Workshop III – Bioinformatics and Viral Genomics

Untargeted Viromics Session

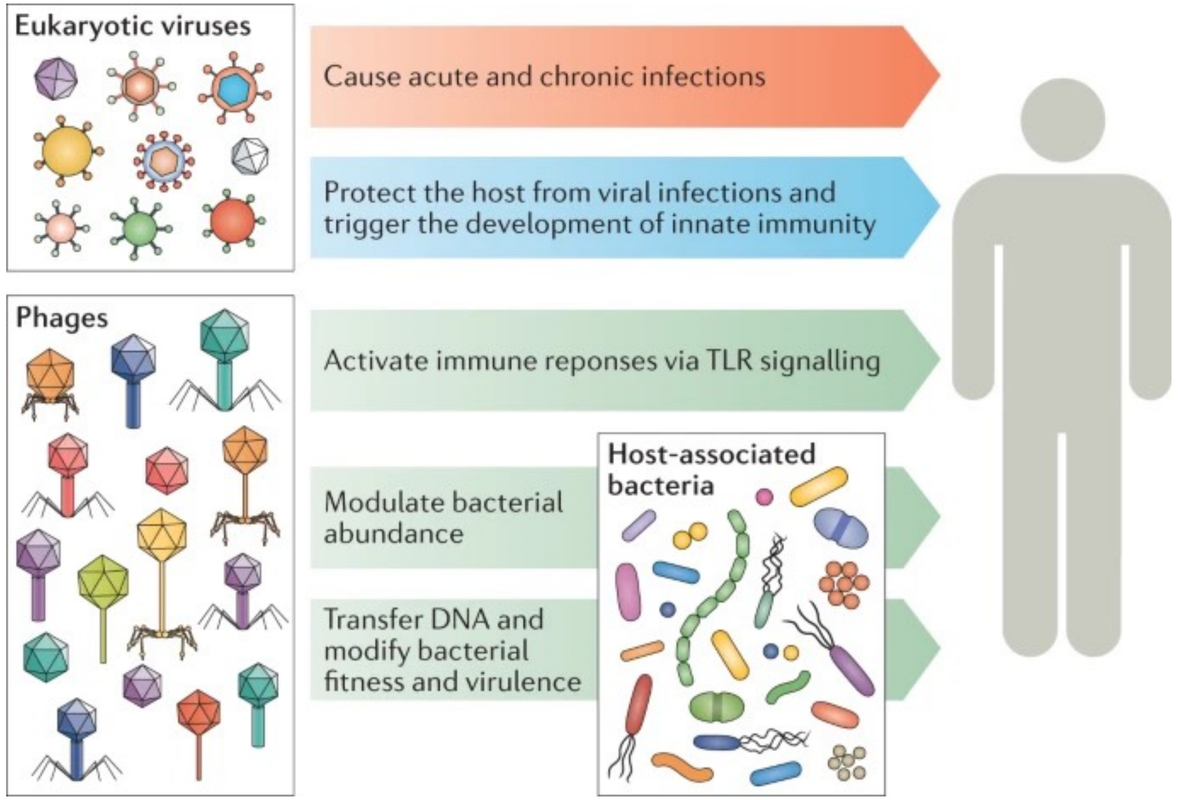
Durban, South Africa
February 27, 2025
Luis Chica

