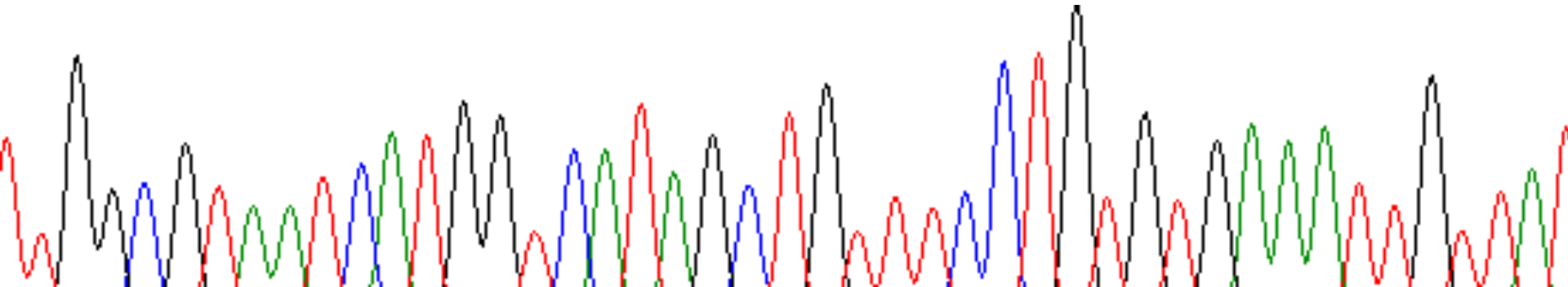
Accounting for Contamination in Sequence Data

Chloe Lee

How do you know what organism your DNA sequence is from?

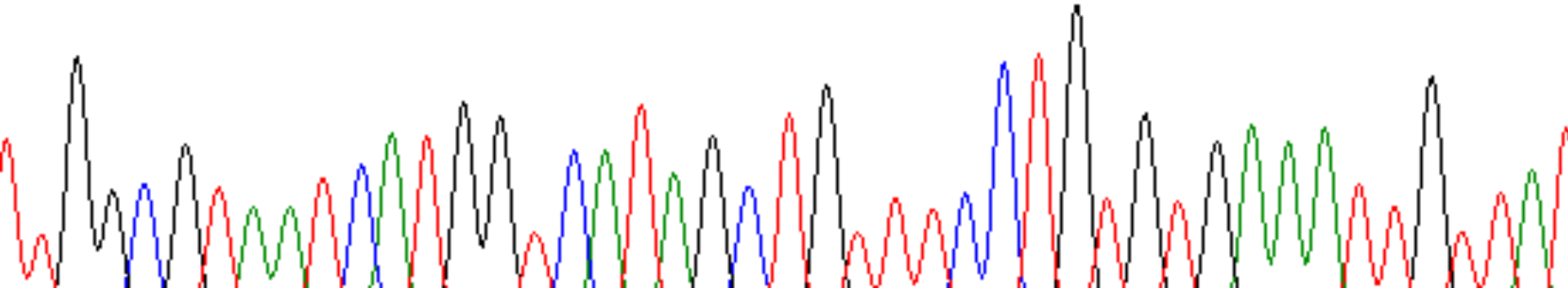


Terminology

- GC content
- GC3
- ORF
- contig
- chimera

How do you know what organism your DNA sequence is from?

- BLAST, but if your sequence is contaminated, what do you get out?
- GC content varies among different organisms
- Translational selection for highly expressed genes influences GC content
- Mutational selection



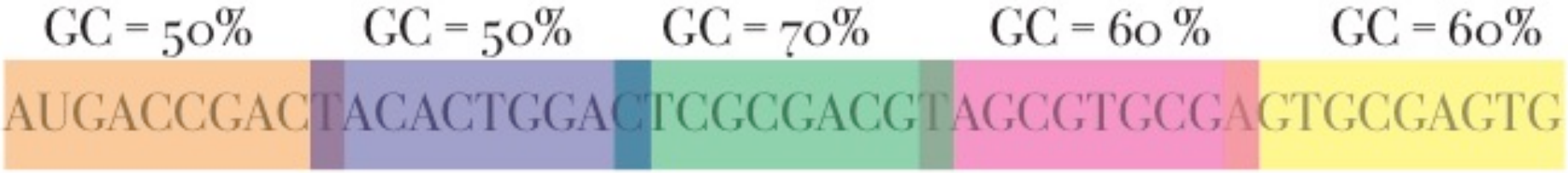
Reasoning for project:

- *Childonella uncinata* as a model organism for genomic architechture and genome evolution
- MAC single gene chromosomes
- Paralogs
- Endosymbionts?

