# Abstract

# Introduction

Metagenomic sequencing data has increased the understanding of the role of the microbiome and given insight into a variety of (bacterial) communities. Korem *et al.* (2015) used the pattern of metagenomic sequencing read coverage around the origin of replication (oriC) to gain insight into the growth dynamics of gut microbiota. Detecting the oriC in bacteria can hence be of great importance for metagenomic analysis. With the oriC being the starting point of bacterial replication, its detection can facilitate the analysis of certain motifs and features around this location or simply be used as a starting point for gene annotation.

A frequently used indicator of the oriC is the so-called skew which is based on the strand asymmetry between the leading and lagging strand. The leadings strand usually is rich in guanine (G) and adenine (A) whereas a higher content of cytosine (C) and thymine (T) can be found in the lagging strand (Touchon and Rocha, 2008). The putative location of the oriC is then indicated by a minimum of the GC skew (G-C)/(G+C) or a maximum in the AT skew respectively. Even though the GC skew is frequently used, the AT skew is able to cover for less pronounced regions in the GC skew (Grigoriev, 1998).

Another measure used for oriC prediction is the DnaA motif. DnaA is the key protein in the initiation of replication. It binds to clusters of DnaA boxes that accumulate around the oriC. The DnaA motif is nine base pairs long and is highly conserved with the consensus sequence

“*TT TNCACA*” (Blaesing *et al*., 2017). Detecting the location of these clusters of DnaA boxes can therefore improve the prediction of a putative oriC.

This project tries to identify putative oriC locations of four different bacteria species (*Escherichia coli, Vibrio cholerae, Salmonella enterica, Thermotoga petrophila*). The approximate region of the oriC is identified using GC and AT skews. In this region, the DnaA motif clusters are then located to further specify the region.

# Material and Methods

## Data

Reference Sequences were downloaded from NCBI Assembly database.

**Table 1:**

|  |  |
| --- | --- |
| **Organism** | **Family** |
| Escherichia coli | Enterobacteriaceae |
| Vibrio cholerae | Vibrionaceae |
| Thermotoga petrophila | Thermotogaceae |
| Salmonella enterica | Enterobacteriaceae |

## Skew Diagrams

First the GC-skew minimum was calculated for reference strains of the bacterial families of *Enterobacteriaceae* and *Vibrionaceae* to obtain an overview of the values to be expected. In the next step the minimum was determined for 10 strains of the four species *Escherichia coli*, *Salmonella enteritidis*, *Vibrio cholerae*, and *Thermotoga petrophila*. Since only one complete genome of *T. petrophila* was available, other Thermotoga *subspecies* were considered too. Determined GC minima were then compared to the minima and the oriC start position given in the DoriC database.

The DnaA motifs were analysed by profile alignment in a 2000 window around the determined GC minimum, including also the reverse complement. Motifs with a score ≥ 10 (to allow some mismatches/variations of the motif) were counted and compared to the number of motifs at the oriC as given in the DoriC database.

The same procedure was applied to *Wigglesworthia glossinidia*. Additionally, for *W. glossinidia* the number of GC minima was read from the plot.

