Project 14: Detection of the Replication Origin in Bacterial Genomes

Annalena Reuß, Antonia Schuster, Kilian Maidhof, Tobias Gresser
Supervisor: Xixi Chen

Abstract

With the increased use of sequencing data, the detection of the origin of replication (oriC) in bacteria can be used to gain insight into the growth dynamics of bacteria or to facilitate the analysis of certain motifs and features around its location.

This study presents a workflow to find the oriC by using two different indicators of its location. Since there is an asymmetry in the GC content of the leading and lagging strand, the cumulative GC skew can be used to calculate the approximate location of the oriC. The region around the oriC is also known to contain DnaA motifs which were used to predict the oriC more precisely.

The calculated minima of the GC skew in test organisms were shown to be fairly accurate. In the vicinity of those locations, DnaA motifs could be found, further confirming the detection of an oriC.

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(A/T)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. In this region DnaA motif occurrences are determined. Additionally, a species specific motif is computed for each of the four species.

Material and Methods

Data

Reference sequence genomes were downloaded from NCBI Assembly database (Geer *et al*., 2010). Table 1 shows the organisms used and their respective families. A list of all strains can be found in the appendix.

Table 1: Species and their respective families

Species	#sequences	Family	#sequences
Reference genomes (variety of species)	118	*various*	-
Known oriC: *Escherichia coli str. K-12 substr. MG1655* *Sinorhizobium meliloti 1021 chromosome* *Chlamydia trachomatis D/UW-3/CX chromosome*	1 each	*Enterobacteriaceae* *Rhizobiaceae* *Chlamydiaceae*	- - -
Escherichia coli	10	*Enterobacteriaceae*	12 (reference)
Salmonella enterica	10		
Vibrio cholerae	10	*Vibrionaceae*	2 (reference)
Thermotoga petrophila	10 (including other species)	*Thermotogaceae*	n/a

Workflow

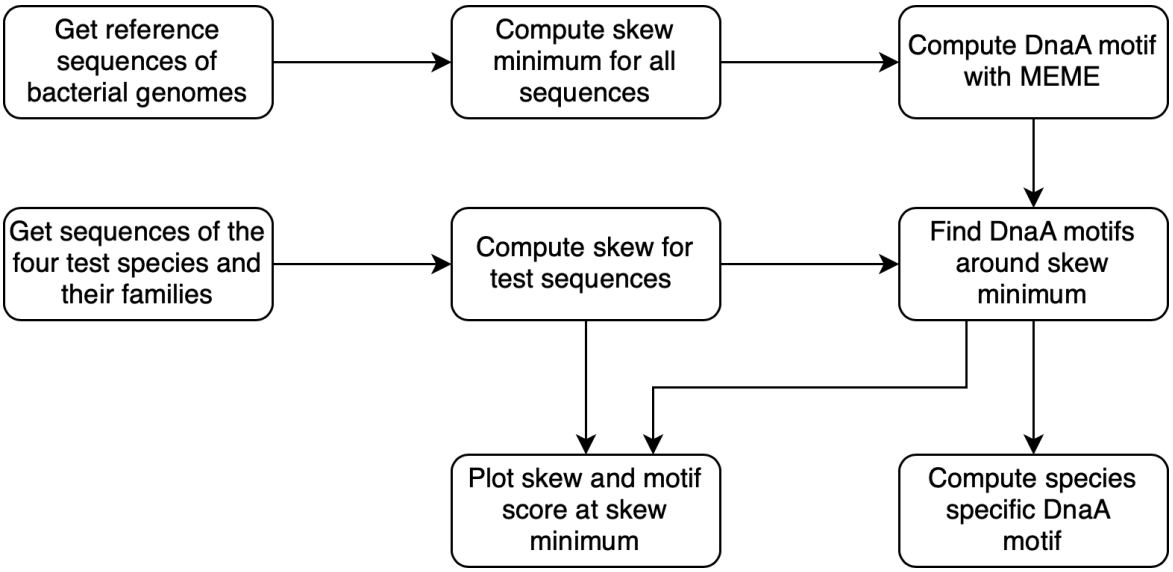


Figure 1: Overview of the project workflow.

Skew Diagrams

First the (cumulative) GC skew and its minimum were 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 enterica*, *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 (Gao and Zhang, 2007) with respect to the genome length. Three sequences with a known, experimentally determined oriC (Sibley *et al.*, 2006; Richardson *et al.*, 2016) 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.

Motif Finding

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 Expectation-maximazation 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 bacteria with MEME. For the execution the common DnaA consensus sequence *TTATCCACA* was given as starting point and a 4th order Markov model was 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 using an information theory based approach. 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 in Figure 3 (bottom).

