JOURNAL OF BACTERIOLOGY, Apr. 2006, p. 2364–2374 Vol. 188, No. 7
0021-9193/06/$08.00+0 doi:10.1128/JB.188.7.2364–2374.2006

Chromosome Evolution in the *Thermotogales*: Large-Scale Inversions and Strain Diversification of CRISPR Sequences

Robert T. DeBoy,* Emmanuel F. Mongodin, Joanne B. Emerson, and Karen E. Nelson

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850

Received 18 October 2005/Accepted 16 January 2006

In the present study, the chromosomes of two members of the *Thermotogales* were compared. A whole-genome alignment of *Thermotoga maritima* MSB8 and *Thermotoga neapolitana* NS-E has revealed numerous large-scale DNA rearrangements, most of which are associated with CRISPR DNA repeats and/or tRNA genes. These DNA rearrangements do not include the putative origin of DNA replication but move within the same replichore, i.e., the same replicating half of the chromosome (delimited by the replication origin and terminus). Based on cumulative GC skew analysis, both the *T. maritima* and *T. neapolitana* lineages contain one or two major inverted DNA segments. Also, based on PCR amplification and sequence analysis of the DNA joints that are associated with the major rearrangements, the overall chromosome architecture was found to be conserved at most DNA joints for other strains of *T. neapolitana*. Taken together, the results from this analysis suggest that the observed chromosomal rearrangements in the *Thermotogales* likely occurred by successive inversions after their divergence from a common ancestor and before strain diversification. Finally, sequence analysis shows that size polymorphisms in the DNA joints associated with CRISPRs can be explained by expansion and possibly contraction of the DNA repeat and spacer unit, providing a tool for discerning the relatedness of strains from different geographic locations.

The advent of genome sequencing has allowed for invaluable insights into the biology of microbial species, particularly as relates to their physiological capabilities. One of the major discoveries brought to the forefront by over a decade of microbial genome sequencing is the extent of gene transfer, now accepted to be far more widespread than originally appreciated, as well as genome rearrangements and gene shuffling, which occur in many microbial species. The mechanisms of genetic exchange can involve mobile genetic elements, such as phages and transposons, which are clearly evident in the genomes of many microbial species (4, 8, 30). However, in some situations, the mechanisms and the reasons why these chromosomal rearrangements happen is not obvious (23, 24).

One of the main advantages of sequencing the genomes of multiple strains from the same species, as well as genomes of closely related species, is that DNA shuffling within the genome, and lateral gene transfer (LGT) events, can be readily identified at the level of a chromosomal replicon. A comparison of closely related genomes therefore enables the identification of chromosomal segments that have undergone DNA rearrangements sometime after the lineages diverged from a common ancestor (6, 7, 17, 19, 32, 39). A computational technique that is commonly used to reveal similarities and differences in the gene order of two microbial genomes is to calculate the pair-wise alignment of their DNA sequences or their translated peptide sequences. The genome alignment can then be visualized as a dot plot in which the x and y coordinates of each position represent similarity between the chromosomes, so that a perfect alignment between two chromosomes would appear as a diagonal line in which $f(x) = x$.

* Corresponding author. Mailing address: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850. Phone: (301) 795-7980. Fax: (301) 838-0208. E-mail: rdeboy@tigr.org.

Comparative genomics of closely related microbial species has revealed an abundance of large-scale genomic changes in the evolution of some species. For example, whole-genome alignments of some closely related species display an "X-shaped" alignment that likely results from numerous chromosomal inversions that pivot around the origin and terminus (9). Whole-genome alignment has also revealed shuffling of chromosomal segments within the same replichore (the half chromosome divided by the replication axis) (39). In the present study, the features examined for two members of the *Thermotogales* are associated with numerous rearrangements within the same replichore.

In the initial analysis of the genome of the hyperthermophilic bacterium *Thermotoga maritima* MSB8, there was evidence that members of this lineage undergo extensive gene transfer, particularly with members of the archaeal domain (24, 27–29). In a more recent study (23), we validated this hypothesis using a comparative genome hybridization (CGH) approach to investigate genome plasticity and LGT in the *Thermotogales*. In this study, numerous gene loss and gain events that have contributed to the metabolic diversity in the members of this species can be seen, and neither mobile elements nor remarkable genomic features such as repeated sequences that could be associated with these genomic rearrangements could be identified. However, our analysis, along with studies of the whole genome, have demonstrated the presence on the chromosome of eight distinct CRISPRs (clustered regularly interspaced short palindromic repeats) (16) that consist of a 30-bp repeat element interspersed with a unique sequence of approximately the same length. These CRISPR elements and their associated group of putative protein-encoding genes (CRISPR-associated sequences [*cas* genes]) have been identified in the genomes of a broad range of microbial species and have been theorized to be involved in the

mobilization of DNA (3, 13, 16). More recently, the intervening spacer sequences in CRISPRs have been shown to have a possible origin from preexisting chromosomal sequences and sometimes from transmissible elements such as bacteriophage and conjugative plasmids. It was also found that these transmissible elements do not reside in cells that carry virus-specific CRISPR spacer sequences but could be found within closely related strains that did not carry these sequences. Thus, a role for CRISPRs in immunity to foreign DNA was proposed (3, 22, 31).

Thermotoga neapolitana strain NS-E, isolated from the Bay of Naples, in Italy (15), has recently been the subject of whole-genome sequencing (Nelson et al., unpublished data). The availability of this additional genome from this lineage of hyperthermophiles has enabled a comprehensive comparative analysis of the chromosomal architecture of the *Thermotogales*. In this report we present a detailed analysis of chromosomal variation and address the features that have contributed to these differences during the evolution of this lineage.

MATERIALS AND METHODS

Whole-genome alignment. Genome sequencing and assembly of *T. neapolitana* NS-E (DSM 4359) was performed as previously described for other microbial genomes sequenced at The Institute for Genomic Research (TIGR) (10, 24, 25). A preliminary draft (8× coverage, 20 contigs representing 1,920,253 bp) of the whole genome sequence of *T. neapolitana* NS-E has been deposited in GenBank under accession no. NC_006811. A whole-genome alignment between *T. neapolitana* strain NS-E and *T. maritima* strain MSB8 (GenBank accession no. AE000512) was performed with the MUMmer package (18) (freely available at http://mummer.sourceforge.net/). The Promer algorithm, part of the MUMmer package, was used to calculate the amino acid percentage of identity for regions containing exact matches of at least 5 amino acids and separated by 30 or fewer amino acids. Gnuplot (www.gnuplot.info/) was then used to visualize the results on a scatterplot that maps each nucleotide position in *T. neapolitana* to its corresponding position in *T. maritima*.

DNA joint assignments. The approximate ends of the rearranged chromosomal segments are evident in the scatterplot of the whole-genome alignment between the two *Thermotoga* species (Fig. 1). Homologous open reading frames (ORFs) at the ends of each rearranged chromosomal segment were identified in the two species by BLASTP analysis (2). The sequence between the putative protein-encoding regions of homologous ORFs at the ends of two adjoining DNA segments is referred to as a "DNA joint." Each of the 15 observed DNA joints was assigned a roman numeral between I and XV (Fig. 1; Table 1) that is based on the order of its appearance in the *T. neapolitana* genome. Also, as depicted in Fig. 1, the DNA joint nomenclature uses a roman numeral (such as X) at one end of a particular DNA segment and a primed roman numeral (such as X′) at the corresponding end of its adjoining DNA segment. The exact position of each rearrangement within these DNA joints was not discerned, because of the relatively low DNA sequence similarity within the intergenic regions of the two species. The DNA joint sequences vary in size from 41 bp to 5,792 bp. Larger DNA joints correspond to regions that encode ORFs that are absent from one of the two species.

Statistical analysis. The Fisher exact test was computed given the null hypothesis that there is no association between the 15 intergenic spaces that contain rearrangements in *T. maritima* and *T. neapolitana* and the intergenic spaces that contain tRNA genes and/or CRISPRs. For *T. maritima*, the test was used to compute the probability that 10 or more instances of tRNA genes and/or CRISPRs are found when 15 intergenic regions are chosen at random from the total pool of 1,095 intergenic spaces, 38 of which are occupied by a tRNA gene and/or a CRISPR. For *T. neapolitana*, the test was used to compute the probability that 9 or more instances of tRNA genes and/or CRISPRs are found when 15 intergenic regions are chosen at random from the total pool of 1,176 intergenic spaces, 37 of which are occupied by a tRNA gene and/or a CRISPR. In both cases, P was <0.001.

