# 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 nucleotide skew which is based on the strand asymmetry between the leading and lagging strand. The leading 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. Skews can hence provide a simple and quick measure to detect strand asymmetries (Touchon and Rocha, 2008).

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 (Mackiewicz et *al*. 2004). The DnaA motif is nine base pairs long and is highly conserved in most bacteria 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 a GC skew. For this region, the DnaA motif is then determined and and its occurrences are counted.

→ [Unser Ziel muss nochmal überarbeitet werden]

# Material and Methods

## Data

Reference sequence genomes were downloaded from NCBI Assembly database. Table 1 shows the organisms used and their respective families. A list of all strains can be found in the appendix.

**Table 1:** Organisms and their respective family

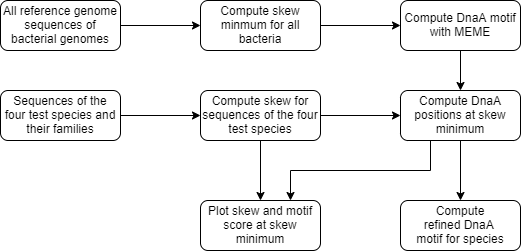
|  |  |  |  |
| --- | --- | --- | --- |
| **Organism** | **#sequences** | **Family** | **#sequences** |
| Escherichia coli | 10 | Enterobacteriaceae | 12 (reference) |
| Vibrio cholerae | 10 | Vibrionaceae | 3 (reference) |
| Thermotoga petrophila | 10 (including other species) | Thermotogaceae | n/a |
| Salmonella enterica | 10 | Enterobacteriaceae | 12 (reference) |

## 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 with respect to the genome length. Three sequences with a known OriC were used as controls.

The DnaA motifs were analysed by profile alignment in a 2000 window around the determined GC minimum, including also the reverse complement (see also section “Motif Finding”). To allow for mismatches/variations, motifs with a score ≥ 10 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 determined.



**Figure 1:** Overview of the project workflow.

## Meme

The Command-Line application MEME of the MEME Suite (Bailey *et al*., 2009) was used to compute relevant sequence motifs of the prokaryotic species and families. MEME Suite contains 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 gapped or shifted sequences.

2000 bases around the GC-skew minimum of bacterial reference sequences were used to compute a more refined motif for all bacterias with MEME. For the execution the common DnaA consensus sequence TTATCCACA were given as starting point and a 4th order markov model were used for the background normalization.

Multiple motifs for species and families were computed with the same approach. But motif discovery with MEME does not deliver satisfying results for few sequences. To estimate a more specific DnaA motif for few sequences of a species another algorithm had to be implemented.

## Motif Finding

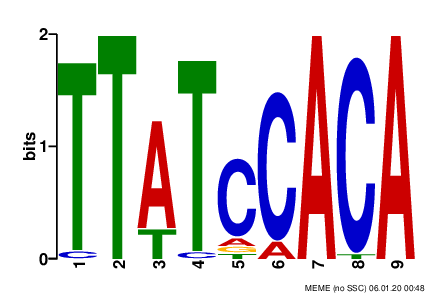
For the scoring and visualization of the DnaA motifs a information theory based approach was used. To find DnaA occurrences in 2000 bases around GC-skew minimum each position of the sequence was scored using Individual Information (Rogan *et al*., 1998). For the underlying position-specific scoring matrix (PSSM) the pseudocounts were set to 1 and uniform background frequencies were assumed.

For the plot only positive bit scores were visualized. All negative scores were ignored. On the other hand matches on the reverse complement were treated as negative values for clarity. The result is visualized at the bottom of the “OriC Analysis” plot.

The scores were also used to create a Weblogo (Crooks *et al*., 2004) for specific species and families. Using the general search positions near sequence GC-skew minima of bacterial species and families matching a more general DnaA motif. From matches at these positions a new motif has been generated with the same approach representing a specific DnaA motif of the species or family. The threshold for a match had to score higher than 8.1 bit which conforms an approximative false positive rate of 0.02 %.

## 

# Results



**Figure 2:** DnaA motif over all bacterial reference genomes computed by MEME.

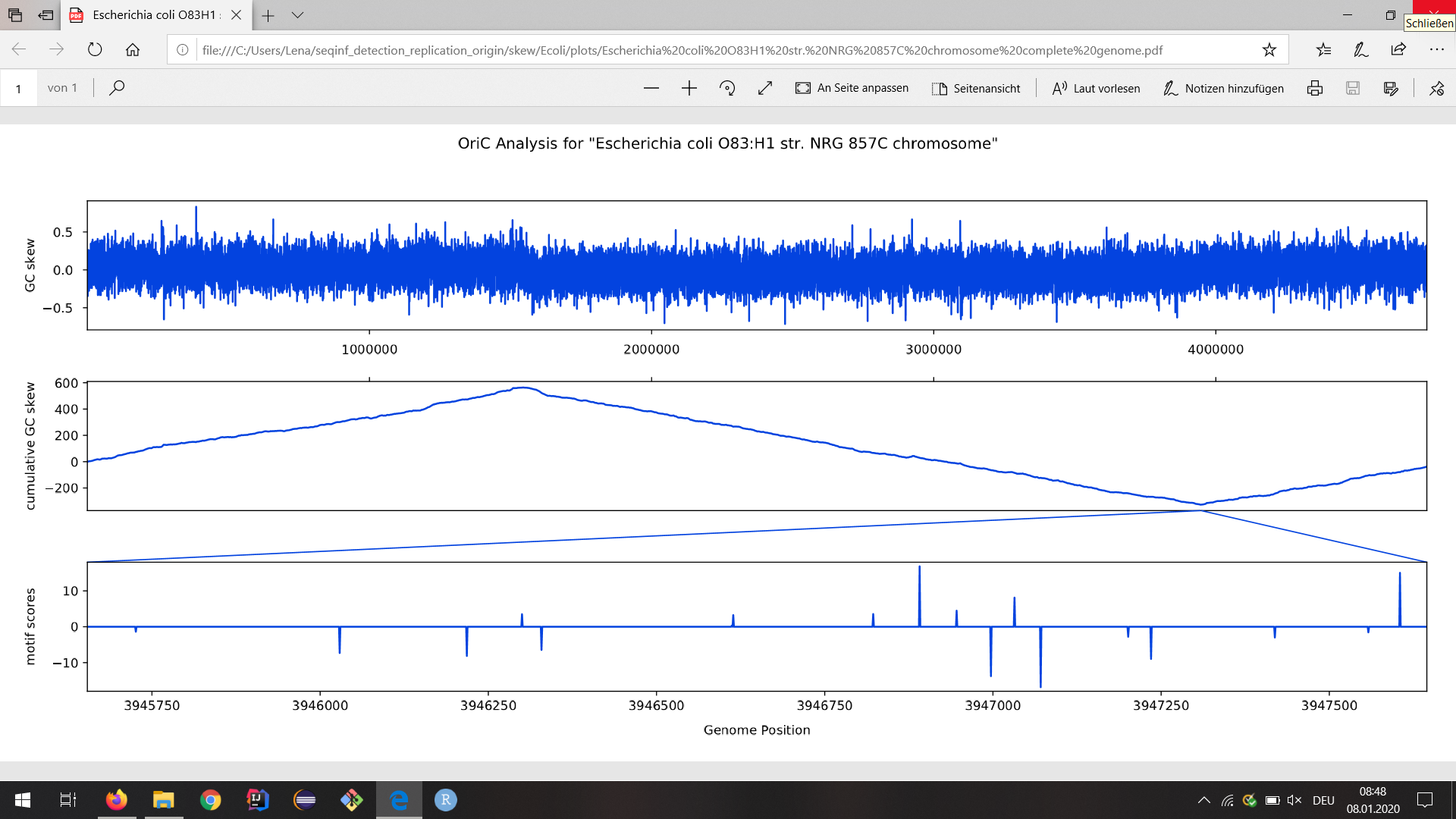
The DnaA motif computed for all bacterial reference genomes at the skew minimum (Figure 2) is not representative for all bacterias. If the skew minimum does not coincide the OriC of the Organism or the motif differed too much from the others it was not taken into account for the DnaA motif. But it delivered a more refined and more precise pattern pattern sequence to search for DnaA motifs than the consensus sequence.

The evaluation of the control strains yielded differences of the minimum to the actual oriC position ranging from 244 to 25634 nucleotides (Table 2).

