.1
Human DNA sequence from clone CH507-146P16 on chromosome 21, complete sequence	804	804	100%	0.0	89%	CT476837.18
Human ribosomal DNA complete repeating unit	798	798	100%	0.0	89%	U13369.1
Homo sapiens clone BAC JH5 genomic sequence	787	787	100%	0.0	88%	MF164266.1
Homo sapiens RNA, 28S ribosomal N3 (RNA28SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146154.1
Homo sapiens RNA, 45S pre-ribosomal N3 (RNA45SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146151.1
Human DNA sequence from clone CH507-528H12 on chromosome 21, complete sequence	784	1707	100%	0.0	88%	FP236383.15

These Guanarito virus sequences are mis-assembled

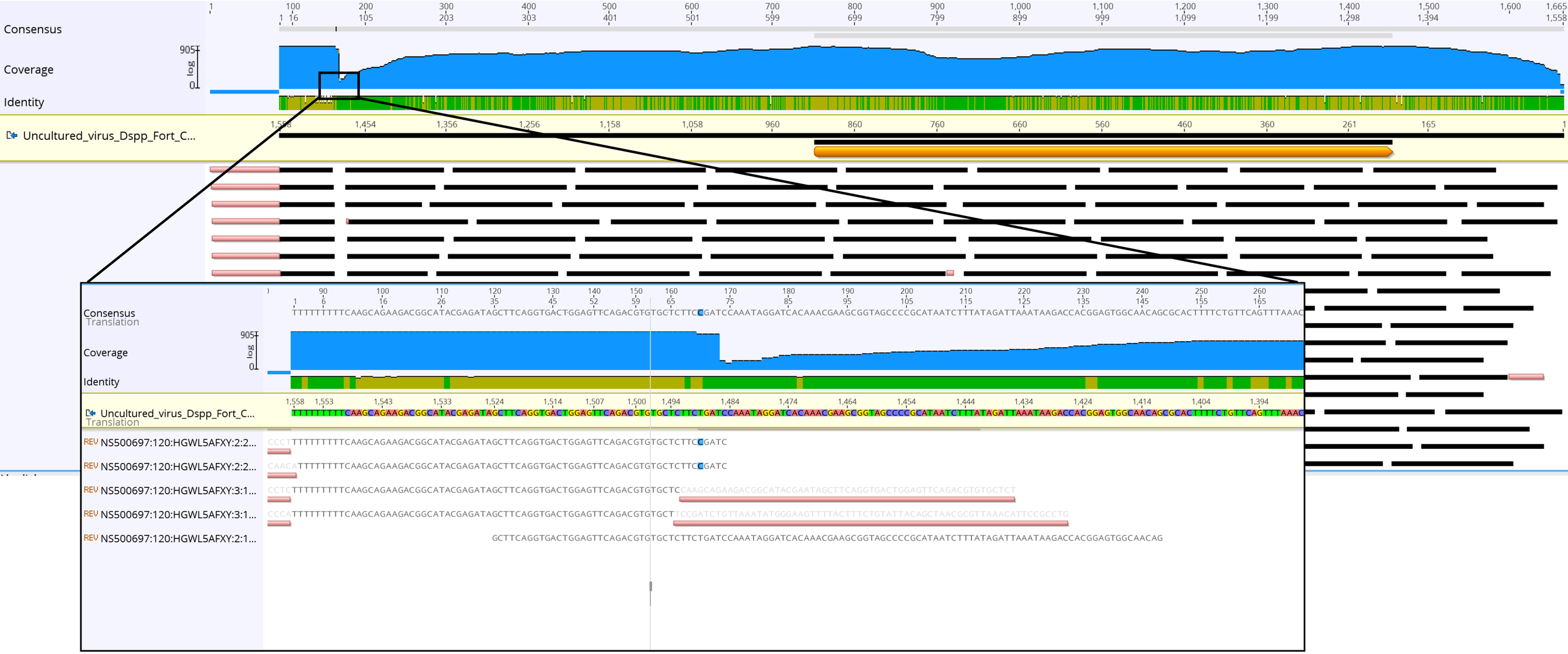


Conclusion: don’t blindly trust database annotation nor the output of analysis software

```
COMMENT     GenBank Accession Numbers KU746283, KU746284 represent sequences
            from the 2 segments of Guanarito mammarenavirus CVH-960201.

            ##Genome-Assembly-Data-START##
            Assembly Method :: Trimmomatic v. 0.32
            SGA v. 0.10.13
            iMetAMOS v. 1.5
            samtools v1.1
            FastQC v. 0.10.0
            Spades v. 3.1.1
            idba v1.1.1
            Pilon v. 1.8
            Quast v. 2.2
            Prokka v. 1.7
            Assembly Name :: GTOV014-SEQ-1-ASM-1
            Genome Coverage :: 6779.96x
            Sequencing Technology :: Illumina HiSeq1500
            ##Genome-Assembly-Data-END##.
FEATURES             Location/Qualifiers
     source          1..8109
                     /organism="Guanarito mammarenavirus"
                     /mol_type="genomic RNA"
```