Sequence composition analysis. For the cumulative GC skew analysis, the G+C composition using a 1-kb window over the entire length of the chromosomes (Fig. 2A and B), or a window of 100 bp for particular *T. neapolitana* and *T. maritima* subsequences (Fig. 2E and F), was quantified with the formula $(G - C)/(G + C)$. A cumulative GC skew was calculated by using successive windows from the beginning to the end of each sequence, and the value of the cumulative GC skew (y axis) was then plotted at its corresponding position on each DNA molecule (x axis).

For the analysis of ORF orientation (Fig. 2C and D), each ORF in the *T. neapolitana* and *T. maritima* genomes was assigned a value of 1 when oriented from left to right (forward orientation) and −1 in the reverse orientation. A running sum (y axis) was calculated from the first to the last ORF in each genome and plotted at its corresponding position in the chromosome (x axis).

PCR assays. PCR assays were used to compare the sizes and structures of the 15 DNA joints (described above) for five strains of *T. neapolitana* that were isolated from different geographical locations (23). The strains used in the present study included the following: *T. neapolitana* strain NS-E, isolated from a shallow submarine hot spring in Naples, Italy; *T. neapolitana* strains LA4 and LA10, isolated from the shore of Lac Abbe, Djibouti; *Thermotoga* sp. strain RQ7, isolated from a geothermal heated seafloor, Ribeira Quente, the Azores; and *Thermotoga* sp. strain VMA1/L2B, isolated from Vulcano Island, Italy. Although several of these were not previously designated *T. neapolitana* strains, their patterns of hybridization in the CGH study of Mongodin and colleagues (23), as well as phylogenetic analyses of their 16S rRNA sequences (23), suggest that they are in fact closely related to *T. neapolitana*. Genomic DNA for these strains was provided by Karl Stetter and Robert Huber from the University of Regensburg, Germany.

PCR primer pairs were designed from the sequences of ORFs flanking each DNA joint (Fig. 1; Table 1) and are as follows for the 15 respective regions: I, ACATGCCCTGTTATCAACTTCAGG; I′, ATCTGCGATTTCCTTTCTTCT TGC; II, CTGCCTGTGAGTTTCAGAAAAACG; II′, GTTCGTCTTGACCA GTTCGTATCC; III, CTTTTCTGTGATCATCGCTTTTGG; III′, TTTCATT CCTTTCAGTGGTTCAGC; IV, GGTACAACGGTTTGATGAACTTGC; IV′, ACGGCAGAGAGTACACTTTTGTGG; V, AATTTCACTTGAATGG GGAGAAGC; V′, GTCCTGTACCTCCCGTTTATTTCC; VI, CCGGAAA AAGAAGCAATTAAGACG; VI′, TTTTCCTACGGCATAGAAACATGG; VII, GGCAGAAAGATCTTCAACATCACC; VII′, CTGATTTCATGGCA AAAGATCACC; VIII, GAACACGGTTTACAACACGAAACG; VIII′, TG CGTACGGATGATATAAGGAAGG; IX, ATGGTGTGCTTCTTCATGAT CTCC; IX′, ATACGTCCCCTCAAGAACAAGACC; X, GGAACGTTGAA CTCCTCAAGAACC; X′, CCTTGCTTTTCAGCAATTCTTTCC; XI, GTC CTTTGTGATGAATCCATAGCC; XI′, TCTGTGAACATCATTTCCCTA CCG; XII, GGTGTTCAAAAAGACGGAAAGAGG; XII′, GGAAGTTCT GGTGAATGGAGAACC; XIII, CTTTGTTTTCAGAAACGGGAATGG; XIII′, GATCTTTTCGGAATTTGTCGAAGG; XIV, AATTTCACTTGAAT GGGGAGAAGC; XIV′, GTCCTGTACCTCCCGTTTATTTCC; XV, AAT CTCTTTCCGTACCCACTTTCG; XV′, GATCTCAGACGACTCAACGTC TCC. In addition, an internal primer pair was designed for walking the relatively large insert of DNA joint XIII: XIIIb, GCACCAGCACACTTTT CTCATAGC; XIIIb′, AAACCGCACACTTAGCCTCTAACC. PCR amplification was performed with TaKaRa *Taq* polymerase, according to the manufacturer's instructions (Chemicon International, Temecula, CA), with the following cycle profile: 98°C for 20 s, 55°C for 20 s, and 68°C for 60 s per kb, for 30 cycles. The resulting PCR products were visually checked by agarose gel electrophoresis and sequenced by walking directly on the PCR product, and the sequences were used for comparative analyses.

Nucleotide sequence accession numbers. The nucleotide sequences of the CRISPR regions that were amplified in the different strains have been deposited in GenBank under accession numbers DQ352545 to DQ352560 and are listed in Table 2. For each group of strains sharing identical CRISPR spacer sequences, a single sequence was deposited in GenBank. One exception is VMA1/L2B region XII, which has its own accession number because it differs in sequence from the other members of the region XII group which also lack a spacer sequence (strains NS-E, LA10, and LA4).

RESULTS

Chromosomal rearrangements of *T. maritima* strain MSB8 versus *T. neapolitana* strain NS-E. A graphical representation of a whole-genome alignment highlighting the level of protein similarity along the chromosomes of *T. neapolitana* strain NS-E and *T. maritima* strain MSB8 is presented in Fig. 1A. A high degree of colinearity and sequence conservation is evident for these two species, with the two proteomes having an average percentage of identity of 83.6% and an average percentage

