Complete genome sequence of *Thermotoga* sp. strain RQ7

Zhaohui Xu1*, Rutika Puranik1, Junxi Hu1,2, Hui Xu1 and Dongmei Han1

**Abstract**

*Thermotoga* sp. strain RQ7 is a member of the family *Thermotogaceae* in the order *Thermotogales*. It is a Gram negative, hyperthermophilic, and strictly anaerobic bacterium. It grows on diverse simple and complex carbohydrates and can use protons as the final electron acceptor. Its complete genome is composed of a chromosome of 1,851,618 bp and a plasmid of 846 bp. The chromosome contains 1906 putative genes, including 1853 protein coding genes and 53 RNA genes. The genetic features pertaining to various lateral gene transfer mechanisms are analyzed. The genome carries a complete set of putative competence genes, 8 loci of CRISPRs, and a deletion of a well-conserved Type II R-M system.

**Keywords:** *Thermotoga*, *T*. sp. strain RQ7, Natural competence, CRISPR, Restriction-modification system, TneDI, CP007633

**Background**

*Thermotoga* species are a group of thermophilic or hyperthermophilic bacteria that can ferment a wide range of carbohydrates and produce hydrogen gas as one of the major final products [1, 2]. Their hydrogen yield from glucose can reach the theoretical maximum: 4 mol of H\(_2\) from each mole of glucose [2, 3], which makes them ideal candidates for biofuel production. Meanwhile, because their enzymes are thermostable by nature, they also hold great prospect in the biocatalyst sector. 16S rRNA gene sequence analyses place *Thermotoga* at a deep branch in the tree of life, and genomic studies also reveal extensive horizontal gene transfer events between *Thermotogales* and other groups, particularly *Archaea* and *Firmicutes* [4]. Controversy over the phylogenetic significance of *Thermotoga* has triggered a prolonged debate on the concepts of species and biogeography, etc. [5].

We have been interested in the genetics of *Thermotoga* over the years and have developed the earliest set of tools to genetically modify these bacteria [6–8]. Strain RQ7 plays an essential role in these studies. This strain possesses the smallest known plasmid, pRQ7 (846 bp) [9], that is absent from most *Thermotoga* strains and serves as the base vector for all *Thermotoga*-E. coli shuttle vectors developed so far. *T*. sp. strain RQ7 is also the first *Thermotoga* strain in which natural competence was discovered [7]. To gain insights into the genetic and genomic features of the strain and to facilitate the continuing effort on developing genetic tools for *Thermotoga*, we set out to sequence the whole genome of *T*. sp. strain RQ7.

**Organism information**

**Classification and features**

*T*. sp. strain RQ7 was isolated from marine sediments of Ribeira Quente, Azores [1]. The strain is a member of the genus *Thermotoga*, the family *Thermotogaceae*, and the order *Thermotogales* (Table 1). Based on 16S rRNA gene sequences, the closest relative of *T*. sp. strain RQ7 is *T. neapolitana* DSM 4359, and these two strains cluster with *T. maritima* MSB8 and *T*. sp. strain RQ2 (Fig. 1). The results are in agreement with previous reports [10].

Like its close relatives *T. neapolitana* DSM 4359 and *T. maritima* MSB8, *T*. sp. strain RQ7 is a strict anaerobe, growing best around 80 °C, utilizing both simple and complex sugars, and producing hydrogen gas. These bacteria grow in both rich and defined media, are free living and non-pathogenic to humans, animals, or plants. Cells are rod-shaped, about 0.5 to 2 μm in length and 0.4 to 0.5 μm in diameter (Fig. 2). The most distinctive feature of *Thermotoga* cells is the “toga” structure that
Table 1 Classification and general features of *Thermotoga* sp. strain RQ7 according to the MIGS recommendations [36]

| MIGS ID    | Property        | Term                  | Evidence code       |
|------------|-----------------|-----------------------|---------------------|
|            | Classification  | Domain *Bacteria*     | TAS [37]            |
|            | Class           | Phylum *Thermotogae*   | TAS [38, 39]        |
|            | Order           | Class *Thermotogae*    | TAS [39, 40]        |
|            | Family          | Order *Thermotogales*  | TAS [39, 41]        |
|            | Genus           | Family *Thermotogae*   | TAS [39, 42]        |
|            | Species         | Genus *Thermotoga*     | TAS [1, 43, 44]     |
|            | strain: RQ7     | Species *T. neapolitana* | IGC, TSA [45, 46] |
| Gram stain | Negative        |                       | TAS [1]             |
| Cell shape | Rod             |                       | IDA, TAS [1]        |
| Motility   | Motile          |                       | IDA, TAS [1]        |
| Sporulation| Not reported    |                       |                     |
| Temperature range | 55–90 °C        |                       | TAS [1]             |
| Optimum temperature | Around 80 °C    |                       | TAS [1]             |
| pH range; Optimum | 5.5–9; 6.5      |                       | IDA, TAS [1]        |
| Carbon source | Mono- and polysaccharides | IDA, TAS [1, 47, 48] |

