Generating %GC over sliding window of 100bp

The resulting bam file `prochloro_R1_R2_B5.bam` was parsed using pySAM for python as seen below into windows of 100bp. For each window, the number of reads occurring in that window, and the % GC was calculated. The resulting csv file `prochloro_R1_R2_B5_parsed.csv` was used for downstream analysis and visualization in R studio.

```
# load packages
import pysam
import pandas as pd

# load data
samfile = pysam.AlignmentFile('prochloro_R1_R2_B5.bam',
  'rb', index_filename = 'prochloro_R1_R2_B5.bam.bai')

window = 100
cols = ['refname', 'window_n', 'start', 'stop', 'n_reads', 'gc_percent']
dat = pd.DataFrame(columns=cols)

for i in range(len(samfile.references)):
    refname = samfile.references[i]
    seqlen = samfile.lengths[i]
    for j in range(1, seqlen, window):
        stop = j+window-1 if j+window-1 < samfile.lengths[i] else samfile.lengths[i]
        window_n = stop/window
        region_set = set()
        sequence = read.query_sequence.upper()
        gc_percent = float(sequence.count('C') + sequence.count('G'))/window*100
        for read in samfile.fetch(refname, j, stop):
            region_set.add(read.query_name)
        dat = dat.append({'refname': refname, 'window_n':(window_n), 'start':j,
            'stop':stop, 'n_reads':len(region_set), 'gc_percent':gc_percent},
            ignore_index =True)
dat.csv('prochloro_R1_R2_B5_parsed.csv')
```

Visualization of enrichment

We visualized the entire genome to evaluate coverage and GC content in areas of enhanced recruitment (Figure 1). In Figure 1 we see that coverage is relatively even across the genome, with spikes of enhanced coverage. To confirm that read recruitment is not even across the genome, we tested the distribution of reads mapped in each 100bp window. We used an Anderson-Darling test for normal distribution. This test was chosed as it is better designed for large data sets such as ours where the number of windows = 17511. The results of our Anderson-Darling test indicate that read coverage is not normally distributed across the genome ($p<<0.05$). This is particularly notable in the region at ~36000 bp where coverage is >8000 reads. This spike clearly has enhanced GC content, however, it is unclear if other enhancement areas are also GC rich.

```
##
##  Anderson-Darling normality test
##
## data:  prochl$n_reads
## A = 6024.9, p-value < 2.2e-16
```

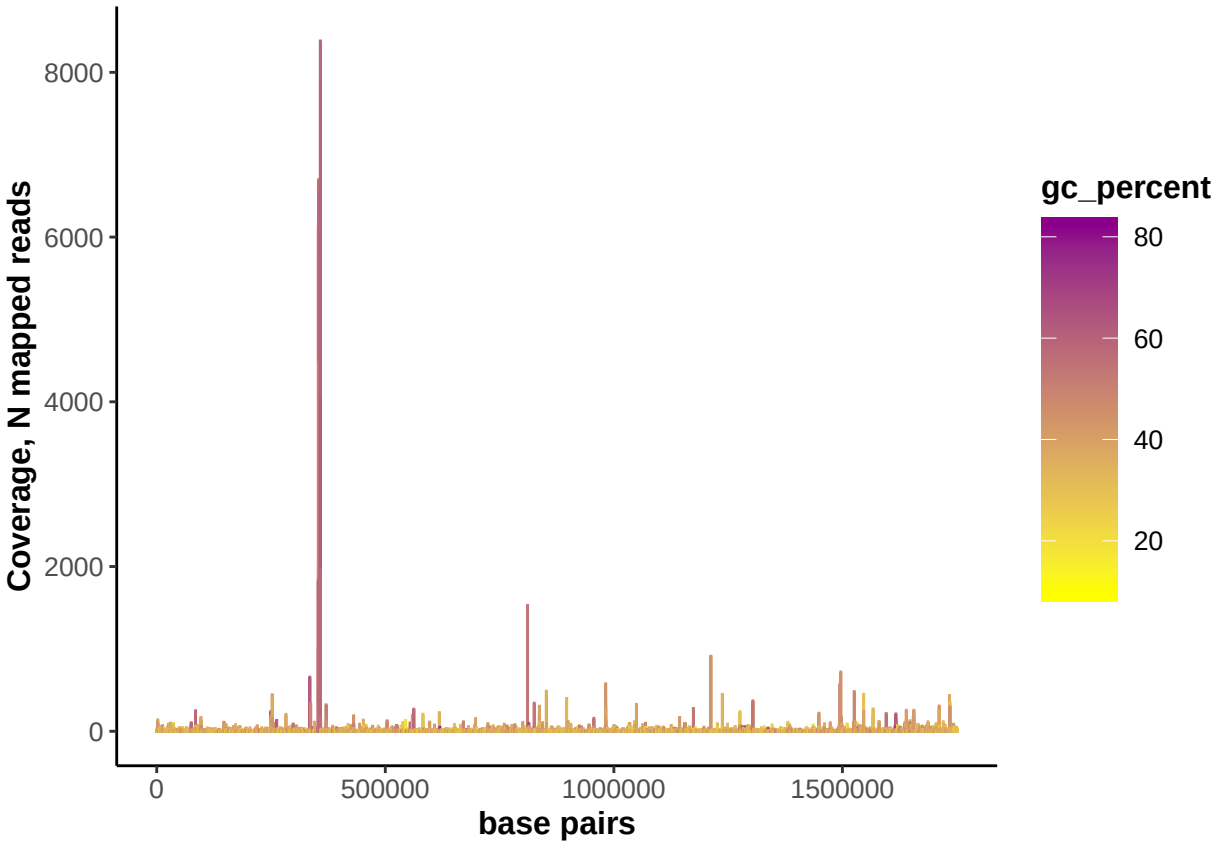

Figure 1: Coverage of Prochlorococcus genome across 100bp windows colored by GC content.

Determining enhanced coverage with high GC content

To determine if uneven coverage occurs with high GC content, we categorized "high" as >50%, and "low" as <50%. Given that coverage is not normally distributed for areas of high or low GC content (Figure 2), we performed a Mann-Whitney test. The null hypothesis is "a given read coverage value is equally likely to occur in high or low GC content areas". Based on results of this test, **we cannot reject the null hypothesis (p = 0.96)**, suggesting that exaggerated recruitment occurs in both high and low GC content as seen in Table 1, and visualized in Figure 2.

Table 1: Table 1: summary of high and low GC content areas

gc_label	window_count	max_coverage	mean	sd	observed_read_occurance
high	1027	8381	110.004869	693.6555	0.4672134
low	16484	4480	7.815518	47.1646	0.5327866

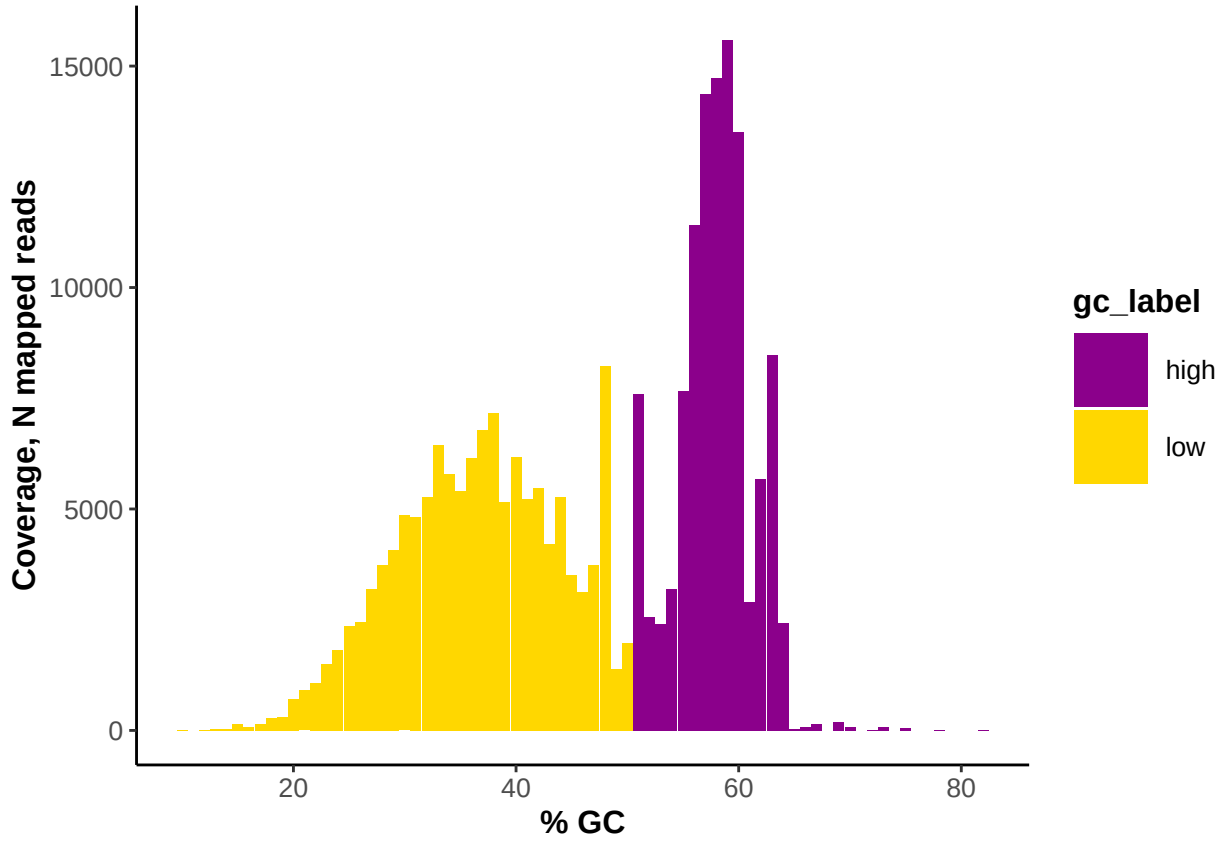

Figure 2: GC % vs. Read coverage.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  n_reads by gc_label
## W = 8471500, p-value = 0.9645
## alternative hypothesis: true location shift is not equal to 0
```