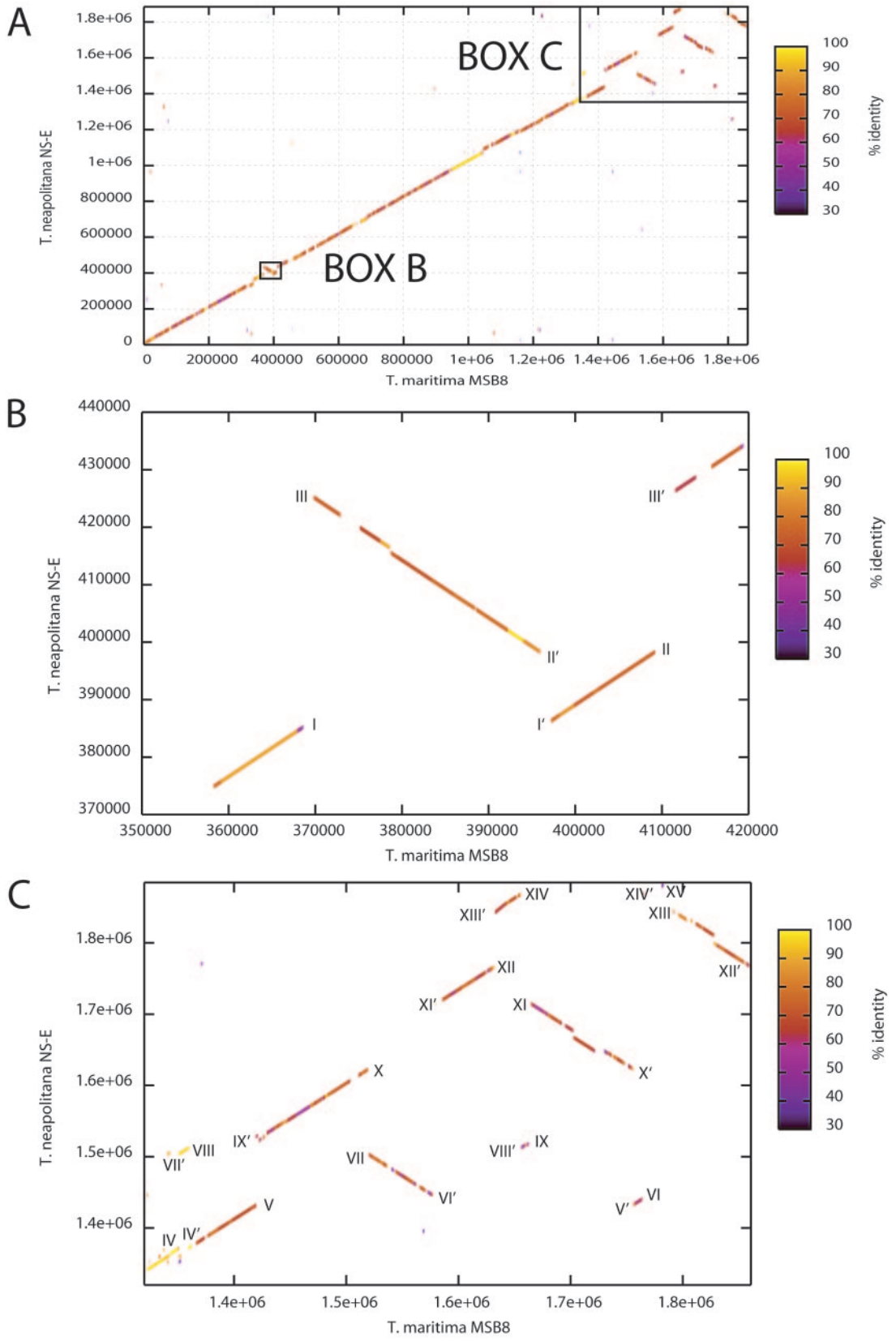

FIG. 1. Whole-genome amino acid alignments between *T. maritima* strain MSB8 and *T. neapolitana* strain NS-E. The Promer algorithm was used to calculate and plot the amino acid percentage identity of maximally unique matching subsequences of at least 5 amino acids between the two genomes. A point (*x*,*y*) indicates a sequence that occurs once within each genome, at location *x* in one genome and at location *y* in the other genome. The matching sequences may occur on either the forward or the reverse strand; in either case, the locations indicate the 5′ end of the sequences. The point 0,0 corresponds to the putative origin of replication for each genome. Panel A shows the alignment at the whole-genome level. Two regions of interest, boxes B and C, are shown in more detail in panels B and C, respectively.

TABLE 1. ORF pairs, which are found at the DNA joints that connect shuffled chromosomal segments in *Thermotoga* spp., and their associated DNA features

DNA joint for indicated strain[a]	ORF at border		Associated feature
	5′ end	3′ end	
T. neapolitana NS-E			
I-I′	GTN0367	GTN0369	CRISPR
II-II′	GTN0379	GTN0380	Unknown
III-III′	GTN0406	GTN0407	CRISPR
IV-IV′	GTN1365	GTN1366	Unknown
V-V′	GTN1429	GTN1430	CRISPR
VI-VI′	GTN1437	GTN1443	tRNA
VII-VII′	GTN1507	GTN1510	Unknown
VIII-VIII′	GTN1516	GTN1517	tRNA
IX-IX′	GTN1525	GTN1527	Unknown
X-X′	GTN1645	GTN1646	CRISPR/tRNA
XI-XI′	GTN1742	GTN1743	tRNA
XII-XII′	GTN1795	GTN1796	CRISPR
XIII-XIII′	GTN1861	GTN1862	CRISPR
XIV-XIV′	GTN1887	GTN1889	Unknown
XV-XV′	GTN1902	GTN0001	Unknown
T. maritima MSB8			
I-III	TM0350	TM0351	CRISPR
II′-I′	TM0376	TM0378	CRISPR
II-III′	TM0389	TM0392	CRISPR
IV-VII′	TM1331	TM1332	Unknown
VIII-IV′	TM1339	TM1339	Duplication
V-IX′	TM1404	TM1405	Unknown
X-VII	TM1523	TM1524	CRISPR
VI′-XI′	TM1588	TM1590	tRNA
XII-XIII′	TM1642	TM1643	CRISPR
XIV-VIII′	TM1668	TM1672	tRNA
IX-XI	TM1682	TM1683	Unknown
X′-V′	TM1778	TM1780	tRNA
VI-XIV′	TM1787	TM1788	tRNA
XV-XIII	TM1814	TM1816	Unknown
XII′-XV′	TM1878	TM0005	CRISPR

[a] Roman numerals were assigned to DNA joints in the order of their appearance in *T. neapolitana* strain NS-E (Fig. 1).

of similarity of 92.4% (percentage of identity range, 23.2 to 100; percentage of similarity range, 42.3 to 100). *T. maritima* strain MSB8 and *T. neapolitana* strain NS-E share 1,726 proteins, out of a total of 1,838 predicted proteins for MSB8 and 1,903 for NS-E. Only 116 proteins (6.3% of the total set of proteins) were found to be unique to *T. maritima*, i.e., lacking a match to a *T. neapolitana* protein with at least 30% similarity over 30% of its length. Also, only 265 proteins (13.9% of the total protein set) were found to be unique to *T. neapolitana*. The full description of the *Thermotoga* core genome, as well as the gene set unique to each strain and the biological implications for each of the *Thermotoga* species, will be further developed in an article describing the *T. neapolitana* genome (Nelson et al., unpublished data).

A large region, delimited by *T. maritima* ORFs TM0939 and TM1016 and covering approximately 80 kb (Fig. 1A, yellow line around coordinate 1000000), is highly conserved between MSB8 and NS-E, with an average percentage of identity of 99.6% between the two strains (compared to 83.6% for the entire proteome). Of the 67 ORFs in this region, 23 encode conserved hypothetical proteins, 5 encode proteins of unknown function, 5 encode proteins involved in ribose metabolism and transport, 1 encodes a protein involved in fucose metabolism, 2 encode putative lipoproteins, 1 encodes a putative membrane protein, and 3 encode putative transcription regulators. This may suggest that some environmental pressure exists to retain some of these ORFs and may explain the preference for the utilization of sugars other than glucose in both species (5). An alternative explanation for the high degree of similarity between the two strains has been proposed recently by Nesbo and coworkers (26). These authors suggest that the high similarity in the TM0939-to-TM1016 region is due to a recent transfer or recombination event between the *T. maritima* and the *T. neapolitana* lineages.

Two chromosomal regions, of approximately 40 kb and 500 kb (Fig. 1B and C), display relatively large rearrangements, including combinations of inverted DNA segments, which have a negative slope (e.g., the segment III-II′ in Fig. 1B), and translocated DNA segments, which have a positive slope and are offset from the otherwise diagonal line (e.g., segment I′-II in Fig. 1B). In total, 15 distinct DNA segments that are rearranged in the chromosome of *T. neapolitana* relative to that of *T. maritima* were identified. The DNA joints that connect these rearranged chromosomal segments, numbered from I to XV in Fig. 1B and C, were assigned as described in Materials and Methods.

Features associated with chromosomal rearrangements. Two predominant types of chromosomal features were identified in the DNA joints between the shuffled DNA segments (Fig. 1B and C; Table 1). In the chromosomes of both species, four DNA joints (VI-VI′, VIII-VIII′, X-X′, and XI-XI′ for *T. neapolitana* and VI′-XI′, XIV-VIII′, X′-V′, and VI-XIV′ for *T. maritima*; Table 1 and Fig. 1B and C) contain one or more tRNA genes. Six DNA joints (I-I′, III-III′, V-V′, X-X′, XII-XII′, and XIII-XIII′ for *T. neapolitana* and I-III, II′-I′, II-III′, X-VII, XII-XIII′, and XII′-XV′ for *T. maritima*) contain one or more copies of a 30-bp DNA repeat (Fig. 3) belonging to a CRISPR element (16). The remaining five DNA joints do not display obvious chromosomal features, with the exception of *T. maritima* TM1339, a conserved hypothetical protein that is duplicated (100% similarity) in *T. neapolitana* (TM1366 and TM1516) where the ORFs are associated with independent, rearranged DNA segments (Table 1). The obvious presence of tRNA genes and CRISPR elements in the DNA joints between shuffled chromosomal segments is unlikely to happen by chance. Statistical analysis, as described in Materials and Methods, shows it to be highly likely that there is an association between these chromosomal features and rearrangements. The Fisher exact test produced a P value of <0.001, given the null hypothesis that the 15 observed rearrangements are not associated with intergenic spaces containing tRNA genes and/or CRISPRs.

