Ethan Hu 2024 Summer at Sym Geno Evo Lab

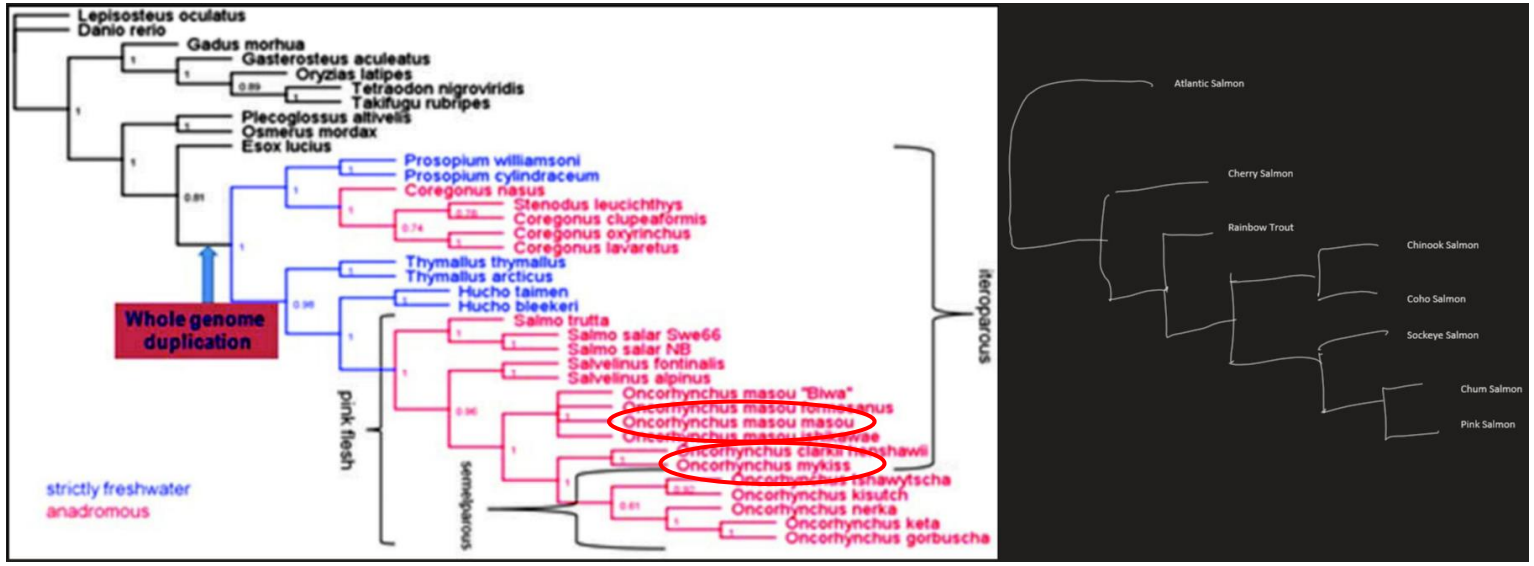
Overview

- Summary of Phylogeny Tree and BRAKER installation
- Literature Reviews
  - Yabin Guo, An overview on the DNA nucleotide compositions across kingdoms
  - James C. Schnable, Genes and gene models, an important distinction
- Macrosynteny Plot
- Results
- Future Work

Summary of Phylogeny Tree and BRAKER install

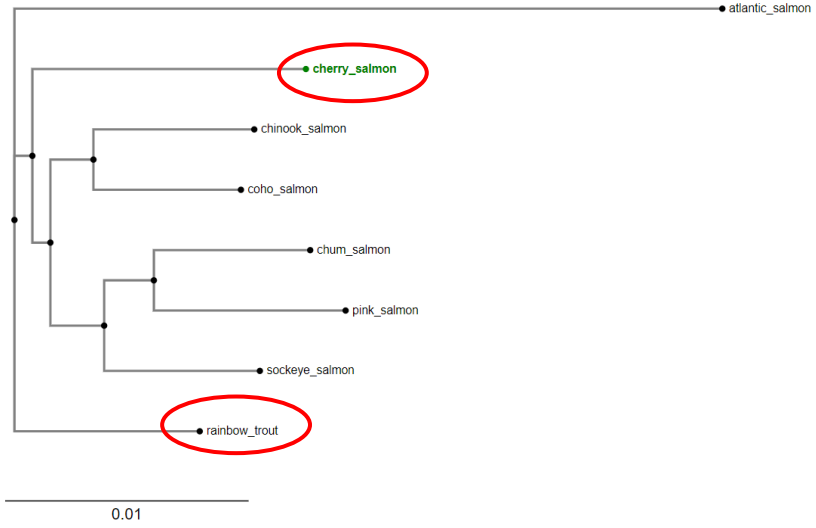
Things of Interest:

- Phylogeny Tree pipeline doesn't work as well for closely related species
- BRAKER3 installation:
  - When installing perl in conda, had to use version 5.26.2, to accommodate for perl scalar util numeric
  - StringTie2, Bedtools, GFFRead, Diamond, SRA Toolkit, samtools all in Genemark-ETP tools
  - When using compleasm in BUSCO mode, delete the Darwin folder in sepp/tools/bundled



Davidson, William S. *Genome*, vol. 56, no. 10, Oct. 2013, pp. 548–550, doi:10.1139/gen-2013-0163.

Orthofinder Tree

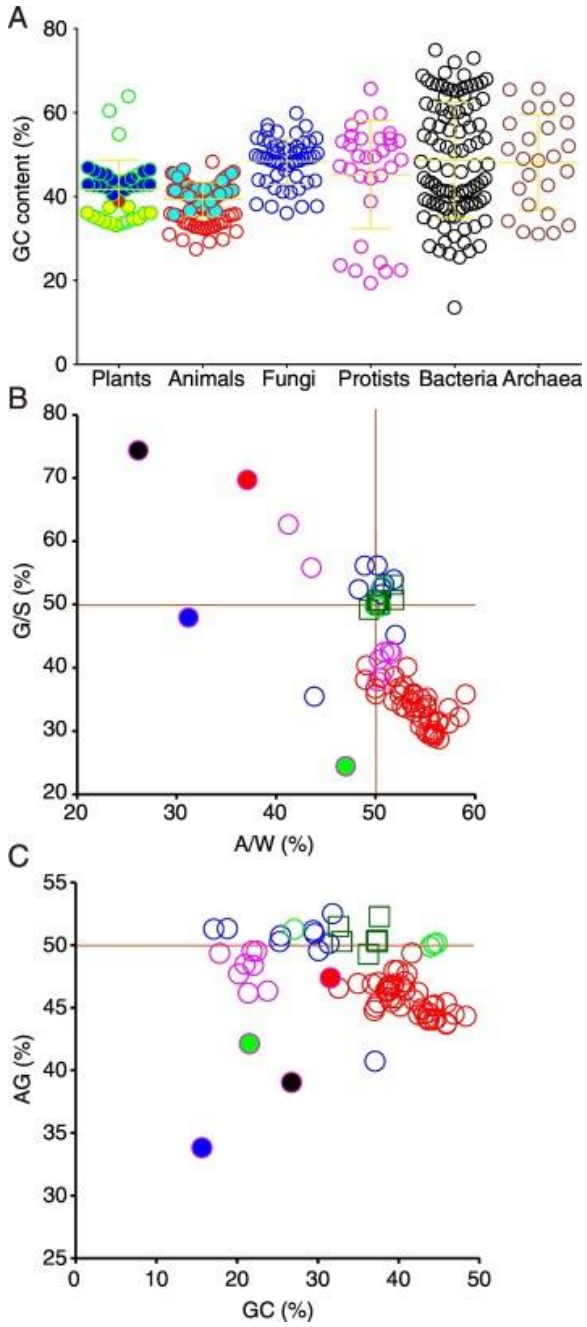


Full Pipeline

Literature Review

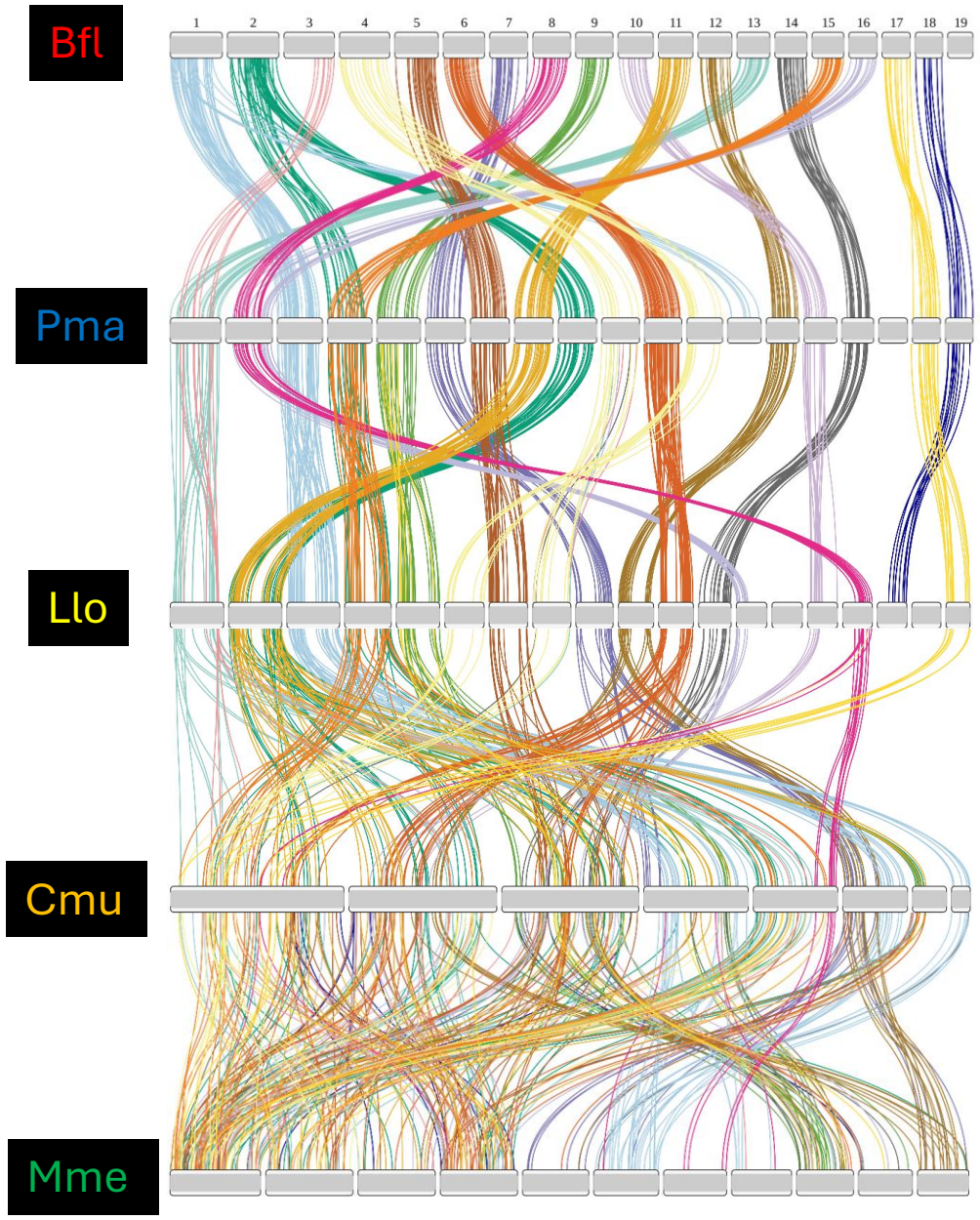
An overview on the DNA nucleotide compositions across kingdoms

- Yabin Guo, Gene Reports
- GC content: Proportion of GC in DNA strand
  - Chargaff's first rule, # of G=# of C, # of A = # of T
- GC skew and AT skew: (G-C)/(G+C) and (A-T)/(A+T), measures asymmetry of the nucleic acids
  - Chargaff's second rule, G/C ~ 1, A/T ~ 1
    - Substitution Rule
    - Symmetry is higher entropy
- Purine content: Proportion of AG in DNA strand
  - Szbalski's rule, DNA template strands have higher purine content <- politeness hypothesis
- Conclusion
  - Chargaff's second rule holds true
  - Some protists have asymmetric chromosomal strands
  - Asymmetric satellite DNA regions occurred in some animal genomes
  - Szybalski's rule is not universal, purine richness of coding regions due to energy requirements
  - Thermophilicity is correlated to both genome GC content and CDS AG% in archaea



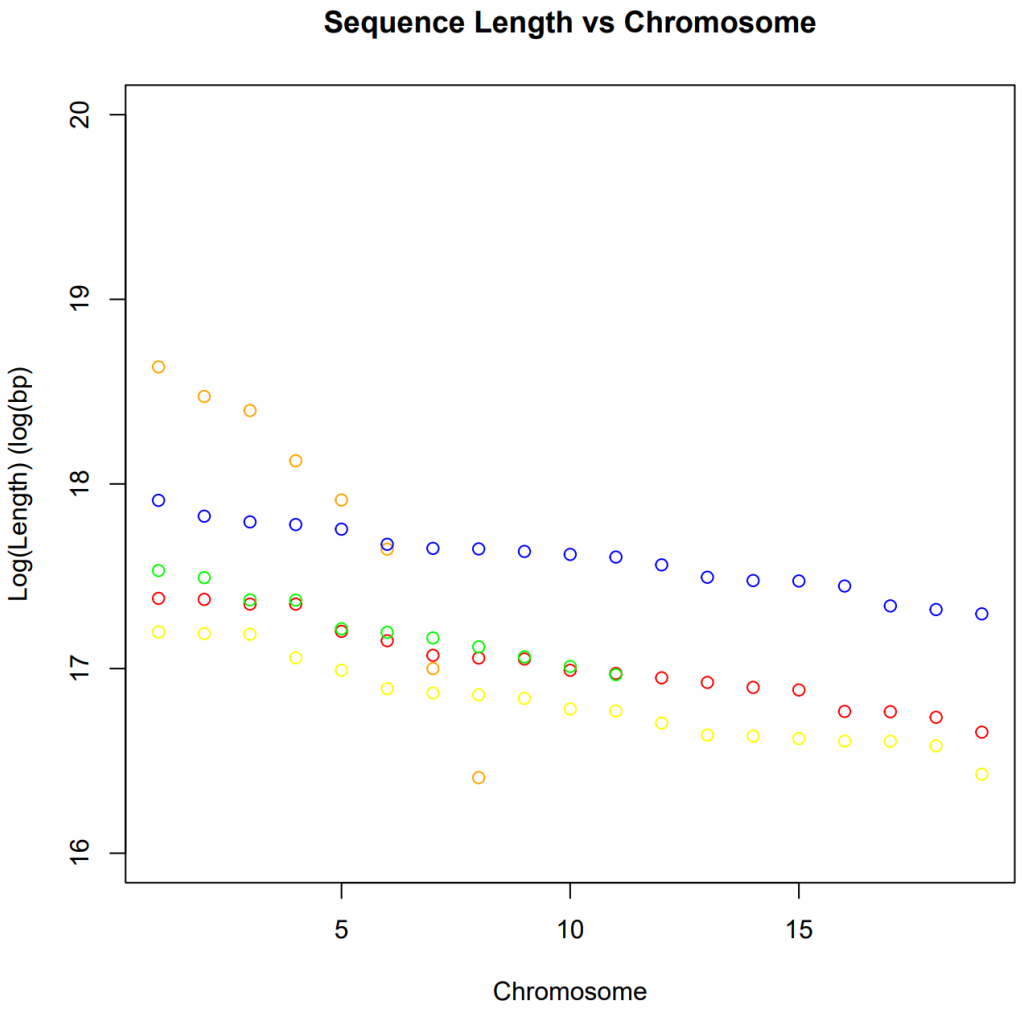
Category	Properties
GC	- Triple Hydrogen Bond
AT	- Double Hydrogen Bond
Purine (AG)	- Double Ring - Heavier
Pyramidine (CT)	- Single Ring - Cheaper to produce