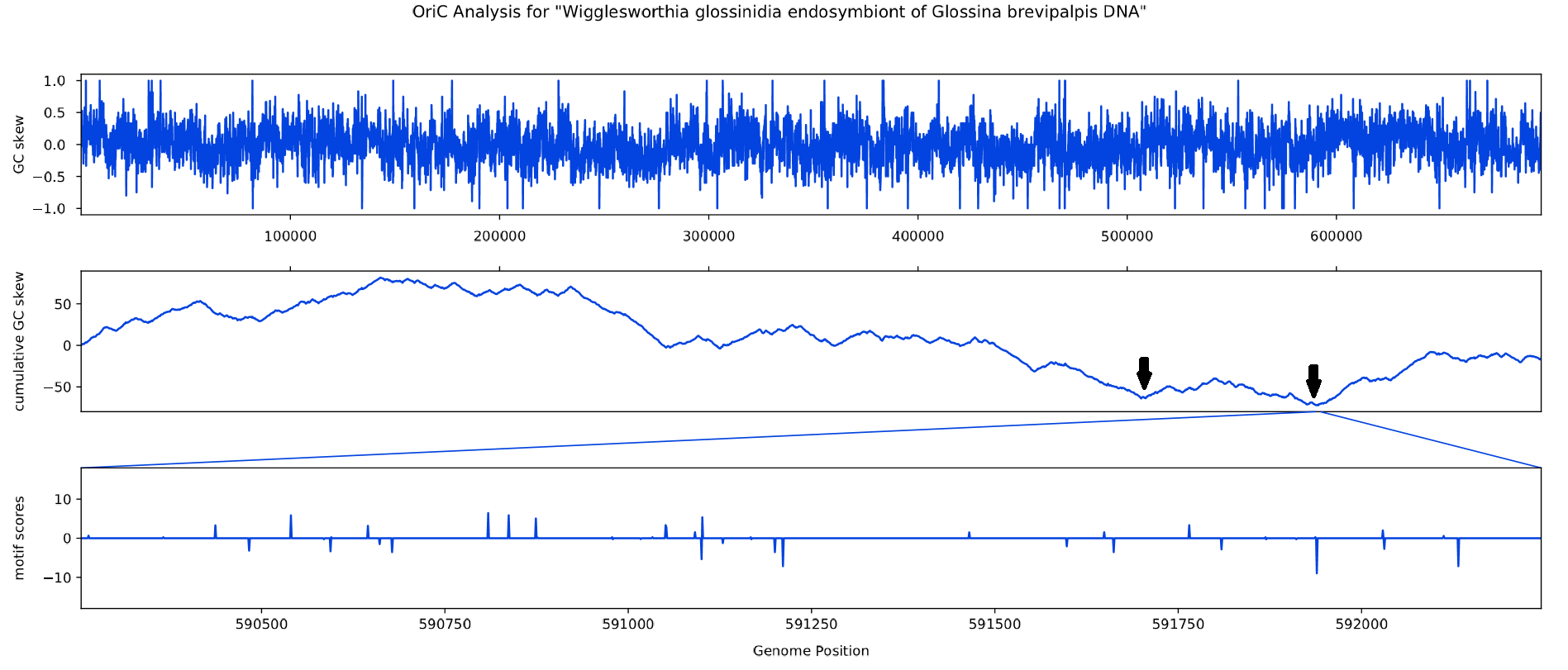
The determined differences to the DoriC database are shown in Table 3 for all families and strains.

The determined minima for the test genomes (*E.coli, V. cholerae, S. enterica, T. petrophila*) differed on average by 132 nucleotides to the values given in DoriC. The average difference to the oriC start position was 990 nucleotides and the determined motif number differed by four.

The GC skew of Wigglesworthia glossinidia shows quite an unclear pattern, however two minimum positions are visible in the plot (Figure 3). In comparison to DoriC, the minimum position differed by 467 and the oriC start by 106176 nucleotides (Table 3). No DnaA motifs were detected.



**Figure 2**: Plot example for *Escherichia coli O83:H1* shows GC skew (top), cumulative GC skew (middle) and the profile alignment motif scores (bottom).