Characterization of inverted chromosomal segments. Four inverted chromosomal segments are revealed by their negative slope in the whole-genome alignment of *T. neapolitana* strain NS-E and *T. maritima* strain MSB8. Each of these four inversions can be tentatively assigned to a particular lineage. That is, it is possible to discern which strain has the original orientation (i.e., corresponding to the orientation of the ancestor) for a particular DNA segment and which strain has the inverted orientation. Previous reports have shown that nucleotide composition is biased in many organisms with respect to the direction of DNA replication: GC skew diagrams (20) and cumu-

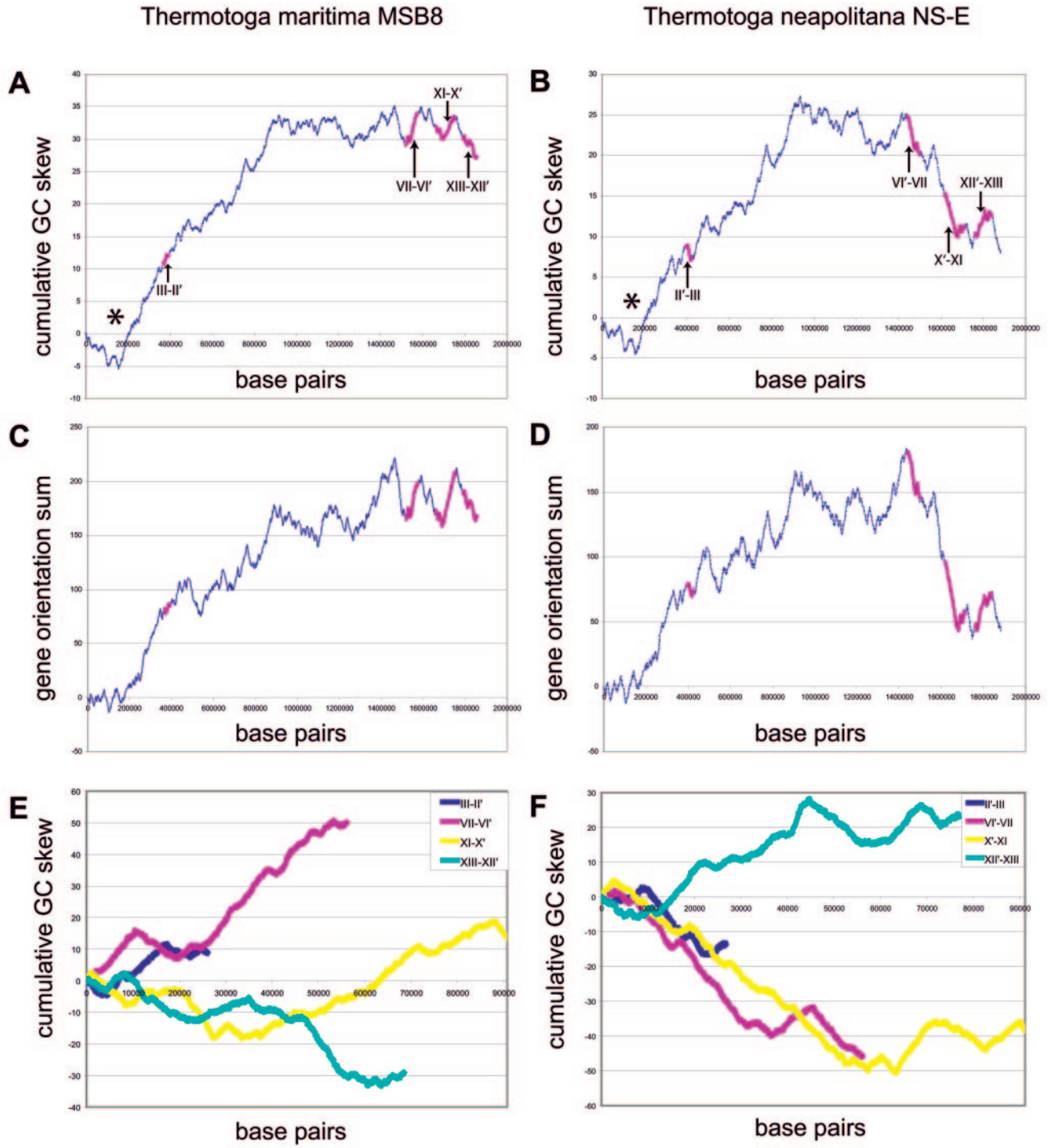

FIG. 2. Cumulative GC skew and ORF orientation in the *T. maritima* strain MSB8 and *T. neapolitana* strain NS-E genomes. (A and B) Plots of cumulative GC skew calculated with 1-kb windows. (C and D) Plots of the running sum of ORF orientation. (E and F) Expanded plots of GC skew for the pink regions displayed in panels A and B and calculated with a 100-bp window. The four regions in *T. maritima* (E) and *T. neapolitana* (F) are putative inverted segments revealed by the whole-genome alignment (Fig. 1). The asterisks in panels A and B correspond to the putative origin of DNA replication, as described by Lopez and coworkers (21). The roman numerals correspond to the nomenclature of the different chromosomal regions displayed in Fig. 1.

lative GC skew diagrams (12) indicate that the leading strand contains more guanine than cytosine residues. Thus, an observed bias for cytosine over guanine in a particular segment of the current leading strand might indicate that a DNA inversion event or a translocation event across the replichore has occurred. Figure 2 depicts the cumulative GC skew for the entire genomes (Fig. 2A and B) or the four putative inverted regions of the completely sequenced genomes of *T. maritima* (Fig. 2E)

TABLE 2. CRISPR spacer consensus sequences found in the DNA joints of five *T. neapolitana* strains

DNA joint	Strain(s) and no.[a]	CRISPR spacer sequence[b]	GenBank accession no.
I	NS-E, LA10		DQ352545
	1	GATTAGTTTTTACCATGTATTGGTAATCTTGTCAAT	
	2	GGGAGTCCTACGTTGGATATCCACAGACGGCAGAGTACA	
	3	GTGGAGAAGGCTTTTATCTCAATGGATATCGTTGGTA	
	4	CGCGAAAGCAAAACTCCACGCCCTCAAAGCGCCTTT	
	5	TCAGTTCGAACACAACCTGGCGACGTTTGTCTCGT	
	LA4	Same as NS-E and LA10 except that it is missing spacer 4	DQ352546
	RQ7		DQ352547
	1	**CGTGCACCTTCTTTGAGAAGTTCAGTGGGATCTTTT**	
	2	**GGAACTGTTAGATGGCTGGGATATGAAGATCAAAAA**	
	3	CTTATTTCGGTGGTCGAACGGAAGCGTGGTTCTTAGG	
	4	TTGTTGAATGTTTGATTTATTGCTTCCATTACAGCGT	
	5	GTACAACAGAGGCACTGGTTCACCGTTGAACAAAGC	
	VMA1/L2B		DQ352548
	1	TGATTCCTCCTATGTACAGCAAAGTATCAGAAACGG	
	2	**CGTGCACCTTCTTTGAGAAGTTCAGTGGGATCTTTT**	
	3	**GGAACTGTTAGATGGCTGGGATATGAAGATCAAAAA**	
III	NS-E, LA10, LA4		DQ352549
	1	CAGGAATTGTTCCAGACGGAGTGGCGGTAGAGACAT	
	2	TCCACGTCAAAGCCGTGCATTTTCAAAGCGAGTCTG	
	3	GAAAAATTGGTTGGTGTTTCTGATGAAATAATTCTGG	
	4	TTGTGAGTTCTGCTCTTTTTGCCTTTTTGTGATATAC	
	5	ACTGGAGGGGGTAGAGGACAGCCTTCTTTTCACCTC	
	6	ATGAATATTTCCCACCAACCCGGCCTTGTACGCAAT	
	RQ7		DQ352550
	1	ATTGTTGTTTGTACTTTCATTTTCTCCCTCCTTTTTCC	
	2	ACTGGAGGGGGTAGAGGACAGCCTTCTTTTCACCTC	
	3	ATGAATATTTCCCACCAACCCGGCCTTGTACGCAAT	
	VMA1/L2B		DQ352551
	1	TAAACCAAACAAGACTCCTCCTTCCCAAGTTTCCAA	
	2	CTGTTAGTGACCTCGCACGCATCACGGGAATCAACAA	
	3	GTCATCACCCCCTTTCCTCTTCCGGAATGTCAACAC	
	4	ATGCCTTCGCTGTAGTGAACATTGAAGGTGGTAAT	
	5	TTTGTGTGCCCTGTAAGATCGACCGCGGCGCTGTACCA	
X	NS-E, LA10, LA4		DQ352552
	1	CTTTATCAGATGATCAAAAGGCTTGAAAAGGAGGATG	
	2	AGTCCAGAGCCAGAGCCATCCACGTTCTGGAGTTTTG	
	3	CCGCATAGTCCAGGTCAGAAAACGGCCCTCCGATGG	
	RQ7		DQ352553
	1	ATGTCCTGACCTTCTTCAACCTGCCTTTTGATATCG	
	2	GTTCATCACGAAGGTGTACACTTACCTGATCATGCA	
	3	CTTTATCAGATGATCAAAAGGCTTGAAAAGGAGGATG	
	VMA1/L2B		DQ352554
	1	TGGTTGAATTTAACAAGTAAATCATGGTCTCTCTCCCT	
	2	ATTATTATTTCTTGCTCTCCTTCCATGCTTTTTAA	
	3	CCTGTTACGGCGCCACCTATGGCATTTGCTATAG	
	4	CAGGCTCTCACCAAGAATGCGCTCGATCTCGTCAGG	
	5	CTACAGGGGGTAATGATACTTCCGCCCCTGTAAGCT	
	6	ATGAAATCCTCATCGAGGAAGGATTCGAAGGAGTCCA	
XII	NS-E, LA10, LA4, VMA1/L2B	No spacer sequences	DQ352555, DQ352557
	RQ7		DQ352556
	1	CACATTCTACGTGGATCGAATCGAGATCCTCAAGAA	
	2	TGTGAGGTCTCCCGCCGGAAGGCCAGACAGCGTGTAG	