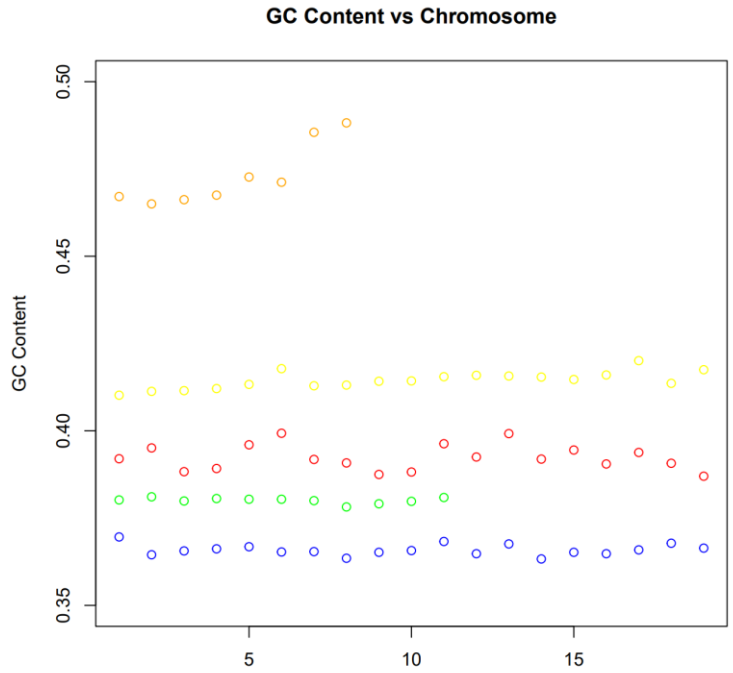
Macrosynteny Plot

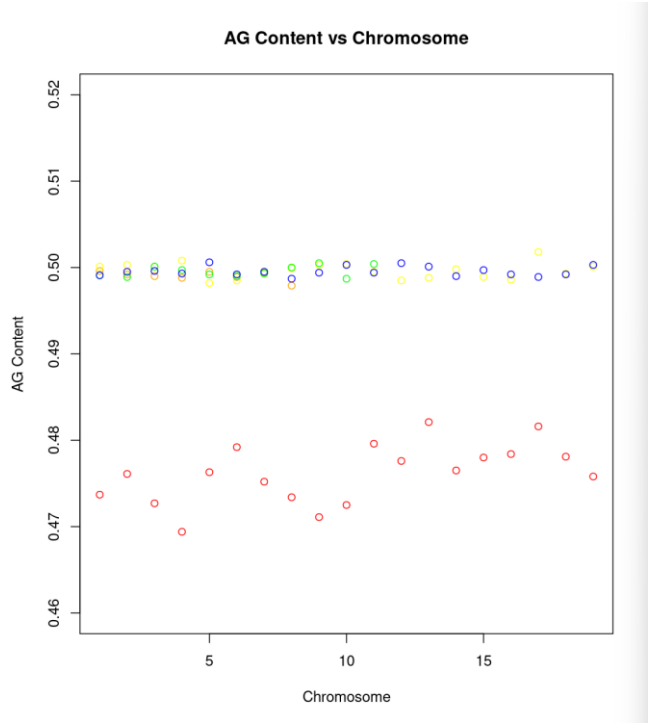


Name	Source
B. floridae, Bfl	Ferdinand Marlétaz's lab + **SyntenyFinder default**
P. maximus, Pma	NCBI
L. longissimus, Llo	NCBI
C. mucedo, Cmu	GEO + Dryad
M. membranacea, Mme	NCBI + Dryad

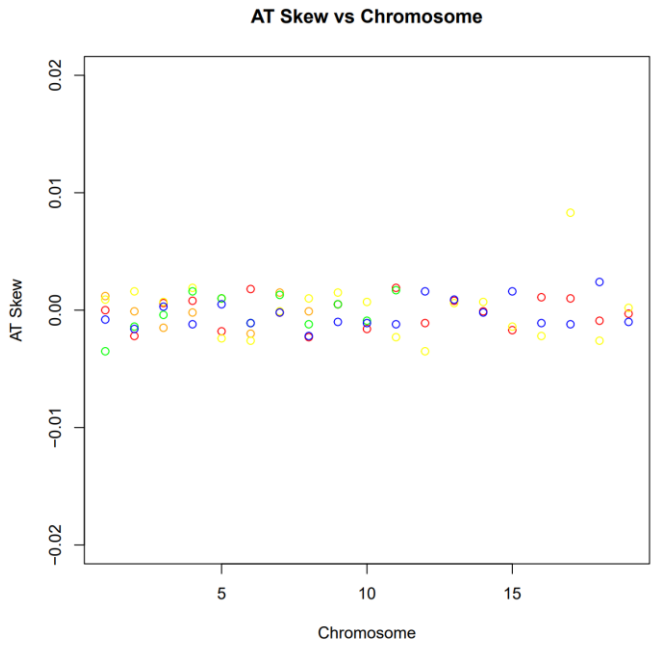
Bfl Pma Llo Cmu Mme

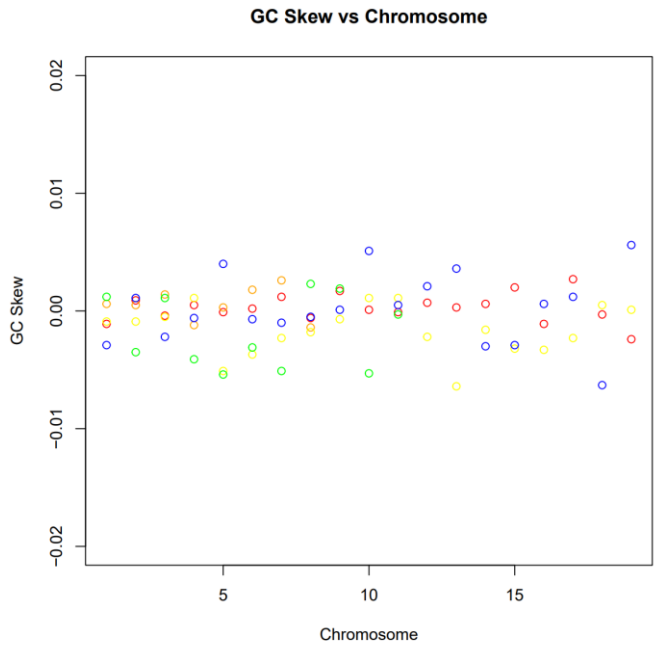


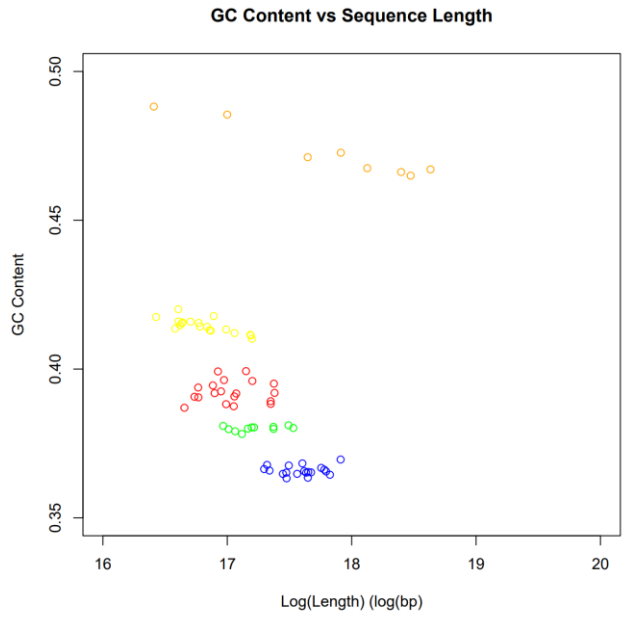




Bfl Pma Llo Cmu Mme

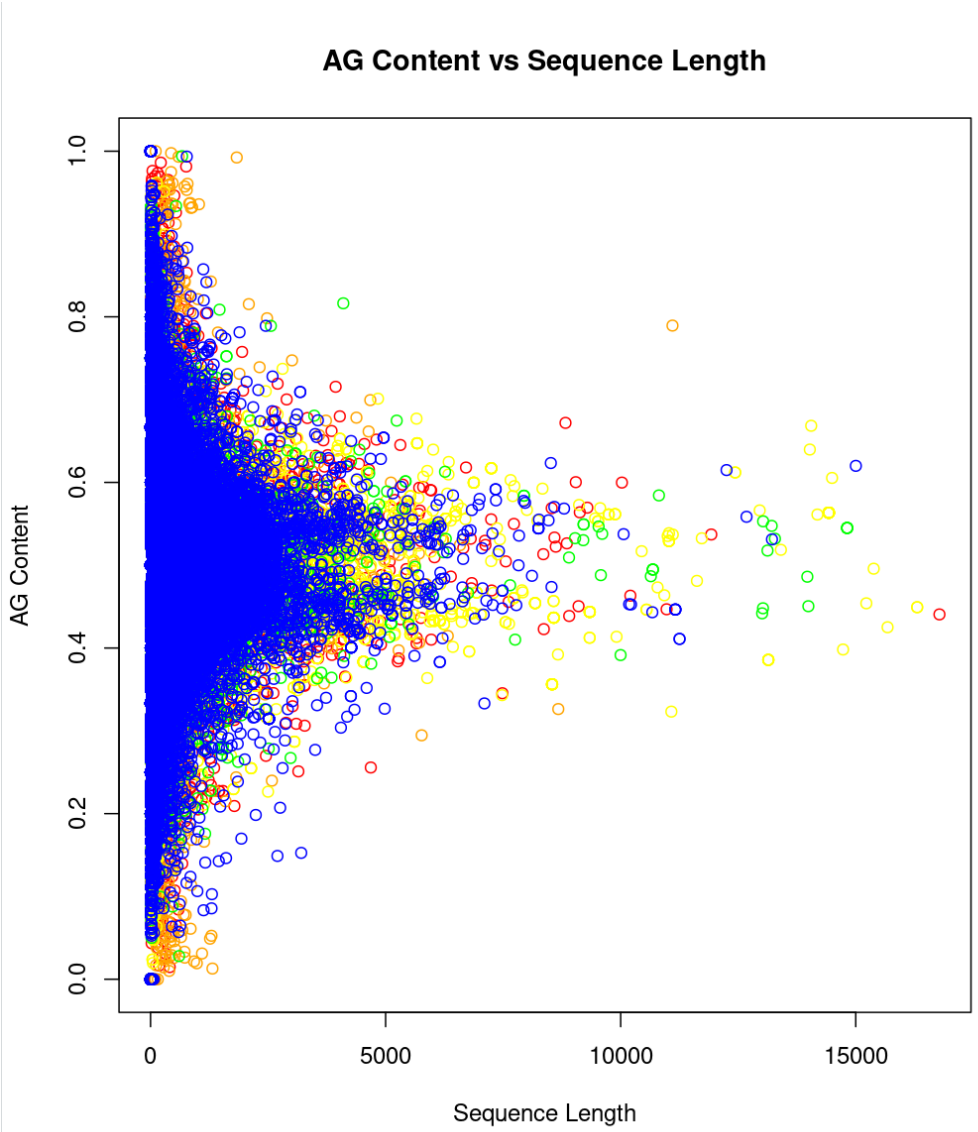


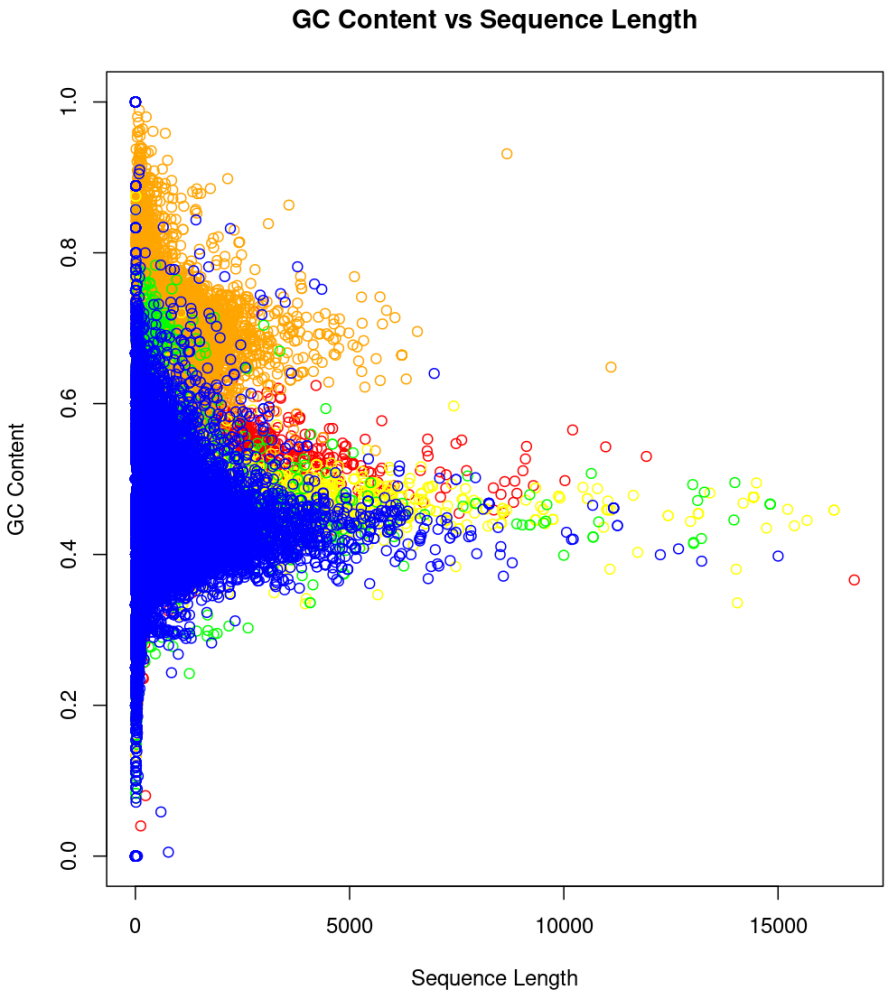




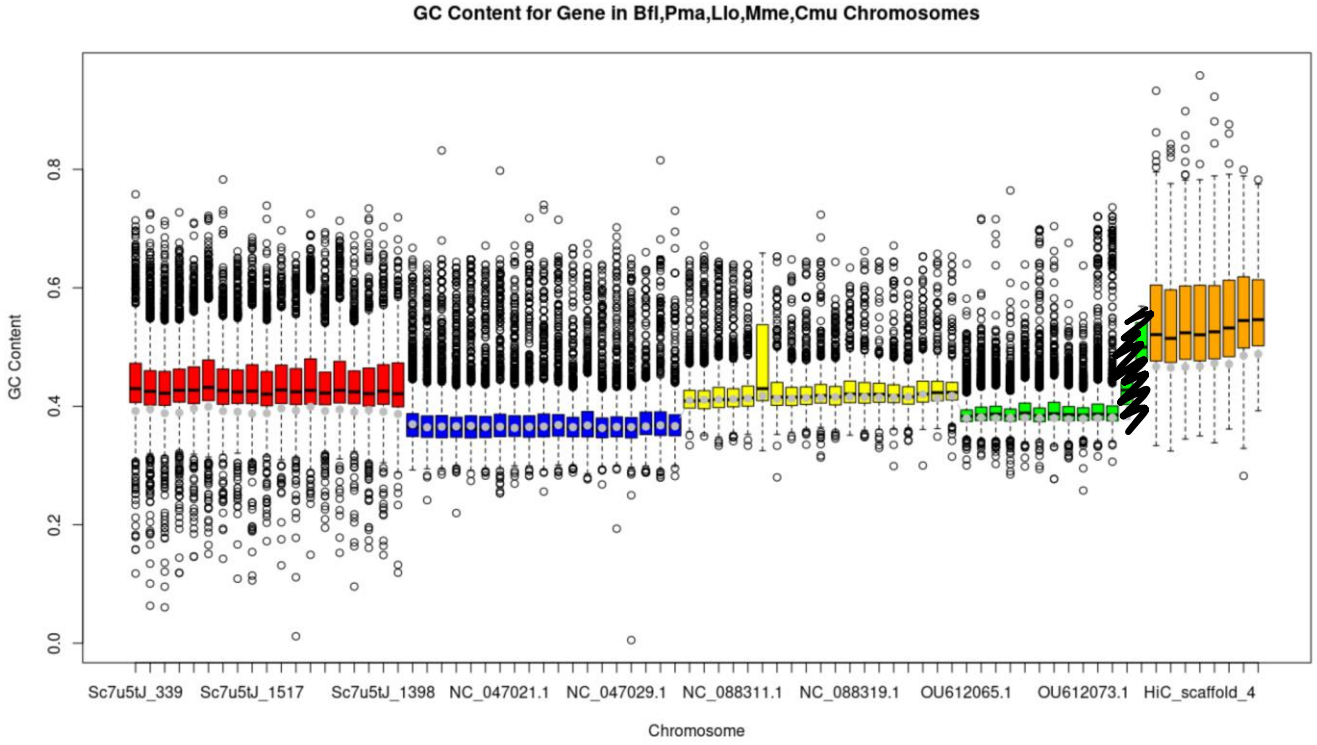
Bfl Pma Llo Cmu Mme

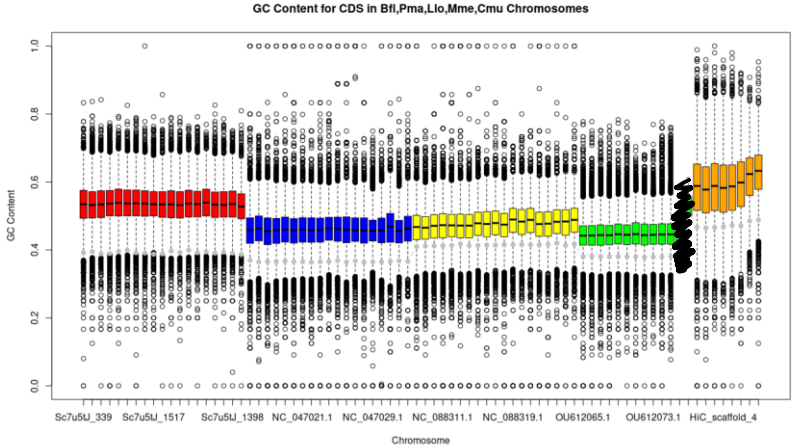
*Only CDS entries

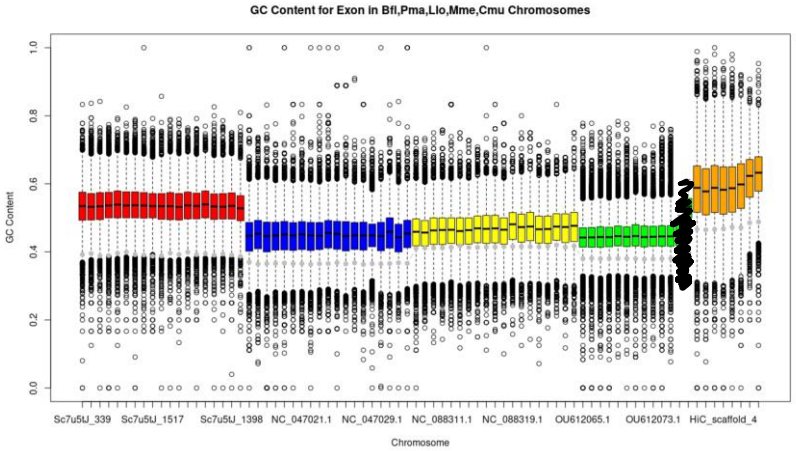




Overview of Work:


