RefSeq genomes GC distribution

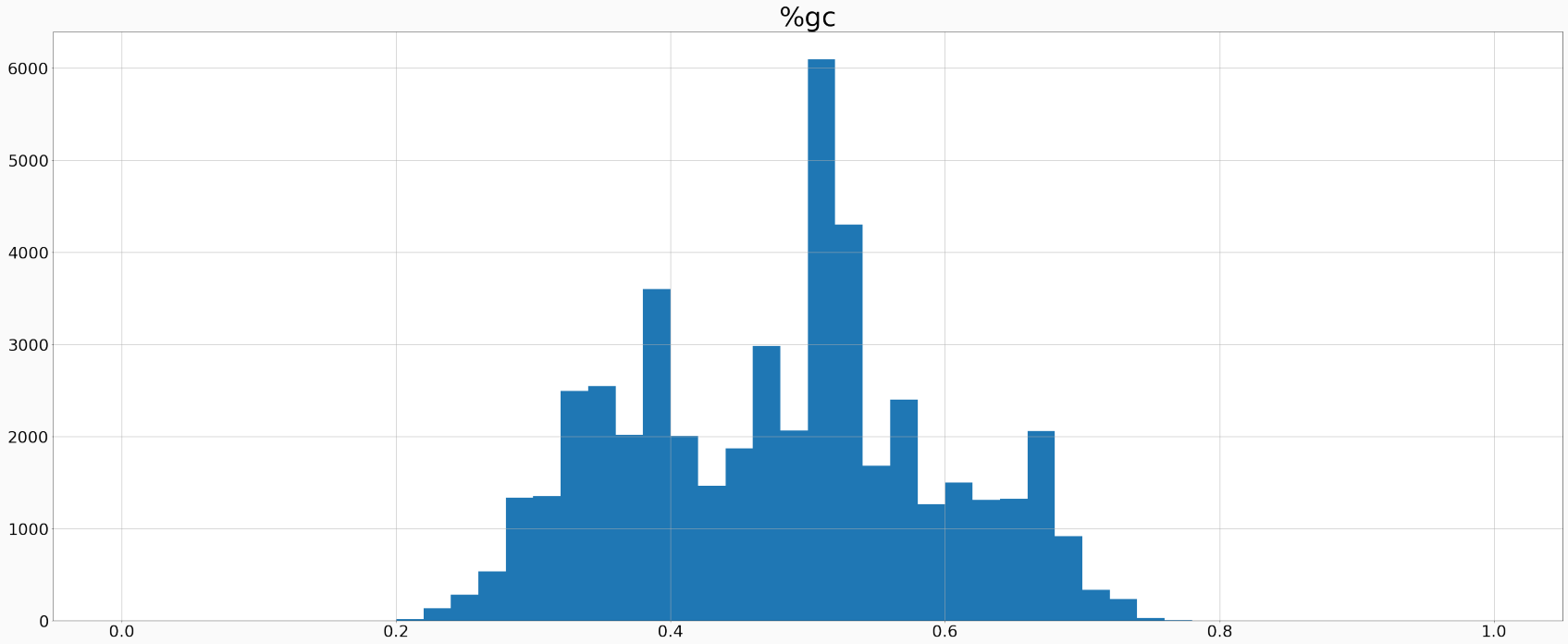
Aniket Mane

April 2022

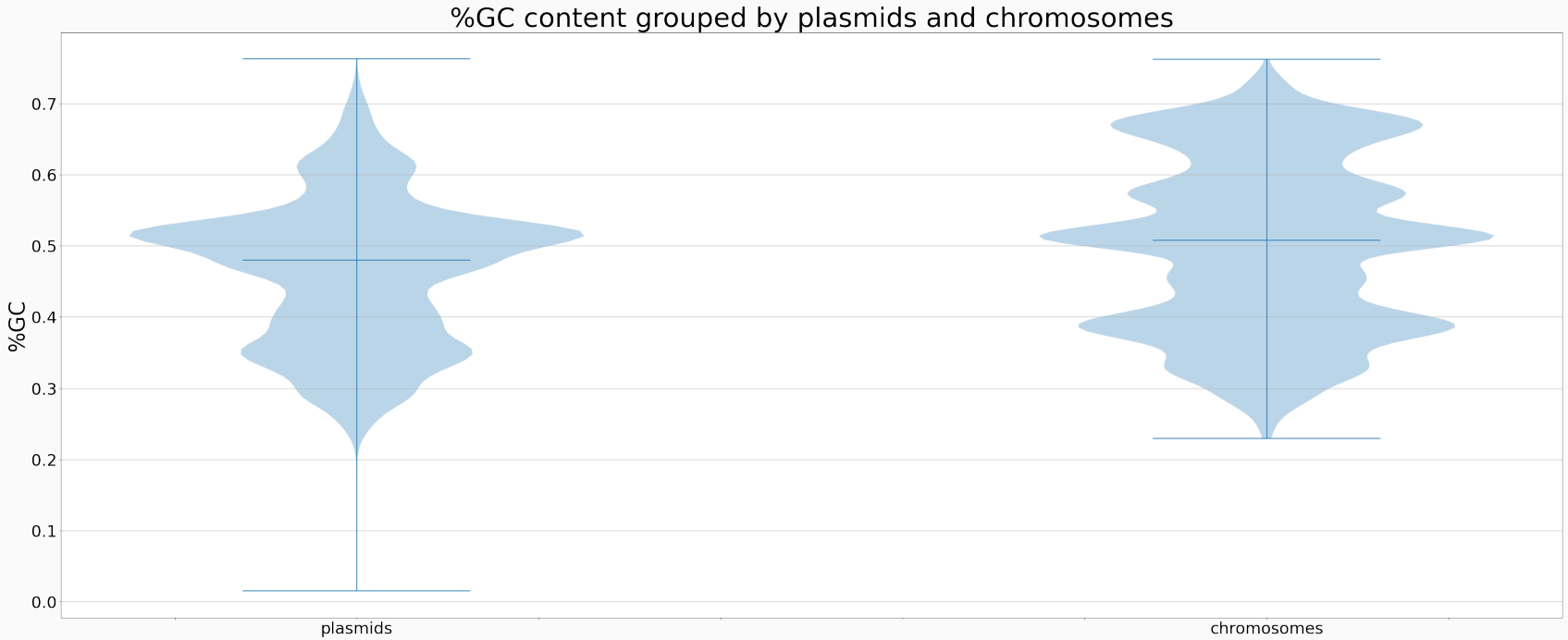
Data used

- Filters used:
  (bacteria[filter] AND (latest[filter] OR "latest refseq"[filter]) AND "complete genome"[filter] AND all[filter] AND "taxonomy check ok"[filter])
- Dataset: 22951 chromosomes, 25181 plasmids and 10 unclassified from 21720 samples
- IDs present in blastdb: 3211 chromosomes, 1850 plasmids and 1 ambiguous

GC distribution - RefSeq overall



GC distribution - Refseq by class

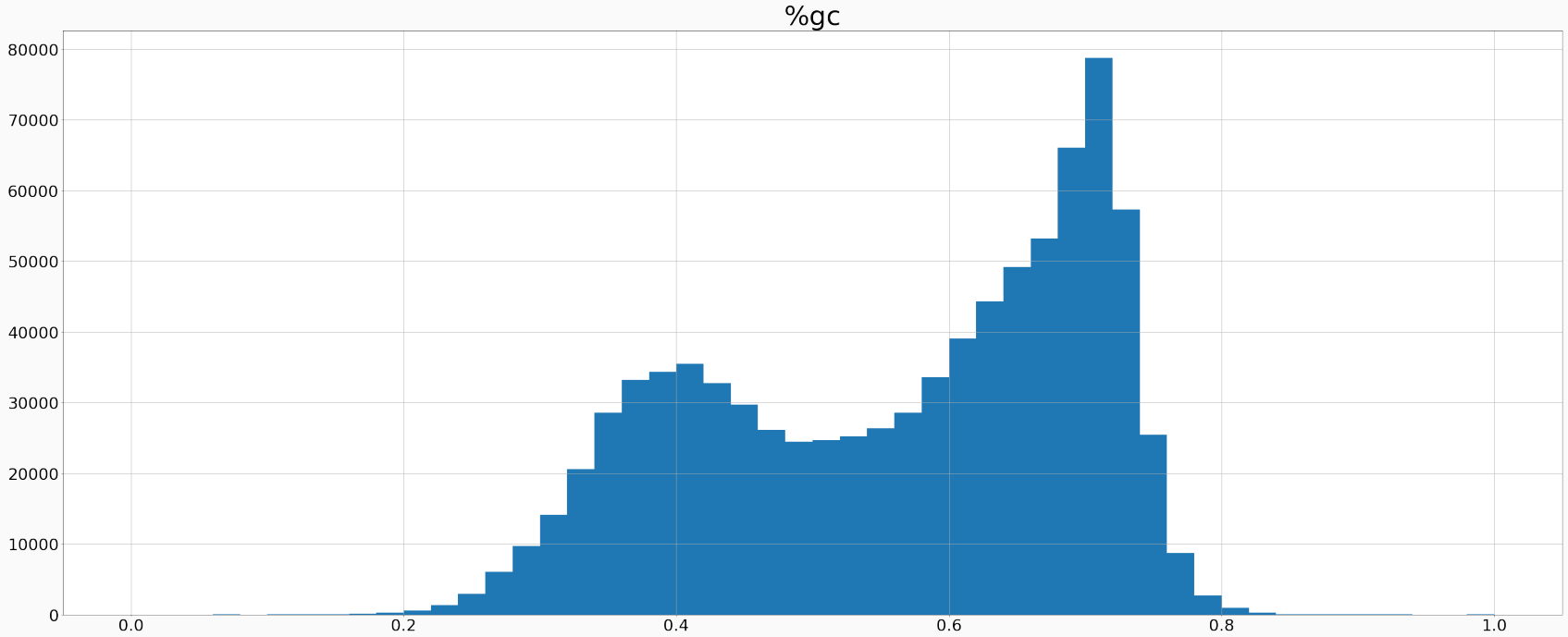


Plasmids show slightly lower GC content. Peaks around 0.35 and 0.5. For chromosomes, additional peak at 0.7.

GC distribution - blast db

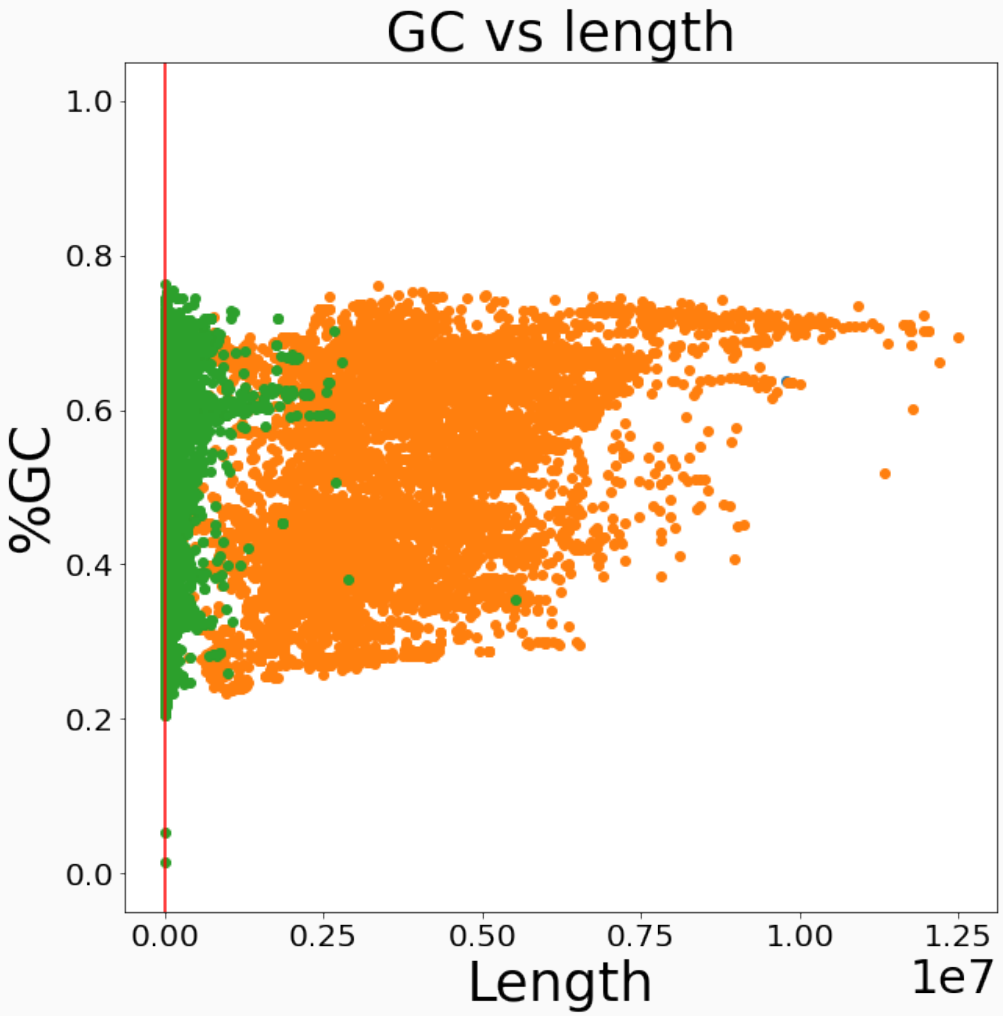
Query used (by Dr. Vinar):

blastdbcmd -entry all -db ref prok rep genomes



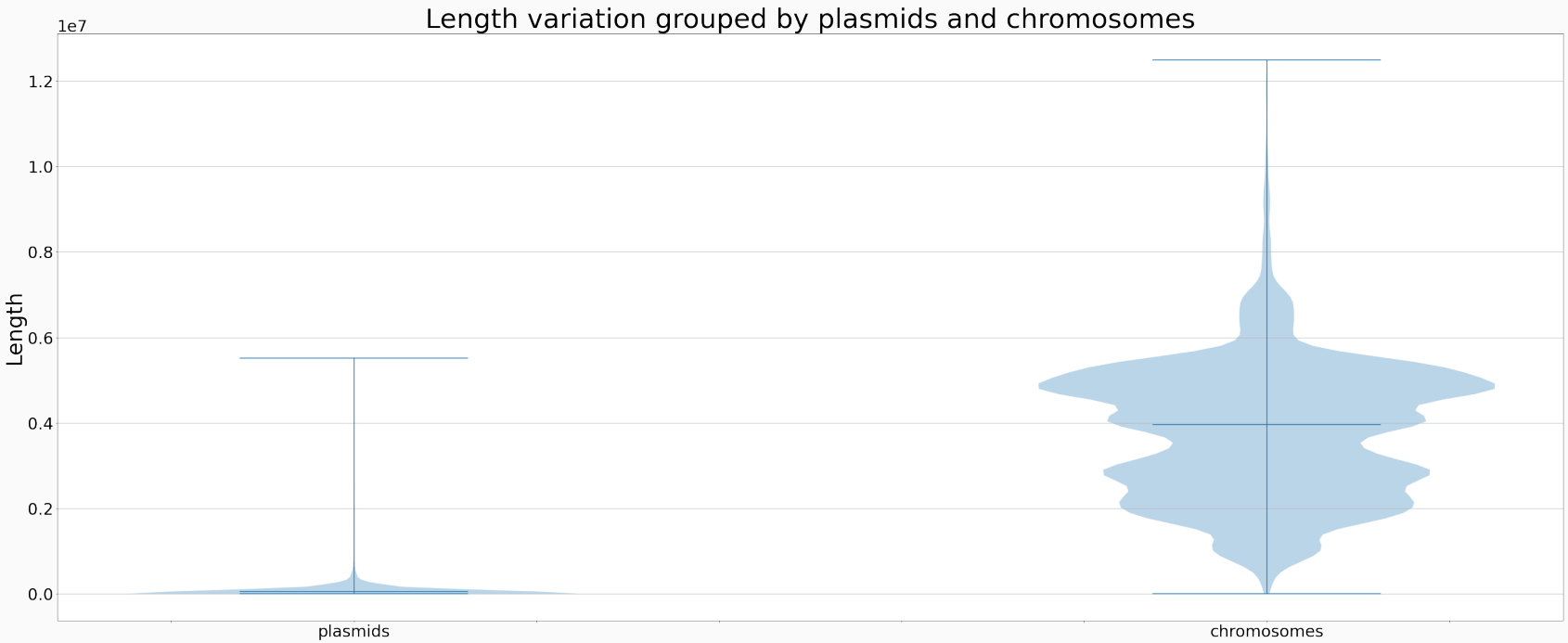
Blastdb contained 865435 entries. Peaks around 0.4 and 0.7

GC vs length

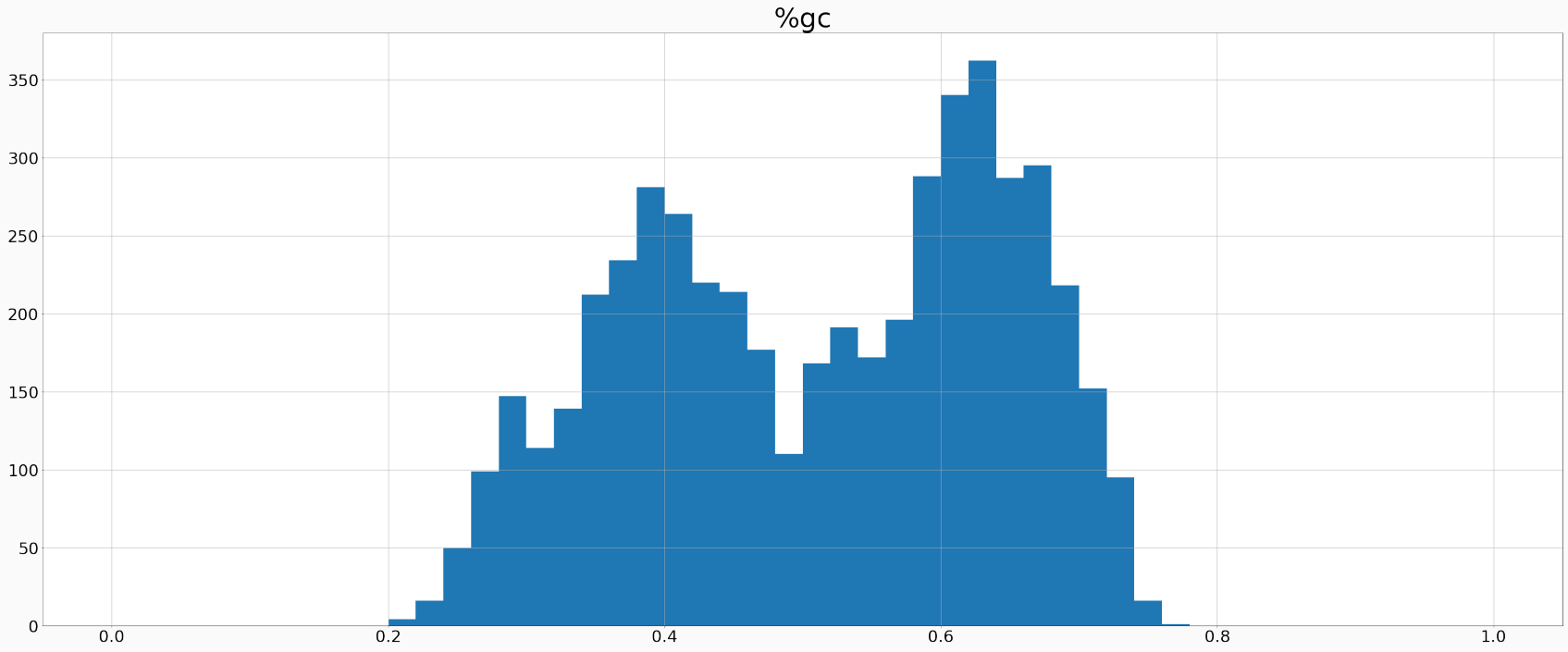


No discernable difference. Although longer plasmids and chromosomes seem to have higher GC content.

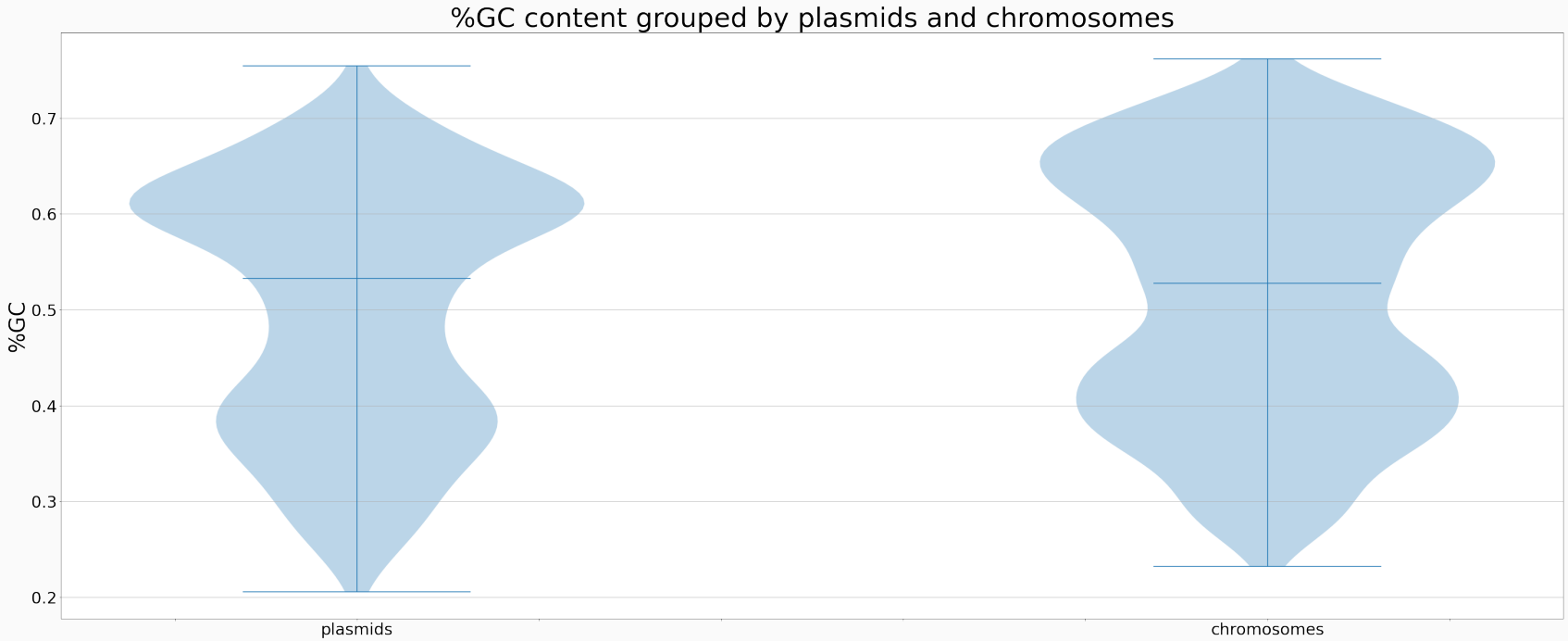
Length distribution



GC distribution - IDs in RefSeq and blastdb

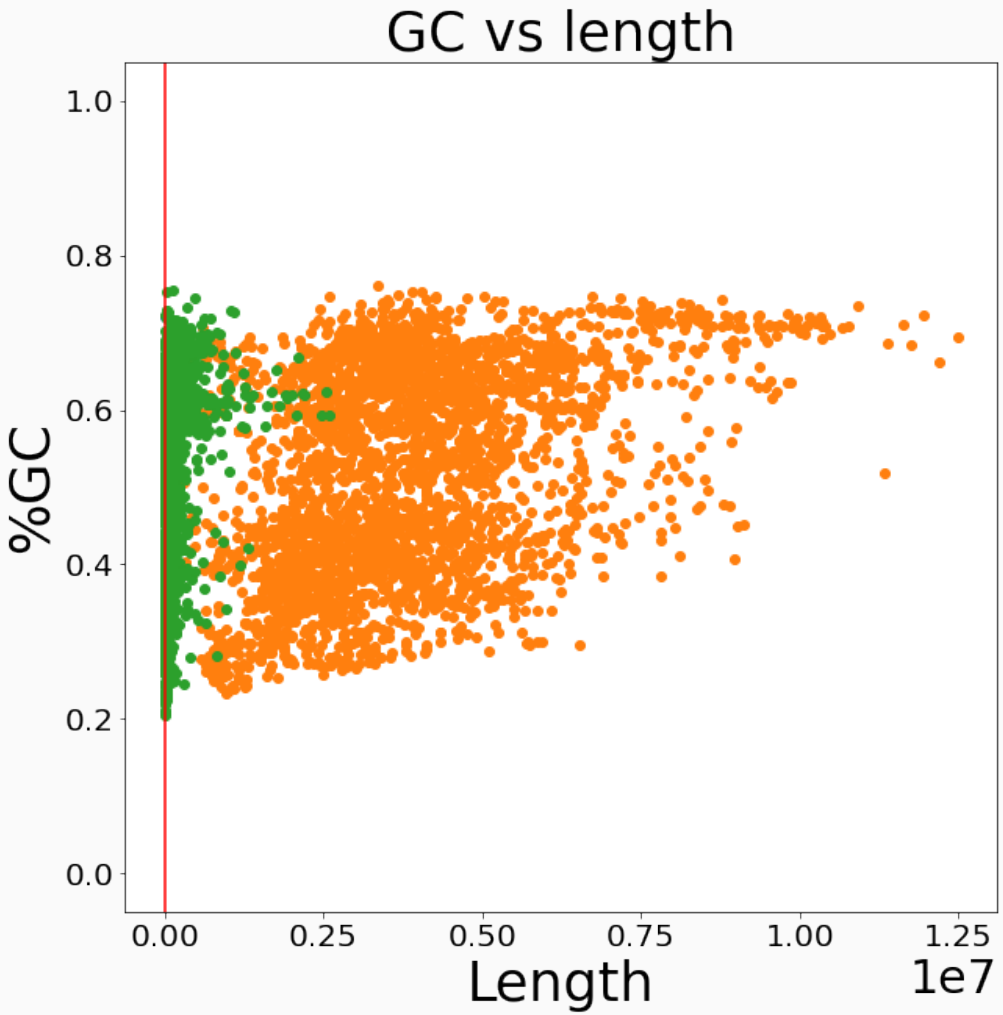


GC distribution - Refseq by class



Peaks around 0.4 for both. For chromosomes, additional peak at 0.7, for plasmids at 0.6.

GC vs length



Similar observation as before.