Continued on following page

TABLE 2—*Continued*

DNA joint	Strain(s) and no.[a]	CRISPR spacer sequence[b]	GenBank accession no.
	3	AACAAGTTCGAACCTCGTAAATTTTCAGGGTTCGCACCT	
	4	GCAGAAAGCGTGTGTACTTTGACGTTCCCGTGAAGAAGC	
	5	AGCGGCACACCTGAGTTGAAGAACTCGGAGAACTTCA	
XIII	NS-E, LA10, LA4	CCGAGCAGTTCCTGGCGACGGTCAAAGACACAGACCT	DQ352558
	RQ7		DQ352559
	1	CGGCAGCAAACCTCAAAGCGCTCGAAAGAAAGGGG	
	2	CGATGTCTGCGTGCTCTAACCATTTGTACAATTCTT	
	3	CAATATCCGCAAAGACGAGGAAGTTTGGATCGAACC	
	4	ACTGCCTCTCTTAAACGACGAAATGCGGAAGAGAG	
	5	AGAGGAGATACATCGCAGAGGAAACAAAGAGAAAGA	
	6	TCGCCTTCCCATACGAGGGTCTGTACGTCGCCCTGG	
	7	TAGCCACGGCGAATGCTTTGTCGTGGCGGGTGAGGT	
	8	TGAAGTACGCCATTTACACGAACCCTAAGGGCGGAT	
	9	ATAAACTCGGCACTTCGAGCATAGACGAGAACTTCGA	
	VMA1/L2B	Same as RQ7 except that it is missing spacer 4	DQ352560

[a] Grouped strains have similar spacer sequences; "no." indicates the order of appearance of the spacer sequences.
[b] Underlining or bold for sequences in the same DNA joint indicates strains that share some but not all spacers.

and *T. neapolitana* (Fig. 2F). A putative origin of DNA replication has been located by Lopez and coworkers around position 157000 on the *T. maritima* chromosome (21) (Fig. 2A); by homology to *T. maritima*, the origin of replication of *T. neapolitana* would therefore be located around position 158000 (Fig. 2B). The leading strand, between the origin of replication and the terminus, is expected to display a net positive GC skew. By inference, DNA segment II′-III is inverted in the leading strand of *T. neapolitana*. A net negative GC skew is expected for the lagging strand between the terminus and the end of the molecule in the graph. The contour of the GC skew for *T. neapolitana* has a negative slope except where there are multiple DNA inversions (one of which, XII′-XIII, was identified in the whole-genome alignment). By inference from the cumulative GC skew, DNA segment XII′-XIII is inverted in the lagging strand. Also, DNA segment VI′-VII and DNA segment X′-XI appear to be oriented correctly in the lagging strand of *T. neapolitana*; that is, these two DNA segments are inverted in *T. maritima*. This conclusion is tempered by the observation that the cumulative GC skew for DNA segment XI-X′ is ambiguous in *T. maritima*. Regardless, one or more inversion events appear to have occurred in each of the *Thermotoga* lineages after they diverged from a common ancestor.

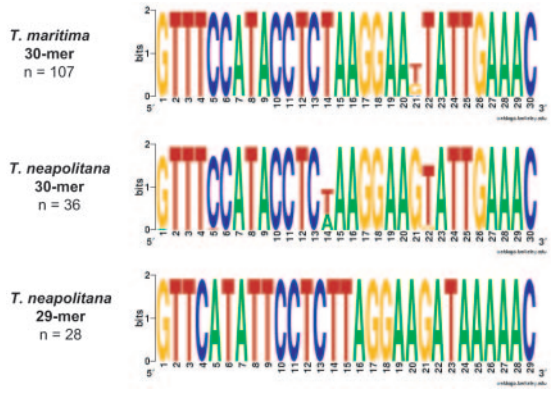

FIG. 3. CRISPR motifs in the genomes of *T. maritima* strain MSB8 and *T. neapolitana* strain NS-E. The CRISPR sequences were generated from three different multiple sequence alignments that were compiled by DNA HMM searches. Note the overlapping identity between the two 30-mers in the different *Thermotoga* species. Also, all occurrences of the 29-mer are located in a single region in *T. neapolitana* which is absent from *T. maritima*.

One obvious exception to the expected trends in the contours of GC skew is the region between ~1.3 Mb and ~1.42 Mb in *T. maritima* and *T. neapolitana*, which was not identified as an inversion in the whole-genome alignment but does display an inverted GC skew. This observation can be explained if these lineages share chromosomal inversions that occurred before the split between *T. maritima* and *T. neapolitana*. The locations of these potential rearrangements appear as common "peaks and valleys" in the contours of the plots in Fig. 2A and B. These two genomes also display a strong correlation between the contours of cumulative GC skew and ORF orientation (Fig. 2C and D). That is, GC content closely reflects ORF orientation, sometimes more so than the position of the origin. One possibility is that many large-scale DNA inversions have shuffled the ORFs with respect to the origin of DNA replication and that the GC content of these displaced ORFs has not had time to ameliorate in their new locations in the chromosome.

Characterization of DNA joints among five strains of *T. neapolitana*. To investigate the prevalence of these chromosomal shuffling events beyond the two completely sequenced strains of *T. maritima* and *T. neapolitana*, and to assess whether or not related but different isolates share these particular DNA joints, PCR assays of four *T. neapolitana* strains from different geographic locales (LA10, LA4, RQ7, and VMA1/L2B, which are described in Materials and Methods) were performed using primer pairs designed to bridge the 15 DNA joints that connect the shuffled chromosomal segments in *T. neapolitana*

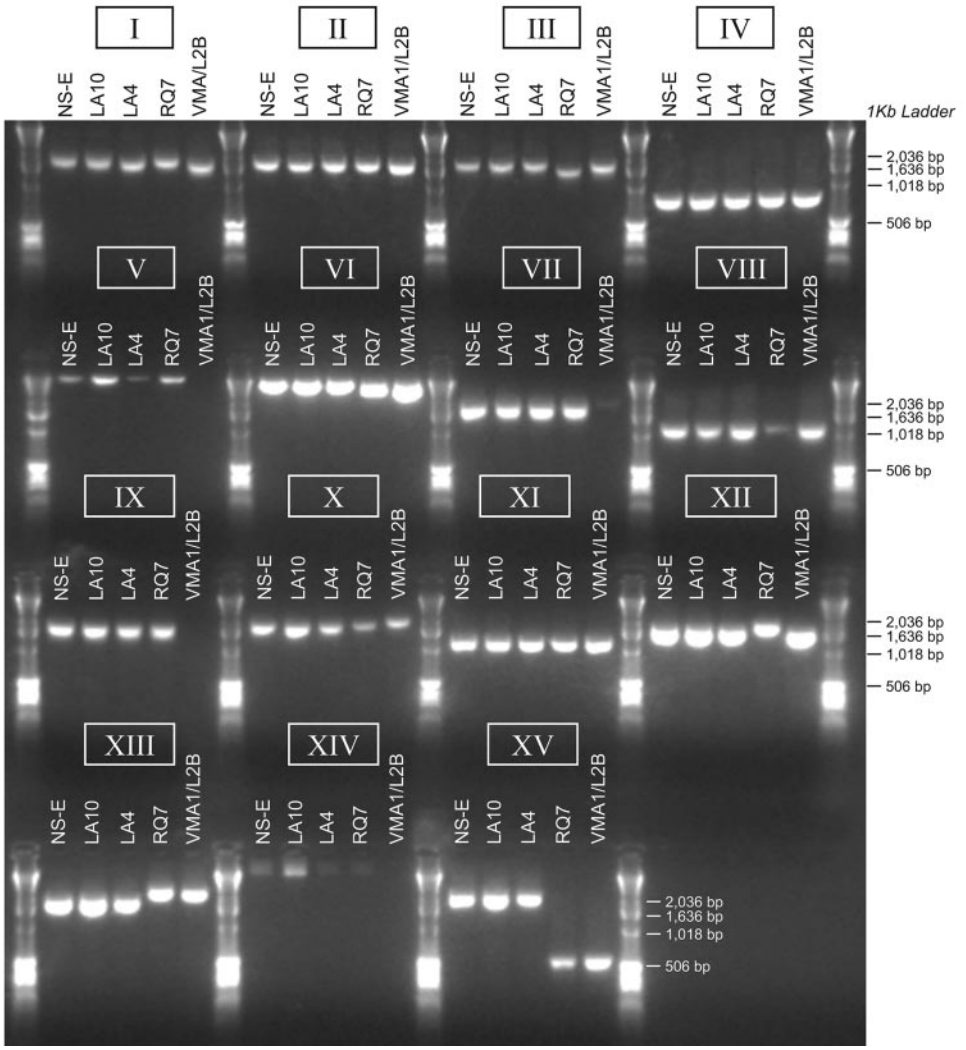

FIG. 4. An ethidium bromide-stained agarose gel of PCR products for each of the *T. neapolitana* species at the 15 DNA joints described in Materials and Methods.