GC Content for Gene in Bfl,Pma,Llo,Mme,Cmu Chromosomes
GC Content
0.0
0.2
0.4
0.6
0.8
Sc7u5tJ_339
Sc7u5tJ_1517
Sc7u5tJ_1398
NC_047021.1
NC_047029.1
NC_088311.1
NC_088319.1
OU612065.1
OU612073.1
HiC_scaffold_4
Chromosome


GC Content for CDS in Bfl,Pma,Llo,Mme,Cmu Chromosomes
GC Content
0.0
0.2
0.4
0.6
0.8
1.0
Sc7u5tJ_339
Sc7u5tJ_1517
Sc7u5tJ_1398
NC_047021.1
NC_047029.1
NC_088311.1
NC_088319.1
OU612065.1
OU612073.1
HiC_scaffold_4
Chromosome

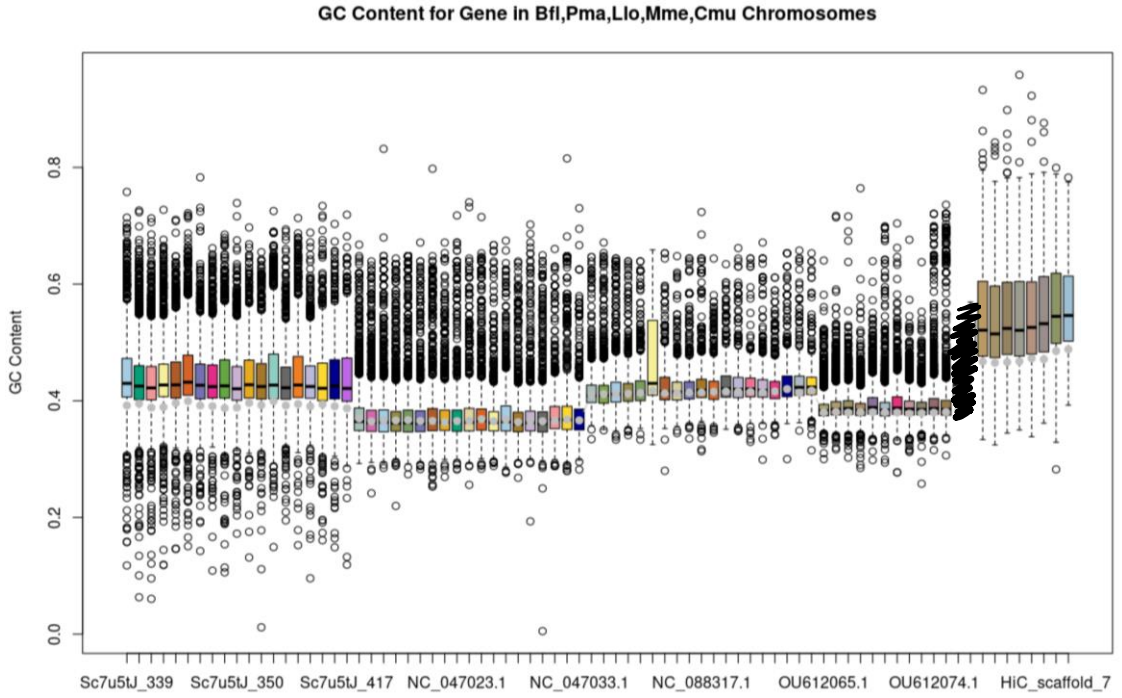

GC Content for Exon in Bfl,Pma,Llo,Mme,Cmu Chromosomes
GC Content
0.0
0.2
0.4
0.6
0.8
1.0
Sc7u5tJ_339
Sc7u5tJ_1517
Sc7u5tJ_1398
NC_047021.1
NC_047029.1
NC_088311.1
NC_088319.1
OU612065.1
OU612073.1
HiC_scaffold_4
Chromosome

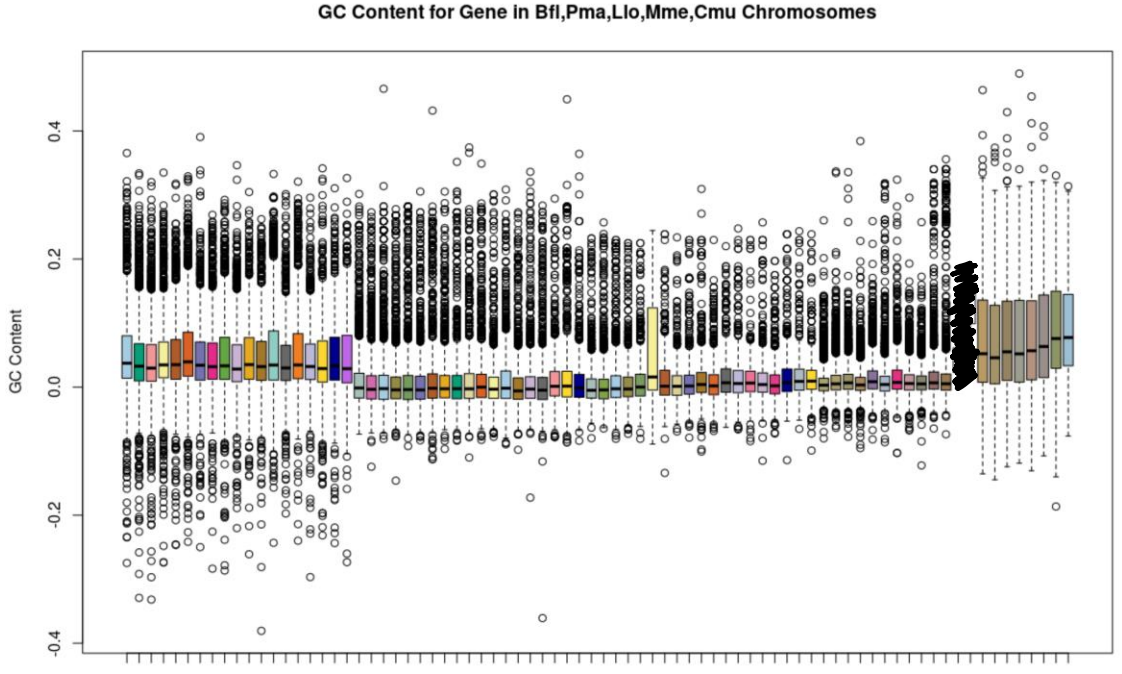
Alg Tracing

- Averaged the color of all the single copy orthologues in the chromosome

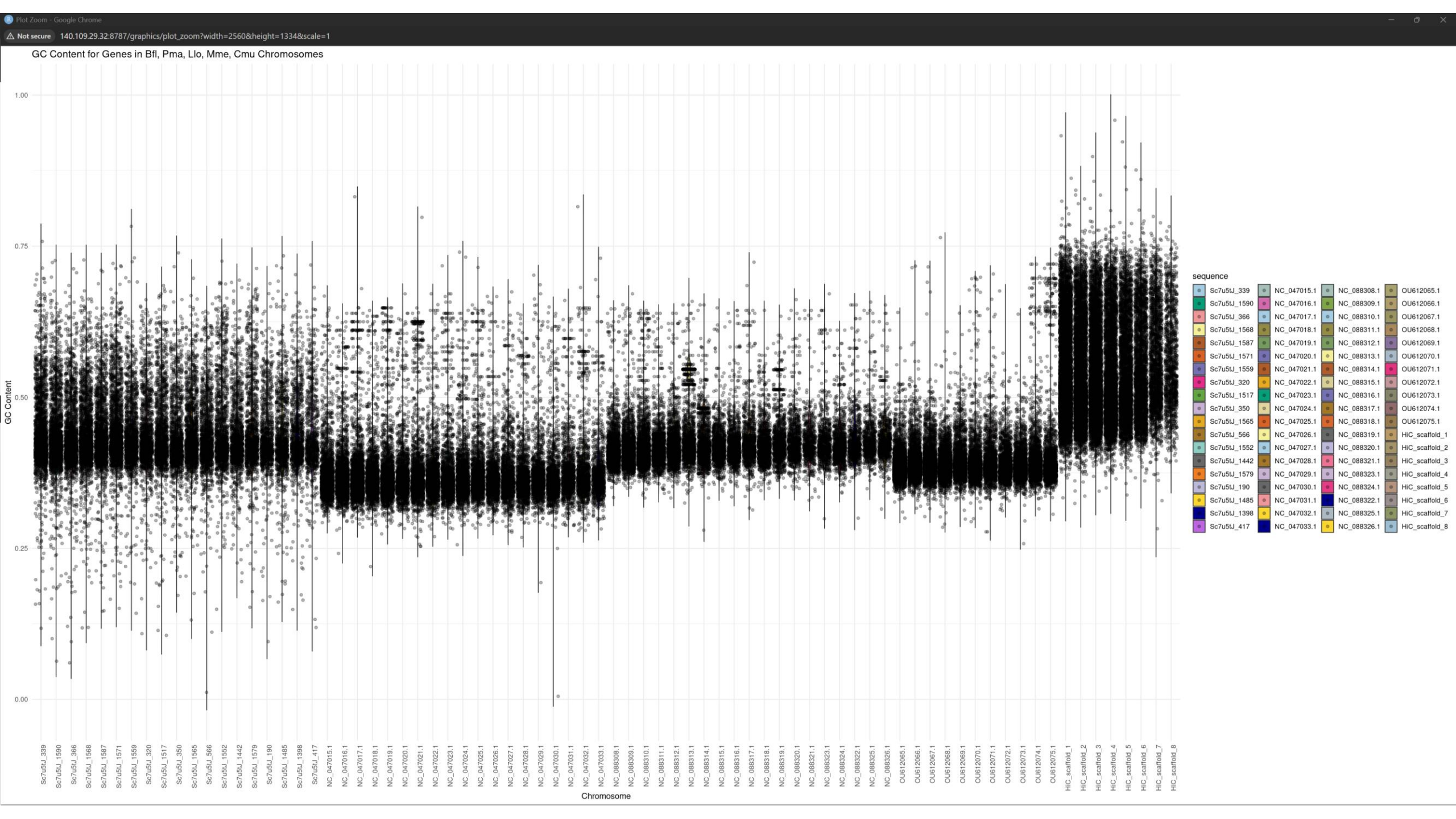
Top: Plotted chromosome average as grey line

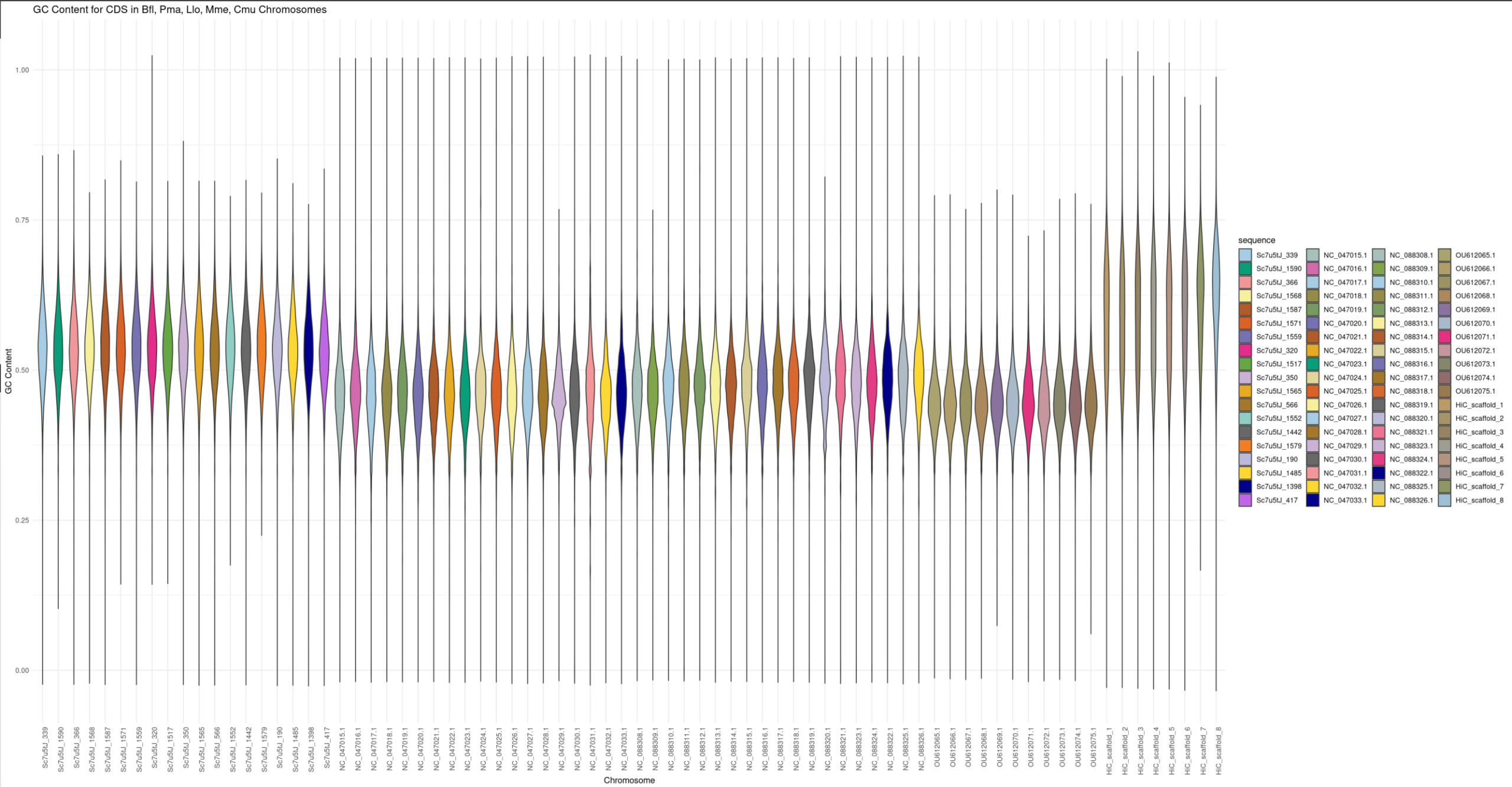
Bottom: Shifted each chromosome by avg GC content of species





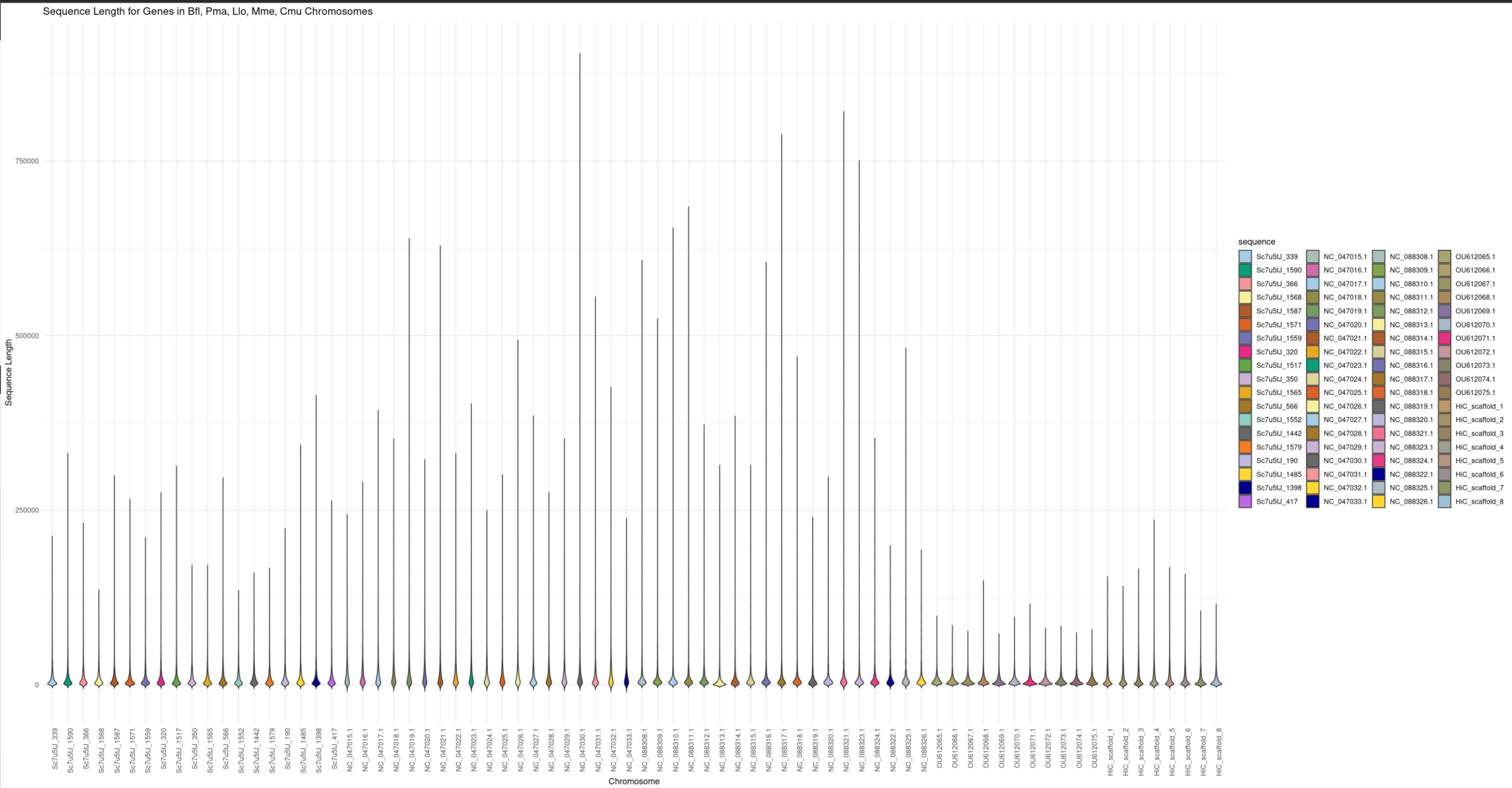

GC Content for Genes in Bfl, Pma, Llo, Mme, Cmu Chromosomes
Sliding Window Analysis for different metrics on the NC_047024.1 scaffold
_1568 scaffold
Sliding Window Analysis for different metrics on the NC_088313.1 scaffold
Sliding Window Analysis for different metrics on the OU612065.1 scaffold
GC Content
GC Skew
AG Content
AT Skew
Null Percent
Position in Sequence (bp)
SC Orthologues
Gene Density
Chromosome


Plot Zoom - Google Chrome
Not secure
140.109.29.32:8787/graphics/plot_zoom?width=2560&height=1334&scale=1
GC Content for Genes in Bfl, Pma, Llo, Mme, Cmu Chromosomes
GC Content
1.00
0.75
0.50
0.25
0.00
Chromosome
sequence
Sc7u5tJ_339
Sc7u5tJ_1590
Sc7u5tJ_366
Sc7u5tJ_1568
Sc7u5tJ_1587
Sc7u5tJ_1571
Sc7u5tJ_1559
Sc7u5tJ_320
Sc7u5tJ_1517
Sc7u5tJ_350
Sc7u5tJ_1565
Sc7u5tJ_566
Sc7u5tJ_1552
Sc7u5tJ_1442
Sc7u5tJ_1579
Sc7u5tJ_190
Sc7u5tJ_1485
Sc7u5tJ_1398
Sc7u5tJ_417
NC_047015.1
NC_047016.1
NC_047017.1
NC_047018.1
NC_047019.1
NC_047020.1
NC_047021.1
NC_047022.1
NC_047023.1
NC_047024.1
NC_047025.1
NC_047026.1
NC_047027.1
NC_047028.1
NC_047029.1
NC_047030.1
NC_047031.1
NC_047032.1
NC_047033.1
NC_088308.1
NC_088309.1
NC_088310.1
NC_088311.1
NC_088312.1
NC_088313.1
NC_088314.1
NC_088315.1
NC_088316.1
NC_088317.1
NC_088318.1
NC_088319.1
NC_088320.1
NC_088321.1
NC_088323.1
NC_088324.1
NC_088322.1
NC_088325.1
NC_088326.1
OU612065.1
OU612066.1
OU612067.1
OU612068.1
OU612069.1
OU612070.1
OU612071.1
OU612072.1
OU612073.1
OU612074.1
OU612075.1
HiC_scaffold_1
HiC_scaffold_2
HiC_scaffold_3
HiC_scaffold_4
HiC_scaffold_5
HiC_scaffold_6
HiC_scaffold_7
HiC_scaffold_8



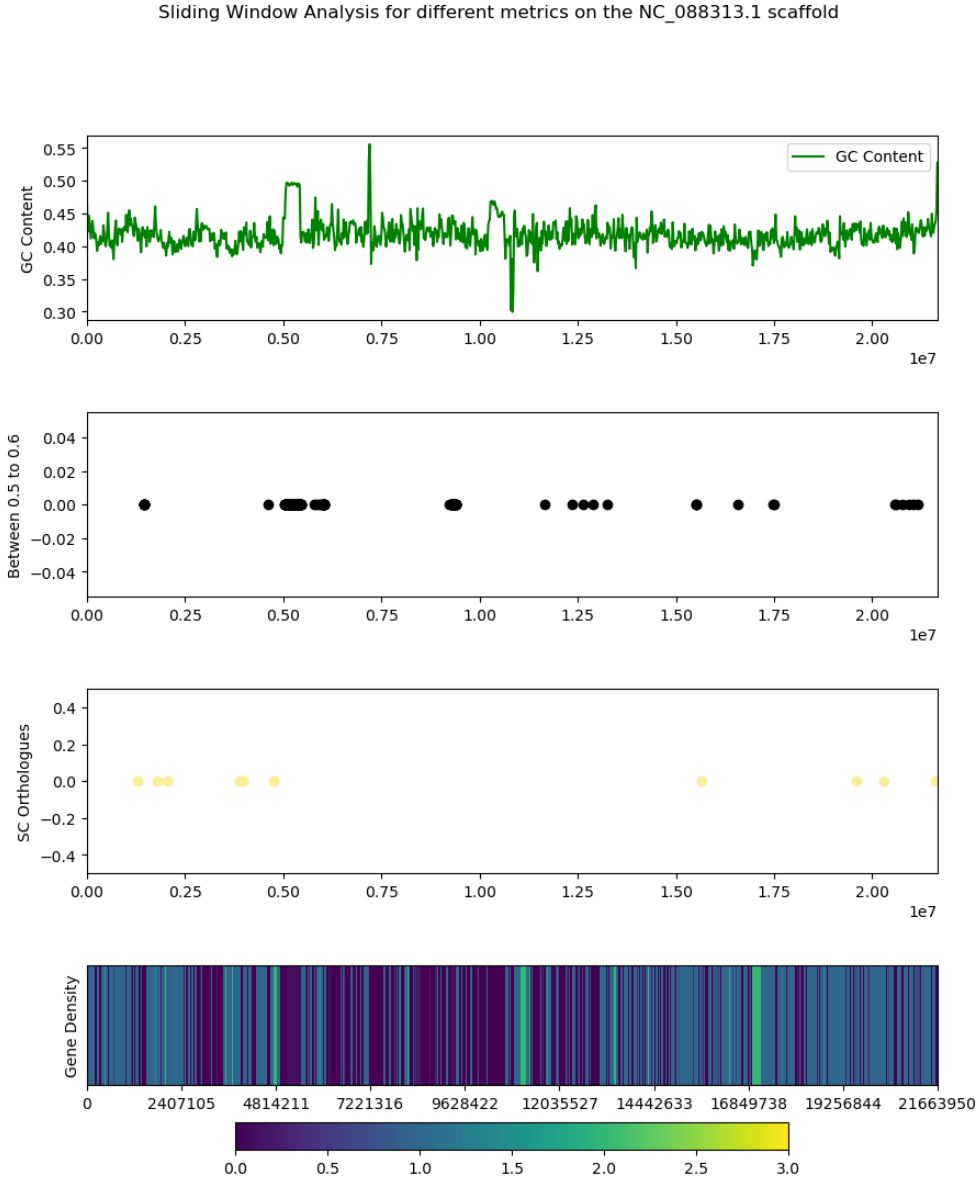
Plot Zoom - Google Chrome

Not secure 140.109.29.32:8787/graphics/plot_zoom?width=2560&height=1334&scale=1



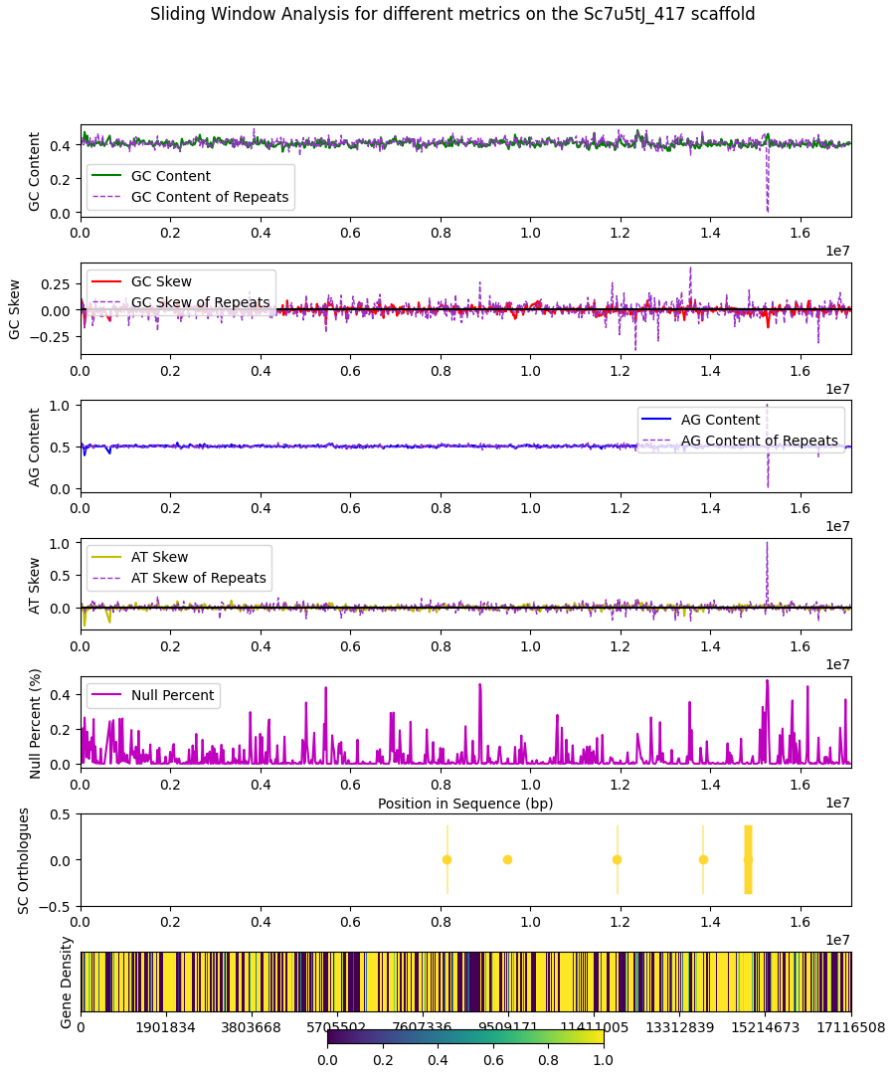
Llo Biomodal Distribution

- High GC genes cluster
- Could be due to large centromeres

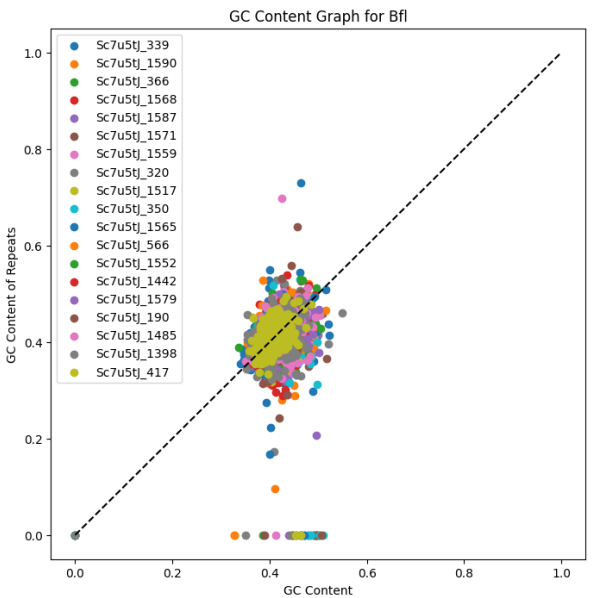


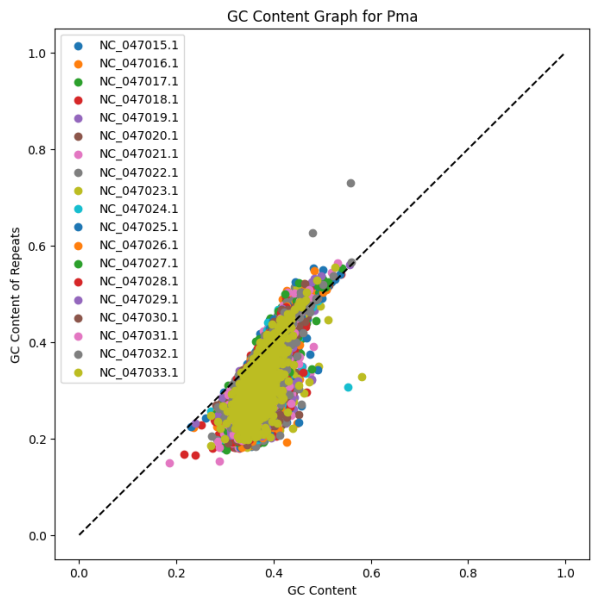
Sliding Window Comparison of Repeats

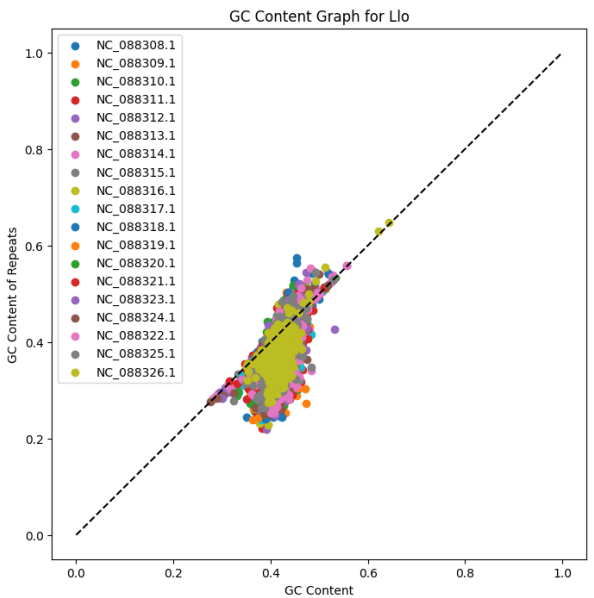
- Repeat content is much more extreme

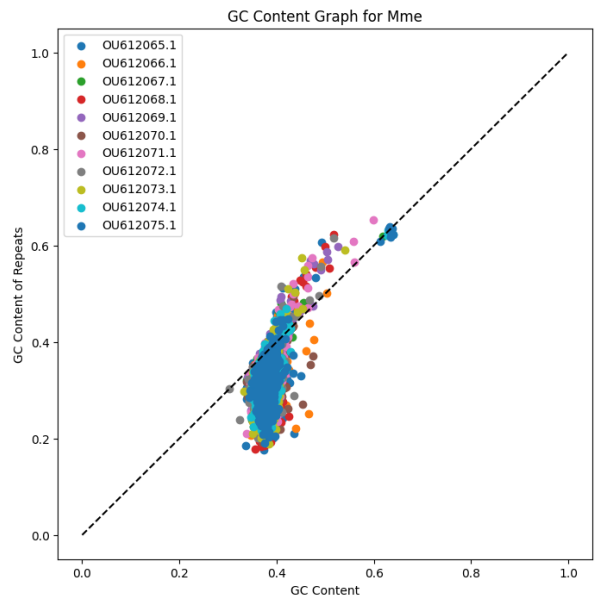


Effects of Repeats





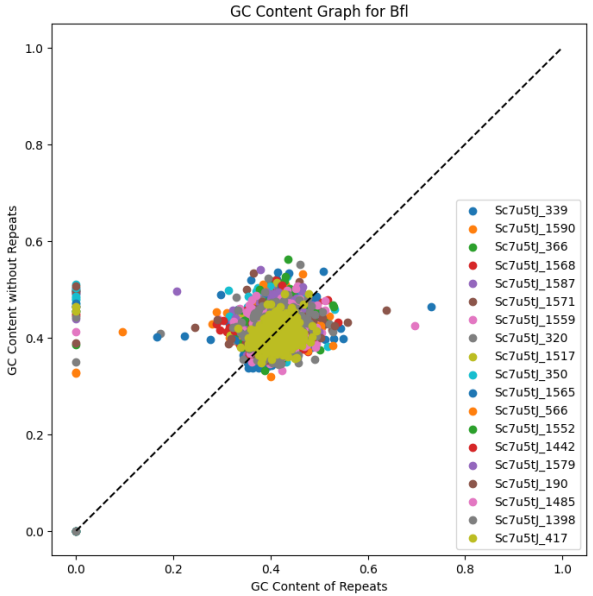


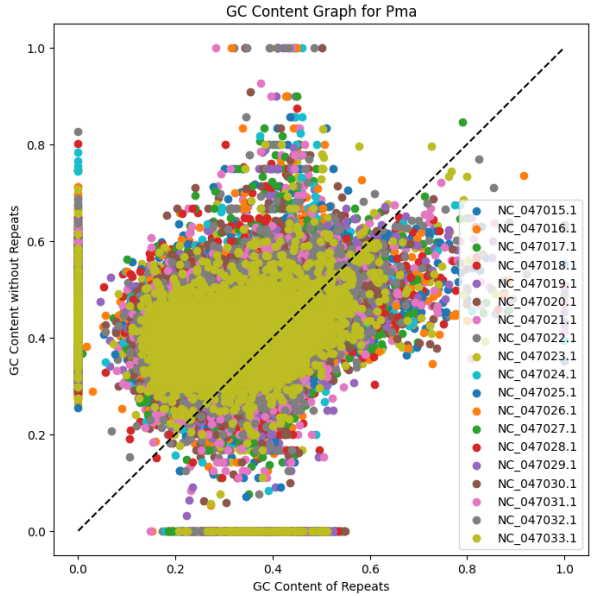


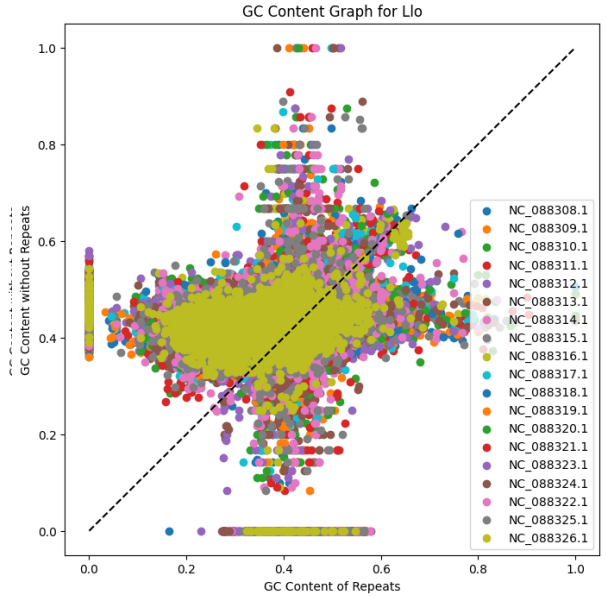
Weak positive correlation between GC content of window and GC content of repeats in window

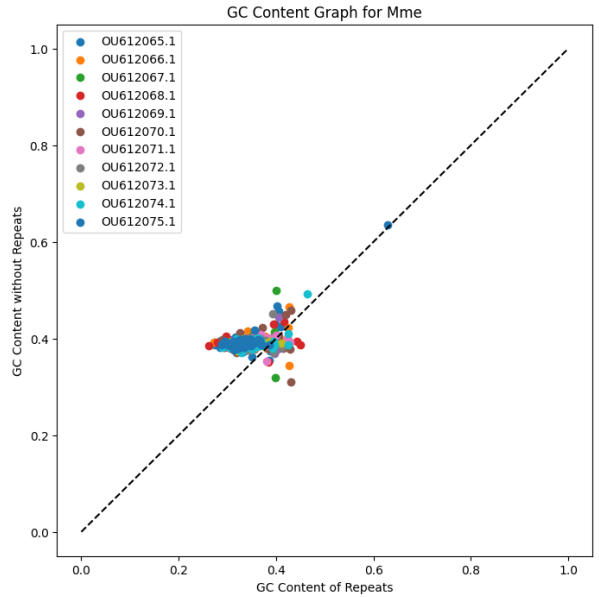
NR v. R

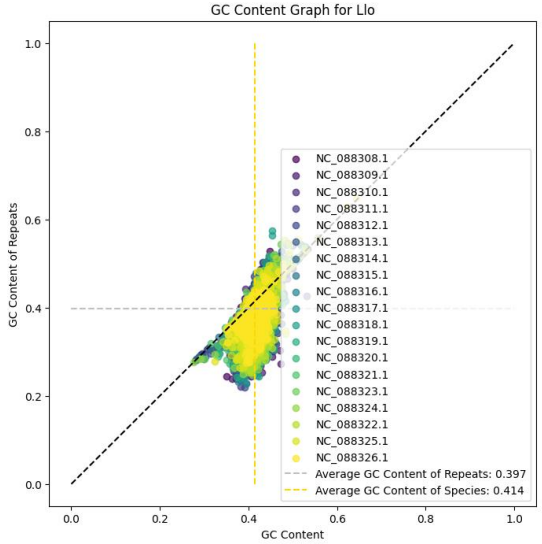
- No correlation
- When window size is too small, crosses appear -> sections of high variance due to small length, while large length has low variance

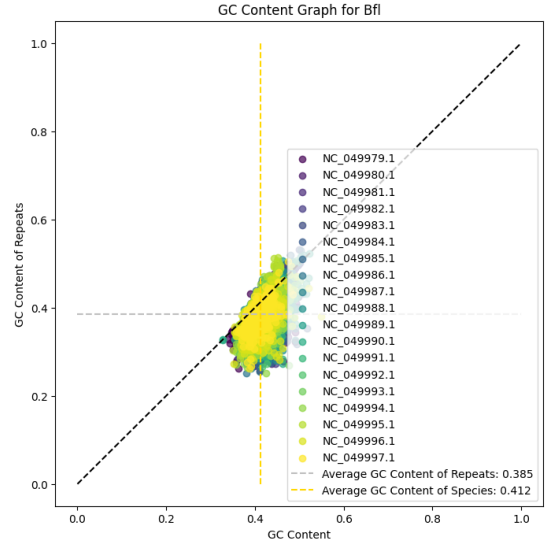


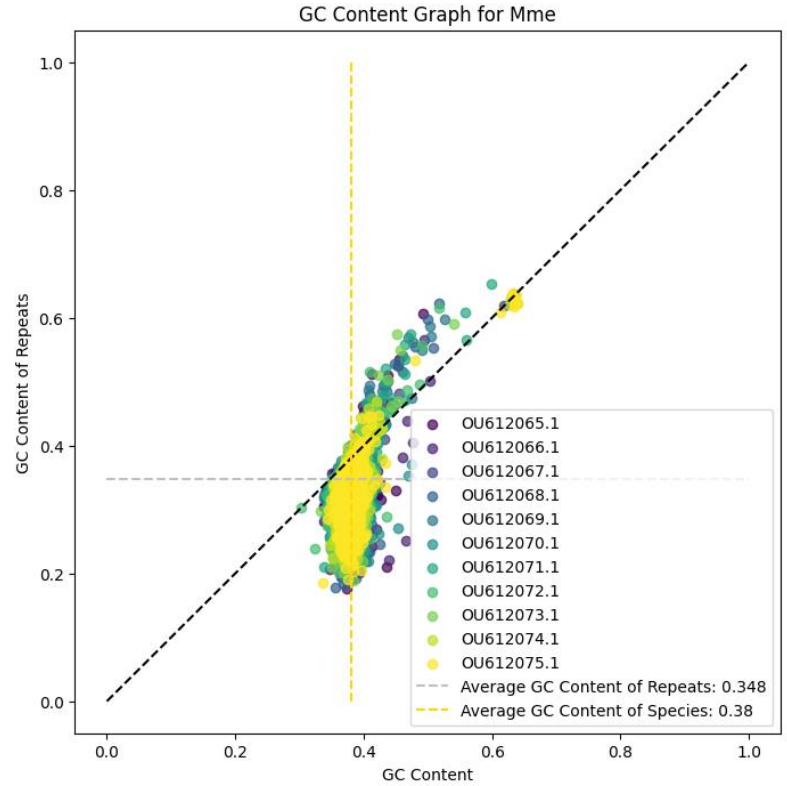


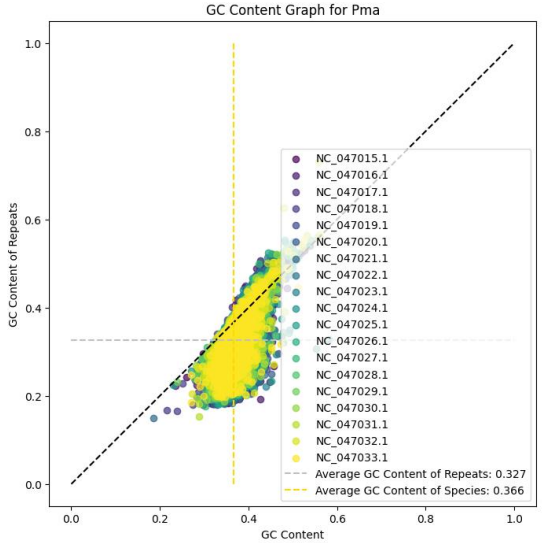


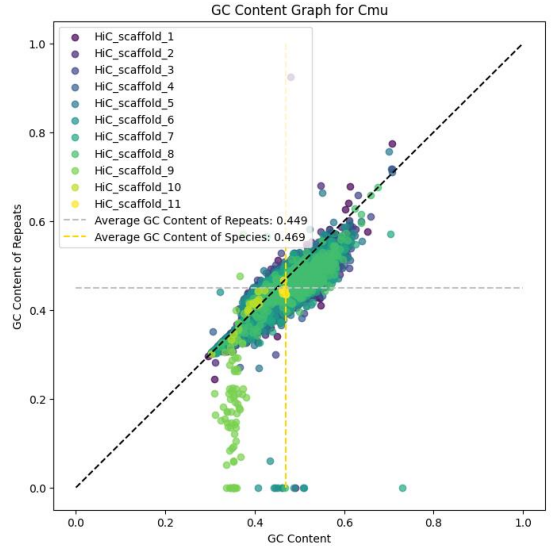



GC Content Graph for Llo
GC Content of Repeats
GC Content
NC_088308.1
NC_088309.1
NC_088310.1
NC_088311.1
NC_088312.1
NC_088313.1
NC_088314.1
NC_088315.1
NC_088316.1
NC_088317.1
NC_088318.1
NC_088319.1
NC_088320.1
NC_088321.1
NC_088323.1
NC_088324.1
NC_088322.1
NC_088325.1
NC_088326.1
Average GC Content of Repeats: 0.397
Average GC Content of Species: 0.414


GC Content Graph for Bfl
GC Content of Repeats
GC Content
NC_049979.1
NC_049980.1
NC_049981.1
NC_049982.1
NC_049983.1
NC_049984.1
NC_049985.1
NC_049986.1
NC_049987.1
NC_049988.1
NC_049989.1
NC_049990.1
NC_049991.1
NC_049992.1
NC_049993.1
NC_049994.1
NC_049995.1
NC_049996.1
NC_049997.1
Average GC Content of Repeats: 0.385
Average GC Content of Species: 0.412

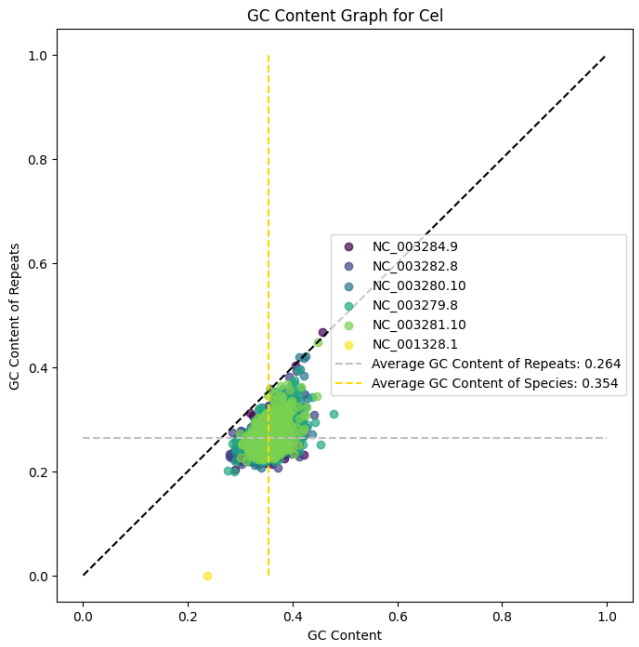

GC Content Graph for Mme
GC Content of Repeats
GC Content
OU612065.1
OU612066.1
OU612067.1
OU612068.1
OU612069.1
OU612070.1
OU612071.1
OU612072.1
OU612073.1
OU612074.1
OU612075.1
Average GC Content of Repeats: 0.348
Average GC Content of Species: 0.38

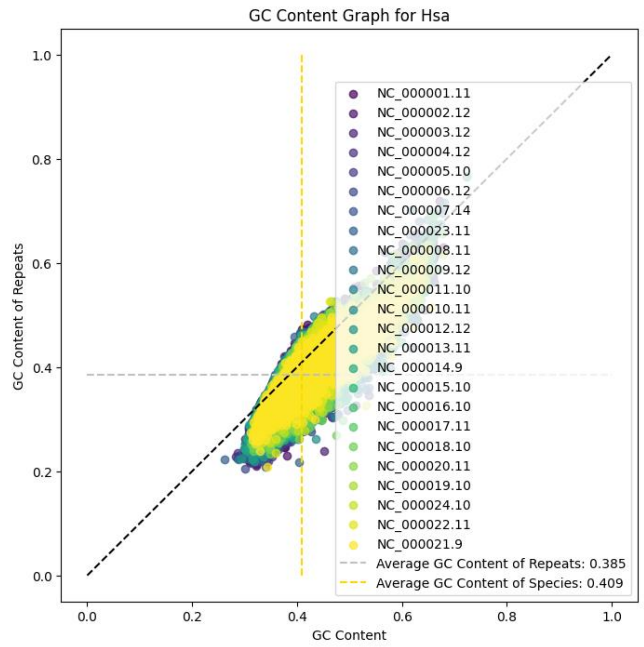

GC Content Graph for Pma
GC Content of Repeats
GC Content
NC_047015.1
NC_047016.1
NC_047017.1
NC_047018.1
NC_047019.1
NC_047020.1
NC_047021.1
NC_047022.1
NC_047023.1
NC_047024.1
NC_047025.1
NC_047026.1
NC_047027.1
NC_047028.1
NC_047029.1
NC_047030.1
NC_047031.1
NC_047032.1
NC_047033.1
Average GC Content of Repeats: 0.327
Average GC Content of Species: 0.366

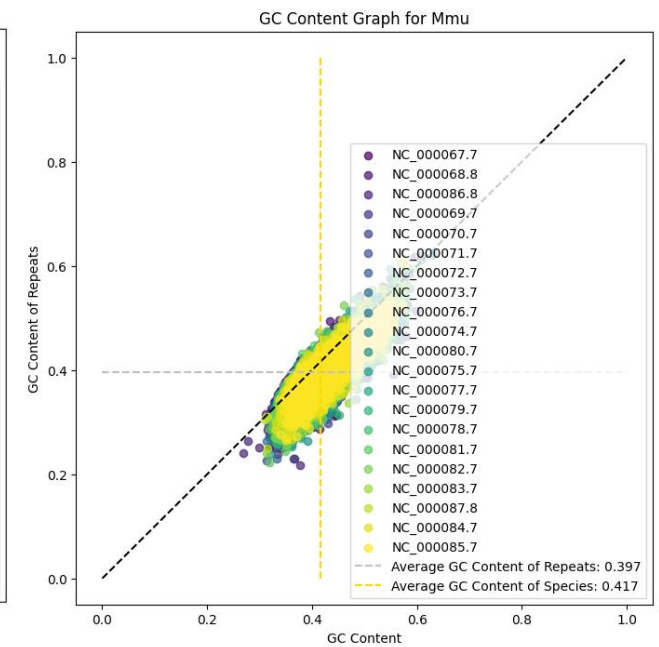

GC Content Graph for Cmu
GC Content of Repeats
GC Content
HiC_scaffold_1
HiC_scaffold_2
HiC_scaffold_3
HiC_scaffold_4
HiC_scaffold_5
HiC_scaffold_6
HiC_scaffold_7
HiC_scaffold_8
HiC_scaffold_9
HiC_scaffold_10
HiC_scaffold_11
Average GC Content of Repeats: 0.449
Average GC Content of Species: 0.469

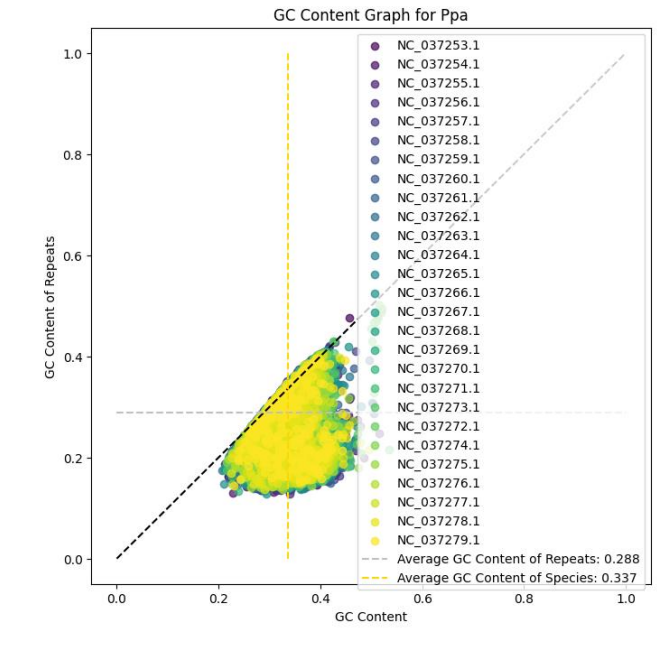
Other Species

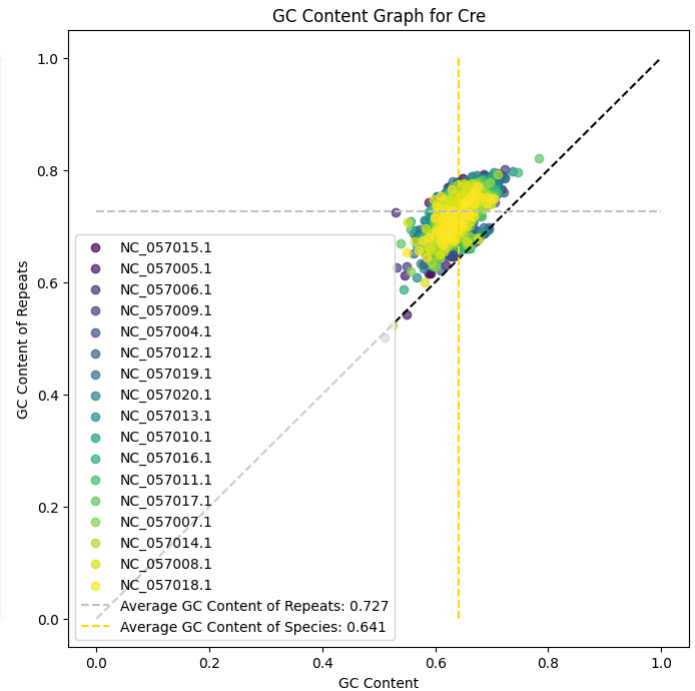
- C. Elegans: GCF_000002985.6
- H. Sapiens: GCF_000001405.40
- M. Musculus: GCF_000001635.27
- P. Patens: GCF_000002425.4
- C. reinhardtii: GCF_000002595.2
- A. Thaliana: GCF_000001735.4

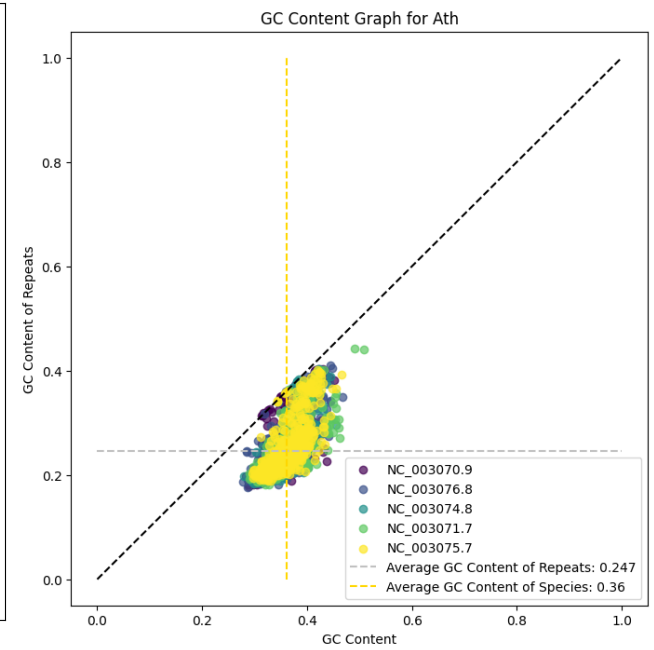

GC Content Graph for Cel
GC Content of Repeats
GC Content
NC_003284.9
NC_003282.8
NC_003280.10
NC_003279.8
NC_003281.10
NC_001328.1
Average GC Content of Repeats: 0.264
Average GC Content of Species: 0.354


GC Content Graph for Hsa
GC Content of Repeats
GC Content
NC_000001.11
NC_000002.12
NC_000003.12
NC_000004.12
NC_000005.10
NC_000006.12
NC_000007.14
NC_000023.11
NC_000008.11
NC_000009.12
NC_000011.10
NC_000010.11
NC_000012.12
NC_000013.11
NC_000014.9
NC_000015.10
NC_000016.10
NC_000017.11
NC_000018.10
NC_000020.11
NC_000019.10
NC_000024.10
NC_000022.11
NC_000021.9
Average GC Content of Repeats: 0.385
Average GC Content of Species: 0.409


GC Content Graph for Mmu
GC Content of Repeats
GC Content
NC_000067.7
NC_000068.8
NC_000086.8
NC_000069.7
NC_000070.7
NC_000071.7
NC_000072.7
NC_000073.7
NC_000076.7
NC_000074.7
NC_000080.7
NC_000075.7
NC_000077.7
NC_000079.7
NC_000078.7
NC_000081.7
NC_000082.7
NC_000083.7
NC_000087.8
NC_000084.7
NC_000085.7
Average GC Content of Repeats: 0.397
Average GC Content of Species: 0.417


GC Content Graph for Ppa
GC Content of Repeats
GC Content
NC_037253.1
NC_037254.1
NC_037255.1
NC_037256.1
NC_037257.1
NC_037258.1
NC_037259.1
NC_037260.1
NC_037261.1
NC_037262.1
NC_037263.1
NC_037264.1
NC_037265.1
NC_037266.1
NC_037267.1
NC_037268.1
NC_037269.1
NC_037270.1
NC_037271.1
NC_037273.1
NC_037272.1
NC_037274.1
NC_037275.1
NC_037276.1
NC_037277.1
NC_037278.1
NC_037279.1
Average GC Content of Repeats: 0.288
Average GC Content of Species: 0.337

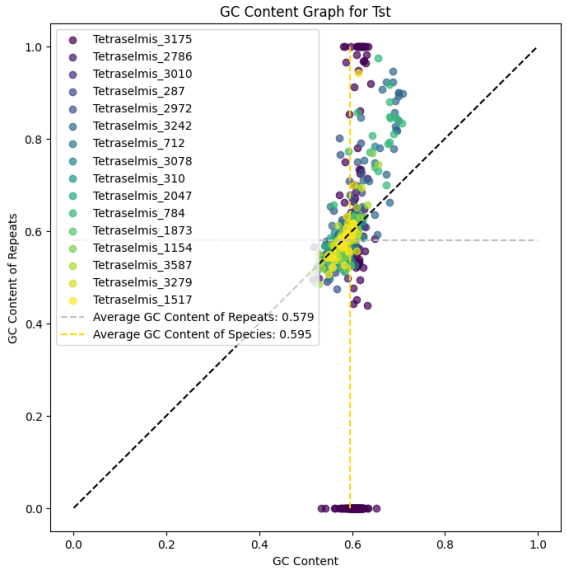

GC Content Graph for Cre
GC Content of Repeats
GC Content
NC_057015.1
NC_057005.1
NC_057006.1
NC_057009.1
NC_057004.1
NC_057012.1
NC_057019.1
NC_057020.1
NC_057013.1
NC_057010.1
NC_057016.1
NC_057011.1
NC_057017.1
NC_057007.1
NC_057014.1
NC_057008.1
NC_057018.1
Average GC Content of Repeats: 0.727
Average GC Content of Species: 0.641

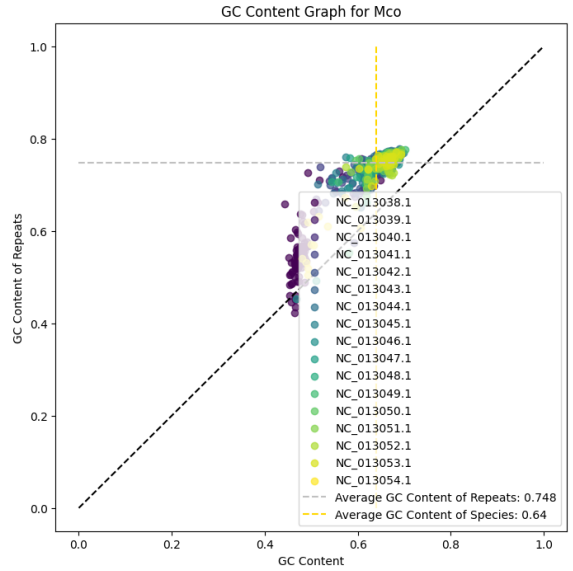

GC Content Graph for Ath
GC Content of Repeats
GC Content
NC_003070.9
NC_003076.8
NC_003074.8
NC_003071.7
NC_003075.7
Average GC Content of Repeats: 0.247
Average GC Content of Species: 0.36

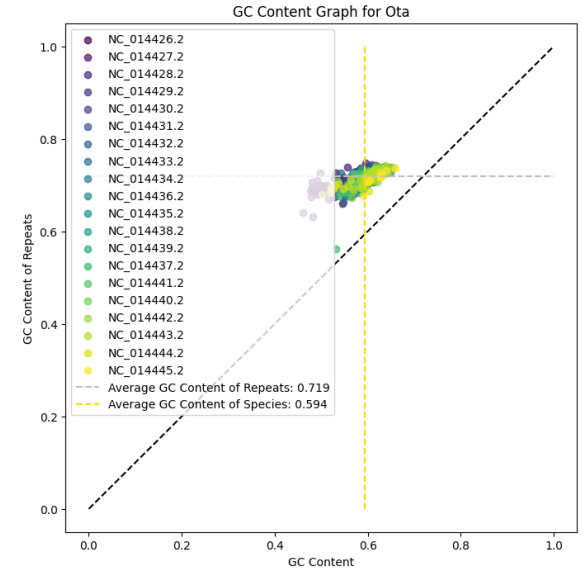
Added in other species

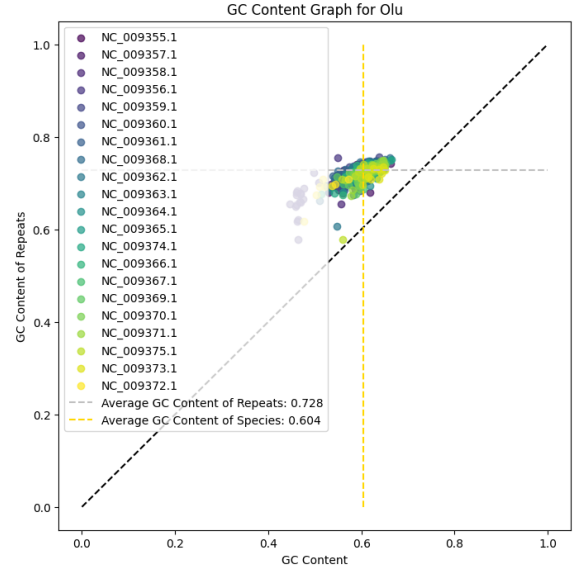
Chlorophyta

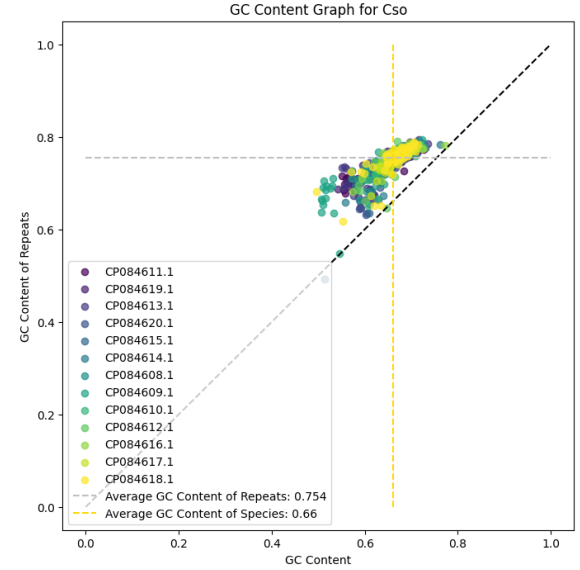
- Micromonas commoda: GCF_000090985.2
- Ostreococcus lucimarinus: GCF_000092065.1
- Bathycoccus prasinos: GCF_002220235.1
- Ostreococcus Tauri: GCF_000214015.3
- Chlorella sorokiniana: GCA_025917655.1
- Tetraselmis Striata: From lab

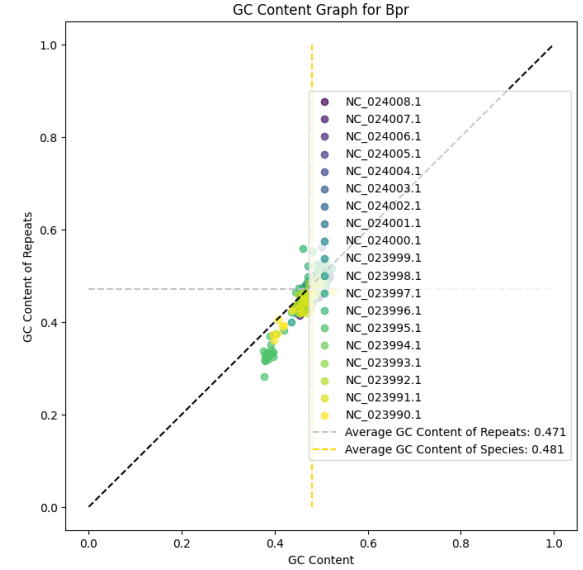

GC Content Graph for Tst
Tetraselmis_3175
Tetraselmis_2786
Tetraselmis_3010
Tetraselmis_287
Tetraselmis_2972
Tetraselmis_3242
Tetraselmis_712
Tetraselmis_3078
Tetraselmis_310
Tetraselmis_2047
Tetraselmis_784
Tetraselmis_1873
Tetraselmis_1154
Tetraselmis_3587
Tetraselmis_3279
Tetraselmis_1517
Average GC Content of Repeats: 0.579
Average GC Content of Species: 0.595
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0


GC Content Graph for Mco
NC_013038.1
NC_013039.1
NC_013040.1
NC_013041.1
NC_013042.1
NC_013043.1
NC_013044.1
NC_013045.1
NC_013046.1
NC_013047.1
NC_013048.1
NC_013049.1
NC_013050.1
NC_013051.1
NC_013052.1
NC_013053.1
NC_013054.1
Average GC Content of Repeats: 0.748
Average GC Content of Species: 0.64
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0


GC Content Graph for Ota
NC_014426.2
NC_014427.2
NC_014428.2
NC_014429.2
NC_014430.2
NC_014431.2
NC_014432.2
NC_014433.2
NC_014434.2
NC_014436.2
NC_014435.2
NC_014438.2
NC_014439.2
NC_014437.2
NC_014441.2
NC_014440.2
NC_014442.2
NC_014443.2
NC_014444.2
NC_014445.2
Average GC Content of Repeats: 0.719
Average GC Content of Species: 0.594
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0

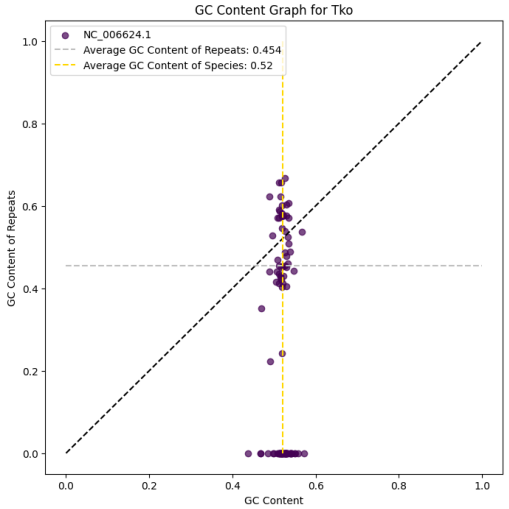

GC Content Graph for Olu
NC_009355.1
NC_009357.1
NC_009358.1
NC_009356.1
NC_009359.1
NC_009360.1
NC_009361.1
NC_009368.1
NC_009362.1
NC_009363.1
NC_009364.1
NC_009365.1
NC_009374.1
NC_009366.1
NC_009367.1
NC_009369.1
NC_009370.1
NC_009371.1
NC_009375.1
NC_009373.1
NC_009372.1
Average GC Content of Repeats: 0.728
Average GC Content of Species: 0.604
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0

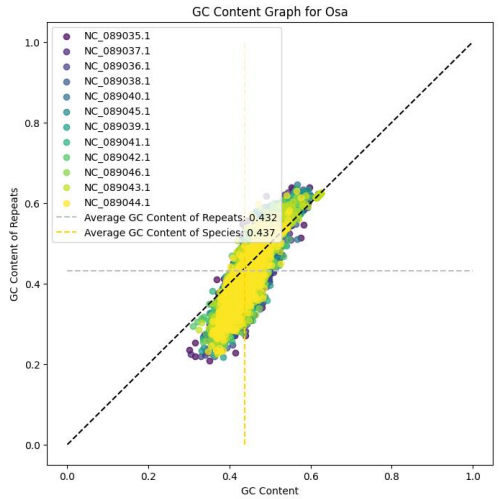

GC Content Graph for Cso
CP084611.1
CP084619.1
CP084613.1
CP084620.1
CP084615.1
CP084614.1
CP084608.1
CP084609.1
CP084610.1
CP084612.1
CP084616.1
CP084617.1
CP084618.1
Average GC Content of Repeats: 0.754
Average GC Content of Species: 0.66
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0

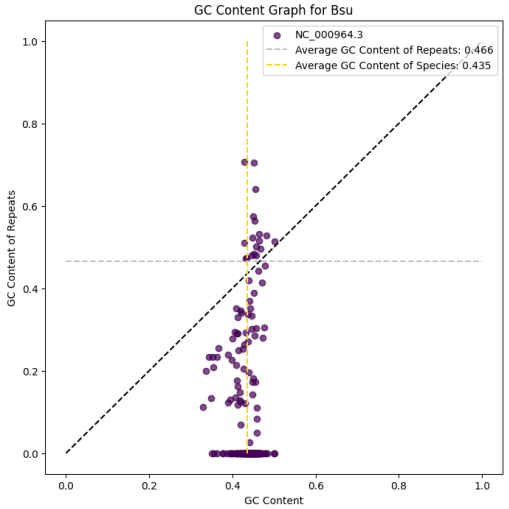

GC Content Graph for Bpr
NC_024008.1
NC_024007.1
NC_024006.1
NC_024005.1
NC_024004.1
NC_024003.1
NC_024002.1
NC_024001.1
NC_024000.1
NC_023999.1
NC_023998.1
NC_023997.1
NC_023996.1
NC_023995.1
NC_023994.1
NC_023993.1
NC_023992.1
NC_023991.1
NC_023990.1
Average GC Content of Repeats: 0.471
Average GC Content of Species: 0.481
GC Content of Repeats
GC Content
0.0
0.2
0.4
0.6
0.8
1.0

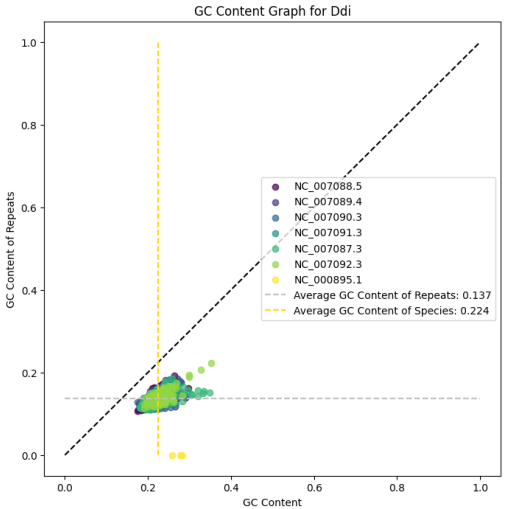
More:

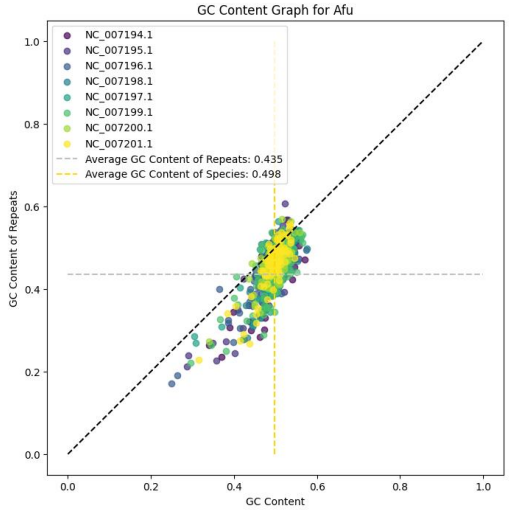
- Archaea: Thermococcus kodakarensis KOD1: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000009965.1/
- Amoebozoa: Dictyostelium discoideum AX4: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000004695.1/
- Mycobacterium tuberculosis H37Rv: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000195955.2/
- Streptomyces Coelicolor A3(2): https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_008931305.1/
- Aspergillus fumigatus: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002655.1/
- Oryza Sativa Japonica Group: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_034140825.1/
- Bacillus subtilis subsp. subtilis str. 168: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000009045.1/
- Haloferax volcanii DS2: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000025685.1/
- Pyrococcus furiosus DSM 3638: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_008245085.1/

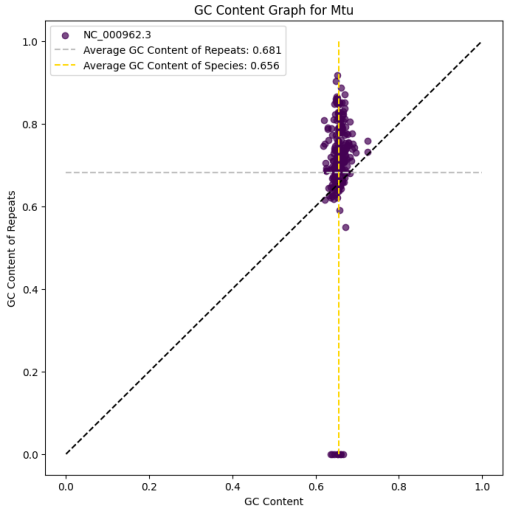

GC Content Graph for Tko
NC_006624.1
Average GC Content of Repeats: 0.454
Average GC Content of Species: 0.52
GC Content of Repeats
GC Content


GC Content Graph for Osa
NC_089035.1
NC_089037.1
NC_089036.1
NC_089038.1
NC_089040.1
NC_089045.1
NC_089039.1
NC_089041.1
NC_089042.1
NC_089046.1
NC_089043.1
NC_089044.1
Average GC Content of Repeats: 0.432
Average GC Content of Species: 0.437
GC Content of Repeats
GC Content


GC Content Graph for Bsu
NC_000964.3
Average GC Content of Repeats: 0.466
Average GC Content of Species: 0.435
GC Content of Repeats
GC Content

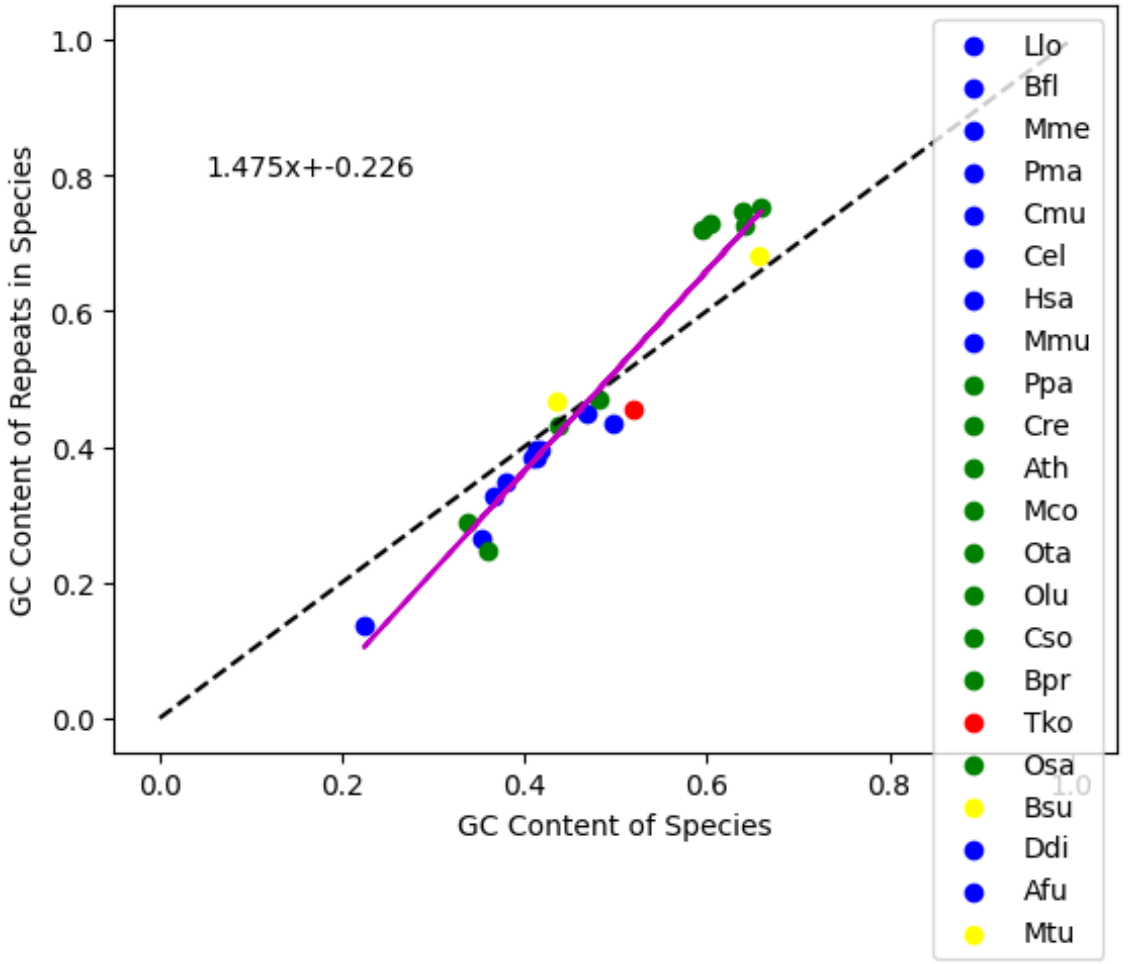

GC Content Graph for Ddi
NC_007088.5
NC_007089.4
NC_007090.3
NC_007091.3
NC_007087.3
NC_007092.3
NC_000895.1
Average GC Content of Repeats: 0.137
Average GC Content of Species: 0.224
GC Content of Repeats
GC Content


GC Content Graph for Afu
NC_007194.1
NC_007195.1
NC_007196.1
NC_007198.1
NC_007197.1
NC_007199.1
NC_007200.1
NC_007201.1
Average GC Content of Repeats: 0.435
Average GC Content of Species: 0.498
GC Content of Repeats
GC Content


GC Content Graph for Mtu
NC_000962.3
Average GC Content of Repeats: 0.681
Average GC Content of Species: 0.656
GC Content of Repeats
GC Content

Repeat Segments are more extreme?

- GC Content in repeats seem to increase more rapidly than the overall GC Content for all species
- Could be due to faster replication rate
  - if mutation bias exists
    - Lack of correlation in data from sliding window analysis contradicts this
  - if selective pressure
    - If so, why?
  - Would not be effected by biased gene conversion, as recombination
- Could be due to weaker constraints
  - Repeats aren't subject to as many selective constraints
- Error: [0.07691182 0.03715188]

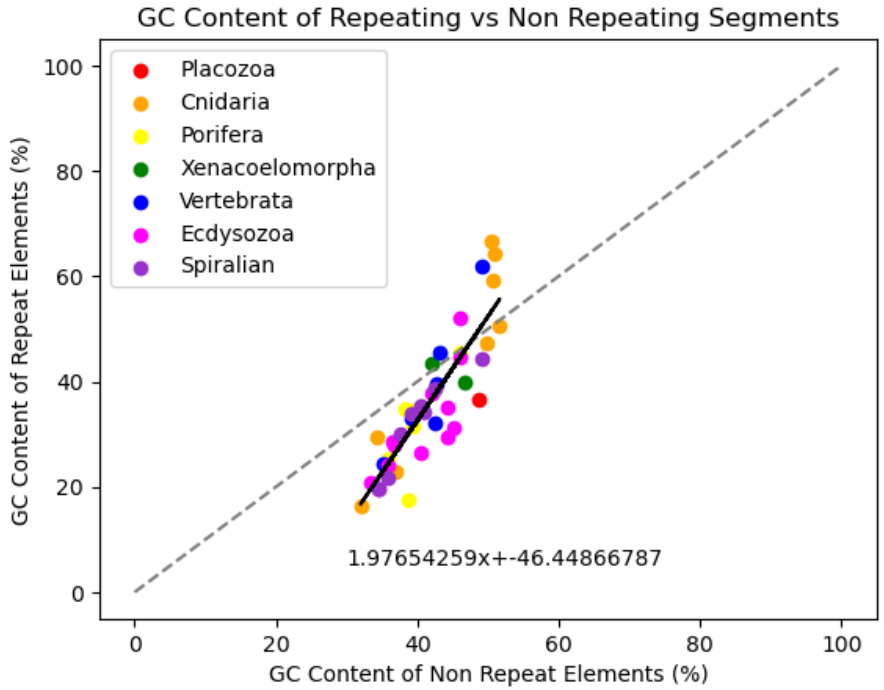


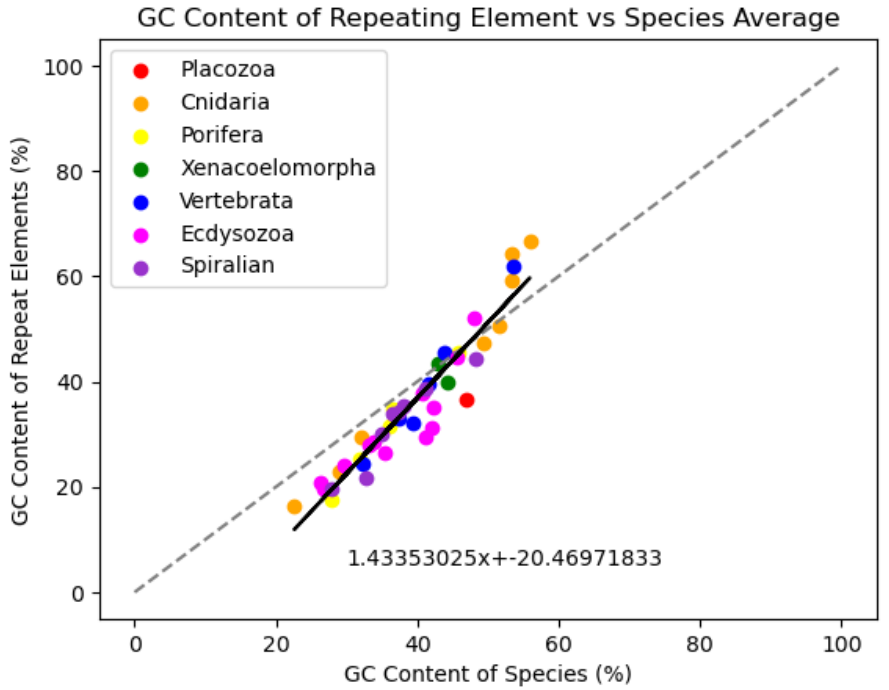
Viruses*

- Sars CoV 2: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009858895.2/
- HIV: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000864765.1/
- Japanese Encephalitis: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000862145.1/

