TITLE: DECREASING THE EFFECTS OF GC BIAS ON *DE NOVO* GENOME ASSEMBLY

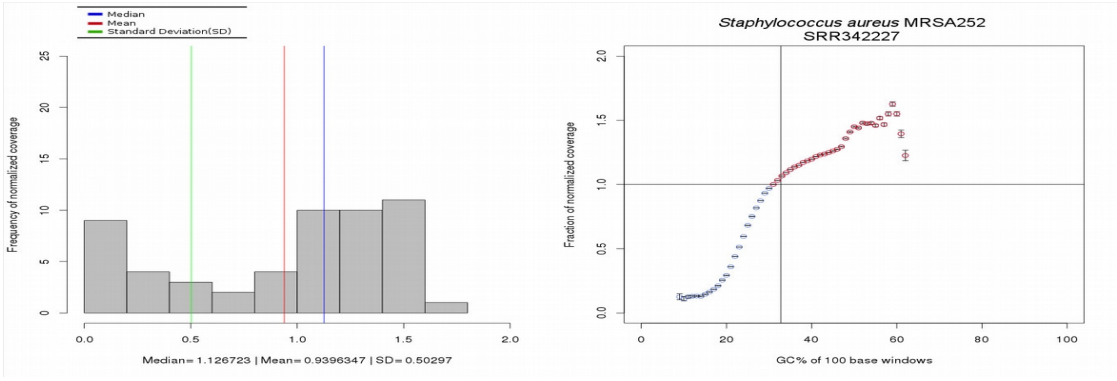
AUTHORS : PINHEIRO, K.C.; MAUES, D.; SILVA, A.; RAMOS, R. T. J.

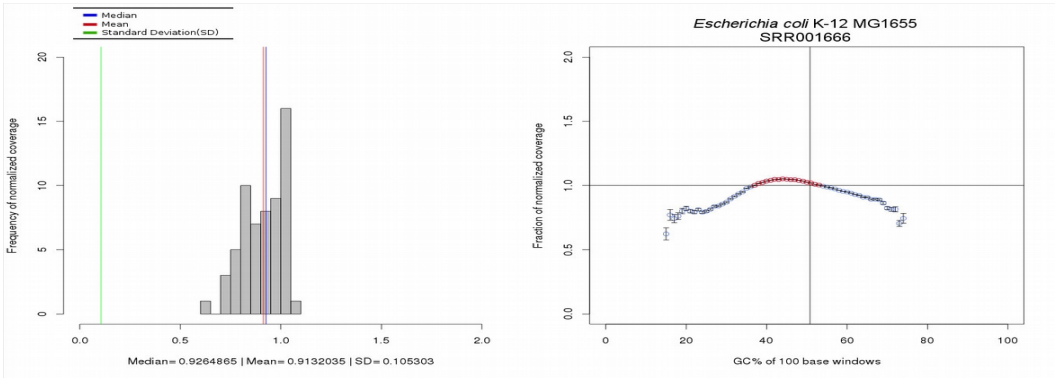
INSTITUTIONS: 1-INSTITUTE OF BIOLOGICAL SCIENCES, LABORATORY OF BIOLOGICAL ENGINEERING, FEDERAL UNIVERSITY OF PARA (R. AUGUSTO CORRÊA, 1 - GUAMÁ, BELÉM - PA, 66075-110)-BRAZIL

