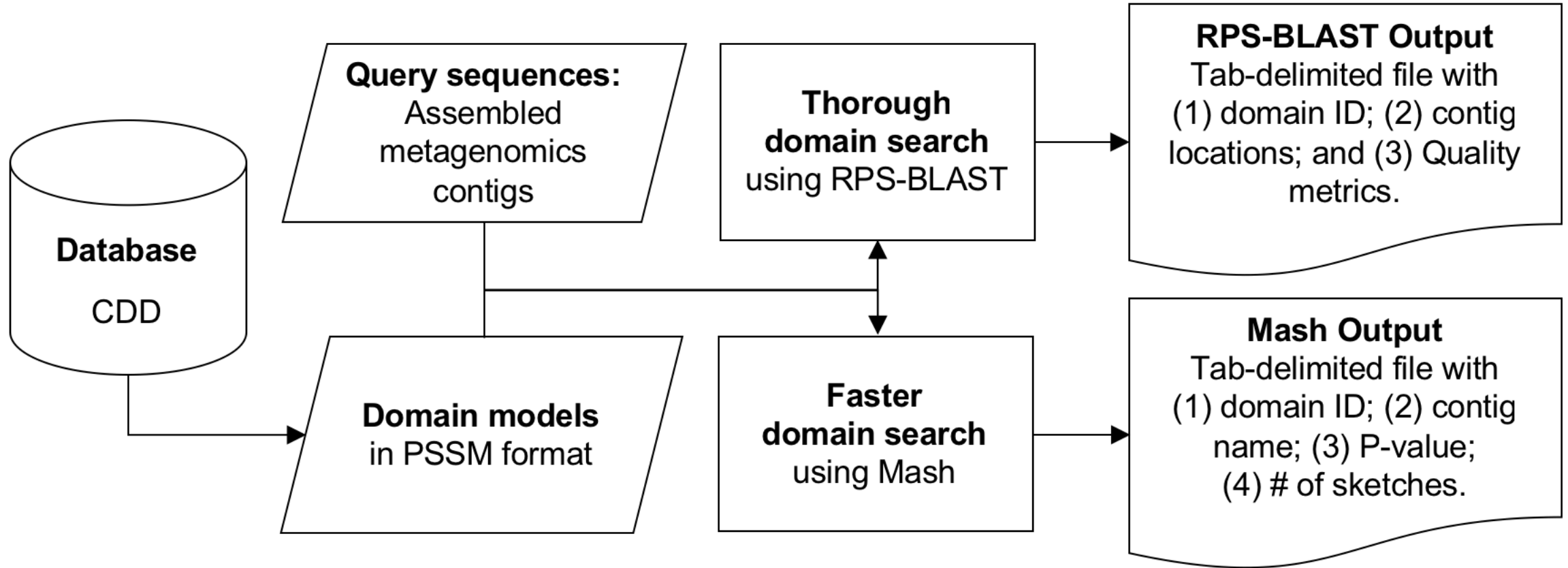


# ***Domain search workflow***



# ***Protein domains search pipeline***

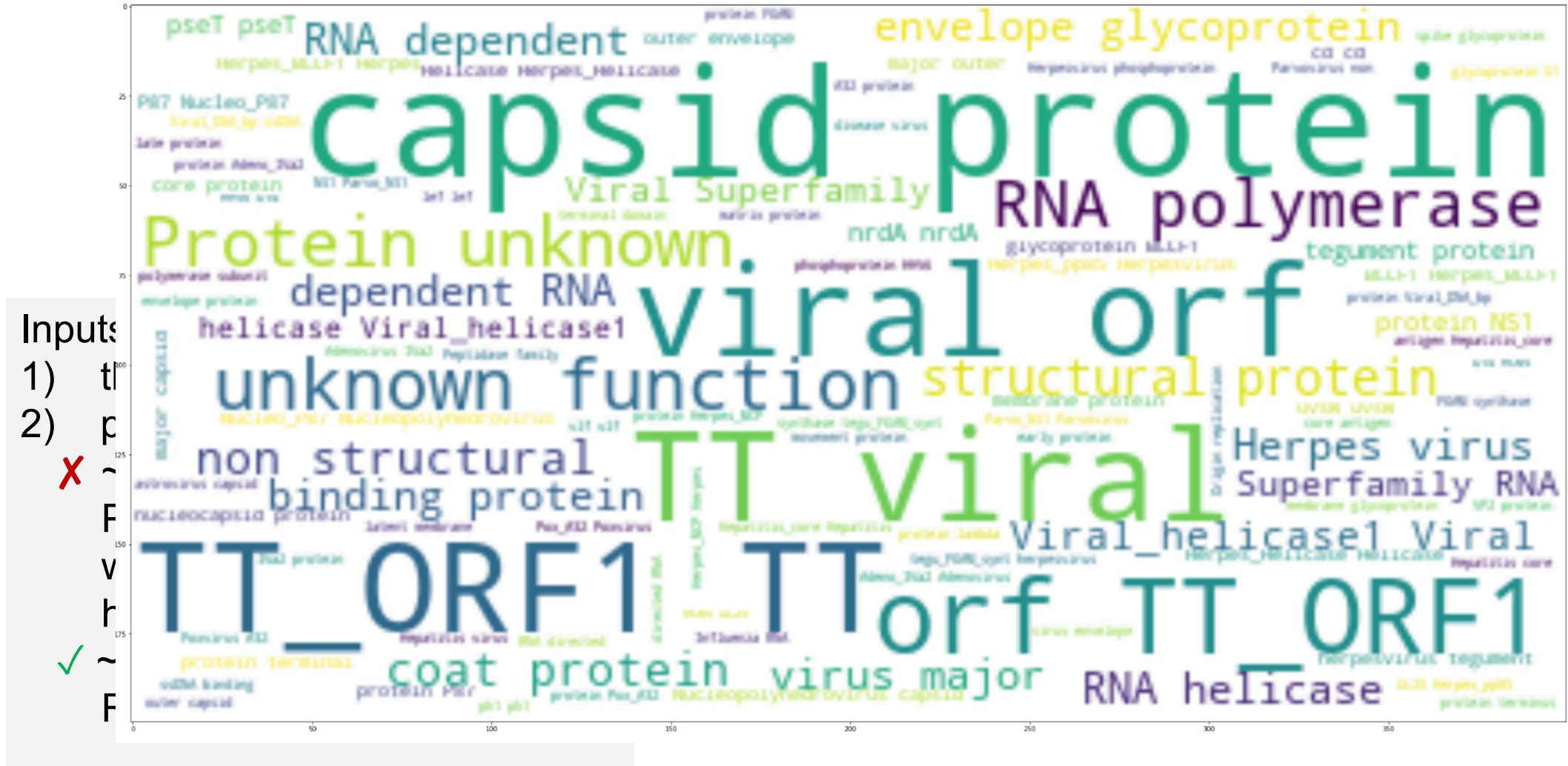
~3k datasets with  
assembled contigs



Inputs:

- 1) the (assembled) contigs
- 2) protein domain models
  - ✗ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
  - ✓ ~2k “viral”-enriched set from Rodney

# Protein domains search pipeline



## General CDD hit stats

Total	1e-3	1e-10
2,997 SRRs	2,745 hits	2,534
55,503,968 contigs	5,606,754 (10%)	278,725 (0.5%)
2,082 CDDs	2,079 CDDs	1,263 CDDs

→ 77,3% of contigs = only 1 CDD hit

# ***Protein domains search pipeline***

~3k datasets with  
assembled contigs



Inputs:

- 1) the (assembled) contigs
- 2) protein domain models
  - ✗ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
  - ✓ ~2k “viral”-enriched set from Rodney

?Dataset clustering?

