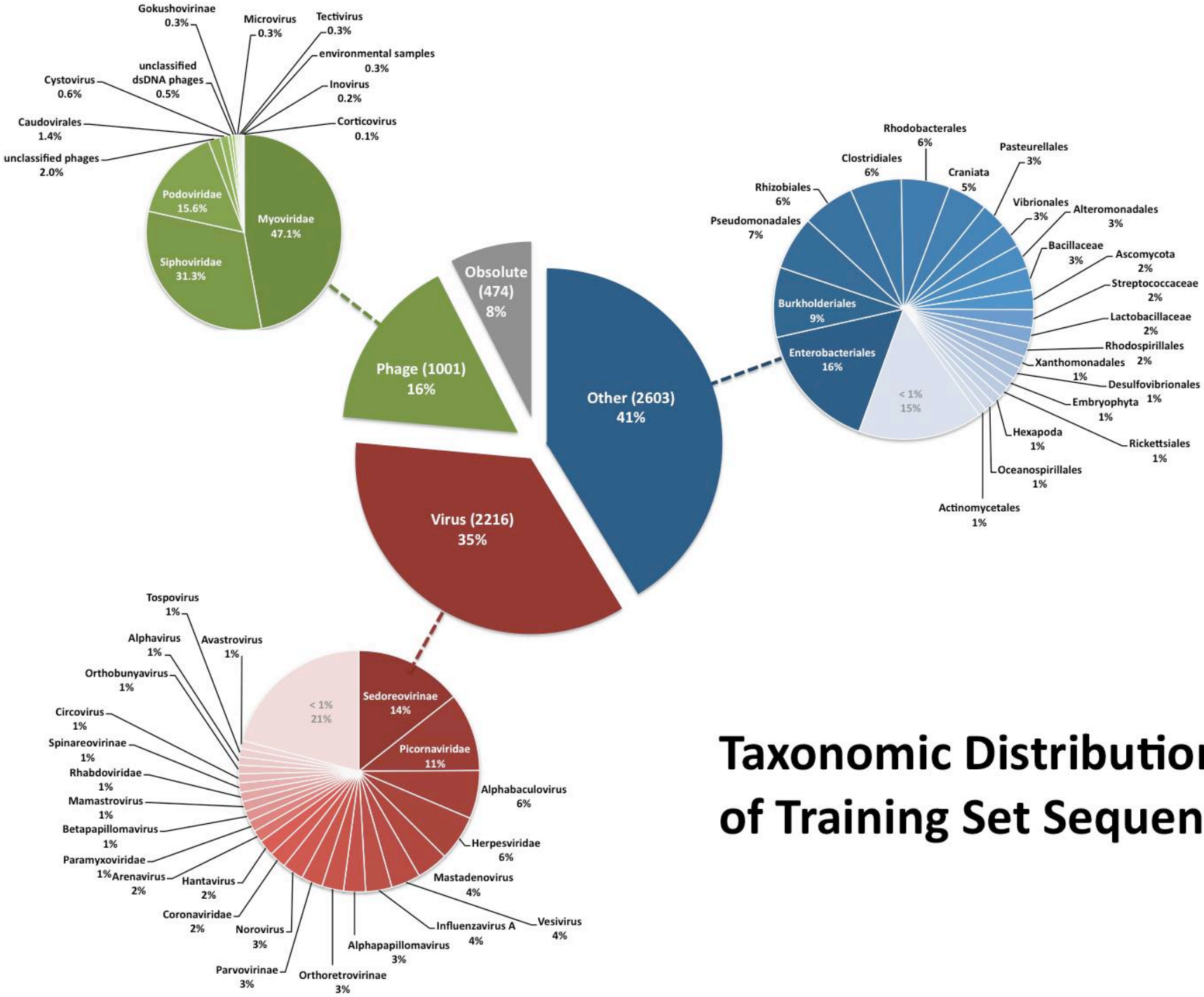
DDM
ANN and Metaproteomics Update

March 8, 2012

SAN DIEGO STATE UNIVERSITY

In Progress

- Neural Network Paper
  - New Figures
  - Status of ANN Interface
- Metaproteomics Paper
  - Analysis of raw reads



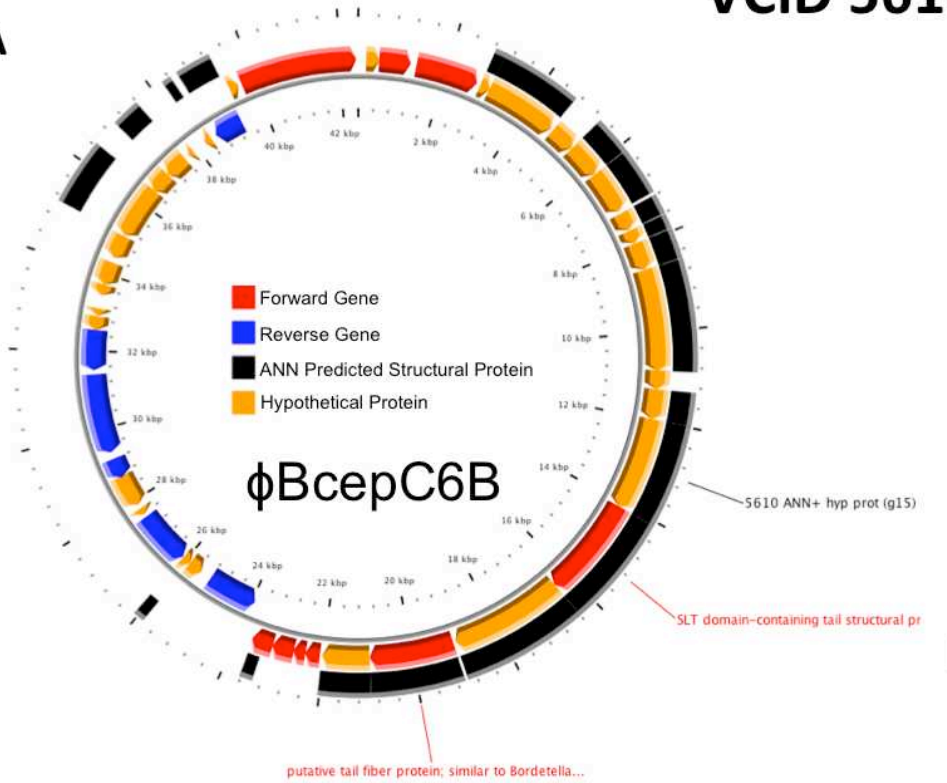
Taxonomic Distribution of Training Set Sequences

Phage Major Capsid Protein (MCP) and Structural Network Predictions of Capsids from Eukaryotic and Archaeal Viruses

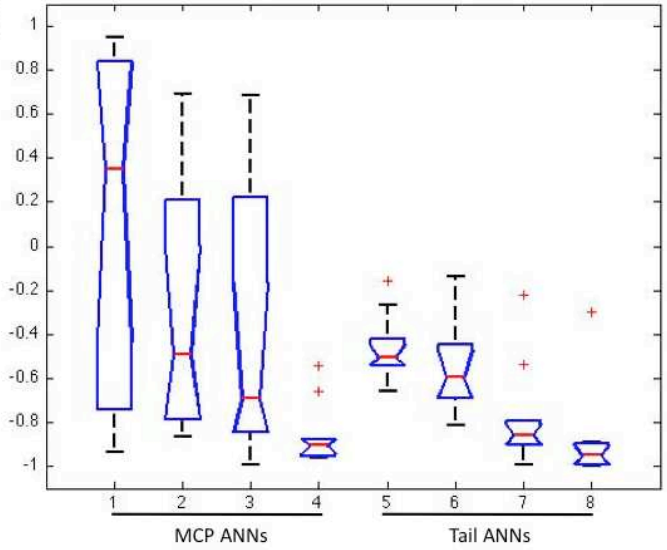
Genome Type (Number of Genomes)	MCP 1:1	MCP 2:1	MCP 3:1	MCP 4:1	MCP 7:1	Structural
dsDNA (123)	15.4%	11.4%	11.4%	10.6%	11.4%	85.9%
dsRNA (44)	4.5%	4.5%	0.0%	2.3%	0.0%	77.3%
ssDNA (325)	0.9%	0.0%	0.0%	0.0%	0.0%	95.1%
ssRNA (338)	16.9%	12.1%	10.9%	6.2%	4.1%	91.4%
Archaea (28)	11.1%	5.6%	5.6%	5.6%	5.6%	71.4%

VCID 5610

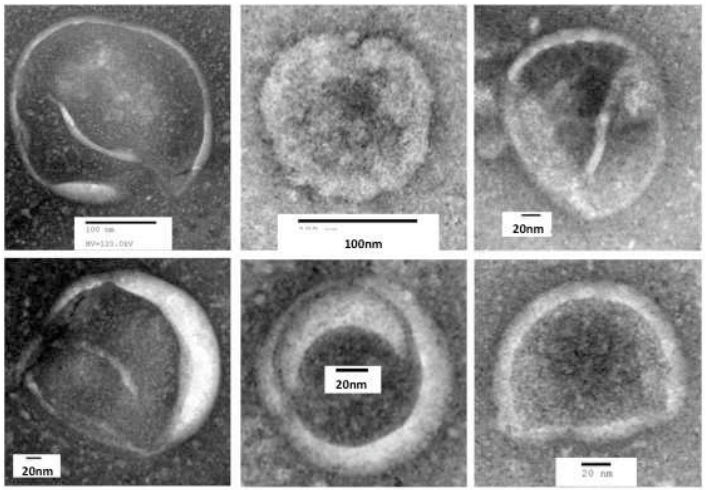
A



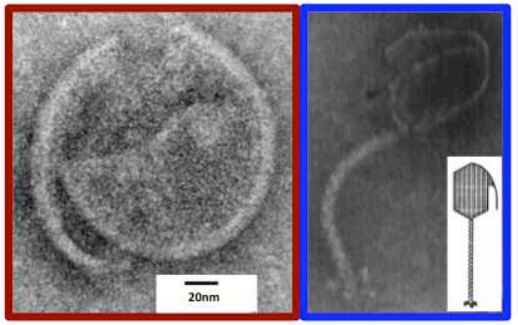
B



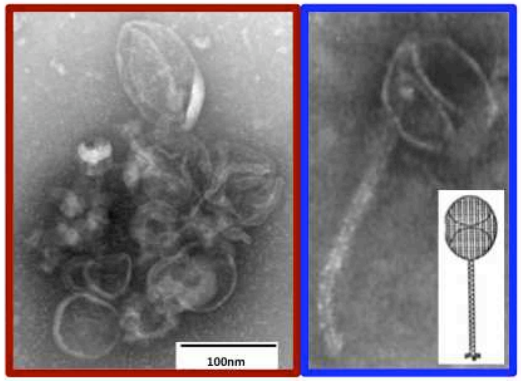
C



D



E

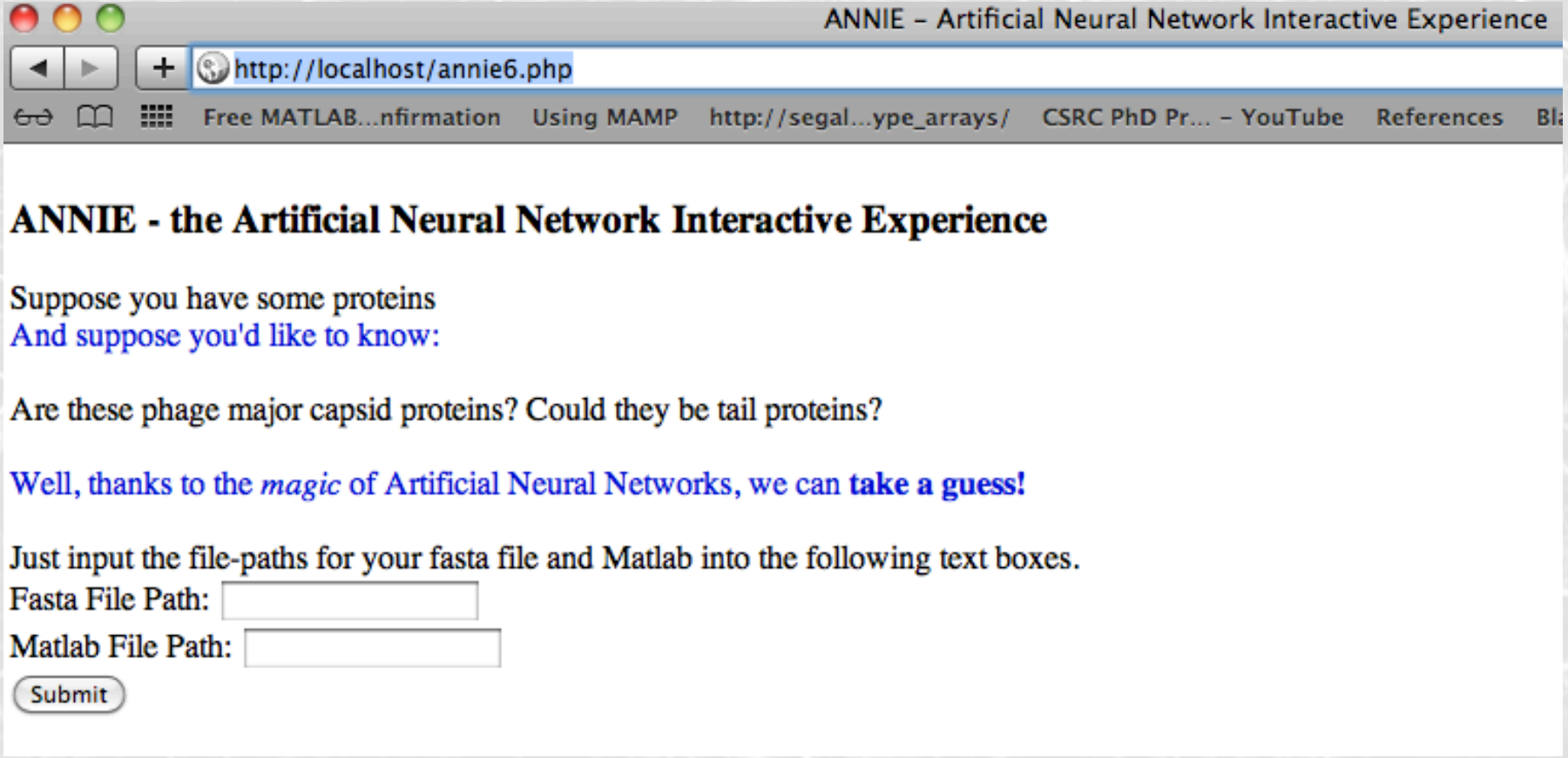


Mike's
Phage ANN Interactive Experience
(Ph-ANNIE)

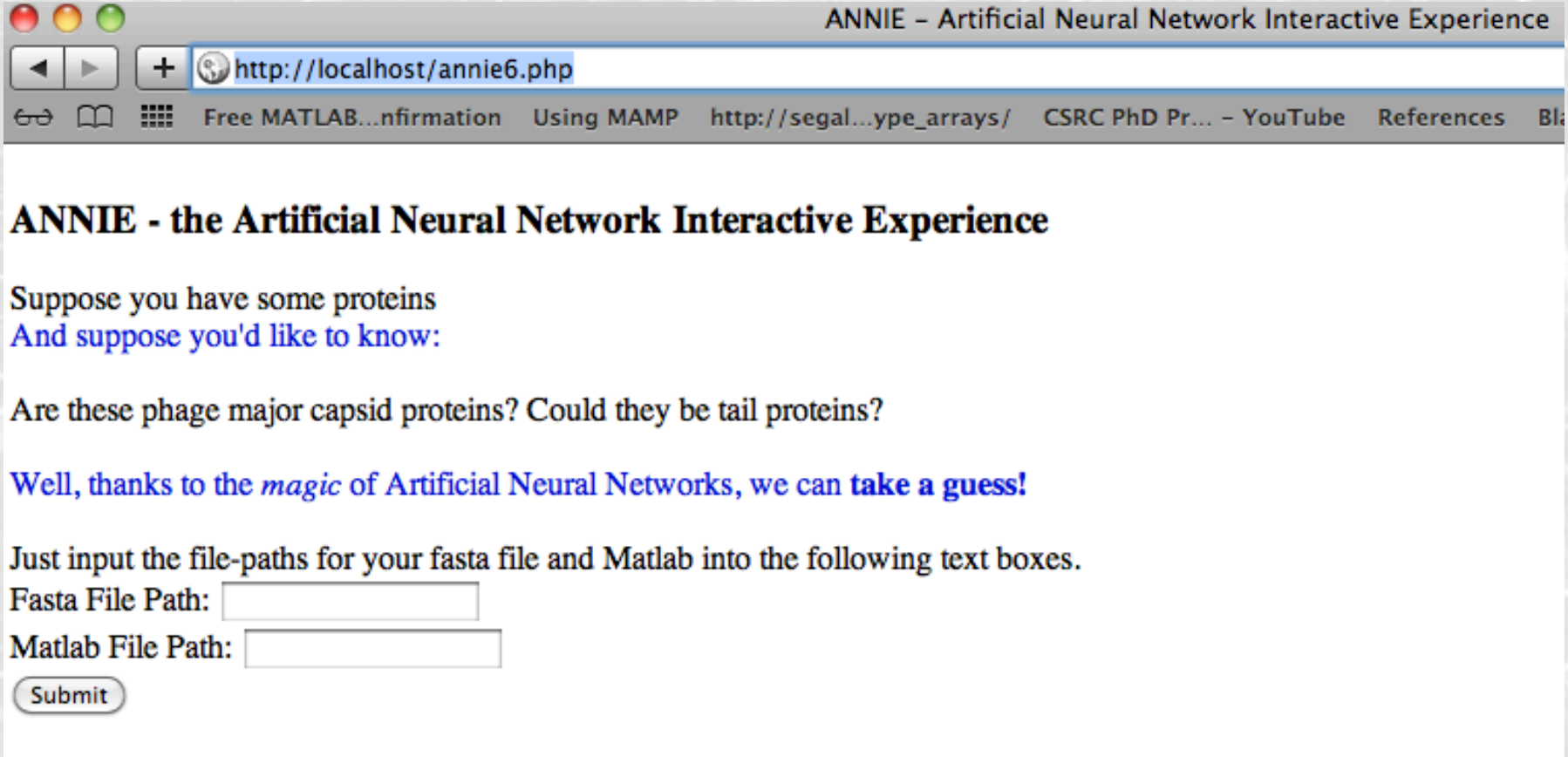
We have developed software which trains Artificial Neural Networks, to recognize proteins according to their Amino Acid Percent Compositions, and Isoelectric Values.

Software must be available to the public, to train new ANNs and to predict unknown proteins.

Currently, a GUI is in development, to predict unknown proteins as phage Major Capsid Proteins, or Tail Proteins



If provided with filepaths for the fasta file and the matlab program on the user's machine, the proteins are converted into AA-comp and pI data…



PCT+PI data:

0.0821 0.0000 0.0075 0.0299 0.0448 0.0896 0.0075 0.0597 0.0448 0.0522 0.0149 0.0896 0.0522 0.0448 0.0373 0.1194 0.0896 0.0746 0.0075 0.0522 10.18
0.1073 0.0000 0.0429 0.0987 0.0365 0.0665 0.0064 0.0644 0.0687 0.1030 0.0236 0.0601 0.0408 0.0408 0.0429 0.0558 0.0579 0.0579 0.0086 0.0172 4.75
0.0755 0.0000 0.0566 0.0943 0.0404 0.0512 0.0027 0.0620 0.0728 0.0916 0.0189 0.0593 0.0296 0.0647 0.0566 0.0458 0.0647 0.0916 0.0081 0.0135 4.88
0.0769 0.0051 0.0513 0.0769 0.0462 0.0872 0.0103 0.0564 0.0410 0.0718 0.0256 0.0769 0.0615 0.0256 0.0359 0.0410 0.0974 0.0667 0.0103 0.0359 4.36
0.0727 0.0038 0.0688 0.0650 0.0229 0.0860 0.0096 0.0650 0.0402 0.0593 0.0153 0.0650 0.0229 0.0593 0.0306 0.0956 0.0975 0.0554 0.0076 0.0574 4.18
0.0655 0.0106 0.0673 0.0549 0.0336 0.0726 0.0177 0.0496 0.0531 0.0779 0.0283 0.0832 0.0584 0.0283 0.0389 0.0655 0.0655 0.0442 0.0283 0.0566 4.93
0.0771 0.0021 0.0899 0.1199 0.0407 0.0857 0.0278 0.0664 0.0343 0.0685 0.0107 0.0428 0.0471 0.0428 0.0471 0.0343 0.0578 0.0642 0.0128 0.0278 4.05
0.0825 0.0063 0.1048 0.1460 0.0222 0.0762 0.0317 0.0603 0.0127 0.0857 0.0032 0.0159 0.0254 0.0444 0.0571 0.0825 0.0540 0.0635 0.0127 0.0127 3.86
0.0885 0.0066 0.1311 0.1344 0.0426 0.0623 0.0328 0.0361 0.0393 0.0525 0.0197 0.0230 0.0492 0.0262 0.0361 0.0492 0.0492 0.0754 0.0164 0.0295 3.84
0.1043 0.0000 0.1257 0.1203 0.0321 0.0722 0.0080 0.0374 0.0455 0.0348 0.0160 0.0267 0.0294 0.0561 0.0455 0.0882 0.0802 0.0401 0.0241 0.0134 3.83
0.0841 0.0000 0.1340 0.1028 0.0249 0.0561 0.0125 0.0654 0.0218 0.0966 0.0218 0.0530 0.0436 0.0156 0.0530 0.0467 0.0779 0.0530 0.0187 0.0187 3.76
0.0892 0.0000 0.1385 0.1127 0.0329 0.0751 0.0117 0.0423 0.0164 0.0563 0.0258 0.0376 0.0399 0.0258 0.0258 0.0681 0.0634 0.0986 0.0164 0.0235 3.43
0.0231 0.0000 0.1769 0.0923 0.0385 0.0692 0.0231 0.0462 0.0000 0.0615 0.0154 0.0000 0.0462 0.0385 0.0462 0.1154 0.1000 0.0692 0.0000 0.0385 3.54
0.0791 0.0000 0.0237 0.0158 0.0356 0.0672 0.0119 0.0632 0.0395 0.0988 0.0277 0.0395 0.0553 0.0277 0.0198 0.0830 0.1107 0.1423 0.0079 0.0514 9.65
0.1992 0.0000 0.0081 0.0407 0.0407 0.0894 0.0041 0.0772 0.0163 0.1748 0.0325 0.0203 0.0488 0.0122 0.0325 0.0407 0.0325 0.0976 0.0122 0.0203 7.52
0.0957 0.0000 0.0319 0.0406 0.0261 0.0638 0.0058 0.0841 0.0348 0.0812 0.0145 0.0551 0.0667 0.0609 0.0348 0.0783 0.0725 0.0754 0.0087 0.0696 6.55
0.1667 0.0000 0.0069 0.0694 0.0278 0.0694 0.0139 0.0347 0.0833 0.0764 0.0278 0.0556 0.0764 0.0486 0.0347 0.0486 0.0417 0.0556 0.0069 0.0556 9.98
0.0708 0.0354 0.0708 0.0708 0.0088 0.0531 0.0177 0.1416 0.0885 0.0619 0.0265 0.0354 0.0265 0.0265 0.0885 0.0531 0.0619 0.0177 0.0000 0.0442 8.92
0.0216 0.0000 0.0791 0.0791 0.0432 0.0647 0.0000 0.1295 0.0360 0.1007 0.0360 0.0072 0.0576 0.0360 0.0935 0.0360 0.0360 0.0791 0.0144 0.0504 4.67
0.1062 0.0000 0.0375 0.0437 0.0625 0.0437 0.0000 0.0625 0.0625 0.1062 0.0312 0.0500 0.0375 0.0250 0.0250 0.0500 0.0813 0.0813 0.0125 0.0813 8.63
0.0364 0.0364 0.0409 0.0318 0.0409 0.0909 0.0136 0.0682 0.0273 0.0773 0.0136 0.0818 0.0545 0.0364 0.0091 0.0955 0.0818 0.0409 0.0091 0.1136 4.35

< M A T L A B (R) > Copyright 1984-2010 The MathWorks, Inc. Version 7.11.0.584 (R2010b) 64-bit (maci64) August 16, 2010 To get started, type one of these: helpwin,

The AA-comp and pI data are provided to a matlab script, which uses pre-trained ANNs, to generate predictions as to protein function.

The Matlab script generates a csv file, in which each predicted protein is represented by row, and each type of ANN is represented by column…

	A	B	C	D	E	F
1	-0.93683	-0.98231	-0.96846	-0.97849	-0.96517	-0.99705
2	-0.99624	-0.99761	-0.99225	-0.99853	-0.99012	-0.99942
3	0.77531	0.7701	0.51864	0.42647	0.46029	-0.081085
4	-0.15756	0.0045422	-0.38005	-0.3447	-0.29112	-0.74178
5	0.71443	0.32266	0.23753	0.25164	-0.032866	-0.25807
6	-0.75167	-0.85966	-0.88113	-0.90151	-0.90575	-0.96731
7	-0.37702	-0.70048	-0.72004	-0.85689	-0.78607	-0.97668
8	0.64417	0.045028	0.1869	-0.16083	-0.34331	-0.64696
9	-0.83235	-0.95784	-0.89619	-0.9012	-0.88052	-0.97162
10	-0.026821	-0.42073	-0.56978	-0.65562	-0.73061	-0.94066
11	-0.92887	-0.97564	-0.97294	-0.97049	-0.98446	-0.9973
12	-0.82382	-0.88519	-0.94914	-0.96145	-0.94392	-0.99408

The scripts still require some changes, which will allow the fasta annotations and ANN type names to be included, in the output file.

	A	B	C	D	E
1	ANN_PREDICTIONS	MCP_ONE	MCP_TWO	MCP_THREE	MCP_FOUR
2	>gi|9634678|ref|NP_038458.1| capsid protein (p38) [Japanese iris necrotic ring virus]	-0.93683	-0.98231	-0.96846	-0.97849
3	>gi|66090971|ref|YP_233104.1| coat protein [Cassia yellow blotch virus]	-0.99624	-0.99761	-0.99225	-0.99853
4	>gi|115350055|ref|YP_762620.1| coat protein [Streptocarpus flower break virus]	0.77531	0.7701	0.51864	0.42647
5	>gi|119964533|ref|YP_950424.1| 18 kDa coat protein [Maracuja mosaic virus]	-0.15756	0.0045422	-0.38005	-0.3447
6	>gi|145651768|ref|YP_001165305.1| coat protein [Phlox virus S]	0.71443	0.32266	0.23753	0.25164
7	>gi|70980517|ref|YP_263307.1| coat protein [Lily virus X]	-0.75167	-0.85966	-0.88113	-0.90151
8	>gi|62326909|ref|YP_224088.1| coat protein [Hydrangea ringspot virus]	-0.37702	-0.70048	-0.72004	-0.85689
9	>gi|94971814|ref|YP_595731.1| coat protein [Daphne virus S]	0.64417	0.045028	0.1869	-0.16083
10	>gi|56692635|ref|YP_164805.1| coat protein [Fragaria chiloensis latent virus]	-0.83235	-0.95784	-0.89619	-0.9012
11	>gi|75750464|ref|YP_319831.1| coat protein [Alstroemeria virus x]	-0.026821	-0.42073	-0.56978	-0.65562
12	>gi|83999993|ref|YP_446996.1| coat protein [Nerine virus X]	-0.92887	-0.97564	-0.97294	-0.97049

Metaproteomics Update

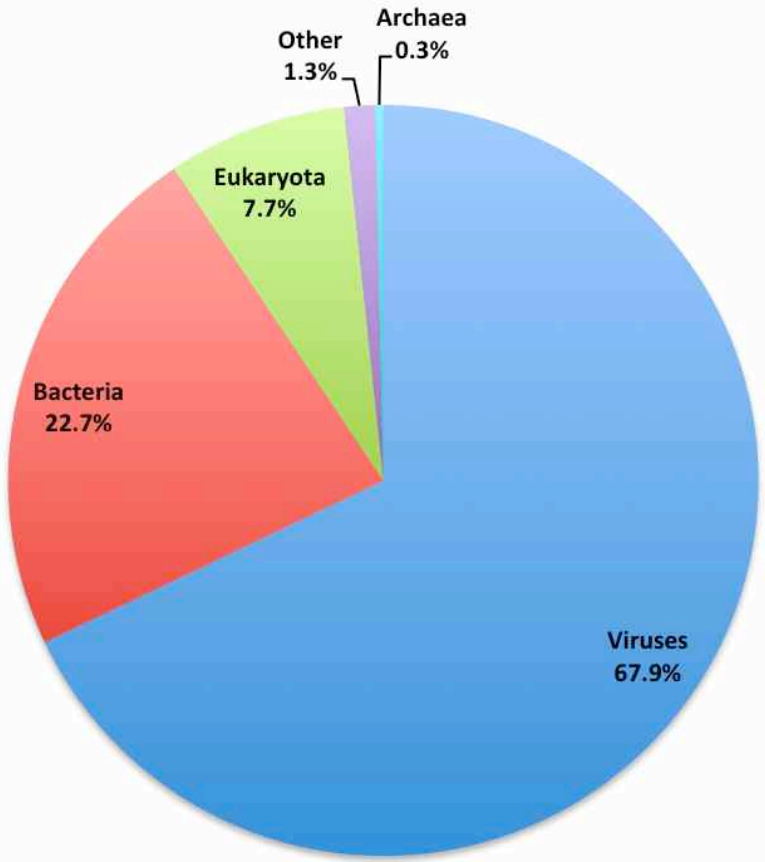
- Analysis of raw reads (MGRAST)

CAR9 (MGRAST)

Analysis Statistics

Upload: Size	266,906,705 bp
Upload: Sequences Count	591,600
Upload: Mean Sequence Length	451 ± 159 bp
Upload: Mean GC percent	45 ± 7 %
Technical Duplicates: Sequence Count	190,428
Post QC: Size	163,602,764 bp
Post QC: Sequences Count	360,246
Post QC: Mean Sequence Length	454 ± 172 bp
Post QC: Mean GC percent	46 ± 7 %
Processed: Predicted Protein Features	217,700
Processed: Predicted rRNA Features	37,836
Alignment: Identified Protein Features	9,117
Alignment: Identified rRNA Features	23
Annotation: Identified Functional Categories	4,953

97.5% Unknown

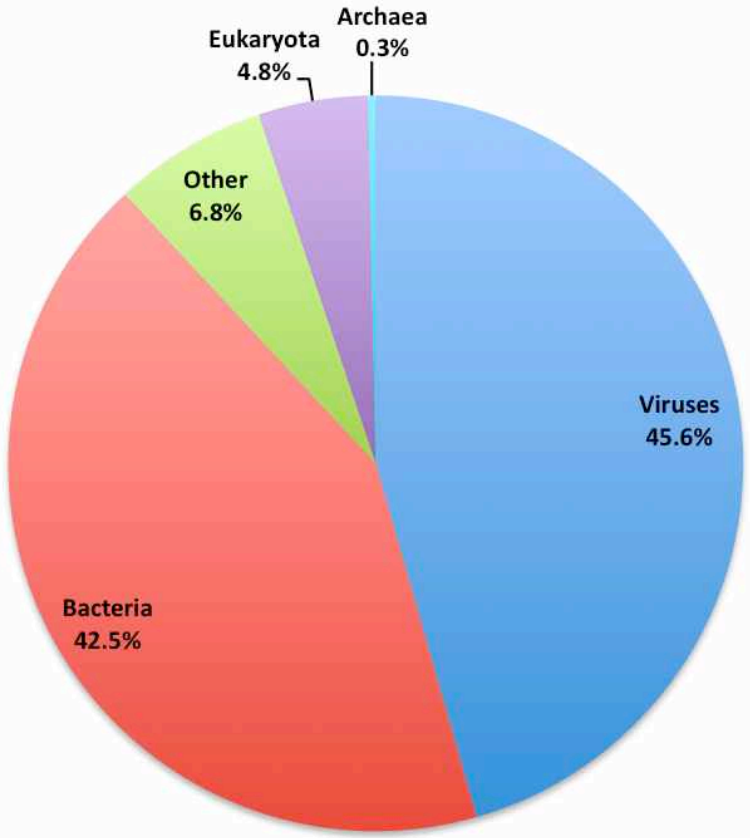


Analysis Statistics

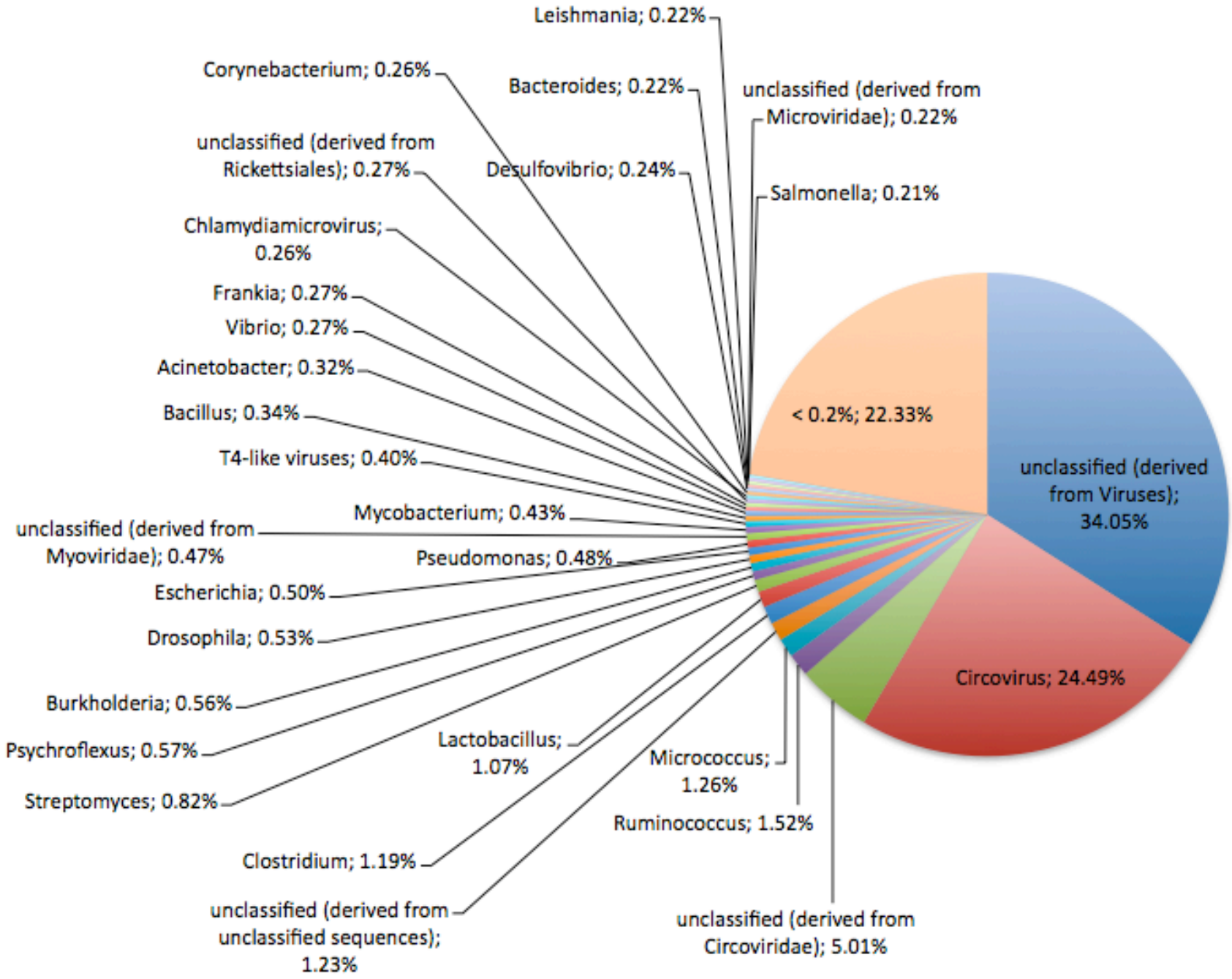
Upload: Size	371,341,354 bp
Upload: Sequences Count	939,311
Upload: Mean Sequence Length	395 ± 131 bp
Upload: Mean GC percent	44 ± 7 %
Technical Duplicates: Sequence Count	359,603
Post QC: Size	222,542,394 bp
Post QC: Sequences Count	579,664
Post QC: Mean Sequence Length	383 ± 147 bp
Post QC: Mean GC percent	42 ± 8 %
Processed: Predicted Protein Features	340,045
Processed: Predicted rRNA Features	57,515
Alignment: Identified Protein Features	42,218
Alignment: Identified rRNA Features	8
Annotation: Identified Functional Categories	18,419

STAR7 (MGRAST)

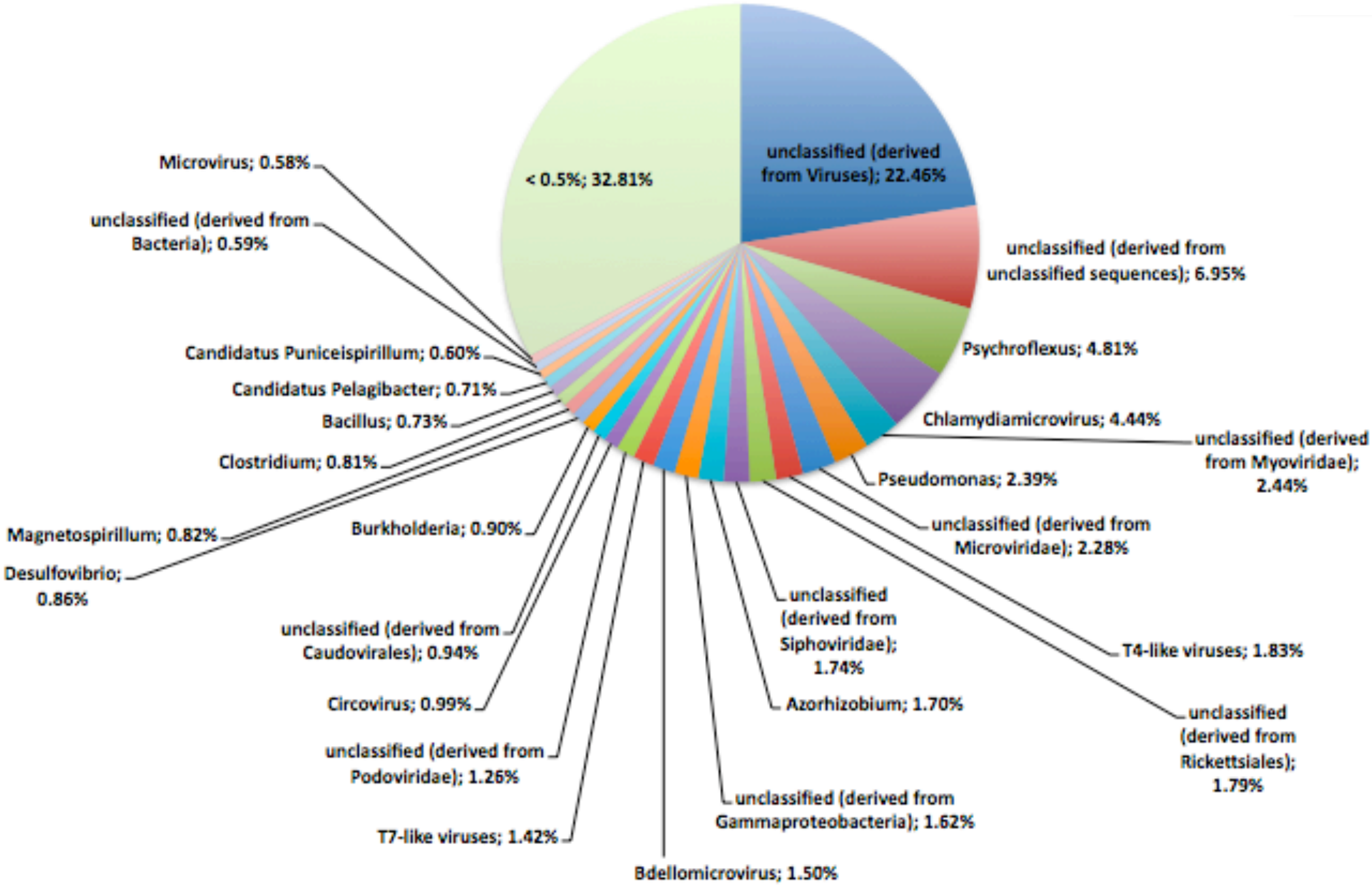
93% Unknown



CAR9 Tax Distribution by Genus



STAR7 Tax Distribution by Genus



To Do (Soon)

- ANN Paper
  - Resubmit next week
- Metaproteomics Paper
  - Add MGRAST results
  - Check accuracy of figures and tables
  - Format manuscript and figures for submission (PLoS Biology???)