**Figure 3:** Plot for Wigglesworthia glossinidia. Arrows indicate two potential minima.

**Table 2:** Control strains with known OriC position. Difference **𝞓** of determined minimum and OriC position were analysed with respect to genome length.

|  |  |  |  |
| --- | --- | --- | --- |
| **Strain** | **Annotated OriStart Position** | **Minimum** | **𝞓** |
| Escherichia coli str. K-12 substr. MG1655 | 3925744 | 3925500 | 244 |
| Sinorhizobium meliloti 1021 chromosome | 1 | 3628500 | 25634 |
| Chlamydia trachomatis D/UW-3/CX chromosome | 719988 | 720300 | 312 |

**Table 3:** 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. Motifs** | **𝞓 Number of Motifs** |
| **Enterobacteriaceae** | 186 ± 207 | 600 ± 645 | 2 ± 1 | 2 ± 1 | 2 |
| **Vibrionaceae** | 126 ± 176 | 103 ± 1 | 2 ± 1 | 2 ± 1 | 2 |
| **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 | 3073 ± 3336 | 1 ± 1 | 1 ± 1 | 5 |
| **Thermotoga sp.** | 70 ± 26 | 237 ± 185 | 0 ± 1 | 1 ± 1 | 7 |
| **Wigglesworthia glossinidia** | 467 | 106176 | 0 | 0 | 0 |

# Discussion

The detected GC minima in the control strains *Escherichia coli str. K-12* and *Chlamydia trachomatis* only show relatively small deviations to the actual origin of replication. For *Sinorhizobium meliloti* a larger deviation was observed, however this can be explained by the three control strains belonging to different bacterial families. Similar variations can also be observed in the different analysed families *Vibrionaceae* and *Enterobacteriaceae*.

In general, the obtained minimum positions of the bacterial families match with the respective test genomes. Furthermore, the minima detected in this project only show relatively small deviations to the ones given in DoriC as well as to the actual oriC start positions. This can be explained by the use of the GC skew, which is not as sensitive as other methods and highly depends on the chosen window size (Touchon and Rocha, 2008). Furthermore, DoriC uses the Z-curve method to determine the oriCs, which considers various nucleotide disparity curves for the detection of the oriC (Gao and Zhang, 2007) and hence also might lead to different results. Applying the Z-curve method could hence improve the results of this project.

Two potential minima were found for *W. glossinidia* which suggests multiple origins of replication. Similar observations were made by Xia (2012) who also detected two minima. Furthermore,, no DnaA motifs were detected in *W. glossinidia*. This was also observed in a study by Akman *et al*. (2002). Replication independent of the oriC can be observed in challenging physiological or genetic conditions, which applies to *W. glossinidia* as an intracellular symbiotic microorganism (Akman *et al*., 2002). Multiple origins of replication are often found in archeal genomes (Xia, 2002). This can be supported by the findings of Robinson *et al*. (2004) who detected two replications origins in the archaeon *Sulfolobus solfataricus*.

# Literature

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., and Aksoy, S.

(2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nature genetics*, **32**(3), 402.

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.

Gao, F. and Zhang, C. T. (2007) DoriC: a database of oriC regions in bacterial genomes.

*Bioinformatics*, **23**(14), 1866-1867.

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.

Mackiewicz, P., Zakrzewska-Czerwińska, J., Zawilak, A., Dudek, M. R., and Cebrat, S. (2004)

Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic*

*acids research*, **32**(13), 3781-3791.

Robinson, N. P., Dionne, I., Lundgren, M., Marsh, V. L., Bernander, R., and Bell, S. D. (2004)

Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. *Cell*, **116**(1), 25-38.

Rogan, P. K., Faux, B. M., and Schneider, T. D. (1998) Information analysis of human splice

site mutations. *Human mutation*, **12**(3), 153-171.

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.

Xia, X. (2012) DNA replication and strand asymmetry in prokaryotic and mitochondrial

genomes. *Current Genomics*, **13**(1), 16-27.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., ... & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic acids research, 37(suppl\_2), W202-W208.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome research, 14(6), 1188-1190.

