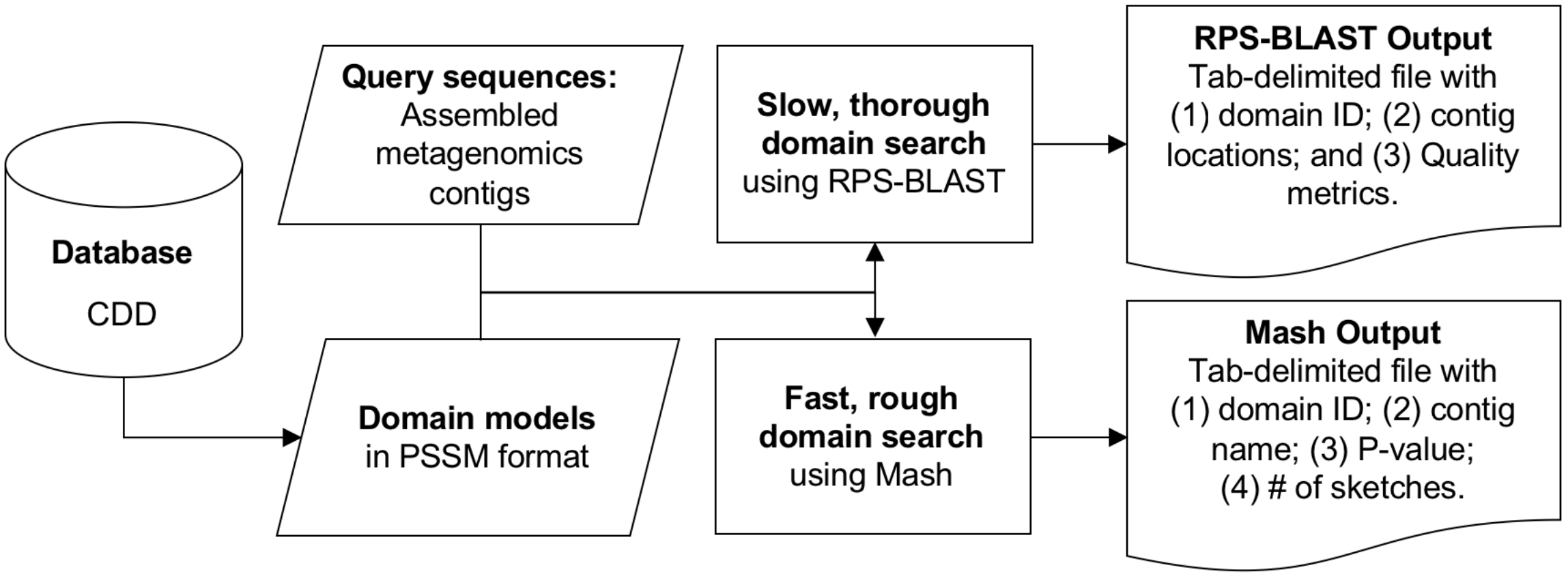
Domain search workflow



Protein domains search pipeline

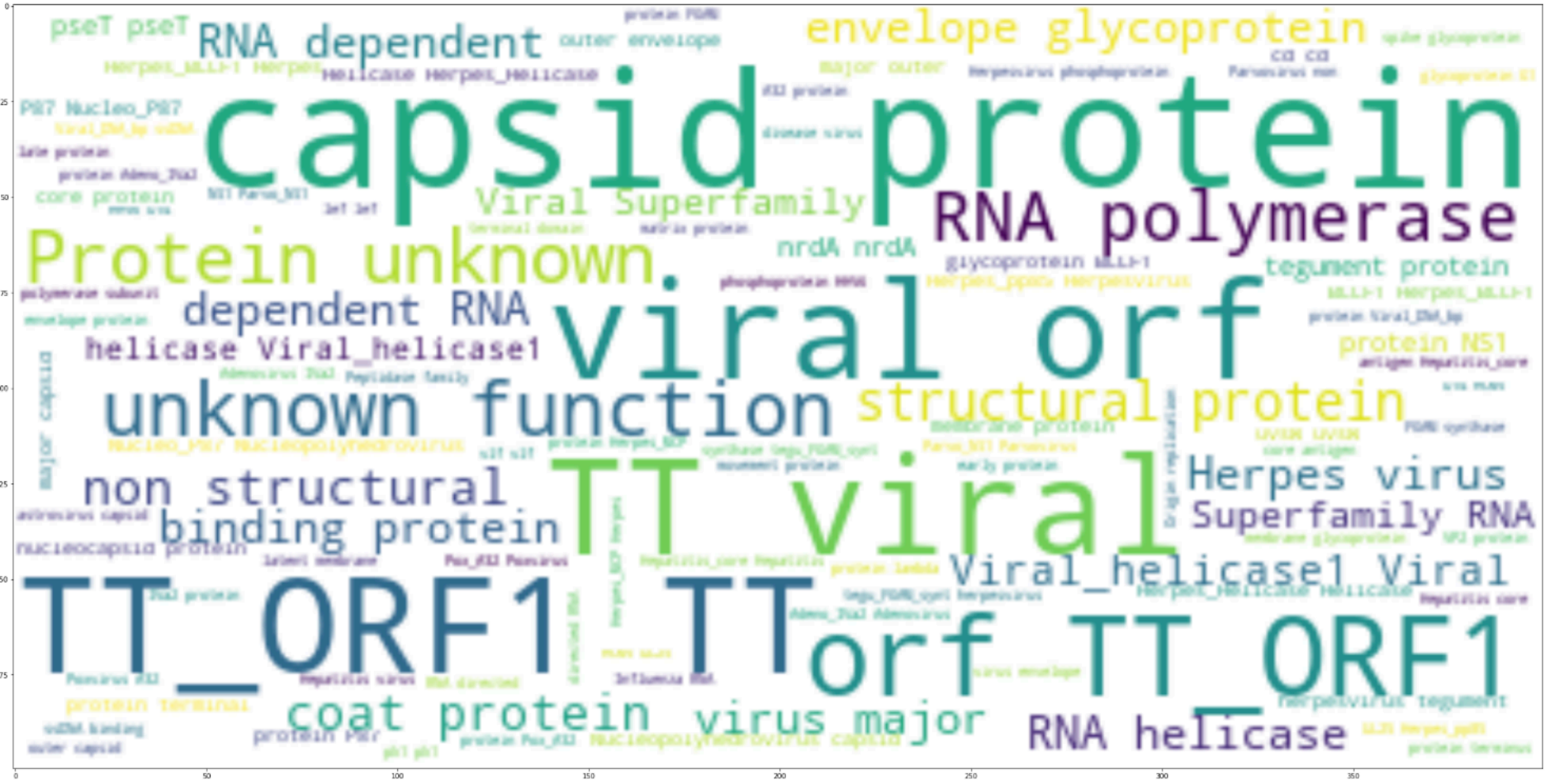
~3k datasets with assembled contigs

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

Inputs
1) t
2) p
✗ ~
F
v
h
✓ ~
F