MIGS-6 Habitat Geothermally heated sediments TAS [1]

MIGS-6.3 Salinity 0.25–3.75% NaCl (w/v) IDA, TAS [1]

MIGS-22 Oxygen requirement Anaerobic IDA, TAS [1]

MIGS-15 Biotic relationship Free-living IDA, TAS [1]

MIGS-14 Pathogenicity Non-pathogen IDA, TAS [1]

MIGS-4 Geographic location Azores, Sao Miguel, Ribeira Quente TAS [1]

MIGS-5 Sample collection 1985 NAS

MIGS-4.1 Latitude Not reported NAS

MIGS-4.2 Longitude Not reported NAS

MIGS-4.4 Altitude About sea level NAS

*Evidence codes - IDA Inferred from Direct Assay, TAS Traceable Author Statement (i.e., a direct report exists in the literature), NAS Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence), IGC Inferred from Genomic Content (i.e., average nucleotide identity, syntenic regions). These evidence codes are from the Gene Ontology project [49]*

Fig. 1 Phylogenetic tree showing the position of *T*. sp. strain RQ7 relative to other species within the order *Thermotogales*. Only species with complete genome sequences are included. The tree was built with 16S rRNA gene sequences, using the Neighbor-Joining method with MEGA7 [50]. *Fervidobacterium nodosum* serves as the outgroup.
balloons out from both ends of the rod [1, 11], an extension of their outer membrane [12].

**Genome sequencing information**

**Genome project history**
The project started in June 2011, and the genome was sequenced by BGI Americas (Cambridge, MA) using the Illumina technology. A total of 400 Mb of clean data were generated, which covered the genome more than 200 fold. The assembled scaffold covers 97.7% of the chromosome. PCR and Sanger sequencing were later used for gap filling. The assembly was finalized in February 2014, and the complete sequence was submitted to the GenBank in April 2014. The sequence was annotated with the NCBI Prokaryotic Genome Annotation Pipeline [13] and the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4) [14]. The project information is summarized in Table 2.

**Growth conditions and genomic DNA preparation**
*T. sp.* strain RQ7 was kindly provided by Drs. Harald Huber and Robert Huber at the University of Regensburg, Germany. It was cultivated in SVO medium [15] at 77 °C, and its genomic DNA was extracted with standard phenol extraction method [16]. Briefly, cells from 250 ml of overnight culture were collected by centrifugation and resuspended in 10 ml of STE solution (10 mM Tris-HCl, 1 mM EDTA, 100 mM NaCl, pH 8.0). SDS and proteinase K were added to a final concentration of 1% (w/v) and 20 μg/ml. The mixture was incubated at 50 °C for 6 h followed by the addition of an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v). After gentle mixing, the mixture was centrifuged at 12,000 g at 4 °C for 15 min. The upper aqueous layer was transferred to a clean tube and mixed with 1/10 volume of 3 M sodium acetate (pH 5.5) and 2 volumes of ice cold 95% (v/v) ethanol. The DNA was spooled out by a glass rod, washed with 70% (v/v) ethanol, air dried, dissolved in 2 ml of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) containing 20 μg/ml RNase A, and stored at −20 °C.

**Genome sequencing and assembly**
The genome of *T. sp.* strain RQ7 was mainly sequenced by BGI Americas using Illumina HiSeq 2000 sequencing platform. Three paired-end libraries, in size of 500, 2000, and 5000 bp, were constructed. The raw data were filtered by a quality control step and generated 400 Mb of clean data, which indicated a coverage of more than 200-fold. The reads were assembled by SOAPdenovo [17] and polished by SOAPaligner [18]. This resulted in a single scaffold of 1,822,593 bp that covered 97.7% of the genome and contained 28 gaps. The gap filling efforts included the integration of the current scaffold with contigs generated by the CLC Genomics Workbench [19] and a small amount of public sequences in GenBank. GapFish [20] was then used to solve a dozen ambiguous regions. Finally, PCR and primer walking were performed to close the remaining gaps, resulting a final assembly of 1,851,618 bp. The entire assembling process integrated wet lab methods with in silico approaches, and the programs used included public software.