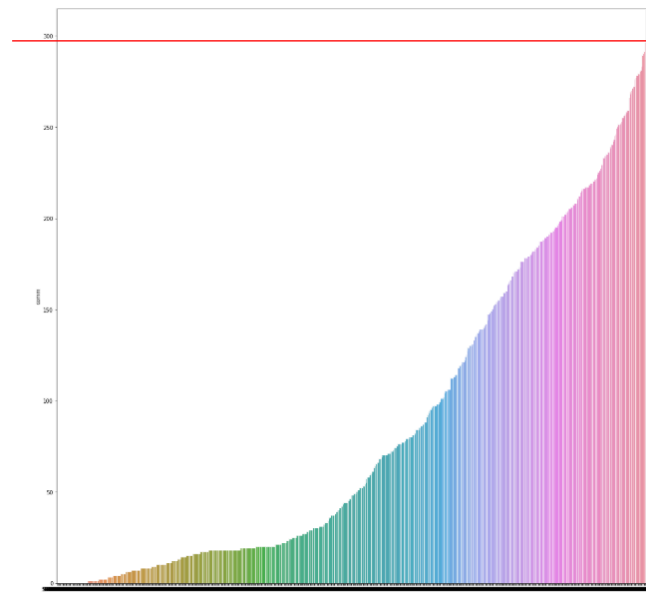
~100k datasets with  
non-assembled reads



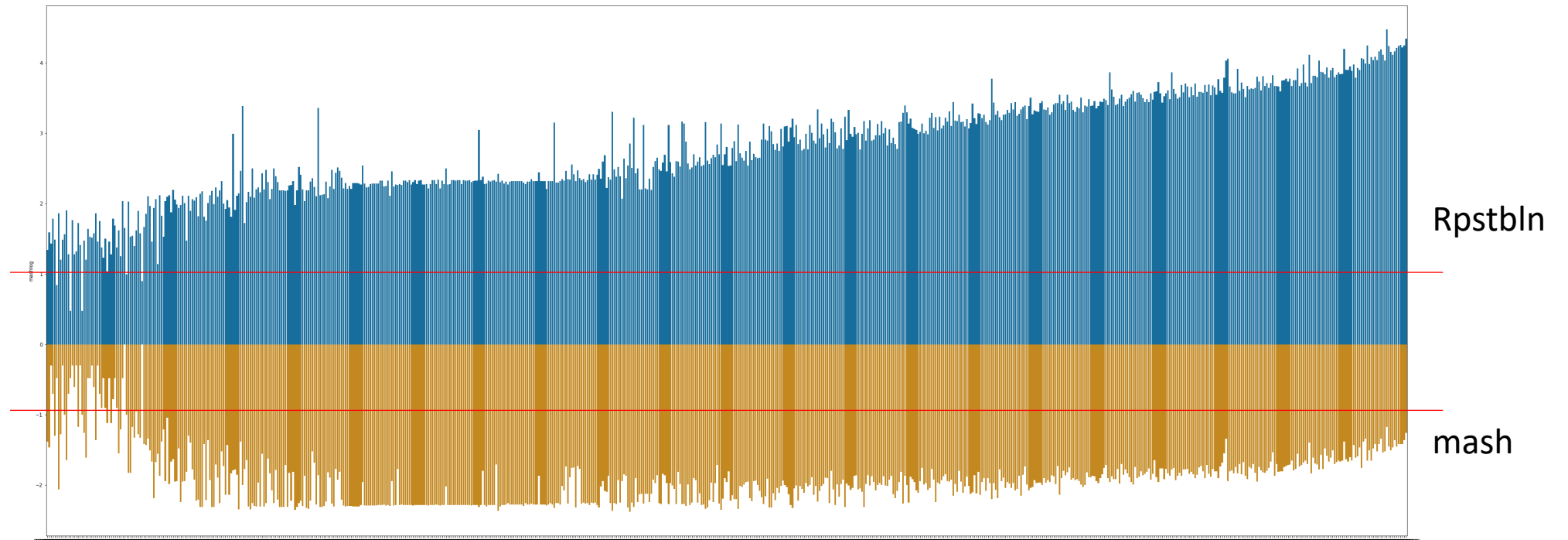
Inputs:

- 1) the (non-assembled) reads
- 2) protein domain models
  - but...how to connect?
    - i. k-mers – build our own method?
    - ii. min-hash, if it would work?