Procedure

- Take assembly from SPAdes
  - Identify contaminants
  - Asses validity of assembly
- Analyze GC content of sequence
- Identify potential protein coding regions
- Determine the effective number of codons for those ORFs

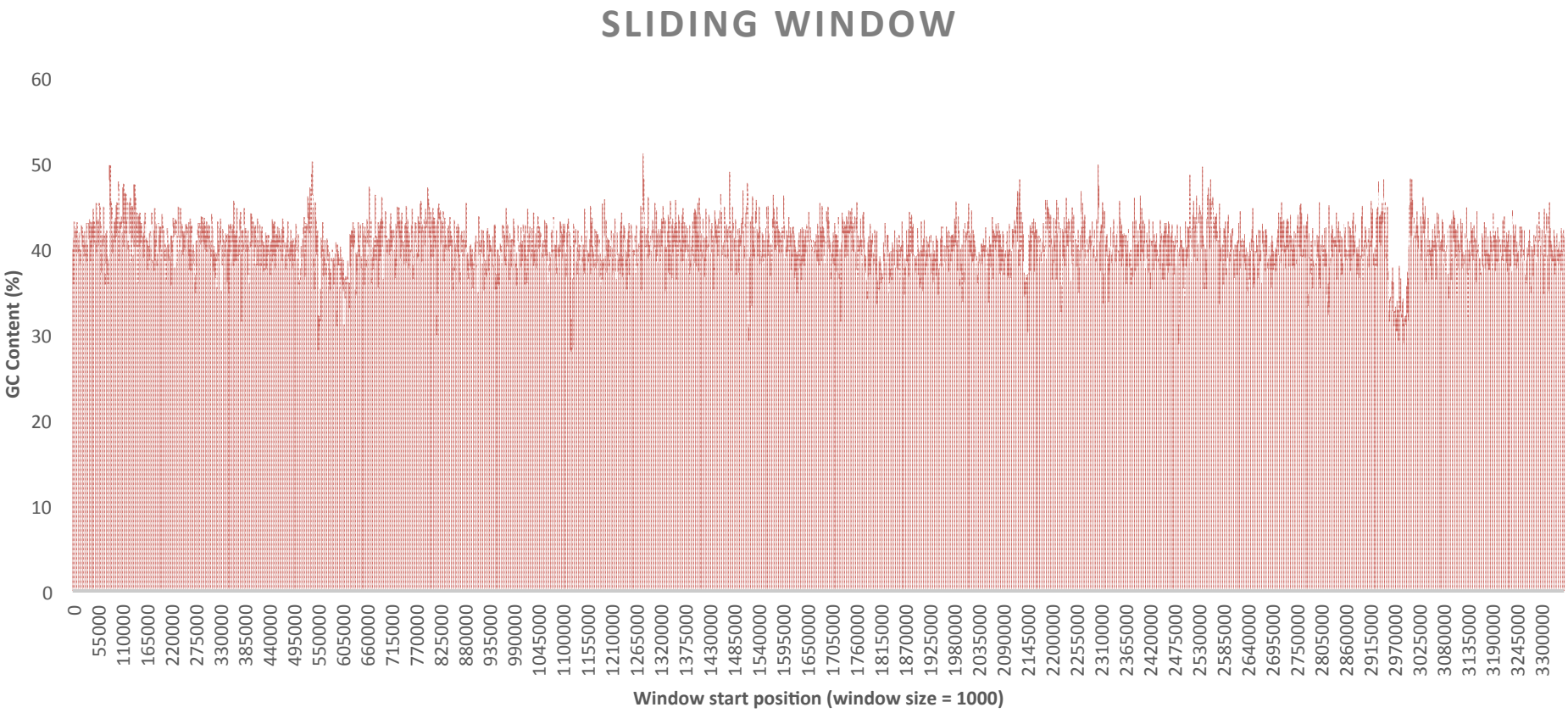
GC content