---

**Table 2** Project information

| MIGS ID | Property   | Term                                    |
|---------|------------|-----------------------------------------|
| MIGS 31 | Finishing quality | Complete                              |
| MIGS-28 | Libraries used | Three Illumina paired-end libraries in sizes of 500, 2000, and 5000 bp |
| MIGS 29 | Sequencing platforms | Illumina and Sanger                    |
| MIGS 31.2 | Fold coverage | > 200x                                  |
| MIGS 30 | Assemblers   | SOAPdenovo [17], SOAPaligner [18], CLC Workbench 5.1 [19], and GapFish [20] |
| MIGS 32 | Gene calling method | GeneMarkS+ [51], Prodigal [52]         |
| Locus Tag |                     | TRQ7 in GenBank; Ga0077854 in JGI-IMG   |
| GenBank ID |               | CP007633, KF798180                       |
| GenBank Date of Release |   | February 4, 2015                         |
| GOLD ID |                | Gp0117593                               |
| BIOPROJECT  |            | PRJNA246218                             |
| MIGS 13 | Source Material Identifier | Personal culture collection (Dr. Harald Huber) |
| Project relevance | | Bioenergy, biotechnology, evolution |
SOAPdenovo and SOAPaligner [17, 18]), a commercial product (CLC Genomics Workbench [19]), and an in-house program GapFish [20]. Details of the assembling process are described in our previous report [20].

**Genome annotation**

The genome was independently annotated by two pipelines, the NCBI Prokaryotic Genome Annotation Pipeline [13] and the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4) [14]. Both pipelines combine a gene-calling algorithm with a similarity-based gene detection approach, even though the algorithms and databases they use are different. For example, PGAAP uses GeneMarkS+ for de novo gene prediction, while MGAP uses Prodigal. Consequently, the two pipelines produced slightly different annotation results. The analyses in this report took into consideration of the results from both pipelines and are assisted with manual curation.

**Genome properties**

The genome of *T*. sp. strain RQ7 is composed of a circular chromosome of 1,851,618 bp with a GC content of 47.05% and a single mini-plasmid of 846 bp with a GC percentage of 39.95 (Fig. 3; Table 3). The plasmid pRQ7 has been characterized [9] and sequenced [6, 21] before. According to the annotation of MGAP, the chromosome carries 1906 putative genes, of which, 1853 are protein coding genes and 53 are RNA genes (Table 4). Among all the genes that are assigned to a COG category (Table 5), a significant portion (~12%, 191 genes) are devoted to carbohydrate utilization, which is typical to *Thermotoga* strains and accords with their versatile use of carbon and energy sources.

**Table 3** Summary of genome: one chromosome and one plasmid

| Label | Size (bp) | Topology | INSDC identifier | RefSeq ID |
|-------|-----------|----------|-----------------|-----------|
| Chromosome | 1,851,618 | Circular | CP007633 | NZ_CP007633 |
| pRQ7 | 846 | Circular | KF798180 | NC_023152 |
Insights from the genome sequence

The chromosomal sequence of *T*. sp. strain RQ7 was compared to those of *T*. maritima MSB8, *T*. neapolitana DSM 4359, and *T*. sp. strain RQ2, with emphases on the genetic elements that have the highest impacts on genetic engineering attempts, such as natural competence genes, CRISPRs, and R-M systems.

Full genome comparison

The alignment of the complete genomic sequence of the four *Thermotoga* strains (Fig. 4) revealed high levels of synteny among their genomes, particularly within the pairs of *T*. sp. strain RQ7–*T*. neapolitana DSM 4359 and *T*. sp. strain RQ2–*T*. maritima MSB8. This is in agreement with their placements in the phylogenetic tree (Fig. 1). The average nucleotide identity between *T*. sp. strain RQ7 and the type strain *T*. neapolitana DSM 4359 is 98.49%, which is higher than the conventional cutoff of 95% for species delineation [22]. Therefore, *T*. sp. strain RQ7 should be considered as a strain of *T*. neapolitana, same as *T*. sp. strain RQ2 to *T*. maritima [23].

| Table 4 Genome statistics according to the MGAP pipeline annotation (chromosome only) |
|-----------------------------------------------|-------------|-----------------|
| Attribute                                  | Value       | % of total      |
| Genome size (bp)                            | 1,851,618   | 100.00          |
| DNA coding (bp)                             | 1,768,561   | 95.51           |
| DNA G + C (bp)                              | 871,250     | 47.05           |
| DNA scaffolds                               | 1           |                 |
| Total genes                                 | 1906        | 100.00          |
| Protein coding genes                        | 1853        | 97.22           |
| RNA genes                                   | 53          | 2.78            |
| Pseudo genes                                | --          | --              |
| Genes in internal clusters                  | 110         | 5.77            |
| Genes with function prediction              | 1522        | 79.85           |
| Genes assigned to COGs                      | 1453        | 76.23           |
| Genes with Pfam domains                     | 1629        | 85.47           |
| Genes with signal peptides                  | 35          | 1.84            |
| Genes with transmembrane helices            | 462         | 24.24           |
| CRISPR repeats                              | 8           |                 |