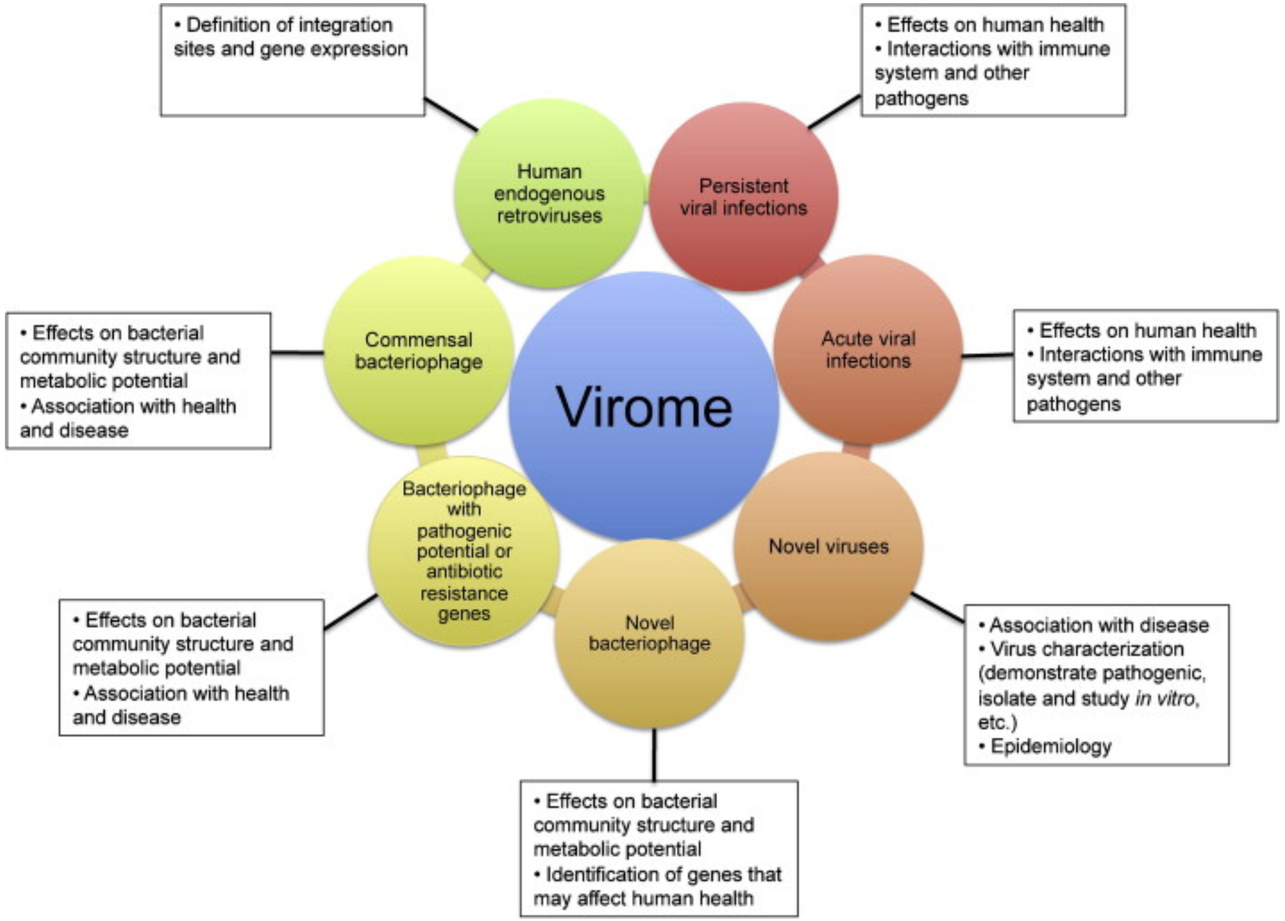
Outline

- Introductory Presentation
  - Virome Background
  - Uncultivated Viral Genomes (UViGs)
  - Key methodologies for identifying and analyzing UViGs
  - Bulk Metagenomics vs. Virus-Like Particle (VLP) Enrichment (Review)
  - Main Approaches for Viral Prediction
  - Pipelines for Analyzing the Virome
- Hands-On Session
- Hands-On Review

Definitions and role of the virome

Spanning eukaryotic viruses to bacteriophages



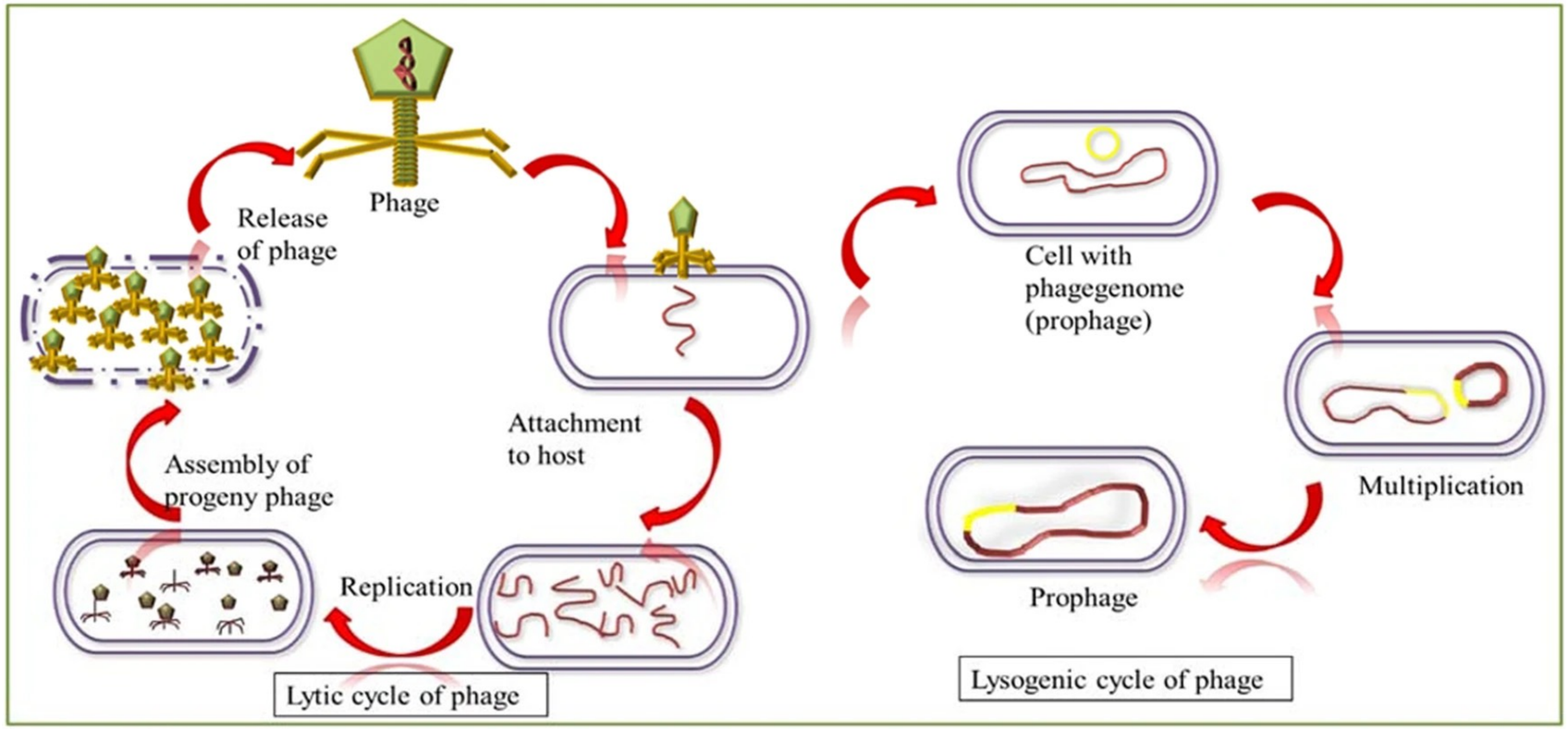


Liang, G., & Bushman, F. (2021)

Wylie, K, et al., (2012)