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

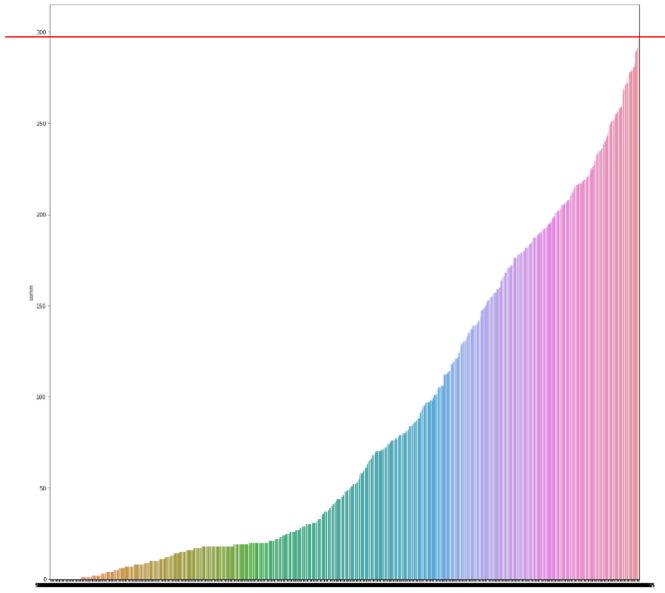
Inputs:
1) the (assembled) contigs
2) protein domain models
   - ✗ ~300k CDD, no PFAM, POGs, etc. to avoid later work in filtering overlapping hits
   - ✓ ~2k “viral”-enriched set from Rodney

