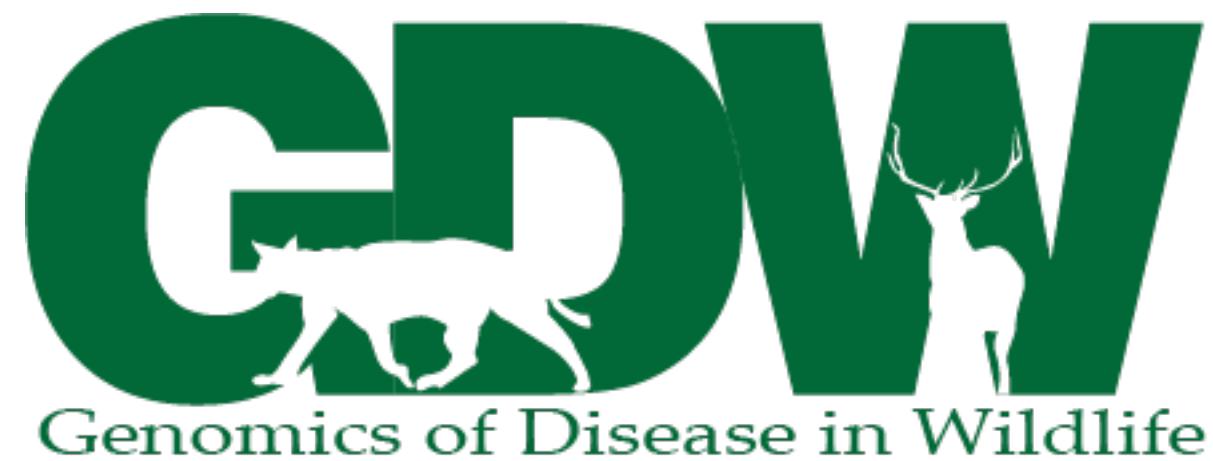


Metagenomics and disease

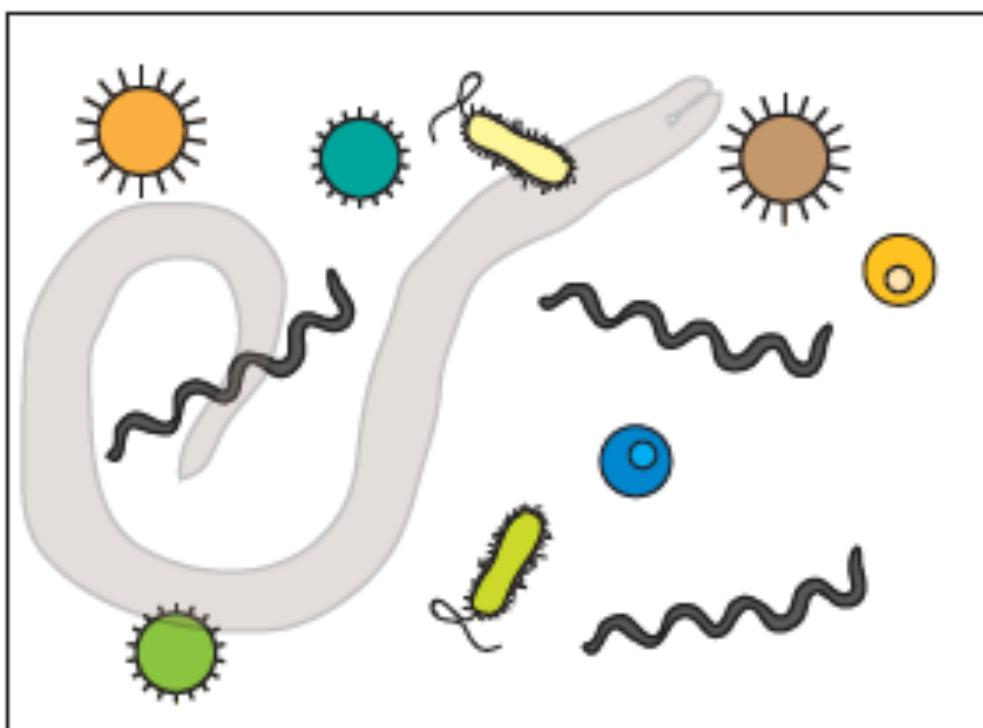
Mark Stenglein, GDW



Metagenomics is the study of >1 genome

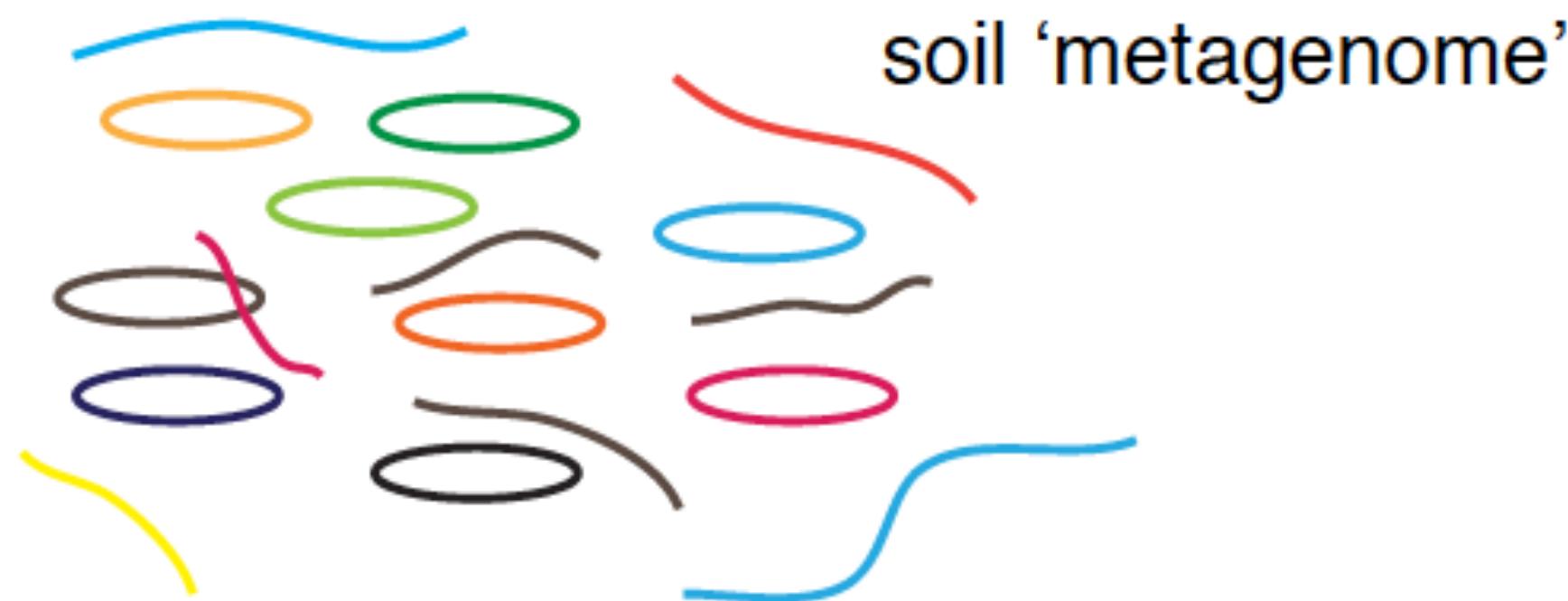
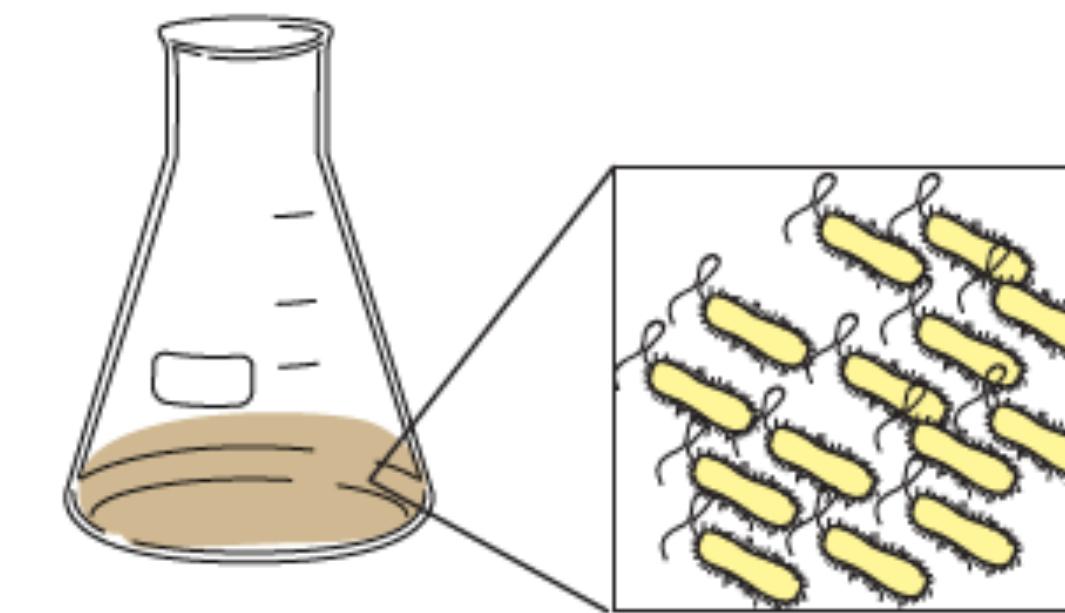
Many genomes