The total is based on the total number of protein coding genes in the genome as annotated by MGAP v.4 [14]

Table 5 Number of genes associated with general COG functional categories

| Code | Value | %age  | Description                                               |
|------|-------|-------|-----------------------------------------------------------|
| J    | 165   | 10.17 | Translation, ribosomal structure and biogenesis          |
| A    | –     | –     | RNA processing and modification                           |
| K    | 75    | 4.62  | Transcription                                             |
| L    | 53    | 3.27  | Replication, recombination and repair                     |
| B    | 1     | 0.06  | Chromatin structure and dynamics                          |
| D    | 19    | 1.17  | Cell cycle control, Cell division, chromosome partitioning|
| V    | 34    | 2.09  | Defense mechanisms                                        |
| T    | 57    | 3.51  | Signal transduction mechanisms                            |
| M    | 74    | 4.56  | Cell wall/membrane biogenesis                            |
| N    | 55    | 3.39  | Cell motility                                            |
| U    | 21    | 1.29  | Intracellular trafficking and secretion                  |
| O    | 66    | 4.07  | Posttranslational modification, protein turnover, chaperones|
| C    | 104   | 6.41  | Energy production and conversion                          |
| G    | 191   | 11.77 | Carbohydrate transport and metabolism                     |
| E    | 169   | 10.41 | Amino acid transport and metabolism                       |
| F    | 65    | 4     | Nucleotide transport and metabolism                       |
| H    | 73    | 4.5   | Coenzyme transport and metabolism                         |
| I    | 42    | 2.59  | Lipid transport and metabolism                            |
| P    | 103   | 6.35  | Inorganic ion transport and metabolism                    |
| Q    | 18    | 1.11  | Secondary metabolites biosynthesis, transport and catabolism|
| R    | 156   | 9.61  | General function prediction only                          |
| S    | 75    | 4.62  | Function unknown                                          |
| –    | 453   | 23.77 | Not in COGs                                              |

The total is based on the total number of protein coding genes in the genome as annotated by MGAP v.4 [14]
A detailed comparison of *T. sp.* strain RQ7 and *T. neapolitana* DSM 4359 found 100 genes belonging only to the former and 120 genes only to the latter. Some of these genes became unique because their counterparts in the other genome have mutated to a pseudogene. However, many of the unique genes seem to have been acquired via recent lateral gene transfer events. The putative functions of these genes are mainly associated to transportation and utilization of carbohydrates and nucleotides. The most notable gene clusters include TRQ7_01555-01655 (nucleotide metabolism), TRQ7_02675-02725 (carbohydrate metabolism), TRQ7_03440-03490 (arabinose metabolism), CTN_0026-0038 (synthesis of antibiotics), CTN_0236-0245 (carbohydrate metabolism), CTN_0355-0373 (ribose metabolism), CTN_1540-1554 (carbohydrate metabolism), and CTN_1602-1627 (ribose metabolism). Follow-up functional genomics studies are needed to validate the predictions on these gene functions and metabolic pathways.

**Natural competence**

*Thermotoga* species are known to undergo lateral gene transfer events. One of the ways this could happen is via natural transformation. Natural competence has been established in *T. sp.* strain RQ7 [7] and *T. sp.* strain RQ2 [8]. Using experimentally characterized competence genes as references, we are able to identify the genes that might play a role in natural competence in *Thermotoga* (Table 6). These genes are widely spread among bacterial genomes, and none of them are clustered into operons. This might imply a primitive form of natural competence that is shared by most, if not all, bacteria. Perhaps, most free-living bacteria are more or less naturally competent during some points of their life. The trick is to identify the right conditions under which the natural competence will be allowed to develop.

