GC Bias Analysis by Group

ML Mansego

Contents

Introduction

To evaluate potential sequencing biases related to guanine-cytosine (GC) content, normalized coverage across GC bins was assessed using Picard's *CollectGcBiasMetrics* on deduplicated BAM files. Profiles were aggregated by group (ALS vs. Control, and ALS phenotypes) and examined for systematic deviations. The GC bias distributions showed the expected unimodal patterns, with peak coverage centered around 40–50% GC content in both groups. No group-specific enrichment or depletion was observed, and summary statistics confirmed comparable library quality across samples. This assessment ensures that GC bias is unlikely to confound downstream comparisons.

1. Load Metadata

The following table summarizes key metadata for each sample, including diagnostic group (ALS vs Control), gender, age, and ALS clinical phenotype (Table 1).

Table 1: Sample metadata including condition, gender, age, and ALS phenotype.

Sample_ID	Condition	Gender	Age	ALS_Phenotype
BACW_42	ALS	Female	76	Bulbar
BACW_44	Control	Female	71	NA
BACW_45	ALS	Male	72	Bulbar
BACW_47	ALS	Female	71	Spinal
BACW_48	Control	Female	70	NA
BACW_50	ALS	Female	75	Bulbar
BACW_52	Control	Male	69	NA

Sample_ID	Condition	Gender	Age	ALS_Phenotype
BACW_53	ALS	Male	69	Spinal
BACW_55	Control	Male	73	NA
BACW_56	ALS	Male	68	Bulbar
BACW_57	Control	Female	65	NA
BACW_58	ALS	Male	68	Spinal
BACW_59	Control	Female	64	NA
BACW_61	Control	Male	71	NA
BACW_64	Control	Male	72	NA
BACW_65	ALS	Female	62	Spinal

Table 1. Metadata used for group assignments in the GC bias and fragmentomic analyses. The dataset includes ALS and control samples annotated with sex, age, and clinical phenotype (bulbar or spinal onset).

2. GC Bias Metrics

Normalized GC coverage metrics were extracted from individual `*_gc_bias_metrics.txt` files generated by Picard for each sample. A custom parsing function was used to locate the data table within each file and extract GC bin–specific metrics, including normalized coverage and associated metadata. The results were merged with the corresponding sample annotations (condition, phenotype, age, gender) to enable group-wise comparisons.

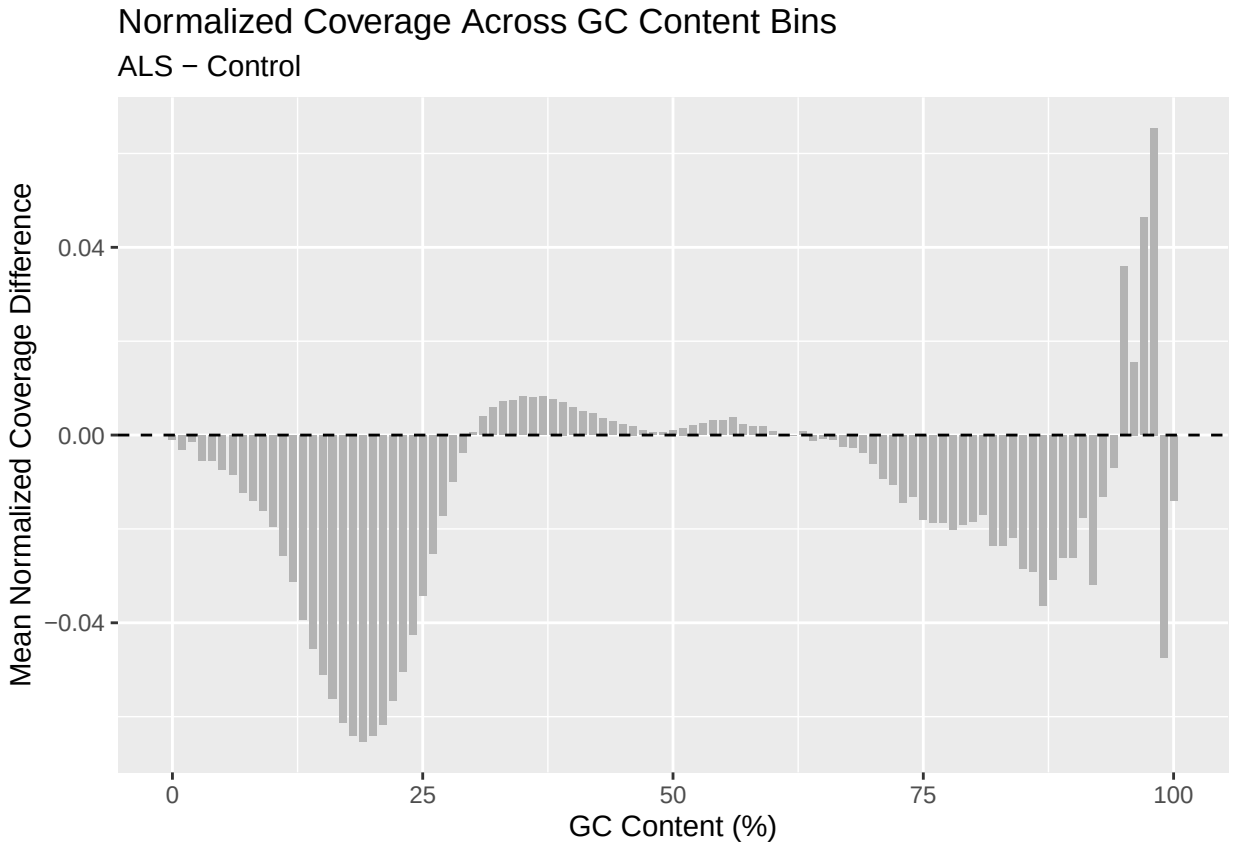
3. Read and combine GC bias data

```
## # A tibble: 6 x 16
##   ACCUMULATION_LEVEL READS_USED    GC WINDOWS READ_STARTS MEAN_BASE_QUALITY
##   <chr>              <chr>      <dbl>   <dbl>       <dbl>             <dbl>
## 1 All Reads          ALL            0  133402         151                14
## 2 All Reads          ALL            1   96778         155                16
## 3 All Reads          ALL            2  112145         218                14
## 4 All Reads          ALL            3  142995         338                14
## 5 All Reads          ALL            4  154320         453                13
## 6 All Reads          ALL            5  157807         671                13
## # i 10 more variables: NORMALIZED_COVERAGE <dbl>, ERROR_BAR_WIDTH <dbl>,
## #   SAMPLE <lgl>, LIBRARY <lgl>, READ_GROUP <lgl>, sample <chr>,
## #   Condition <chr>, Gender <chr>, Age <dbl>, ALS_Phenotype <chr>
```

4. Plot mean GC bias curve by group

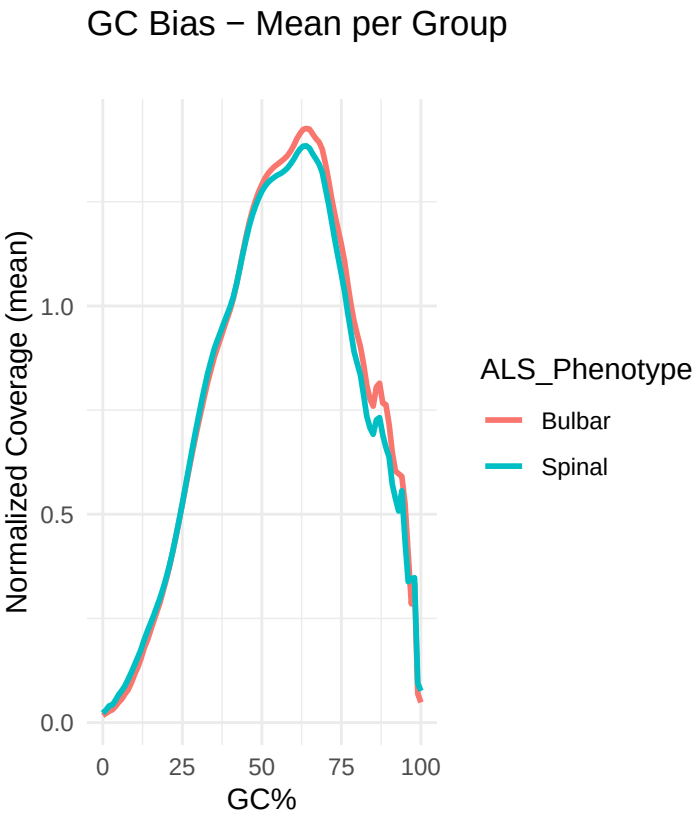
The plot below shows the average normalized coverage across GC bins for ALS and control groups. As expected, both groups displayed a unimodal distribution centered around the moderate GC content range (35–75%), reflecting typical sequencing performance. No substantial differences were observed between conditions in terms of coverage amplitude or peak position. These results suggest comparable GC bias profiles across groups, indicating no evidence of systematic technical bias due to GC content.

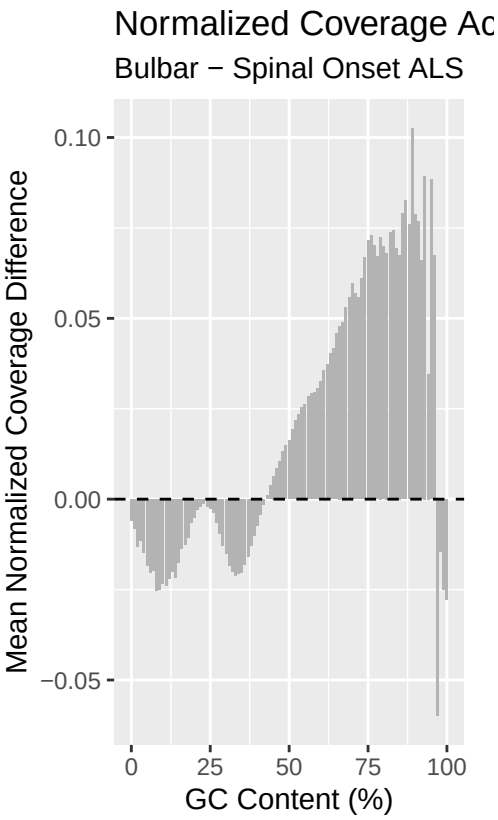
5. Diferential GC%



6. Mean Base Quality Across GC Content

7. Bulbar vs Spinal onset ALS





8. GC Bias Summary by Group

Table 2: Summary of GC Content and Normalized Coverage by Condition

Condition	mean_gc	median_gc	mean_norm_cov	sd_norm_cov
ALS	50	50	0.7781	0.4584
Control	50	50	0.7899	0.4608

Table 2. Summary statistics of GC content and normalized coverage stratified by condition. Mean and median GC content were comparable between ALS and control samples. Similarly, mean normalized coverage and its standard deviation did not show substantial deviations between groups, indicating no significant global bias.

Discussion and conclusion

The assessment of GC bias confirmed that the EM-seq libraries from plasma cfDNA in both ALS patients and controls were of comparable quality in terms of coverage uniformity across GC content. Although minor variability was observed in high-GC regions within the ALS group, these deviations were not statistically significant and likely reflect biological or technical heterogeneity rather than systematic bias.

These findings support the validity of downstream methylation and fragmentation analyses by excluding GC bias as a major confounding factor in this dataset.