?Dataset clustering?

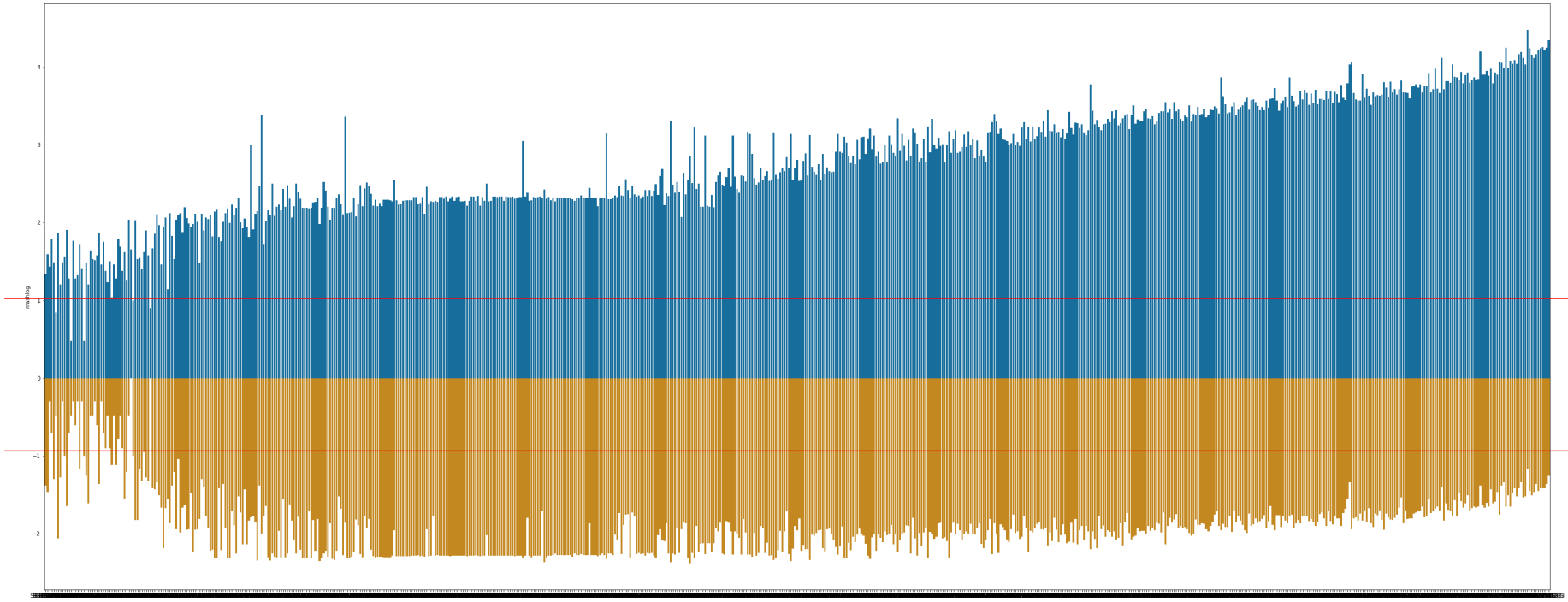
~100k datasets with non-assembled reads

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

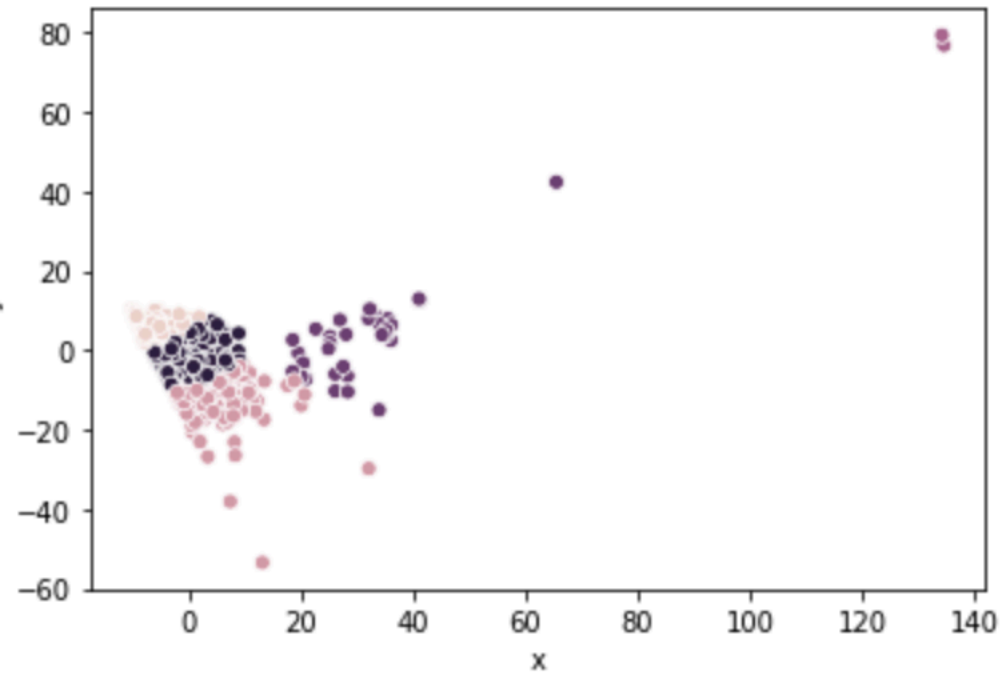


Common (very low …)

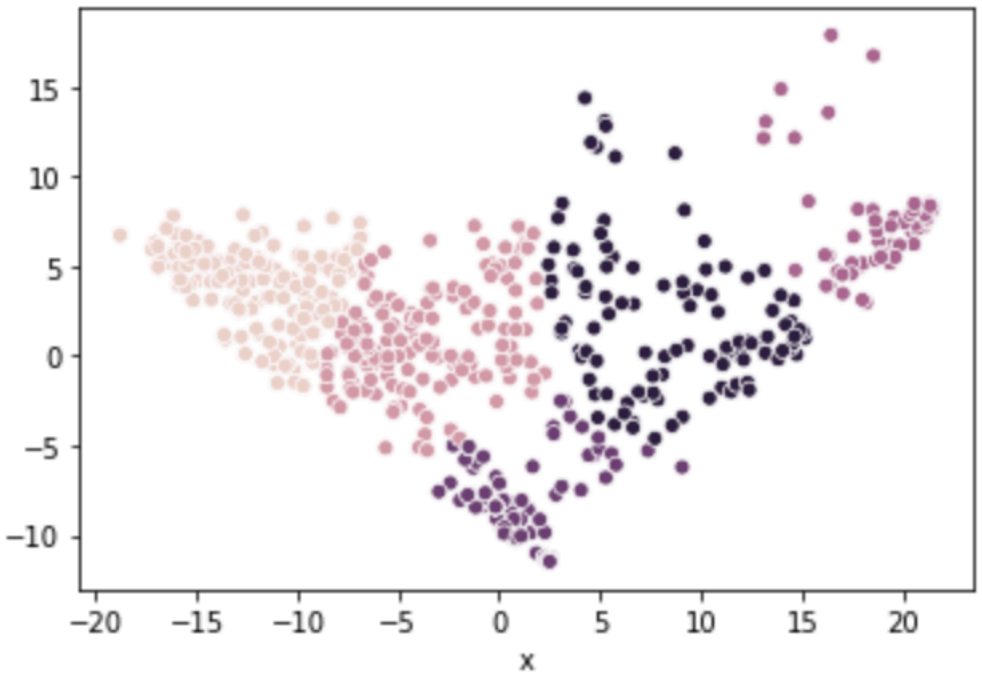


Kmer-6

rpstbln



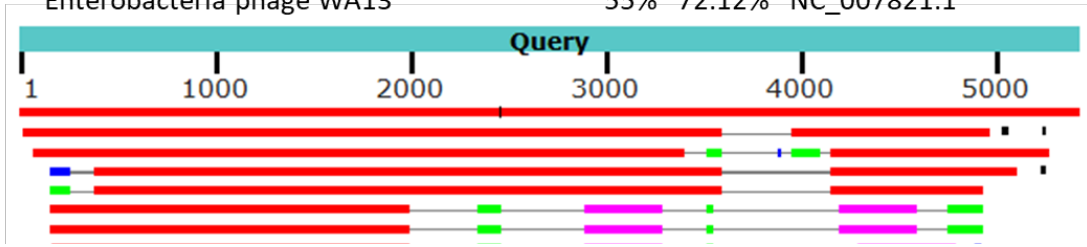
Mash



Selected 3 contigs with highest number of CDD hits

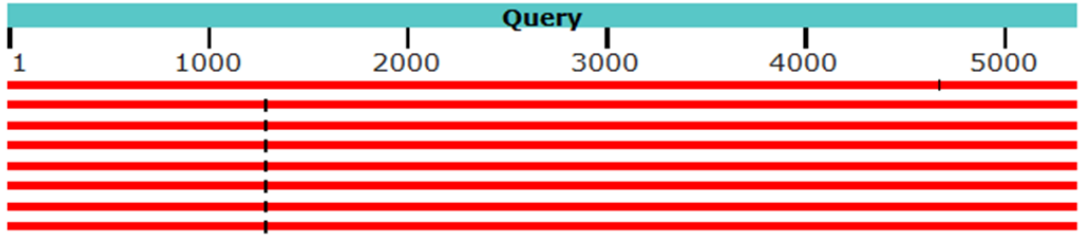
BLASTN vs viral RefSeqs only
Contig_1_390.199_Circ:1.5386

Hit	Cover	%ID	Acc.
Coliphage phi-X174	100%	99.90%	NC_001422.1
Enterobacteria phage ID18	84%	71.85%	NC_007856.1
Enterobacteria phage ID2 Moscow/ID/2001	86%	72.04%	NC_007817.1
Coliphage ID52	78%	71.88%	NC_007825.1
Genome of phage G4 (coliphage)	75%	71.38%	NC_001420.2
Coliphage WA45	55%	72.49%	NC_007822.1
Enterobacteria phage WA13	55%	72.12%	NC_007821.1



BLASTN vs all viral seqs
Contig_41_104.343:1.5343

Hit	Cover	%ID	Acc.
Shigella phage SGF3	99%	95.48%	MN266305.1
Enterobacteria phage phiX174	100%	100.00%	CP004084.1
Escherichia virus phiX174	100%	99.93%	MN385565.1
Coliphage phi-X174	100%	99.93%	NC_001422.1
Enterobacteria phage phiX174 isolate JACSK	100%	99.88%	GU385905.1
Coliphage phiX174 isolate Anc	100%	99.88%	AF176034.1



BLASTN vs all viral seqs
Contig_1478_31.5683:1.15688

Hit	Cover	%ID	Acc.
Wolbachia endosymbiont wVitB of Nasonia vitripennis phage WOVitB segment 2	76%	94.46%	HQ906666.1
Wolbachia phage WO	76%	94.56%	KX522565.1
Wolbachia endosymbiont wVitA of Nasonia vitripennis phage WOVitA1	76%	94.56%	HQ906662.1
Satyrvirus sp. clone Satyrvirus_5 genomic sequence	2%	68.24%	MK072441.1
Flamingopox virus FGPVKD09	2%	70.00%	MF678796.1