**CRISPRs**

CRISPRs provide prokaryotes a form of adaptive immunity against invading phages and plasmids in a sequence specific manner [24, 25]. The system utilizes non-coding CRISPR RNA and a set of CRISPR-associated proteins to target invading nucleic acid, including both DNA and RNA. CRISPRs have been reported to prevent natural transformation [26, 27]. They have been noticed before in *Thermotoga* and are credited for large scale chromosomal recombination events in these species [28, 29]. NCBI’s PGAAP pipeline identified 6 loci of CRISPR arrays in *T. sp.* strain RQ7, whereas JGI-IMG’s MGAP pipeline and a manual analysis using CRISPRFinder [30] recognized a total of 8 loci (Table 7). Among these eight CRISPR loci, #1 and #3 are the ones not considered by PGAAP. Two clusters of *cas* genes are also found. The *cas6-cas2* cassette is sandwiched between loci #3...
and #4, and the cas6-csm1 cassette is located 2285 bp upstream of locus #3 (Fig. 5, Table 7).

Although analysis with CRISPRFinder revealed the same number of CRISPR loci in the four close relatives, i.e. T. sp. strain RQ7, T. neapolitana DSM 4359, T. maritima MSB8, and T. sp. strain RQ2, the total number of spacers they carry vary dramatically, as 95, 60, 106, and 129 spacers are found respectively. T. maritima MSB8 and T. sp. strain RQ2 also harbor RNA-targetting cmr genes in addition to DNA-targetting cas genes [31]. These differences may affect the efficiency of lateral gene transfer events among the strains.

### Table 6: Manually curated competence genes

| RQ7     | Gene name | Putative function                          | Tn  | Tm  | RQ2   |
|---------|-----------|--------------------------------------------|-----|-----|-------|
| DNA uptake and translocation |           |                                            |     |     |       |
| TRQ7_00110 | pilZ (Pa, Vc) | Type IV pilus biogenesis and twitching motility [54–56] | CTN_1670 | TM0905 | TRQ2_0022 |
| TRQ7_00455 | pilB (Pa, Vc) | Type II secretion system (T2SS), Type IV fimbrial assembly NTase [57–59] | CTN_1739 | TM0837 | TRQ2_0090 |
| TRQ7_01410 | pilQ (Nm, Tt) | Secretin, forms gated channel for extrusion of assembled pilin [60–62] | CTN_1450 | TM1117 | TRQ2_1699 |
| TRQ7_04530 | comEC (Bs) | Putative channel protein, Transports DNA across the cell membrane [63, 64] | CTN_0598 | TM_0094 | TRQ2_0853 |
| TRQ7_08710 | pilC (Ps, Ng) | Type II secretory pathway, component PulF / Type IV fimbrial assembly protein [65–67] | CTN_0883 | TM1696 | TRQ2_1138 |
| TRQ7_05855 | pilG (Vv, Ng) | Type IV prepilin peptidase, processes N-terminal leader peptides for prepilins [68–69] | CTN_0965 | TM1775 | TRQ2_1049 |
| TRQ7_06260 | comEC (Bs) | Putative channel protein, Transports DNA across the cell membrane [63, 64] | CTN_1168 | TM1584 | TRQ2_1247 |
| TRQ7_07315 | comF (Hi) | Phosphoribosyltransferase [70, 71] | CTN_0965 | TM1775 | TRQ2_1049 |
| TRQ7_07650 | pilT (Ng) | Motility protein [72] | CTN_1229 | TM1362 | TRQ2_1467 |
| TRQ7_07980 | pilE (Nm, Tt) | Type IV pilin; major structural component of Type IV pilus [73, 74] | CTN_1301 | TM1271 | TRQ2_1548 |
| TRQ7_09065 | comEA (Bs) | High affinity DNA-binding periplasmic protein [75–78] | CTN_1515 | TM1052 | TRQ2_1756 |

Post-translocation:

| RQ7     | Gene name | Putative function                          | Tn  | Tm  | RQ2   |
|---------|-----------|--------------------------------------------|-----|-----|-------|
| TRQ7_02260 | comM (Hi) | Promotes the recombination of the donor DNA into the chromosome [79] | CTN_0158 | TM0513 | TRQ2_0424 |
| TRQ7_03645 | dprA (Hi) | DNA protecting protein [80, 81] | CTN_0436 | TM0250 | TRQ2_0698 |

A Gene names are given after the experimentally characterized genes of the species in parentheses. Pa Pseudomonas aeruginosa, Vc Vibrio cholerae, Nm Neisseria meningitidis, Tt Thermus thermophilus, Ps Pseudomonas stutzeri, Nc Neisseria gonorrhoeae, Vv Vibrio vulnificus, Bs Bacillus subtilis, Hi influenza, RQ7 T. sp. strain RQ7, Tn T. neapolitana DSM 4359, Tm T. maritima MSB8, RQ2 T. sp. strain RQ2.