Definitions and role of the virome

Concept of bacteriophage lifecycles

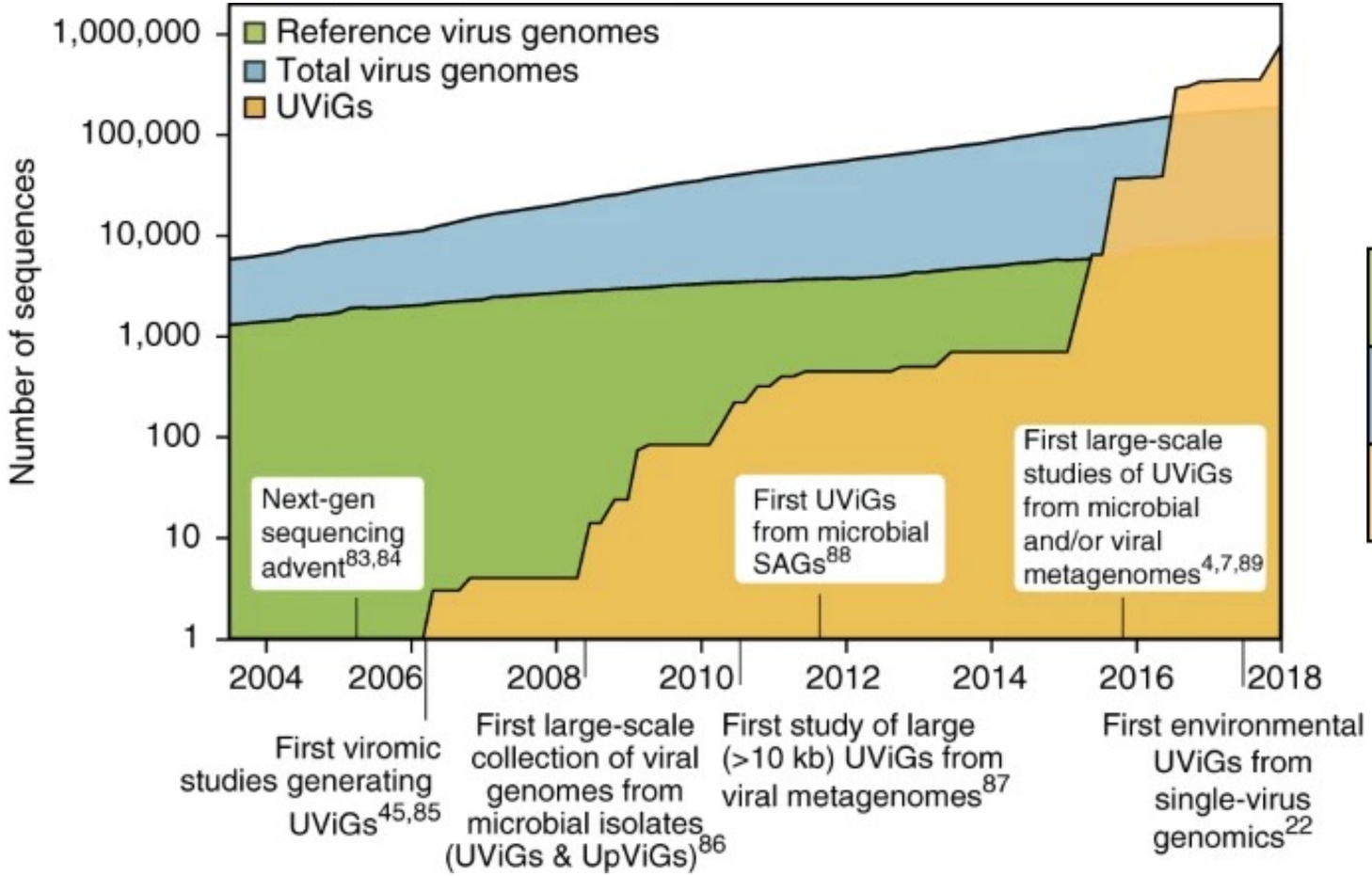


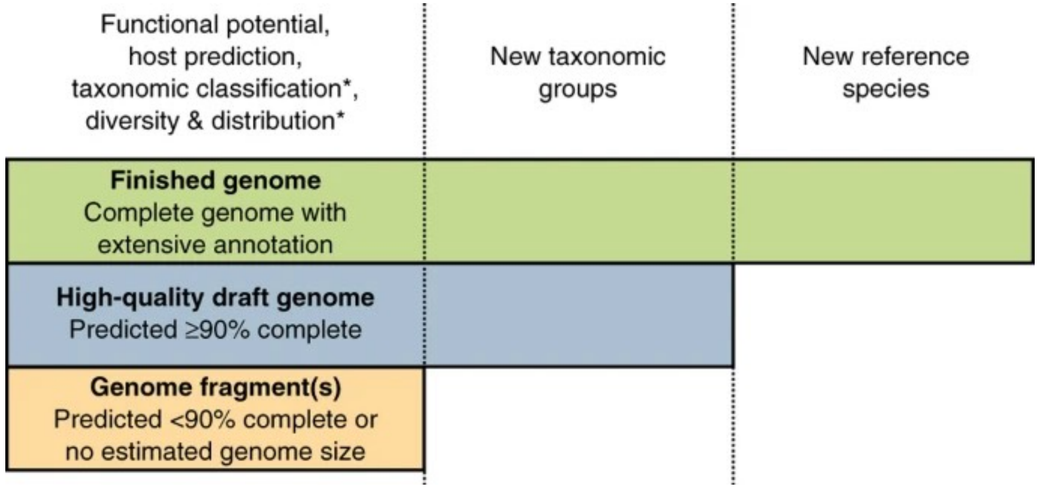
- Lysogenic phages can integrate to the bacterial genome and replicate as long as the bacteria replicates.
- Lysogenic phages = temperate phages.
- Prophages: Stage in which the phage is integrated in the genome.

Ravieer et al., 2024

Uncultivated viruses (uViGs)

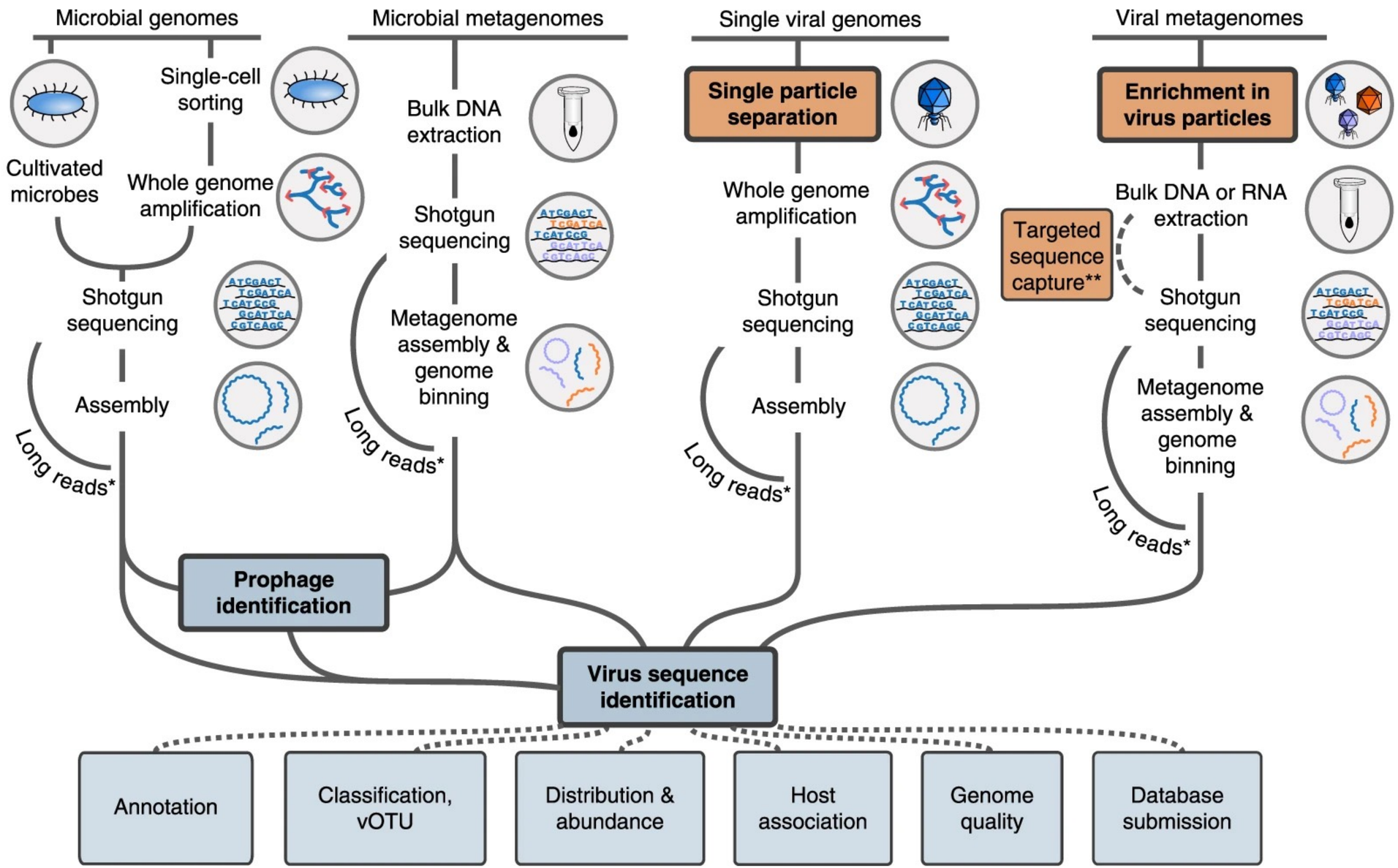
Drastic increase of the number of uViGs deposited in databases over the past years.





Roux et al., 2023

Main techniques for analysing uncultivated viruses



Roux et al., 2023

Bulk metagenomics vs VLP enrichment

Depending on the approach your viral results change

- Comprehensive Sampling: Allow us to relate the different microbial communities within the environment.
- More accurate identification of prophages and their host.
- Lower Viral Specificity: hard to detect low-abundance viruses.
- Complex Data Analysis
- Higher Background Noise: potentially mask of viral signals.

- Enhanced Viral Detection: More sensitive method for detecting viruses.
- Reduced Background Noise: Eliminates most non-viral genetic material.
- Possibility of detecting RNA viruses
- Misses prophages and latent viruses
- Complex and costly sample preparation.

Main approaches for viral prediction

COMPUTATIONAL TOOLS FOR PREDICTING VIRAL SEQUENCES

Homology	Comparing sequences using viral databases and local alignment	**Pros**: High accuracy for known viruses. Allows distant homologous detection **Cons**: Dependent on the quality and completeness of reference databases. Slow
K-mers	Dividing sequences into subsequences and comparing against DBs	**Pros:** Fast and scalable. **Cons**: Detection is limiled to high identity relatives within databases
Machine learning	Training model with viral genomic features	**Pros**: Can detect novel viruses. High accuracy with well-trained models. **Cons**: computational intesnive

Main approaches for viral prediction

COMPUTATIONAL TOOLS FOR PREDICTING VIRAL SEQUENCES

Phylogenetic approaches

Using phylogenetic defined references to located the query sequences

Pros: Provides evolutionary context. Useful for novel virus discovery.
Cons: Computationally intensive. Requires high-quality alignments.

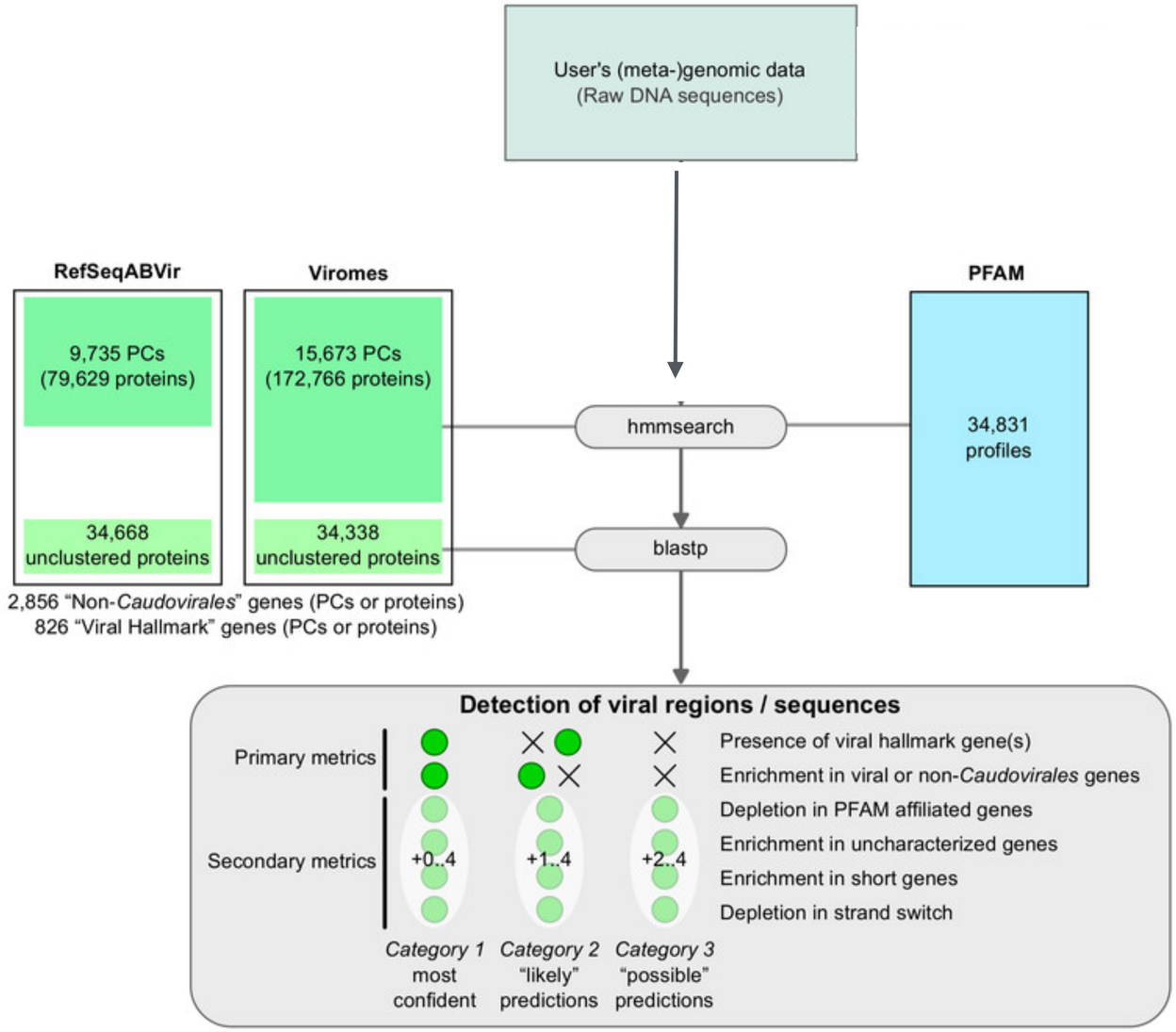
Hybrid

Normally, it combines machine learning and homology approaches

- k-mer analysis typically relies on reads as input.
- Homology-based approaches can use either reads or contigs as input.
- Machine learning and hybrid approaches usually require contigs as input, along with several genomic features for accurate prediction.
- Phylogenetic analysis can use contigs or specific genes extracted after assembly.

Main approaches for viral prediction

Homology based approaches

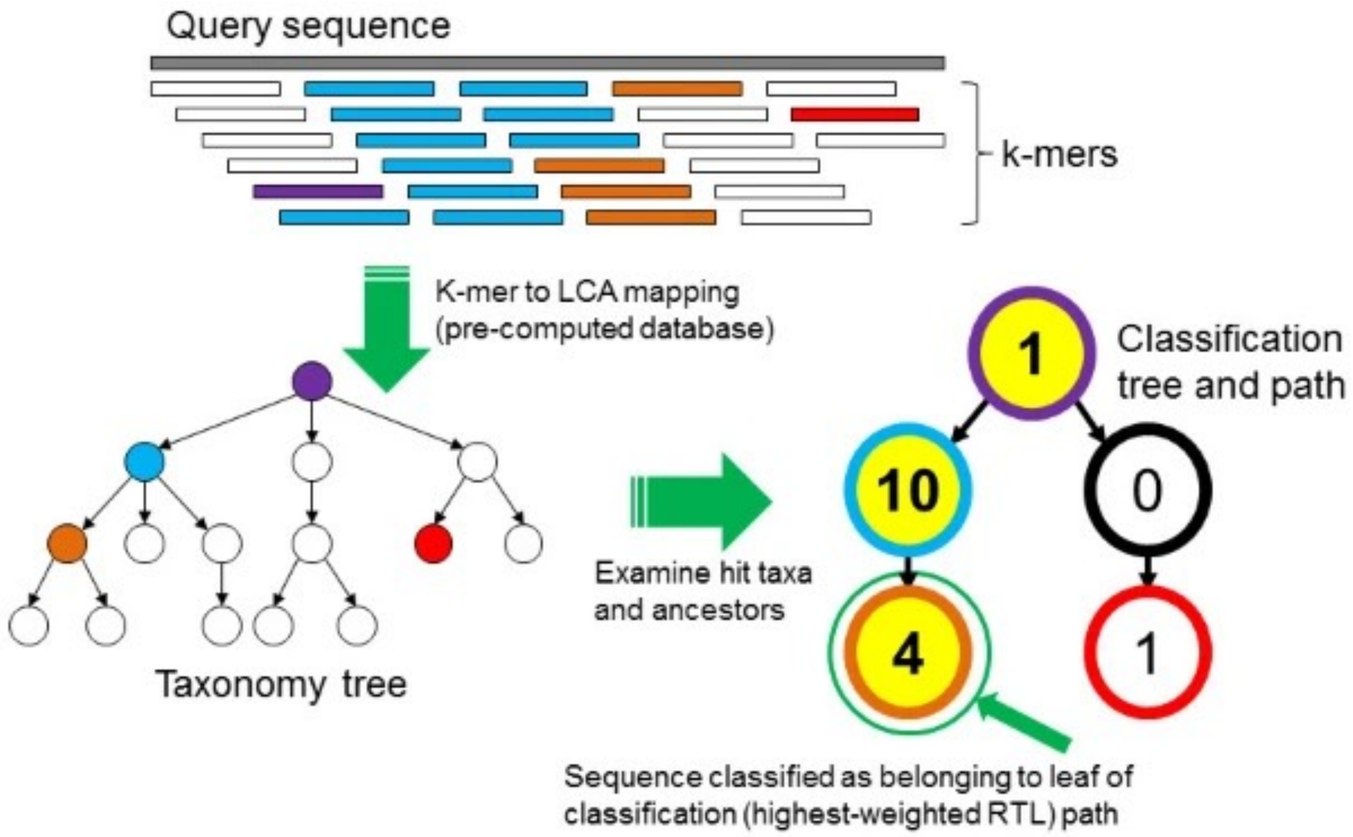


- Use of local alignments and Hidden Markov Models (HMMs) against specific viral databases for precise identification of viral sequences.
- Use of sliding windows to classify viral gene-rich regions and identify regions enriched in viral genes.

Roux et al., 2015

Main approaches for viral prediction

K-mer based approaches

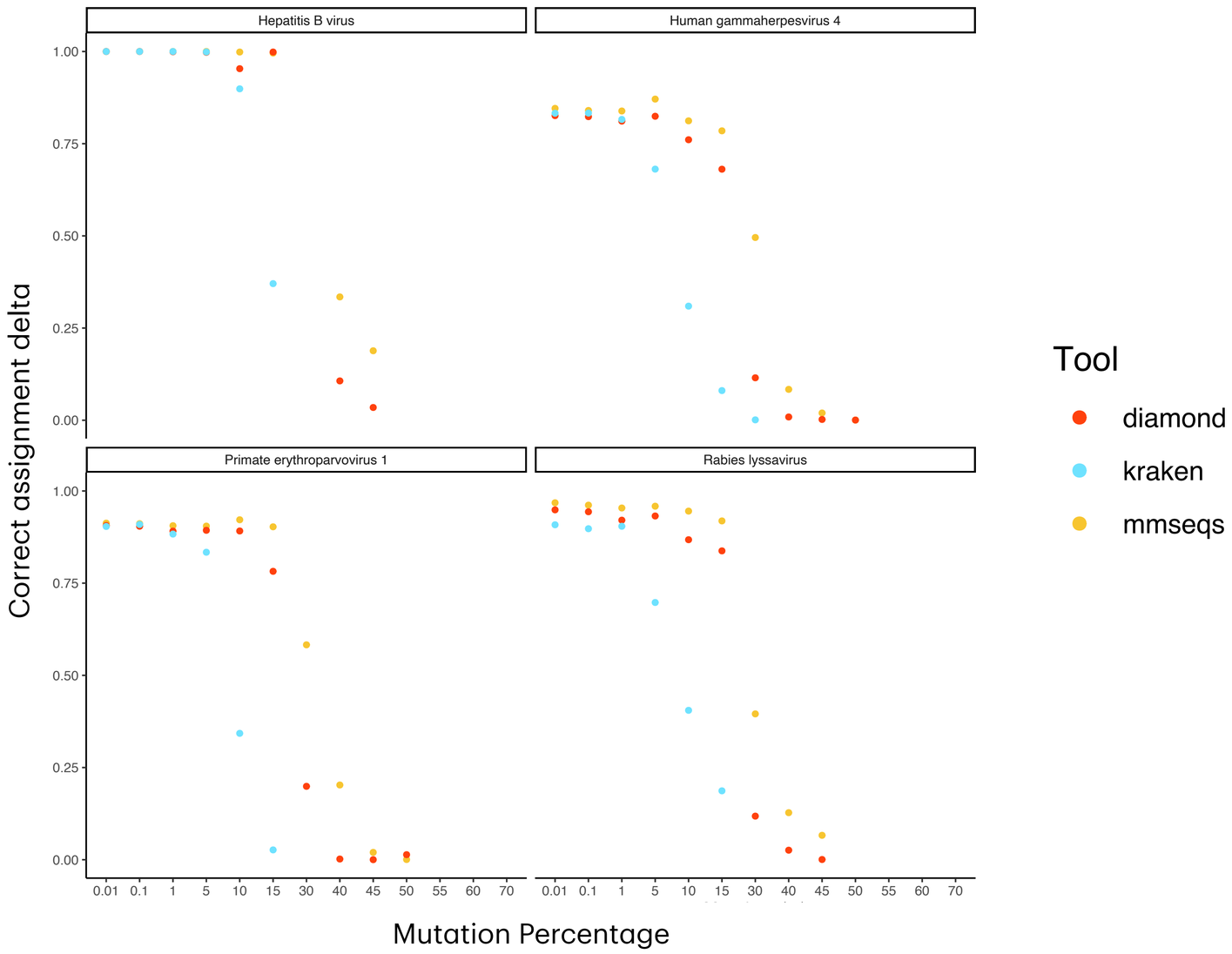


- Use of k-mer frequency analysis to identify unique compositional patterns in viral sequences.
- Comparison of k-mer profiles against curated databases or reference genomes for viral classification.

Wood & Salzberg, 2014

When to use homology or k-mers for read based analysis

How well-known is the system you are working on and data set size as key factors

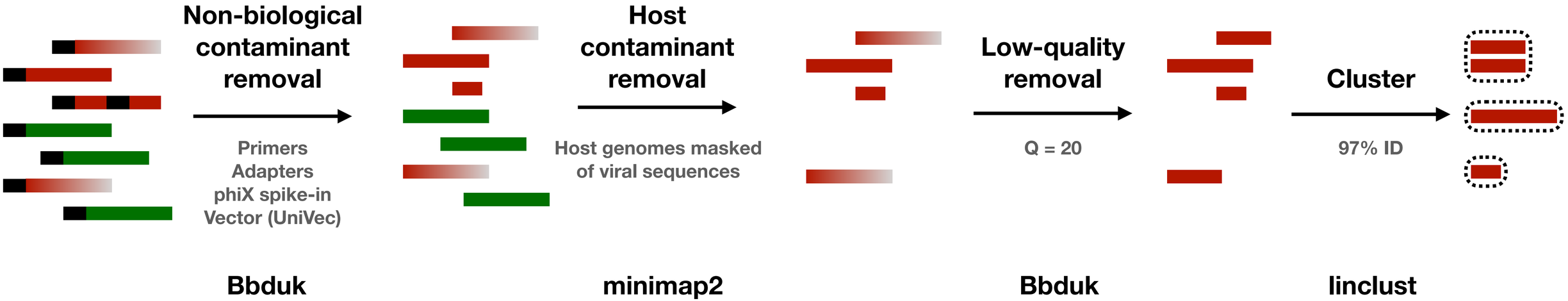


- An increase in mutations can significantly affect the performance of k-mer-based tools.
- This is particularly problematic for novel viruses or viruses without close representatives in the reference database, as they may be missed or misclassified.

Complete workflows for viral analysis: Read and contig-based approaches

Quality control for viral genomics

Key step for avoiding false positives and reducing data sizes

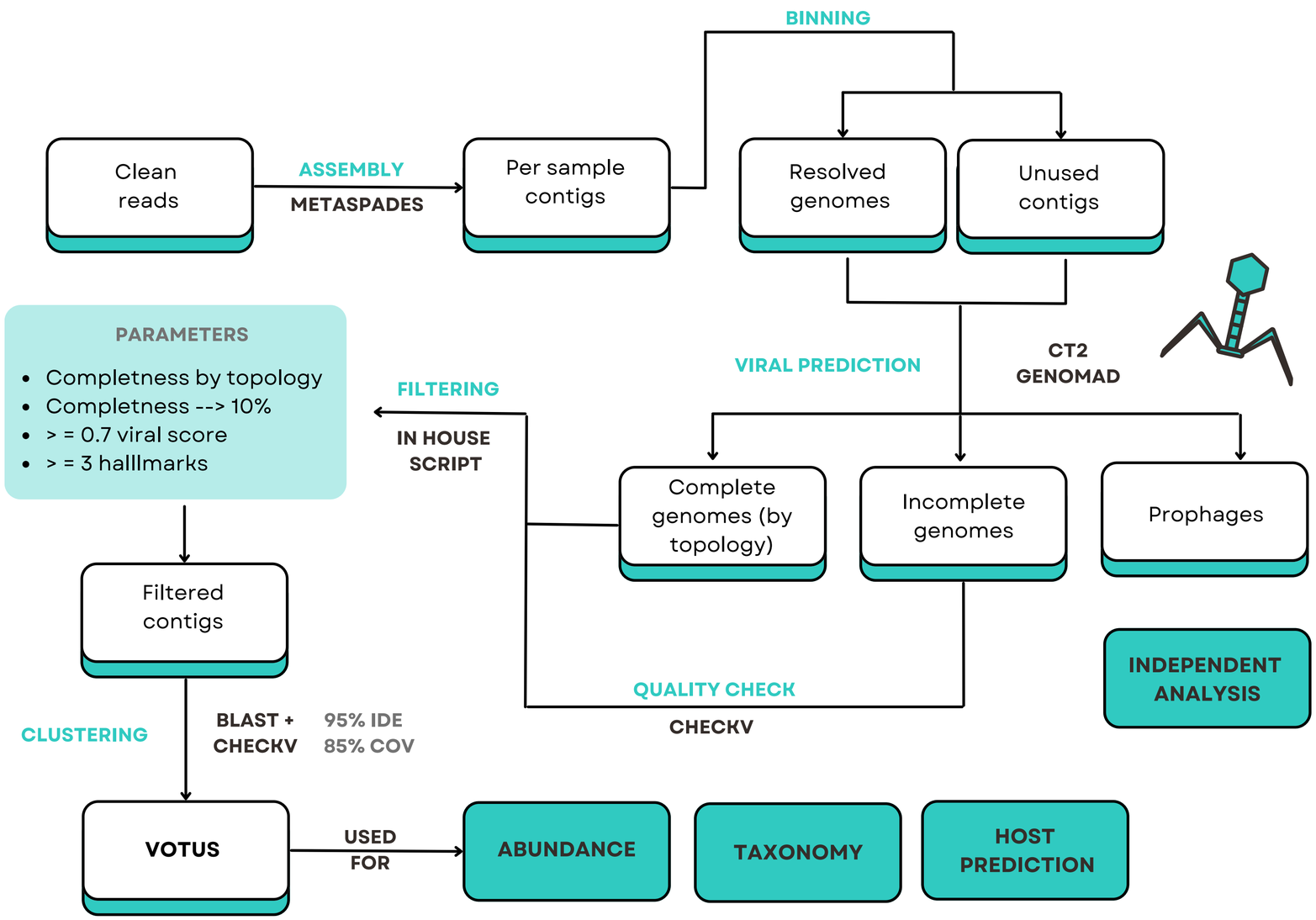


Virus masked reference genomes:

- Human
- Mouse
- Rat
- Macaque
- Pig
- Bat
- Cow
- Camel
- Cat
- Dog
- Mosquito
- Tick
- C. elegans

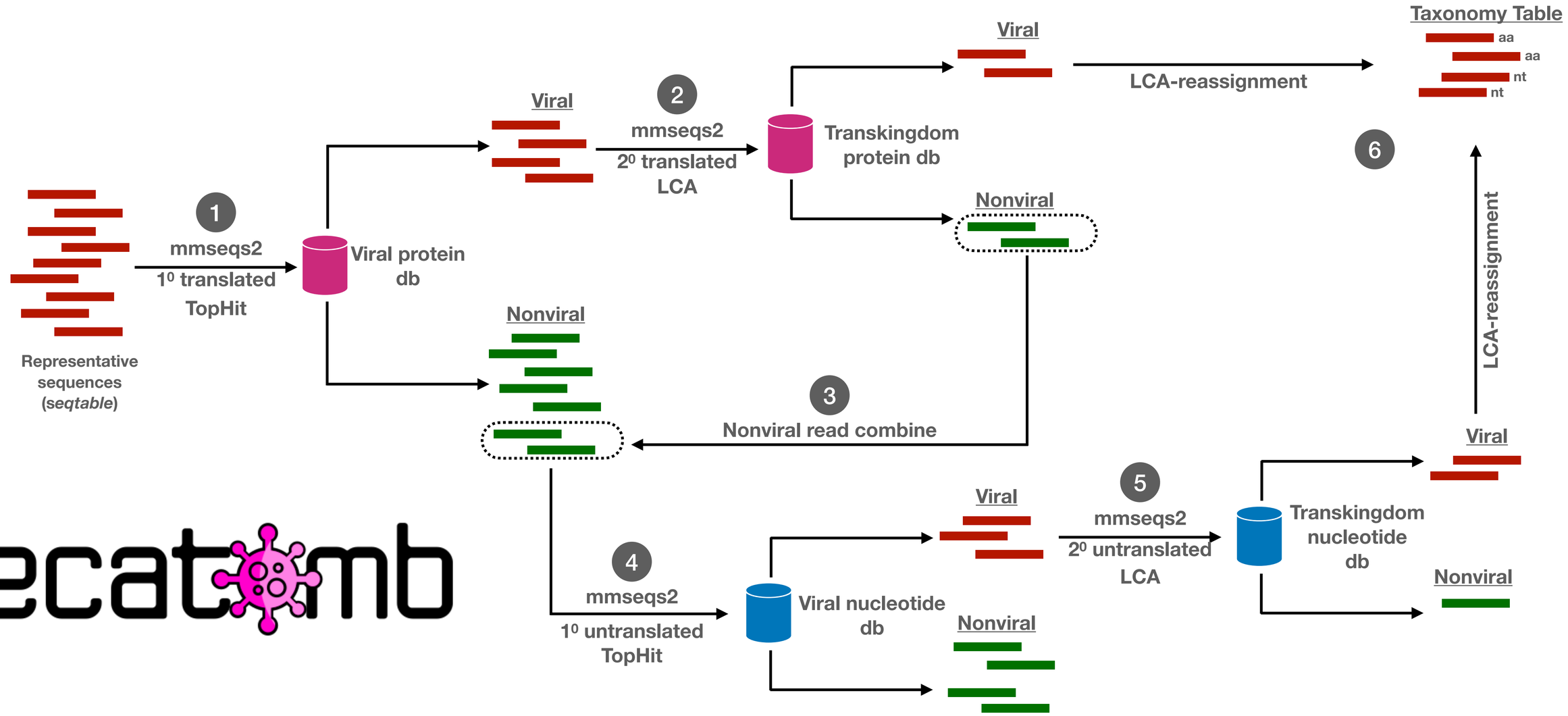
Workflow for viral analysis using contigs

No gold standard available



- Optional steps:
  - Binning
  - Clustering

Workflow for viral analysis using reads



Hecatomb

Hecatomb main output

Bigtable

```
                           seqID               sampleID count        CPM alnType   targetID                                evalue pident fident
1: A13-256-115-06_GTTTCG:8:66010  A13-256-115-06_GTTTCG     8  6.7270249      aa A0A1L2BK65 0.00000000000000000000000000003293   67.5  0.675
2: A13-256-115-06_GTTTCG:1:66034  A13-256-115-06_GTTTCG     1  0.8408781      aa A0A1W5PXG9 0.00000000000000000000000009094000   70.2  0.702
3: A13-256-115-06_GTTTCG:1:66041  A13-256-115-06_GTTTCG     1  0.8408781      aa A0A345BQH3 0.00000000000014019999999999999835   59.6  0.596
4: 183-06-02-24-12_GGCTAC:1:5285 183-06-02-24-12_GGCTAC     1  2.1246354      aa A0A345N1W4 0.00000000000000000000601100000000   53.9  0.539
5: 183-06-02-24-12_GGCTAC:4:5288 183-06-02-24-12_GGCTAC     4  8.4985414      aa A0A345MXJ8 0.00000000000290199999999999995334   58.8  0.588
6: 183-06-02-24-12_GGCTAC:6:5291 183-06-02-24-12_GGCTAC     6 12.7478122      aa A0A1L2BL70 0.00000000548699999999999976705536   58.0  0.580
   nident mismatches  qcov  tcov qstart qend qlen tstart tend tlen alnlen bits                                 targetName taxMethod kingdom
1:     50         24 0.949 0.186     11  232  234    257  330  398    222  118                       A0A1L2BK65_9VIRU VP4       LCA Viruses
2:     52         22 0.949 0.247     11  232  234      7   80  299    222  108                                      Cap       LCA Viruses
3:     37         20 0.701 0.257    239   69  244    164  225  241    186   73  A0A345BQH3_9VIRU Zonula occludens toxin       LCA Viruses
4:     41         35 0.979 0.563    229    2  233     43  118  135    228   94 A0A345N1W4_9VIRU Uncharacterized protein       LCA Viruses
5:     30         21 0.654 0.073    233   81  234    646  696  696    153   68                       Major capsid protein       LCA Viruses
6:     29         21 0.600 0.362    245   96  250     76  125  138    150   60                       A0A1L2BL70_9VIRU VP3       LCA Viruses
                        phylum                     class                     order                     family
1:                Phixviricota         Malgrandaviricetes               Petitvirales                Microviridae
2:             Cressdnaviricota              Arfiviricetes               Cremevirales                Smacoviridae
3:               Hofneiviricota             Faserviricetes              Tubulavirales                  Inoviridae
4: unclassified Viruses phylum unclassified Viruses class unclassified Viruses order unclassified Viruses family
5:                Phixviricota         Malgrandaviricetes               Petitvirales                Microviridae
6:                Phixviricota         Malgrandaviricetes               Petitvirales                Microviridae
                           genus                                 species baltimoreType baltimoreGroup
1: unclassified Microviridae genus       unclassified Microviridae species         ssDNA             II
2:                Porprismacovirus Macaca mulatta feces associated virus 1         ssDNA             II
3:   unclassified Inoviridae genus                          Inoviridae sp.         ssDNA             II
4:      unclassified Viruses genus                       Bacteriophage sp.          <NA>           <NA>
5: unclassified Microviridae genus                        Microviridae sp.         ssDNA             II
6: unclassified Microviridae genus                    Gokushovirus WZ-2015a         ssDNA             II
```

- Main output of taxonomic assignment
- It combines the seqtable IDs with their sampleID, counts, normalized counts, alignment information, taxonomic assignments and Baltimore classification.
- It is a big file, but is designed to make merging with sample metadata, plotting, and statistical interrogation as easy as possible.

Summary

- **Sequencing Methodology**:
  - The choice of sequencing approach (e.g., bulk metagenomics, VLP enrichment) depends on the main scope of the project.
  - Different methodologies impact the detection of specific viruses and the estimation of viral abundance.
- **Viral Detection:**
  - Viral detection should be performed carefully, applying multiple tools to increase confidence in predictions.
  - Combining complementary approaches enhances robustness.
- **Tool Selection:**
  - For read-based analysis, tool selection should consider the novelty of the system.
  - k-mer-based approaches may have limitations in understudied systems due to the lack of reference data.
  - Homology-based approaches are time- and resource-intensive but are particularly useful for unknown or novel systems.

Acknowledgements

Washington University in St. Louis
School of Medicine

Handley Lab
Scott Handley
Leran Wang
Kathie Mihindukulasuriya
Lindsay Droit
Dhoha Abid
Megan Johnson
Martina Moore

Dave Wang

Center for Genome Sciences
Jessica Hoisington-Lopez
MariaLynn Crosby

Ragon Institute
MASS GENERAL | MIT | HARVARD

Kwon Lab
Doug Kwon
Joseph Elsherbini
Sarah Eisa
Cameron Reitan