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 underlying 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

The DnaA motif computed for all bacterial reference genomes at the skew minimum (Figure 2) is not representative for all bacteria. If the skew minimum does not coincide with 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 sequence to search for DnaA motifs than the consensus sequence.

The motifs computed for the different species and their families (Table 2) are in accordance with the consensus sequence. No motif was found for *Thermotoga species* and *W. glossinidia*.

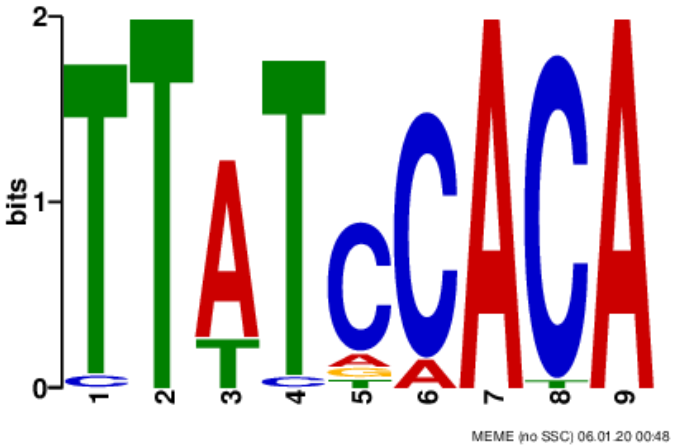
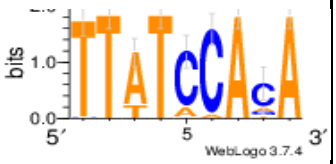
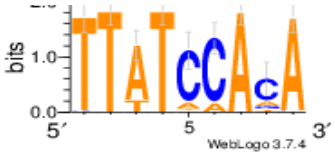
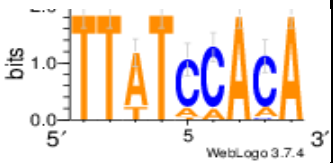
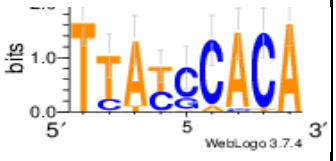
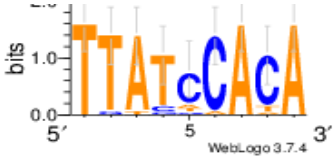


Figure 2: DnaA motif over all bacterial reference genomes computed by MEME.

Table 2: Weblogo motifs for species and families

Species	Motif	Family	Motif
Escherichia coli		**Enterobacteriaceae**	
Salmonella enterica			
Vibrio cholerae		**Vibrionaceae**	
Thermotoga sp.	No DnaA boxes above threshold	n/a	n/a
Wigglesworthia glossinidia	strongly deviating, motif was not found	n/a	n/a

The evaluation of the control strains yielded differences of the minimum to the actual oriC position ranging from 244 to 25634 nucleotides (Table 3).

The determined differences to the DoriC database are shown in Table 4 for all families and strains. The computed 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 4). In comparison to DoriC, the minimum position differed by 467 and the oriC start by 106176 nucleotides (Table 4). No DnaA motifs above the threshold were detected.

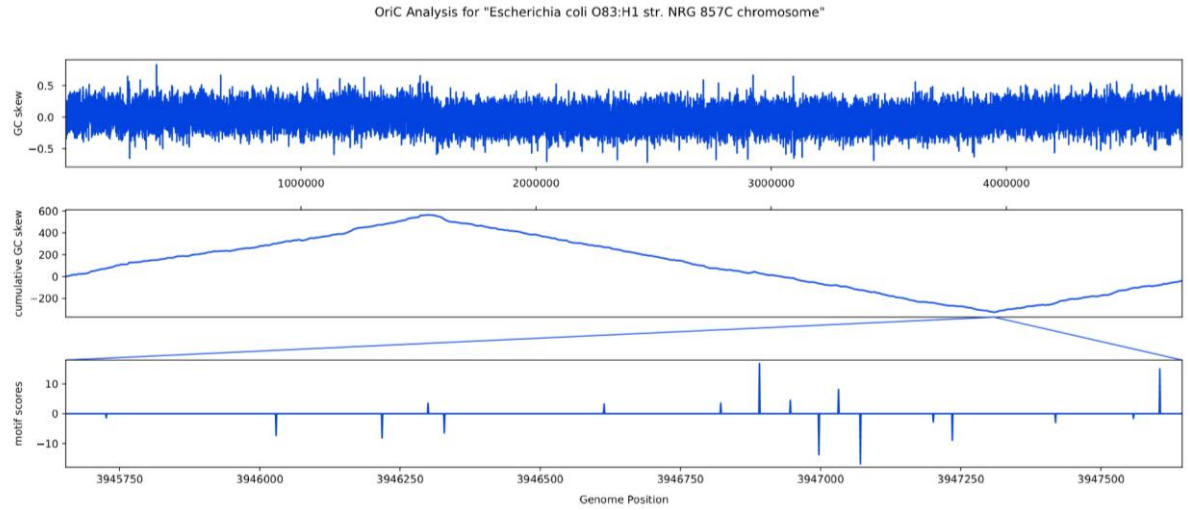


Figure 3: Plot example for *Escherichia coli O83:H1* shows GC skew (top), cumulative GC skew (middle) and the profile alignment motif scores (bottom).