### Table 7: Summary of CRISPR loci in T. sp. strain RQ7

| Locus | Repeats | Coordinates | No. of spacers | Cas genes |
|-------|---------|-------------|----------------|-----------|
| 1     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA CTGTTAACATCTCTCTATTAGATTGATTGAAACCA | 553,849-554,014 | 2  | No  |
| 2     | TTTCCATATACCTCTTAAAGGATTGAAACCA GATTTACCATCTCTTAAAGGATTGAAACCA | 594,500-594,927 | 6  | No  |
| 3     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 975,191-975,420 | 3  | Yes |
| 4     | GTTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 983,596-986,955 | 51 | Yes |
| 5     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 1,011,410-1,012,101 | 10 | No  |
| 6     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 1,090,312-1,090,681 | 6  | No  |
| 7     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 1,233,649-1,233,878 | 3  | No  |
| 8     | GTTTCATACCTCTCTATTAGATTGATTGAAACCA GATTTACCATCTCTCTATTAGATTGATTGAAACCA | 1,242,811-1,243,509 | 10 | No  |

aCoordinates as documented in JGI-IMG. The start coordinates in GenBank are 20 bp smaller because the chromosome is linearized at a site 20 bp downstream of what JGI-IMG uses.
Type II R-M system TneDI

R-M systems are other defense mechanisms that prokaryotes have developed to protect the integrity of their genetic materials. The Type II R-M system TneDI has been characterized in *T. neapolitana* DSM 4359 and overexpressed in *Escherichia coli* [32, 33]. The nuclease R.TneDI cleaves at the center of the recognition site (CG↓CG), and the methylase M.TneDI modifies one of the cytosines. The TneDI system has been found in many members of the Thermotogaceae family, including *T. maritima* MSB8 and *T. sp. strain RQ2* [32]. However, it is absent from *T. sp. strain RQ7*, although the neighborhood is still highly conserved (Fig. 6). To exclude the possibility of an assembling error, primers spanning the region in question were designed, and the PCR results confirmed the deletion (Fig. 7). The absence of the TneDI system makes the DNA of *T. sp. strain RQ7* susceptible to R.TneDI, and in vitro treatment with M.TneDI provides complete protection to its genomic DNA (Fig. 8).

M.TneDI has been predicted to be a m\(^4\)C methylase based on sequence analysis [32]. It has also been noticed that m\(^4\)C methylation is more common than m\(^5\)C in thermophiles, probably due to a reduced risk of deamination [34]. The speculation of M.TneDI being a m\(^4\)C methylase is further supported by the observation that the genomic DNA of TneDI-bearing species is still susceptible to BstUI (Fig. 9), which is an isoschizomer of R.TneDI and known to be blocked by m\(^5\)C methylation [35].
Conclusions

The genome of *T*. sp. strain RQ7 shares large regions of synteny with those of its close relatives, namely, *T. neapolitana* DSM 4359, *T. maritima* MSB8, and *T*. sp. strain RQ2. They all have a complete set of putative competence genes, although natural transformation has yet to be established in *T. neapolitana* DSM 4359 and *T. maritima* MSB8. The same number of CRISPR loci are found in all four genomes, even though the number of spacers vary. The most noticeable difference among the strains is the absence of the TneDI R-M system in *T*. sp. strain RQ7, which partially explains why this strain is more amenable to genetic modifications than others. In general, this work sheds light on the genetic features of *T*. sp. strain RQ7, promoting genetic and genomic studies of *Thermotoga* spp.

Abbreviations

Cas: CRISPR associated; CRISPR: Clustered regularly interspaced short palindromic repeats; R-M: Restriction-modification

Acknowledgements

We are grateful to Drs. Harald Huber and Robert Huber at the University of Regensburg, Germany, for kindly providing *T*. sp. strain RQ7.

Funding

This work was supported by the BGSU Commercialization Catalyst Award and the BGSU Building Strength Award to ZX. BGSU plays no role in designing or conducting the study, collecting or analysing the data, or writing the manuscript.

Authors’ contributions

ZX conceived and coordinated the study, participated in all aspects of data analysis and drafted the manuscript. RP participated in most parts of the work and helped in writing the manuscript. JX produced the SEM photo and the phylogenetic tree. HX contributed to the R-M study. DH initiated the sequencing project and the R-M study. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

1Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43403, USA. 2School of Life Sciences, Minnan Normal University, 36 Xuanqianzhi Street, Zhangzhou, Fujian 363000, China.
Received: 4 October 2016 Accepted: 21 September 2017
Published online: 11 October 2017