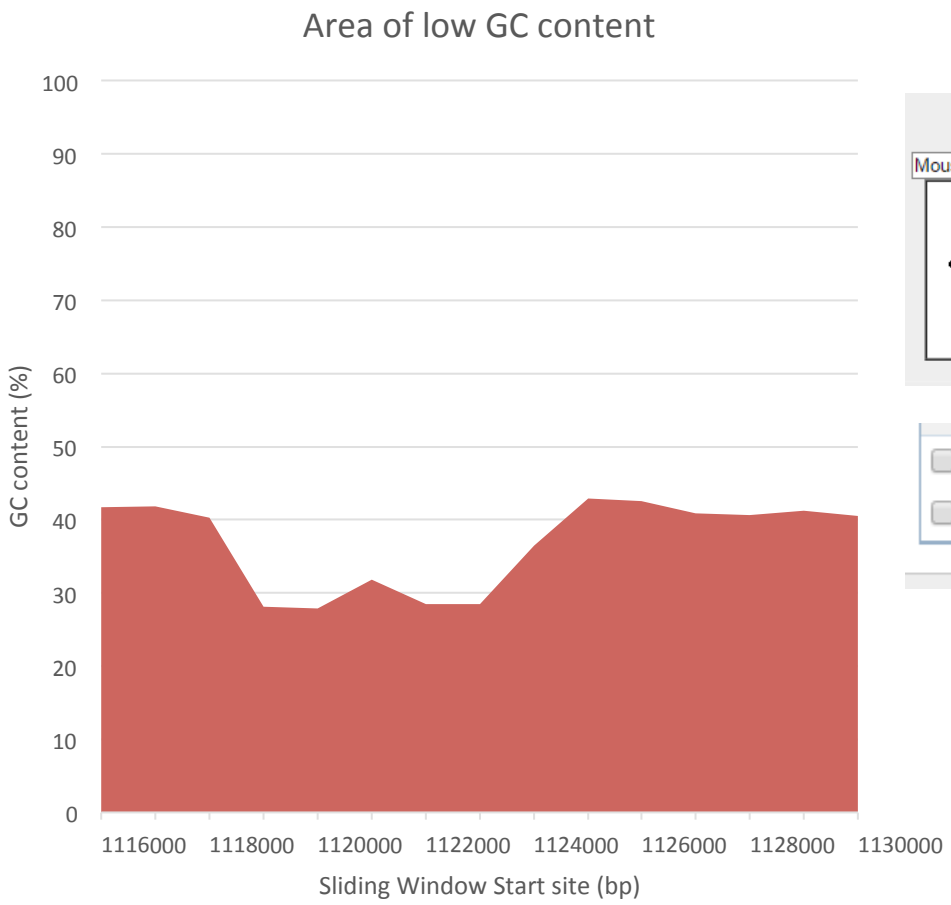
- Genes are characterized as having higher GC content than genome as a whole
- Longer coding sequences associated with higher GC content
- GC content variable among organisms

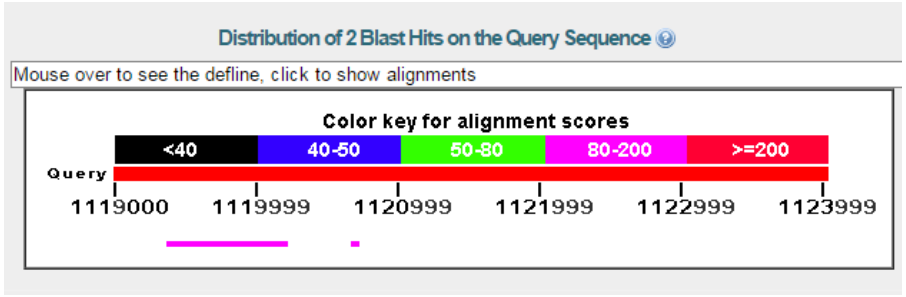
Implementation:

