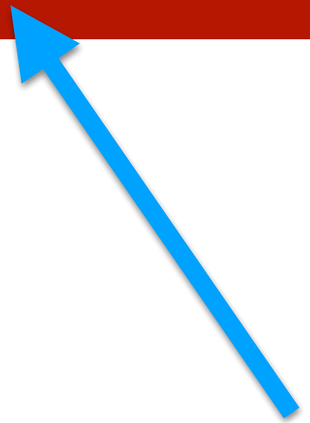
Metagenomics and Viromics



Scott A. Handley, PhD
Associate Professor
Washington University School of Medicine
Department of Pathology and Immunology &
The Edison Family Center for Genome Sciences & Systems Biology

Metagenomics and Viromics

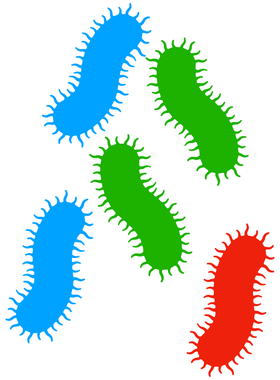
Metagenomics is just one of two popular ways to study a microbiome

The study of the collection of viruses in a sample (e.g. soil, stool, pond water)

So what is a microbiome?

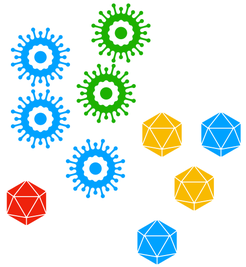
What is a microbiome

A collection of all microorganisms within a given environment



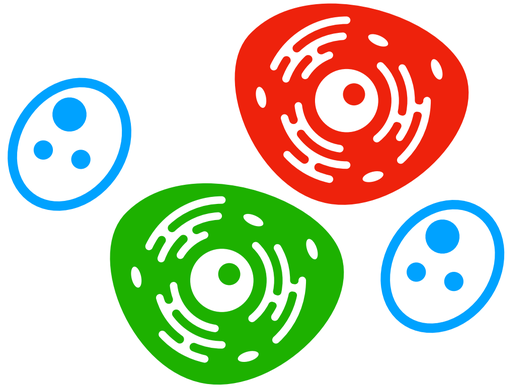
Bacteria and Archaea

(This is what people typical think about with microbiome studies)



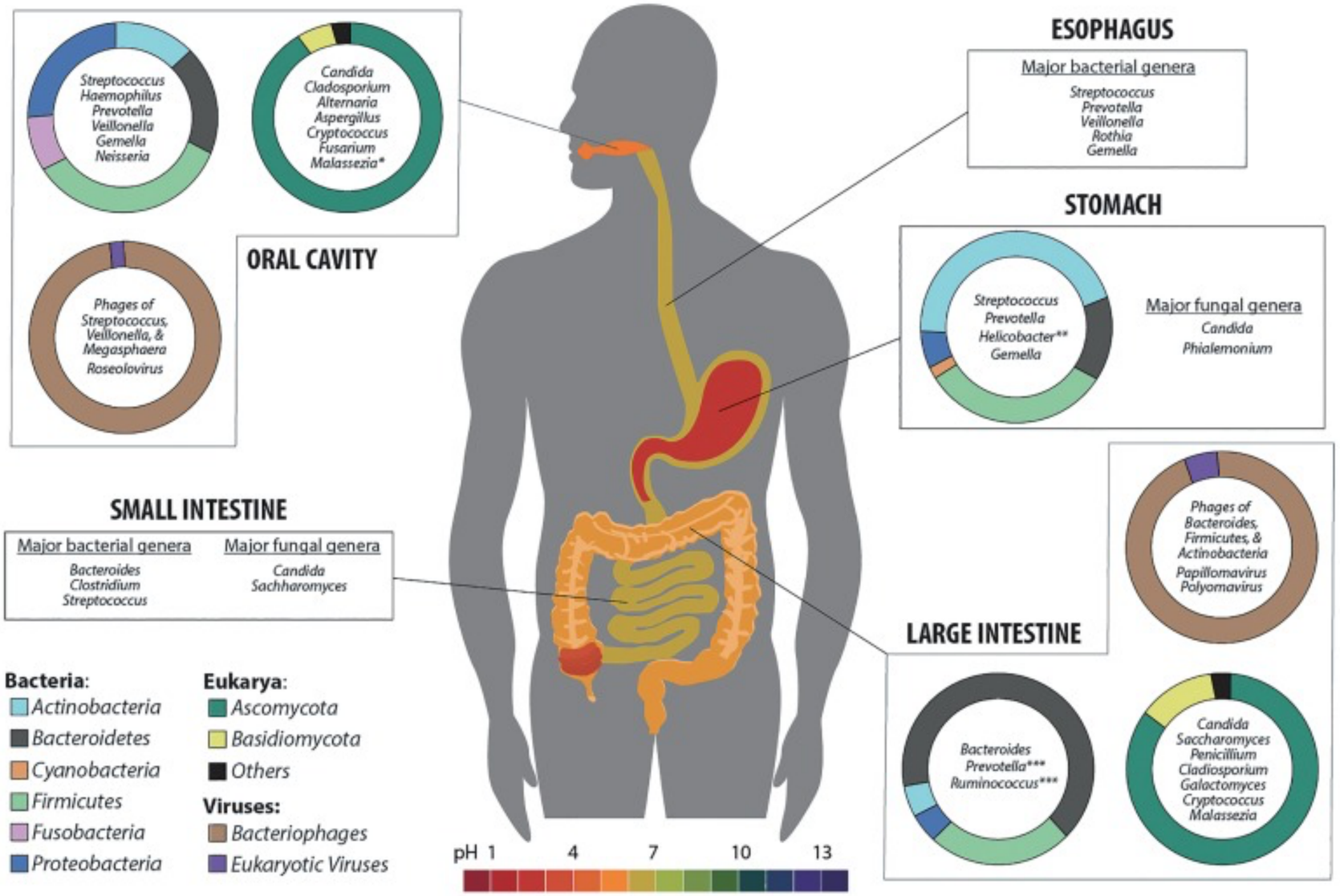
Viruses

- Bacteriophage
  - Bacteria and Archaea
- Eukaryotic viruses
  - Plant viruses
  - Vertebrate viruses



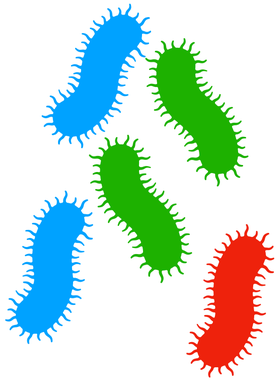
Fungi / Microeukaryotes

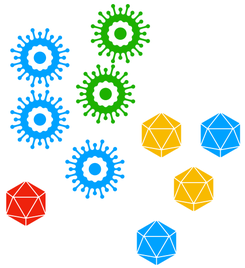
Human Associated Microbiomes

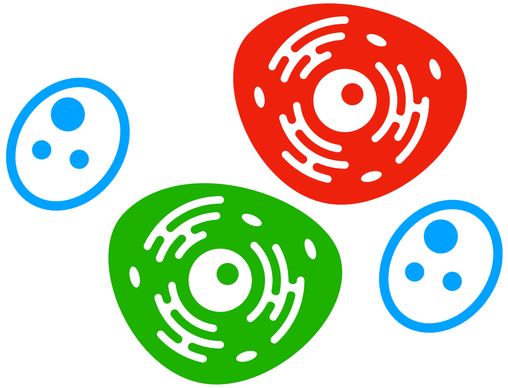


Tillman, E.T. et al. Microbes Environ. 2017

How do we study these microbial groups?







Bacteria and Archaea

- Amplicon (16S rRNA) surveys
- Metagenomics (aka shotgun approaches)

Viruses

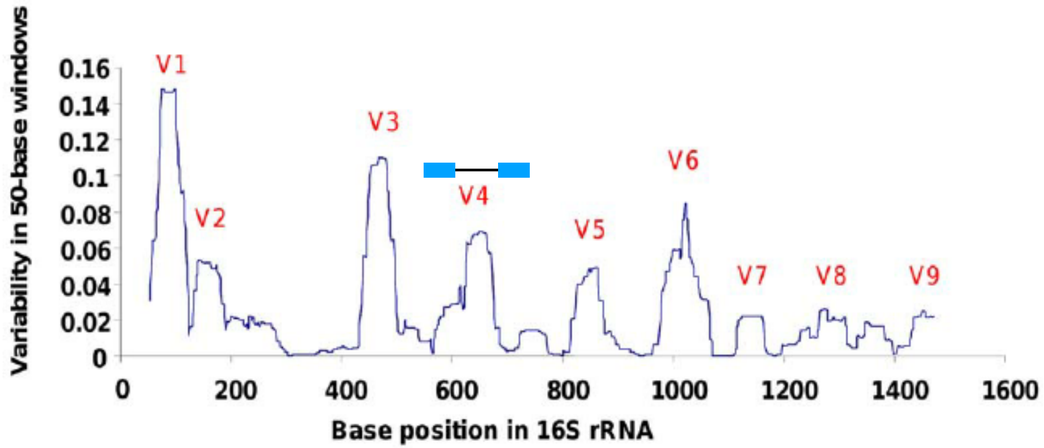
- Metagenomics (aka shotgun approaches)

Fungi and Microeukaryotes

- Amplicon (ITS/18S) surveys
- Metagenomics (aka shotgun approaches)

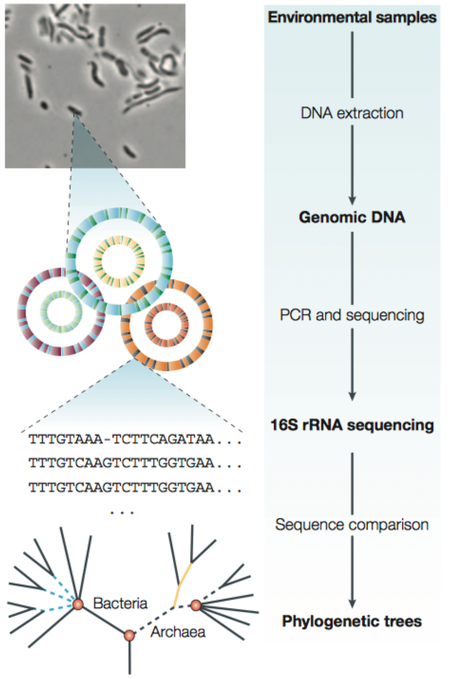
Just Shotgun Everything?

Amplicon Based Approaches



- Primers bind to conserved regions that flank a variable region
- Primers are barcoded (1 barcode per sample) for parallel multiplexing

16S rRNA Amplicon Survey



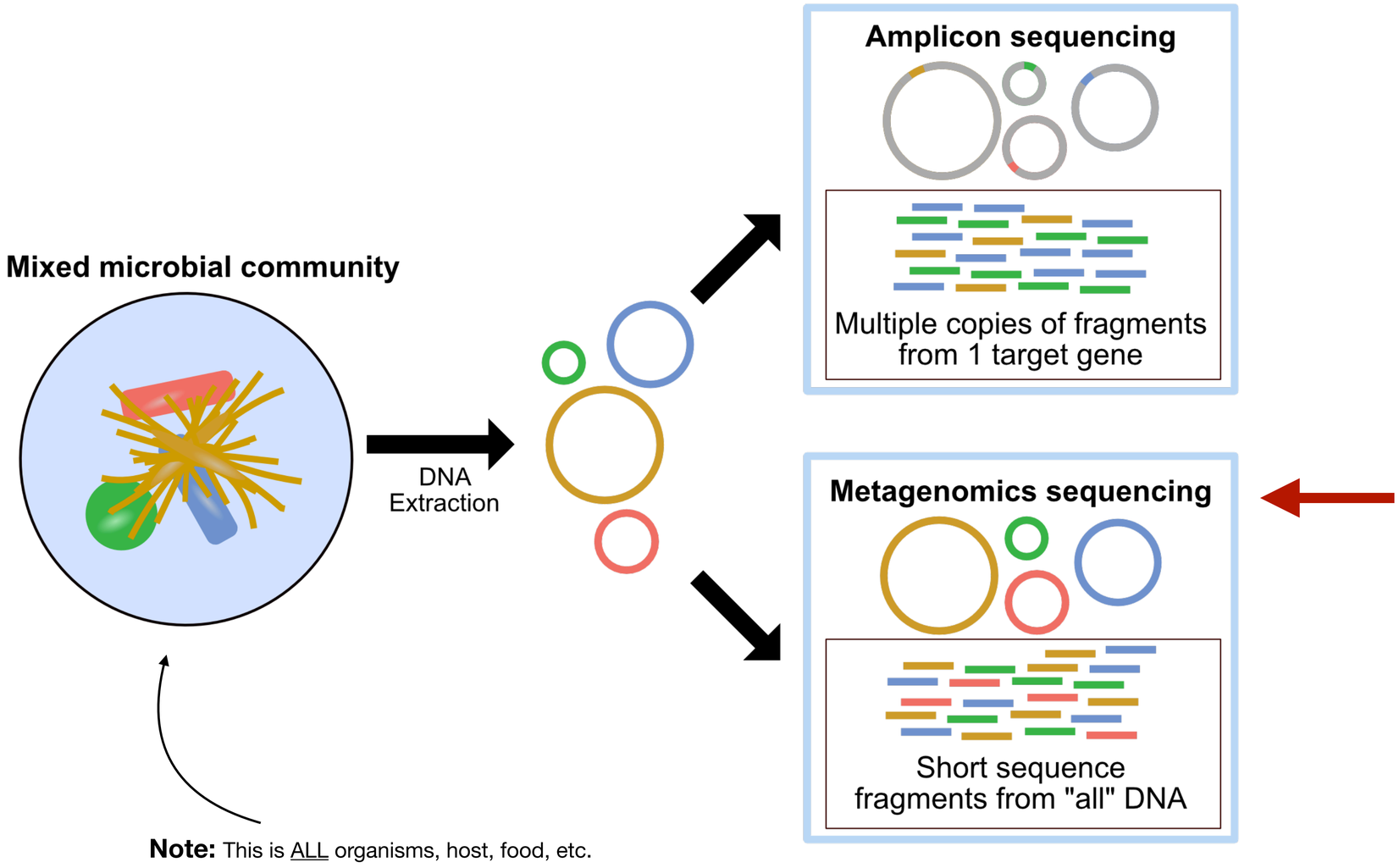
Study Design

Laboratory

Bioinformatics, Ecological Analysis and Statistics

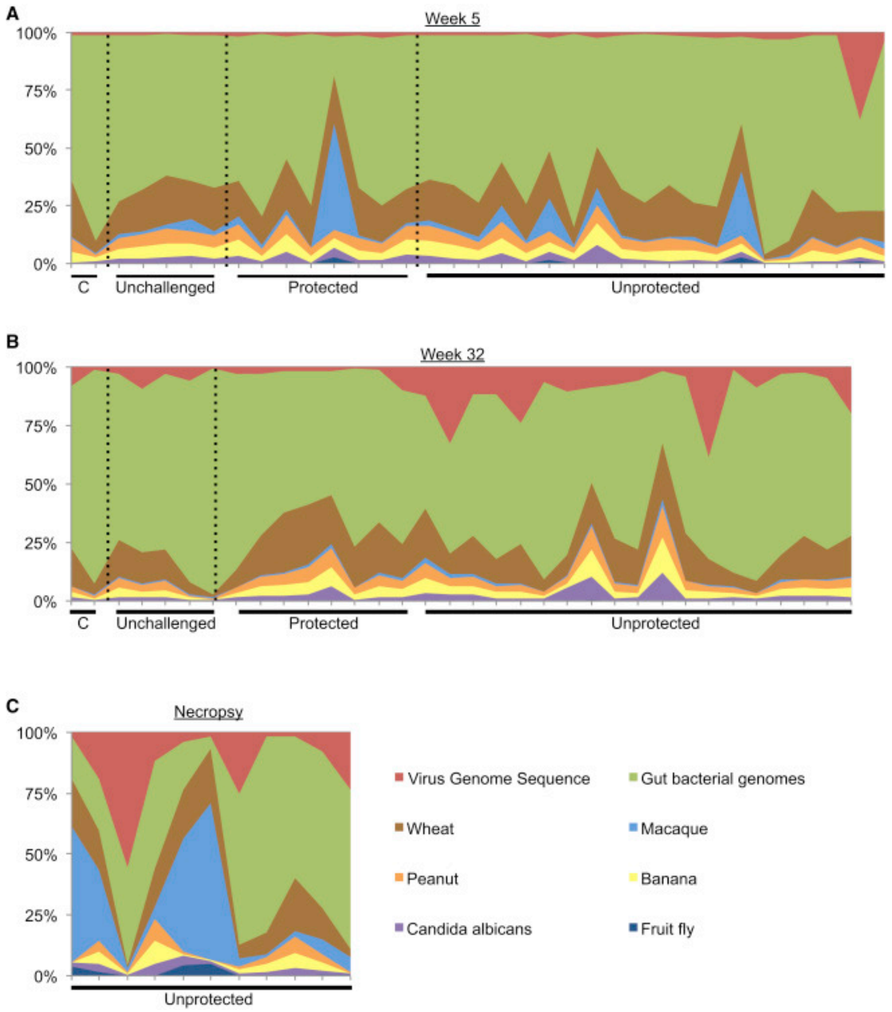
Tringe, S.G., Rubin, E.M. Nat Rev Genet. 2005 Nov;6(11):805-14