soil community

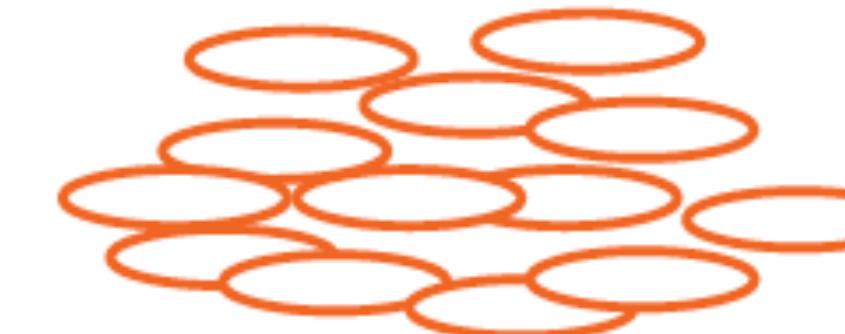


1 genome

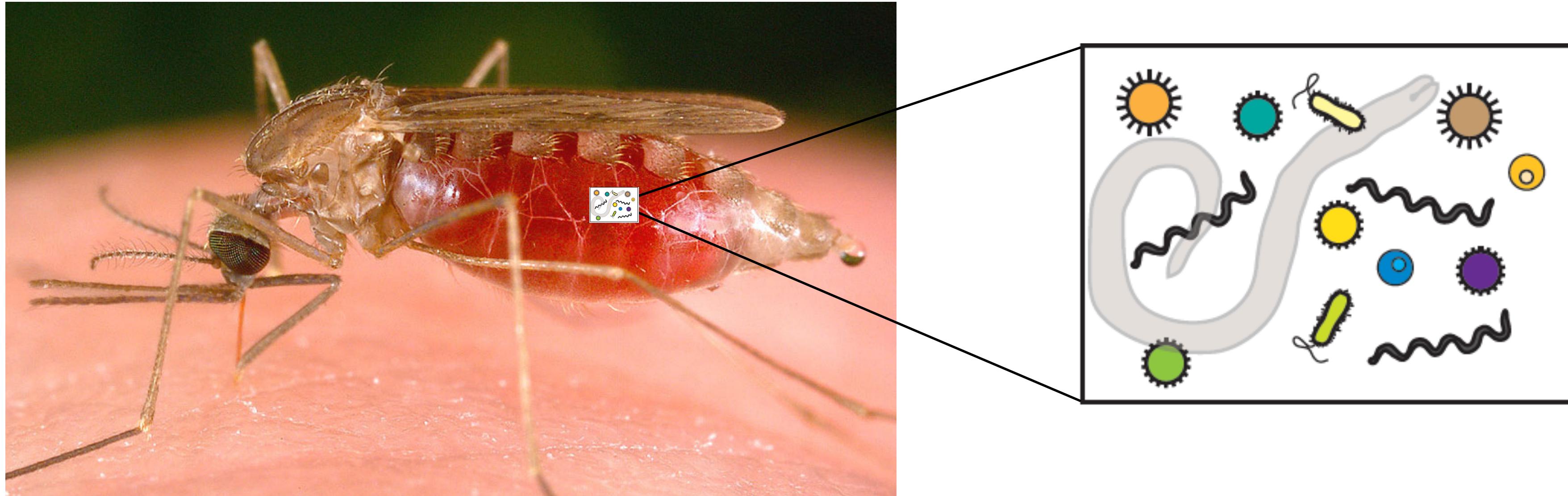
bacterial isolate



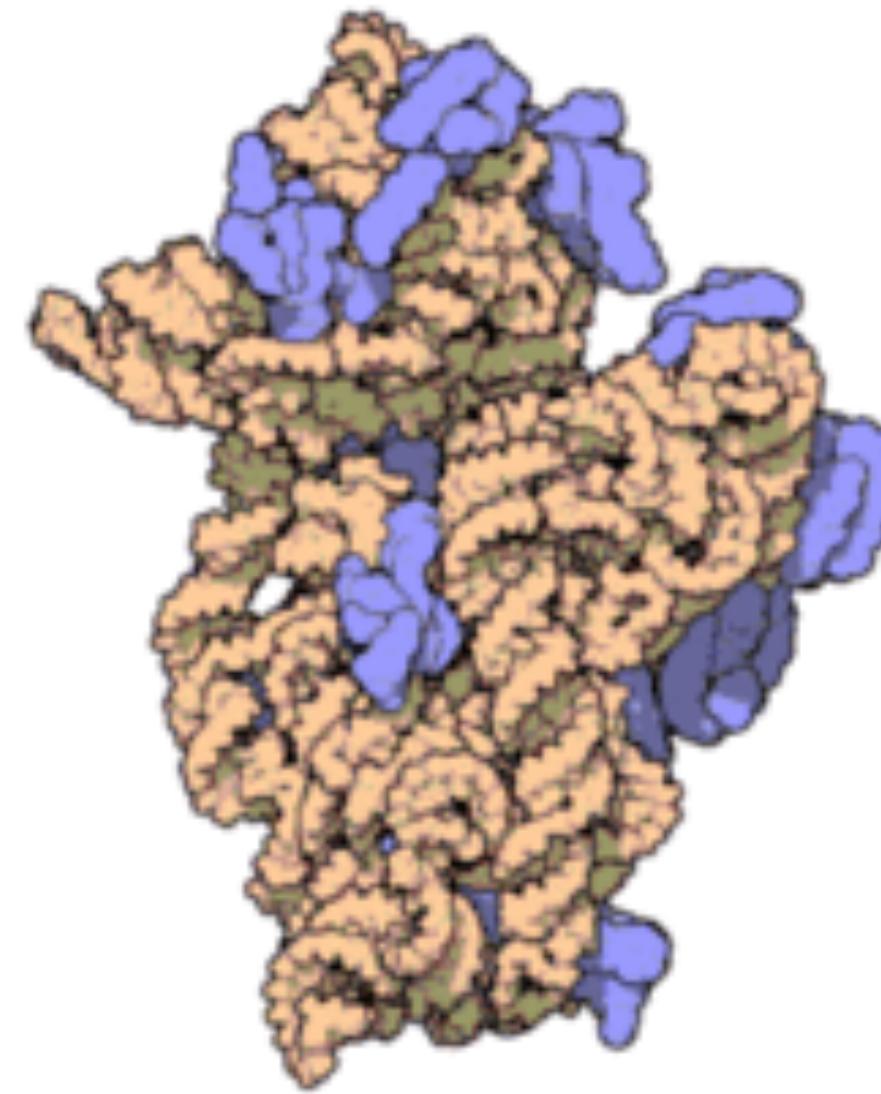
bacterial genome



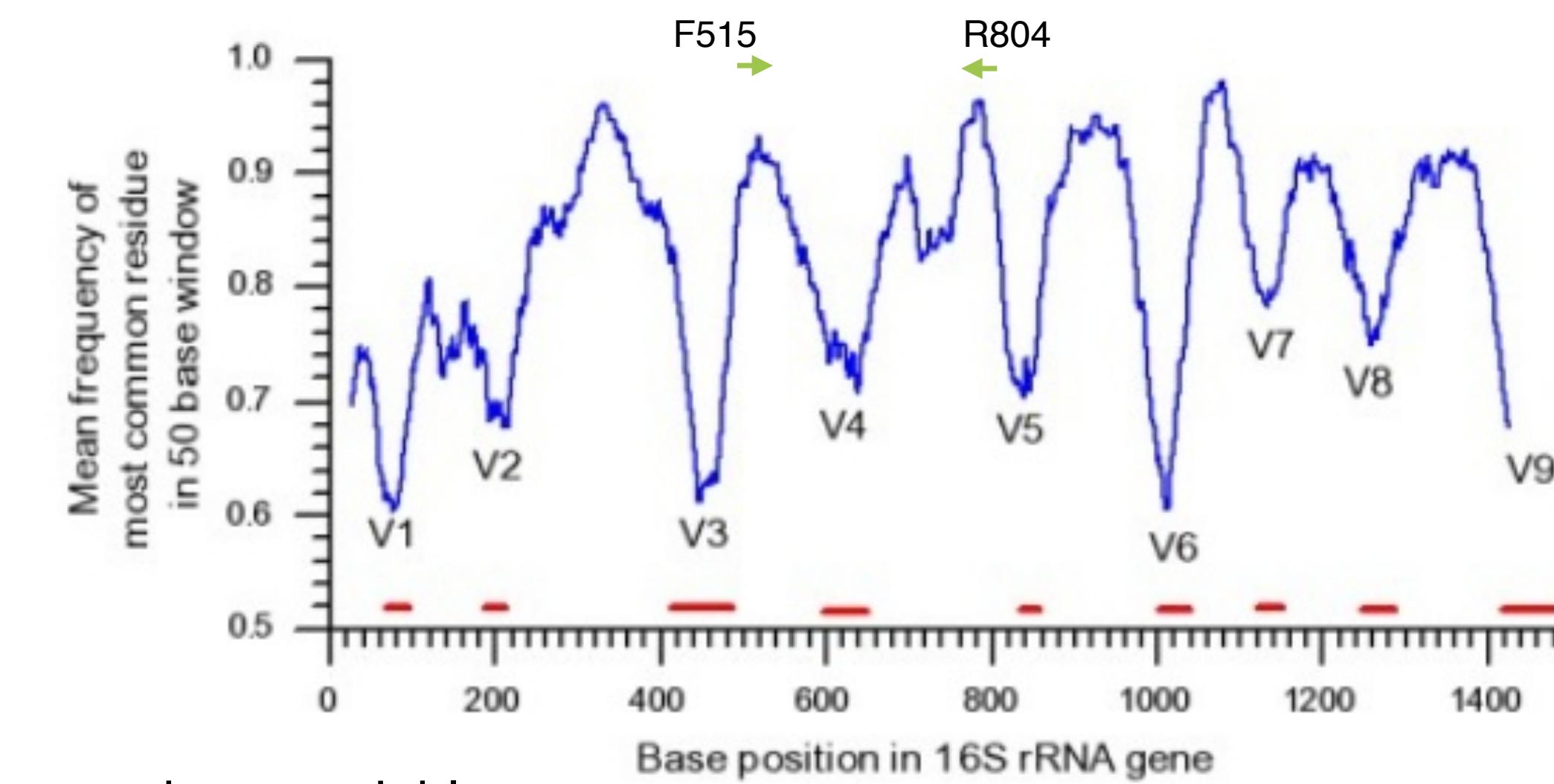