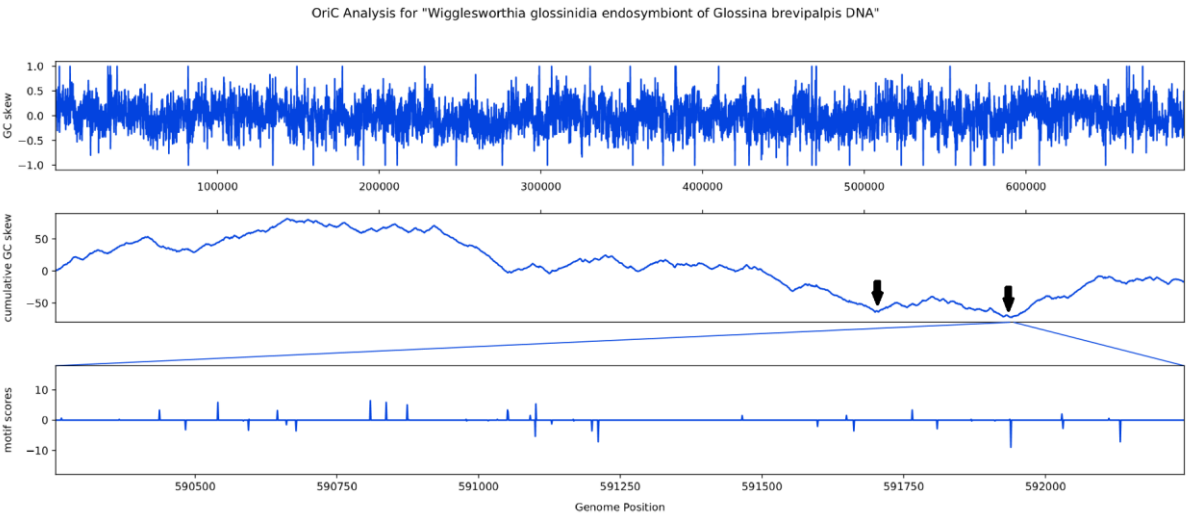


Figure 4: Plot for *Wigglesworthia glossinidia*. Arrows indicate two potential minima.

Table 3: 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 4: 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 for the 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*.

The computed DnaA motif logos are in accordance with the consensus sequence "*TT(A/T)TNCACA*" (Blaesing *et al*., 2017) and highly conserved throughout the species and bacterial families. For Thermotoga no DnaA boxes above the used threshold were present and hence no motif could be determined. This result can be confirmed by a study by Lopez *et al*. (2000), who discovered that instead of a 9 bp long motif, the 12 bp repeat "AAACCTACCACC" is present in Thermotoga.

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.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. and Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, **37**(suppl_2), W202-W208.

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.

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

Gao, F. and Zhang, C. T. (2007) DoriC: a database of oriC regions in bacterial genomes. *Bioinformatics*, **23**(14), 1866-1867.

Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W. and Bryant, S. H. (2009) The NCBI biosystems database. *Nucleic acids research*, **38**(suppl_1), D492-D49.

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.

Lopez, P., Forterre, P., le Guyader, H., and Philippe, H. (2000) Origin of replication of Thermotoga maritima. *Trends in Genetics*, **16**(2), 59-60.

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.

Richardson, T. T., Harran, O., and Murray, H. (2016) The bacterial DnaA-trio replication origin element specifies single-stranded DNA initiator binding. *Nature*, **534**(7607), 412.

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.

Sibley, C. D., MacLellan, S. R., and Finan, T. (2006) The Sinorhizobium meliloti chromosomal origin of replication. *Microbiology*, ***152***(2), 443-455.

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.