Metagenomic Approach



Happy Belly Bioinformatics (https://astrobiomike.github.io/misc/amplicon_and_metagen)

Metagenomics Sequences Everything



Handley, S.A. Cell Host & Microbe. 2016

Practical Considerations*

Metagenomic Approaches

Challenges

- Nucleic acid concentration/integrity
    - Low concentration/degraded samples can be more challenging to consistently build metagenomic libraries
- Some unique bioinformatics challenges

Advantages

- Sampling from the whole genome, so functional analysis is possible
- Better taxonomic resolution due to gathering a better array of phylogenetically informative characters
- Assembly is also a possibility
- Can capture trans-kingdom data

Amplicon Based Approaches

Challenges

- Assembly is **not** a possibility
- Function only by inference
- Low taxonomic resolution

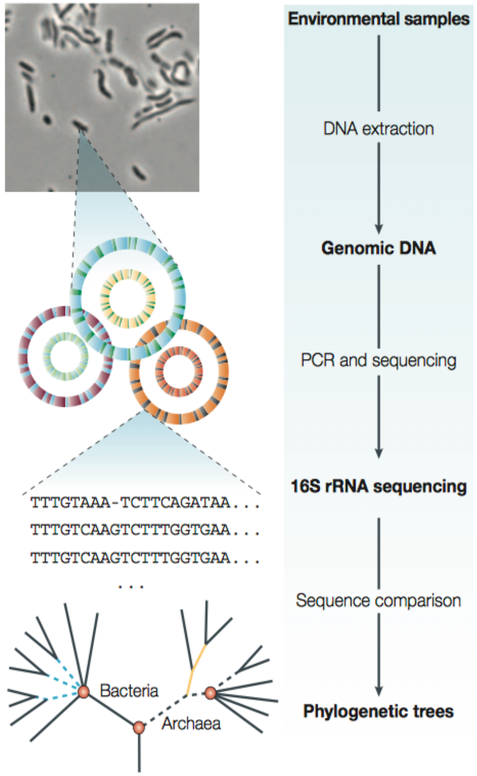
Advantages

- Can be done on samples with extremely low concentrations or degraded samples
- Bioinformatics is more evolved (kind of)
- Curated taxonomic reference databases

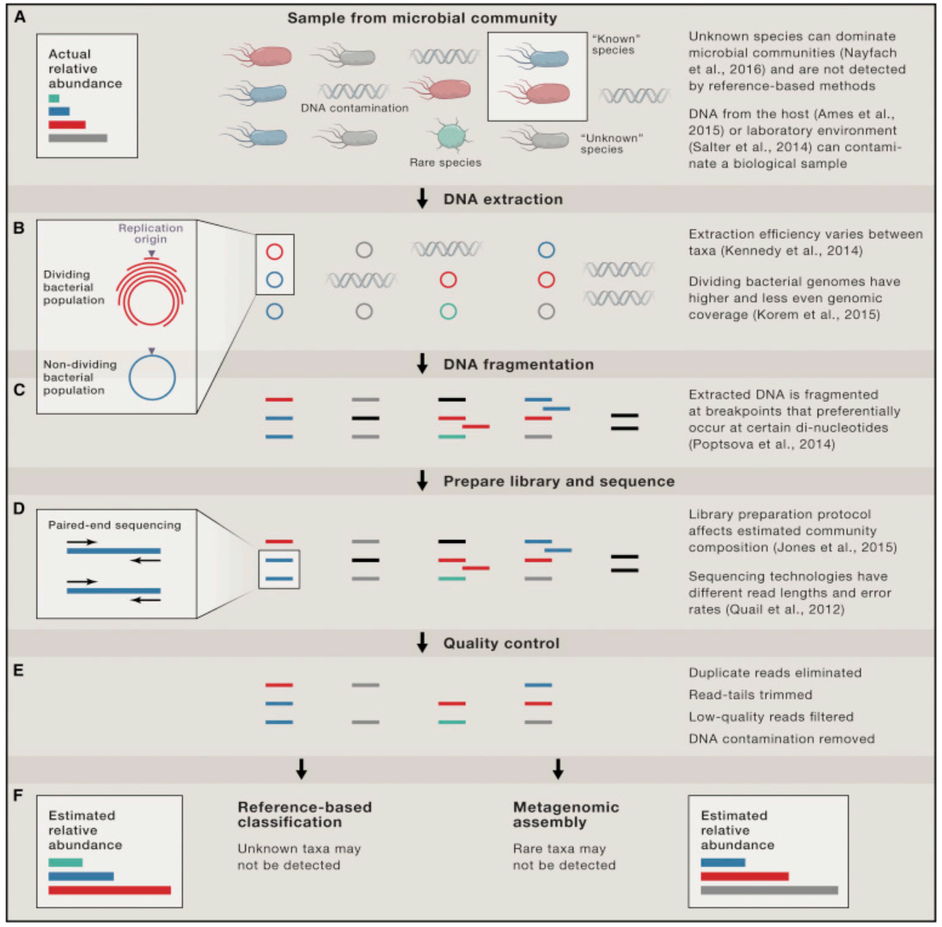
*** This is not comprehensive and is nuanced !!! Don't @ me!**

Bioinformatic Considerations

QIIME
Phyloseq
MOTHUR
…



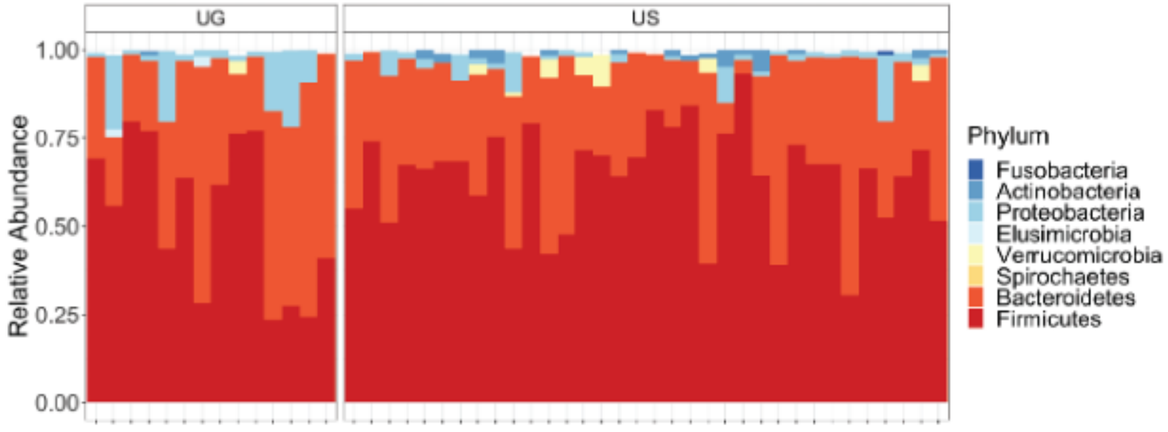
Tringe, S.G., Rubin, E.M. Nat Rev Genet. 2005 Nov;6(11):805-14

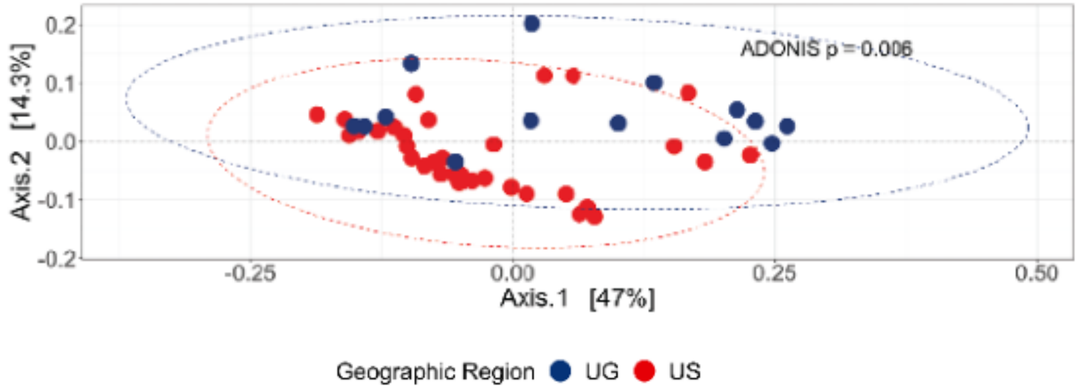


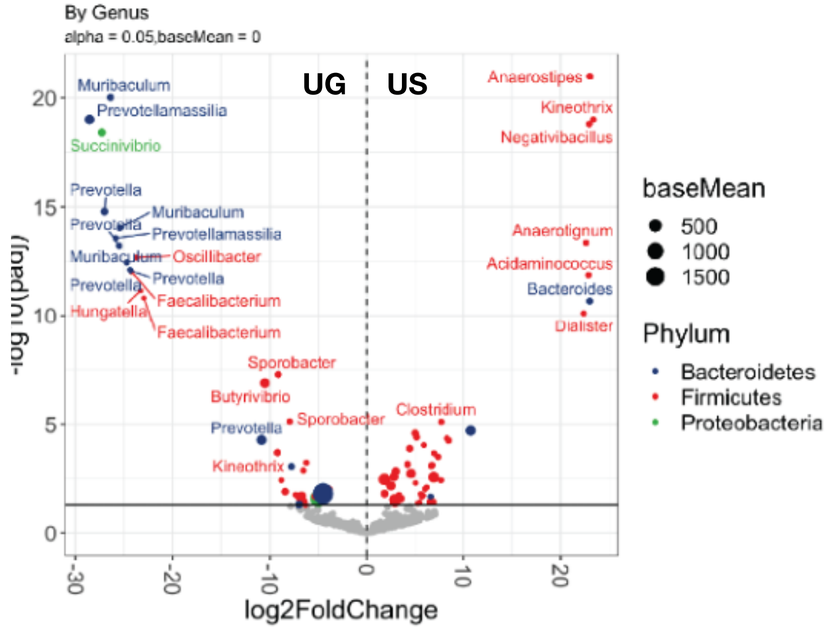
Nayfach S., Pollard KS. Cell. Aug 25;166(5):1103-16

MetaPhlAn
Assembly / Binning approaches
…

What do you get?







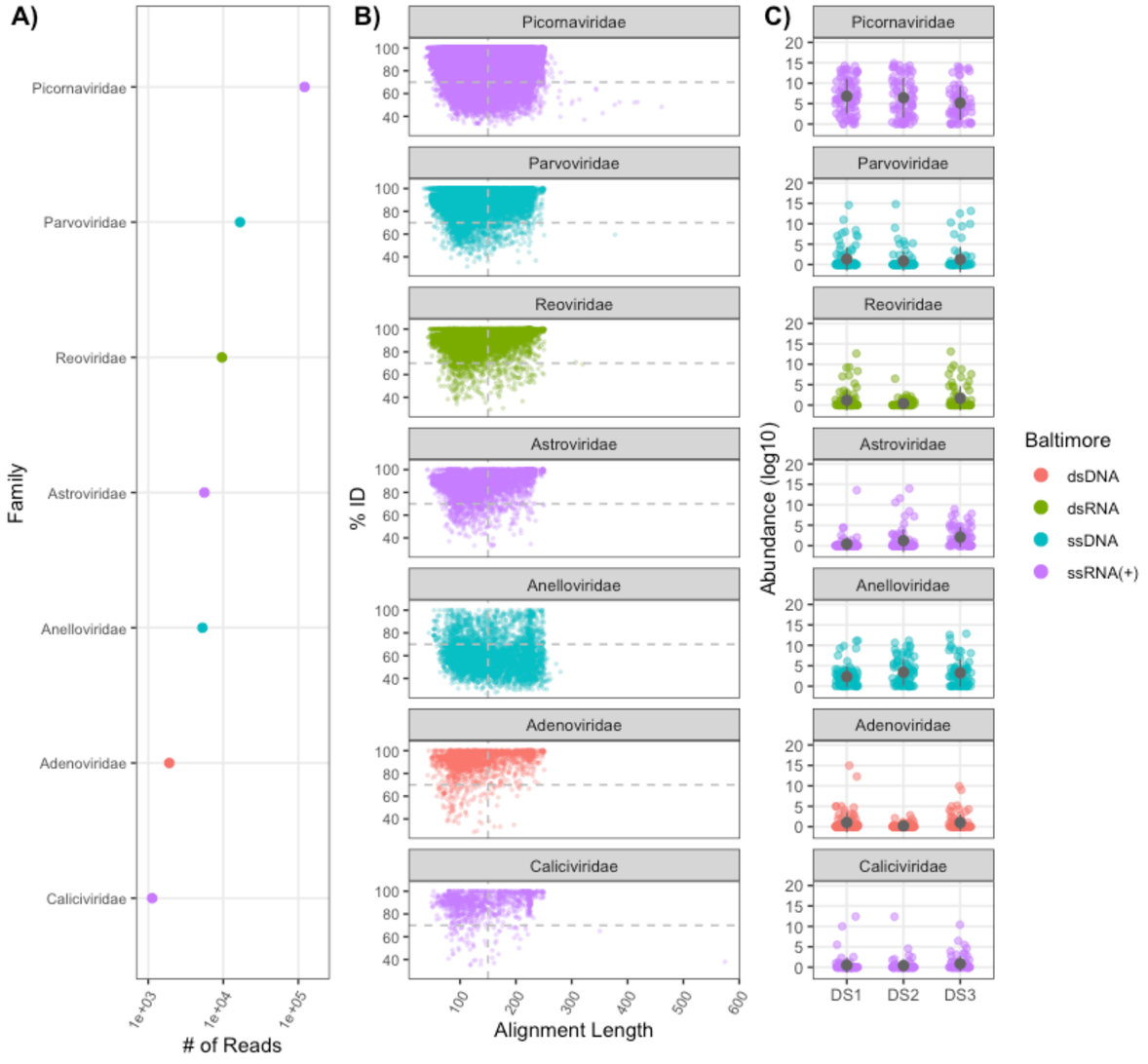
Summary

Both amplicon and metagenomics approaches are valid techniques!

- Sample type (biomass, quality, etc.)
- Cost per sample (metagenomics tends to still be more expensive)
  - Mixed model: Amplicon survey on all samples, deep metagenomics on a subset of samples
- Are your goals taxonomic or functional?
  - If taxonomic will amplicon approaches give you the resolution you require?
- Computational
  - A lot of metagenomics generates larger data sets and *can* be computationally more expensive to run than amplicon based approaches
  - Do appropriate reference databases exist for your sample type?

Virome Analysis: Hecatomb

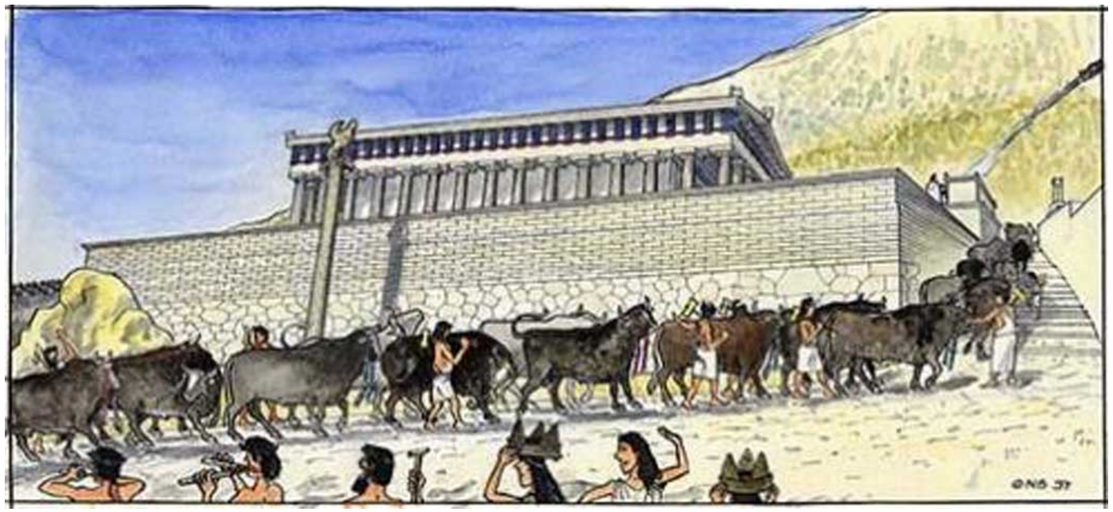
https://github.com/shandley/hecatomb



History

Etymology

- ***Hecatomb*** 1: An ancient Greek and Roman sacrifice of 100 oxen or cattle. 2: the sacrifice or slaughter of many victims
    - Two '*accepted*' pronunciations: heh-kuh-***towm*** (American English) heh-kuh-***toom*** (British English)
- *hecatomb* is used to describe the sacrifice or destruction by fire, tempest, disease or the sword of any large number of persons or animals; and also of the wholesale destruction of inanimate objects, and even of mental and moral attributes
    - In our case, we are trying to destroy false-positive viral calls using bioinformatics



Its also the name of a card game!