*Can’t mask, sequences too short

Focusing on Eukaryotes





Error for R vs NR: [0.17515199 7.34616462]

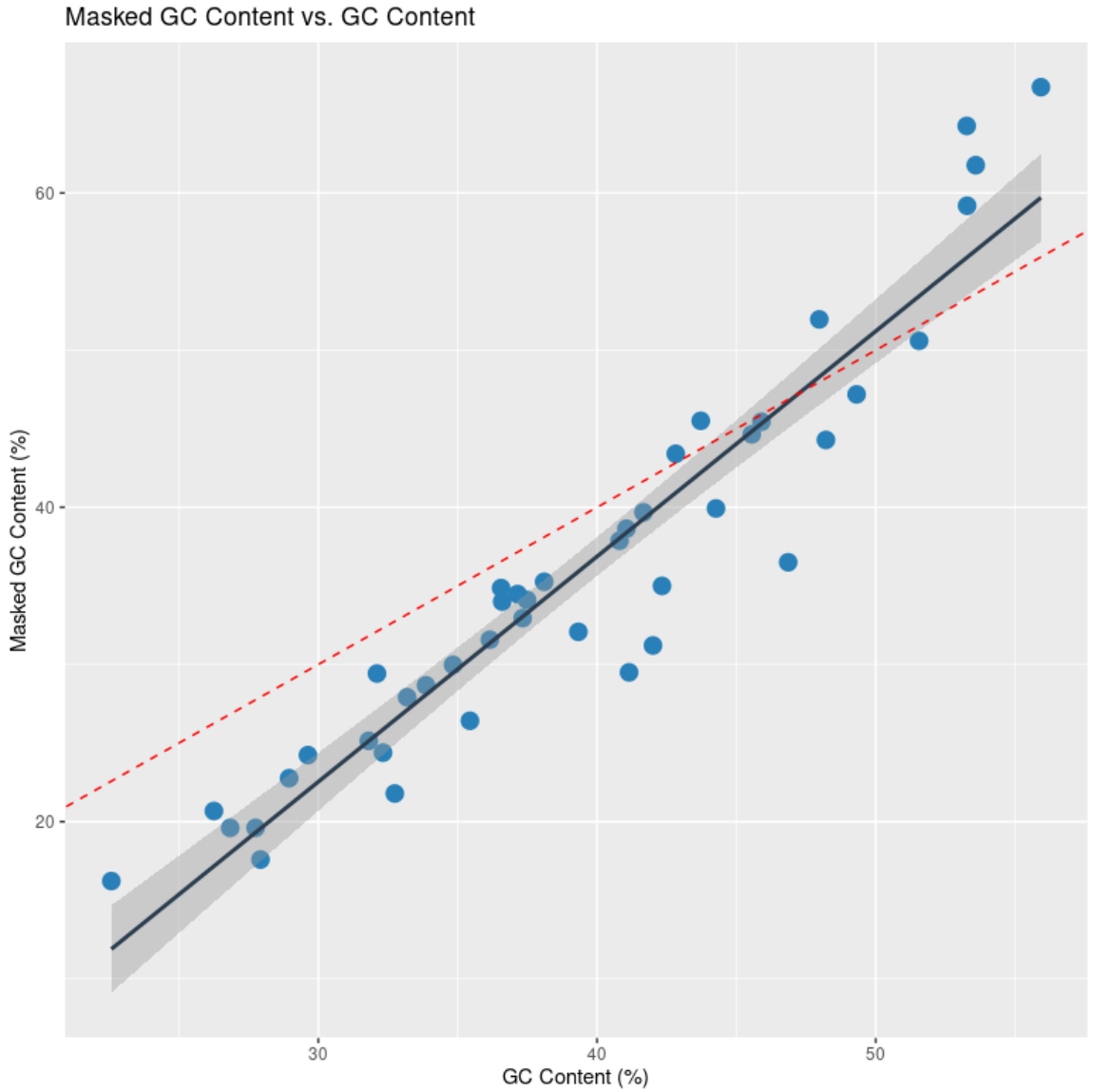
Error for R vs Species: [0.07408874 2.96851225]

If $x = y, \frac{46.44866787}{1.97654359-1} = 47.56440563$

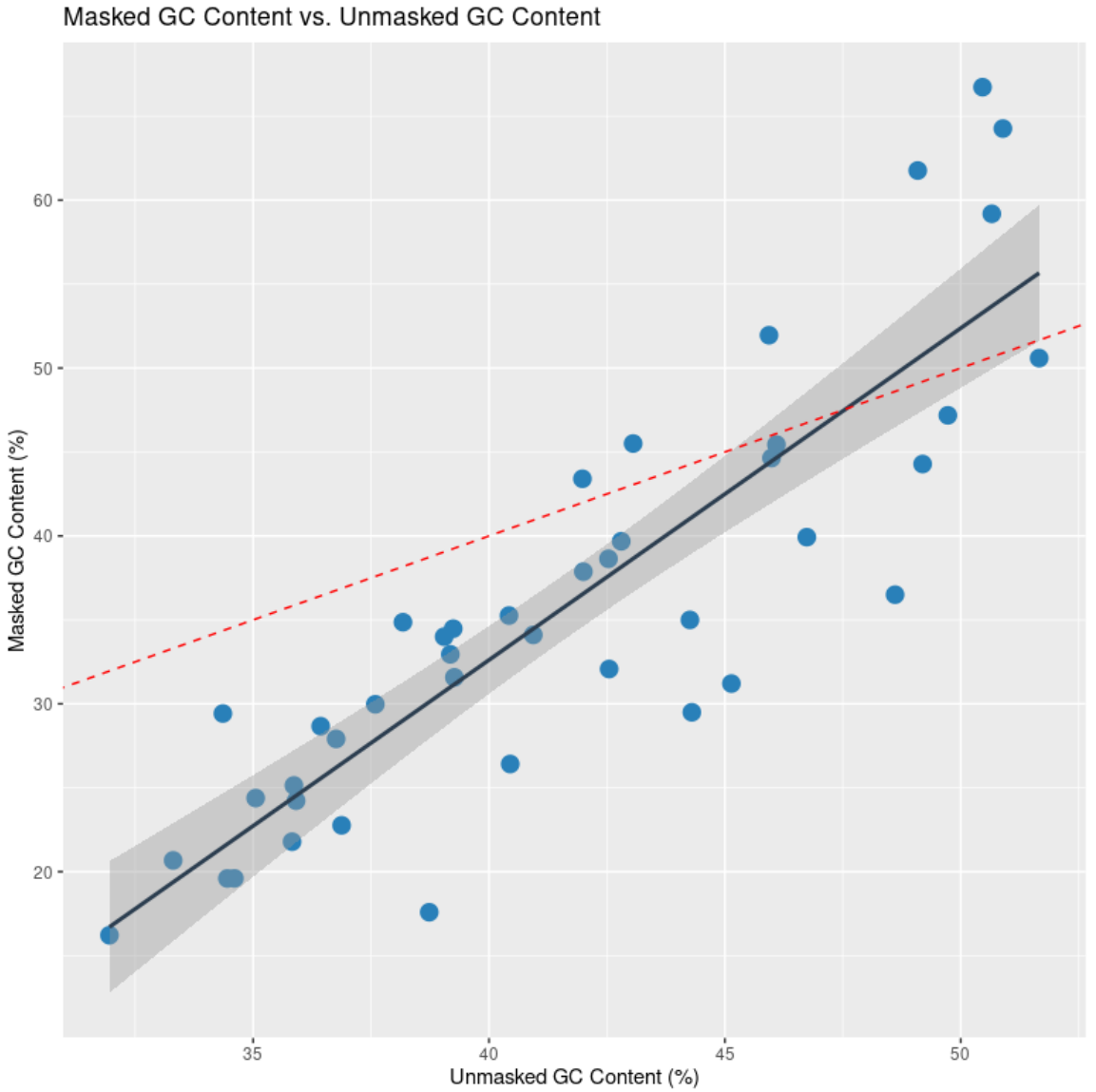
If $x = y, \frac{20.46971833}{1.43353025-1} = 47.21635533$

Error: $47.56440563\sqrt{\left(\frac{0.17515199}{1.97654259-1}\right)^2 + \left(\frac{7.34616462}{46.44866787}\right)^2} = 11.37408807$

Error: $47.21635533\sqrt{\left(\frac{0.07408874}{1.43353025-1}\right)^2 + \left(\frac{2.96851225}{20.46971833}\right)^2} = 10.58281436$

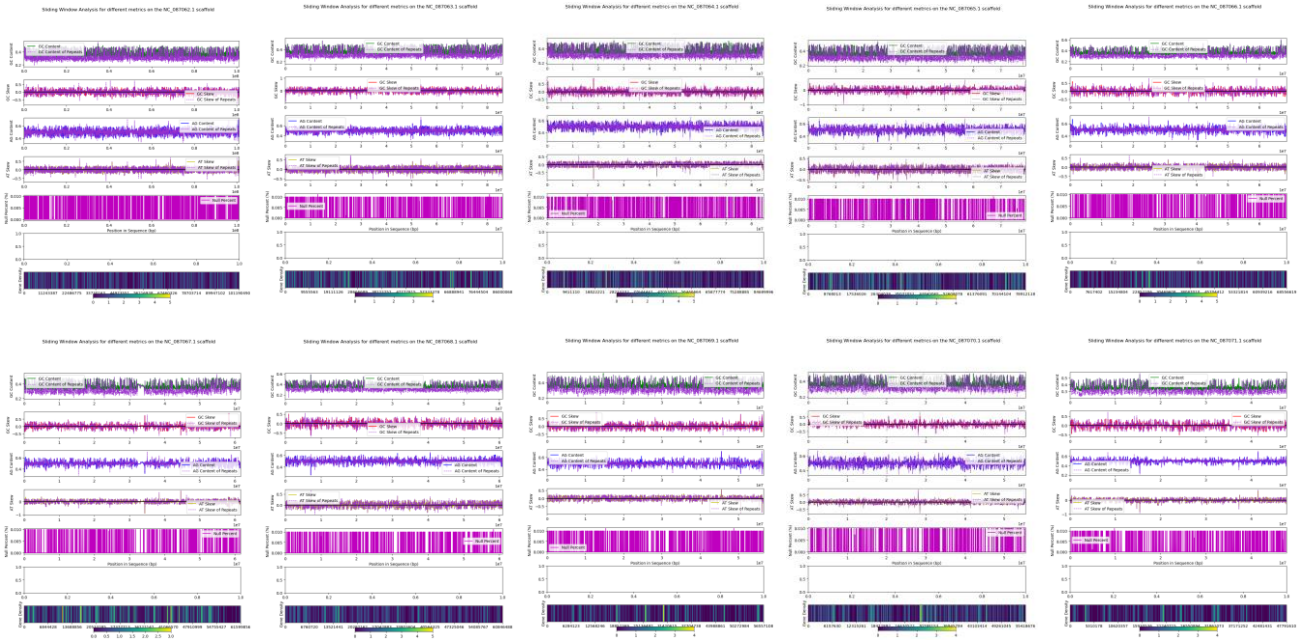


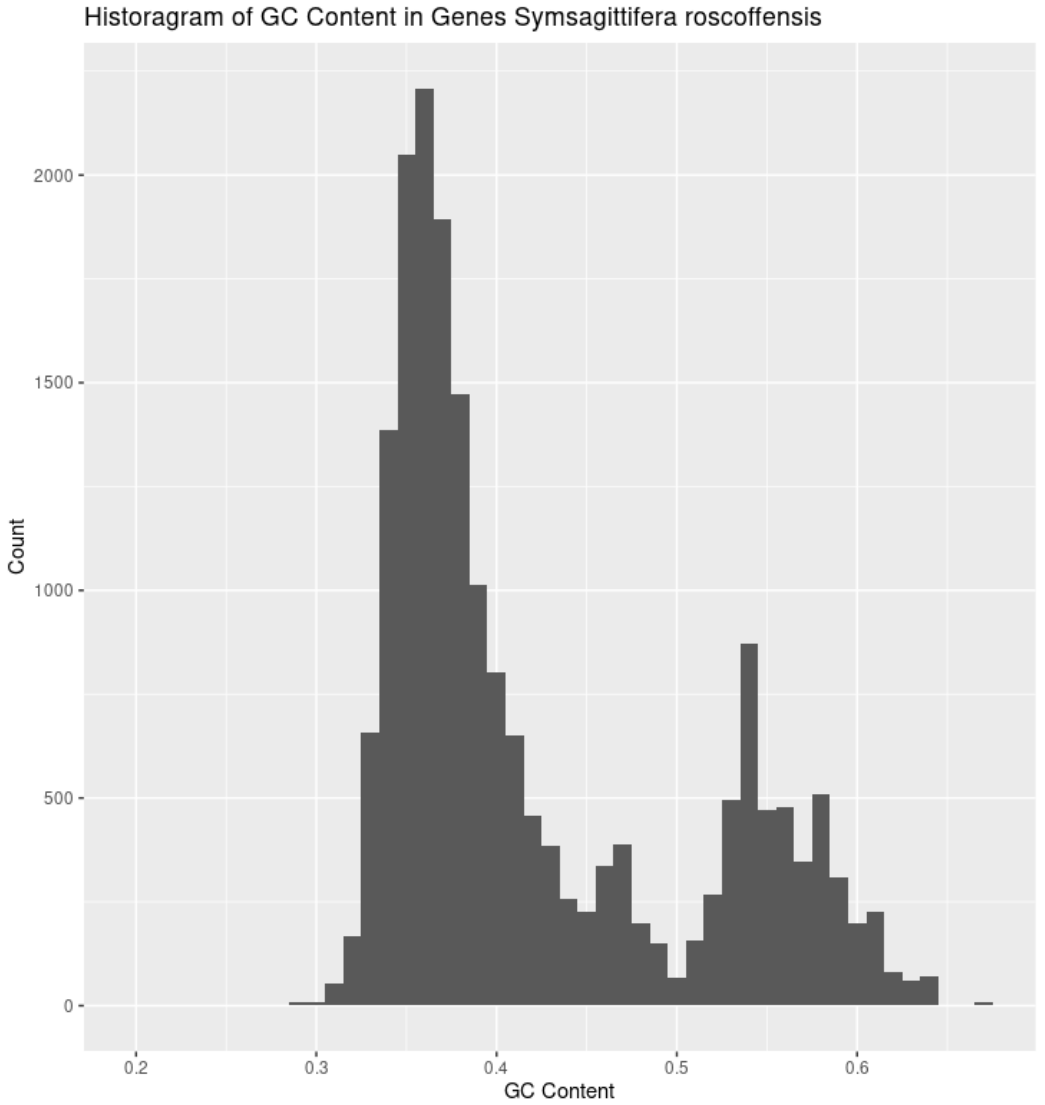
Call: lm(formula = masked_gc_content ~ gc_content, data = df) Residuals: Min 1Q Median 3Q Max -10.2002 -2.4081 0.4221 2.3565 8.3815 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -20.46972 2.96851 -6.896 2.31e-08 *** gc_content 1.43353 0.07409 19.349 < 2e-16 *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 3.997 on 41 degrees of freedom Multiple R-squared: 0.9013, Adjusted R-squared: 0.8989 F-statistic: 374.4 on 1 and 41 DF, p-value: < 2.2e-16



Call: lm(formula = masked_gc_content ~ unmasked_gc_content, data = df) Residuals: Min 1Q Median 3Q Max -13.1258 -4.1637 0.7848 3.1862 13.4383 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -46.4487 7.3462 -6.323 1.50e-07 unmasked_gc_content 1.9765 0.1752 11.285 3.76e-14 (Intercept) *** unmasked_gc_content *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 6.278 on 41 degrees of freedom Multiple R-squared: 0.7565, Adjusted R-squared: 0.7505 F-statistic: 127.3 on 1 and 41 DF, p-value: 3.761e-14

Histograms for detecting HGT?





Initial Observations

- Compositional Domains exist in S. roscoffensis
- Small subset of high GC genes seem to exist
- Don’t know if horizontal gene transfer would conserve GC content to an extent where you can use it as a test