# Kmer-6

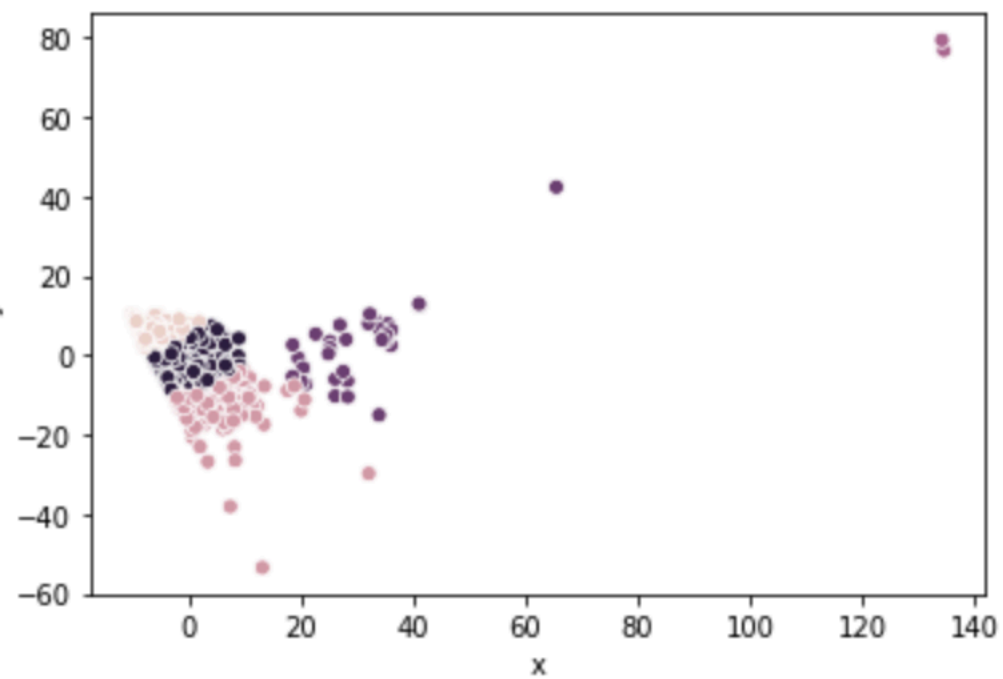


Common (very low ...)