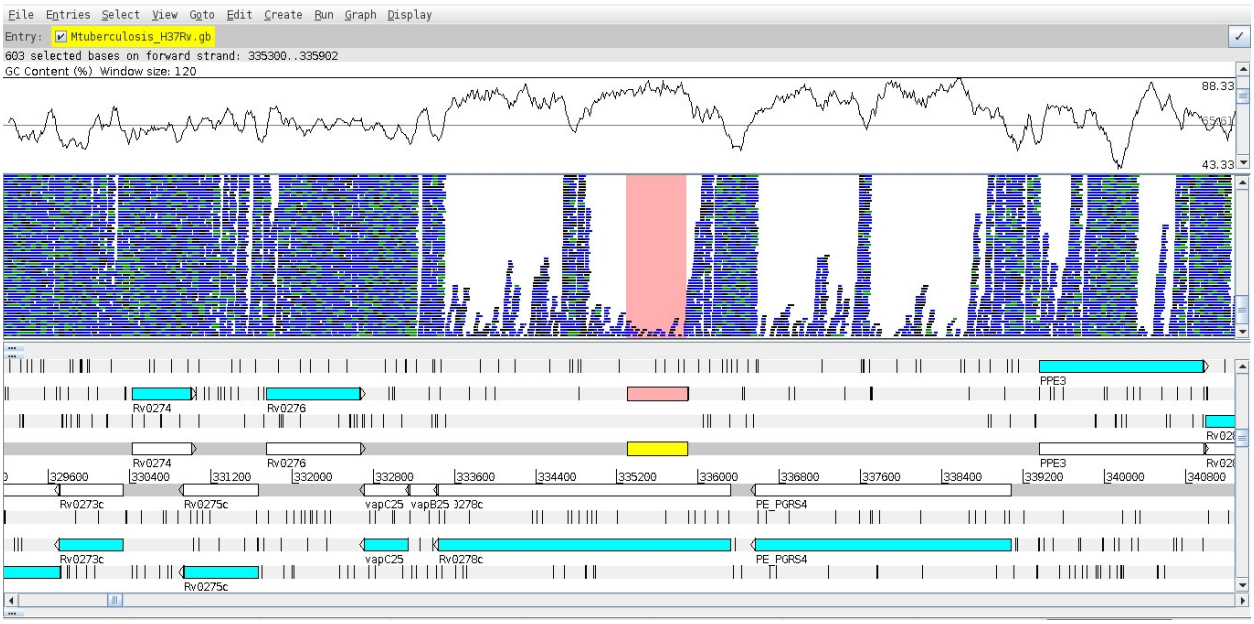
ABSTRACT:
The emergence of high throughput sequencing (HTS) platforms increased the amount of data making feasible to obtaining complete genomes. Despite the advantages and the throughput produced by these platforms, the different genomic coverage in the regions of the genome can be related to GC content. This GC bias may affect genomic analyzes and the genomic/transcriptomic analysis based on *de novo* and reference approach. In addition, the ways to evaluate the GC bias should be fit to data with different profiles of the relationship GC vs coverage, such as linear and non-linear. We used the data of 5 species: *Pseudomonas fluorescens*, *Shewanella amazonensis*, *Escherichia coli*, *Staphylococcus aureus* and *Mycobacterium tuberculosis* sequenced by Illumina platform. This paper proposes, the use of a new metric to measure GC bias regardless of pattern observed between GC content and coverage, then, based on a degree of GC Bias, use a new approach to identify and reduce gaps from such regions by run a sliding window in a reference genome and separate the reads by GC Content range, creating groups, to submit them to a tools to discover the best k for a k-mer *de novo* assembly. Our analysis showed that the new metric called “Median – IQR” (difference between Median and Interquartile range - IQR) is able measure different bias patterns (monotonic and non-monotonic) and also could capture different degrees of bias in different samples. Also, the new method to assembly data with high degree of GC bias could represent regions that would not be identified in data assembled with default parameters.

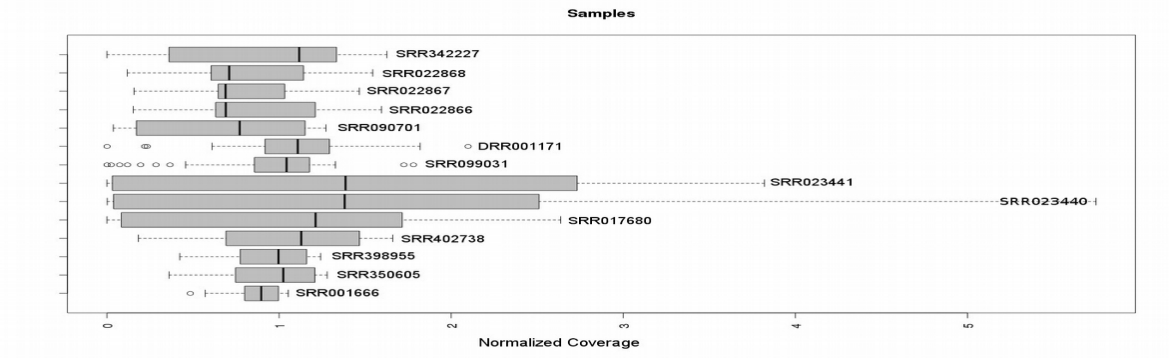
Keywords: GC Bias, Median, Interquartile range, assembly