… and a death metal band!

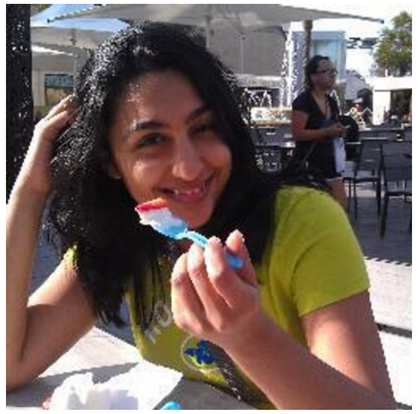
Development

Washington University (USA)

Kathie Mihindukulasuriya

Leran Wang

Barry Hykes (past-member)



Chandni Desai (past-member)

Flinders University (Australia)

Michael Roach

Rob Edwards

Computational and Experimental Resources for Virome Analysis in Inflammatory Bowel Disease (CERVAID) - RC2 DK116713
Emerging infections: surveillance, epidemiology and pathogenesis (U01 AI151810)

Luigi Marongiu

Luigi Marongiu
To: Handley, Scott

Thursday, Nov 28, 2019, 9:26 AM

Dear Scott,
I am running BlastX but so far it has been an hecatomb: of the 300 sequences ran, only 18 were identified as viruses. I'd be lucky if I'll have 30 viruses out of 700 initially identified.
Just to be sure, the pipeline I have done was:

1. blastn of the reads and discard those that had lower e-value for the human genome
2. get all the reads for each patient/tissue that mapped on a specific virus then generate a cluster as
    a. if the reads were overlapping, merge them into a contig using a consensus generated with EMBOSS cons from a clustalX alignment
    b. reads that did not overlap were given as a separate contig
    c. the contigs mapping on the same virus were concatenated with an NNNNN string in between
3. run blastX and retrieve the top 10 hits
4. those that have all hits as bacteria are discarded (which are alarmingly about 97% of the hits!)
5. manually check all the others (since they are few, I can do that)

Is this pipeline acceptable? is this failure rate normal or is there something weird in the data?
Thank you
Luigi

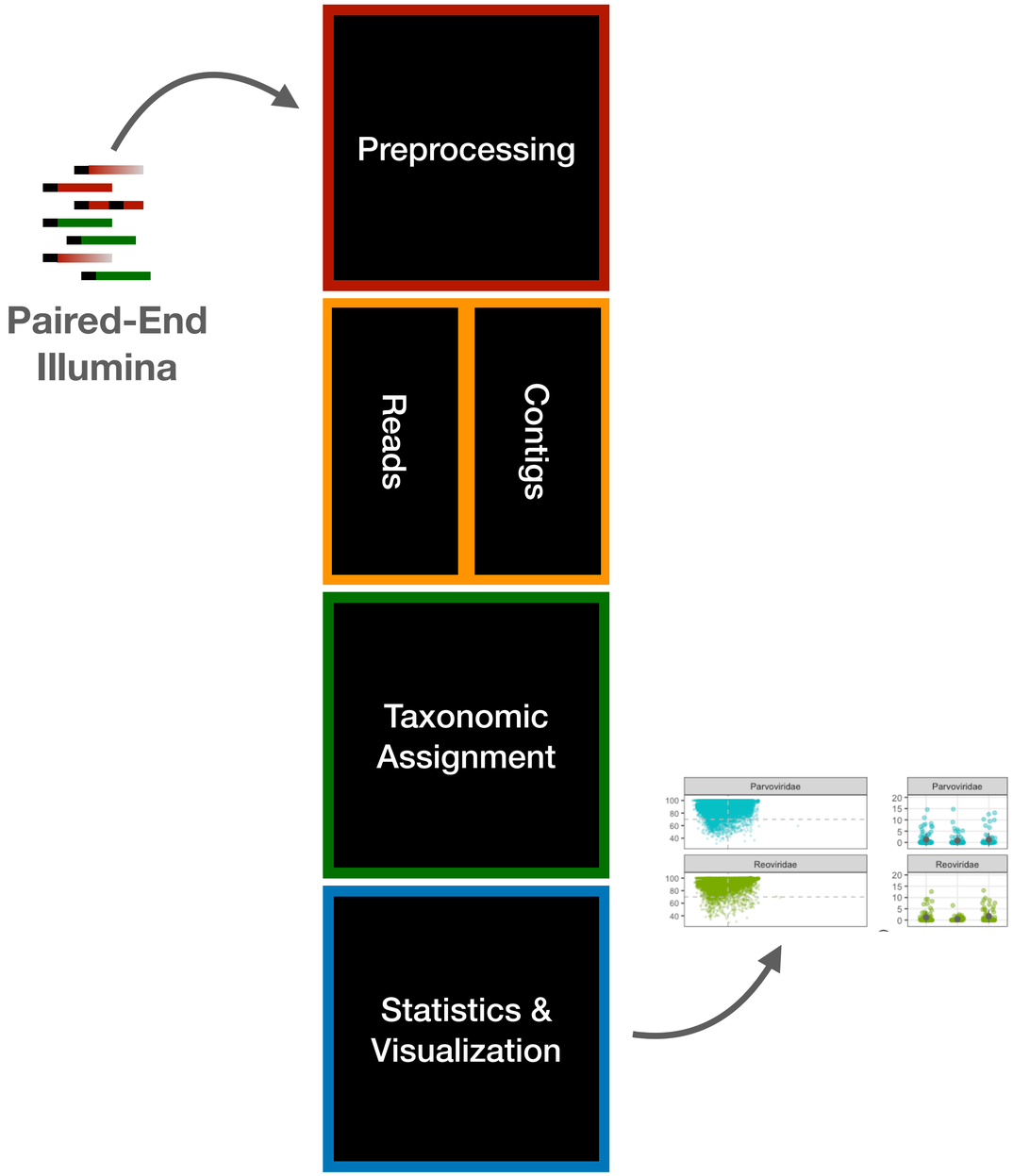
Reply | Forward | Quick Reply

Other Software

- **VirusSeeker** - https://github.com/guoyanzhao/VirusSeeker-Virome
  - Gold standard for removal of false-positives
  - Challenging to run under different compute architectures
  - Results are difficult to integrate with other data types
- **IdSeq** - https://idseq.net
  - Cloud-based
  - “All” microorganisms, just not virus
  - No phage analysis
  - ‘Complicated’ terms-of-service
- **VirScan** / **VirFinder** / **DeepVirFinder** / **cenotetaker2**
  - Viral contig annotation only

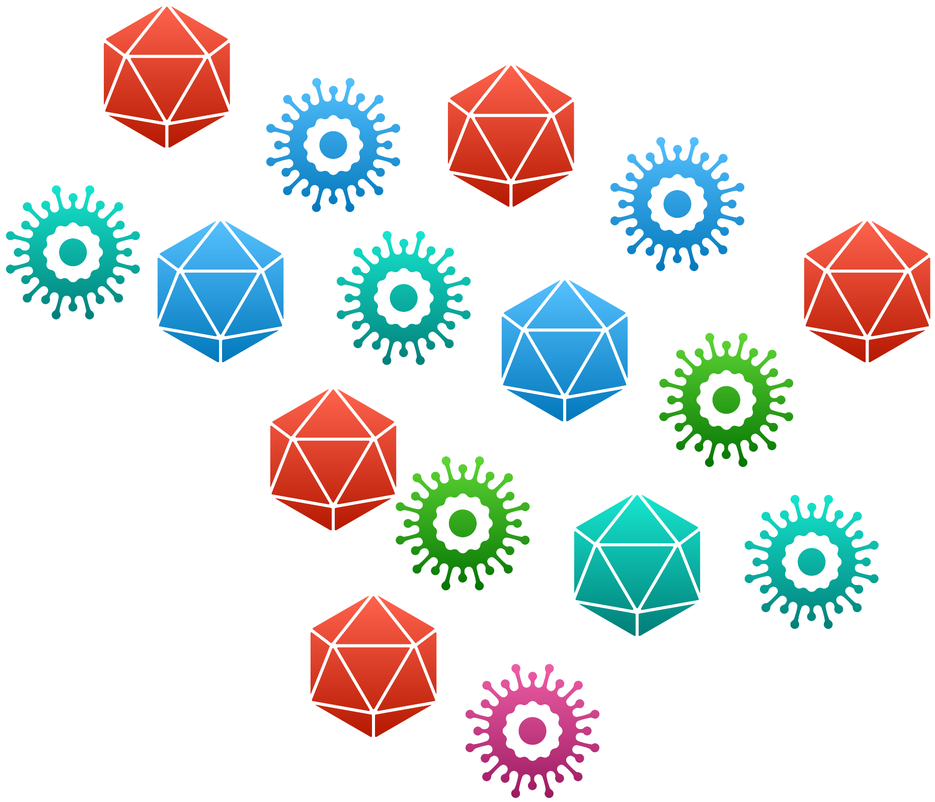
What hecatomb *is*?

- Broadly:
  - Virome analysis software
- Specifically:
  - Computational workflow to detect and annotate viral sequences from metagenomic sequences
  - Can detect and analyze both phage and eukaryotic viral sequences
  - Works on individual reads and contigs
  - Integrates taxonomy, counts, sample data and external data sources into a single R object
  - Workflow management with Snakemake
  - Dependency management with Conda
  - Recognizes ***resource imperfection*** and balances it with ***data integration*** and ***investigator tools***



What hecatomb *is not*?

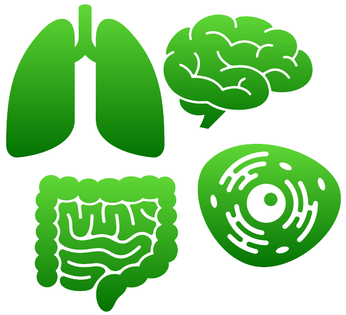
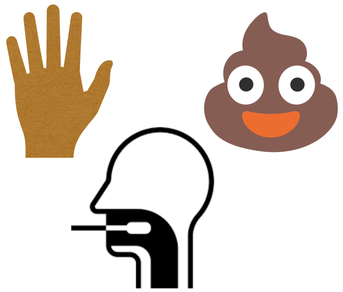
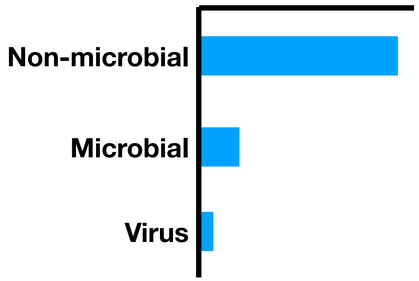
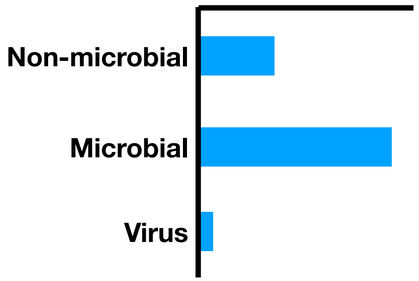
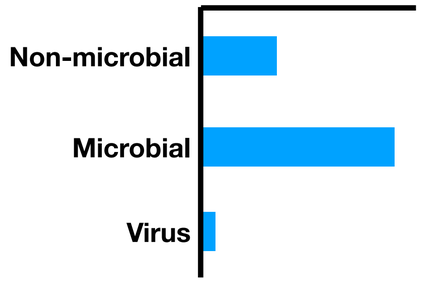
- A bacterial, fungal or other organism analysis tool
- Overly-opinionated
    - Settings are typically set to *annotate* instead of *remove/filter* data
- A ‘push-button’ tool
    - Data production (e.g. quality-control, taxonomic assignment) is relatively well-automated
    - Data analysis is meant to be interactive and managed by an invested researcher
- ***Perfect***



Virome Analysis Challenges

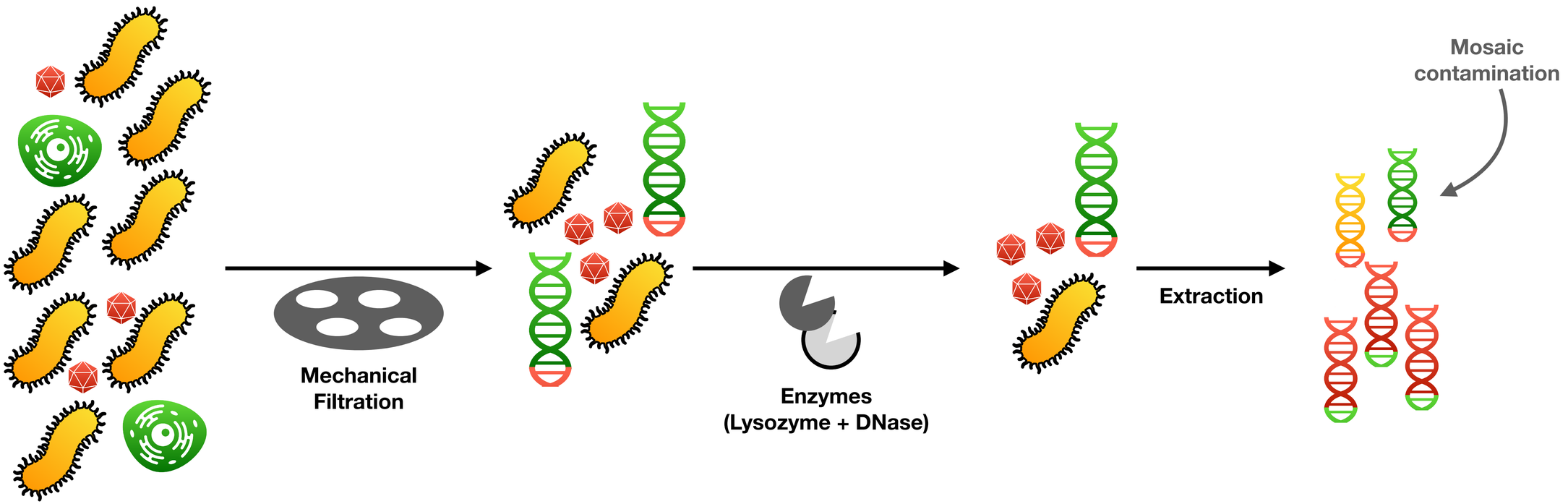
1) Sample types
2) Contamination
3) Genetic mosaicism
4) Viral genome complexity

Challenge 1: Sample Type Variety

	Host Dominant	Microbe Dominant / Host Associated	Microbe Dominant / Environmental
			
Examples	• Tissue biopsy • Cultured cells • Whole animal	• Stool Samples • Skin Swab • Oral Wash	• Water samples • Soil • Surfaces
Properties			
Challenges	• Host background • Viral sequences in host and microbe	• Microbial background • Viral sequences in microbes and host	• Microbial background • Viral sequences in microbes

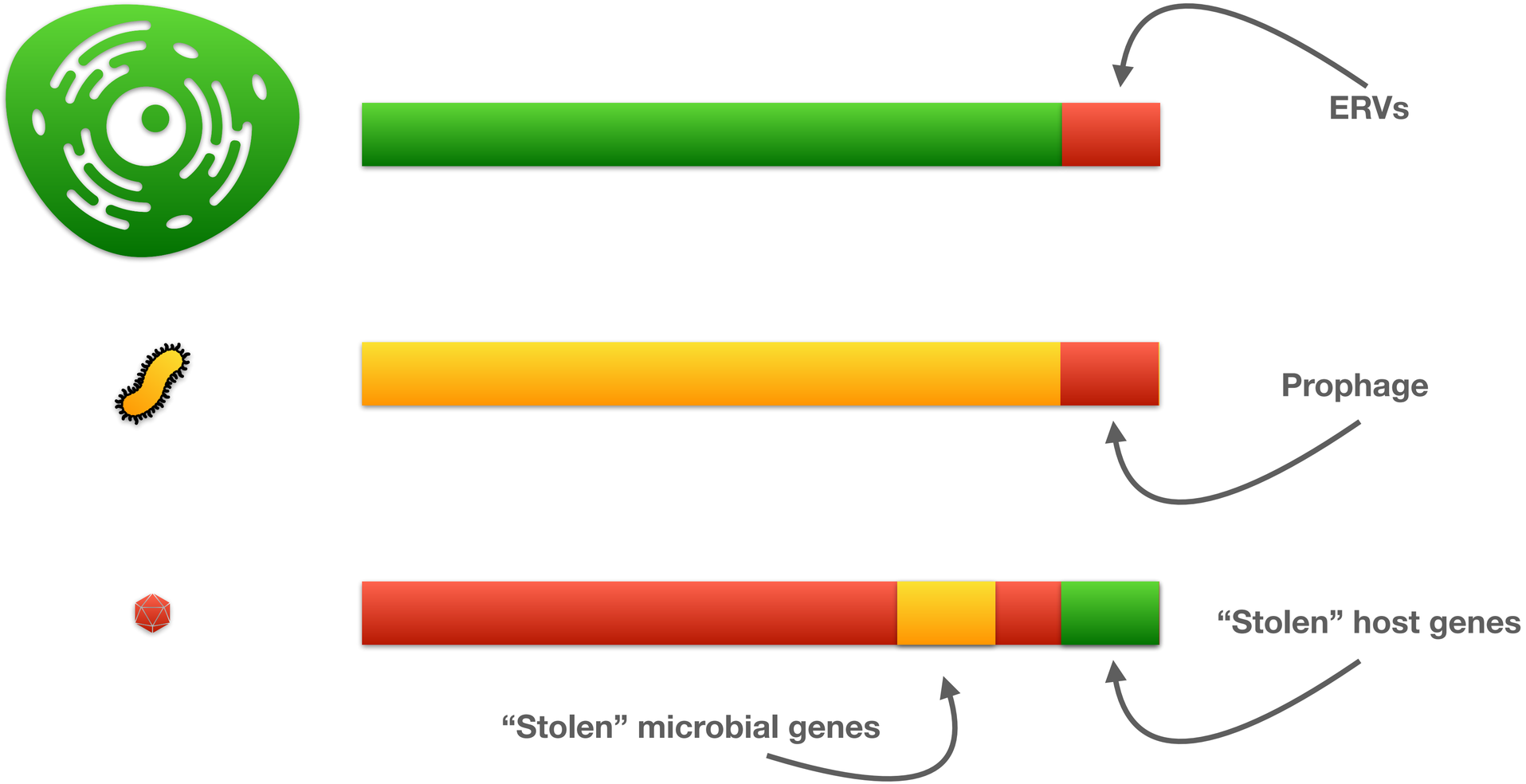
Challenge 2: Contamination

Enrichment ≠ Purification



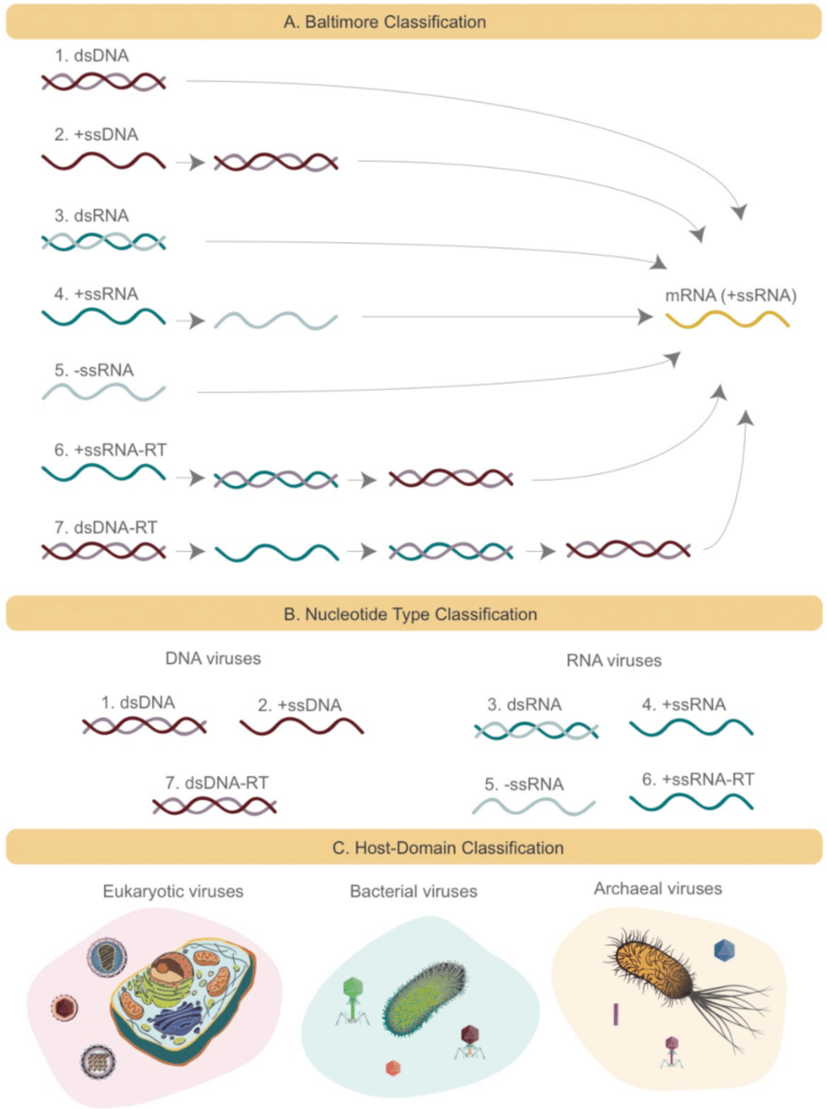
Challenge 3: Genetic Mosaicism

Genetic Mosaicism

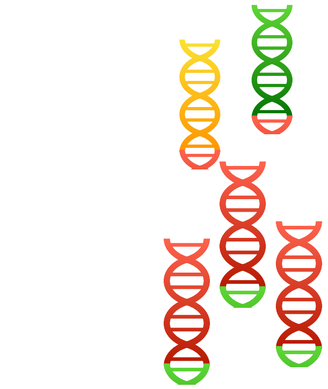


Challenge 4: Viral Genome Complexity

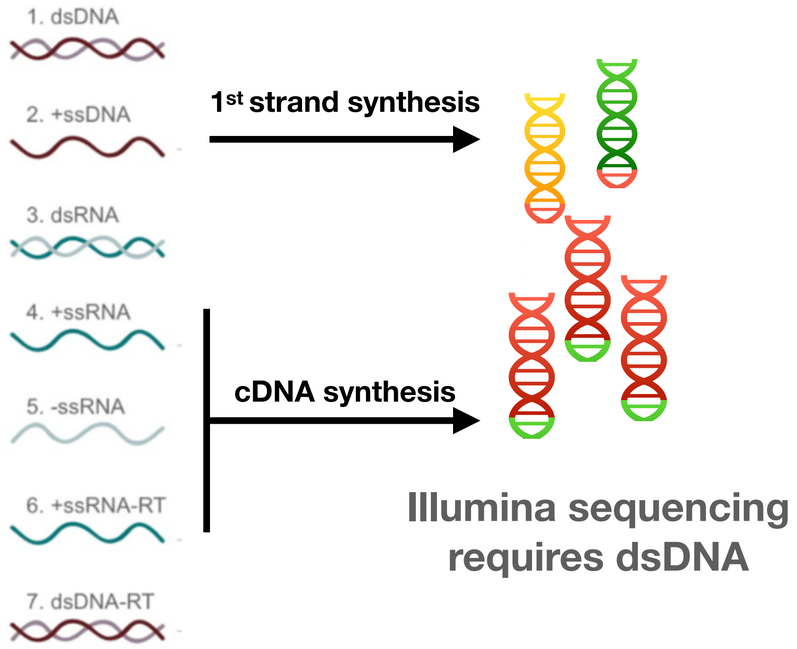
Viral Genome Architectures



Mahmoudabadi G, Phillips R. *Elife*. 2018 doi: 10.7554/eLife.31955



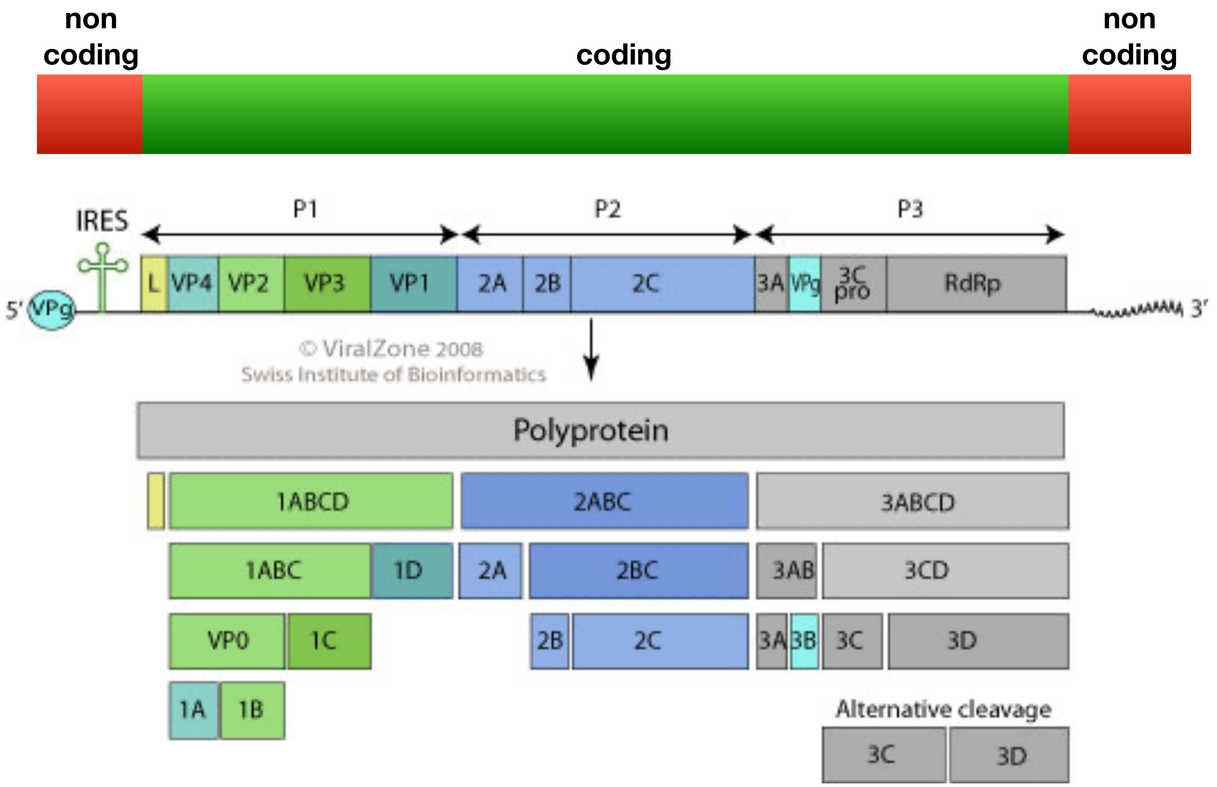
We don't actually start with this



We start with this

Illumina sequencing requires dsDNA

Coding and Non-coding



- **Noncoding regions will not be represented in amino acid databases**

How Hecatomb Works

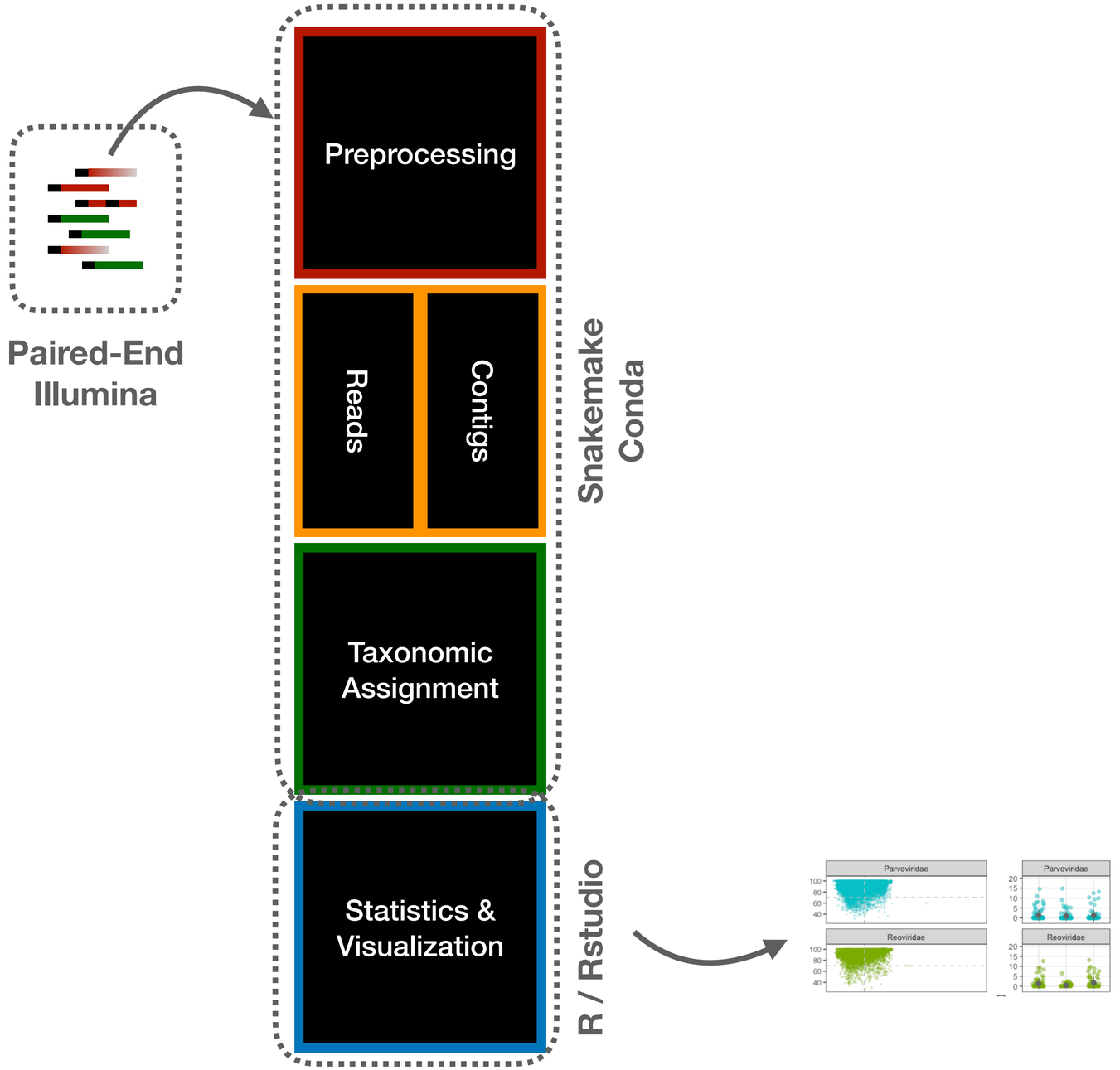
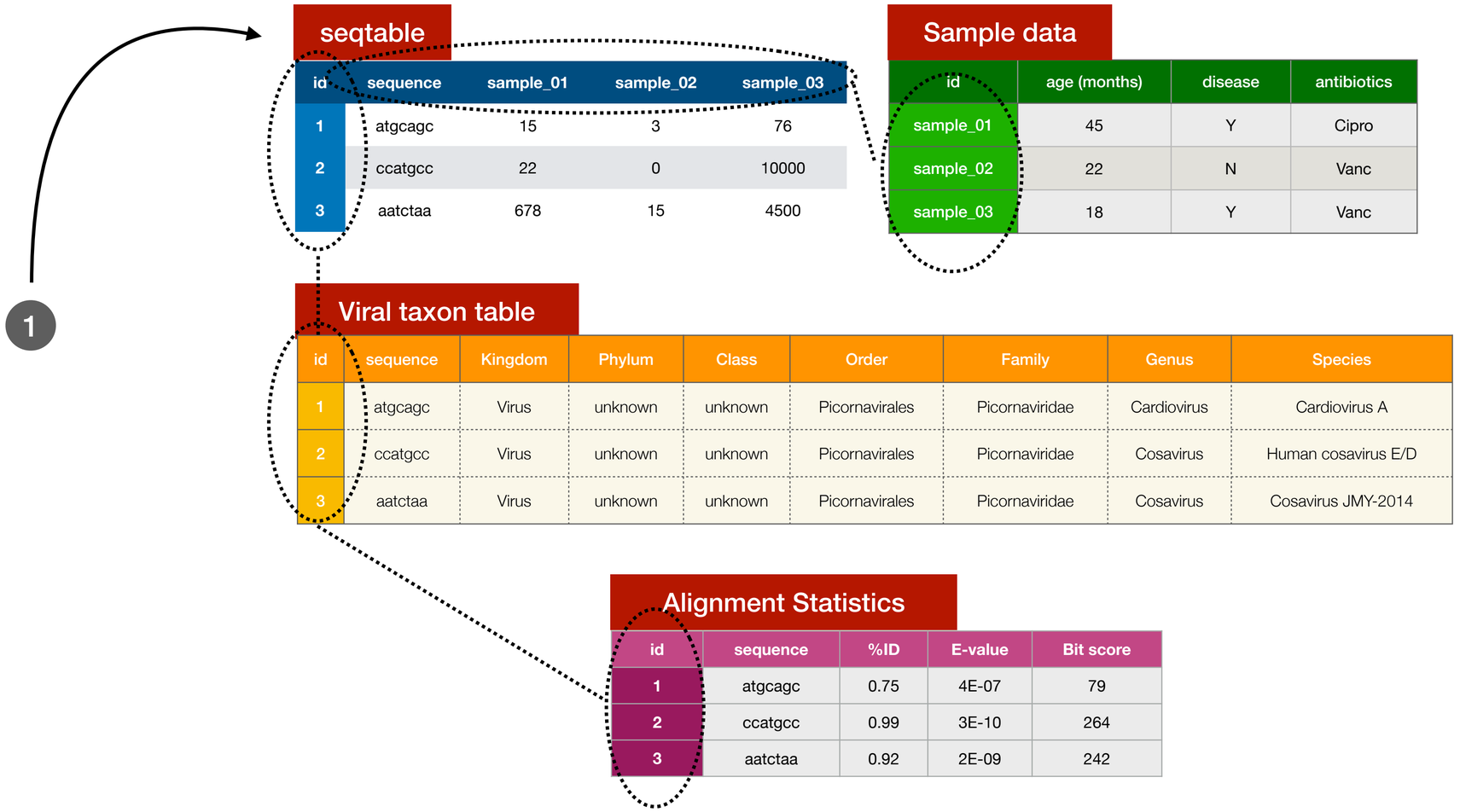

Paired-End
Illumina
Preprocessing
Reads
Contigs
Taxonomic
Assignment
Statistics &
Visualization
Snakemake
Conda
R / Rstudio
Parvoviridae
Reoviridae
Parvoviridae
Reoviridae

Table Building

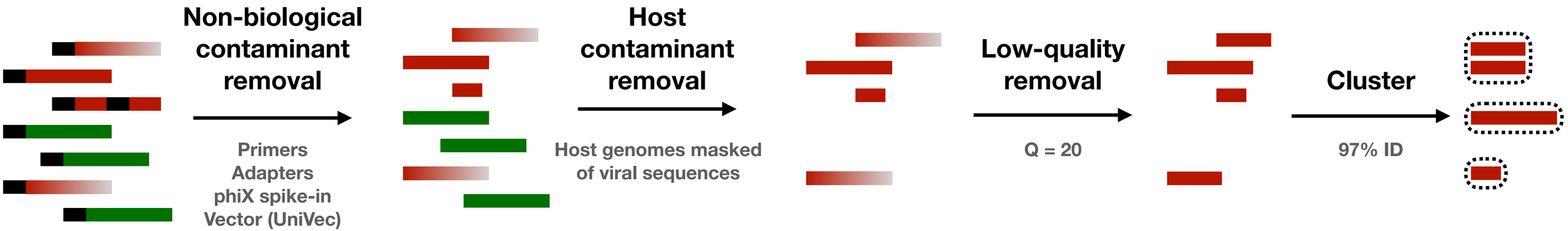


Sequence Table (*seqtable*)

- Every row is a unique sequence
- Every column is a unique sample
- Each cell is the occurrence of each sequence in each sample
- Created by clustering quality-controlled sequences and counting the size of each cluster

id	sequence	sample_01	sample_02	sample_03
1	atgcagc	15	3	76
2	ccatgcc	22	0	10000
3	aatctaa	678	15	4500

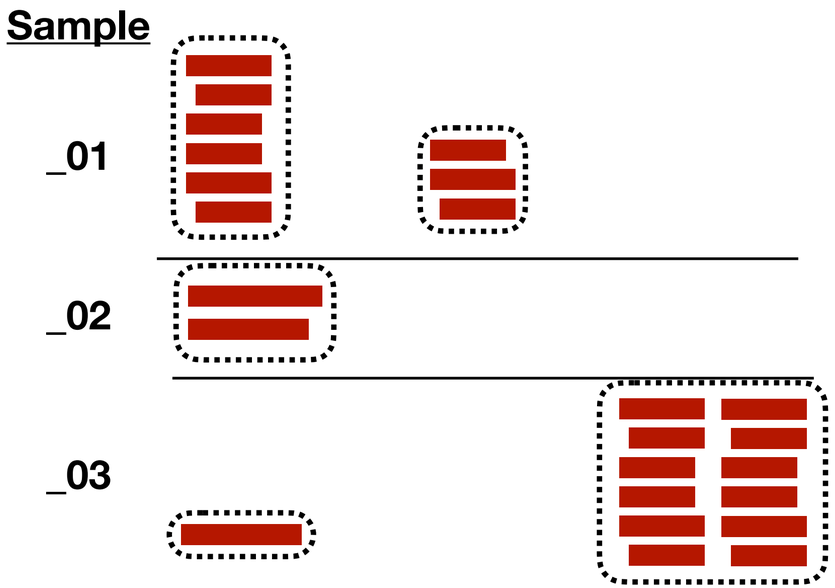
Preprocessing



Seqtable: Clustering & Counting

1

Samples are clustered an counted independently



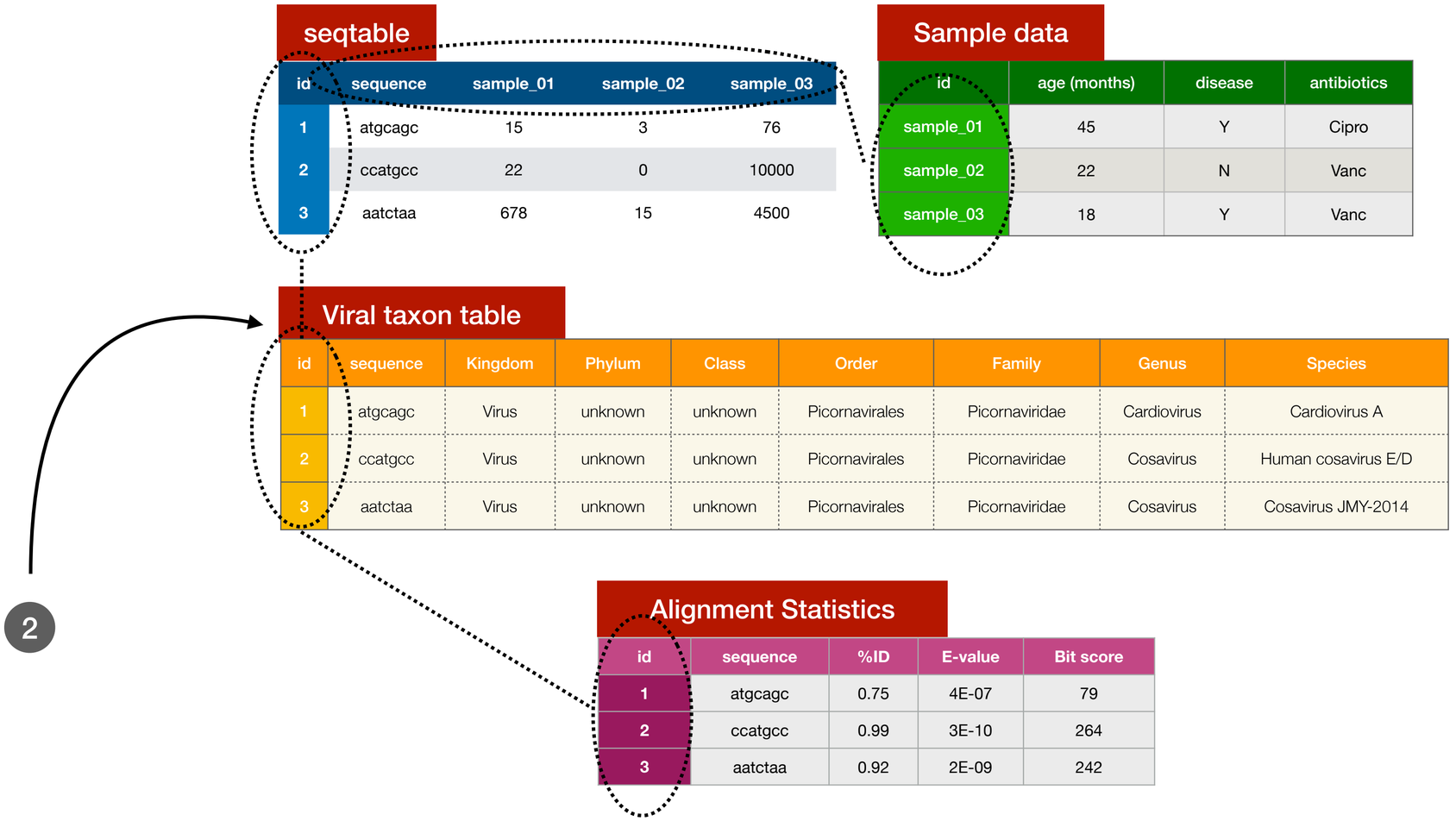
2

Individual sample counts are combined into a single table and fast file

id	sequence	sample_01	sample_02	sample_03
1	atgcagc	6	3	0
2	ccatgcc	2	0	0
3	aatctaa	1	0	12

Representative Sequence
Only needs to be searched once (3 vs. 24)

Taxonomy Table



Taxonomy Table

- Every row is a unique sequence
- Every column is a taxonomic rank
  - Classical *linnean* only (although full is reserved)
- Each cell is the rank appropriate lineage name for each sequence
- Created by searching (mmseqs2) all representative clustered sequences

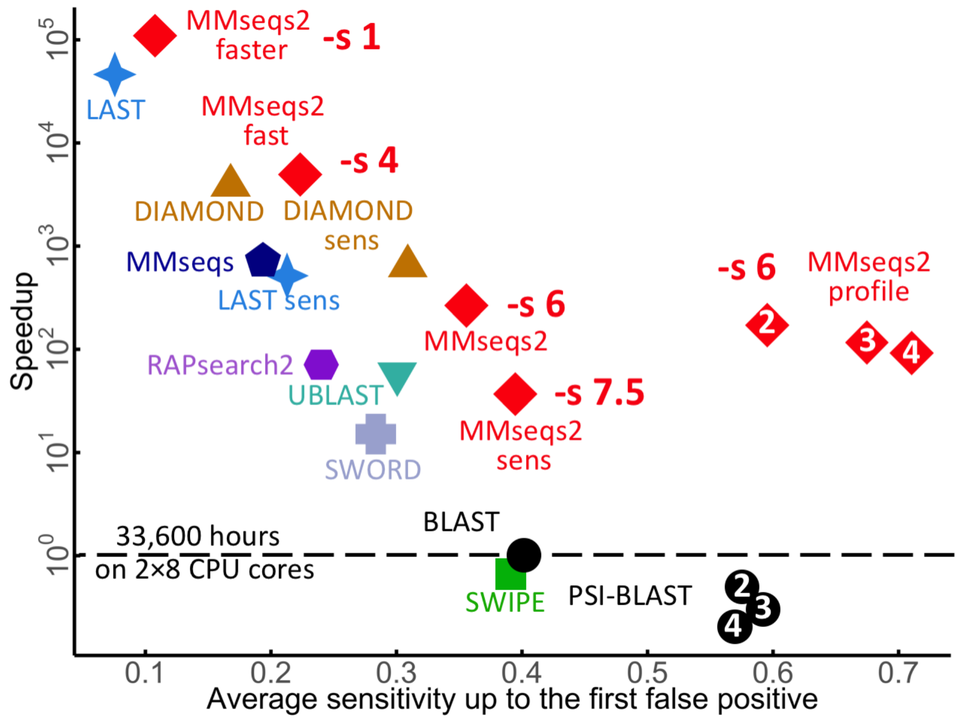
id	sequence
1	atgcagc
2	ccatgcc
3	aatctaa

id	sequence	Kingdom	Phylum	Class	Order	Family	Genus	Species
1	atgcagc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cardiovirus	Cardiovirus A
2	ccatgcc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	Human cosavirus E/D
3	aatctaa	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	Cosavirus JMY-2014

  - Iterative search process
  - Amino acid and nucleotide databases

mmseqs2

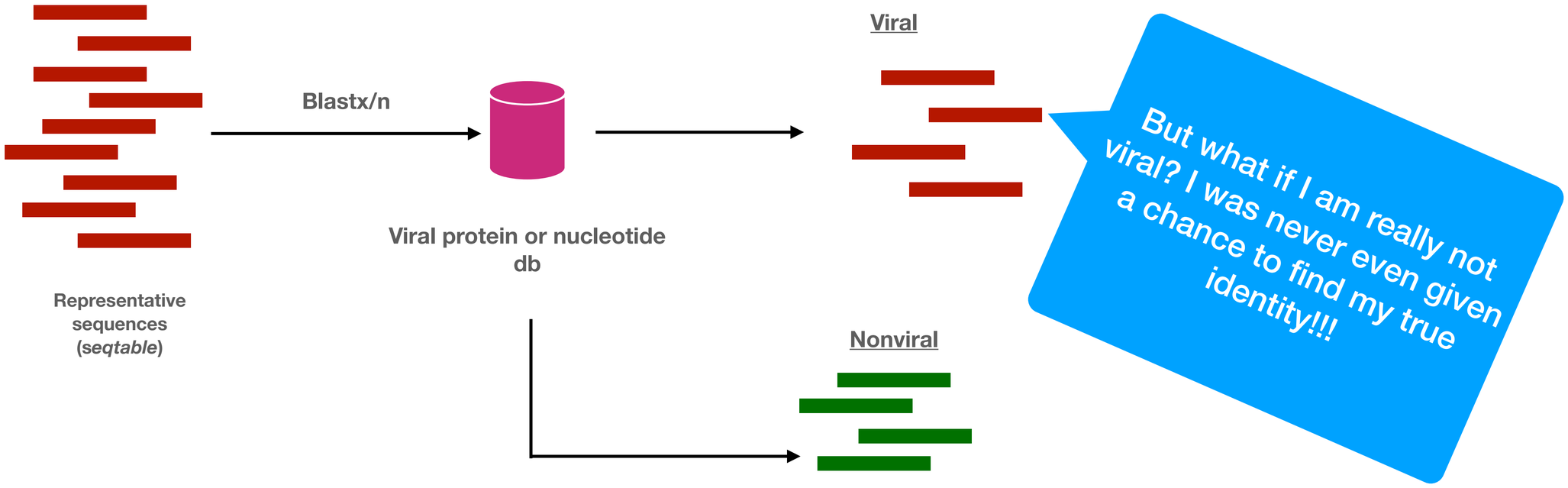
- Fast query of sequences against reference databases
- Untranslated (blastn) and translated (blastx) searching
- Integrated taxonomy modules
  - Multiple lowest common ancestor (LCA) algorithms
- Hecatomb uses iterative searching from -s 1 to -s 7 (see this link for thorough explanation)



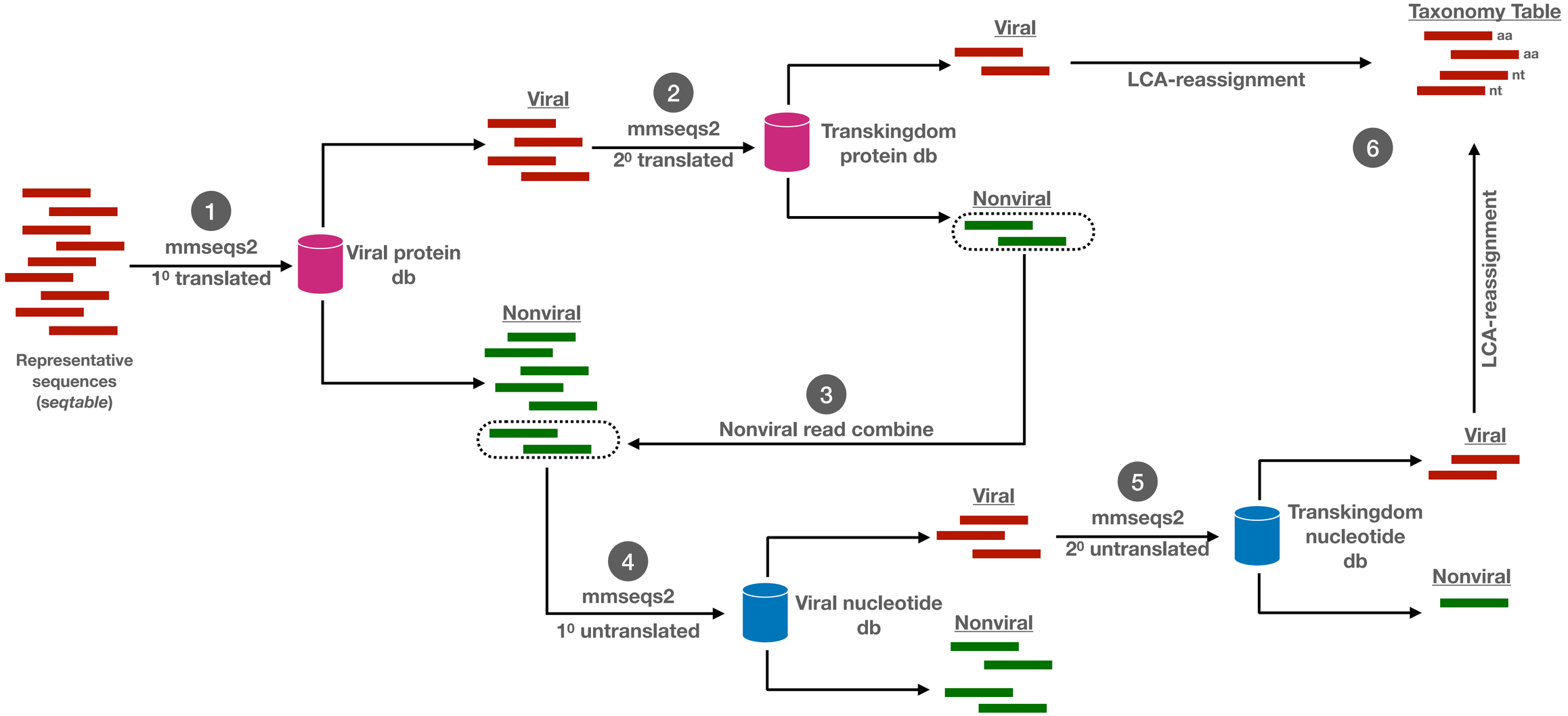
https://github.com/soedinglab/MMseqs2

Steinegger M, Söding J. Nat Biotechnol. 2017 Nov;35(11):1026-1028. doi: 10.1038/nbt.3988. Epub 2017 Oct 16.

Fast (flawed) Approach



Taxonomic Assignment

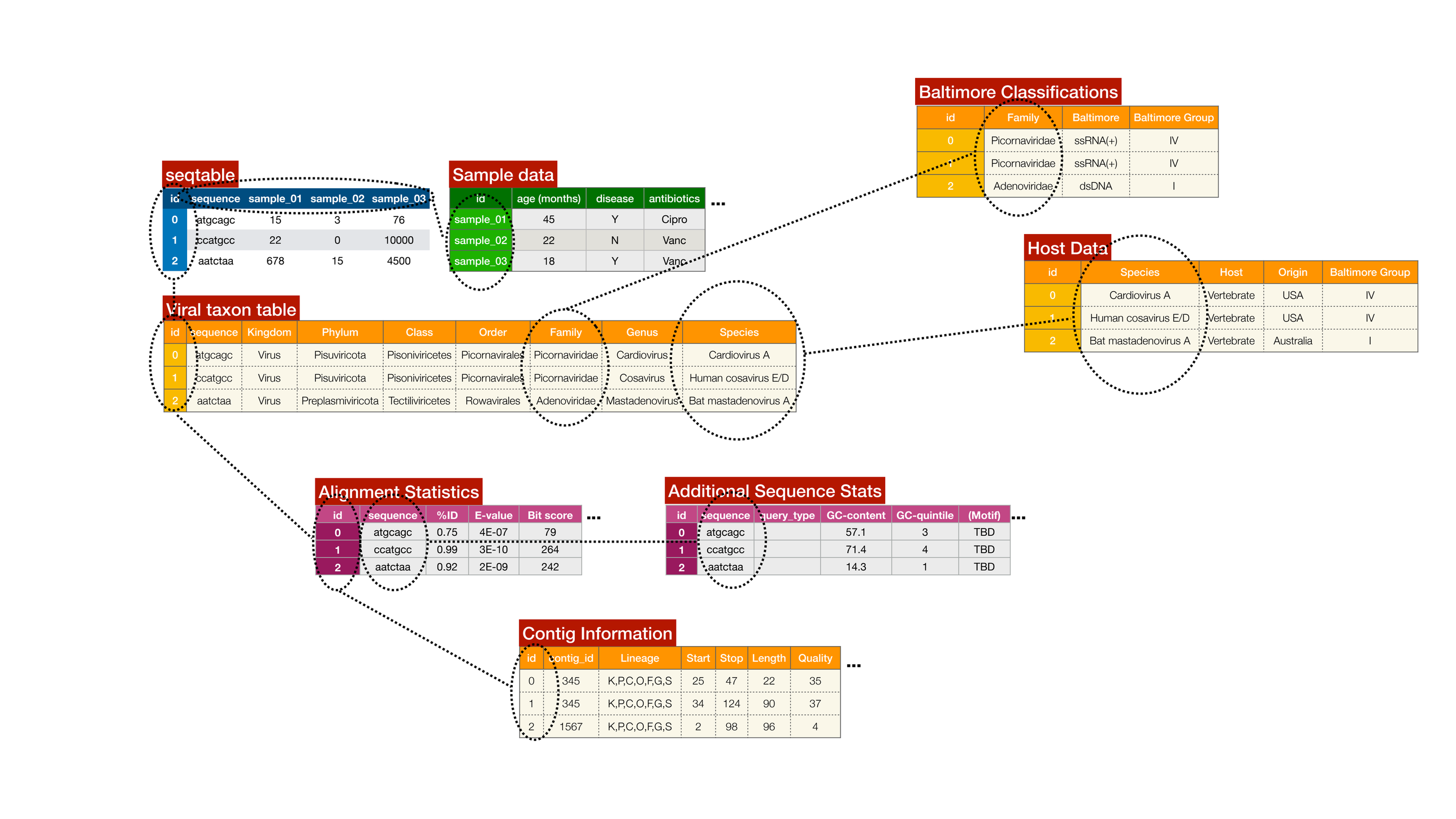


Extended Taxonomy Table

id	sequence	Kingdom	Phylum	Class	Order	Family	Genus	Species	query_type	e_value
1	atgcagc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cardiovirus	Cardiovirus A	translated	1E-02
2	ccatgcc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	Human cosavirus E/D	translated	1E-03
3	aatctaa	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	Cosavirus JMY-2014	untranslated	1E-41

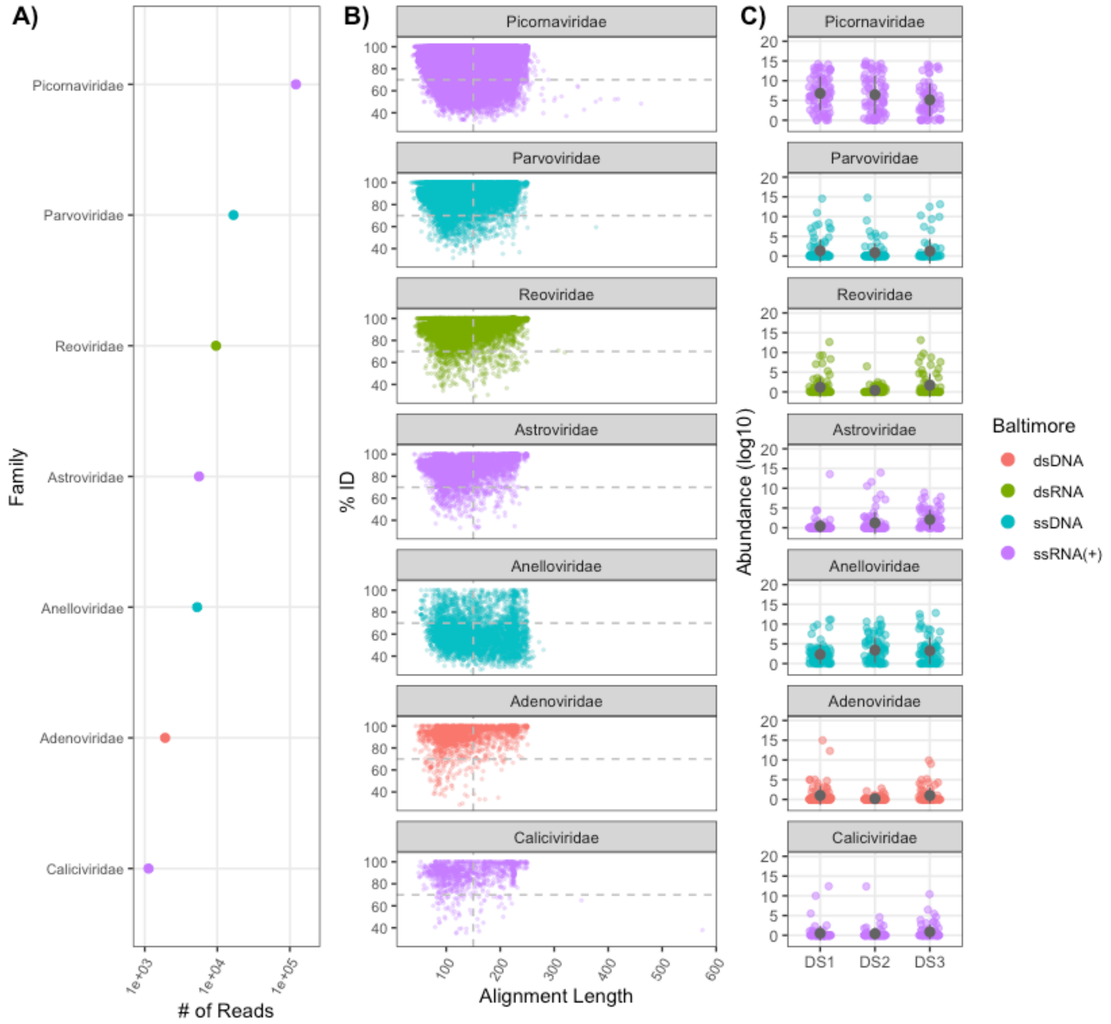
id	sequence	sample_01	sample_02	sample_03
1	atgcagc	6	3	0
2	ccatgcc	2	0	0
3	aatctaa	1	0	12

Data Integration

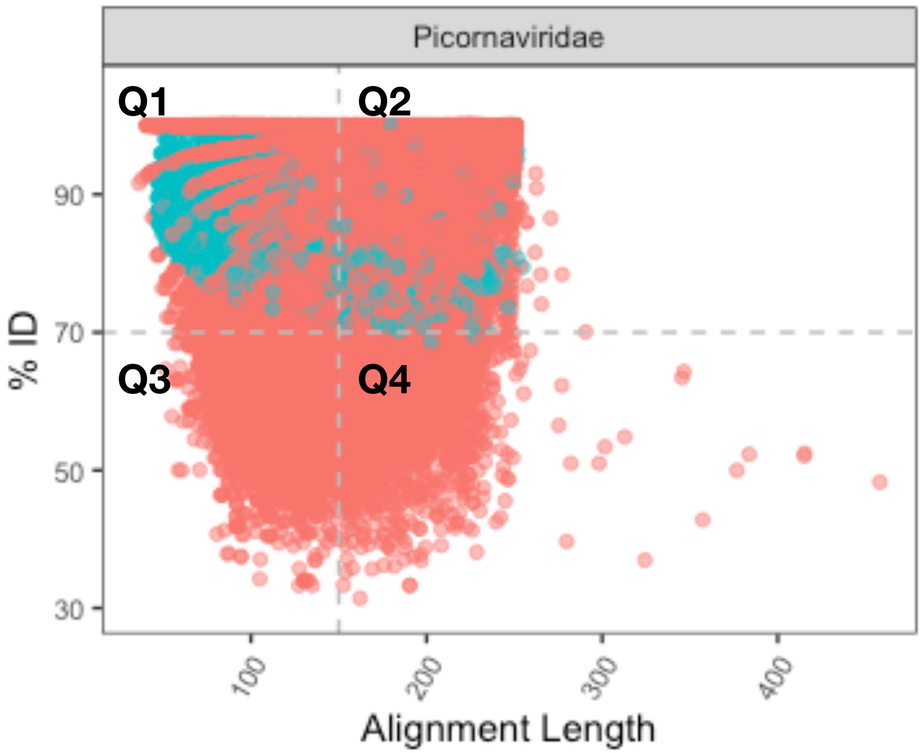

seqtable
id sequence sample_01 sample_02 sample_03
0 atgcagc 15 3 76
1 ccatgcc 22 0 10000
2 aatctaa 678 15 4500
Sample data
id age (months) disease antibiotics ...
sample_01 45 Y Cipro
sample_02 22 N Vanc
sample_03 18 Y Vanc
Baltimore Classifications
id Family Baltimore Baltimore Group
0 Picornaviridae ssRNA(+) IV
1 Picornaviridae ssRNA(+) IV
2 Adenoviridae dsDNA I
Host Data
id Species Host Origin Baltimore Group
0 Cardiovirus A Vertebrate USA IV
1 Human cosavirus E/D Vertebrate USA IV
2 Bat mastadenovirus A Vertebrate Australia I
Viral taxon table
id sequence Kingdom Phylum Class Order Family Genus Species
0 atgcagc Virus Pisuviricota Pisoniviricetes Picornavirales Picornaviridae Cardiovirus Cardiovirus A
1 ccatgcc Virus Pisuviricota Pisoniviricetes Picornavirales Picornaviridae Cosavirus Human cosavirus E/D
2 aatctaa Virus Preplasmiviricota Tectiliviricetes Rowavirales Adenoviridae Mastadenovirus Bat mastadenovirus A
Alignment Statistics
id sequence %ID E-value Bit score ...
0 atgcagc 0.75 4E-07 79
1 ccatgcc 0.99 3E-10 264
2 aatctaa 0.92 2E-09 242
Additional Sequence Stats
id sequence query_type GC-content GC-quintile (Motif) ...
0 atgcagc 57.1 3 TBD
1 ccatgcc 71.4 4 TBD
2 aatctaa 14.3 1 TBD
Contig Information
id contig_id Lineage Start Stop Length Quality ...
0 345 K,P,C,O,F,G,S 25 47 22 35
1 345 K,P,C,O,F,G,S 34 124 90 37
2 1567 K,P,C,O,F,G,S 2 98 96 4

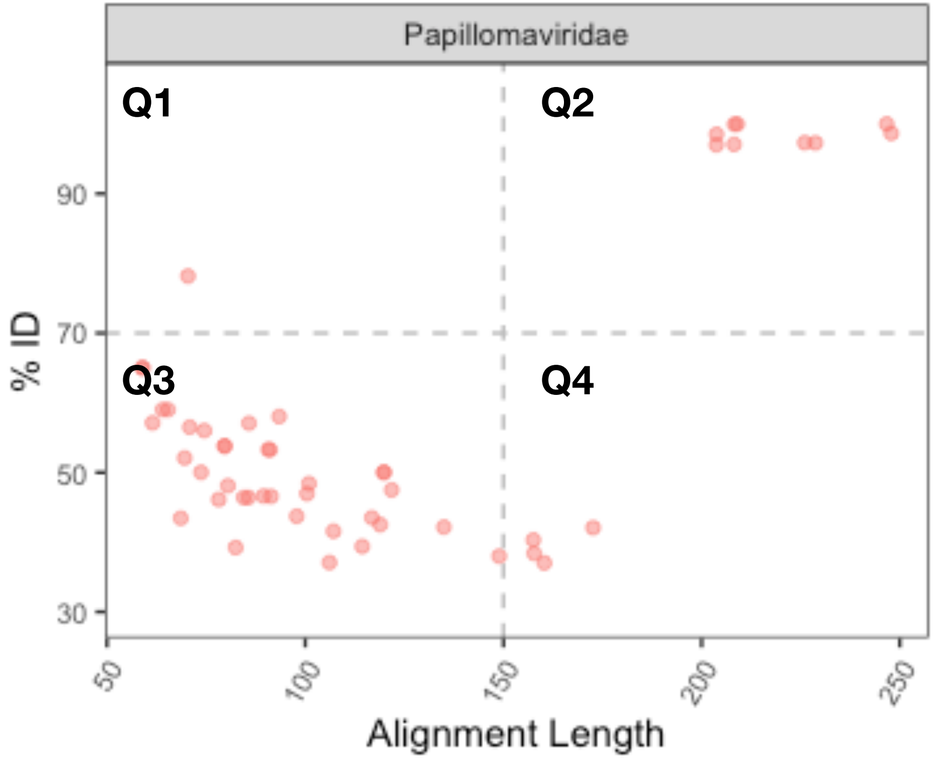
Example Analysis

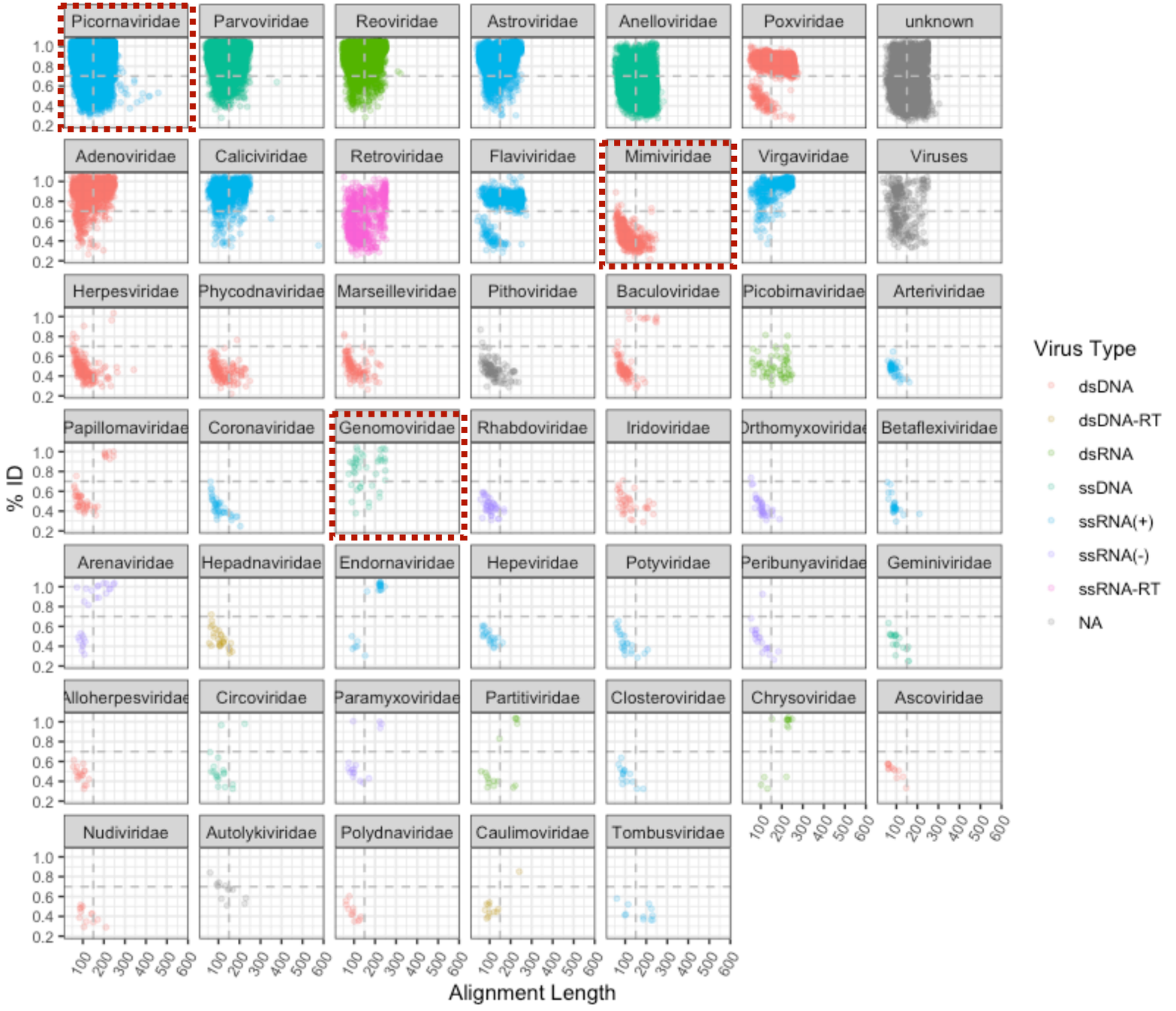
- Pediatric stools collected from infants in Ghana
- Three time-points coinciding with 3 dose vaccine schedule with the oral rotavirus vaccine (rotarix)



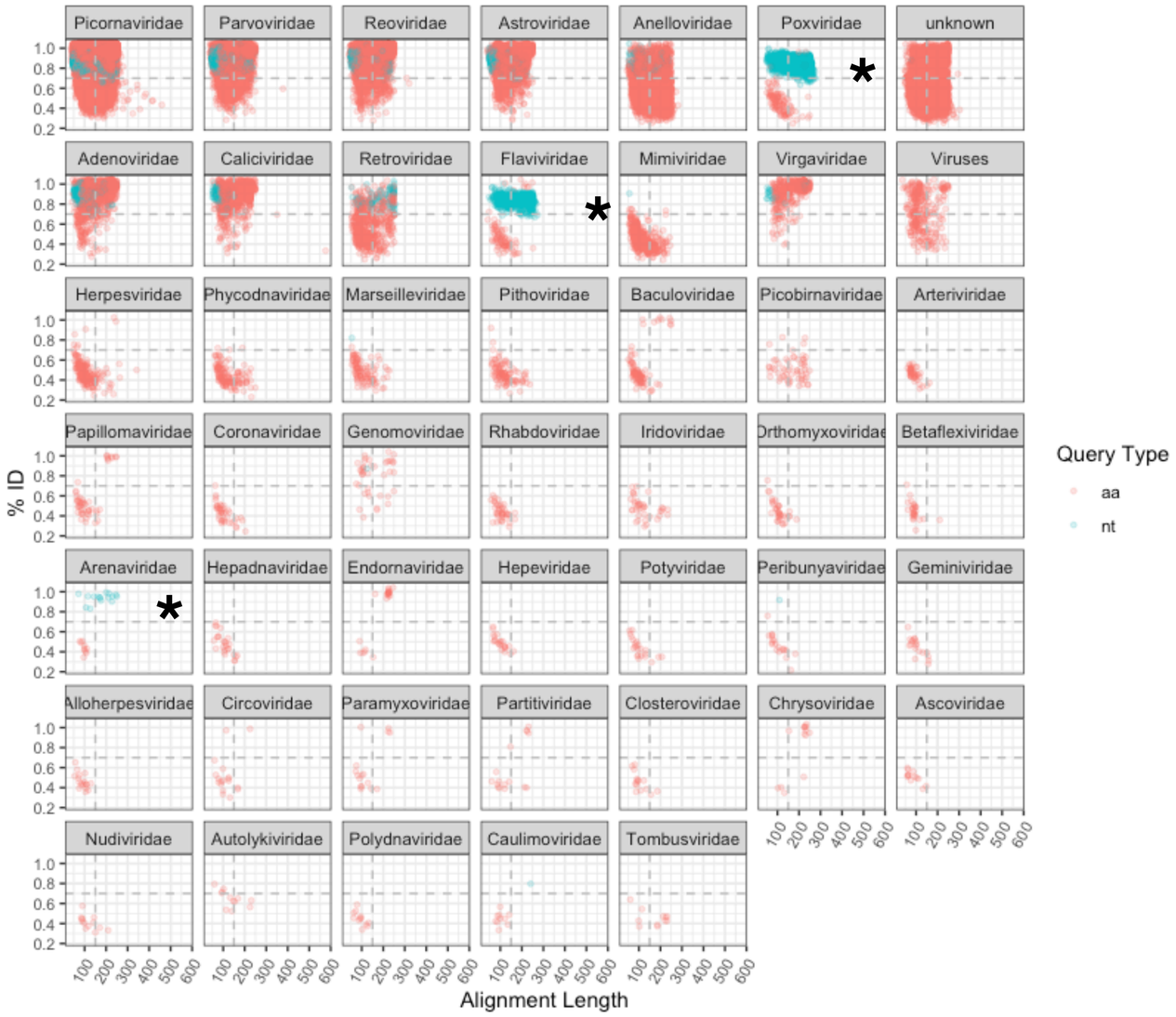
Read Stat Quadrant Analysis



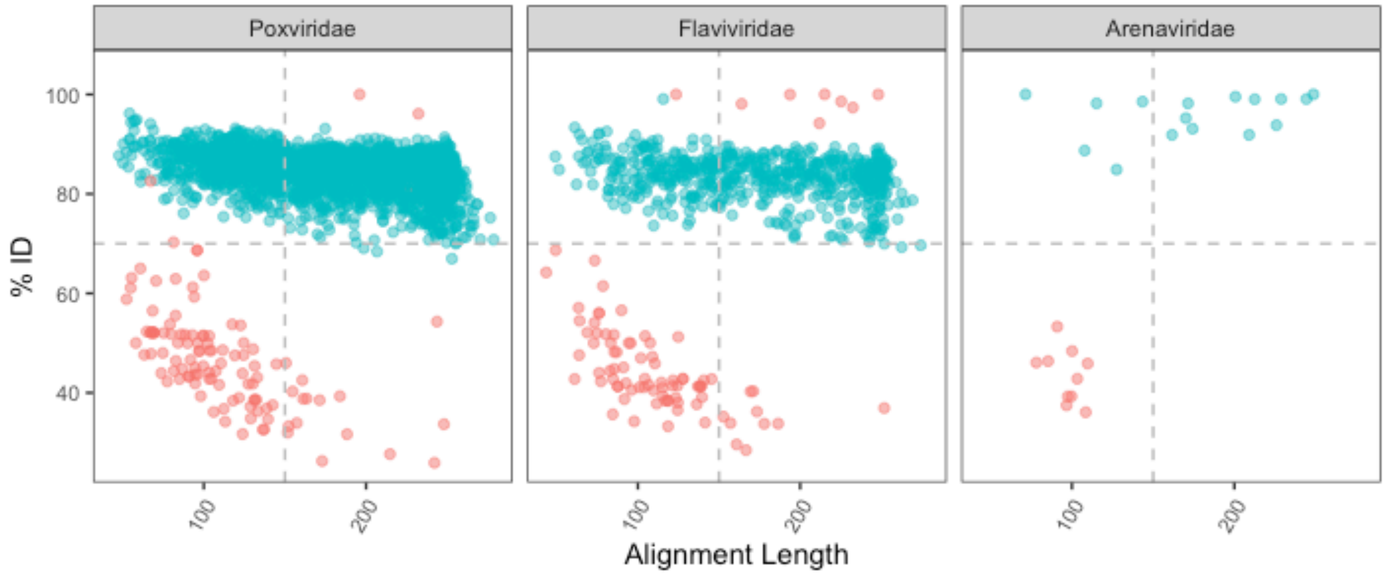


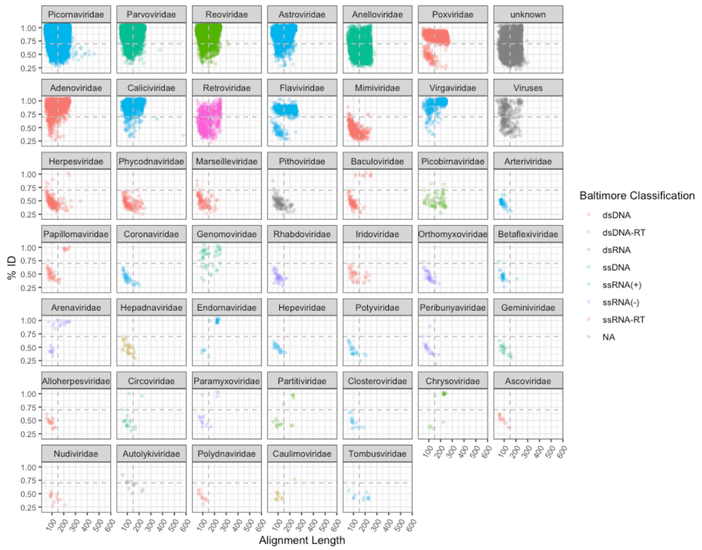

Picornaviridae
Parvoviridae
Reoviridae
Astroviridae
Anelloviridae
Poxviridae
unknown
Adenoviridae
Caliciviridae
Retroviridae
Flaviviridae
Mimiviridae
Virgaviridae
Viruses
Herpesviridae
Phycodnaviridae
Marseilleviridae
Pithoviridae
Baculoviridae
Picobirnaviridae
Arteriviridae
Papillomaviridae
Coronaviridae
Genomoviridae
Rhabdoviridae
Iridoviridae
Orthomyxoviridae
Betaflexiviridae
Arenaviridae
Hepadnaviridae
Endornaviridae
Hepeviridae
Potyviridae
Peribunyaviridae
Geminiviridae
Alloherpesviridae
Circoviridae
Paramyxoviridae
Partitiviridae
Closteroviridae
Chrysoviridae
Ascoviridae
Nudiviridae
Autolykiviridae
Polydnaviridae
Caulimoviridae
Tombusviridae
% ID
1.0
0.8
0.6
0.4
0.2
100
200
300
400
500
600
Alignment Length
Virus Type
dsDNA
dsDNA-RT
dsRNA
ssDNA
ssRNA(+)
ssRNA(-)
ssRNA-RT
NA

Power of Integrated Data

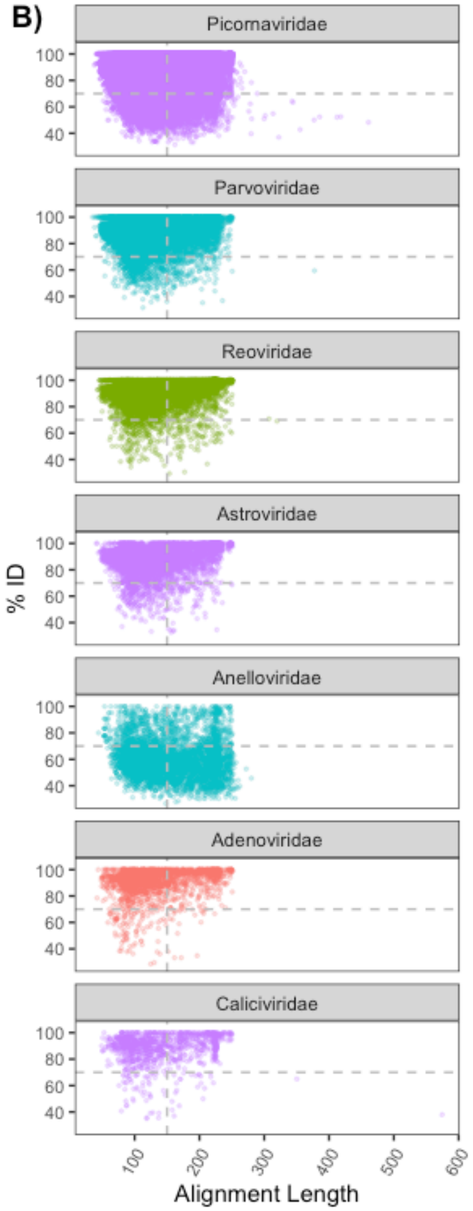


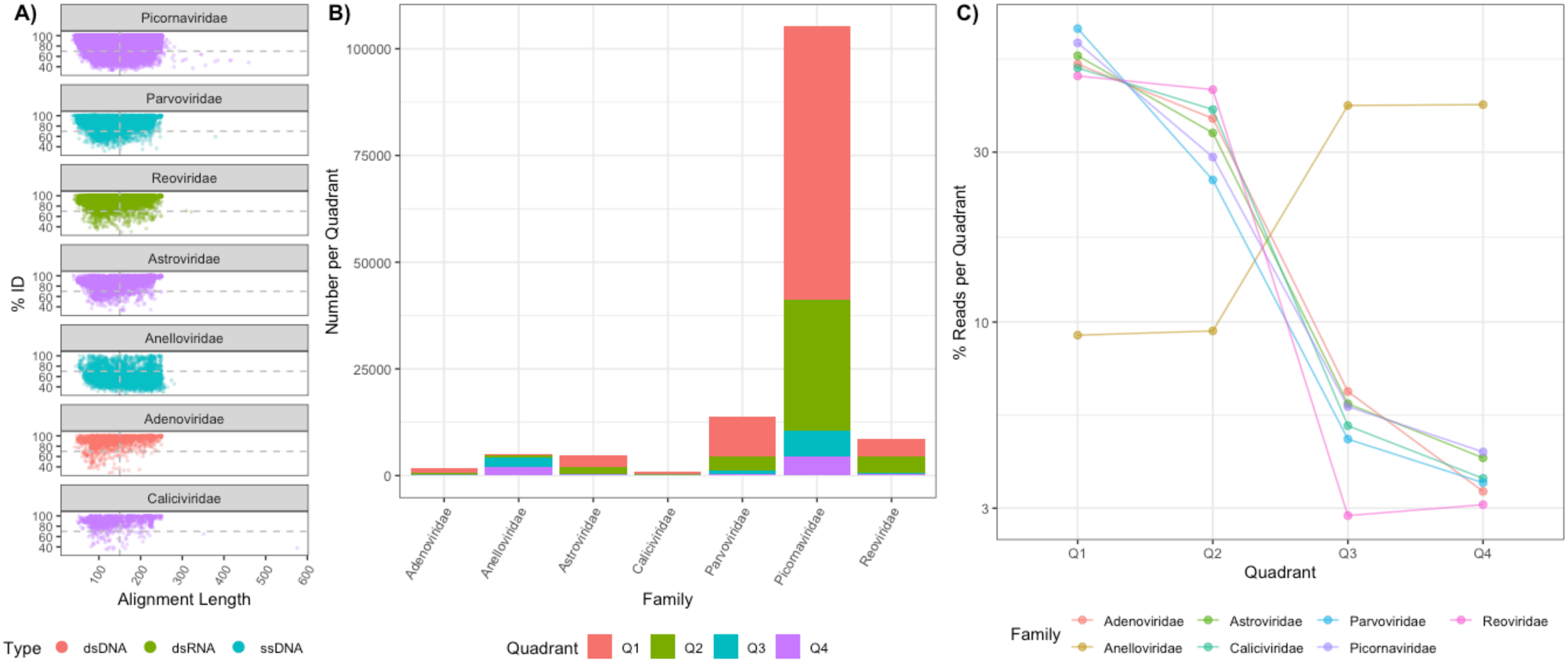
Power of Integrated Data





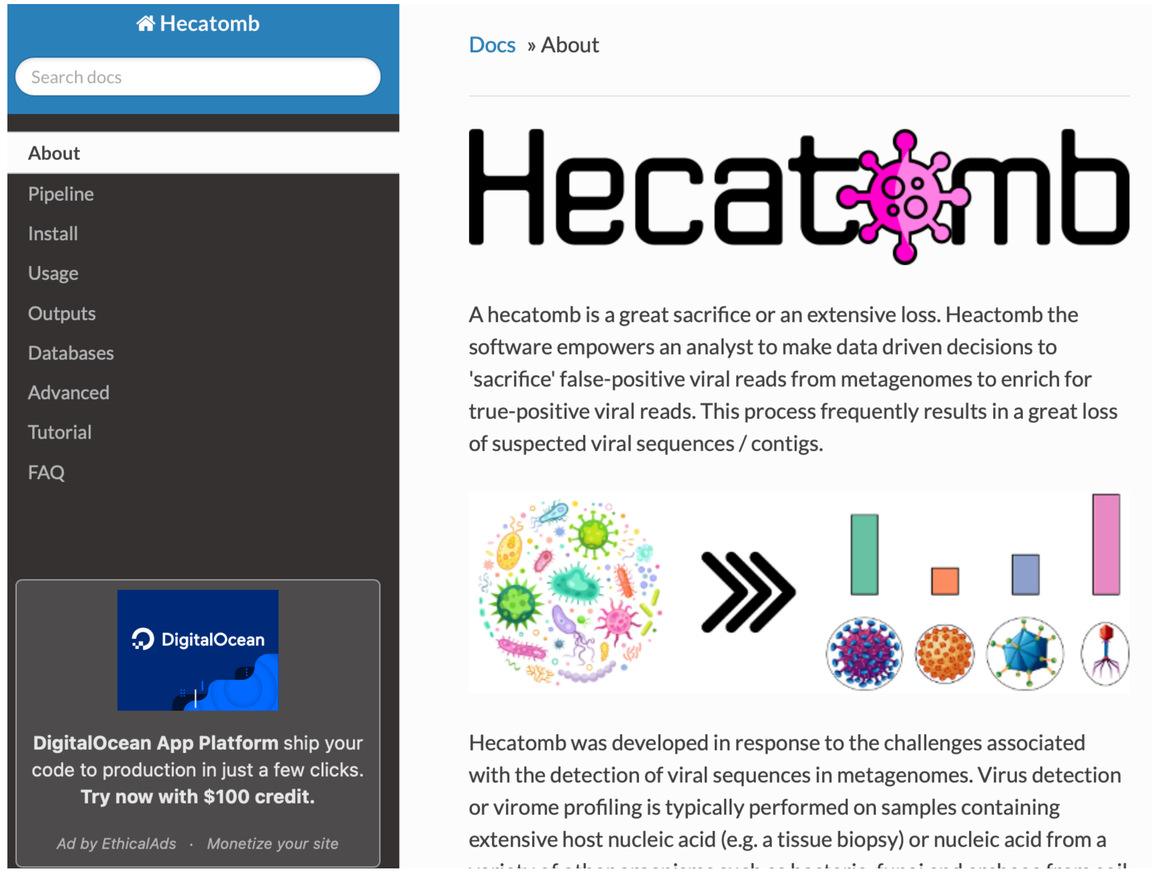
Families with evidence of high-quality alignments




A)
Picornaviridae
Parvoviridae
Reoviridae
Astroviridae
Anelloviridae
Adenoviridae
Caliciviridae
% ID
100
80
60
40
100
200
300
400
500
600
Alignment Length
Type
dsDNA
dsRNA
ssDNA
B)
Number per Quadrant
100000
75000
50000
25000
0
Adenoviridae
Anelloviridae
Astroviridae
Caliciviridae
Parvoviridae
Picornaviridae
Reoviridae
Family
Quadrant
Q1
Q2
Q3
Q4
C)
% Reads per Quadrant
30
10
3
Q1
Q2
Q3
Q4
Quadrant
Family
Adenoviridae
Anelloviridae
Astroviridae
Caliciviridae
Parvoviridae
Picornaviridae
Reoviridae

Summary

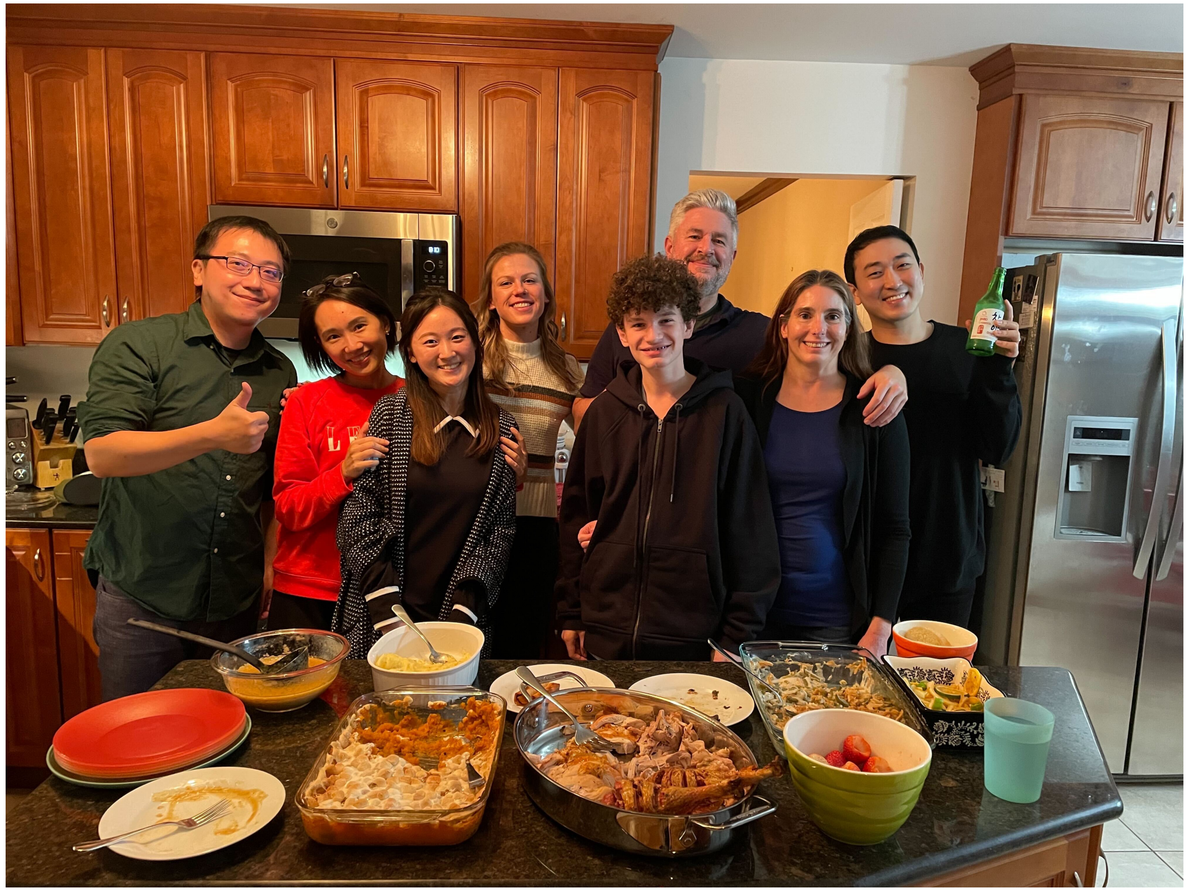
- Virome analysis has a number of unique challenges
  - Sample types
  - Viral genomic structure
  - Mosaicism
  - False-positives
- Hecatomb provides a rigorous taxonomic assignment framework and data integration to enable strategic decision making and ‘true positive’ viral selections



https://hecatomb.readthedocs.io/en/latest/

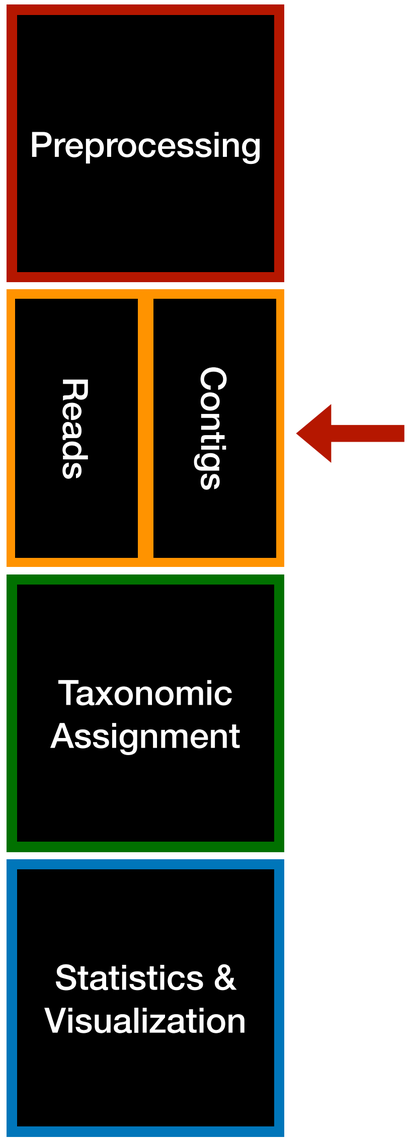
Acknowledgements

- **Washington University**
  - Dave Wang
  - Mike Diamond
- **Flinders University**
  - Rob Edwards
  - Liz Dinsdale
  - Mike Roach
- **San Diego State University**
  - Anca Segall



Reads vs. Contigs

Reads vs. Contigs



Reads

- Less informative
- Useful for detection
  - You will not obtain a contig for every virus
- Useful for determining if you *believe* a virus is present
  - Query statistic evaluation
- Rule of thumb: Most useful for eukaryotic virus detection / analysis

Contigs

- More informative
- Useful for population analysis
  - You don’t need to obtain a contig for every virus
- Genes, ORFs, etc.
- Rule of thumb: Most useful for bacteriophage analysis

Reads + Contigs

Preprocessing

Reads | Contigs

Taxonomic Assignment

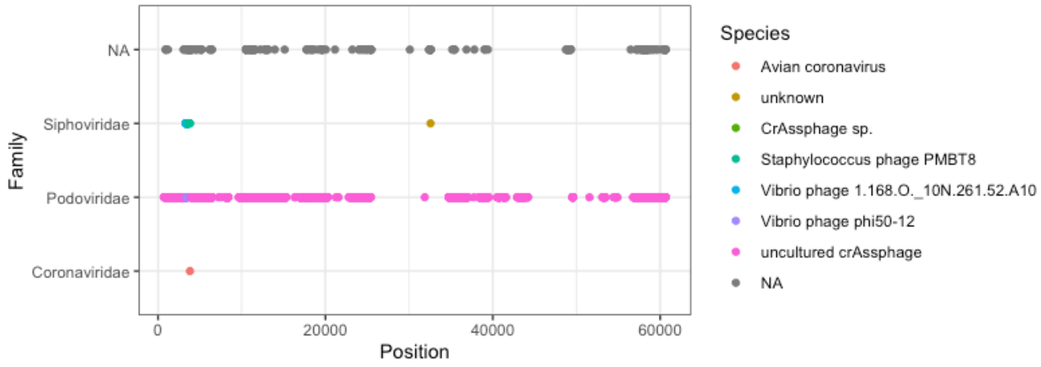
Statistics & Visualization

Viral taxon table

id	sequence	Kingdom	Phylum	Class	Order	Family	Genus	Species
0	atgcagc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cardiovirus	Cardiovirus A
1	ccatgcc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	Human cosavirus E/D
2	aatctaa	Virus	Preplasmiviricota	Tectiliviricetes	Rowavirales	Adenoviridae	Mastadenovirus	Bat mastadenovirus A

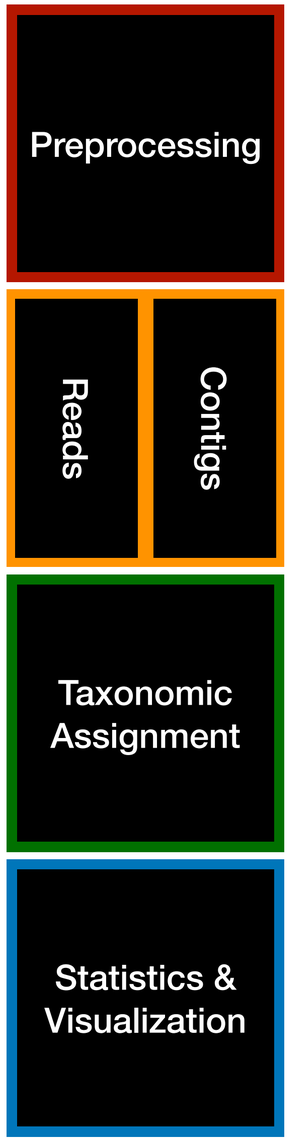
Contig Information

id	contig_id	Lineage	Start	Stop	Length	Quality
0	345	K,P,C,O,F,G,S	25	47	22	35
1	345	K,P,C,O,F,G,S	34	124	90	37
2	1567	K,P,C,O,F,G,S	2	98	96	4



- Contig confirmation of read-based taxonomic assignment
- Recursive read-based taxonomic assignment basic on contig annotation
- Built-in contig annotation

Running Hecatomb



- https://github.com/shandley/hecatomb
  - More information on the wiki: https://github.com/shandley/hecatomb/wiki
- Dependencies
  - Snakemake
  - Conda
  - R
  - RStudio (technically not necessary, but very helpful)
- Run
  - snakemake --snakefile ./Snakefile --configfile ../config/my_config.yaml --resources mem_mb=100000 --cores 64 --use-conda --conda-frontend mamba

config.yaml

```yaml
Paths:
  # The base database directory
  # You can install them using download_databases.snakefile

  Databases: /mnt/data1/databases/hecatomb

  # The reads directory has your input fastq files
  # Note: All of your results will go into this directory

  Reads: ../../test_data

  # Where do you want the results stored?
  # Recommended that you make a specific based dir (e.g. heactomb_runs) followed by project specific (e.g.
  test_data) for all of your runs
  #This should not be a subdirectory to where your Reads are located!

  Results: test_data_results

  # Host directory name
  # e.g. human, mouse, dog, etc.
  # Needs to be the name of the directory containing the masked reference
  # If your reference is not available post an issue on GitHub requesting it to be added

  Host: macaque

  # Temp is a temporary directory. By default we make
  # subdirectories in here for each application

  Temp: .tmp

System:
  # How much memory you want to allocate to java (required for bbtools steps)

  # This is in gigabytes of memory (e.g 2GB would use 2, 128GB would use 128)

  Memory: 100

  # Number of threads to use

  Threads: 64

############################
# Optional Rule Parameters #
############################
```

- Centralized file for all system and run specific options
  - Database directory
  - Read (input) directory
  - Results directory
  - Host name
  - Memory
  - Threads
  - Other options
    - QC threshold
    - E-value thresholds