If you sequence total nucleic acid from an intact multicellular organism,
you are doing metagenomics



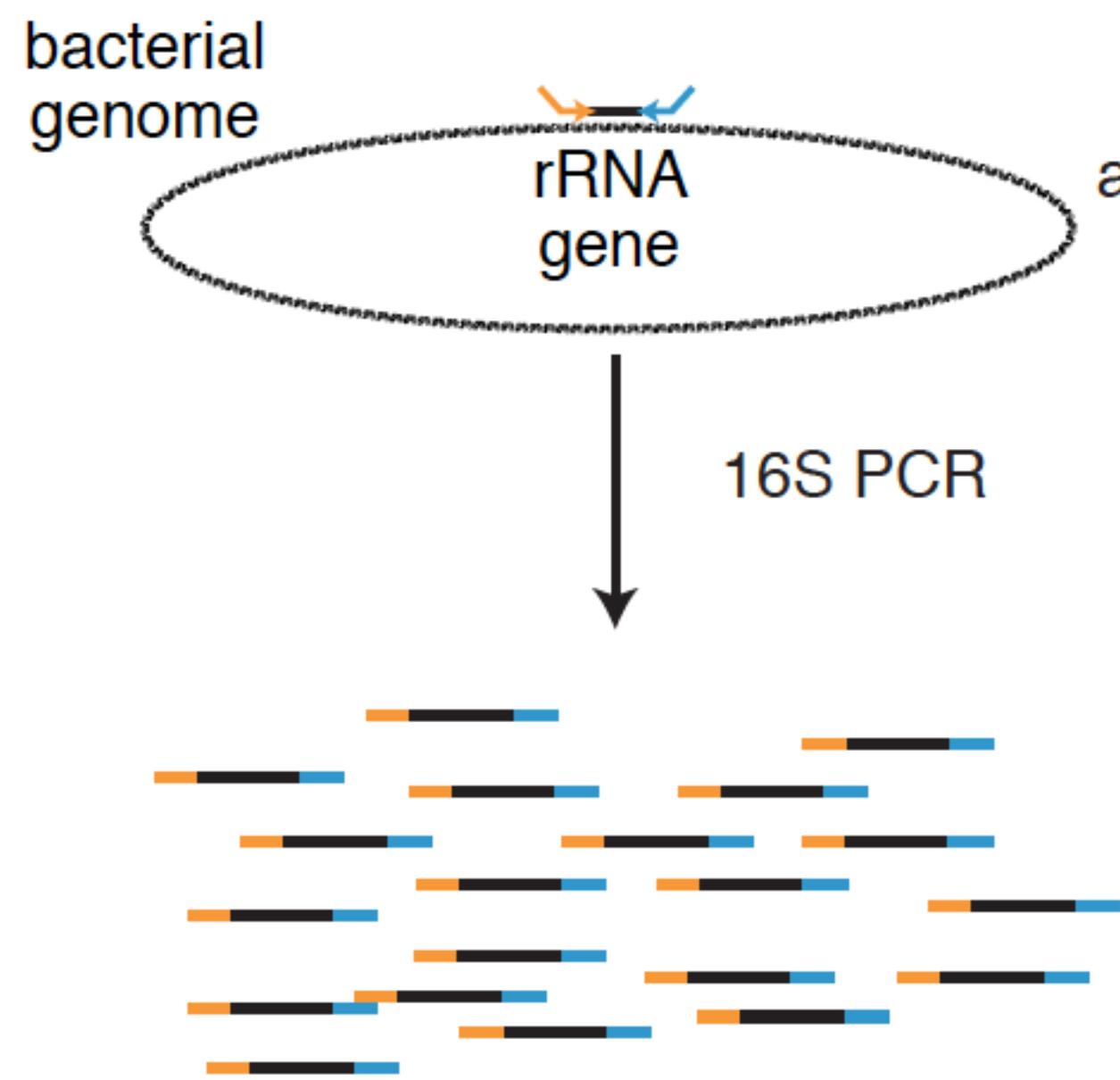
16S sequencing (microbiome sequencing) is a form of metagenomics



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

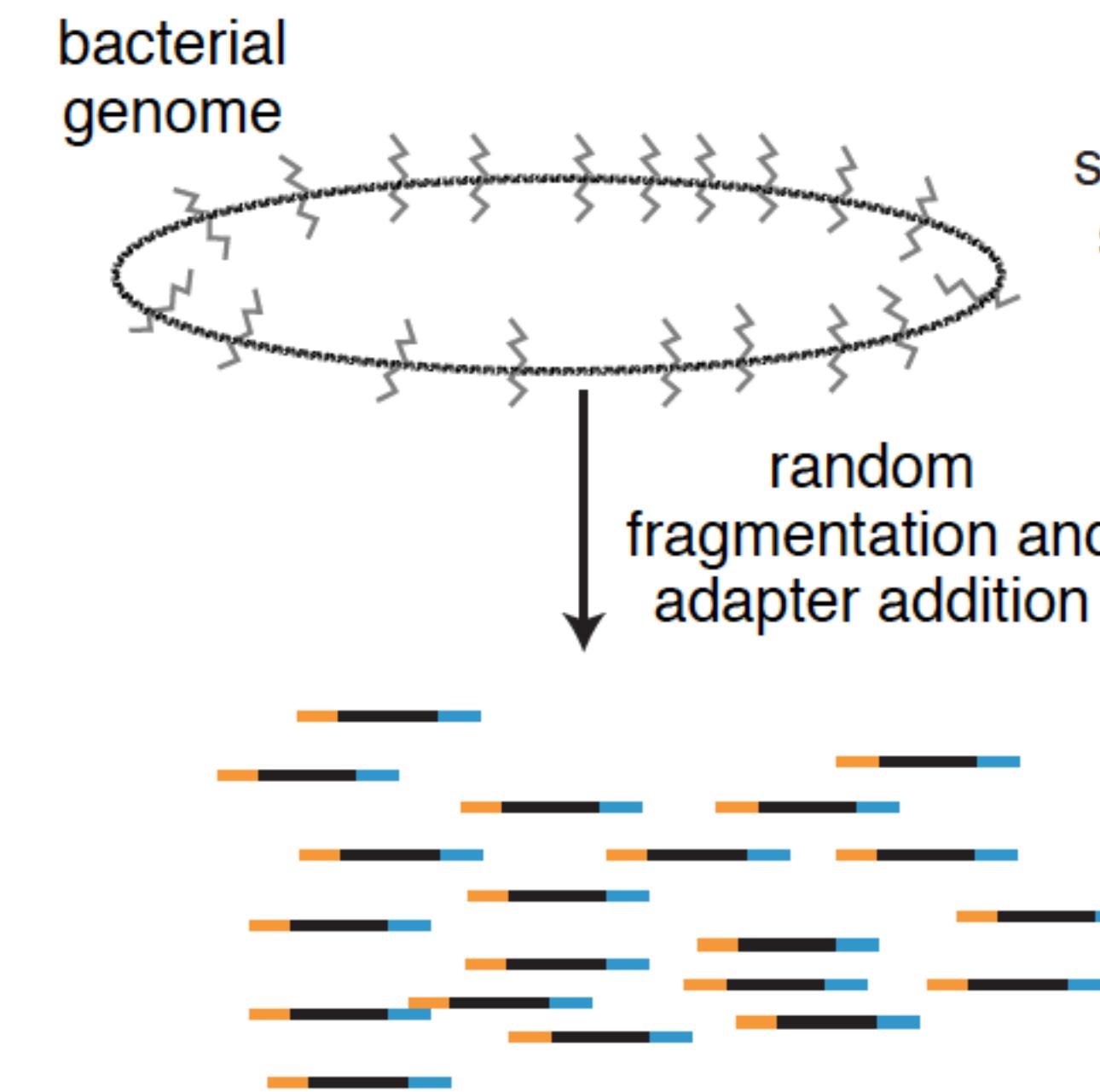


16S sequencing vs. shotgun metagenomics



16S PCR amplifies ~0.01% of a bacterial genome

library molecules contain 16S sequences from one or more genomes



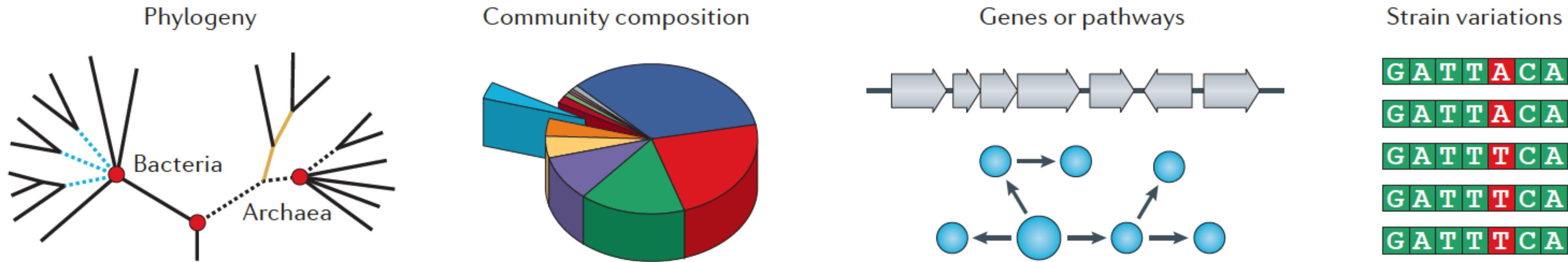
shotgun library molecules contain random bits of the genome

- 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

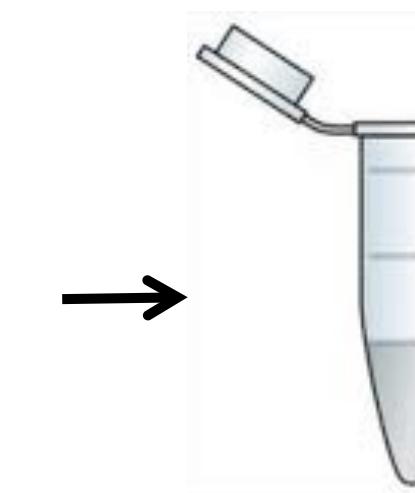
- Community composition analysis (environmental samples, microbiomes, ...)
- Characterization of diet, bloodmeal composition
- Identification of genes of interest: AMR genes, industrially useful enzymes
- Pathogen detection and discovery



Pathogen detection and discovery using metagenomic sequencing



case and control
tissues



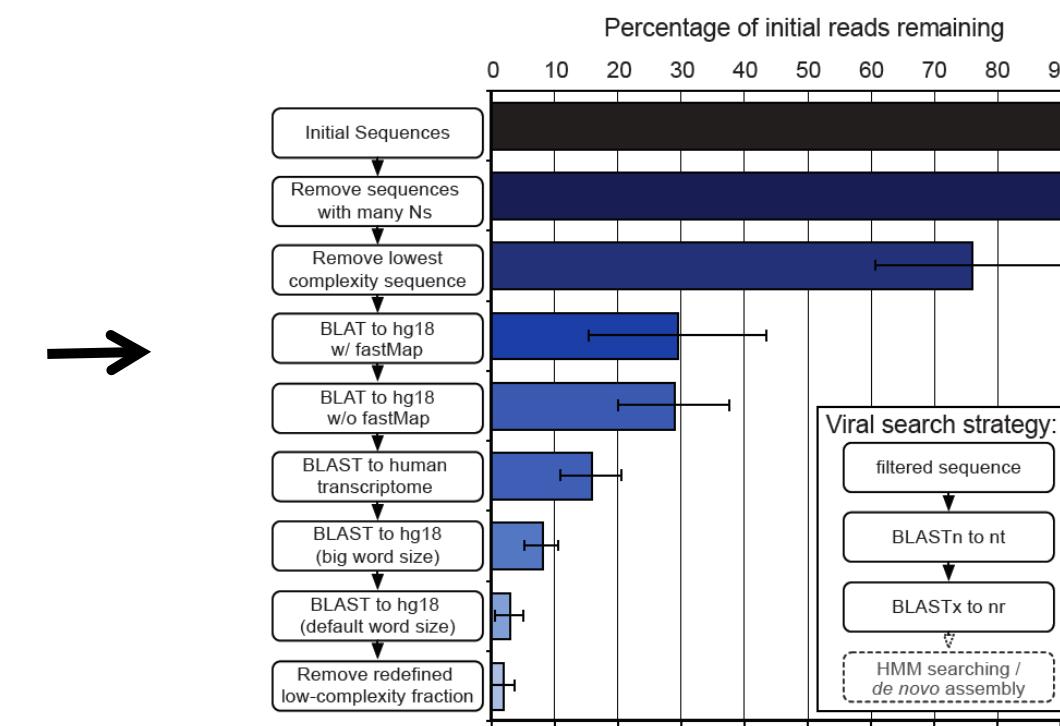
Nucleic acid



Library prep
/ barcode



Illumina
sequencing



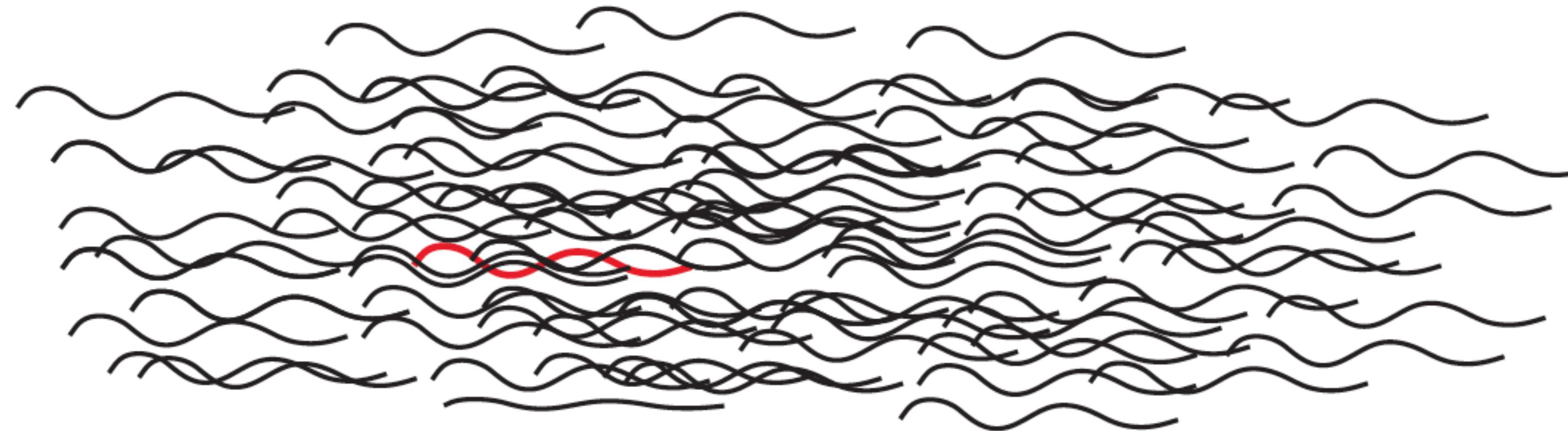
Computational
Analysis



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??TATACT??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 close)



Identification of this pathogen completely consistent with histopathology

case had been tested for *ovine herpesvirus 2*



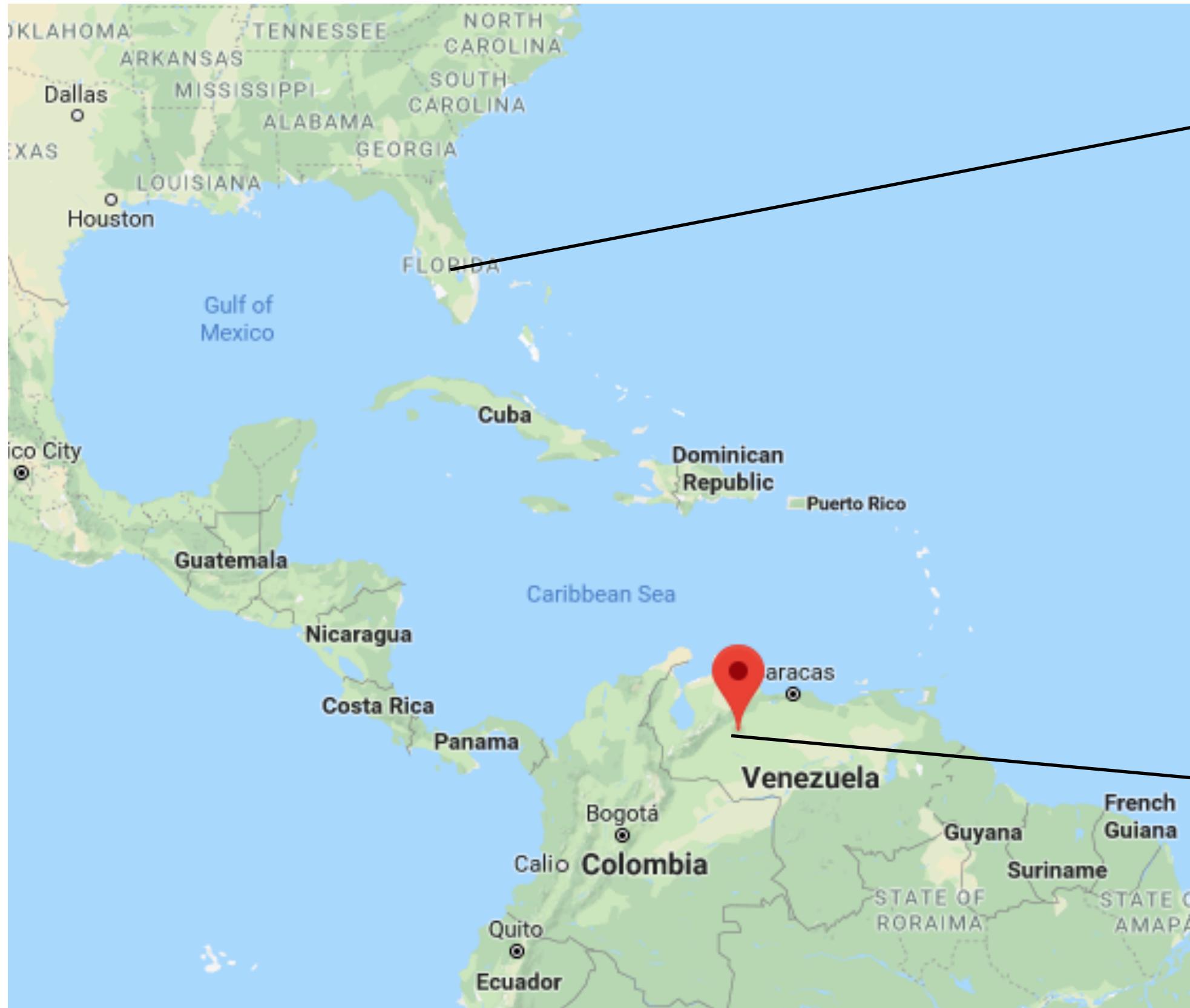
A single read pair aligning to **caprine herpesvirus-2** amongst ~0.5M mule deer reads

PCR is generally more sensitive than NGS for targeted pathogen detection

Laura Hoon-Hanks, DVM

Samples from: Karen Fox DVM, CO Parks & Wildlife

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

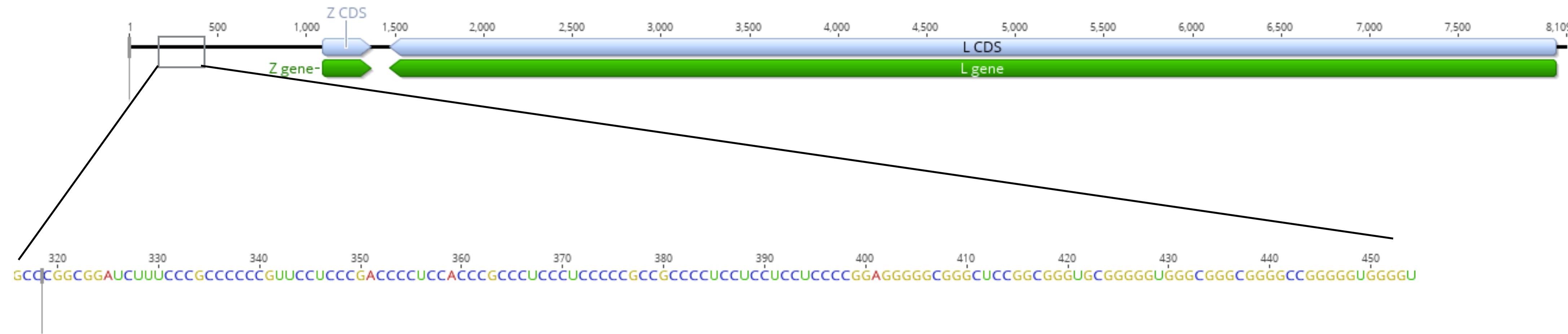
```
>NODE_274_length_640_cov_11681_ID_547
GCGGGGGTGGCGGGCGGGCCGGGGTGGGTCGGCGGGGACCGTCCCCGACCGGCGACCGGCCGCGCCGGC
GCATTTCCACCGCGGCGGTGCGCCGCGACCGGCTCCGGACGGCTGGGAAGGCCC GGCGGGAAAGGTGGCTCGGGGG
GCCCGTCCGCCCGCTCTCCCCCGCCCGTCCTCCCCCGGGAGGGCGCGGGTCGGGCGGC GGCGGTGGC
GGCGGGACCACCCCCCGAGTGTTACAGCCCCCGGCAGCAGCACTGCCGAATCCCGGGCCGAGGGAGCGAGACCC
GTCGCCGCGCTCTCCCCCTCCCGGCCACCCCCCGCGGGGCCCGGGGTCCCCCGCGGGGCGCGC
CCGGCGGTCTCGTGGGGGCCACCCCTCCCACGGCGCGACCGCTCTCCCACCCCCCTCCCCGCACCCCCGGC
GACGGGGCCCGCGCGGGTGGGGCGGGCGGACTGTCCCCAGTGCGCCCCGGCGGTGCGCCGTCGGGCCGG
GGGTTCTCTCGGGGCCACGCGCGTCCCTCGAAGAGGGGACGGCGAGCGAGCGCACGGGTGGCGCGATGT
CGGCTACCCACCCGACCGTCTTG
```

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
<input type="checkbox"/>	Guanarito mammarenavirus isolate CVH-960201 segment L, complete sequence	1081	1081	94%	0.0	99%	KU746283.1
<input type="checkbox"/>	Guanarito mammarenavirus isolate CVH-950801 segment S, complete sequence	1064	1064	90%	0.0	99%	KU746280.1
<input type="checkbox"/>	Chimpanzee 28S ribosomal RNA gene fragment	826	826	100%	0.0	89%	M30950.1
<input type="checkbox"/>	Gorilla 28S ribosomal RNA gene fragment	817	817	99%	0.0	89%	M30951.1
<input type="checkbox"/>	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
<input type="checkbox"/>	Homo sapiens clone BAC JH1 genomic sequence	808	1612	100%	0.0	89%	MF164269.1
<input type="checkbox"/>	Homo sapiens RNA, 45S pre-ribosomal N2 (RNA45SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146144.1
<input type="checkbox"/>	Homo sapiens RNA, 28S ribosomal N2 (RNA28SN2), ribosomal RNA	804	804	100%	0.0	89%	NR_146148.1
<input type="checkbox"/>	Human DNA sequence from clone CH507-146P16 on chromosome 21, complete sequence	804	804	100%	0.0	89%	CT476837.18
<input type="checkbox"/>	Human ribosomal DNA complete repeating unit	798	798	100%	0.0	89%	U13369.1
<input type="checkbox"/>	Homo sapiens clone BAC JH5 genomic sequence	787	787	100%	0.0	88%	MF164266.1
<input type="checkbox"/>	Homo sapiens RNA, 28S ribosomal N3 (RNA28SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146154.1
<input type="checkbox"/>	Homo sapiens RNA, 45S pre-ribosomal N3 (RNA45SN3), ribosomal RNA	784	784	100%	0.0	88%	NR_146151.1
<input type="checkbox"/>	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

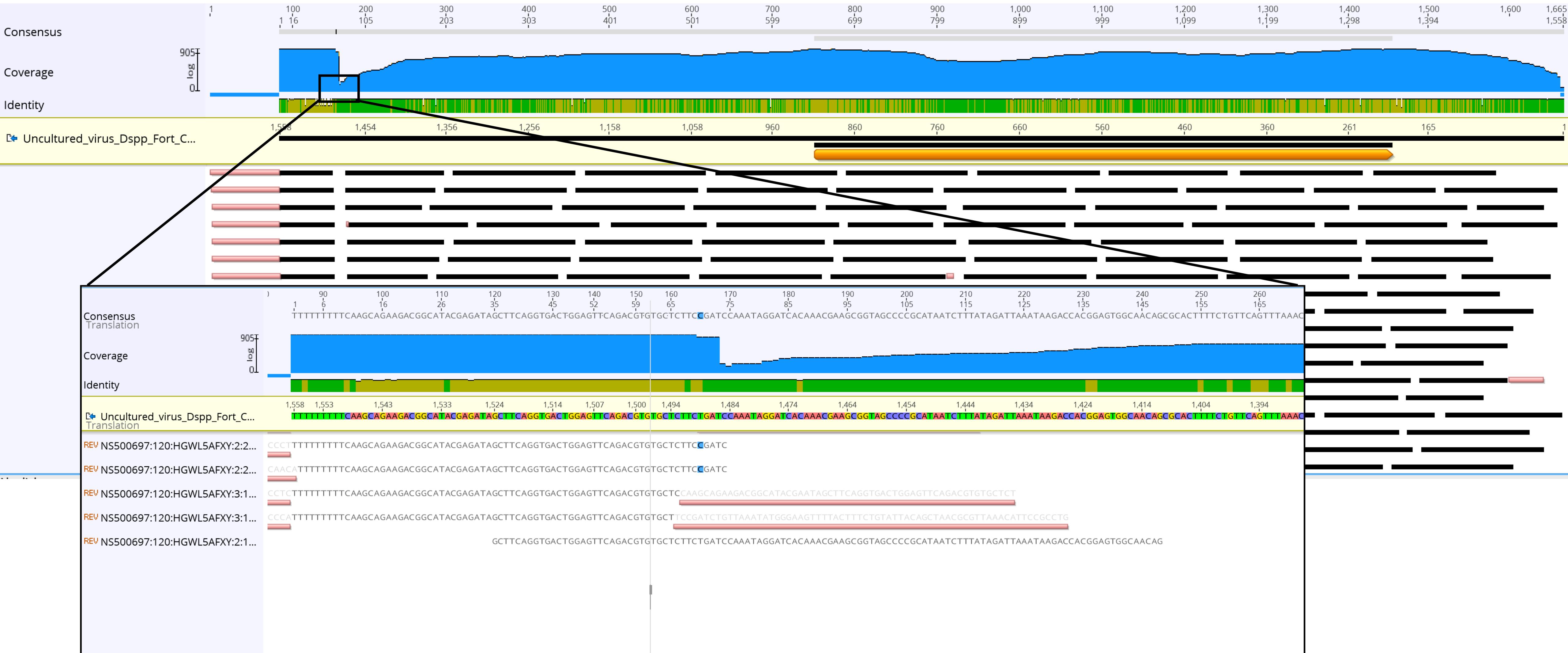


Conclusion: don't blindly trust database annotation nor the output of analysis software

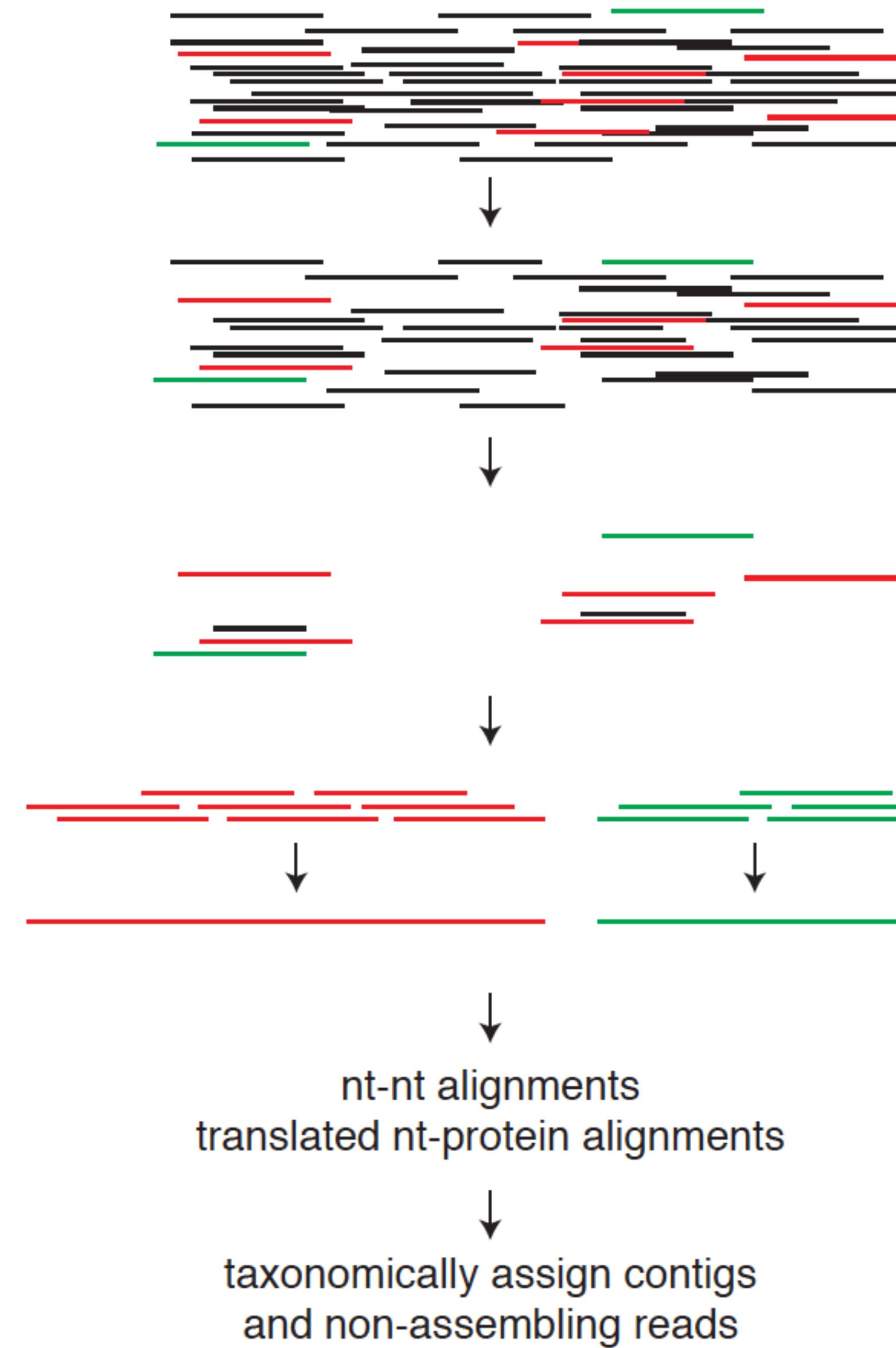
COMMENT
GenBank Accession Numbers KU746283, KU746284 represent sequences from the 2 segments of Guanarito mammarenavirus CVH-960201.
##Genome-Assembly-Data-START##
Assembly Method :: Trimmomatic v. 0.32
SGA v. 0.10.13
iMetAMOS v. 1.5
samtools v1.1
FastQC v. 0.10.0
Spades v. 3.1.1
idba v1.1.1
Pilon v. 1.8
Quast v. 2.2
Prokka v. 1.7
Assembly Name :: GT0V014-SEQ-1-ASM-1
Genome Coverage :: 6779.96x
Sequencing Technology :: Illumina HiSeq1500
##Genome-Assembly-Data-END##.