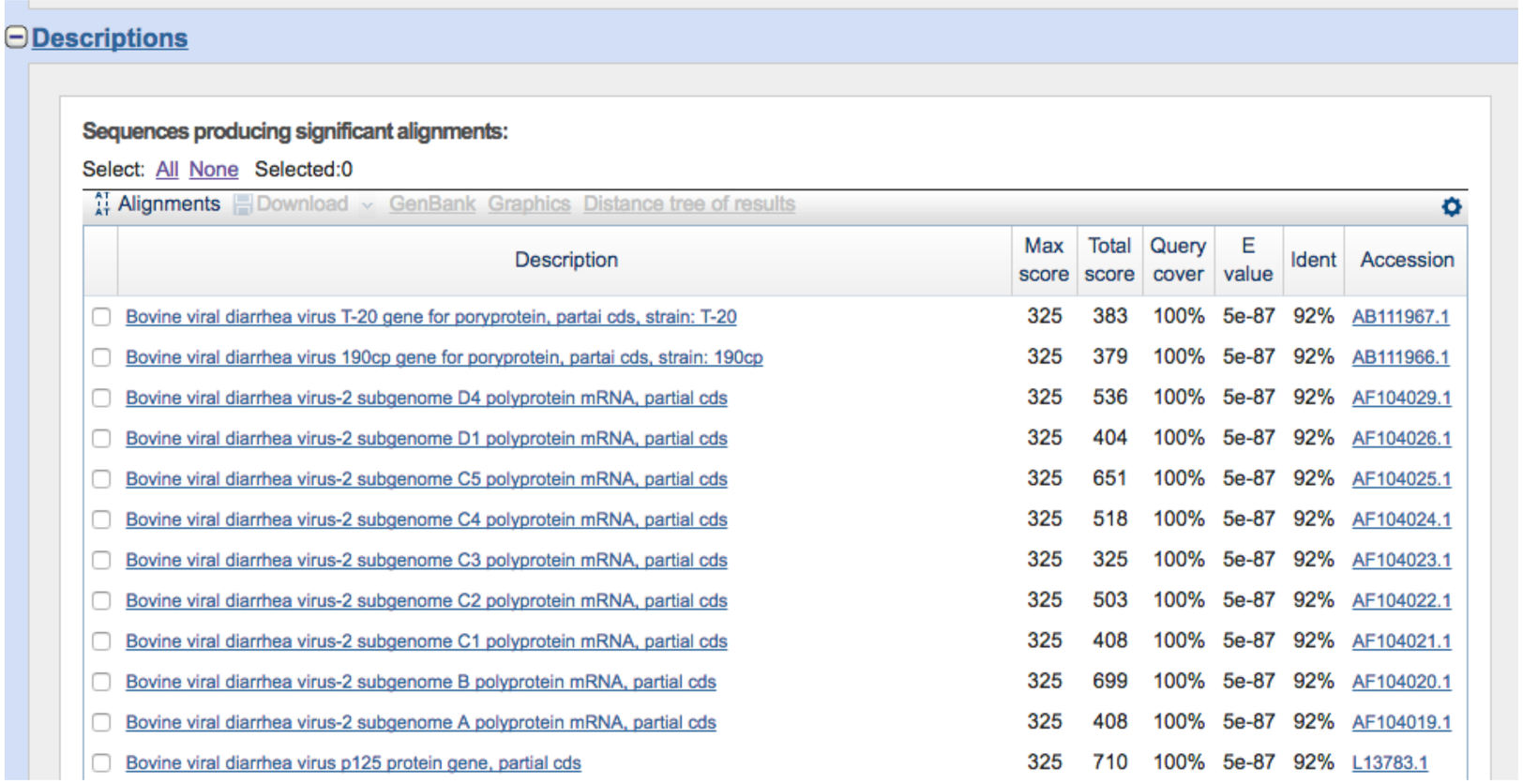
A single read (a PCR chimera?) triggered a similar missassembly



Another caveat: using smaller databases (e.g. all viral genomes in RefSeq) is faster but it can produce misleading results

Here: a read was BLASTed against all of the virus nucleotide sequences in Genbank

```
>a_sequence
ATGCAGATCTTCGTGAAGACTCTGACTGGTAAGACCATCACCCTCGAGGTTGAGCC...
```



Descriptions

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
Bovine viral diarrhea virus T-20 gene for poryprotein, partai cds, strain: T-20	325	383	100%	5e-87	92%	AB111967.1
Bovine viral diarrhea virus 190cp gene for poryprotein, partai cds, strain: 190cp	325	379	100%	5e-87	92%	AB111966.1
Bovine viral diarrhea virus-2 subgenome D4 polyprotein mRNA, partial cds	325	536	100%	5e-87	92%	AF104029.1
Bovine viral diarrhea virus-2 subgenome D1 polyprotein mRNA, partial cds	325	404	100%	5e-87	92%	AF104026.1
Bovine viral diarrhea virus-2 subgenome C5 polyprotein mRNA, partial cds	325	651	100%	5e-87	92%	AF104025.1
Bovine viral diarrhea virus-2 subgenome C4 polyprotein mRNA, partial cds	325	518	100%	5e-87	92%	AF104024.1
Bovine viral diarrhea virus-2 subgenome C3 polyprotein mRNA, partial cds	325	325	100%	5e-87	92%	AF104023.1
Bovine viral diarrhea virus-2 subgenome C2 polyprotein mRNA, partial cds	325	503	100%	5e-87	92%	AF104022.1
Bovine viral diarrhea virus-2 subgenome C1 polyprotein mRNA, partial cds	325	408	100%	5e-87	92%	AF104021.1
Bovine viral diarrhea virus-2 subgenome B polyprotein mRNA, partial cds	325	699	100%	5e-87	92%	AF104020.1
Bovine viral diarrhea virus-2 subgenome A polyprotein mRNA, partial cds	325	408	100%	5e-87	92%	AF104019.1
Bovine viral diarrhea virus p125 protein gene, partial cds	325	710	100%	5e-87	92%	L13783.1

Cool, looks like a flavivirus! Right?