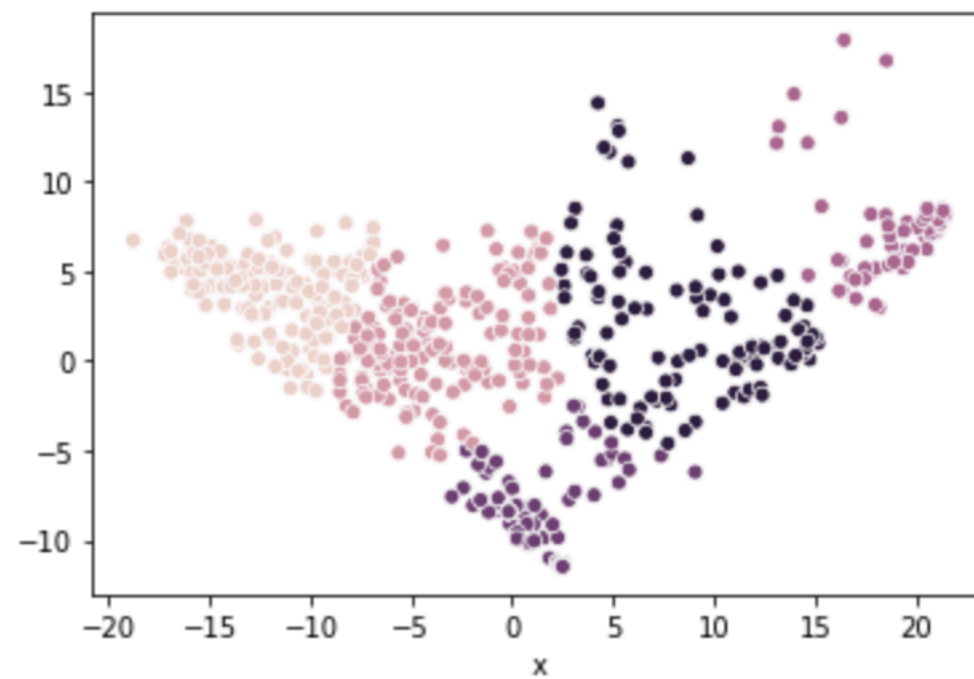


# Kmer-6

rpstbln



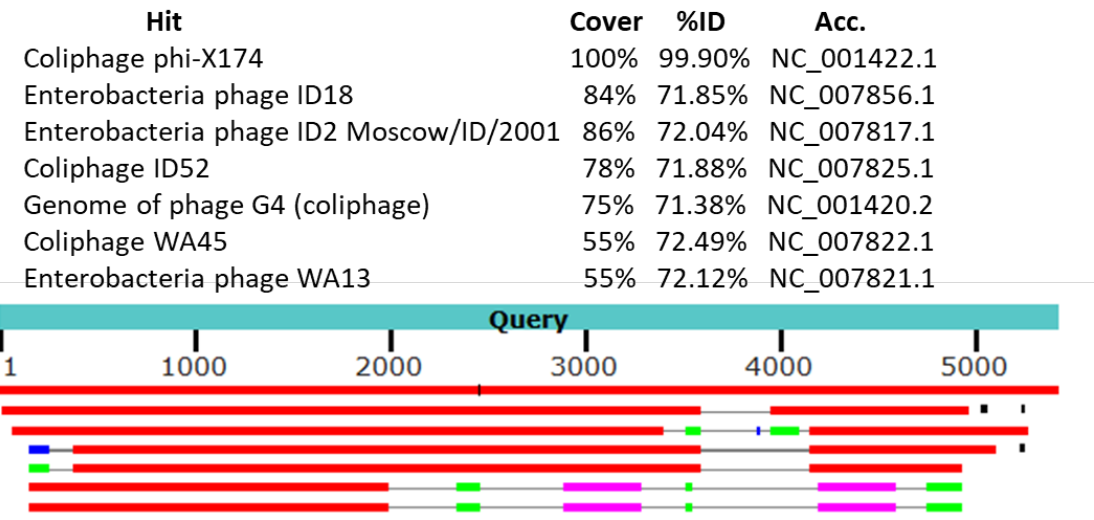
Mash



# Selected 3 contigs with highest number of CDD hits

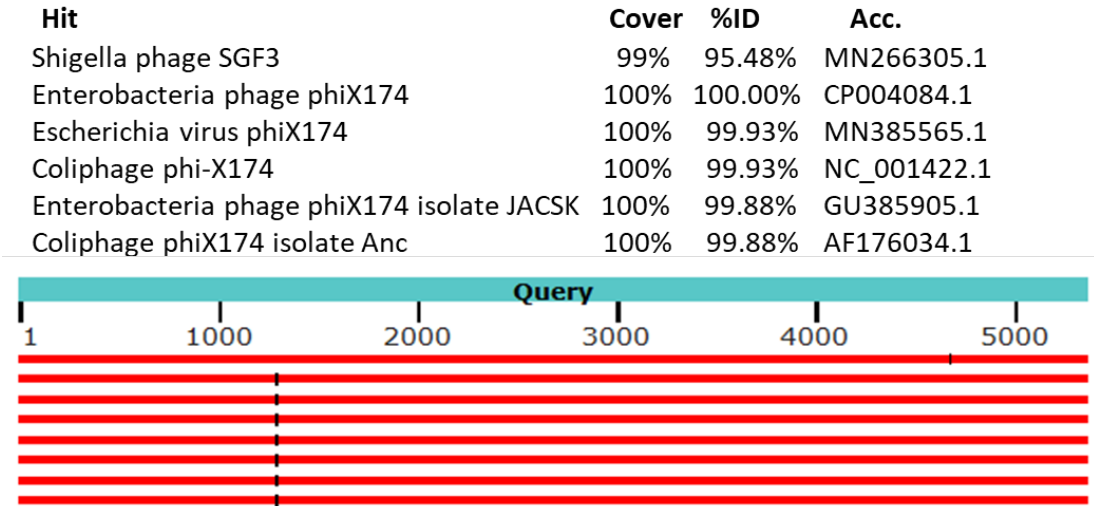
BLASTN vs viral RefSeqs only

Contig\_1\_390.199\_Circ:1.5386



BLASTN vs all viral seqs

Contig\_41\_104.343:1.5343



BLASTN vs all viral seqs

Contig\_1478\_31.5683:1.15688