FEATURES
source
Location/Qualifiers
1..8109
/organism="Guanarito mammarenavirus"
/mol_type="genomic RNA"

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



A typical pathogen discovery analysis workflow



Quality filter /
remove PCR duplicates
[cutadapt](#)
[cd-hit-est](#)

Remove host sequences
[bowtie2](#)

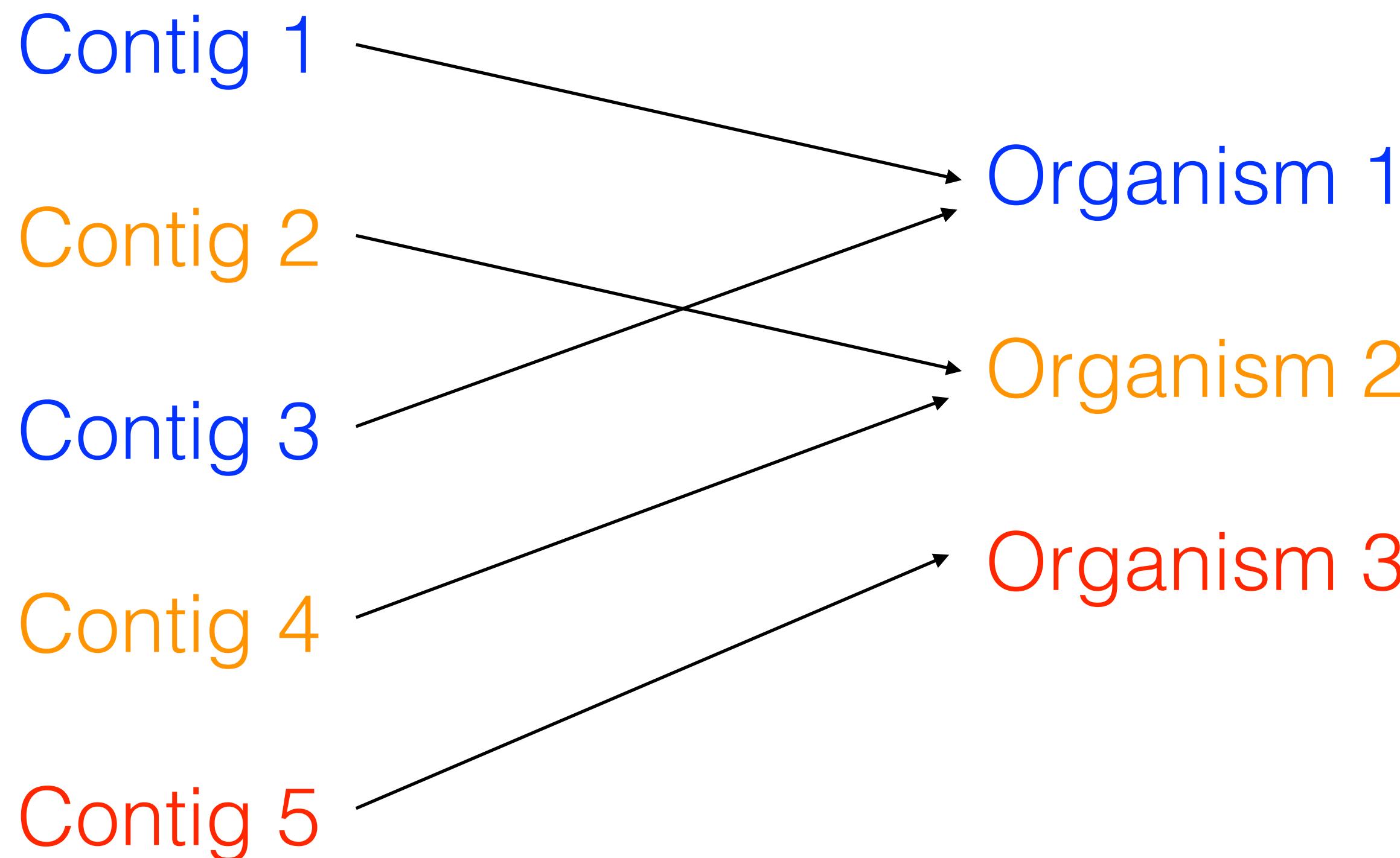
Assemble remaining reads into
contigs
[SPAdes](#)

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

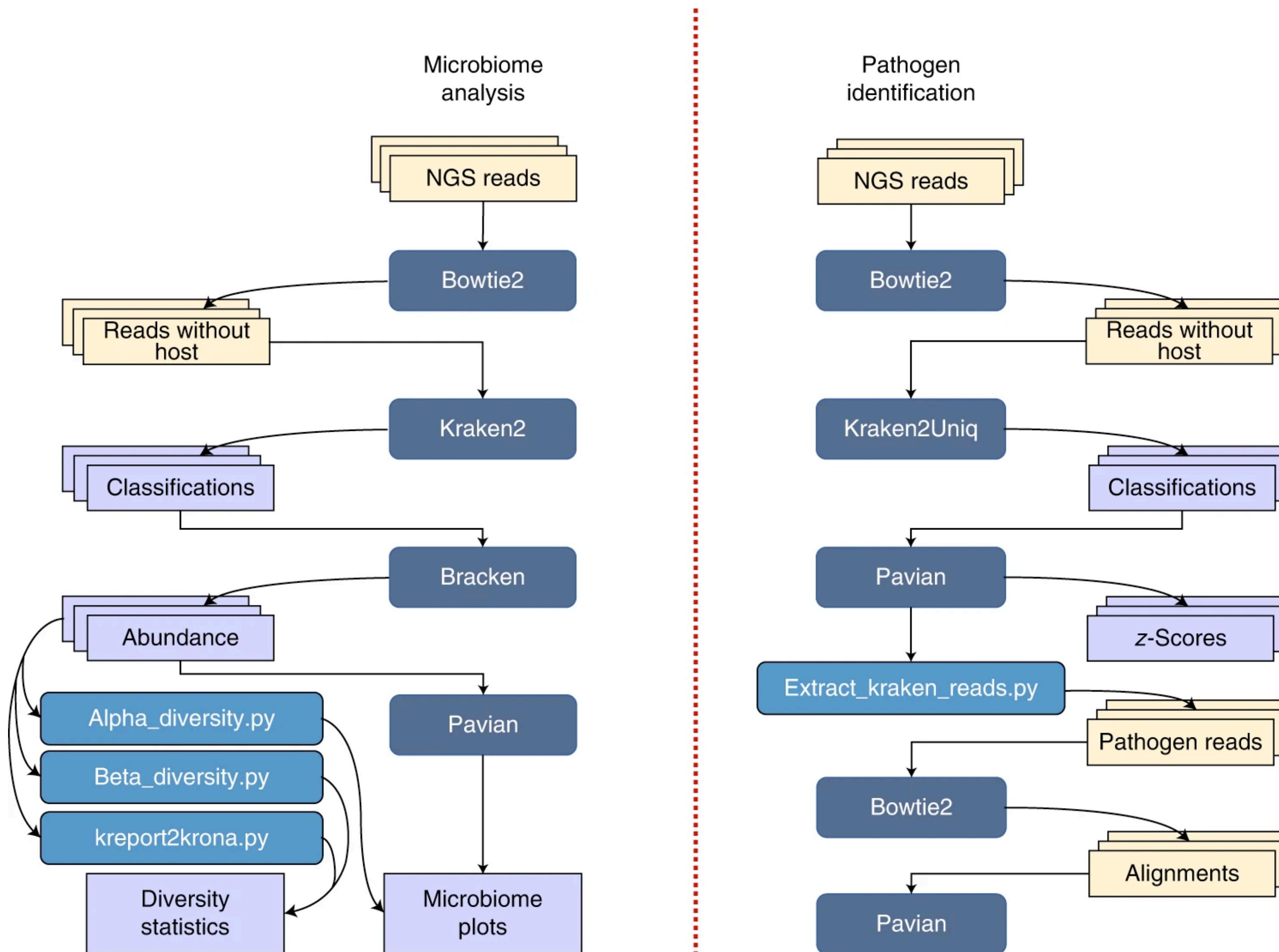
The goal of metagenomic classification software is to map sequence information to taxonomic information.



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. Miller¹

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

The screenshot shows a web browser window for <https://czid.org>. The page features a dark header with the CZ ID logo and navigation links for Impact, Resources, and Sign in. A prominent blue banner at the top announces a new feature for detecting antimicrobial resistance genes. Below the banner, the main heading reads "Real-time Pathogen Detection" and describes Chan Zuckerberg ID as a free, cloud-based metagenomics platform for researchers. A registration form is displayed, asking for an email address and a "Register Now" button.

Most Visited: blastn, blastx, blastp, NCBI Tax, Genbank, SRA, PubMed, FoCo W, ggplot theme, Primer3, Nextflow :: Anacond..., CC Search, RC, Nanopore, Illumina, Other Bookmarks

! New! CZ ID users can now detect and analyze antimicrobial resistance genes in sequencing data. [Learn More.](#)

cZ ID

Impact Resources [Sign in](#)

Real-time Pathogen Detection

Chan Zuckerberg ID: The free, cloud-based metagenomics platform for researchers

Your email address [Register Now >](#)



The snake sample we analyzed run through this tool

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

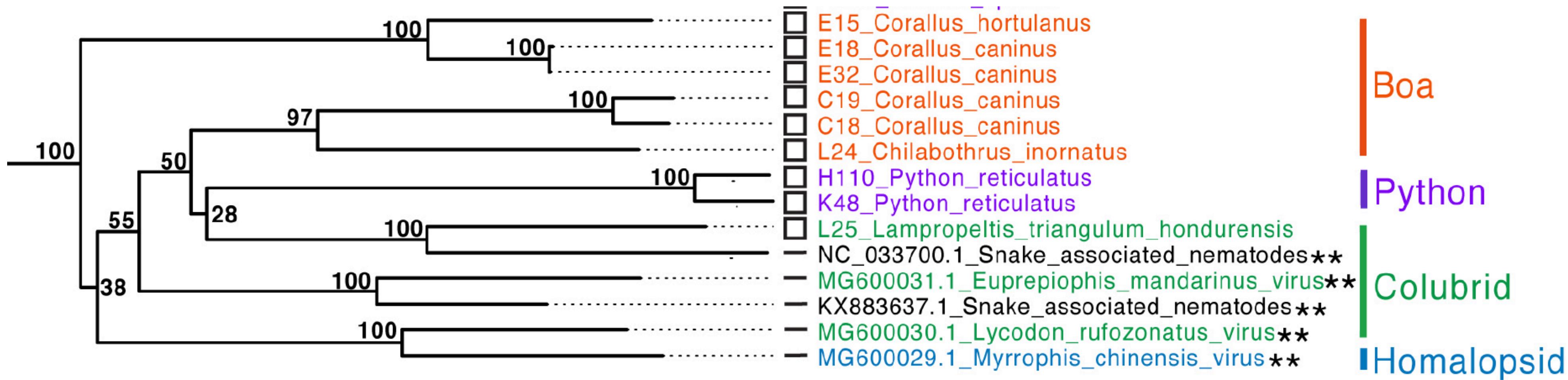
The screenshot shows a web browser window with the URL https://github.com/stnglein-lab/taxonomy_pipeline in the address bar. The page title is "README.md". The main content is a large heading "Stnglein lab taxonomic assessment pipeline" followed by a descriptive paragraph: "This is a nextflow implementation of the pipeline used in the [Stnglein lab](#) to taxonomically classify sequences in NGS datasets. It is mainly designed to identify virus sequences in a metagenomic dataset but it also performs a general taxonomic classification of sequences. A [previous bash-based version of this pipeline](#) has been reported in a number of [publications](#) from our lab." Below this, there is a section titled "How to run the pipeline".

- Sharing pipelines on places like GitHub facilitates reproducibility

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

Metagenomic sequencing only gives you sequences

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



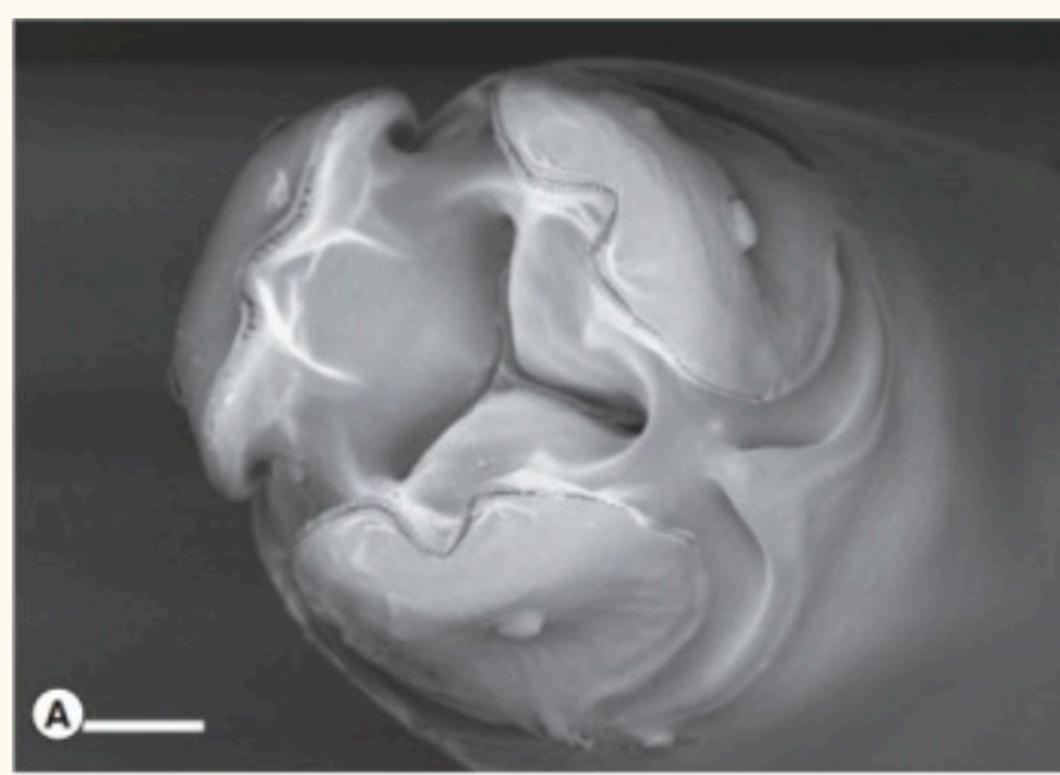
Are these related viruses infecting both nematodes and snakes?

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