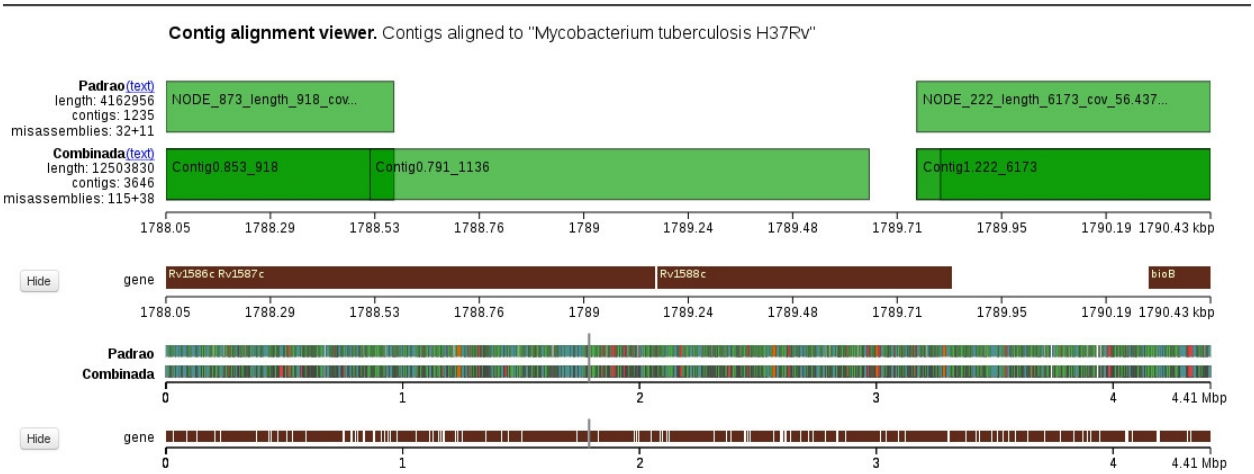
Development Agency: CNPq, CAPES


Median
Mean
Standard Deviation(SD)
Frequency of normalized coverage
Median= 1.126723 | Mean= 0.9396347 | SD= 0.50297
Staphylococcus aureus MRSA252
SRR342227
Fraction of normalized coverage
GC% of 100 base windows


Median
Mean
Standard Deviation(SD)
Frequency of normalized coverage
Median= 0.9264865 | Mean= 0.9132035 | SD= 0.105303
Escherichia coli K-12 MG1655
SRR001666
Fraction of normalized coverage
GC% of 100 base windows


File Entries Select View Goto Edit Create Run Graph Display
Entry: Mtuberculosis_H37Rv.gb
603 selected bases on forward strand: 335300..335902
GC Content (%) Window size: 120
88.33
43.33
PPE3
Rv0274
Rv0276
Rv0274
Rv0276
PPE3
329600
330400
331200
332000
332800
333600
334400
335200
336000
336800
337600
338400
339200
340000
Rv0273c
Rv0275c
vapC25
vapB25
Rv0278c
PE_PGRS4
Rv0273c
vapC25
Rv0278c
PE_PGRS4
Rv0275c


Samples
SRR342227
SRR022868
SRR022867
SRR022866
SRR090701
DRR001171
SRR099031
SRR023441
SRR023440
SRR017680
SRR402738
SRR398955
SRR350605
SRR001666
0
1
2
3
4
5
Normalized Coverage


Contig alignment viewer. Contigs aligned to "Mycobacterium tuberculosis H37Rv"
Padrao (text)
length: 4162956
contigs: 1235
misassemblies: 32+11
Combinada (text)
length: 12503830
contigs: 3646
misassemblies: 115+38
NODE_873_length_918_cov...
NODE_222_length_6173_cov_56.437...
Contig0.853_918
Contig0.791_1136
Contig1.222_6173
1788.05
1788.29
1788.53
1788.76
1789
1789.24
1789.48
1789.71
1789.95
1790.19
1790.43 kbp
Hide
gene
Rv1586c Rv1587c
Rv1588c
bioB
Padrao
Combinada
0
1
2
3
4
4.41 Mbp