strain NS-E. In the majority of the strains, PCR products were obtained for all 15 of the DNA joints (Fig. 4). Also, for at least eight DNA joints (regions I, III, VI, VII, X, XII, XIII, and XV), the sizes of the PCR products appear to vary between strains, with strains RQ7 and VMA1/L2B most often associated with size changes or the absence of a PCR product (see below). Despite these differences in size, perhaps the most significant result is that 15 PCR products were obtained for all of the *T. neapolitana* test strains, with the exception of three DNA joints (regions V, IX, and XIV) in strain VMA1/L2B. At region XIV, for example, PCR products of varying abundance are present for strains NS-E, LA10, LA4, and RQ7 but not for VMA1/L2B, which might have failed to produce PCR products because of sequence divergence at the corresponding primer binding sites. That is, the architecture of the shuffled chromosomal segments was established in a common ancestor of the thermophilic *T. neapolitana* lineages that were isolated from disparate geographical locations.

Table 2 summarizes the sequence results for five of the six DNA joints (I, III, X, XII and XIII) that are associated with CRISPRs. None of the sequenced CRISPR spacer sequences matched preexisting sequences in GenBank. The common theme for these variable-length DNA joints is the expansion and possibly the contraction of CRISPR repeat and spacer units. The uniqueness of the CRISPR spacer also allows for strain comparisons. For example, strain NS-E (isolated from Naples, Italy) and strains LA10 and LA4 (both isolated from Lac Abbe, Djibouti) appear most similar to one another; they share four out of five spacer sequences in region I, six spacer sequences in region III, three spacer sequences in region X, and one spacer sequence in region XIII. Likewise, strain RQ7 (isolated from Ribeira Quente, the Azores) and strain VMA1/L2B (isolated in Vulcano Island, Italy) appear similar to each other, sharing eight out of nine spacer sequences in region XIII and two spacer sequences in region I, where both strains have also diversified with novel spacer sequences. However, strain RQ7 also has similarities to strains NS-E, LA10, and LA4; they share two spacer sequences in region III and one spacer sequence in region X. Again, the CRISPR sequences have diversified beyond the shared spacer

sequences in these two regions. Conversely, RQ7 displays a novel expansion of CRISPR spacers in region XII, which the other strains are lacking. Thus, in addition to its unique sequences, strain RQ7 appears to host a mosaic of CRISPR sequences that are found in both the VMA1/L2B strain and the NS-E, LA10, and LA4 group of strains. Based on this observation, it seems reasonable to propose that these strains are derived from an RQ7-like ancestor and that strain diversification yielded at least three lineages: RQ7, VMA1/L2B, and the group comprised of NS-E, LA10, and LA4.

DISCUSSION

Comparative genomics is a valuable approach for reconstructing evolutionary relationships between related organisms. It has commonly been used to perform pairwise comparisons of closely related pathogenic and nonpathogenic species, with the goal of identifying novel genes involved in virulence or genes specific to a particular serotype/phenotype of infection (10, 11). Very few studies, however, have used comparative genomics to investigate microbial genome plasticity and chromosome evolution. The results of the study presented here highlight the value of applying whole-genome sequencing and comparative genomic analysis to closely related species. The data gathered from the direct comparison of the *T. neapolitana* and *T. maritima* genomes has revealed important information about chromosome shuffling-driven evolution of the chromosomes of these two species and the strong association of CRISPR sequences and tRNA genes with the observed large-scale rearrangements.

Global genomic rearrangements, such as duplications, inversions, and translocations, contribute significantly to the evolution of species. Pair-wise comparisons of organisms such as *Helicobacter*, *Chlamydia*, *Mycobacterium*, *Vibrio cholerae*, *Escherichia coli*, and *Pyrococcus* (9, 39) have shown that genome rearrangements occurred mainly via replication-directed translocation across an axis defined by the origin and the terminus of replication. Since matching sequences tend to occur at the same distance from the origin (but not necessarily on the same side of the origin), whole-genome alignments display "X-shaped" patterns that are symmetric about the origin of replication of the two genomes being compared (9, 39). The origin of replication of *T. maritima* still remains unknown, mostly because the classical approaches, such as GC ratio, GC skew (20), and asymmetric distribution of oligomers along the genome (35), have failed to unambiguously detect it. In the 1999 publication of the *T. maritima* genome, Nelson and coworkers (24) assigned bp 1 of the genome to the beginning of the longest stretch (2.6 kb) of 30-bp repeats, which was characterized later as one of the eight CRISPR loci present in the chromosome. In a 2000 publication, Lopez and colleagues (21) used tetramer skews and subsequent identification of DNA repeats having similarity to DnaA boxes to predict that the origin of DNA replication is located between coordinates 156960 and 157518. Although the typical features of bacterial origins of replication, such as a local minimum in a plot of cumulative GC skew, seem to be in agreement with this prediction, it has not been experimentally confirmed.

Early genetic studies of chromosomal inversions in *Salmonella enterica* (36) and *E. coli* (33) found that these rearrangements can use endpoints encompassing the origin of DNA replication or endpoints contained within a replichore. Different explanations are proposed for the constraints observed at some chromosomal segments (reviewed in reference 34). The results of the present study suggest that *Thermotogales* species, including *T. maritima* and *T. neapolitana*, favor inversion/translocation events within a replichore. In the example below, we propose a model in which a succession of simple inversion events produces the mosaic of chromosomal rearrangements displayed in the whole-genome alignment of the two *Thermotoga* species. From the cumulative GC skew analysis presented in Fig. 2, DNA segment II′-III of *T. neapolitana* appears to be inverted, and from the whole-genome alignment, the adjoining DNA segment, I′-II, appears to be translocated. In this simple scenario, the *T. maritima* sequence can be represented as the DNA string ABCD and the *T. neapolitana* sequence can be represented as the DNA string ACB′D, where C is the translocated segment I′-II and B′ is the inverted segment II′-III (Fig. 5). In a series of two inversion events, the *T. maritima* sequence can be rearranged into the *T. neapolitana* sequence. In one possible path (Fig. 5, green arrows), the segment BC flips once, producing the segment AC′B′D. A second inversion occurs, flipping the segment C′ to produce the final *T. neapolitana* sequence ACB′D. An alternative way (Fig. 5, red arrows) of producing the same result would be to flip segment C, producing the sequence ABC′D, and then flip the segment BC′ to produce the final *T. neapolitana* sequence ACB′D. Although less straightforward, a similar transformation via a series of inversion events might be responsible for the more

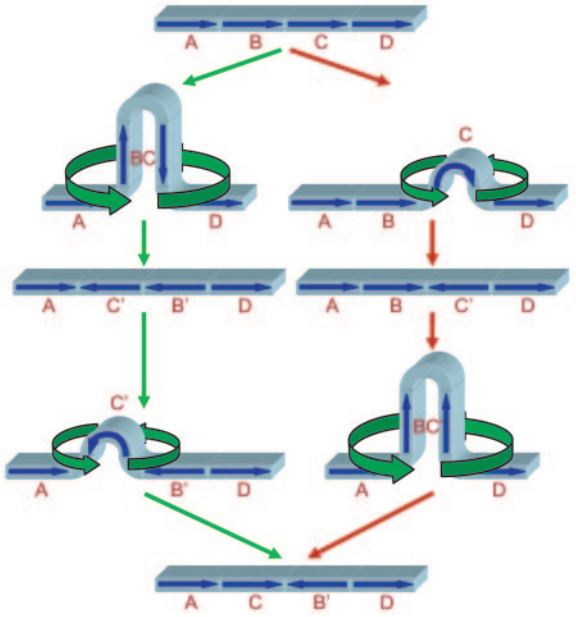

FIG. 5. Two-step model of successive inversions to produce an inversion of one DNA segment (B to B′) and a concomitant translocation of an adjoining DNA segment (C). Two alternative pathways, which differ in the order of the larger and smaller inversions, are proposed: the green path begins with a large inversion, and the red path ends with a large inversion. In both pathways, segment C is inverted in the first step, and it is proposed that there was positive selective pressure favoring a subsequent cell population in which segment C is restored to its original orientation after the second step.

complex pattern of rearrangements observed in the 500-kb region between 1.3 Mb and the end of the chromosome for these two strains.

It is difficult to know the biological significance of the observed DNA rearrangements within these *Thermotoga* strains. In this study, we have identified four chromosomal segments that have been inverted in either *T. maritima* or *T. neapolitana*. One view speculates that one or more of the DNA inversions confer a metabolic advantage. Gene expression of physiologically important pathways, for example, might differ in the two orientations that are observed in *T. neapolitana* and *T. maritima*, so that these species would differ in their metabolic or growth profiles, with no significant genetic differences between the two organisms. All members of the *Thermotogales* possess the ability to produce hydrogen (H_2), but *T. neapolitana* has been shown to produce considerably greater quantities than the rest (37, 38). However, metabolic studies, as well as genomic data, have shown that both *T. maritima* and *T. neapolitana* contain the same pathways for transporting, hydrolyzing, and utilizing a range of poly- and monosaccharides. It is also clear that the genes unique to each species do not seem to account for the observed differences in H_2 production. Thus, it is possible that the same genes and pathways placed under different regulatory conditions (e.g., constitutive versus inducible activation), combined with critical mutations of key genes, would greatly improve a particular metabolic ability of one species compared to another. Alternatively, it is possible to speculate that the inverted DNA segments themselves are not metabolically significant. In the modeled inversion pathways described above, for example, flipping BC in ABCD may lead to a "weakening" of genes in segment C′. Thus, a subsequent inversion of segment C′, producing ACB′D, would restore the metabolic and growth profile. In this scenario, the biologically important sequences are located outside the inverted segment.

The importance of the involvement of CRISPR elements and tRNA genes in these chromosomal rearrangements is currently unknown. Our study shows an association of these sequence features with the location of DNA joints for the potential DNA inversions. Unfortunately, as modeled by the alternative pathways described above, the precise sequence of DNA inversions is also unknown. It is therefore difficult to discern which particular CRISPR and tRNA genes might be involved. In the above two-step inversion example that represents the transformation of ABCD to ACB′D (Fig. 5), the DNA joint adjoining either segment A or segment D might be involved in both steps, depending on which path (green or red arrows) is used. Thus, it is possible that the biologically important sequence feature resides at one or the other location.

The targeted PCR and sequence analysis of additional strains provided information that validates a previous relatedness study of these two members of the *Thermotogales* (23). First and foremost, the five strains of *T. neapolitana* examined here have remarkably similar gross chromosome architectures; thus, their common rearrangements appear to have occurred before strain diversification. Even the apparent lack of 3 of 15 PCR products for strain VMA1/L2B might be explained by sequence divergence of the PCR primer sites. That is, VMA1/L2B might have the same chromosome architecture as the other four strains of *T. neapolitana* at all 15 DNA joints. Alternatively, the chromosome architecture of strain VMA1/L2B might differ at three DNA joints that did not produce PCR products, which might suggest that a VMA1/L2B-like ancestor diverged relatively early, before subsequent chromosomal rearrangements gave rise to an ancestor of the other *T. neapolitana* strains having all 15 DNA joints.

Based on a more detailed sequence analysis of the short CRISPR spacer sequences at five loci, a prediction of this study is that the five strains of *T. neapolitana* can be clustered into three different groups: strain RQ7, strain VMA1/L2B, and the group comprised of strains NS-E, LA10, and LA4. This conclusion agrees with the previous conclusions of a hierarchical clustering of CGH data for *T. neapolitana* strains compared to *T. maritima* strain MSB8 and differs somewhat from a phylogenetic comparison of 16S rRNA genes for the same strains (23). Thus, CRISPR spacer sequence analysis appears to add information to 16S rRNA analysis for reconstructing the relatedness of strains. From a comparison of their spacer sequences, strain RQ7 appears to share components of two DNA joints with strain VMA1/L2B and two DNA joints with the group of NS-E, LA10, and LA4 strains. Thus, an RQ7-like ancestor appears to be the common link between these three *T. neapolitana* strain groups.

The rich diversity of CRISPR spacer sequences in the *Thermotogales* examined so far hints that a treasure trove of horizontally transferred genetic elements exists in the extreme environment these organisms live in. Previous tallies of CRISPR sequences in *T. maritima* identified 105 unique spacer sequences in strain MSB8 and 39 more sequences in locus I of additional strains (23). In this study, 49 unique spacer sequences were identified in the five CRISPR regions examined in the five strains of *T. neapolitana*. Mojica and colleagues (22) suggested that CRISPR spacer sequences could be involved in conferring specific immunity against foreign DNA, such as plasmids and phages; e.g., CRISPR spacers get added in response to foreign DNA. However, with the exception of the relatively small pRQ7-like plasmids (1, 14), these types of horizontally transferred genetic elements have not been identified in the *Thermotogales*. Thus, there is potentially a significant cache of genetic elements awaiting isolation and study. Some of the CRISPR spacer sequences in the five strains of *T. neapolitana* were observed to vary by geographic locale. However, other spacer sequences were found to be shared at four loci by members of the NS-E, LA10, and LA4 strain group, suggesting that these particular CRISPR spacer sequences might have originated in a common ancestor and the three lineages subsequently migrated to their current locales. Future studies examining strains from more diverse locations, and more strains at the same locations, might answer the question of whether or not there is a correlation between CRISPR spacer sequences and geographic location.

ACKNOWLEDGMENTS

We thank Karl Stetter and Robert Huber for providing genomic DNA samples, Derrick E. Fouts for useful discussions about chromosomal rearrangements, Nirmal Bhagabati for help with statistical analysis, and Ioana Hance for laboratory assistance.

This project was supported by U.S. Department of Energy Office of Biological Energy Research Co-Operative Agreements DE-FC02-95ER61962 and DE-FG02-01ER63133.

REFERENCES

1. **Akimkina, T., P. Ivanov, S. Kostrov, T. Sokolova, E. Bonch-Osmolovskaya, K. Firman, C. F. Dutta, and J. A. McClellan.** 1999. A highly conserved plasmid from the extreme thermophile *Thermotoga maritima* MC24 is a member of a family of plasmids distributed worldwide. Plasmid **42:**236–240.
2. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
3. **Bolotin, A., B. Quinquis, A. Sorokin, and S. D. Ehrlich.** 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151:**2551–2561.
4. **Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, and H. Brussow.** 2003. Phage as agents of lateral gene transfer. Curr. Opin. Microbiol. **6:**417–424.
5. **Chhabra, S. R., K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly.** 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima* J. Biol. Chem. **278:**7540–7552.
6. **Chinen, A., I. Uchiyama, and I. Kobayashi.** 2000. Comparison between *Pyrococcus horikoshii* and *Pyrococcus abyssi* genome sequences reveals linkage of restriction-modification genes with large genome polymorphisms. Gene **259:**109–121.
7. **Diruggiero, J., D. Dunn, D. L. Maeder, R. Holley-Shanks, J. Chatard, R. Horlacher, F. T. Robb, W. Boos, and R. B. Weiss.** 2000. Evidence of recent lateral gene transfer among hyperthermophilic archaea. Mol. Microbiol. **38:**684–693.
8. **Dobrindt, U., B. Janke, K. Piechaczek, G. Nagy, W. Ziebuhr, G. Fischer, A. Schierhorn, M. Hecker, G. Blum-Oehler, and J. Hacker.** 2000. Toxin genes on pathogenicity islands: impact for microbial evolution. Int. J. Med. Microbiol. **290:**307–311.
9. **Eisen, J. A., J. F. Heidelberg, O. White, and S. L. Salzberg.** 2000. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol. **1:**RESEARCH0011.1–0011.9. [Online.] http://genomebiology.com/2000/1/6/RESEARCH/0011.
10. **Fouts, D. E., E. F. Mongodin, R. E. Mandrell, W. G. Miller, D. A. Rasko, J. Ravel, L. M. Brinkac, R. T. DeBoy, C. T. Parker, S. C. Daugherty, R. J. Dodson, A. S. Durkin, R. Madupu, S. A. Sullivan, J. U. Shetty, M. A. Ayodeji, A. Shvartsbeyn, M. C. Schatz, J. H. Badger, C. M. Fraser, and K. E. Nelson.** 2005. Major structural differences and novel potential virulence mechanisms from the genomes of multiple *Campylobacter* species. PLoS Biol. **3**(1)**:**e15.
11. **Glaser, P., L. Frangeul, C. Buchrieser, C. Rusniok, A. Amend, F. Baquero, P. Berche, H. Bloecker, P. Brandt, T. Chakraborty, A. Charbit, F. Chetouani, E. Couve, A. de Daruvar, P. Dehoux, E. Domann, G. Dominguez-Bernal, E. Duchaud, L. Durant, O. Dussurget, K. D. Entian, H. Fsihi, F. Garcia-del Portillo, P. Garrido, L. Gautier, W. Goebel, N. Gomez-Lopez, T. Hain, J. Hauf, D. Jackson, L. M. Jones, U. Kaerst, J. Kreft, M. Kuhn, F. Kunst, G. Kurapkat, E. Madueno, A. Maitournam, J. M. Vicente, E. Ng, H. Nedjari, G. Nordsiek, S. Novella, B. de Pablos, J. C. Perez-Diaz, R. Purcell, B. Remmel, M. Rose, T. Schlueter, N. Simoes, A. Tierrez, J. A. Vazquez-Boland, H. Voss, J. Wehland, and P. Cossart.** 2001. Comparative genomics of *Listeria* species. Science **294:**849–852.
12. **Grigoriev, A.** 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. **26:**2286–2290.
13. **Haft, D. H., J. Selengut, E. F. Mongodin, and K. E. Nelson.** 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. **1**(6)**:**e60.
14. **Harriott, O. T., R. Huber, K. O. Stetter, P. W. Betts, and K. M. Noll.** 1994. A cryptic miniplasmid from the hyperthermophilic bacterium *Thermotoga* sp. strain RQ7. J. Bacteriol. **176:**2759–2762.
15. **Jannasch, H. W., R. Huber, S. Belkins, and K. O. Stetter.** 1988. *Thermotoga neapolitana* sp. nov. of the extremely thermophilic, eubacterial genus Thermotoga. Arch. Microbiol. **150:**103–104.
16. **Jansen, R., J. D. Embden, W. Gaastra, and L. M. Schouls.** 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. **43:**1565–1575.
17. **Kobayashi, I.** 2001. Genome comparison: involvement of restriction modification genes in genome rearrangements. Tanpakushitsu Kakusan Koso **46:**2393–2399. (In Japanese.)
18. **Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg.** 2004. Versatile and open software for comparing large genomes. Genome Biol. **5:**R12. [Online.] http://genomebiology.com/2004/5/2/R12.
19. **Lecompte, O., R. Ripp, V. Puzos-Barbe, S. Duprat, R. Heilig, J. Dietrich, J. C. Thierry, and O. Poch.** 2001. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res. **11:**981–993.
20. **Lobry, J. R.** 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13:**660–665.
21. **Lopez, P., P. Forterre, H. le Guyader, and H. Philippe.** 2000. Origin of replication of *Thermotoga maritima*. Trends Genet. **16:**59–60.
22. **Mojica, F. J., C. Diez-Villasenor, J. Garcia-Martinez, and E. Soria.** 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. **60:**174–182.
23. **Mongodin, E. F., I. R. Hance, R. T. DeBoy, S. R. Gill, S. Daugherty, R. Huber, C. M. Fraser, K. Stetter, and K. E. Nelson.** 2005. Gene transfer and genome plasticity in *Thermotoga maritima*, a model hyperthermophilic species. J. Bacteriol. **187:**4935–4944.
24. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, C. M. Fraser, et al.** 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399:**323–329.
25. **Nelson, K. E., D. E. Fouts, E. F. Mongodin, J. Ravel, R. T. DeBoy, J. F. Kolonay, D. A. Rasko, S. V. Angiuoli, S. R. Gill, I. T. Paulsen, J. Peterson, O. White, W. C. Nelson, W. Nierman, M. J. Beanan, L. M. Brinkac, S. C. Daugherty, R. J. Dodson, A. S. Durkin, R. Madupu, D. H. Haft, J. Selengut, S. Van Aken, H. Khouri, N. Fedorova, H. Forberger, B. Tran, S. Kathariou, L. D. Wonderling, G. A. Uhlich, D. O. Bayles, J. B. Luchansky, and C. M. Fraser.** 2004. Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. Nucleic Acids Res. **32:**2386–2395.
26. **Nesbo, C. L., M. Dlutek, and W. F. Doolittle.** 2006. Recombination in *Thermotoga*: implications for species concepts and biogeography. Genetics, doi: 10.1534/genetics.105.049312. [Epub ahead of print.]
27. **Nesbo, C. L., and W. F. Doolittle.** 2003. Targeting clusters of transferred genes in *Thermotoga maritima.* Environ. Microbiol. **5:**1144–1154.
28. **Nesbo, C. L., S. L'Haridon, K. O. Stetter, and W. F. Doolittle.** 2001. Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. **18:**362–375.
29. **Nesbo, C. L., K. E. Nelson, and W. F. Doolittle.** 2002. Suppressive subtractive hybridization detects extensive genomic diversity in *Thermotoga maritima*. J. Bacteriol. **184:**4475–4488.
30. **Paulsen, I. T., L. Banerjei, G. S. Myers, K. E. Nelson, R. Seshadri, T. D. Read, D. E. Fouts, J. A. Eisen, S. R. Gill, J. F. Heidelberg, H. Tettelin, R. J. Dodson, L. Umayam, L. Brinkac, M. Beanan, S. Daugherty, R. T. DeBoy, S. Durkin, J. Kolonay, R. Madupu, W. Nelson, J. Vamathevan, B. Tran, J. Upton, T. Hansen, J. Shetty, H. Khouri, T. Utterback, D. Radune, K. A. Ketchum, B. A. Dougherty, and C. M. Fraser.** 2003. Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. Science **299:**2071–2074.
31. **Pourcel, C., G. Salvignol, and G. Vergnaud.** 2005. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151:**653–663.
32. **Read, T. D., G. S. Myers, R. C. Brunham, W. C. Nelson, I. T. Paulsen, J. Heidelberg, E. Holtzapple, H. Khouri, N. B. Federova, H. A. Carty, L. A. Umayam, D. H. Haft, J. Peterson, M. J. Beanan, O. White, S. L. Salzberg, R. C. Hsia, G. McClarty, R. G. Rank, P. M. Bavoil, and C. M. Fraser.** 2003. Genome sequence of *Chlamydophila caviae* (*Chlamydia psittaci* GPIC): examining the role of niche-specific genes in the evolution of the Chlamydiaceae. Nucleic Acids Res. **31:**2134–2147.
33. **Rebollo, J. E., V. Francois, and J. M. Louarn.** 1988. Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA **85:**9391–9395.
34. **Rocha, E. P.** 2004. Order and disorder in bacterial genomes. Curr. Opin. Microbiol. **7:**519–527.
35. **Salzberg, S. L., A. J. Salzberg, A. R. Kerlavage, and J. F. Tomb.** 1998. Skewed oligomers and origins of replication. Gene **217:**57–67.
36. **Schmid, M. B., and J. R. Roth.** 1983. Selection and endpoint distribution of bacterial inversion mutations. Genetics **105:**539–557.
37. **Van Ooteghem, S. A., S. K. Beer, and P. C. Yue.** 2002. Hydrogen production by the thermophilic bacterium *Thermotoga neapolitana*. Appl. Biochem. Biotechnol. **98–100:**177–189.
38. **Van Ooteghem, S. A., A. Jones, D. Van Der Lelie, B. Dong, and D. Mahajan.** 2004. H(2) production and carbon utilization by *Thermotoga neapolitana* under anaerobic and microaerobic growth conditions. Biotechnol. Lett. **26:**1223–1232.
39. **Zivanovic, Y., P. Lopez, H. Philippe, and P. Forterre.** 2002. *Pyrococcus* genome comparison evidences chromosome shuffling-driven evolution. Nucleic Acids Res. **30:**1902–1910.