References
1. Huber R, Langworthy TA, Konig H, Thomm M, Woese CR, Sleyter UB, Stetter KO. Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 degrees C. Arch Microbiol. 1986;144(4):234–33.
2. Schroder C, Sellis M, Schoneit P. Glucose Fermentation to Acetate, CO2 and H2 in the Anaerobic Hyperthermophilic Eubacterium Thermotoga Maritima: Involvement of the Embden-Meyerhof Pathway. Arch Microbiol. 1994;161(6):460–70.
3. Takahata Y, Nishijsima M, Hoaki T, petrophila MTT. sp. nov. and Thermotoga naphthophila sp. nov., two hyperthermophilic eubacteria from the Kubiki oil reservoir in Niigata, Japan. Int J Syst Evol Microbiol. 2001;51:1901–9.
4. Zhaxybayeva O, Swoffers KS, Lapierre P, Fournier GP, Bickat DA, DeBot RT, Nelson KE, Niebo CL, Doolittle WF, Gogarten JP, et al. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci U S A. 2009;106(14):5865–70.
5. Niebo CL, Doolittle WF, Wightman W. Recombination in thermotoga: Implications for species concepts and biogeography. Genetics. 2006;172(2):579–66.
6. Han D, Norris SM, Xu Z. Construction and transformation of a Thermotoga-E. coli shuttle vector. BMC Biotechnol. 2012;12.
7. Han D, Xu H, Puranik R, Xu Z. Natural transformation of Thermotoga sp. strain RQ7. BMC Biotechnol. 2014;14:39.
8. Xu H, Han D, Xu Z. Expression of Heterologous Cellulases in Thermotoga sp. Strain RQ2. Biomed Res Int. 2015;2015:304523.
9. Hartott OT, Huber R, Stetter KO, Sets PW, Noll KM. A Cryptic Miniplasmid from the Hyperthermophilic Bacterium Thermotoga Sp Strain Rq7. J Bacteriol. 1994;176(9):2759–62.
10. Frock AD, Notey JS, Kelly RM. The genus Thermotoga: recent developments. J Bacteriol. 1997;179(22):7161–71.
11. Windberger E, Huber R, Trincone A, Fricke H, Stetter KO. Thermotoga thermarum sp. nov. and Thermotoga nepalitana occuring in African continental solfataric springs. Arch Microbiol. 1989;151(6):596–12.
12. Rachel R, Engel AM, Huber R, Stetter KO, Baumeister W. A Porin-Type Protein Is the Main Constituent of the Cell-Envelope of the Anaerobic Eubacterium Thermotoga-Maritima. FEBS Lett. 1990;262(1):164–8.
13. Angiuoli SV, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira CD, Kyrpides N, Madupu R, Markowitz V, et al. Toward an online repository of Standard Operating Procedures (SOPs) for (metagenomic annotation). OMICS. 2008;12(2):137–41.
14. Huntemann M, Ivanova NN, Mavromatis K, Tripp HW, Geer PD, Witherspoon S, et al. Genomescale comparative analysis of extremely thermophilic bacteria. Int J Syst Evol Microbiol. 2007;57(Pt 7):1881–91.
15. Han D, Norris SM, Xu Z. Thermotoga maritima, a model hyperthermophilic species. J Bacteriol. 2005;187(14):4935–44.
16. DeBoy RT, Montgomery EF, Emerson JB, Nelson KE. Chromosome evolution in the Thermotogales: Large-scale inversions and strain diversification of CRSPr sequences. J Bacteriol. 2006;188(7):2364–74.
17. Gissi I, Vergnaud G, Pourcel C. CRSPrFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res. 2007;35(Web Server issue)W52–W7.
18. Haft DH, Selengut J, Montgomery EF, Nelson KE. A guild of 45 CRSPr-associated (Cas) protein families and multiple CRSPr/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. 2005;1(6):e60.
19. Xu Z, Han D, Cao J, Saini U. Cloning and characterization of the Tn631 restriction modification system of Thermotoga nepalitana. Extremophiles. 2001;5(3):665–72.
20. Xu H, Han D, Xu Z. Overexpression of a lethal methylase, M.Tn631, in E. coli BL21(DE3). Biotechnol Lett. 2014;36(9):1853–9.
21. Ehrlich M, Gama-Sosa MA, Carreha LH, Ljungdahl LG, Kuo KC, Gehrkke CW. DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. Nucleic Acids Res. 1985;13(4):1399–412.
22. Jn Sc Kadam S, Pfefer GP. Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Res. 2010;38(11):e125.
23. Field D, Garrity G, Gray T, Morrison N, Selengut J, Stenker P, Tatusova T, Thomson N, Allen MJ, Argioli SV, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol. 2008;26(5):541–7.
24. Woese CR, Kandler O, Whiles ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87(13):4576–9.
25. Reysenbach A-L, Class I. Thermotogae class, nov. In: Garrity GM, Boone DR, Castenholz RW, editors. Bergey’s manual of systematic bacteriology, vol. 1. 2nd ed. New York: Springer; 2001. p. 369–87.
26. Reysenbach A-L, Bé P. Thermotogae phy.nov. In: Garrity GM, Boone DR, Castenholz RW, editors. Bergey’s manual of systematic bacteriology, vol. 1. 2nd ed. New York: Springer; 2001. p. 369–87.
27. Reysenbach A-L, Family I. Thermotogaceae fam. nov. In: Garrity GM, Boone DR, Castenholz RW, editors. Bergey’s manual of systematic bacteriology, vol. 1. 2nd ed. New York: Springer; 2001. p. 370–87.
28. Bhandari V, Gupta RS. Molecular signatures for the phyllum (class) Thermotogae and a proposal for its division into three orders (Thermotogales, Kosmotogales ord. nov. and Petrotogales ord. nov.) containing four families (Thermotogaceae, Fervidobacteriaceae fam. nov., Kosmotogaceae fam. nov. and Petrotogaceae fam. nov.) and a new genus Pseudothermotoga gen. nov. with five new combinations. Antonie Van Leeuwenhoek. 2010;105(1):143–68.
29. List Editor. Validation of the Publication of New Names and New Combinations Previously Effectively Published Outside the USB. List No. 85. Int J Syst Bacteriol. 2002;52(3):950–90.
30. List Editor. Validation of the Publication of New Names and New Combinations Previously Effectively Published Outside the USB. List No. 22. Int J Syst Bacteriol. 1986;36(4):573–6.
31. List Editor. Validation of the Publication of New Names and New Combinations Previously Effectively Published Outside the USB. List No. 28. Int J Syst Bacteriol. 1989;39(1):193–4.
Xu et al. Standards in Genomic Sciences (2017) 12:62

47. Chhabra SR, Shockley KR, Conners SB, Scott KL, Wolfinger RD, Kelly RM. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium Thermotoga maritima. J Biol Chem. 2003;278(9):7450–2.
48. Yu X, Drapcho CM. Hydrogen production by the hyperthermophilic bacterium Thermotoga neapolitana using agricultural-based carbon and nitrogen sources. Biol Eng Trans. 2011;4(2):101–12.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
50. Kumar S, Steger G, Tamura K. MEGAS: molecular evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33(7):1870–4.
51. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29(12):2607–18.
52. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
53. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14(7):1394–403.
54. Alm RA, Bodero AJ, Free PD, Mattick JS. Identification of a novel gene, pilZ, essential for type 4 fimbrial biogenesis in Pseudomonas aeruginosa. J Bacteriol. 1996;178(1):46–53.
55. van Schaik EJ, Giltner CL, Audette GF, Keizer DW, Bautista DL, Slupsky CM, Sykes BD, Irvin RT. DNA binding: a novel function of Pseudomonas aeruginosa type IV pil. J Bacteriol. 2005;187(4):1394–45.
56. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
57. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
58. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
59. Nunn DN, Lory S. Product of the Pseudomonas aeruginosa gene pilD is a prepilin leader peptidase. Proc Natl Acad Sci U S A. 1991;88(8):3281–5.
60. Nykyri J, Imajime G, Koozov Y, Dubnau D. Characterization of comE, a late competence operon of Bacillus subtilis required for the binding and uptake of transforming DNA. Mol Microbiol. 1993;10(9):1599–111.
61. Draskovic I, Dubnau D. Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulfide bonds. Mol Microbiol. 2005;55(3):881–96.
62. Bakalatz LO, Baker BD, Jurcisek JA, Harrison A, Novotny LA, Bookwalter JE, Munger R, Munson RS Jr. Demonstration of Type IV pilus expression and a twitching phenotype by Haemophilus influenzae. Infect Immun. 2005;73(3):1635–43.
63. Laron TG, Goodgal SH. Sequence and transcriptional regulation of com101A, a locus required for genetic transformation in Haemophilus influenzae. J Bacteriol. 1991;173(15):4683–91.
64. Wolfsberg T, Laufer P, Park HS, Brossay L, Hebert J, Koomen M. PIPT mutations lead to simultaneous defects in competence for natural transformation and twitching motility in pilliated Neisseria gonorrhoeae. Mol Microbiol. 1998;29(1):321–30.
65. Xue et al. Standards in Genomic Sciences (2017) 12:62

Submit your next manuscript to BioMed Central and we will help you at every step:
- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit