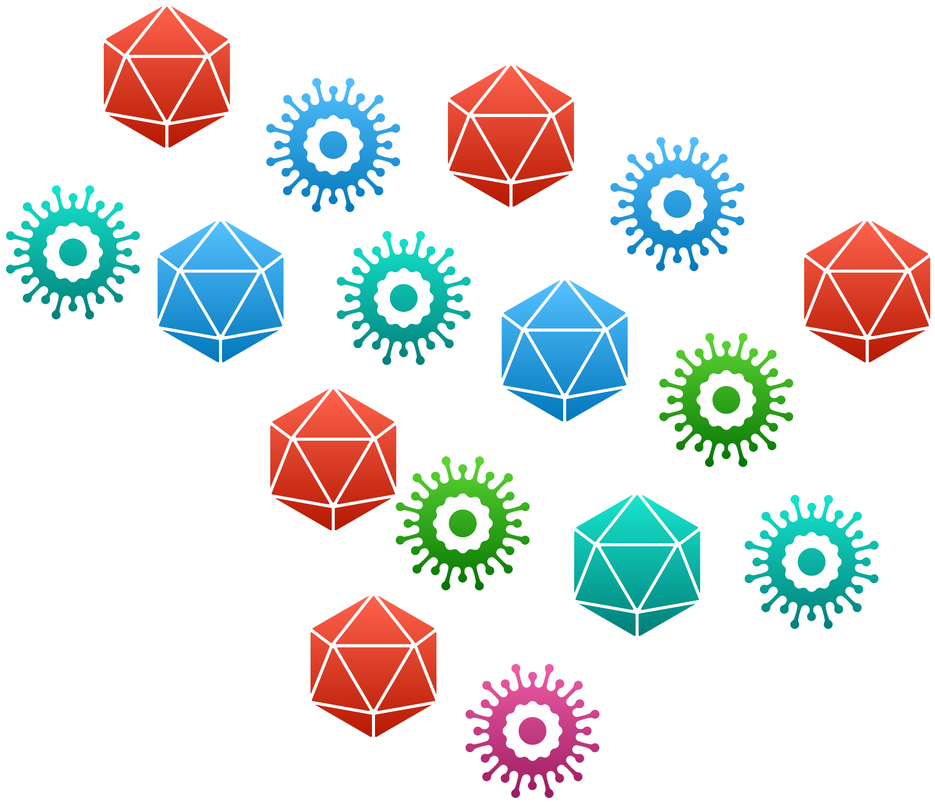
Virome Analysis Challenges

https://github.com/shandley/hecatomb

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

Overview

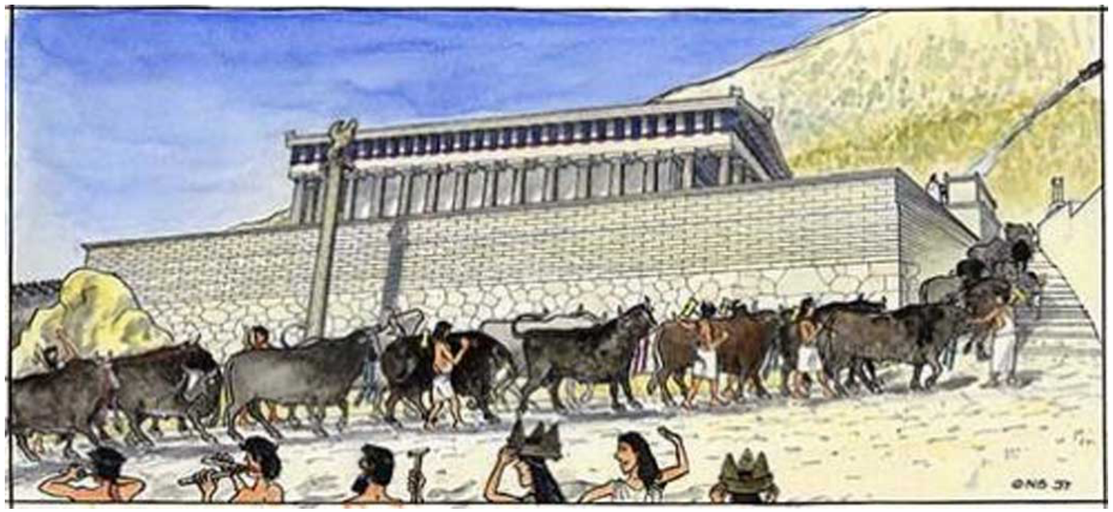
- History
- What is hecatomb?
- What problems is hecatomb designed to address?
- How hecatomb works
  - Virus sequence counts
  - Virus taxonomic assignment
- Data integration
- Reads vs. Contigs
- Running Hecatomb
- Examples



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!

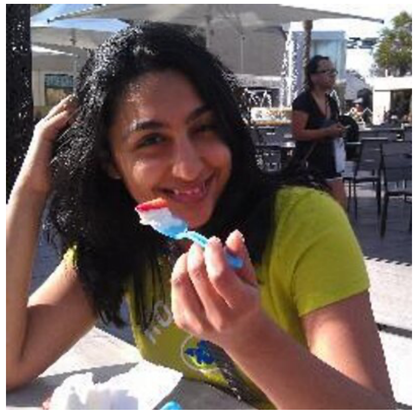
Development

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 is hecatomb?

Is this sequence viral?

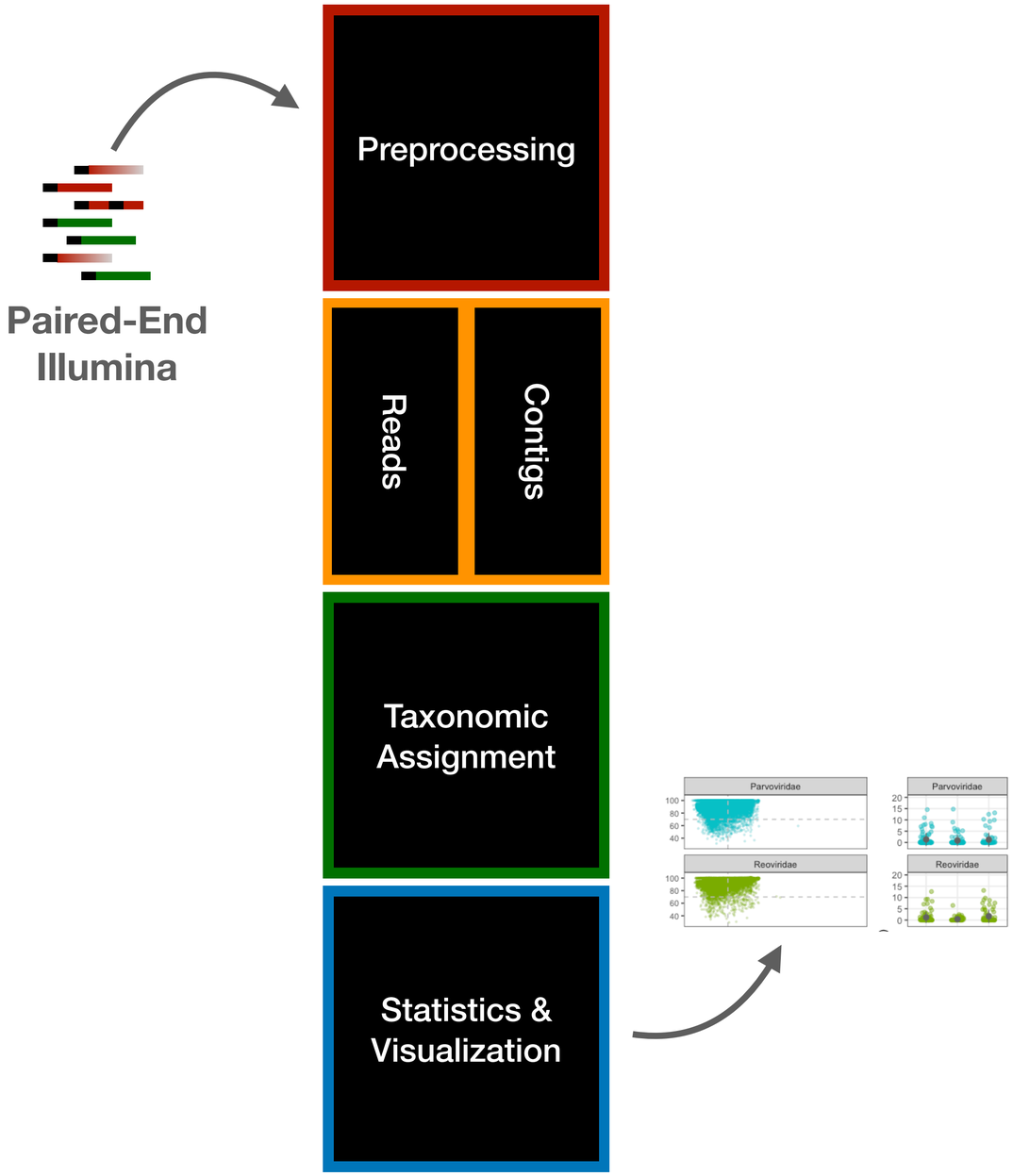
ATCAGCATCGTGATCGTGACTAGCTTACGTACGTA

ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
ATCAGCATCGTGATCGTGACTAGCTTACGTACGTA
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
AGCTACTGTCGTATGTTTTATCGATCTGTACGTTTCAG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
AAACATCTAGCGATGCTGTACTGCTTACGTAGTCTATC
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
CCCAGCTGATGCAGCTTCATCGTAGCTCATGCTGAC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG


ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
Adenovirus
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
ATCAGCATCGTGATCGTGACTAGCTTACGTACGTA
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
AGCTACTGTCGTATGTTTTATCGATCTGTACGTTTCAG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
AAACATCTAGCGATGCTGTACTGCTTACGTAGTCTATC
Picornavirus
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
CCCAGCTGATGCAGCTTCATCGTAGCTCATGCTGAC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
ATCAGCATCGTGATTCAAACTAGCTTACGTACGTA
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
Coronavirus
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG
TTTTACGATCTAGCTACTTGCATGCTTGAGCC
GGGAGTCGTAGCTTACTGTCATCGTATCGATCGATGCT
CTAGATCGTAGCTGCTGTCATGTAGCTAGCTGCTACG

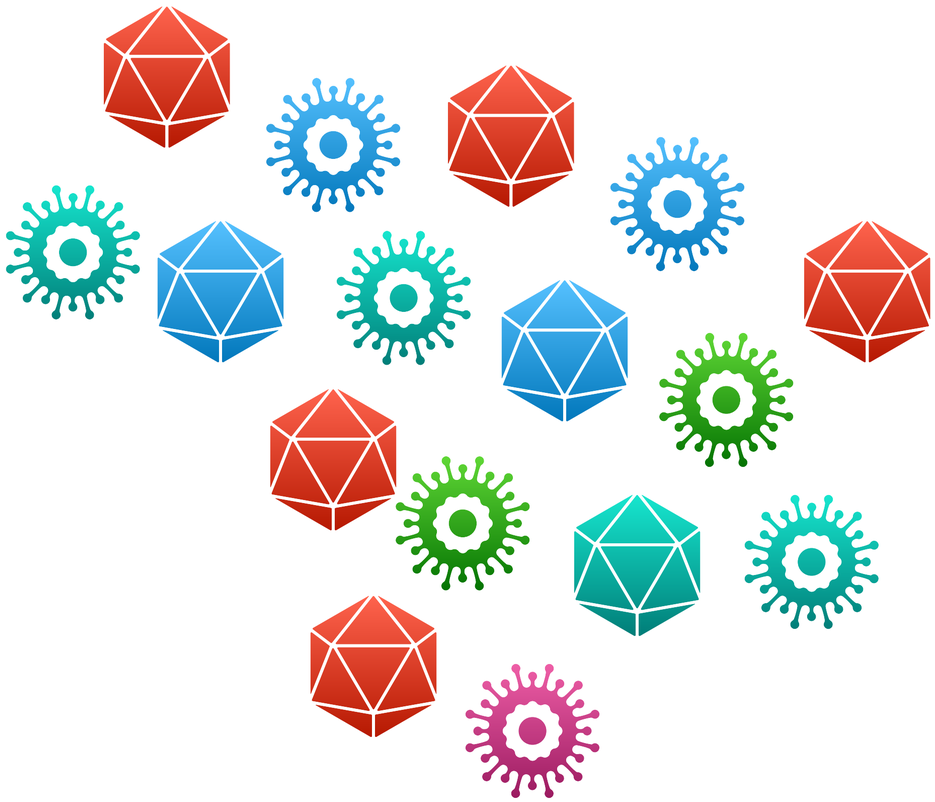
What hecatomb *is*?

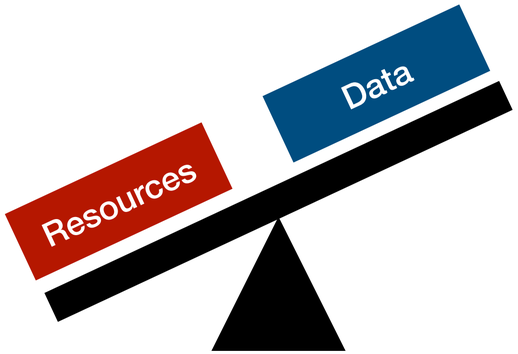
- Broadly:
  - Virome analysis software
- Specifically:
  - Computational workflow to detect and annotate viral sequences from metagenomic sequences
  - Can detect and analyze both phage and eukaryotic viral sequences
  - Works on individual reads and contigs
  - Integrates taxonomy, counts, sample data 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***

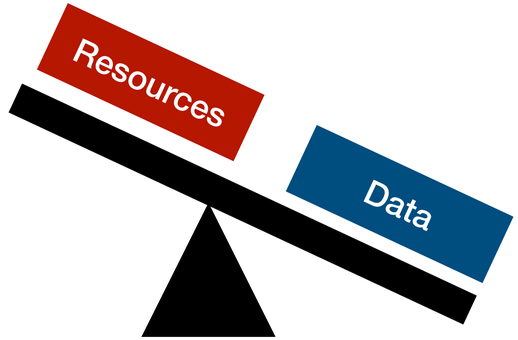




Low quality data
Low-sensitivity search
Poor search strategy
Information poor (small) databases
Compute = small

Lots of reads

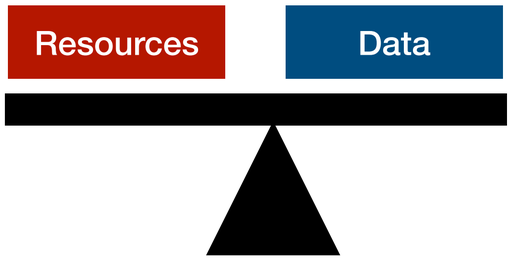
Most '*homebrew*' approaches



High-quality data
Deep search
Complex search strategy
Extensive (large) databases
Compute = Large

Poor data integration
Few investigator tools

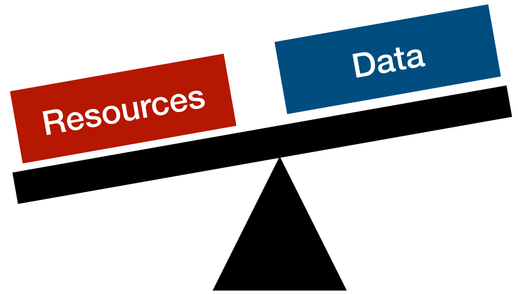
- VirusSeeker
- IdSeq



High-quality data
Deep search
Complex search strategy
Extensive (large) databases
Compute = Large

Deep data integration
Extensive investigator tools

Theoretical optimal approach



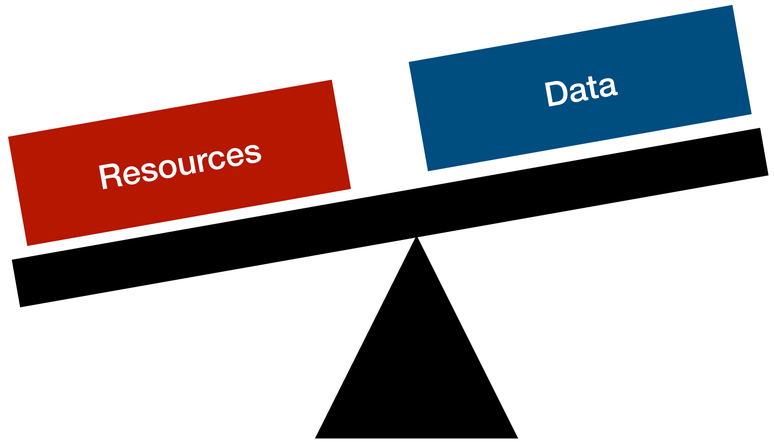
High-quality data
Deep search
Moderate databases
Compute = Medium

Deep data integration
Extensive investigator tools

Hecatomb

Hecatomb Philosophy

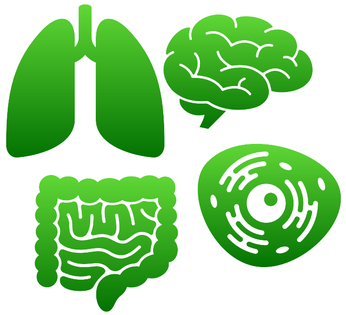
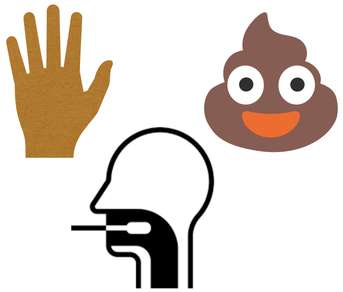
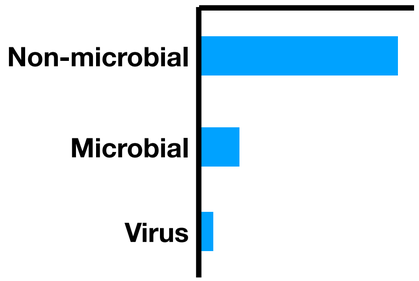
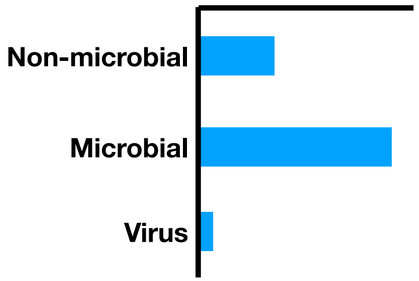
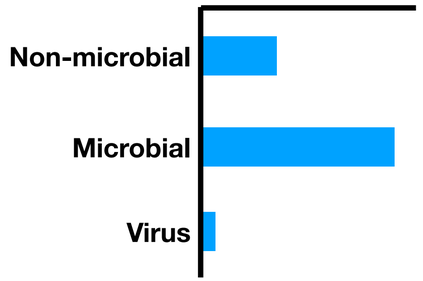
- There is no ***perfect*** search strategy or ***perfect*** reference database
    - There will always be true/false positives and negatives
- Hecatomb's approach is to do a ***good enough*** (quite good actually) job at data cleaning, assembly, taxonomic assignment
    - Does not require:
        - monolithic databases (nr/nt)
        - Super computing
- Instead, hecatomb is designed to provide maximum information for investigator decision making, statistics and visualization
    - This off-sets the need for perfect/exhaustive resources



What challenges is hecatomb designed to address?

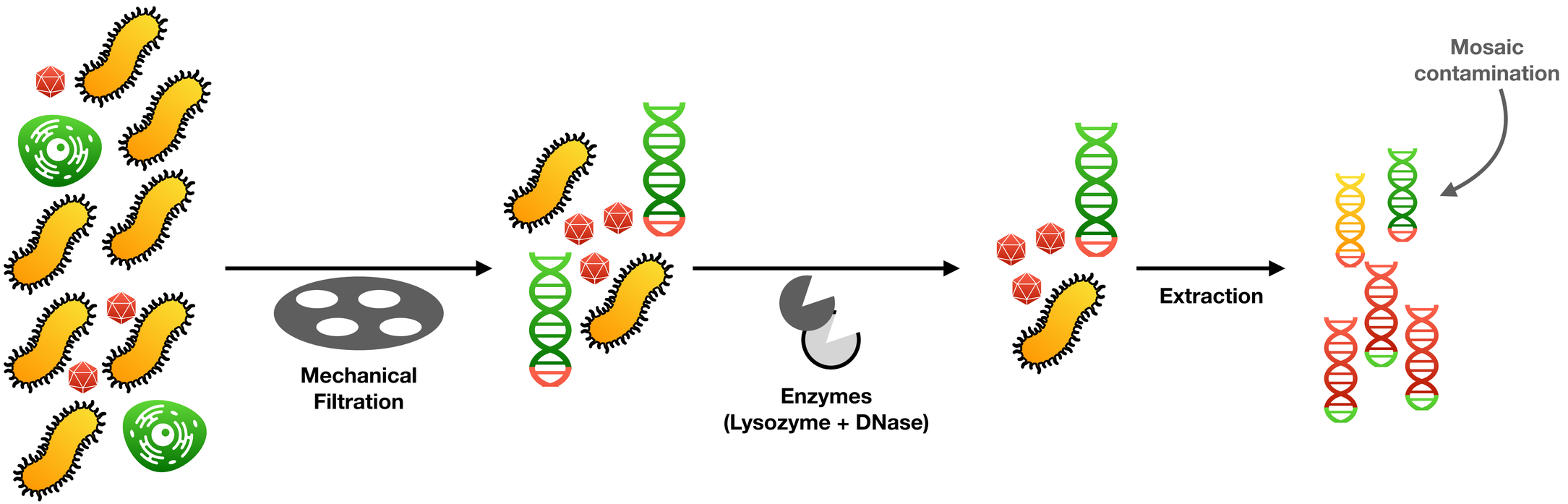
1) Sample types
2) Contamination
3) Genetic mosaicism
4) Viral genome complexity

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

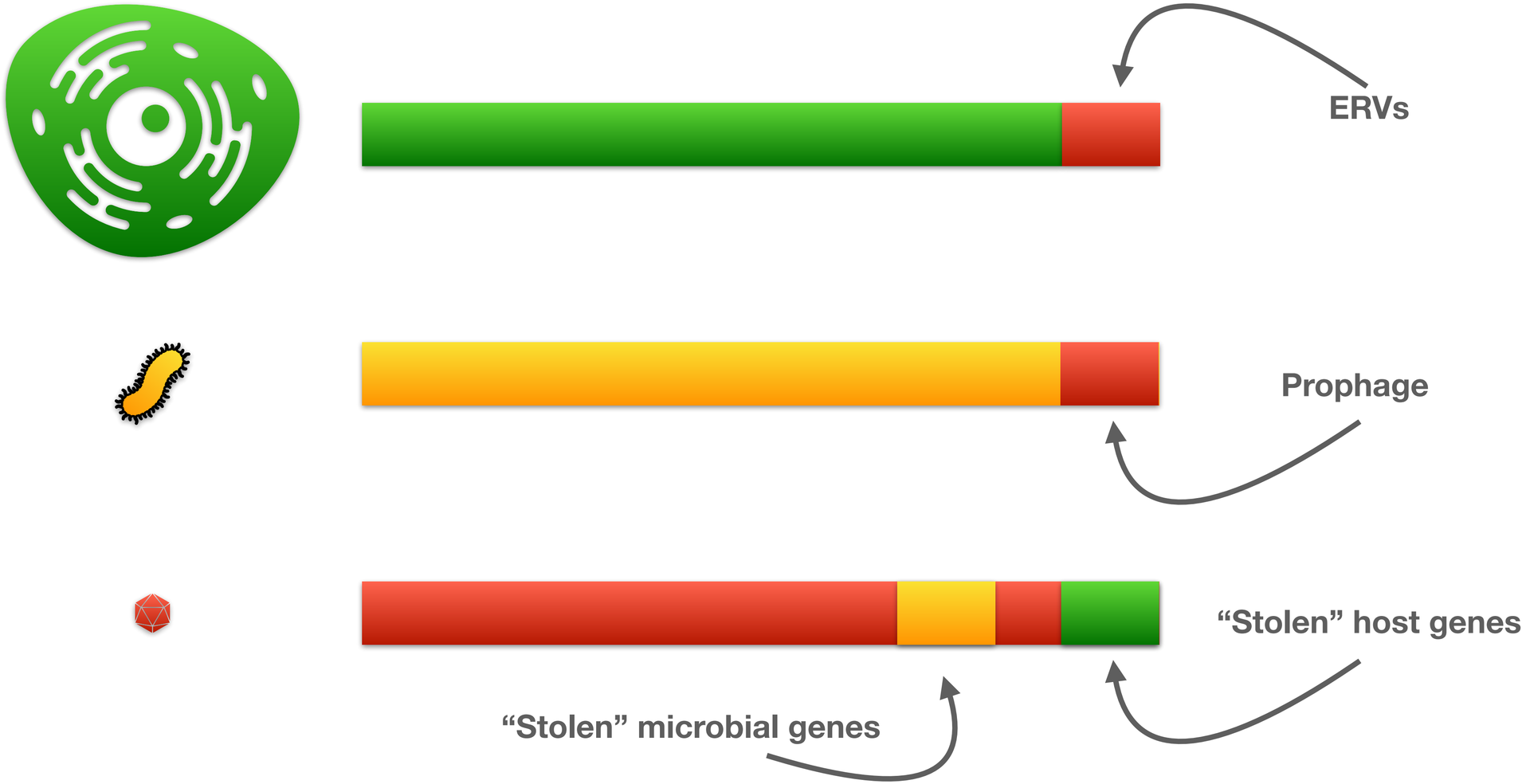
Challenge 2: Contamination

Enrichment ≠ Purification



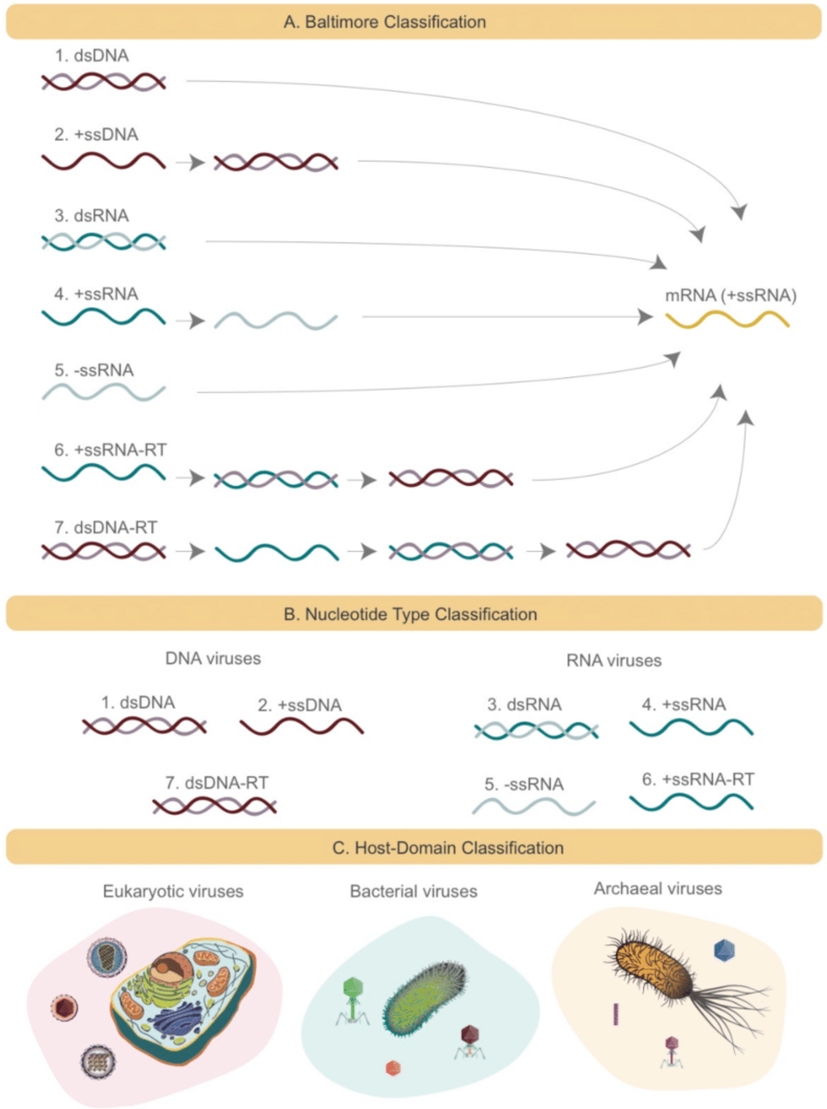
Challenge 3: Genetic Mosaicism

Genetic Mosaicism

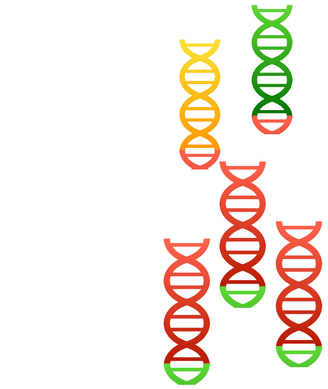


Challenge 4: Viral Genome Complexity

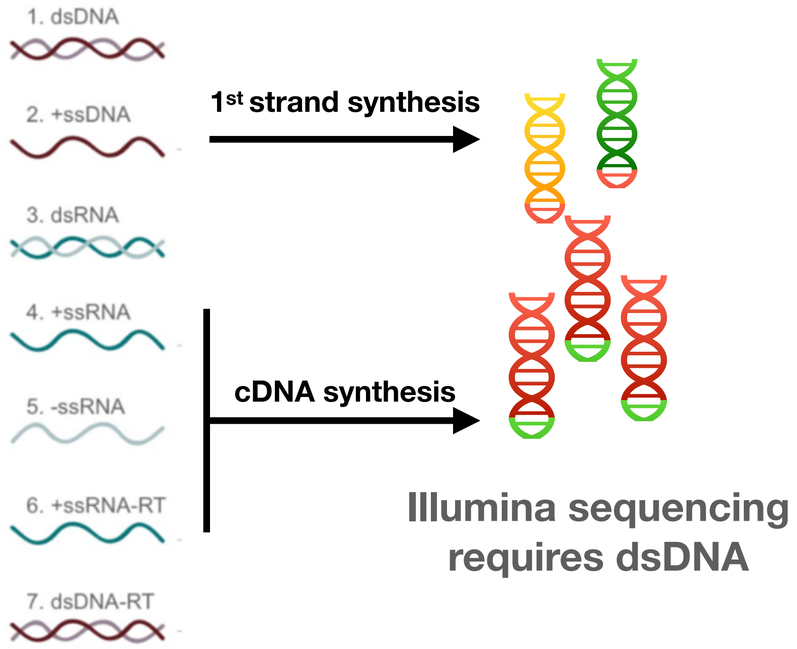
Viral Genome Architectures



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



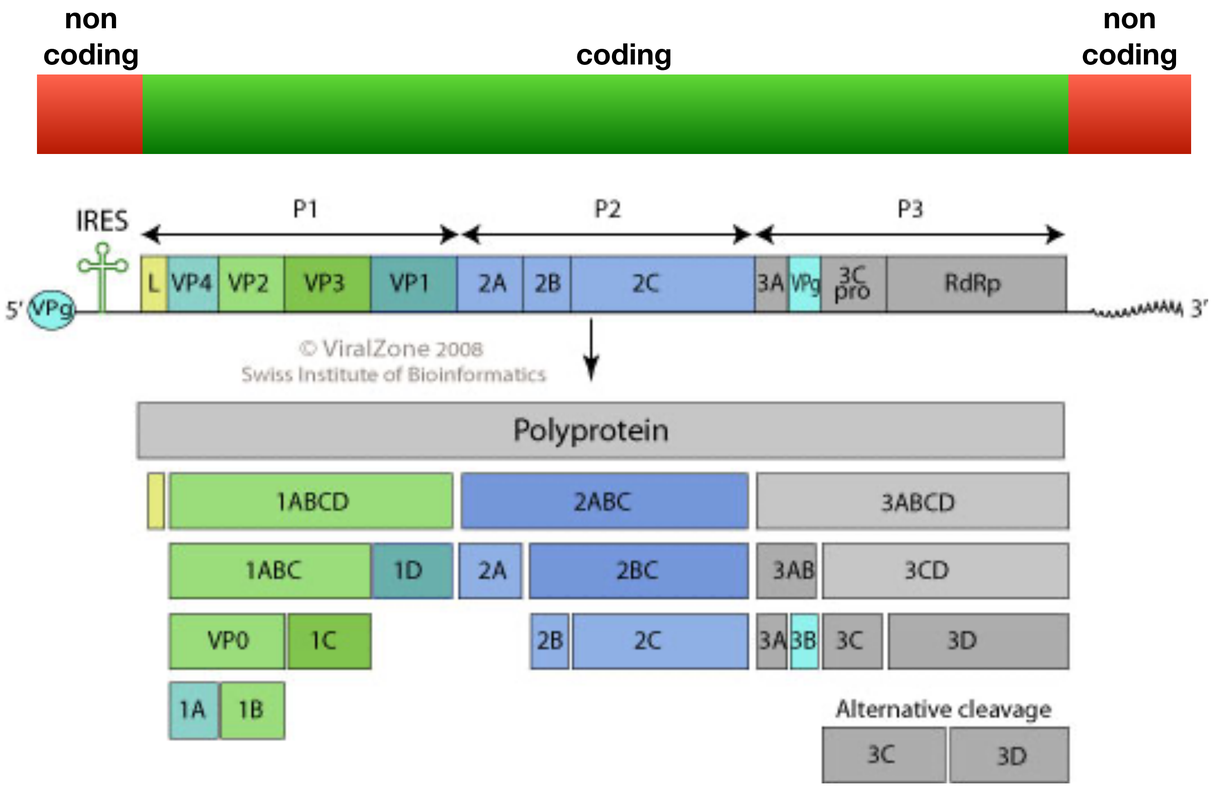
We don't actually start with this



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

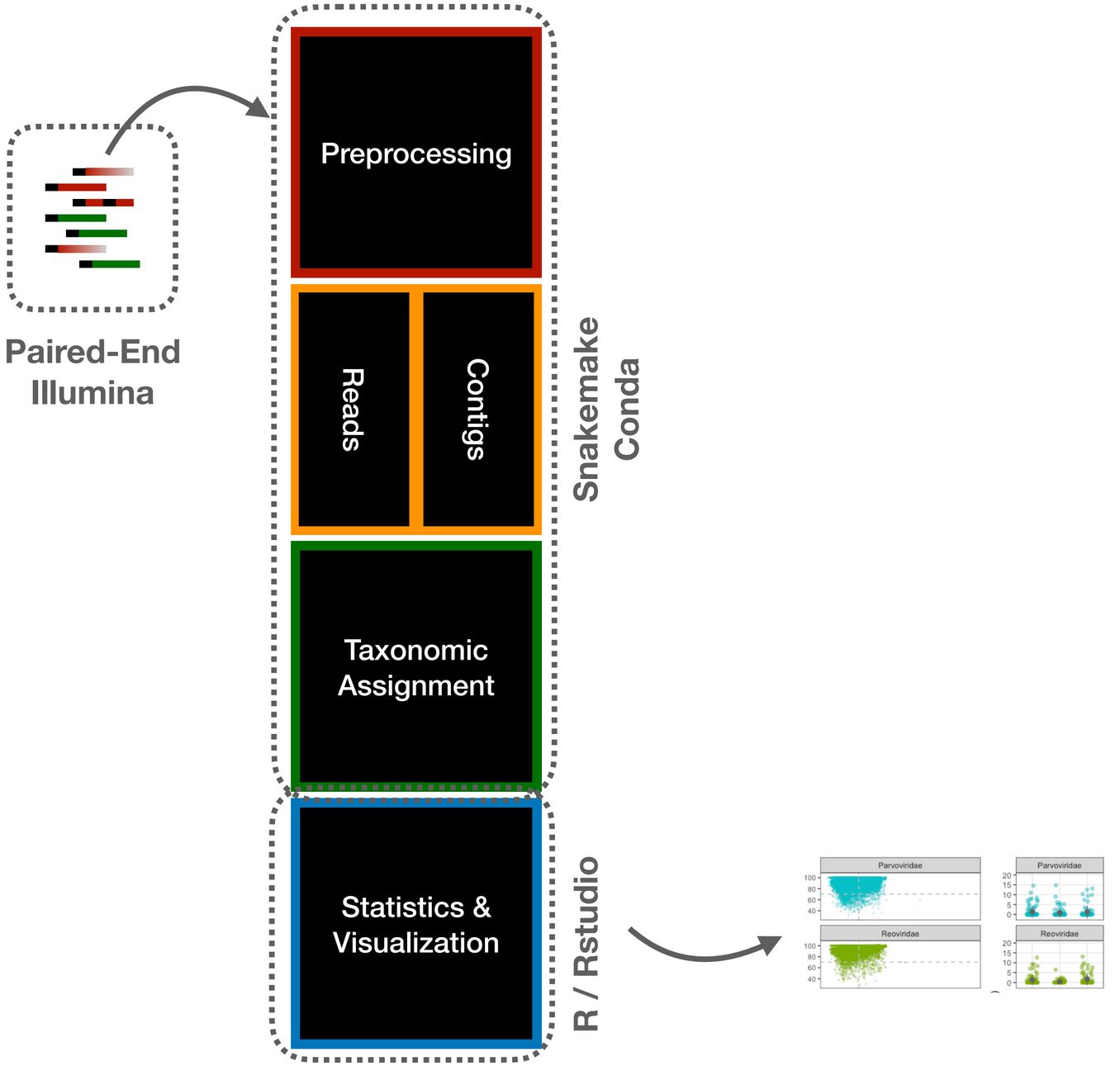
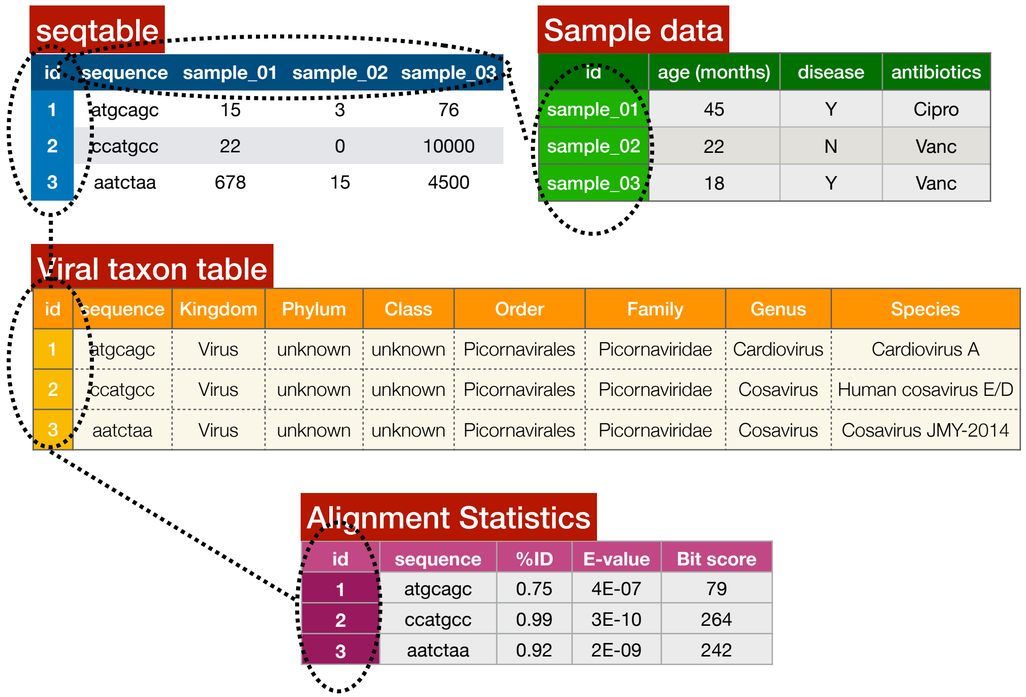

Paired-End
Illumina
Preprocessing
Reads
Contigs
Taxonomic
Assignment
Snakemake
Conda
Statistics &
Visualization
R / Rstudio
Parvoviridae
Reoviridae
Parvoviridae
Reoviridae

Table Building

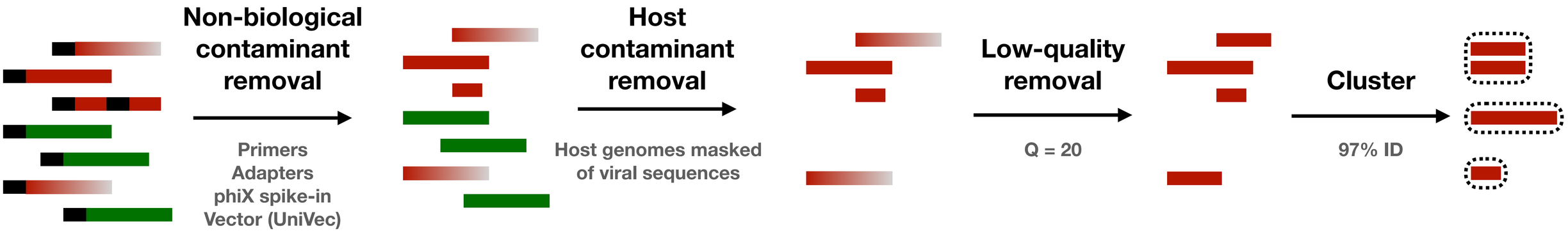


Sequence Table (*seqtable*)

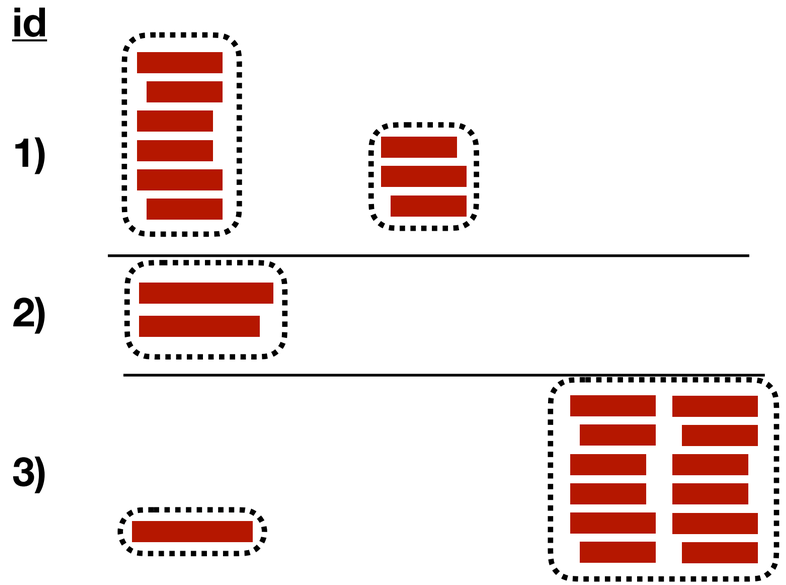
- Every row is a unique sequence
- Every column is a unique sample
- Each cell is the occurrence of each sequence in each sample
- Created by clustering quality-controlled sequences and counting the size of each cluster

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

Quality Control



Clustering & Counting



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

- 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

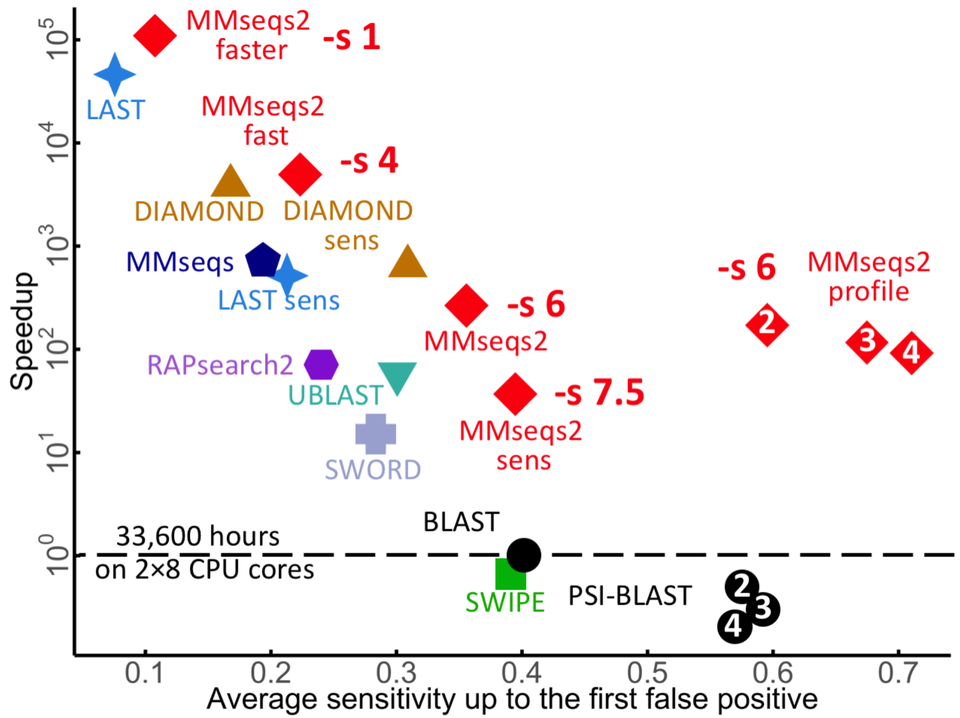
id	sequence
1	atgcagc
2	ccatgcc
3	aatctaa

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

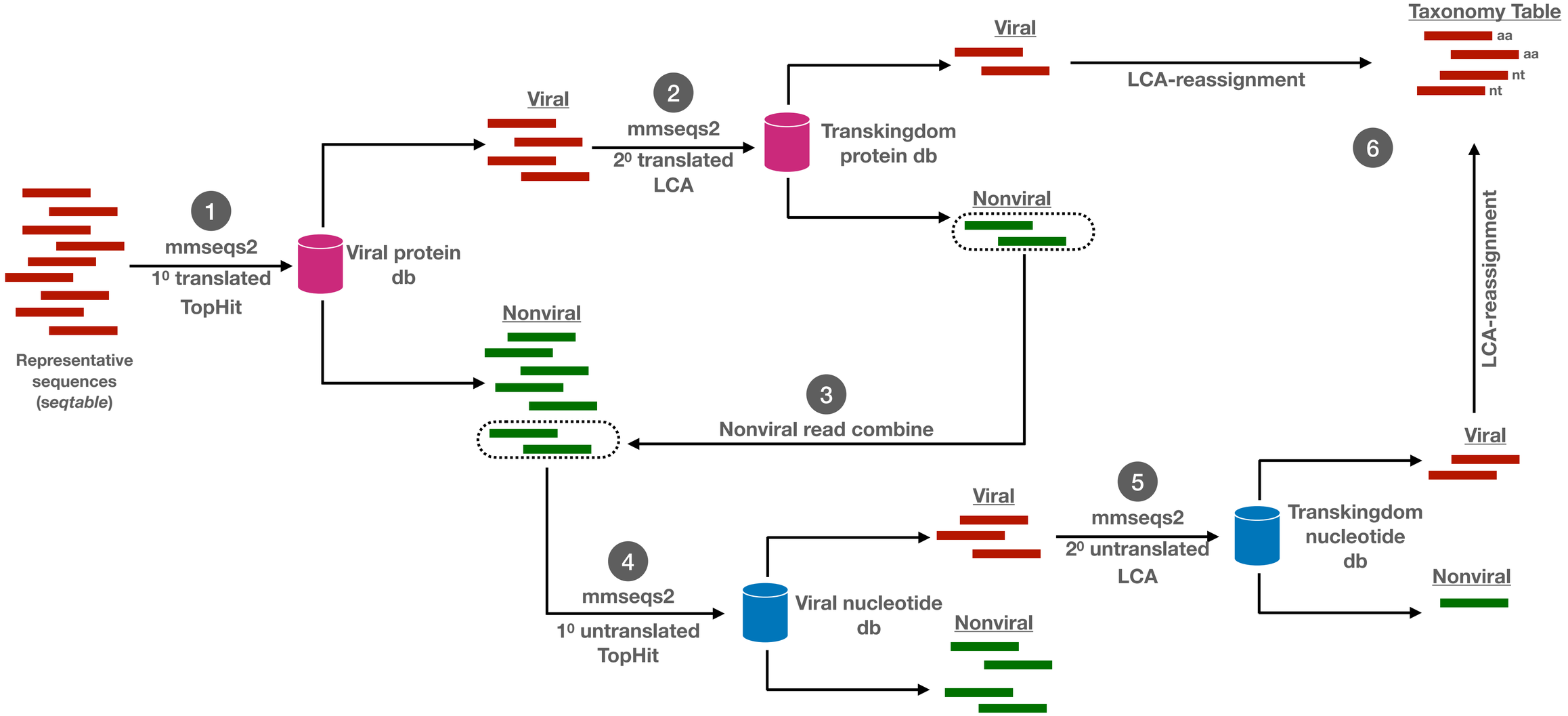
- Fast query of sequences against reference databases
- Untranslated (blastn) and translated (blastx) searching
- Integrated taxonomy modules
  - Multiple lowest common ancestor (LCA) algorithms
- Hecatomb uses iterative searching from -s 1 to -s 7 (see this link for thorough explanation)



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.

Taxonomic Assignment



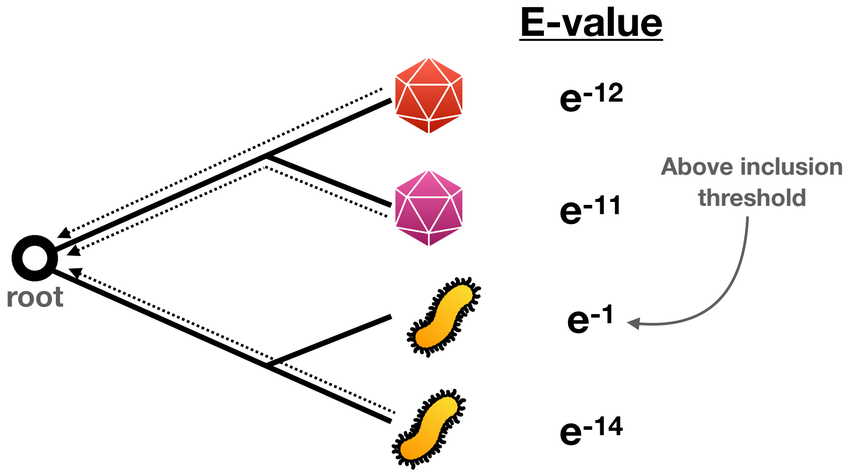
Extended Taxonomy Table

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

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

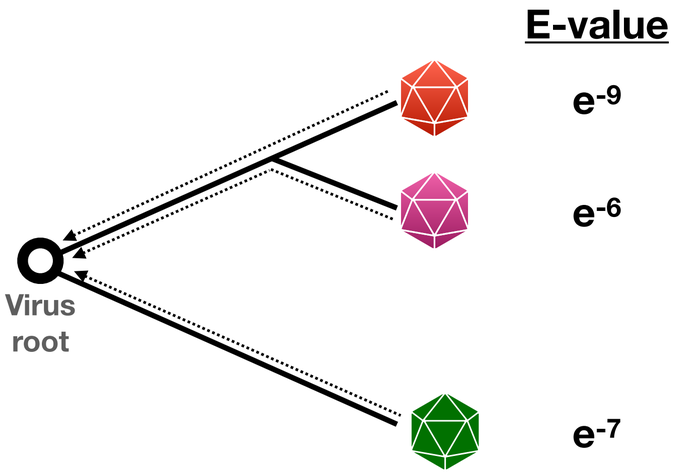
LCA Challenges

Transkingdom Rooting



- Transkingdom LCA will assign '*root*' taxonomy
- Requires tree and some inclusion threshold

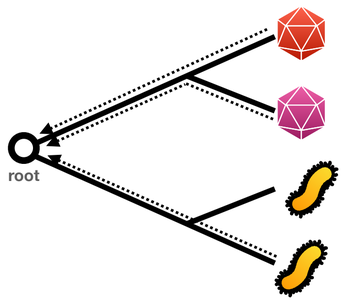
Virus Rooting

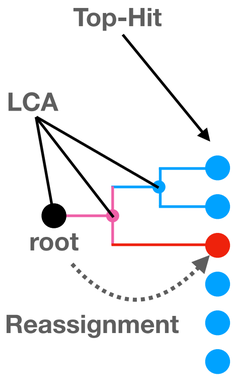


- Transviral LCA will assign 'virus *root*' taxonomy
- Requires tree and some inclusion threshold

Transkingdom Rooting Example

Transkingdom Rooting





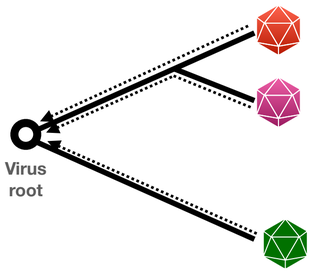
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
hypothetical protein [Lachnospiraceae bacterium]	Lachnospiraceae bacterium	60.8	60.8	99%	4e-09	44.59%	189	MBE5925732.1
hypothetical protein [Holdemanella biformis]	Holdemanella biformis	60.5	60.5	95%	6e-09	49.30%	190	MBD9053365.1
hypothetical protein [Siphoviridae sp.]	Siphoviridae sp.	58.2	58.2	95%	4e-08	41.43%	193	AXF52477.1
hypothetical protein [Firmicutes bacterium]	Firmicutes bacterium	55.8	55.8	74%	3e-07	51.72%	201	MBE6148201.1
TPA: hypothetical protein [Eubacterium sp.]	Eubacterium sp.	54.3	54.3	63%	9e-07	59.57%	180	HAH18714.1
hypothetical protein [Clostridiales bacterium]	Clostridiales bacterium	47.4	47.4	62%	3e-04	47.83%	160	NLV87667.1

LCA inclusion Threshold (adjustable)

- Hecatomb reformats trans kingdom LCA rooting by assigning the TopHit Virus instead
  - In this example, instead of '*root*' it would assign *Siphoviridae sp.* (TopHit Virus)
  - *Flags* the sequence as being TopHit reassigned
- Can we use this information for host prediction?

Virus Rooting Example

Virus Rooting



Top-Hit Classification

```
107483  tr|Q65YV7|Q65YV7_9VIRU Holin OS=Lactobacillus phage phigaY OX=272757 GN=hol PE=4 SV=1
```

LCA Classification

```
107483  10239   superkingdom    Viruses Viruses;uc_Viruses;uc_Viruses;uc_Viruses;uc_Viruses;uc_Viruses;uc_Viruses
```

Organism	Blast Name	Score	Number of Hits	Description
Viruses	viruses		14	
. Caudovirales	viruses		13	
. . unclassified Myoviridae	viruses		6	
. . . Lactobacillus phage phi jlb1	viruses	71.2	2	Lactobacillus phage phi jlb1 hits
. . . Lactobacillus phage KC5a	viruses	65.9	2	Lactobacillus phage KC5a hits
. . . Lactobacillus prophage Lj771	viruses	60.8	2	Lactobacillus prophage Lj771 hits
. . Lactobacillus prophage Lj928	viruses	53.9	4	Lactobacillus prophage Lj928 hits
. . Lactobacillus phage JNU_P11	viruses	48.1	1	Lactobacillus phage JNU_P11 hits
. . Lactobacillus prophage Lj965	viruses	37.0	2	Lactobacillus prophage Lj965 hits
. Lactobacillus phage phigaY	viruses	70.9	1	Lactobacillus phage phigaY hits

- Hecatomb reformats LCA virus-rooting by assigning the TopHit Virus instead
  - In this example, instead of '*uc_Viruses*' it would assign *Lactobacillus phage phigaY.* (TopHit Virus)
  - *Flags* the sequence as being TopHit reassigned?

Taxon Summary

id	sequenc	Kingdom	Phylum	Class	Order	Family	Genus	Species	query_type	Classification	sample_01	sample_02
1	atgcagc	Virus	Pisuviricota	Pisoniviricete	Picornavirale	Picornaviridae	Cardiovirus	**Cardiovirus A**	translated	LCA	0	10
2	ccatgcc	Virus	Pisuviricota	Pisoniviricete	Picornavirale	Picornaviridae	Cosavirus	**Human cosavirus E/D**	translated	LCA-corrected	2	5
3	aatctaa	Virus	Pisuviricota	Pisoniviricete	Picornavirale	Picornaviridae	Cosavirus	**Cosavirus JMY-2014**	untranslated	LCA	100	25
4	gatctta	Virus	Pisuviricota	Pisoniviricete	Picornavirale	Picornaviridae	Cosavirus	**Cosavirus JMY-2014**	untranslated	LCA	1500	10
5	gctctag	Virus	Pisuviricota	Pisoniviricete	Picornavirale	Picornaviridae	Cosavirus	**Cosavirus JMY-2014**	untranslated	LCA	800	40

↓ **Species Counts**

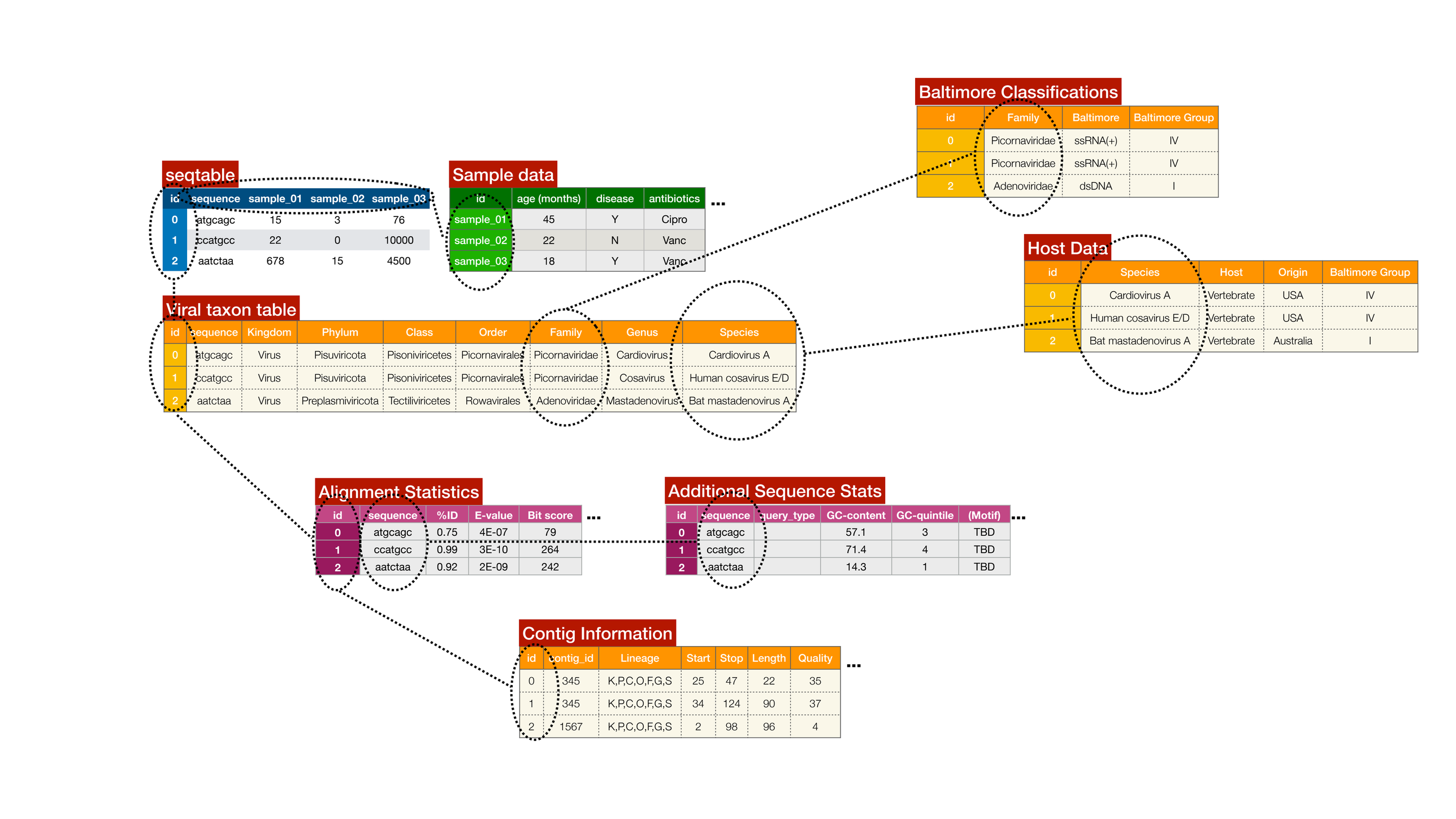
id	sequenc	Kingdom	Phylum	Class	Order	Family	Genus	Species	query_type	Classification	sample_01	sample_02
1	atgcagc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	**Cardiovirus**	Cardiovirus A	translated	LCA	0	10
2	ccatgcc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	**Cosavirus**	Human cosavirus E/D	translated	LCA-corrected	2	5
3	aatctaa	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	**Cosavirus**	Cosavirus JMY-2014	untranslated	LCA	2400	75

↓ **Genus Counts**

id	sequence	Kingdom	Phylum	Class	Order	Family	Genus	query_type	Classification	sample_01	sample_02
1	atgcagc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cardiovirus	translated	LCA	0	10
2	ccatgcc	Virus	Pisuviricota	Pisoniviricetes	Picornavirales	Picornaviridae	Cosavirus	translated	LCA-corrected	2402	80

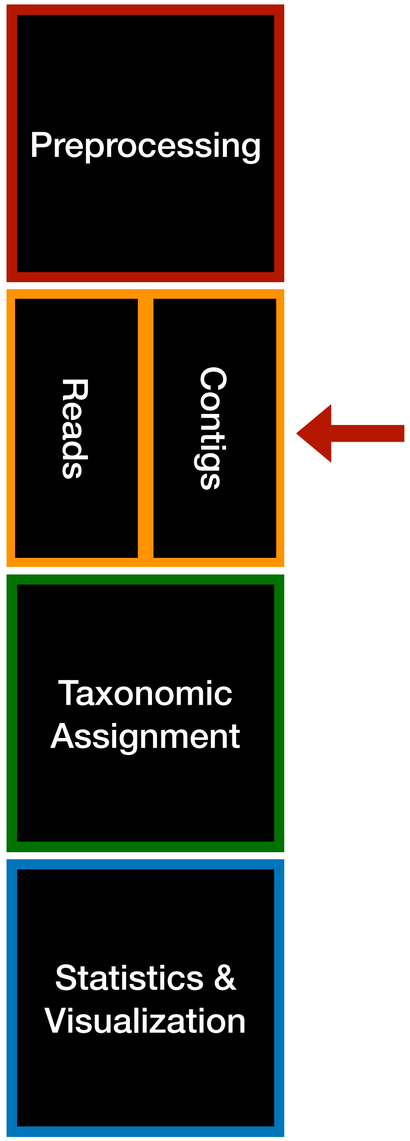
↓ **So on …**

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

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

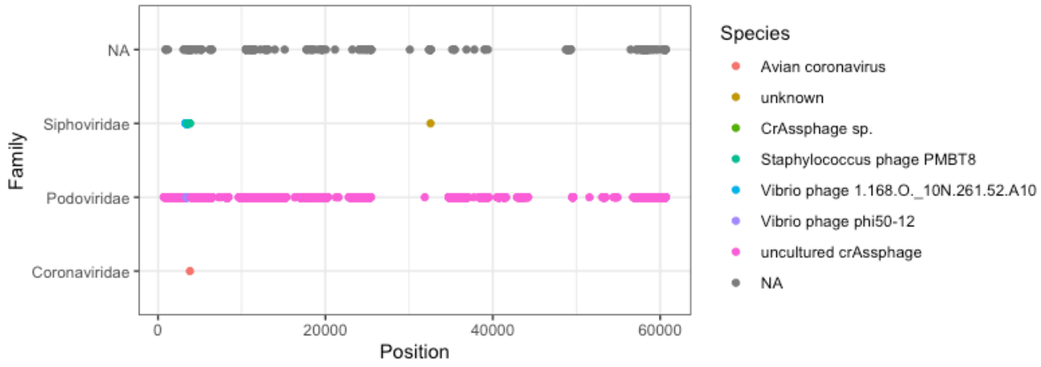
Statistics & Visualization

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

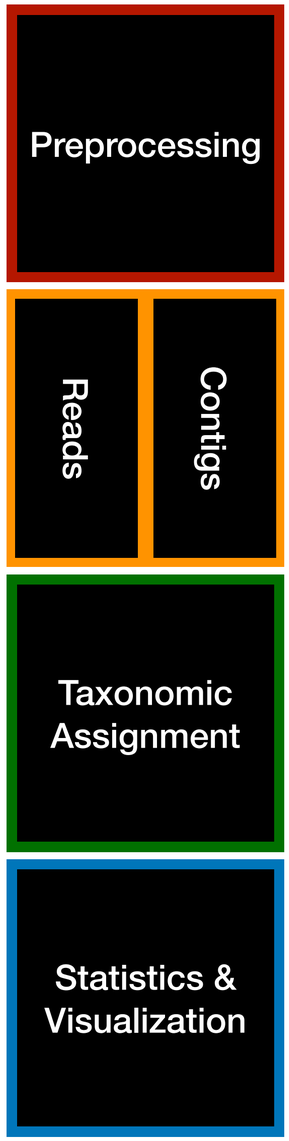
Contig Information

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



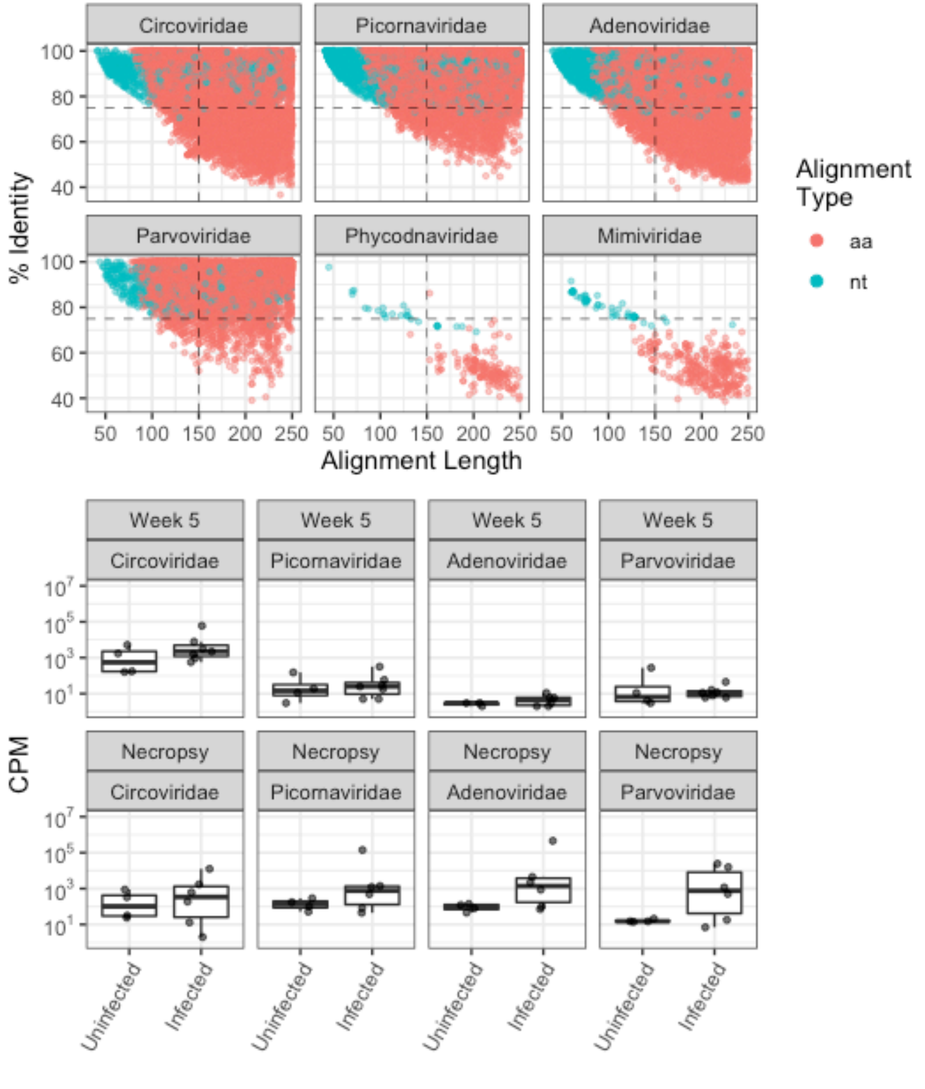
- Prepublication: https://www.biorxiv.org/content/10.1101/2022.05.15.492003v2
- Download: https://github.com/shandley/hecatomb
  - Read the docs: https://hecatomb.readthedocs.io/en/latest/
- Dependencies
  - Snakemake
  - Conda
  - R / RStudio (technically not necessary, but very helpful)

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

Example Analysis



Example Analysis