## Motif Finding

position-specific scoring matrix (PSSM) calculated

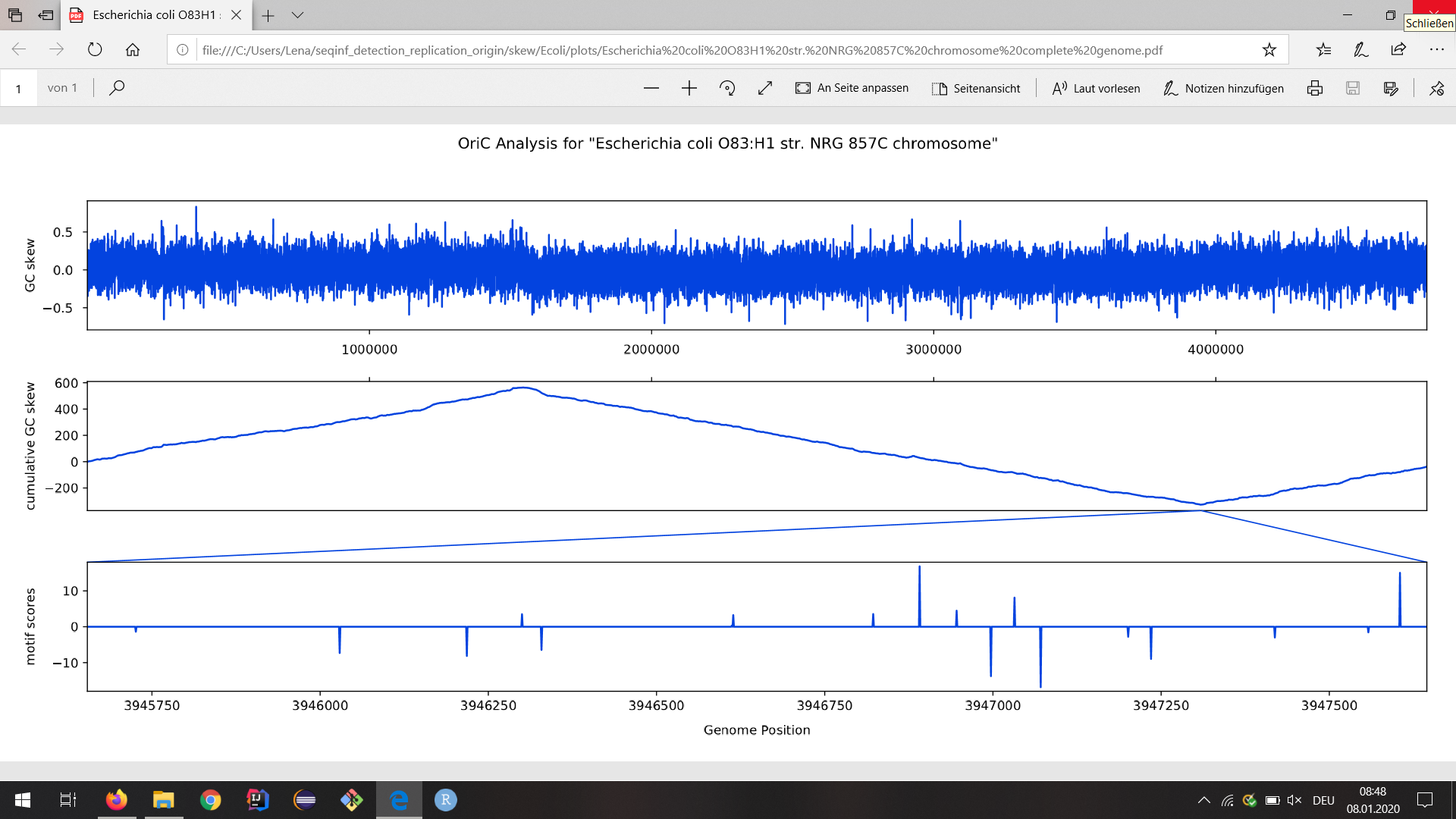
PSSM score for each position of the OriC region sequences

## Meme

Command-Line version of MEME Suite[[1]](#footnote-0) is used to compute relevant sequence motifs of the prokaryotic species and families. MEME Suite is an implementation of the MEME algorithm (Multiple Em for Motif Elicitation) which besides Gibbs sampling is a motif discovery algorithm. MEME is based on the concept of expectation–maximization and position dependent probability based matrices. Therefore meme does not support gaped or shifted sequences.

## 

# Results



**Figure 1**: Plot example for *Escherichia coli O83:H1* shows GC skew (top), cumulative GC skew (middle) and the profile alignment motif scores (bottom).

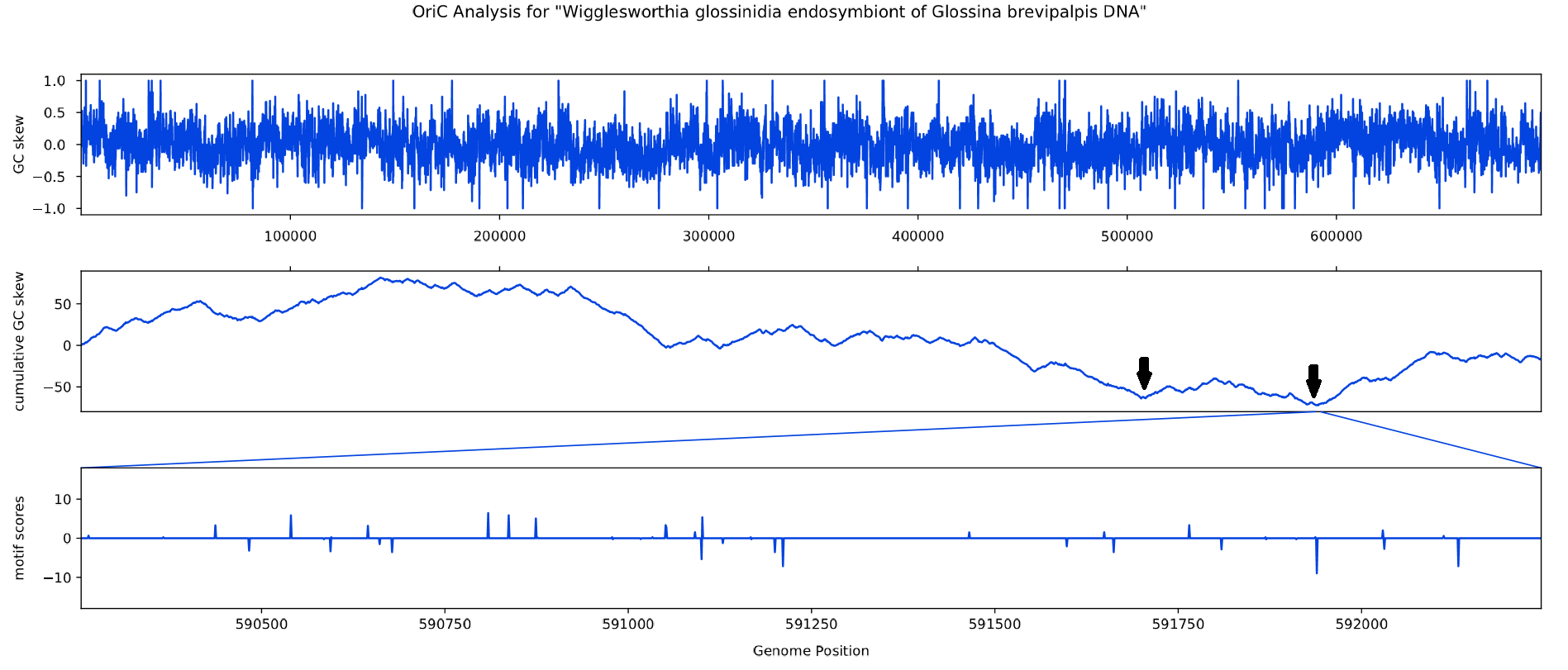


Figure 2: Plot for Wigglesworthia glossinidia. Arrows indicate two potential minima(?).

**Table 2:** Average distances of determined minima to minima (𝞓 Minimum) and oriC start position (𝞓 oriStart) in DoriC as well as observed number of (reverse complement) motifs with scores >= 10 and their average distance to the number of DnaA boxes (𝞓 Number of Motifs) in the oriC region as given by DoriC.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Strain/Family** | **𝞓 Minimum** | **𝞓 oriStart** | **Number of Motifs** | **Number of Rev. Comp.** | **𝞓 Number of Motifs** |
| **Known OriC** | 105 ± 28 | 1085 ± 1073 | 1 ± 1 | 1 ± 1 | 2 |
| **References (E.coli)** | 164 ± 189 | 523 ± 881 | 2 ± 0 | 2 ± 0 | 2 |
| **Enterobacteriaceae** | 186 ± 207 | 600 ± 645 | 2 ± 1 | 2 ± 1 | 2 |
| **Vibrionaceae** | 1099008 ± 1895994 | 2086124 ± 1806388 | 2 ± 1 | 2 ± 1 | 1 |
| **Escherichia coli** | 191 ± 250 | 177 ± 227 | 2 ± 1 | 2 ± 0 | 2 |
| **Salmonella enterica** | 152 ± 117 | 534 ± 165 | 3 ± 1 | 2 ± 1 | 1 |
| **Vibrio cholerae** | 115 ± 98 | 3497 ± 3361 | 1 ± 1 | 1 ± 1 | 5 |
| **Thermotoga sp.** | 70 ± 26 | 237 ± 185 | 0 ± 1 | 1 ± 1 | 7 |
| **Wigglesworthia glossinidia** | 467 | 106176 | 0 | 0 | 0 |

[Lena] Known OriC als “Kontrolle”, References evtl löschen. Vibrio Erklärung hinzufügen

# Discussion

Z-curve method would likely improve the OriC localization.[[2]](#footnote-1)

# Literature

Blaesing, F., Weigel, C., Welzeck, M. and Messer, W. (2017) Analysis of the DNA‐binding

domain of Escherichia coli DnaA protein. *Molecular microbiology*, **36**(3), 557-569.

Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic acids*

*research*, **26**(10), 2286-2290.

Korem, T., Zeevi, D., Suez, J., Weinberger, A., Avnit-Sagi, T., Pompan-Lotan, M., Matot, E.,

Jona, G., Harmelin, A., Cohen, N., Sirota-Madi, A., Pevsner-Fischer, M., Sorek, R.,

Xavier, R., Elinav, E. and Segal, E. (2015) Growth dynamics of gut microbiota in health

and disease inferred from single metagenomic samples. *Science*, **349**(6252), 1101-1106.

Touchon, M., and Rocha, E. P. (2008) From GC skews to wavelets: a gentle guide to the

analysis of compositional asymmetries in genomic data. ​*Biochimie*, ​**90**(4), 648-659.

1. [↑](#footnote-ref-0)
2. "Identification of replication origins in archaeal ... - NCBI." <https://www.ncbi.nlm.nih.gov/pubmed/15876567>. Aufgerufen am 7 Jan.. 2020. [↑](#footnote-ref-1)