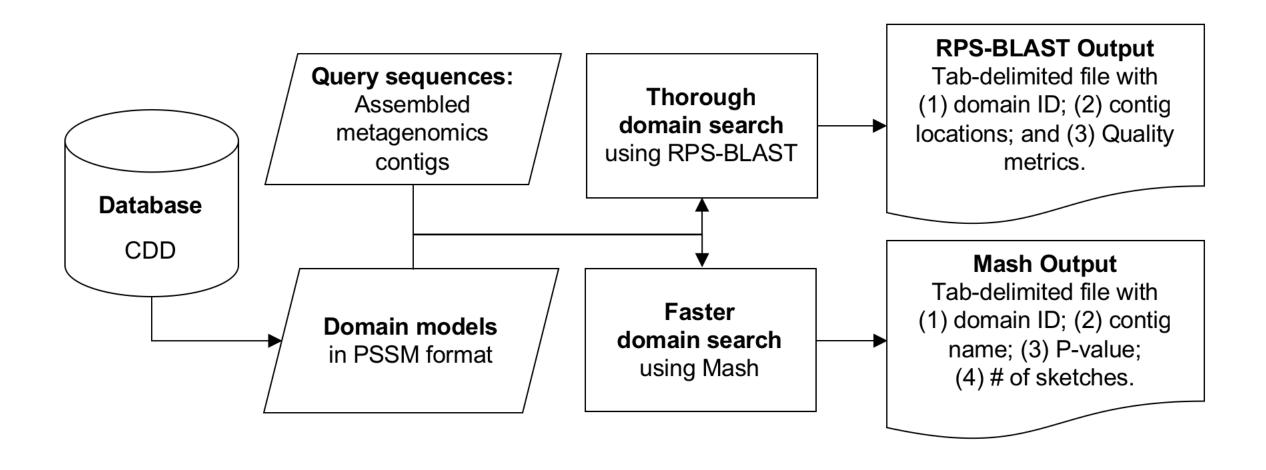
### Domain search workflow



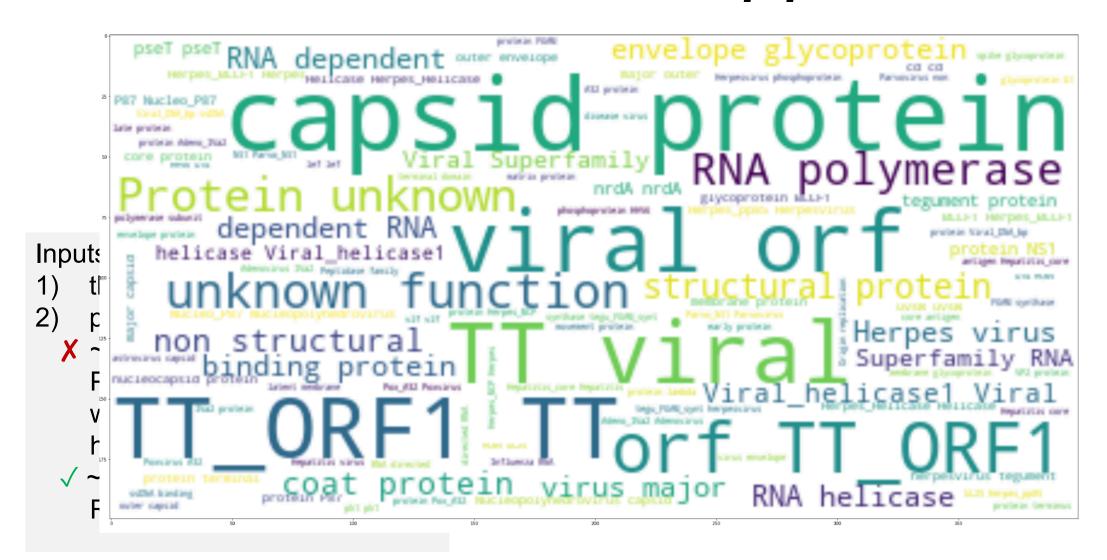
# Protein domains search pipeline

~3k datasets with assembled contigs

#### Inputs:

- 1) the (assembled) contigs
- 2) protein domain models
  - X ~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 SRR	s 2,745 hits	2,534
55,503,96 contigs	8 5,606,754 (10%)	278,725 (0.5%)
2,082 CDD	s 2,079 CDDs	1,263 CDDs

## Protein domains search pipeline

~3k datasets with assembled contigs

#### Inputs:

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

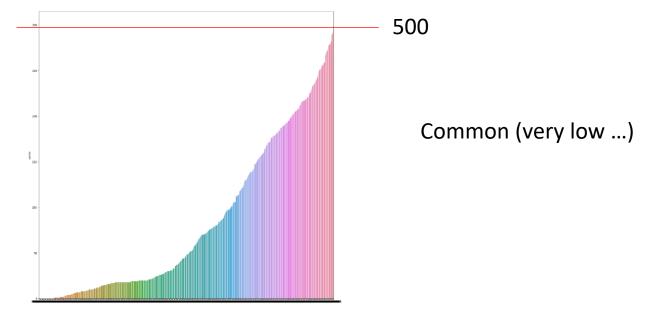
~100k datasets with non-assembled reads

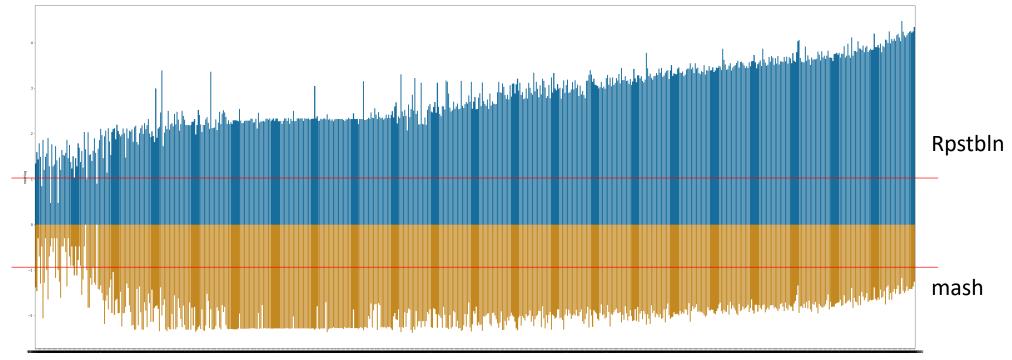
#### Inputs:

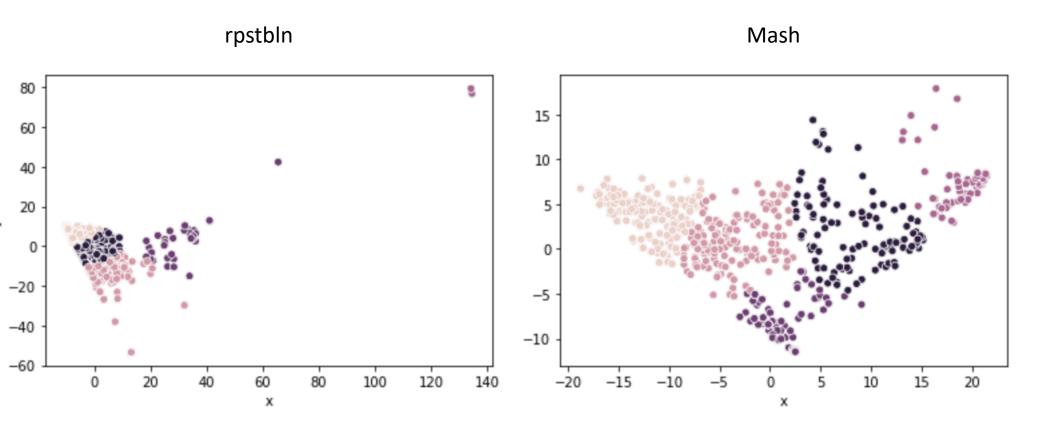
- 1) the (non-assembled) reads
- 2) protein domain models
  - but...how to connect?
    - i. k-mers build our own method?
    - ii. min-hash, if it would work?

?Dataset clustering?



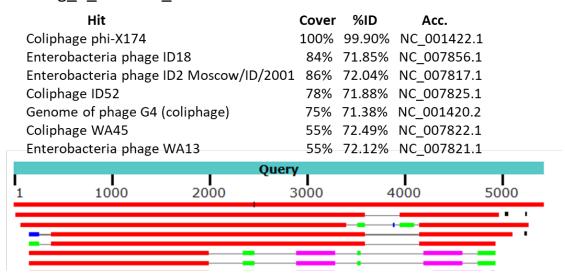






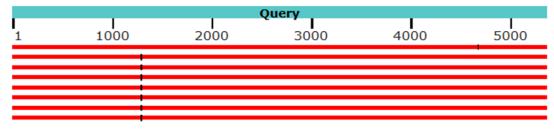
## Selected 3 contigs with highest number of CDD hits

BLASTN vs viral RefSeqs only Contig 1 390.199 Circ:1.5386



BLASTN vs all viral seqs Contig\_41\_104.343:1.5343

05.1
34.1
65.1
122.1
05.1
84.1



BLASTN vs all viral seqs

Contig_1478_31.5683:1.15688								
	Hit				Cove	r %ID	Acc.	
Wolb	Wolbachia endosymbiont wVitB of Nasonia vitripennis phage WOVitB segment 2				ent 2 769	6 94.46%	HQ906666.1	
Wolbachia phage WO					769	6 94.56%	KX522565.1	
Wolbachia endosymbiont wVitA of Nasonia vitripennis phage WOVitA1					769	6 94.56%	HQ906662.1	
Satyrvirus sp. clone Satyrvirus 5 genomic sequence					2%	68.24%	MK072441.1	
Flami	ngopox virus FGPVKD	009			29	6 70.00%	MF678796.1	
			Query					
1	3000	6000	9000	12000	15000			
		1						