Appendix

REFERENCES

GCF_000011365.1 ASM1136v1
GCF_000196115.1 ASM19611v1
GCF_000007805.1 ASM780v1
GCF_000011065.1 ASM1106v1
GCF_000007825.1 ASM782v1
GCF_000025565.1 ASM2556v1
GCF_000215745.1 ASM21574v1
GCF_000018865.1 ASM1886v1
GCF_000007145.1 ASM714v1
GCF_000069185.1 ASM6918v1
GCF_000146165.2 ASM14616v2
GCF_000006945.2 ASM694v2
GCF_000014805.1 ASM1480v1
GCF_000011385.1 ASM1138v1
GCF_000195955.2 ASM19595v2
GCF_000013085.1 ASM1308v1
GCF_000009045.1 ASM904v1
GCF_000008765.1 ASM876v1
GCF_000008625.1 ASM862v1
GCF_000011445.1 ASM1144v1
GCF_000008685.2 ASM868v2
GCF_000008305.1 ASM830v1
GCF_000196835.1 ASM19683v1
GCF_000203835.1 ASM20383v1
GCF_000015005.1 ASM1500v1
GCF_000185905.1 ASM18590v1
GCF_000006765.1 ASM676v1
GCF_000007565.2 ASM756v2
GCF_000012245.1 ASM1224v1
GCF_000008865.2 ASM886v2
GCF_000240185.1 ASM24018v2
GCF_000009925.1 ASM992v1
GCF_000299455.1 ASM29945v1
GCF_000318015.1 ASM31801v1
GCF_000008505.1 ASM850v1
GCF_000008165.1 ASM816v1
GCF_000007845.1 ASM784v1
GCF_000026345.1 ASM2634v1
GCF_000317935.1 ASM31793v1
GCF_000195995.1 ASM19599v1
GCF_000183345.1 ASM18334v1
GCF_000009065.1 ASM906v1
GCF_000005845.2 ASM584v2
GCF_000009345.1 ASM934v1
GCF_000007985.2 ASM798v2
GCF_000023405.1 ASM2340v1
GCF_000195755.1 ASM19575v1
GCF_000013045.1 ASM1304v1
GCF_000011705.1 ASM1170v1
GCF_000008485.1 ASM848v1
GCF_000011325.1 ASM1132v1
GCF_000008185.1 ASM818v1
GCF_000092025.1 ASM9202v1
GCF_000008565.1 ASM856v1
GCF_000007325.1 ASM732v1
GCF_000006985.1 ASM698v1
GCF_000020985.1 ASM2098v1
GCF_000008545.1 ASM854v1
GCF_000021645.1 ASM2164v1
GCF_000091545.1 ASM9154v1
GCF_000024905.1 ASM2490v1
GCF_000007925.1 ASM792v1
GCF_000006925.2 ASM692v2
GCF_000012005.1 ASM1200v1
GCF_000195835.2 ASM19583v2
GCF_000092565.1 ASM9256v1
GCF_000009205.2 ASM920v2
GCF_000195715.1 ASM19571v1
GCF_000011545.1 ASM1154v1
GCF_000022005.1 ASM2200v1
GCF_000006905.1 ASM690v1
GCF_000018545.1 ASM1854v1
GCF_000063585.1 ASM6358v1
GCF_000191145.1 ASM19114v1
GCF_000017145.1 ASM1714v1
GCF_000017045.1 ASM1704v1
GCF_000006965.1 ASM696v1
GCF_000203855.3 ASM20385v3
GCF_000196095.1 ASM19609v1
GCF_000195855.1 ASM19585v1
GCF_000007785.1 ASM778v1
GCF_000006745.1 ASM674v1
GCF_000196035.1 ASM19603v1
GCF_000011805.1 ASM1180v1
GCF_000014525.1 ASM1452v1
GCF_000064305.2 ASM6430v2
GCF_000013425.1 ASM1342v1
GCF_000223375.1 ASM22337v1

GCF_000174395.2 ASM17439v2
GCF_000013105.1 ASM1310v1
GCF_000011345.1 ASM1134v1
GCF_000007645.1 ASM764v1
GCF_000014205.1 ASM1420v1
GCF_000006865.1 ASM686v1
GCF_000008805.1 ASM880v1
GCF_000007525.1 ASM752v1
GCF_000165905.1 ASM16590v1
GCF_000007265.1 ASM726v1
GCF_000006845.1 ASM684v1
GCF_000027165.1 ASM2716v1
GCF_000026745.1 ASM2674v1
GCF_000007045.1 ASM704v1
GCF_000007465.2 ASM746v2
GCF_000007765.2 ASM776v2
GCF_000011985.1 ASM1198v1
GCF_000008985.1 ASM898v1
GCF_000006785.2 ASM678v2
GCF_000027305.1 ASM2730v1
GCF_000008925.1 ASM892v1
GCF_000008525.1 ASM852v1
GCF_000159155.2 ASM15915v2
GCF_000009085.1 ASM908v1
GCF_000008745.1 ASM874v1
GCF_000195735.1 ASM19573v1
GCF_000008725.1 ASM872v1
GCF_000068585.1 ASM6858v1
GCF_000027345.1 ASM2734v1
GCF_000009605.1 ASM960v1

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