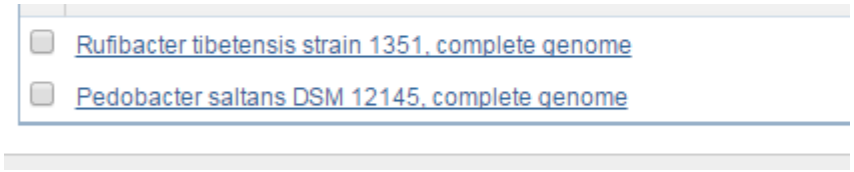
- Sliding window script in python


SLIDING WINDOW
GC Content (%)
60
50
40
30
20
10
0
0
55000
110000
165000
220000
275000
330000
385000
440000
495000
550000
605000
660000
715000
770000
825000
880000
935000
990000
1045000
1100000
1155000
1210000
1265000
1320000
1375000
1430000
1485000
1540000
1595000
1650000
1705000
1760000
1815000
1870000
1925000
1980000
2035000
2090000
2145000
2200000
2255000
2310000
2365000
2420000
2475000
2530000
2585000
2640000
2695000
2750000
2805000
2860000
2915000
2970000
3025000
3080000
3135000
3190000
3245000
3300000
Window start position (window size = 1000)

ZOOMING IN ON AREAS OF LOW OR HIGH GC







Glimmer

Gene Locating Interpolated Markov Model

- Linear combination of 8 markov chains
- 3 different Markov Models
- Sequence score is based on probabilities of bases in sequence
- Designed for working with gene rich sequences

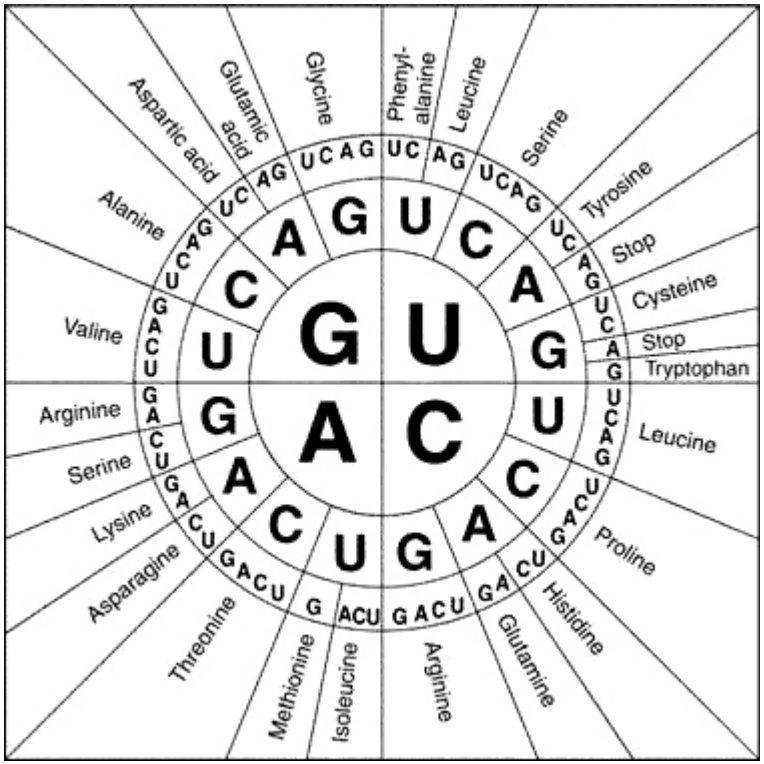
Glimmer output file:

- **ORF ID ▪ Start point ▪ End point ▪ Frame ▪ Raw Score**

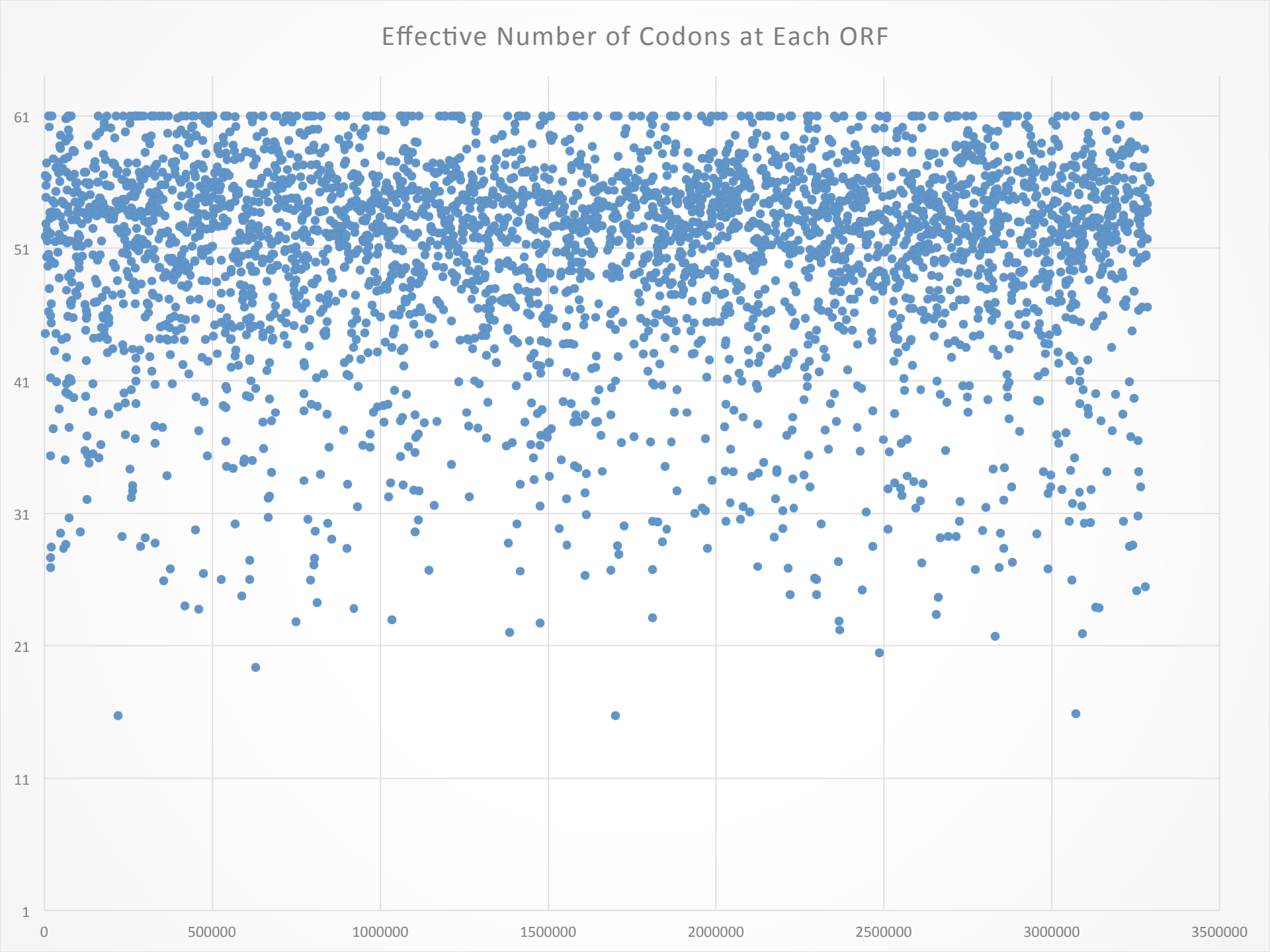
```
>NODE_1_length_3297883_cov_269.809_ID_1
orf00001  3295745         9  +1     14.89
orf00002      153       881  +3     12.32
orf00004     1893       997  -1     17.83
orf00005     2349      1900  -1     18.73
orf00006     3615      2359  -1     13.95
orf00008     3811      3602  -2     16.59
orf00010     3846      5303  +3     15.28
orf00011     5371      5955  +1     16.80
orf00013     5989      7014  +1     19.15
orf00014     7017      7799  +3     17.83
orf00016     7822      8442  +1     17.23
orf00018     8445      9629  +3     15.11
orf00021     9645     10415  +3     17.59
orf00022    10419     11012  +3     16.08
```

Nc

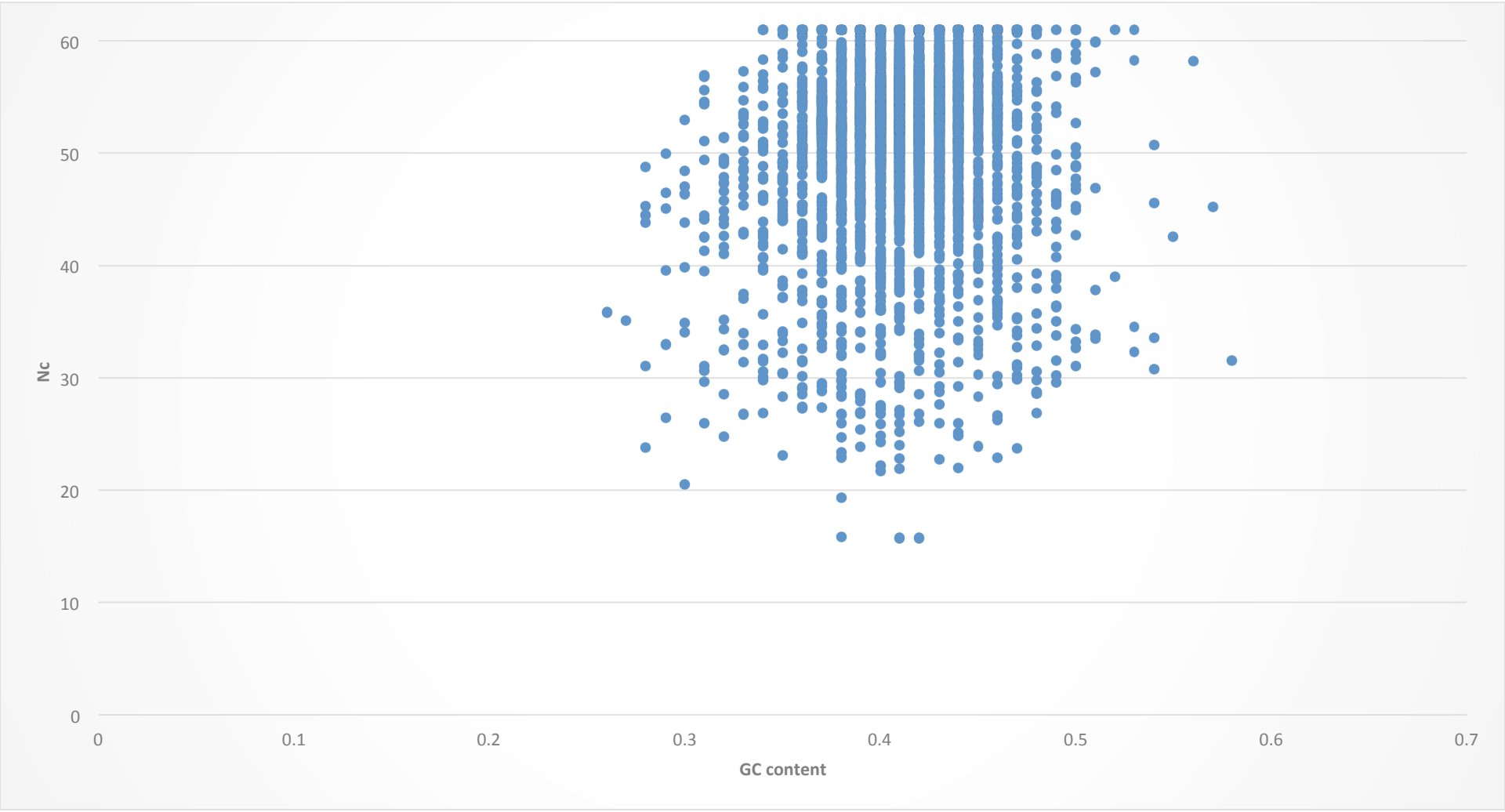
Effective number of codons



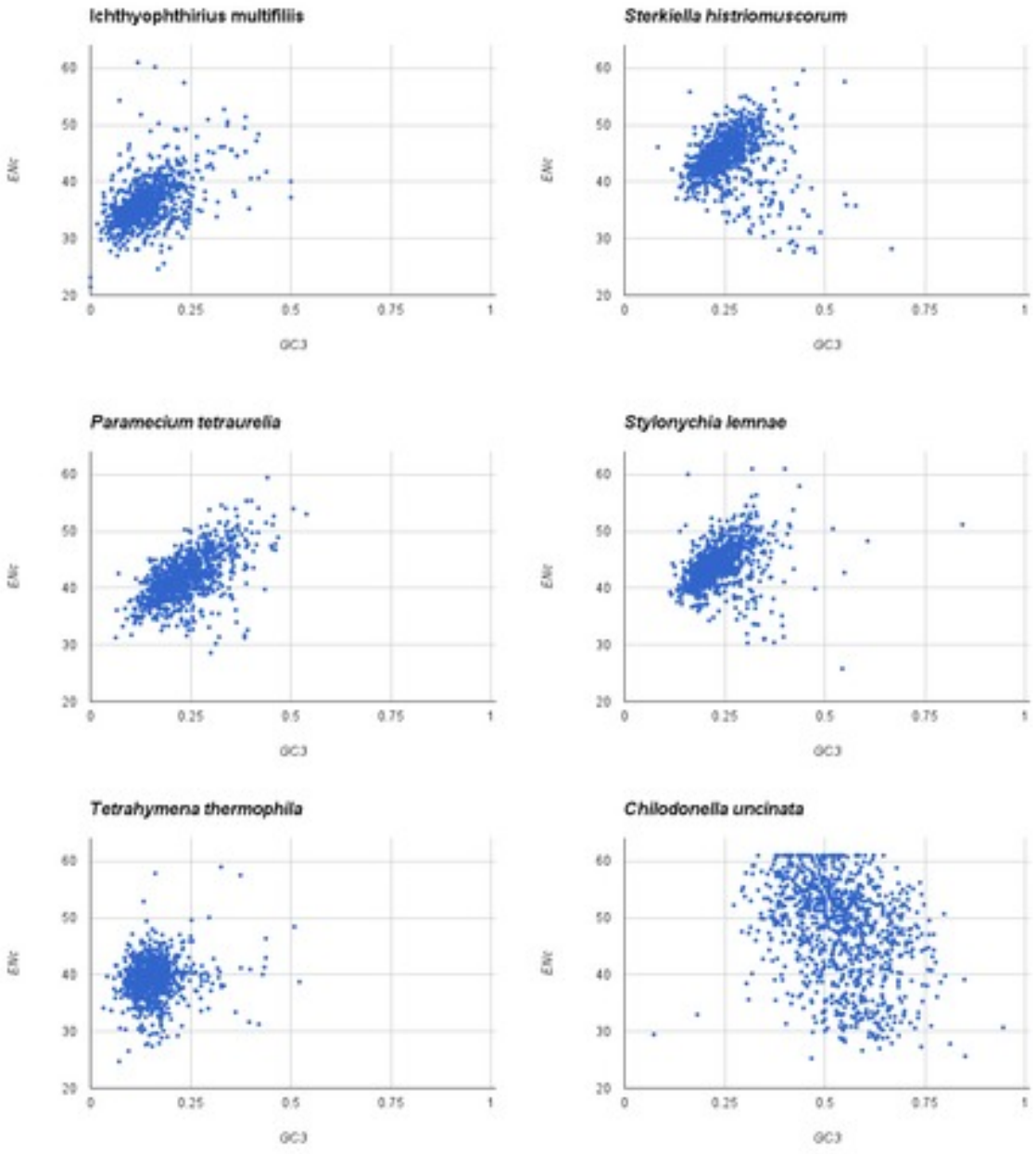
- Evaluated using the program chips on the ORFs found by glimmer
- low values in highly biased genes
- high values in lowly biased genes


Effective Number of Codons at Each ORF
61
51
41
31
21
11
1
0
500000
1000000
1500000
2000000
2500000
3000000
3500000

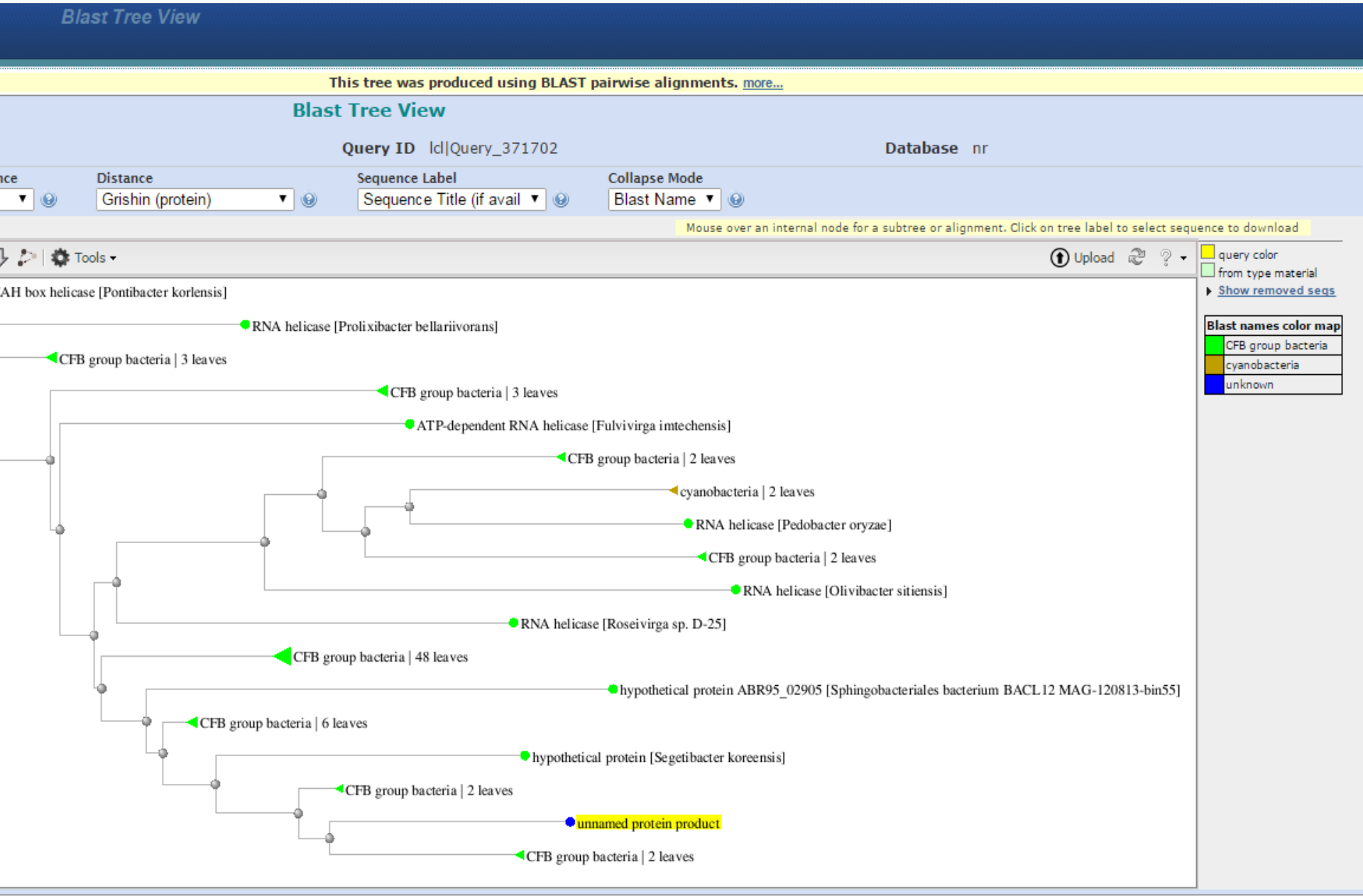
Effective number of codons vs. GC content of glimmer predicted orfs



Examples


Ichthyophthirius multifiliis
Sterkiella histriomuscorum
Paramecium tetraurelia
Stylonychia lemnae
Tetrahymena thermophila
Chilodonella uncinata
ENc
GC3
20
30
40
50
60
0
0.25
0.5
0.75
1

Blast of ORFs with high raw score



Goals

- Make scripts user-friendly
- Learn how to use matplotlib
- Calculate GC3 instead of GC
- Plot results of different contigs together
- Train Glimmer using both *Chilo* and contaminant clade predictions
- Investigate Glimmer-MG
- BLAST results – convert to protein sequences
- Isolate *Chilodonella uncinata* DNA from the assembly.

Sources

Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. Nucleiv Acids Res. 1998 Jan 15; 26(2):544-8.

NCBI BLAST

Glimmer 3.02

Emboss Chips

CodonW Sourceforge

Supek, Frank et al. Translational Selection is Ubiquitous in Prokaryotes. PLOS Genetics. 2010 June 25.

Xyrus Maurer-Alcala, Laura Katz

Cusp

#CdsCount: 50065

#Coding GC 41.09%
#1st letter GC 41.07%
#2nd letter GC 41.31%
#3rd letter GC 40.90%

#Codon	AA	Fraction	Frequency	Number
GCA	A	0.344	16.901	17035
GCC	A	0.236	11.604	11696
GCG	A	0.140	6.874	6929
GCT	A	0.280	13.730	13839
TGC	C	0.506	16.912	17046
TGT	C	0.494	16.510	16641
GAC	D	0.251	6.241	6291
GAT	D	0.749	18.647	18795
GAA	E	0.734	21.273	21442
GAG	E	0.266	7.718	7779