Keep analyses as unbiased as possible

The same read was BLASTed against all the nucleotide sequences in Genbank (the ‘nt’ database):

```
>a_sequence
ATGCAGATCTTCGTGAAGACTCTGACTGGTAAGACCATCACCCTCGAGGTTGAGCC...
```

Sequences producing significant alignments:

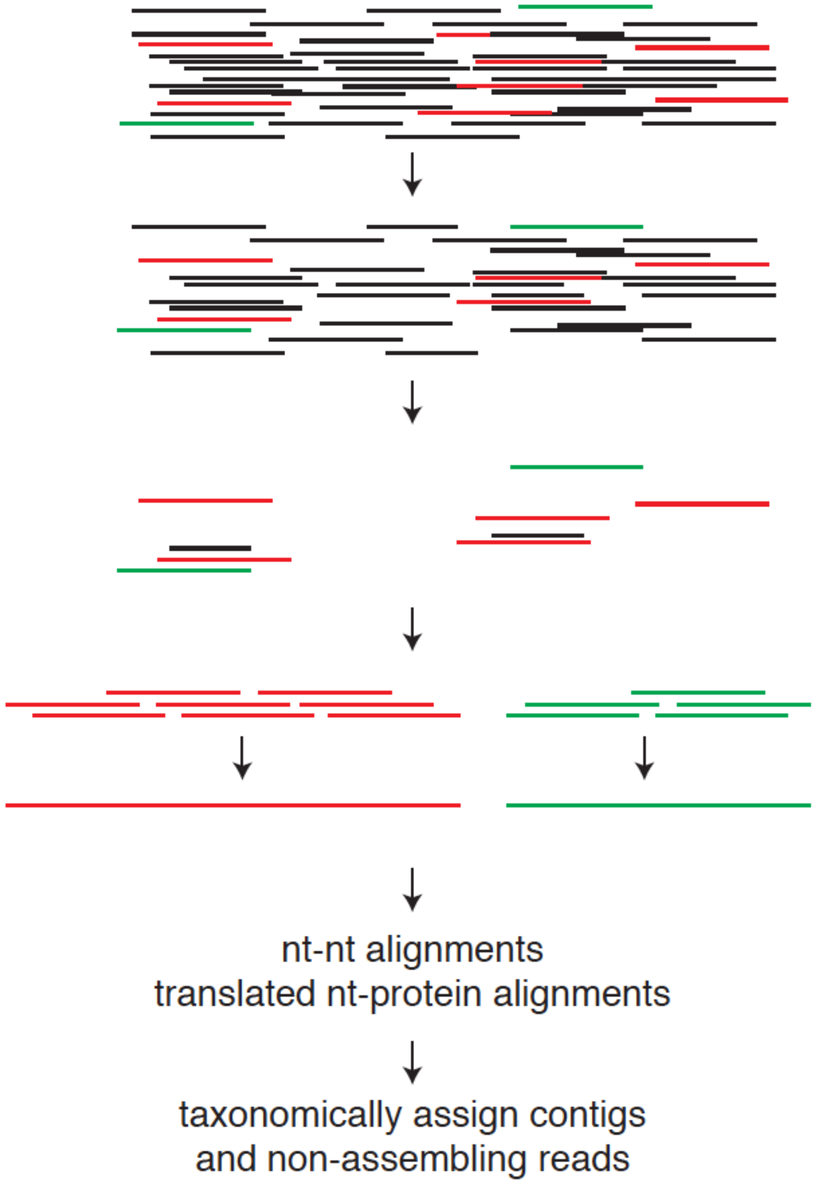
Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens ubiquitin C (UBC), RefSeqGene on chromosome 12	412	3261	100%	2e-111	100%	NG_027722.2
Homo sapiens ubiquitin C (UBC), mRNA	412	3261	100%	2e-111	100%	NM_021009.6
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKJ9	412	3241	100%	2e-111	100%	AB643790.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: DJL8	412	2881	100%	2e-111	100%	AB643789.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: DJL9	412	3266	100%	2e-111	100%	AB643788.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKS7	412	2558	100%	2e-111	100%	AB643787.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: MKS9	412	3257	100%	2e-111	100%	AB643786.1
Homo sapiens UbC gene for ubiquitin C, complete cds, clone: BHP7	412	2549	100%	2e-111	100%	AB643785.1
Pan troglodytes mRNA for ubiquitin, complete cds, clone: PtsC-51-5_D12	412	1833	100%	2e-111	100%	AK306071.1

Some BVDV genomes contain ubiquitin homologs

A typical pathogen discovery analysis workflow



Quality filter /
remove PCR duplicates
cutadapt
cd-hit-est

Remove host sequences
bowtie2

Assemble remaining reads into contigs
SPAdes

nt-nt alignments
translated nt-protein alignments

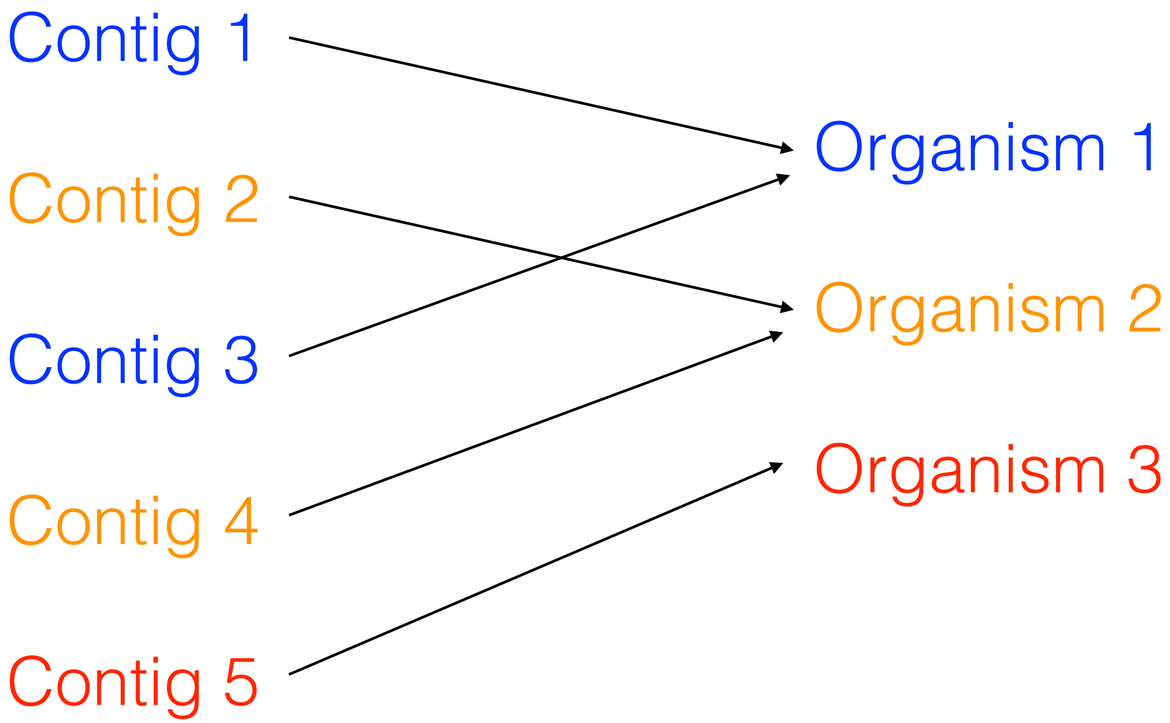
taxonomically assign contigs and non-assembling reads

Identify most similar sequences in NCBI databases
BLASTN (gsnap),
BLASTX (diamond)

~40 min for a dataset w/ 6M reads

caveat: for a dataset where host filtering removed almost all of the reads

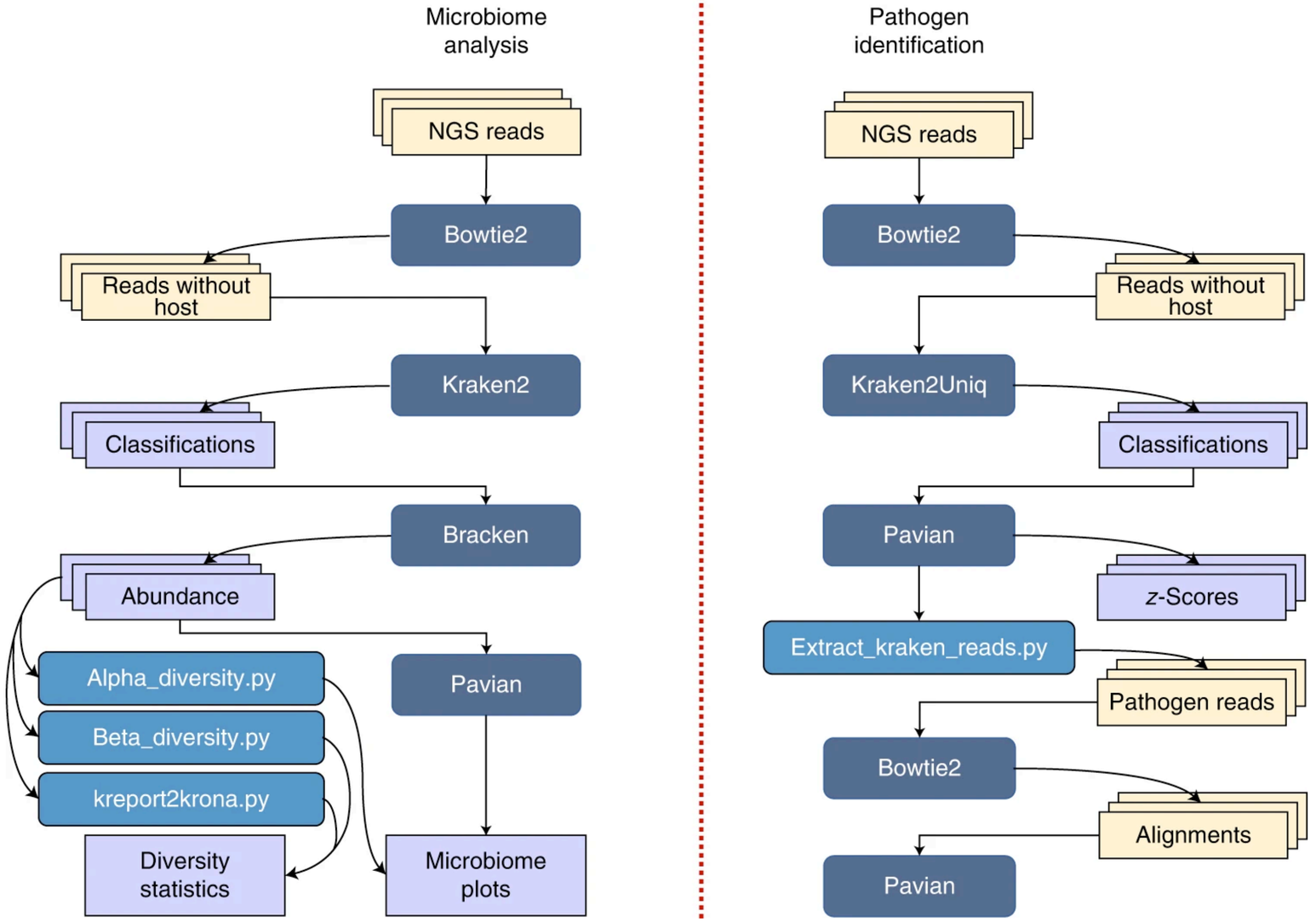
The goal of metagenomic classification software is to map sequence information to taxonomic information.



Earlier, we did this by BLASTing several contigs on the NCBI website. This is not a practical approach for many contigs.

Nucleotide-level similarity identifies closely related organisms (blastn-like (blastn))
Protein-level similarity discovers ‘new’ organisms (blastx-like (diamond))

There are also metagenomic classifiers that skip assembly and classify reads



Kraken2: Lu et al: https://www.nature.com/articles/s41596-022-00738-y

Metagenomic classification can be challenging

Resource intensive
Large databases
Large assemblies
Memory and storage intensive

Bioinformatics challenges

User-friendly bioinformatics software for analysis of mNGS data is not currently available. Thus, customized bioinformatics pipelines for analysis of clinical mNGS data[56,109–111] still require highly trained programming staff to develop, validate and maintain the pipeline for clinical use. The laboratory can either host computational servers locally or move the bioinformatics analysis and data storage to cloud platforms. In either case, hardware and software setups can be complex, and adequate measures

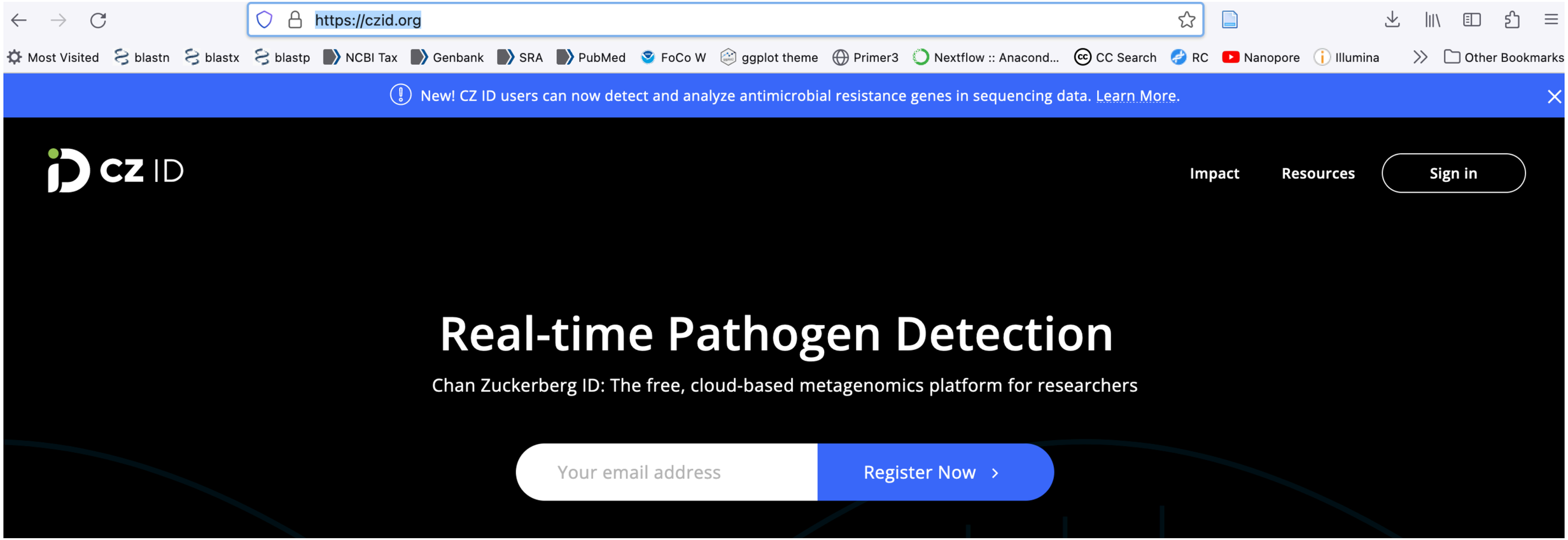
Clinical metagenomics

Charles Y. Chiu[1,2]* and Steven A. Miller[1]

350 | JUNE 2019 | VOLUME 20

Nat Rev Genetics

czid.org is a new web-based tool that does metagenomic classification



CHAN ZUCKERBERG BIOHUB

The snake sample we analyzed run through this tool

https://idseq.net/samples/18731?currentTab=Metagenomic

IDSEQ | My Data | Public | Upload | Mark Stenglein

Metagenomic Pipeline v3.7, NT/NR: 2018-12-01 | processed 2 years ago

Snake_7 >

SRR1984309

Sample Details

Share | Download | ?

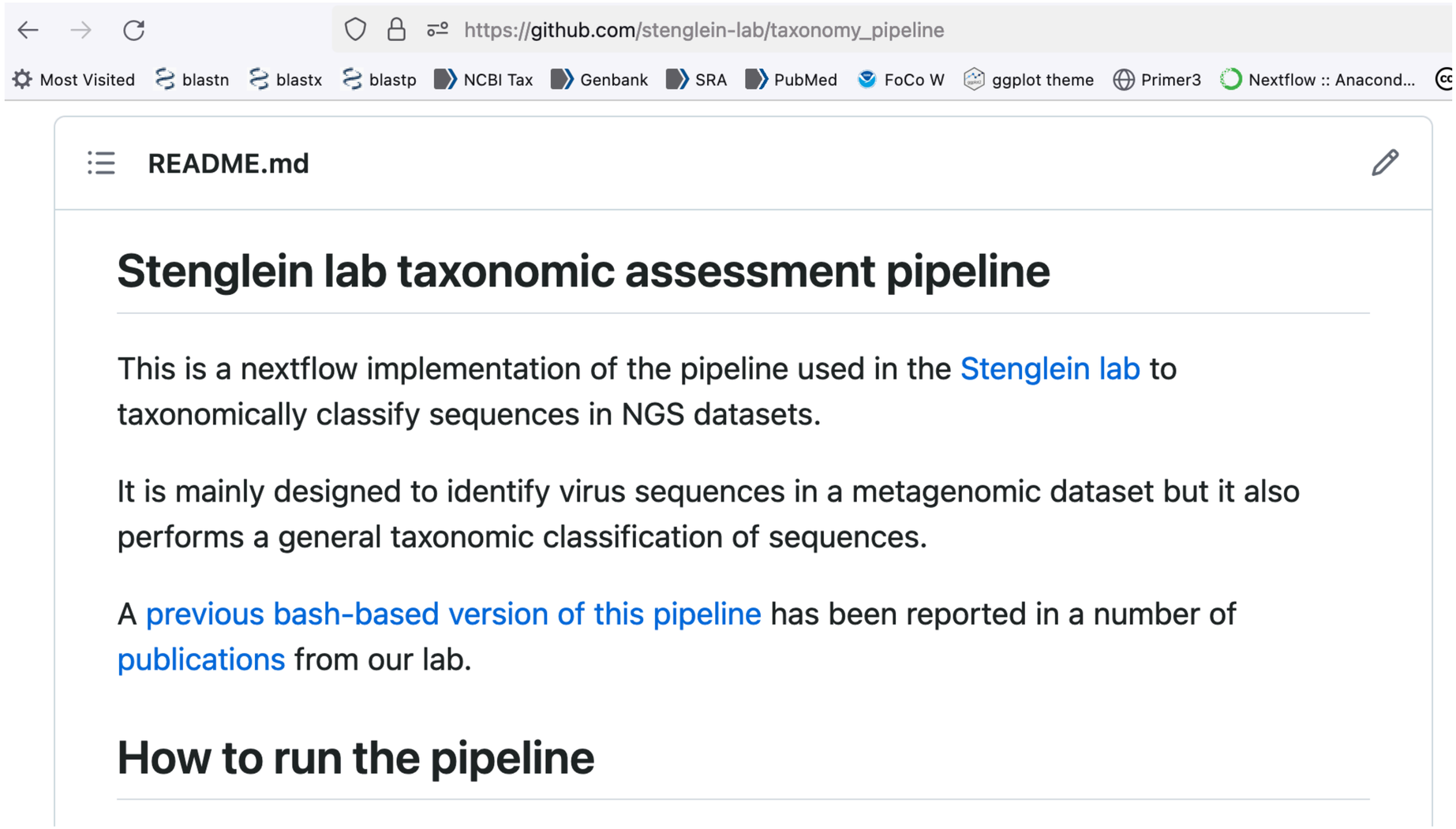
Metagenomic | Antimicrobial Resistance

Taxon name | **Name Type:** Scientific | **Background:** NID Human CSF v3 | **Categories** | **Threshold Filters** | **Read Specificity:** All

6153 total rows.

Taxon	Score	Z Score	rPM	r	contig	contig r	%id	L	E value	NT NR
Reptarenavirus *(11 viral species)*	**1,042,963,754**	**99.0** 99.0	**57,382.5** 50,860.0	**16,601** 14,714	**14** 20	**12,126** 12,215	**98.8** 98.2	**1,195.7** 386.3	**10^{-233}** 10^{-266}	
Clostridium *(35 bacterial species)* ●● 2	**115,659,009**	**76.3** 99.0	**916.0** 11,921.7	**265** 3,449	**0** 1	**0** 3,412	**99.1** 36.1	**26.1** 58.8	**10^{-6}** 10^{-4}	
Pogonomyrmex *(2 eukaryotic species)*	**39,646,877**	**100.0** 0.0	**3,971.6** 0.0	**1,149** 0	**0** 0	**0** 0	**100.0** 0.0	**22.0** 0.0	**10^{-3}** 0	
Solanum *(7 eukaryotic species)*	**28,728,424**	**99.0** -100.0	**4,704.4** 0.0	**1,361** 0	**0** 0	**0** 0	**99.0** 0.0	**22.9** 0.0	**10^{-4}** 0	
Caenorhabditis *(6 eukaryotic species)*	**27,474,943**	**99.0** -100.0	**3,563.7** 0.0	**1,031** 0	**0** 0	**0** 0	**96.9** 0.0	**24.3** 0.0	**10^{-4}** 0	

Our lab's pipeline is available on GitHub if you want to see how we do it.

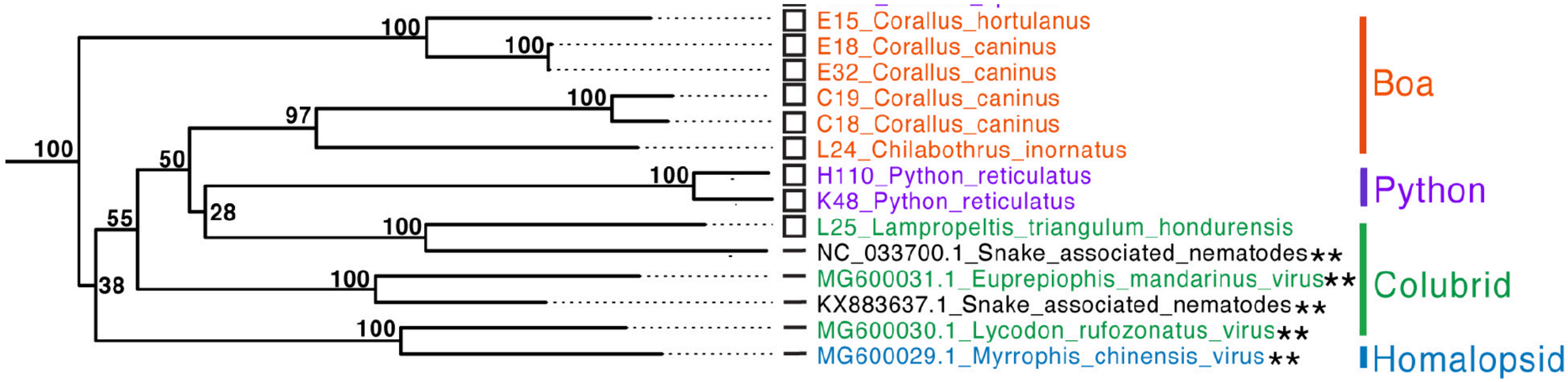


- Sharing pipelines on places like GitHub facilitates reproducibility

https://github.com/stenglein-lab/taxonomy_pipeline

Metagenomic sequencing only gives you sequences

Serpentoviruses detected in snakes with respiratory disease and also snake-associated nematodes

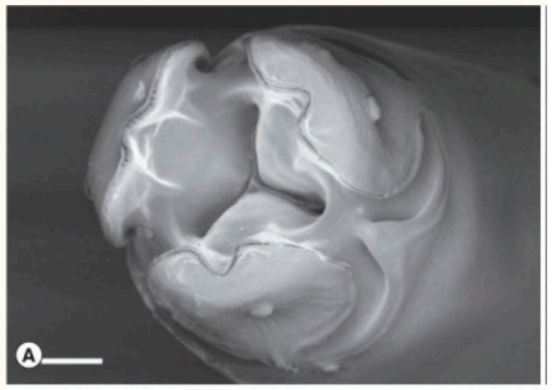


Are these related viruses infecting both nematodes and snakes?

I’d bet that the ‘nematode’ viruses really infect snakes

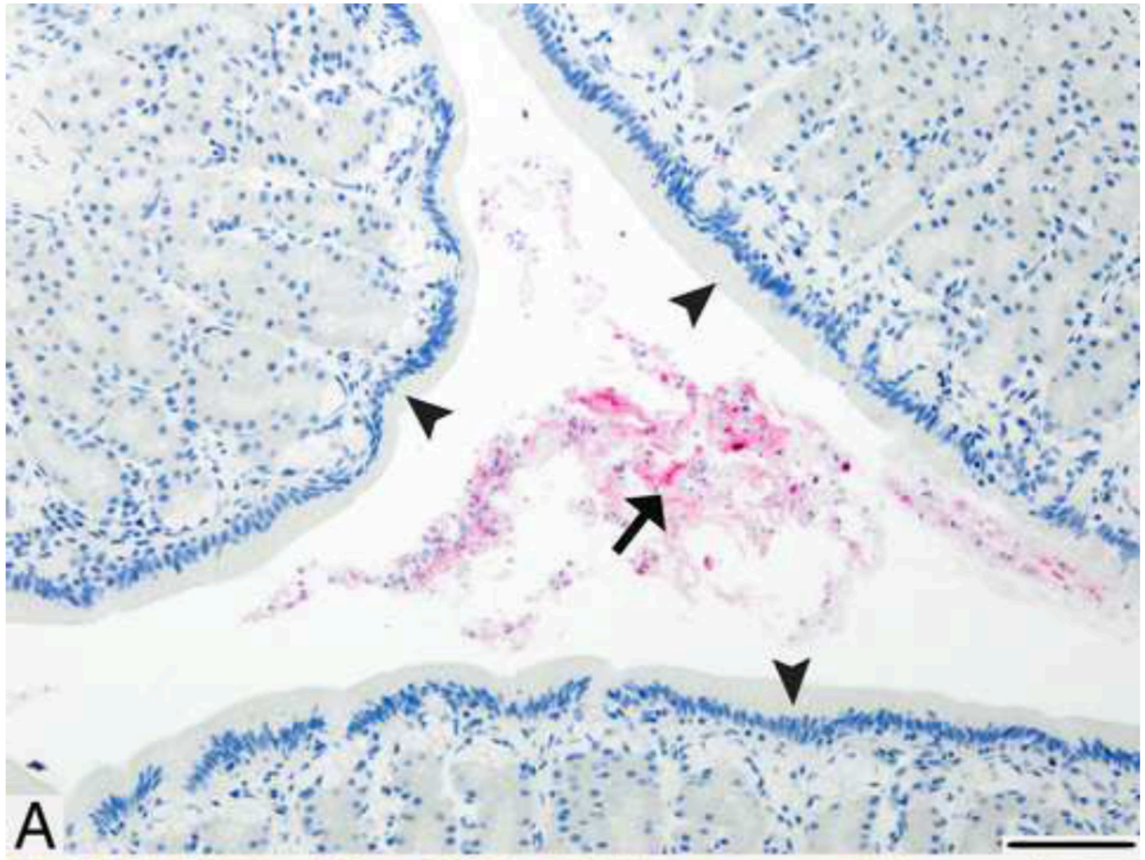
Laura Hoon-Hanks

SEM of a snake nematode



Choe et al (2016)

Serpentovirus antigen detected in python intestinal lumen



Astroviruses associated with fatal gastroenteritis in rabbits: likely the cause but need additional proof

A rabbit facility in TN experienced an outbreak of fatal gastroenteritis

image: wikipedia

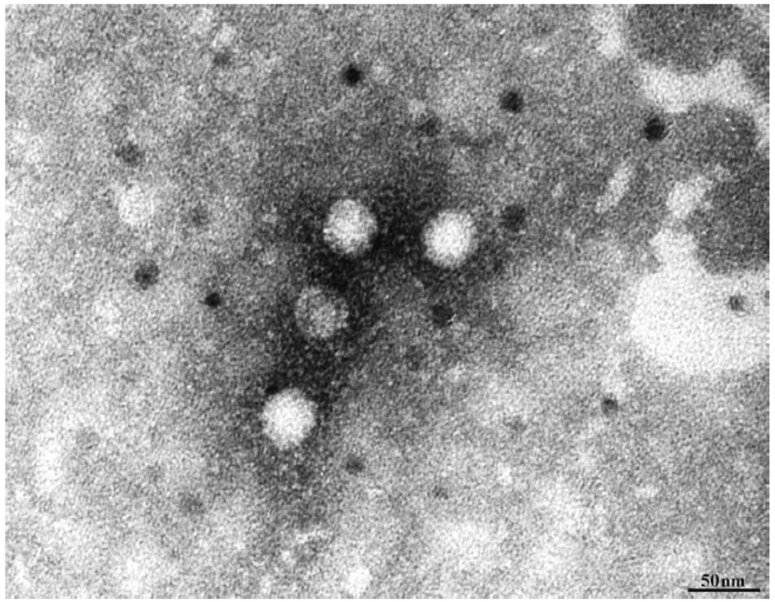


Figure 1 Electron micrograph of virus like particles in the stool of one animal (Table 1). Scale bar indicates 50 nm.

astrovirus sequences in the stool samples from sick rabbits

(Meta)genomics is useful for hypothesis generation but experiments must be done

Astrovirus particles

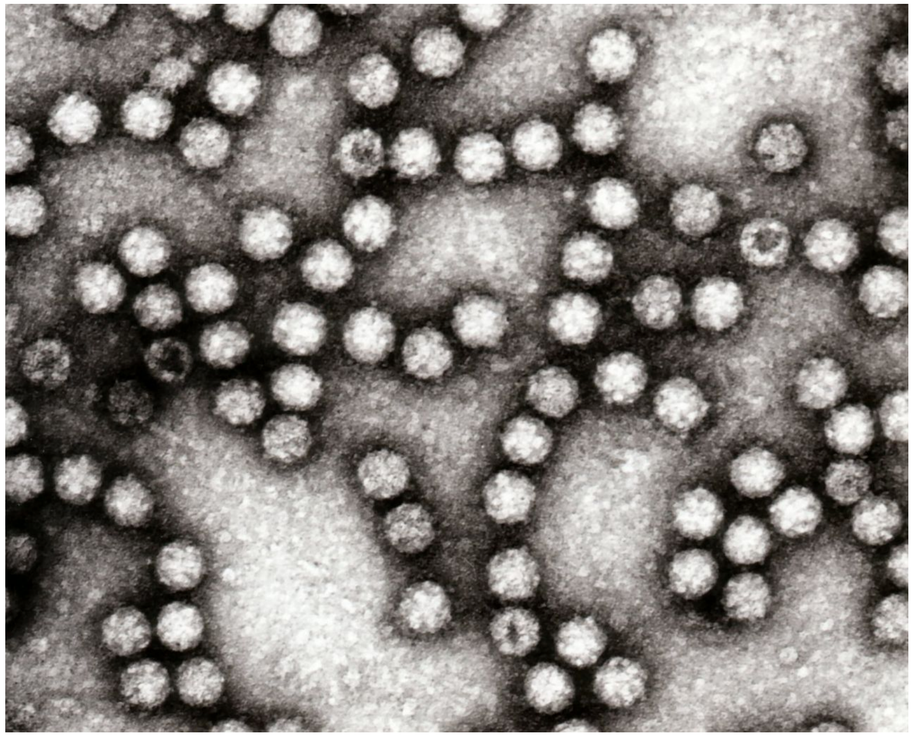


image: wikipedia

JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 1993, p. 955–962
0095-1137/93/040955-08$02.00/0

Vol. 31, No. 4

Characterization and Seroepidemiology of a Type 5 Astrovirus Associated with an Outbreak of Gastroenteritis in Marin County, California

KAREN MIDTHUN,[1]†* HARRY B. GREENBERG,[1]‡ JOHN B. KURTZ,[2] G. WILLIAM GARY,[3] FENG-YING C. LIN,[4] AND ALBERT Z. KAPIKIAN[1]

RESULTS

Volunteer study. Nineteen adult volunteers were orally administered a filtrate prepared from a 0.1% suspension of stool from one of the ill individuals in the original Marin County outbreak. None of 17 volunteers who received a 1-ml inoculum became ill. Because of this, the amount of inoculum was increased to 20 ml. Of two volunteers who received the larger inoculum, one developed a gastrointestinal illness characterized by nausea, vomiting, diarrhea, and malaise.