**Appendix**

**KNOWN ORICS**

NC\_000913.3, Escherichia coli str. K-12 substr. MG1655

NC\_003047.1, Sinorhizobium meliloti 1021 chromosome

NC\_000117.1, Chlamydia trachomatis D/UW-3/CX chromosome

**FAMILIES**

Enterobacteriaceae:

NC\_000913.3 Escherichia coli str. K-12 substr. MG1655

NC\_002695.2 Escherichia coli O157:H7 str. Sakai DNA

NC\_011750.1 Escherichia coli IAI39 chromosome

NC\_017634.1 Escherichia coli O83:H1 str. NRG 857C chromosome

NC\_018658.1 Escherichia coli O104:H4 str. 2011C-3493 chromosome

NC\_004337.2 Shigella flexneri 2a str. 301 chromosome

NC\_007606.1 Shigella dysenteriae Sd197 chromosome

NC\_003197.2 Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 NC\_003198.1 Salmonella enterica subsp. enterica serovar Typhi str. CT18

NC\_014121.1 Enterobacter cloacae subsp. cloacae ATCC 13047 chromosome NC\_015663.1 Enterobacter aerogenes KCTC 2190 chromosome

NC\_016845.1 Klebsiella pneumoniae subsp. pneumoniae HS11286 chromosome

Vibrionaceae:

NC\_002505.1 Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I

NC\_004603.1 Vibrio parahaemolyticus RIMD 2210633 chromosome 1

**TEST GENOMES**

Escherichia coli:

NC\_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome

NC\_011750.1 Escherichia coli IAI39 chromosome, complete genome

NC\_017634.1 Escherichia coli O83:H1 str. NRG 857C chromosome, complete genome NC\_018658.1 Escherichia coli O104:H4 str. 2011C-3493 chromosome, complete genome NC\_002695.2 Escherichia coli O157:H7 str. Sakai DNA, complete genome

NC\_004431.1 Escherichia coli CFT073, complete genome

NC\_007946.1 Escherichia coli UTI89, complete genome

NC\_008253.1 Escherichia coli 536, complete genome

NC\_008563.1 Escherichia coli APEC O1, complete genome

NC\_009800.1 Escherichia coli HS, complete genome

Vibrio cholerae:

NZ\_CP010811 Vibrio cholerae strain 1154-74, complete genome

NZ\_CP010812 Vibrio cholerae strain 10432-62, complete genome

NC\_009456.1 Vibrio cholerae O395 chromosome 1

NC\_009457.1 Vibrio cholerae O395 chromosome 2

NC\_012578.1 Vibrio cholerae M66-2 chromosome 1

NC\_012580.1 Vibrio cholerae M66-2 chromosome 2

NC\_012668.1 Vibrio cholerae MJ-1236 chromosome 1

NC\_012667.1 Vibrio cholerae MJ-1236 chromosome 2

NC\_016445.1 Vibrio cholerae O1 str. 2010EL-1786 chromosome chromosome 1 NC\_016446.1 Vibrio cholerae O1 str. 2010EL-1786 chromosome chromosome 2

Salmonella enterica:

NC\_003198.1 Salmonella enterica subsp. enterica serovar Typhi str. CT18

NC\_006511.1 Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150 NC\_010102.1 Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7 NC\_016854.1 Salmonella enterica subsp. enterica serovar Typhimurium str. D23580 NC\_016810.1 Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344 NC\_011080.1 Salmonella enterica subsp. enterica serovar Newport str. SL254

NC\_011083.1 Salmonella enterica subsp. enterica serovar Heidelberg str. SL476 NC\_011294.1 Salmonella enterica subsp. enterica serovar Enteritidis str. P125109 NZ\_CP007245 Salmonella enterica subsp. enterica serovar Enteritidis str. EC20120008 NZ\_CP015574 Salmonella enterica strain FORC\_038

Thermotoga family:

NC\_009486.1 Thermotoga petrophila RKU-1

NC\_000853.1 Thermotoga maritima MSB8 NC\_009828.1 Pseudothermotoga lettingae TMO NC\_010483.1 Thermotoga sp. RQ2

NC\_011978.1 Thermotoga neapolitana DSM 4359

NC\_013642.1 Thermotoga naphthophila RKU-10

NZ\_CP003408 Thermotoga sp. 2812B

NZ\_CP007633 Thermotoga sp. RQ7

NZ\_CP010967 Thermotoga maritima strain Tma200 NZ\_CP011108 Thermotoga maritima strain Tma100

Wigglesworthia glossinidia:

NC\_004344.2 Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis