Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes

Matteo Cossu, Violette Da Cunha, Claire Toffano-Nioche, Patrick Forterre, Jacques Oberto

Biochimie 118 (2015) 313–321

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes

Matteo Cossu, Violette Da Cunha, Claire Toffano-Nioche, Patrick Forterre, Jacques Oberto*

Institute of Integrative Cellular Biology, CEA, CNRS, Université Paris Sud, 91405 Orsay, France

ARTICLE INFO

Article history:
Received 30 March 2015
Accepted 8 July 2015
Available online 10 July 2015

Keywords:
Archaea
Thermococcales
Genome evolution
Mobile elements
Bioinformatics
Chromosomal landmarks

ABSTRACT

The genomes of the 21 completely sequenced Thermococcales display a characteristic high level of rearrangements. As a result, the prediction of their origin and termination of replication on the sole basis of chromosomal DNA composition or skew is inoperative. Using a different approach based on biologically relevant sequences, we were able to determine *oriC* position in all 21 genomes. The position of *dif*, the site where chromosome dimers are resolved before DNA segregation could be predicted in 19 genomes. Computation of the core genome uncovered a number of essential gene clusters with a remarkably stable chromosomal position across species, in sharp contrast with the scrambled nature of their genomes. The active chromosomal reorganization of numerous genes acquired by horizontal transfer, mainly from mobile elements, could explain this phenomenon.

1. Introduction

The discovery of anaerobic hyperthermophilic microbes by Karl Stetter and Wolfram Zillig extended the limits of life beyond environmental barriers commonly considered as insuperable. Inhospitable habitats such as saline thermal pools and deep sea hydrothermal vents have been remarkably colonized by these extremophilic life forms. The organisms whose optimal growth temperature approaches or exceeds that of boiling water, belong exclusively to the third domain of life: the Archaea. A significant proportion of microorganisms thriving at the fringe of life in terms of temperature belong to the taxonomic order Thermococcales, ranked in the Euryarchaeaota phylum [1]. Thermococcales are divided into three principal genera: *Pyrococcus*, *Thermococcus* and *Palaeococcus*, and grow chemoorganoheterotrophically at temperatures ranging from 80 °C to 100 °C [2]. They require a source of protein and present variable amino acid requirements; several species such as *Pyrococcus furiosus* and *Thermococcus kodakarensis* are able to use chitin as a carbon source [3]. Thermococcales grow easily in the laboratory in complete or synthetic media under strict anoxia. To produce energy, these Archaea prefer anaerobic respiration using S° as terminal electron acceptor to produce hydrogen sulfide. Alternatively, they are able to ferment pyruvate to produce hydrogen [2]. Such unique growth parameters prompted several teams to investigate biosynthetic pathways in Thermococcales. The central metabolism differs quite notably from previously known pathways. The pentose pathway is absent, the TCA cycle is incomplete and glycolysis uses a number of enzymes remarkably different from the canonical view [2]. Even if the net energy balance is still subject to debate, it appears that these Archaea are geared towards an extremely conservative use of energy [2]. Despite their extreme growth conditions, low energetic efficiency and simplified biochemistry, Thermococcales display a very short generation time as low as 23 min [4]. This doubling interval is remarkably similar to that of the fast growing model microbe *Escherichia coli*, grown under the much more favorable conditions of aerobic respiration [5]. Growth efficiency of Thermococcales is in sharp contrast with an apparent disorganization of their chromosome. Indeed it has been reported that these genomes are subjected to a shuffling-driven evolution [6]. This apparent paradox prompted us to investigate, in this work, the process of fast cell growth and rapid chromosome replication by analyzing genomic organization and replication patterns of the completely sequenced Thermococcales.

2. Material and methods

2.1. Genomic data files retrieval and formatting

GenBank genomic data files corresponding to the 21 Thermococcales species were retrieved locally from the NCBI repository

Abbreviations: dif, chromosome dimer resolution site; NCBI, National Center for Biotechnology Information; nt, nucleotide; ORB, origin recognition boxes; *oriC*, origin of replication; PSSM, position-specific scoring matrix; TCA, tricarboxylic acid; RPKM, reads per kilobase per million mapped reads.

* Corresponding author.

http://dx.doi.org/10.1016/j.biochi.2015.07.008

using four sequential commands from NCBI Entrez Programming Utilities (E-Utilities). This redundant procedure was defined in order to guarantee retrieval of the main chromosome of complete genomes exclusively. The first command allows retrieval of the species-specific bioproject:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=bioproject&term=[speciesname]

The second command permits to examine the 'Sequencing_-Status' flag for completeness:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=bioproject&id=[bioproject]

The third command retrieves the unique and chromosome-specific GenBank Identification (GI) number:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nuccore&term=[bioproject]

The fourth command retrieves locally the organism-specific data file in GenBank format:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=[GI]&rettype=gbwithparts

The Thermococcales protein sequences were extracted in Fasta format from these GenBank files using an in-house c# parsing script retaining only the actual amino acid sequence and the unique genomic identification number (GI). All proteins were merged into a single database which was converted to binary format using the NCBI executable 'makeblastdb'. The same script generated a separate indexed file where each individual protein was represented using the following fields: ORF genomic orientation, ORF starting and ending coordinates, gene name, unique protein GI identifier, protein function and source organism name.

2.2. Thermococcales phylogenetic tree

DNA sequence corresponding to the 16S ribosomal RNA genes were retrieved using the BAGET web service at http://archaea.u-psud.fr/bin/baget.dll [7]. PhyML phylogeny was computed using web service http://phylogeny.lirmm.fr/ [8].

2.3. Thermococcales origin of replication prediction

Replication origin predictions with GC skew or Z-curve methods were performed using software Ori-Finder 2 available at http://tubic.tju.edu.cn/Ori-Finder2/ [9]. In a second predictive method, we used the mini-ORBs sequences identified in *Pyrococcus abyssi* by Matsunaga et al. [10] as a matrix for *oriC* prediction using FITBAR available at http://archaea.u-psud.fr/fitbar [11]. In this case, the search algorithm parameters were log-odds PSSM, with a local Markov Model to compute the p-value of the newly predicted ORB site and the investigation was made in intergenic regions only. We have considered as putative replication origin, intergenic regions where more than 4 mini-ORBs can be predicted using FITBAR, with p-values < 0.005. These results were compared to those obtained with Ori-Finder 2 using as ORBs sequences, the three motifs predicted for *Thermococcaceae*. These three conserved motifs of ORBs sequences were obtained from the comparison of Thermococcales replication origin indicated in the DoriC database [12]. The conserved ORB motifs were calculated from the Thermococcales records in DoriC, with the MEME tool (Multiple EM for Motif Elicitation) used to discover conserved patterns in related DNA sequences [13].

2.4. Thermococcales dif site prediction

The identification of *dif* sites on the 21 sequenced Thermococcales chromosomes was performed using a consensus sequence deduced from the alignment of predicted dif sites in *P. abyssi, Pyrococcus horikoshii, P. furiosus* and *Thermococcus kodakaraensis* [14]. This consensus was then used to perform *dif* site prediction using FITBAR with the same search algorithm parameters as described above for ORBs prediction but on the whole chromosome. Progressively, every newly predicted sequence was added to the consensus to improve detection sensitivity.

2.5. Homology searches of XerA recombinase

Thermococcales XerA orthologs were searched by BLASTp analysis using the amino acid sequence of *P. abyssi* XerA (NP_126073.1). A second predictive method was performed using SYNTTAX web service [15] available at http://archaea.u-psud.fr/synttax.

2.6. Core genome procedure

The core genome procedure was conducted as follows. We designed a c# script to construct protein orthologous groups by non-redundant bi-directional BLASTs. Every BLAST score was normalized to the alignment of query and hit proteins to themselves. Proteins showing normalized bi-directional BLASTs $> 30\%$ were considered orthologous as recommended by Lerat et al. [16]. A c# script was designed to query the orthologous groups and define the core genome which consists of all protein genes present at least once in the whole dataset. A 'single core' dataset was derived for this core genome by excluding orthologous classes containing more than a single representative per genome.

2.7. Core genome chromosomal positioning

For each gene composing the single core, we calculated the mean distance to the predicted origin of replication and its standard deviation (SD) using an in house c# script. The core genes were then successively ranked by mean distance and SD to highlight the presence of clusters.

2.8. P. abyssi genome expression

In order to quantify the expression level of every gene in *P. abyssi*, we used RNA-seq data obtained across several growth phases as described in Ref. [17]. As the sequencing was produced in a directed way, the reads alignment respects the strand of the DNA molecule. The CompareOverlapping tool from the S-mart toolbox [18] was used (with the -c option to respect strand constraint) in order to define the number of overlapping reads for every CDS feature defined into the NC_000868.1 entry from the NCBI repository. For each gene, the RPKM measurement defined by Ref. [19] was computed based on the number of overlapping reads, a read size of 40nt, and a total of 5587560 aligned reads. We have used the RPKM measure for each gene as an estimation of their respective expression level.

3. Results

3.1. Thermococcales genomic dataset

At the time of writing, 21 Thermococcales genomes have been completely sequenced and annotated. They are publicly available at the NCBI repository and consist of 13 *Thermococcus*, 7 *Pyrococcus* and 1 *Palaeococcus* (Table 1). Thermococcales carry a single ~2 Mb chromosome and encode an average of 2100 proteins. Evolutionary relationships among the various species are illustrated by a phylogenetic tree of their 16S ribosomal RNA genes (Fig. 1). Genomic sequences were retrieved as described in Materials and Methods. The comparative genomic analysis presented here is based on this entire dataset. The first step of this analysis consisted in the identification of chromosomal landmarks such as the origin and terminus of DNA replication followed in a second step by the comparison of the protein content at the genomic level.

3.2. Prediction of Thermococcales DNA replication origins

The duplication and transmission of genetic information without loss is of fundamental importance for living cells. Cell division must be accompanied by DNA replication executed with appropriate timing and frequency. In all organisms, replication initiates at specific region(s) of the genome known as the origin of replication (*oriC*) site(s). Eukaryotic DNA replication is initiated at multiple origins at different times across linear chromosomes. In eukaryotes, the origin recognition complex (ORC) contains six separate polypeptides, Orc1-6. Comparative genomic analysis of whole archaeal genome sequences show that the archaeal machinery responsible for DNA replication is largely homologous to that of eukaryotes and is clearly distinct from its bacterial counterpart [20,21]. It has been shown experimentally that the archaeal origin binding protein is homologous to the related eukaryotic Orc1 and Cdc6 proteins [22]. The fine mapping of the three replication origins in *Sulfolobus solfataricus* led to the identification of origin recognition boxes (ORBs) and mini-ORBS [23]. ORBs are repeated sequences located on both sides of A/T rich regions and were shown to be the binding site for Cdc6 proteins [23]. ORBs from different species share sequence similarity with a consensus sequence referred to as mini-ORB. It was shown that mini-ORBs are sufficient to bind Cdc6 proteins and that Cdc6 from one organism (*Cdc6-1* of *S. solfataricus*) can bind ORBs from other species *in vitro* (*P. furiosus*, *Halobacterium* NRC1) [23]. ORBs sites are well conserved across many archaeal species and specific binding of ORB sequences by Cdc6 is likely to be a common mechanism for origin recognition in Archaea [22,24–26]. Several archaeal species such as *S. solfataricus, Sulfolobus acidocaldarius*, *Haloferax volcanii* and *Aeropyrum pernix* possess multiple *oriC* per chromosome [23,27–29]. Multiple chromosomal replication origins might have arisen by capture of viral or plasmidic replication origins and their respective associated initiator factor [21]. On the other hand, single origins were found in *Methanothermobacter thermautotrophicus* [24] and mapped precisely in the Thermococcales genus *Pyrococcus* [22,30]. In order to compare our genomic dataset, it was fundamental to identify a common and unique genomic feature shared by all 21 Thermococcales genomes under study. Since the origin of replication was shown to be unique in these genomes, we proceeded with a computational prediction of their respective locations. Several bioinformatics techniques have been used to locate origins of replication in prokaryotic genomes: they are based on the measure of asymmetric nucleotide compositions on leading and lagging strands. Cumulative GC-skew plots are commonly used for this purpose [31–34]. Thermococcales *oriC* for species *P. abyssi*, *P. horikoshii* and *P. furiosus* have been located using other skewed sequences such as GGTT and GGGT [6,30]. However, these two particular skews and the remaining 254 tetranucleotide combinations failed to reliably predict *Thermococcus* origins (data not shown). Alternative scoring methods such as Z-curve calculation have been used successfully for the archaea *Methanocaldococcus jannaschii* and *Methanosarcina mazei*, *Halobacterium* sp. strain NRC-1 and *S. solfataricus* P2 [9]. Cumulative GC skew and Z-curve methods were tested on Thermococcales genomes using the Ori-Finder 2 web service [9], and the results obtained with four representative genomes are shown in Supplemental Fig. S1. Our results show that the cumulative GC skew method fails to locate replication origins in Thermococcales. The Z-curve approach is positive for few genomes such as *P. abyssi* and *T. kodakarensis* but does not provide a prediction for the remaining genomes. Clearly, methods based on Z-curve and DNA composition bias or skew were inoperative for the robust prediction or replication origins in Thermococcales. Therefore, in order to map the position of the replication origins we adopted a different approach based on the systematic detection of biological sequences associated with the initiation of DNA synthesis. As shown above these repeated sequences called ORB are clustered at or near the replication origin and often closely associated with the Cdc6 genes encoding a protein involved in the initiation or replication [10]. All Thermococcales encode a unique Cdc6 gene except *Thermococcus* sp. CL1 which encodes a second putative Cdc6-related protein encoded by gene CL1_0695. Using the published archaeal mini-ORB sequences [10], the web service FITBAR [11] was used to build consensus sequence and detect its occurrences genome wide, as described in Materials and Methods. A unique *oriC* could be detected unambiguously in all Thermococcales from the dataset with a p-value < 0.005 (Table 2 and Suppl. Fig. S2). No putative ORB sequence could be found near the second Cdc6-related gene of *Thermococcus* sp. CL1 and this observation is in agreement with Ori-Finder 2 predictions (data not shown). The association between *oriC* and Cdc6 was found in all genomes except *Thermococcus litoralis* and *Thermococcus sibiricus* where the *oriC*-Cdc6 distance is respectively 453 kb and 349 kb. Synteny analysis using the SYNTTAX web service [15] indicated that in *Thermococcus* and *Palaeococcus* genera, *oriC* is located between Cdc6 and Rad51-ortholog RadA (Suppl. Fig. S3A). Like its bacterial *recA* and eukaryal Rad51 orthologs, RadA in involved not only in double strand break repair but also in DNA replication by rescuing collapsed replication forks [35]. In *Pyrococcus* genus, Cdc6 and *oriC* are also immediately adjacent whereas RadA is not syntenic (Suppl. Fig. S3B). In all cases, the origin of replication is located in extended non-translated regions or overlaps small computer-predicted orphan genes (Suppl. Fig. S3A&B). A prediction of clustered ORB sequences obtained with the FITBAR web service [11] was used to localize *oriCs* as shown in Supplemental Table S1. Our analysis indicates that the most robust *oriC* predictions are those based solely on mini-ORB clusters. The positions of these clusters were therefore considered as *bona fide oriC* (Table 2, column 2). Replication origin positioning was then used as the first common reference to align and orient all genomes in the dataset (Supp. Fig. S2).

3.3. Prediction of Thermococcales DNA replication termination sites

As shown above, the cumulative GC-skew cannot be used reliably to predict the location of *terC* where Thermococcales terminate bidirectional DNA replication. So far, *terC* sites have received much less attention than *oriC*. To our knowledge, neither biological nor sequence data are available to define where replication forks meet. In accordance with the bacterial paradigm, archaeal DNA replication forks are believed to terminate in the vicinity of *dif* sites [14,36]. These *dif* sites are present in a single copy per genome and

Table 1
List of Thermococcales species with a complete genome sequence available.

Species	Bioproject	GI	Genes	Size (Mb)	GC%	Optimum T°C	Habitat	Reference
Palaeococcus pacificus DY20341	PRJNA207495	664800204	2046	1.86	43.0	80 °C	Aquatic	[57]
Pyrococcus abyssi GE5	PRJNA62903	14518450	1875	1.77	44.71	103°C/90 °C	Aquatic	[58]
Pyrococcus furiosus DSM 3638	PRJNA57873	18976372	2225	1.90	40.77	100°C/90 °C	Aquatic	[59]
Pyrococcus furiosus COM1	PRJNA169620	397650687	2113	1.91	40.79	100 °C	Aquatic	[60]
Pyrococcus horikoshii OT3	PRJNA57753	14589963	2000	1.73	41.88	98°C/95 °C	Aquatic	[61]
Pyrococcus sp. NA2	PRJNA66551	332157643	2028	1.86	42.74	93 °C	Aquatic	[62]
Pyrococcus sp. ST04	PRJNA167261	389851449	1839	1.73	42.30	95 °C	Aquatic	[63]
Pyrococcus yayanosii CH1	PRJNA68281	337283511	1952	1.72	51.64	98 °C	Aquatic	[64]
Thermococcus barophilus MP	PRJNA54733	315229765	2257	2.01	41.76	85 °C	Aquatic	[65]
Thermococcus eurythermalis strain A501	PRJNA251677	700302025	2183	2.12	53.47	85 °C	Aquatic	[66]
Thermococcus gammatolerans EJ3	PRJNA59389	240102057	2210	2.05	53.56	88 °C	Aquatic	[67]
Thermococcus guaymasensis DSM11113	PRJNA230529	744793172	2170	1.92	52.86	88 °C	Aquatic	Zhang,X. et al., 2015
Thermococcus kodakarensis KOD1	PRJNA58225	57639935	2358	2.09	52.00	85 °C	Aquatic	[68]
Thermococcus litoralis DSM 5473	PRJNA82997	530547444	2575	2.22	43.09	83 °C	Aquatic	[69]
Thermococcus nautili strain 30-1	PRJNA237737	589908590	2288	1.97	54.84	87.5 °C	Aquatic	[70]
Thermococcus onnurineus NA1	PRJNA59043	212223144	2026	1.85	51.27	80 °C	Terrestrial	[71]
Thermococcus sibiricus MM 739	PRJNA59399	242397997	2107	1.85	40.20	78 °C	Oil	[72]
Thermococcus sp. 4557	PRJNA70841	341581088	2181	2.01	56.08	ND	Aquatic	[73]
Thermococcus sp. AM4	PRJNA54735	350525682	2279	2.08	54.78	80 °C	Aquatic	[74]
Thermococcus sp. CL1	PRJNA168259/PRJNA167371	390960176	2090	1.95	55.82	85 °C	Aquatic	[75]
Thermococcus sp. ES1	PRJNA230233	573023865	2090	1.95	40.30	82 °C	Aquatic	[76]

are used by a Xer-like recombinase to resolve chromosome dimers, a critical step before their segregation into daughter cells [37]. The 28-nt dif site is composed by two inverted repeats of 11 base pairs (each one specific for one of the two Xer recombinase) separated by a central hexanucleotide; the XerCD/*dif* recombination system is widespread in the bacterial domain [38]. The efficiency of the archaeal XerA/*dif* system has been demonstrated *in vitro* [14]. By sequence homology search, XerA orthologs were found in single copy in all Thermococcales (data not shown). In order to identify *dif* sites in our dataset, we followed the same methodology used for *oriC*, as described above. The biological *dif* sites proposed by Cortes et al. [14] were used to build a consensus for genome wide searching using FITBAR [11]. *Bona fide* unique *dif* sites could be identified for 19 genomes out of 21 (Table 2 and Suppl. Fig. S2). The *dif* site position of *Pyrococcus* sp. NA2 and *Pyrococcus* sp. ST04 were estimated to be opposite from their respective predicted *oriC*.

3.4. Core genome

Early chromosomal alignments demonstrated the high level of recombinations and rearrangements in Thermococcales genomes [6]. These observations indicate that these genomes evolve rapidly which might suggest that their genetic content is also highly variable among species. In order to quantify this genomic drift, we submitted our dataset to a recursive systematic comparison of the predicted protein sequences they encode. Each Thermococcales genome encodes an average of 2100 proteins. All the corresponding sequences were compared as described in Material and Methods in order to rank them into orthologous groups. These groups could then be queried to extract common proteins, defined as 'core genome' as well as species-specific or genus-specific proteins and their combinations (Fig. 2). We have used two genetic subsets to define the core: a distinction was made between the 'general core'

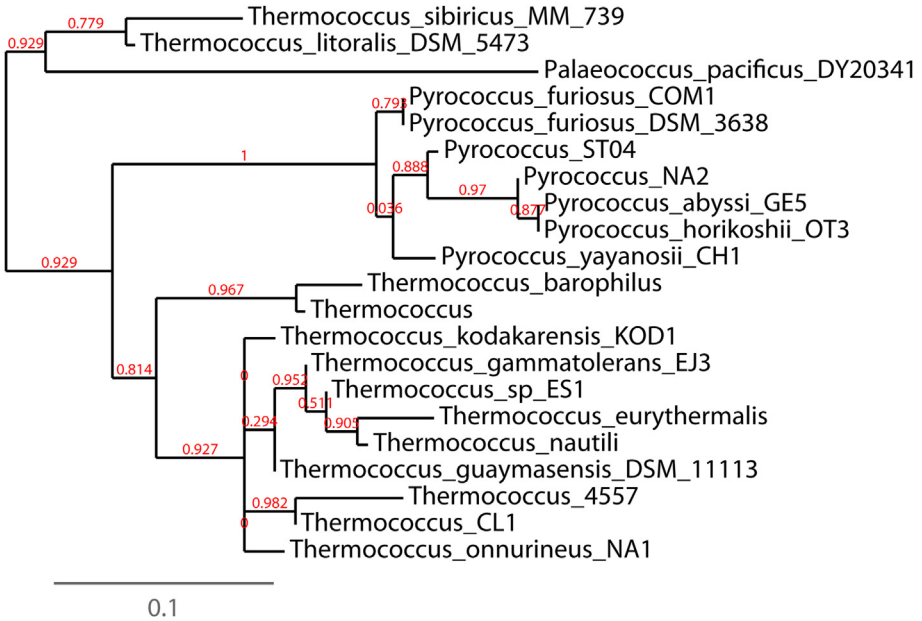

Fig. 1. Phylogenetic tree of the 21 sequenced Thermococcales. The phylogeny of the Thermococcales dataset was calculated with PhyML using the 16S ribosomal RNA genes as described in Material and Methods.

Table 2
Prediction of *oriC* and *dif* in Thermococcales.

Species	Putative *oriC* characteristics		Putative *dif* characteristics				
	Position on chromosome (Orb cluster coord.)	Cdc6 coord.	Sequence (28 bp)			Position on chromosome	Intergenic location
			Left arm	Spacer	Right arm		
Palaeococcus pacificus DY20341	1858353..0	583..1839	TTTGGA**TATAA**	TCAACA	**TTATAT**CTA**AA**	1158048	Yes
Pyrococcus abyssi GE5	122701..123499	121402..122700	ATTGGA**TATAA**	TCGGCC	**TTATAT**CTA**AA**	1220264	Yes
Pyrococcus furiosus DSM 3638	15355..16235	16236..17498	TTAGA**TATAA**	TCAGCC	**TTATAT**CTA**AA**	659548	Yes
Pyrococcus furiosus COM1	1479769..1480649	1478506..1479768	TTAGA**TATAA**	TCAGCC	**TTATAT**CTA**AA**	462638	Yes
Pyrococcus horikoshii OT3	110790..111561	109476..110789	TTAGA**TATAA**	TCAGCC	**TTATAT**CTA**AA**	736581	Yes
Pyrococcus sp. NA2	579324..580109	578064..579323		ND			
Pyrococcus sp. ST04	227904..228761	228762..230021		ND			
Pyrococcus yayanosii CH1	1426398..1427171	1427172..1428431	TTAGA**TATAA**	TGATCC	**TTATAT**CTA**AA**	1058381	Yes
Thermococcus barophilus MP	1672620..1673707	1670448..1671713	TTGTCA**TATAA**	TATGCC	**TTATAT**CTA**AA**	880625	Yes
Thermococcus eurythermalis strain A501	425720..426421	423614..424867	TTAGA**TATAA**	TGTACC	**TTATAT**CTA**AA**	1862025	Yes
Thermococcus gammatolerans EJ3	126739..127591	125431..126738	TTTGGA**TATAA**	TGTACC	**TTATAT**CTA**AA**	1457065	Yes
Thermococcus guaymasensis DSM11113	813701..814368	1594403..1595665	TTAGA**TATAA**	TGTGCC	**TTATAT**CTC**AA**	100930	Yes
Thermococcus kodakarensis KOD1	1711251..1712157	1712158..1713405	TTTTGA**TATAA**	TGTACC	**TTATAT**GAC**AA**	483614	Yes
Thermococcus litoralis DSM 5473	974680..975085	1594403..1595665	TTTGGA**TATAA**	TGTGCC	**TTATAT**GAC**AA**	1867166	No
Thermococcus nautili strain 30-1	1603522..1604207	1605068..1606321	TTGAGA**TATAA**	TGTACC	**TTATAT**CTA**AA**	772784	Yes
Thermococcus onnurineus NA1	1510250..1510926	1508116..1509363	TTAGA**TATAA**	TGTGTC	**TTATAT**CTA**AA**	854799	Yes
Thermococcus sibiricus MM 739	1783451..1784177	1434100..1435362	TTGTCA**TATAA**	TAAGCC	**TTATAT**CTA**AA**	689121	No
Thermococcus sp. 4557	1373703..1374410	1376165..1377412	TTTTCC**TATAA**	TGTGCC	**TTATAT**CTA**AA**	97343	Yes
Thermococcus sp. AM4	1530315..1531266	1529070..1530314	TTTGGA**TATAA**	TGTGCC	**TTATAT**CCA**AA**	849102	Yes
Thermococcus sp. CL1	1018000..1018309	1020367..1021614	TTTGGA**TATAA**	TGTACC	**TTATAT**CCA**AA**	1704316	Yes
Thermococcus sp. ES1	1754560..1755481	1752377..1753639	TTTAGA**TATAA**	TGAATC	**TTATAT**GAC**AA**	1028150	Yes
Thermococcales dif consensus			W**T**KDSM**TATAA**	TVDDYM	**TTATAT**SHM**AA**		

which contains proteins orthologs and paralogs in every genome and a more restrictive 'single core' which regroups only single copy orthologs shared by all genomes. The general core and single core amount to 790 and 668 proteins respectively (Fig. 2 and Suppl. Table S2A&B). A detailed gene list of the 668 core genome is presented in Supplemental Table S3. The same procedure allowed the identification of genus-specific proteins as well. *Pyrococcus* and *Palaeococcus* genera encoded respectively 19 and 116 specific proteins whereas a single *Thermococcus*-specific protein was found. As shown in Table 3, these proteins could be ranked into functional groups as defined in the archaeal clusters of orthologous genes (ArCOGS) [39]. The core genome comprises proteins of the following classes: information storage and processing (32%), metabolism (30%), poorly characterized (27%) and cellular processes and signaling (11%). This high conservation is in sharp contrast with the very limited chromosomal alignment observed to these organisms [6]. Thus it seemed important to analyze whether this genomic conservation would be clustered to particular chromosomal locations.

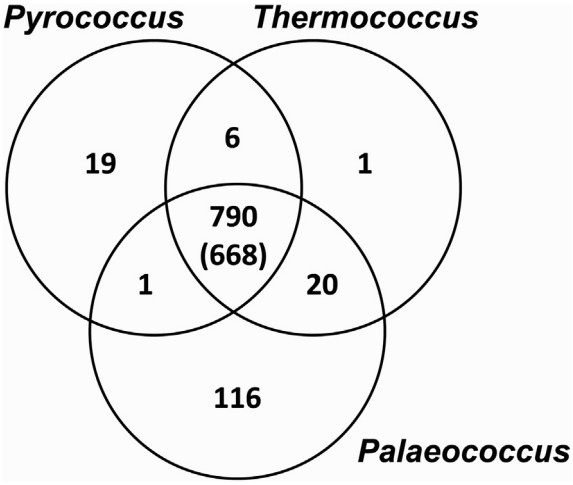

Fig. 2. Venn diagram for core and genus-specific proteins counting. Core, genus-specific proteins and their combinations were computed as described in Materials and Methods.

3.5. Core genome positioning

In Eukarya, genes involved in related and essential functions often cluster on the chromosome and are co-expressed, which correlates with elevated expression rates [40,41]. In Archaea and Bacteria, these genes belong to single transcription units or operons, which provide tight co-regulation in addition to expression polarity [42]. Furthermore, bacterial genomes display a non-random gene organization at a higher level such as macro-domains [43] or with multiple scales [44]. Additional chromosomal structuring involves positioning of essential genes preferentially on the leading strand [45] and clustering of transcription and replication genes in the proximity of the bacterial origin of replication [46]. The archaeal chromosome organization has not been investigated in depth with the exception of a few Crenarcheota. It was shown that *S. solfataricus* and *S. acidocaldarius* are equipped with three origins or replication surrounded by a higher density of core or essential genes; furthermore, these same regions are more highly expressed [36]. These reports prompted us to investigate the genomic architecture of the Euryarchaeota Thermococcales. For each genome in the dataset, we constructed a detailed physical map indicating the position of each gene. We have used our *oriC* and *dif* sites predictions to determine the polarity of each gene respective to the orientation of the replication forks (Fig. 3 and

Table 3
ArCOG assignment of the Thermococcales core genes.

ArCOG class	Function	790 core	668 core
Information storage and processing **32%** (34%)	Translation, ribosomal structure and biogenesis	**149**	140
	RNA processing and modification	**0**	0
	Transcription	**52**	43
	Replication, recombination and repair	**51**	45
	Chromatin structure and dynamics	**0**	0
Cellular processes and signaling **11%** (10%)	Cell cycle control, cell division, chromosome partitioning	**11**	8
	Nuclear structure	**0**	0
	Defense mechanisms	**11**	8
	Signal transduction mechanisms	**5**	4
	Cell wall/membrane/envelope biogenesis	**14**	12
	Cell motility	**7**	5
	Cytoskeleton	**0**	0
	Extracellular structures	**0**	0
	Intracellular trafficking, secretion, and vesicular transport	**8**	8
	Posttranslational modification, protein turnover, chaperones	**31**	22
	Mobilome: prophages, transposons	**0**	0
Metabolism **30%** (27%)	Energy production and conversion	**52**	28
	Carbohydrate transport and metabolism	**33**	30
	Amino acid transport and metabolism	**45**	36
	Nucleotide transport and metabolism	**28**	25
	Coenzyme transport and metabolism	**41**	36
	Lipid transport and metabolism	**12**	12
	Inorganic ion transport and metabolism	**25**	11
	Secondary metabolites biosynthesis, transport and catabolism	**5**	4
Poorly characterized **27%** (29%)	General function prediction only	**128**	115
	Function unknown	**82**	76

Bold numbers in columns 1 & 3 refer to 790 core genes.

Suppl. Fig. S2). These maps could be used to calculate the proportion of genes whose transcription is collinear with the orientation of DNA replication. Out of the 19 genomes where *dif* could be predicted, 16 display a higher proportion of genes encoded on the leading strand (Suppl. Table S4). Plotting of 'single core' genes onto the same circular physical maps indicated an even higher proportion of leading strand-encoded genes for 16 genomes (Suppl. Table S4). Since previous studies have shown that essential *Sulfolobus* genes are clustered near the origin or replication [36], we investigated whether this is the case in Thermococcales as well. We therefore calculated the genomic distance to the respective predicted *oriC* for each single core ortholog (Suppl. Table S3). Computation of their mean distance and standard deviation allowed the definition of 17 genes clusters whose distance to *oriC* remains relatively invariable across species (Table 4). The locations of these clusters for each Thermococcales are shown in Supplemental Fig. S2; they often correlate with GC-skew variations.

3.6. Expression of core genes and conserved gene clusters

Recent experiments have shown that core genes are more strongly expressed in the model organism *E. coli* [47]. It was therefore important to verify this observation in Thermococcales. The next logical step consisted in the analysis of the correlation between gene position and level of gene expression. We have used the pangenomic gene expression data which was measured recently in *P. abyssi* using RNA-seq [17]. As shown in Table 4, the mean expression level of the 17 gene clusters described above indicates that they are more transcribed than single core genes which in turn are also more expressed than non-core genes. The largest clusters 8, 9 and 10 were found to be the most highly expressed; they contain genes encoding RNA polymerase subunits and ribosomal proteins. Remarkably, these clusters are positioned at one-quarter of the genome length suggesting that a high selective pressure is acting to constrain them at this particular favorable location.

3.7. Localization of organism-specific genes

The positioning of the 'single core' on the chromosomal maps revealed, for all genomes, a number or large area devoid of core

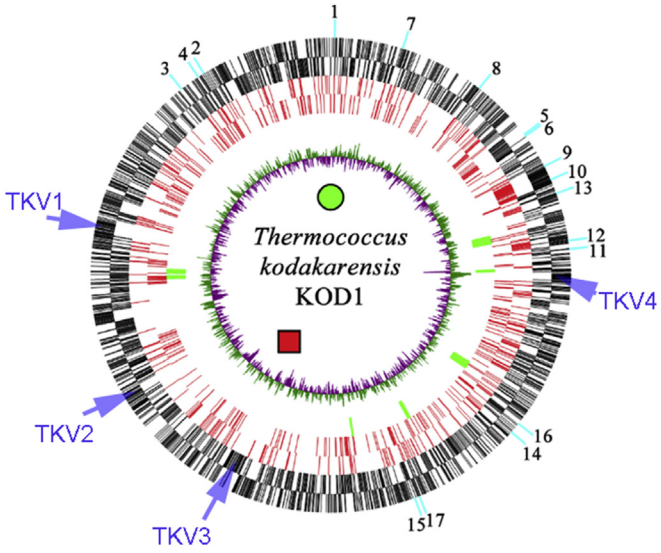

Fig. 3. Graphical correlation between core-free genomic regions and integration of mobile elements in *Thermococcus kodakarensis*. The physical map corresponding to *Thermococcus kodakarensis* was drawn proportionally. The outermost numbered cyan bars indicate the clusters of core genes. Each black bar positions a single gene of the entire genome: the outer bars correspond to genes transcribed in the same polarity as DNA replication; the inner bars refer to the opposite orientation. Similarly, red bars correspond to single 'core genes' with the same orientation convention as above. Bright green bars indicate the location of clusters of species-specific genes (integrated mobile elements). Purple and green bars correspond to GC skew values calculated in windows of 1000bp, shifted 500bp with the purple and green bars indicating values below and above average genomic GC skew, respectively. Predicted origins of replication and *dif* sites are show as green circles and red squares, respectively. The positions of the four integrated elements (TKV1 to TKV4) as well as the predicted dark matter islands are represented in blue color.

Table 4
Thermococcales conserved clusters characteristics.

Cluster	oriC distance		Number of genes	Mean expression level pangenomic: 668.5 single core: 896.7 clusters: 1978.8	Relevant encoded protein(s)
	Mean (%)	Standard deviation (%)			
01	0.33	0.44	3	478.9	Hypothetical
02	2.69	1.91	2	221.1	Molybdopterin converting factor, subunit 2
03	5.17	3.42	2	2551.7	Hypothetical
04	5.39	3.23	3	557.2	KEOPS complex KAE1
05	7.36	4.34	7	877.6	V-type ATP synthase, 7 subunits
06	8.25	3.41	3	268.2	Preprotein translocase
07	9.14	4.67	2	357.5	Oligopeptide transporters
08	12.94	5.18	5	2926.0	RNA polymerase
09	17.76	3.90	27	3626.6	Ribsosomal proteins
10	20.89	3.63	10	2234.8	Ribosomal proteins – RNA polymerase
11	22.40	5.77	5	482.4	Thymidylate kinase
12	23.46	4.47	3	1011.2	DNA primase
13	24.62	5.45	3	234.9	Mevalonate kinase
14	26.50	5.92	7	1535.2	Ribosomal proteins - RNA polymerase
15	33.34	6.01	2	486.7	Glutamyl-tRNA(Gln) amidotransferase
16	34.14	5.44	2	840.6	Translation initiation factor IF-2
17	38.58	5.63	2	1685.0	Ribosomal protein

genes (Fig. 3 and Suppl. Fig. S2). We observed that clusters containing 3 or more species-specific genes could overlap these blank regions. Since species-specific clusters correspond very likely to the integration of mobile elements such as plasmids or viruses, we can extrapolate the nature of these blank regions as being integrated mobile elements shared by several genomes. Contrarily to what was observed in Sulfolobales [48], the integration of mobile elements in Thermococcales is not confined to a specific location and seems to occur randomly on the chromosome (Suppl. Fig. S2). To confirm this observation, we have mapped on the *T. kodakarensis* genomic map the four known integrated elements (TKV1 to TKV4) [49] and predicted dark matter islands [50]; all are located in core-free regions (Fig. 3).

4. Discussion

With the exception of three methanogens, all archaeal genomes sequenced to date encode at least one Cdc6/Orc1 protein which initiates chromosomal DNA replication at one or more *oriC* origins [51].

In most prokaryotes including several Archaea, chromosomal *oriCs* can be predicted on the basis of DNA composition using GC-skew [52] or Z-curve algorithms [53]. The comparative genomics analysis presented here confirms the initial observation that Thermococcales chromosomes are highly rearranged. In these genomes, DNA sequence scrambling has reached such a high level that commonly observed prokaryotic chromosomal landmarks such as *oriC* and *terC* are no longer readily identifiable by measuring DNA composition biases. It was indeed reported that pure *in silico* approaches can be unreliable due to frequent genome rearrangements [54]. Nevertheless, the regions corresponding to the origin and termination of replication could be predicted by the means of biological sequence sites determined either biochemically or by analogy to bacterial systems. In most Archaea, replication initiates at ORB sites specifically recognized and bound by Cdc6 [22]. Using the well documented ORB sequences [10], unique origins of replication could be predicted unambiguously for all 21 genomes. They are located in close proximity to RadA which corresponds also to the genomic context of Cdc6 in 19 genomes out of 21. The chromosomal location of *terC* was identified by the means of the XerC binding site (*dif*) as defined by Cortez et al. [14]. A unique corresponding site could be identified with high confidence in 19 genomes out of 21. The locations of *oriC* and *dif* in each genome define the respective replichores which appear asymmetrical in most Thermococcales and extremely asymmetrical in *Pyrococcus yayanosii*. This observation raises the question whether *terC* and *dif* are co-localized. By analogy to bacterial systems, it is commonly accepted that DNA replication termination and *dif* sites coincide [14,36]. On the other hand, an extensive computational analysis based on bacterial genomes has shown a lack of correlation between *dif* position and the degree of GC skew suggesting that replication termination does not occur strictly at *dif* sites [55]. However it is quite difficult to extrapolate replication features between Archaea and Bacteria since they use such different replication proteins. Recent evidence has shown that in the Crenarchaeota *S. solfataricus*, replication termination and dimer resolution are temporally and spatially distinct processes [56]. Since this organism carries three functional *oriCs* whereas a single one is found in Thermoccocales, it is once again difficult to transpose replication features across archaeal phyla. In the absence of experimental data and of a functional cumulative GC skew in Thermococcales, we cannot prove nor disprove that *terC* and *dif* positions are distinct.

To assess whether the observed genomic rearrangement could be reflected at the protein level as well, we conducted an extensive ranking of each protein into orthologous groups using a discriminant threshold of 30% similarity. This procedure permitted to characterize the core genome of Thermococcales as well as genus- and species-specific proteins. The 21 genomes considered here share 790 orthologs which corresponds to ~40% of their total proteins. From the core genome, we isolated the subset of proteins found only once per genome. The genes encoding these 668 'single core' proteins were plotted onto circular chromosome maps which revealed several interesting features. First, the 'single core' genes are not evenly distributed along the chromosome: a number of very extensive areas without core genes are readily observable in all 21 genomes. This phenomenon can be interpreted as the result of recent acquisitions of (non essential) genetic information through horizontal transfer. In a further analysis we were indeed able to show that clusters of strain-specific genes, which correspond presumably to integrated mobile elements, are precisely located within these regions. A second feature consists in the conservation of clusters of core genes in particular location of the chromosome, across Thermococcales. A series of 17 clusters could be identified with a standard deviation of mean distance to origin $\leq$6%. Despite a high level of genomic rearrangements, the absolute distance between these clusters and the origin of replication remains

remarkably constant. These clusters are not confined to *oriC*-proximal regions but are scattered along the entire chromosome. It is interesting to note that the individual clusters do not belong to the same replichore in every organism; however, their distance to *oriC* is maintained in a mirrored fashion. The size of each cluster is variable and ranges from 2 to 27 genes often expressed in operons. The largest clusters group essential genes involved in protein translation (cluster 9, 27 genes), gene transcription and protein translation (cluster 10, 10 genes; cluster 14, 7 genes) and energy metabolism (cluster 5, 7 genes). A third feature of the 'single core' consists in its enrichment of genes encoded on the leading strand. This is particularly true with the largest clusters for which a net variation in GC skew is also readily apparent and is very likely to reflect a gene orientation bias of the genes composing the clusters. Indeed, we computed that in 16 organisms out of 19, the core genome is enriched in genes expressed in the same orientation as DNA replication. We were able to show that most of the large clusters display a significantly higher expression rate which further correlates conserved gene position with essential biological functions. The positional conservation of essential genomic subregions is found in the three domains of life [40–42]. This work has shown that this property is particularly relevant in Archaea Thermococcales due to the highly level of rearrangements of their chromosomes. These small and heavily scrambled genomes were able to maintain highly expressed key genes in the most favorable chromosomal positions and transcribe them in a polarity compatible with DNA replication. We would like to hypothesize that genome shuffling is instrumental to better adapt to challenging extreme environments.

5. Conclusion

5.1. Evolution considerations

All the above observations indicate that a remarkable degree of 'order' has been maintained across Thermococcales even if they display highly scrambled chromosomes. Nevertheless, these organisms display an astonishingly short cell cycle in extreme and resource-deficient environments. This apparent paradox motivated our analysis. The data we presented here led us to propose that Thermococcales chromosome shuffling introduces an increased genome variability which is being actively used by natural selection: (1) to maintain highly expressed key essential genes in favorable and invariant chromosomal positions (2) continuously adapt and optimize the positioning of the constant flow of new genes acquired by horizontal transfer, in order to allow allopatric speciation. The molecular mechanism by which Thermococcales rearrange their chromosomes is presently being investigated.

Acknowledgements

This work was funded by the European Research Council under the European Union's Seventh Framework Program (FP/2007-2013)/Project EVOMOBIL - ERC Grant Agreement no. 340440.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.biochi.2015.07.008.

References

[1] L. Achenbach-Richter, R. Gupta, W. Zillig, C.R. Woese, Rooting the archaebacterial tree: the pivotal role of *Thermococcus celer* in archaebacterial evolution, Syst. Appl. Microbiol. 10 (1988) 231–240.
[2] C. Brasen, D. Esser, B. Rauch, B. Siebers, Carbohydrate metabolism in archaea: current insights into unusual enzymes and pathways and their regulation, Microbiol. Mol. Biol. Rev. MMBR 78 (2014) 89–175.
[3] T. Oku, K. Ishikawa, Analysis of the hyperthermophilic chitinase from *Pyrococcus furiosus*: activity toward crystalline chitin, Biosci. Biotechnol. Biochem. 70 (2006) 1696–1701.
[4] A. Gorlas, K. Alain, N. Bienvenu, C. Geslin, *Thermococcus prieurii* sp. nov., a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent, Int. J. Syst. Evol. Microbiol. 63 (2013) 2920–2926.
[5] G. Sezonov, D. Joseleau-Petit, R. D'Ari, *Escherichia coli* physiology in Luria-Bertani broth, J. Bacteriol. 189 (2007) 8746–8749.
[6] Y. Zivanovic, P. Lopez, H. Philippe, P. Forterre, Pyrococcus genome comparison evidences chromosome shuffling-driven evolution, Nucleic Acids Res. 30 (2002) 1902–1910.
[7] J. Oberto, BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence, Bioinformatics 24 (2008) 424–425.
[8] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.M. Claverie, O. Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Res. 36 (2008) W465–W469.
[9] H. Luo, C.T. Zhang, F. Gao, Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes, Front. Microbiol. 5 (2014) 482.
[10] F. Matsunaga, A. Glatigny, M.H. Mucchielli-Giorgi, N. Agier, H. Delacroix, L. Marisa, P. Durosay, Y. Ishino, L. Aggerbeck, P. Forterre, Genomewide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in Pyrococcus sp, Nucleic Acids Res. 35 (2007) 3214–3222.
[11] J. Oberto, FITBAR: a web tool for the robust prediction of prokaryotic regulons, BMC Bioinform. 11 (2010) 554.
[12] F. Gao, H. Luo, C.T. Zhang, DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes, Nucleic Acids Res. 41 (2013) D90–D93.
[13] T.L. Bailey, N. Williams, C. Misleh, W.W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, Nucleic Acids Res. 34 (2006) W369–W373.
[14] D. Cortez, S. Quevillon-Cheruel, S. Gribaldo, N. Desnoues, G. Sezonov, P. Forterre, M.C. Serre, Evidence for a Xer/dif system for chromosome resolution in archaea, PLoS Genet. 6 (2010) e1001166.
[15] J. Oberto, SyntTax: a web server linking synteny to prokaryotic taxonomy, BMC Bioinform. 14 (2013) 4.
[16] E. Lerat, V. Daubin, N.A. Moran, From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria, PLoS Biol. 1 (2003) E19.
[17] C. Toffano-Nioche, A. Ott, E. Crozat, A.N. Nguyen, M. Zytnicki, F. Leclerc, P. Forterre, P. Bouloc, D. Gautheret, RNA at 92 degrees C: the non-coding transcriptome of the hyperthermophilic archaeon *Pyrococcus abyssi*, RNA Biol. 10 (2013) 1211–1220.
[18] M. Zytnicki, H. Quesneville, S-MART, a software toolbox to aid RNA-seq data analysis, PloS one 6 (2011) e25988.
[19] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-seq, Nat. Methods 5 (2008) 621–628.
[20] D.R. Edgell, W.F. Doolittle, Archaea and the origin(s) of DNA replication proteins, Cell 89 (1997) 995–998.
[21] K. Raymann, P. Forterre, C. Brochier-Armanet, S. Gribaldo, Global phylogenomic analysis disentangles the complex evolutionary history of DNA replication in archaea, Genome Biol. Evol. 6 (2014) 192–212.
[22] F. Matsunaga, P. Forterre, Y. Ishino, H. Myllykallio, In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin, Proc. Natl. Acad. Sci. U. S. A. 98 (2001) 11152–11157.
[23] N.P. Robinson, I. Dionne, M. Lundgren, V.L. Marsh, R. Bernander, S.D. Bell, Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*, Cell 116 (2004) 25–38.
[24] A.I. Majernik, J.P. Chong, A conserved mechanism for replication origin recognition and binding in archaea, Biochem. J. 409 (2008) 511–518.
[25] K.K. Ojha, D. Swati, Mapping of origin of replication in Themococcales, Bioinformation 5 (2010) 213–218.
[26] Z. Wu, H. Liu, J. Liu, X. Liu, H. Xiang, Diversity and evolution of multiple orc/cdc6-adjacent replication origins in haloarchaea, BMC Genomic 13 (2012) 478.
[27] M. Lundgren, A. Andersson, L. Chen, P. Nilsson, R. Bernander, Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 7046–7051.
[28] C. Norais, M. Hawkins, A.L. Hartman, J.A. Eisen, H. Myllykallio, T. Allers, Genetic and physical mapping of DNA replication origins in *Haloferax volcanii*, PLoS Genet. 3 (2007) e77.
[29] N.P. Robinson, S.D. Bell, Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 5806–5811.
[30] H. Myllykallio, P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe, P. Forterre, Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon, Science 288 (2000) 2212–2215.
[31] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998) 2286–2290.
[32] P. Lopez, P. Forterre, H. le Guyader, H. Philippe, Origin of replication of *Thermotoga maritima*, Trends Genet. TIG 16 (2000) 59–60.
[33] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.

[34] J.R. Lobry, A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria, Biochimie 78 (1996) 323–326.
[35] S. Haldenby, M.F. White, T. Allers, RecA family proteins in archaea: RadA and its cousins, Biochem. Soc. Trans. 37 (2009) 102–107.
[36] A.F. Andersson, E.A. Pelve, S. Lindeberg, M. Lundgren, P. Nilsson, R. Bernander, Replication-biased genome organisation in the crenarchaeon Sulfolobus, BMC Genomics 11 (2010) 454.
[37] A.C. Lindas, R. Bernander, The cell cycle of archaea, Nat. Rev. Microbiol. 11 (2013) 627–638.
[38] C. Carnoy, C.A. Roten, The dif/Xer recombination systems in proteobacteria, PloS One 4 (2009) e6531.
[39] Y.I. Wolf, K.S. Makarova, N. Yutin, E.V. Koonin, Updated clusters of orthologous genes for Archaea: a complex ancestor of the Archaea and the byways of horizontal gene transfer, Biol. Direct 7 (2012) 46.
[40] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, Nat. Genet. 31 (2002) 180–183.
[41] E.J. Williams, D.J. Bowles, Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*, Genome Res. 14 (2004) 1060–1067.
[42] C. Pal, L.D. Hurst, Evidence against the selfish operon theory, Trends Genet. TIG 20 (2004) 232–234.
[43] E. Esnault, M. Valens, O. Espeli, F. Boccard, Chromosome structuring limits genome plasticity in *Escherichia coli*, PLoS Genet. 3 (2007) e226.
[44] T.E. Allen, N.D. Price, A.R. Joyce, B.O. Palsson, Long-range periodic patterns in microbial genomes indicate significant multi-scale chromosomal organization, PLoS Comput. Biol. 2 (2006) e2.
[45] E.P. Rocha, A. Danchin, Essentiality, not expressiveness, drives gene-strand bias in bacteria, Nat. Genet. 34 (2003) 377–378.
[46] E. Couturier, E.P. Rocha, Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes, Mol. Microbiol. 59 (2006) 1506–1518.
[47] M. Vital, B.L. Chai, B. Ostman, J. Cole, K.T. Konstantinidis, J.M. Tiedje, Gene expression analysis of *E. coli* strains provides insights into the role of gene regulation in diversification, Isme J. 9 (2015) 1130–1140.
[48] M.L. Reno, N.L. Held, C.J. Fields, P.V. Burke, R.J. Whitaker, Biogeography of the *Sulfolobus islandicus* pan-genome, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 8605–8610.
[49] M. Krupovic, D.H. Bamford, Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota, Virology 375 (2008) 292–300.
[50] K.S. Makarova, Y.I. Wolf, P. Forterre, D. Prangishvili, M. Krupovic, E.V. Koonin, Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes, Extremophiles 18 (2014) 877–893.
[51] E.R. Barry, S.D. Bell, DNA replication in the archaea, Microbiol. Mol. Biol. R. 70 (2006) 876.
[52] A. Grigoriev, Analyzing genomes with cumulative skew diagrams, Nucleic Acids Res. 26 (1998) 2286–2290.
[53] C.T. Zhang, R. Zhang, H.Y. Ou, The Z curve database: a graphic representation of genome sequences, Bioinformatics 19 (2003) 593–599.
[54] P. Lopez, H. Philippe, H. Myllykallio, P. Forterre, Identification of putative chromosomal origins of replication in Archaea, Mol. Microbiol. 32 (1999) 883–886.
[55] N. Kono, K. Arakawa, M. Tomita, Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes, BMC Genomics 12 (2011) 19.
[56] I.G. Duggin, N. Dubarry, S.D. Bell, Replication termination and chromosome dimer resolution in the archaeon *Sulfolobus solfataricus*, EMBO J. 30 (2011) 145–153.
[57] X. Zeng, X. Zhang, L. Jiang, K. Alain, M. Jebbar, Z. Shao, *Palaeococcus pacificus* sp. nov., an archaeon from deep-sea hydrothermal sediment, Int. J. Syst. Evol. Microbiol. 63 (2013) 2155–2159.
[58] G.N. Cohen, V. Barbe, D. Flament, M. Galperin, R. Heilig, O. Lecompte, O. Poch, D. Prieur, J. Querellou, R. Ripp, J.C. Thierry, J. Van der Oost, J. Weissenbach, Y. Zivanovic, P. Forterre, An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*, Mol. Microbiol. 47 (2003) 1495–1512.
[59] F.T. Robb, D.L. Maeder, J.R. Brown, J. DiRuggiero, M.D. Stump, R.K. Yeh, R.B. Weiss, D.M. Dunn, Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology, Methods Enzym. 330 (2001) 134–157.
[60] S.L. Bridger, W.A. Lancaster, F.L. Poole 2nd, G.J. Schut, M.W. Adams, Genome sequencing of a genetically tractable *Pyrococcus furiosus* strain reveals a highly dynamic genome, J. Bacteriol. 194 (2012) 4097–4106.
[61] Y. Kawarabayasi, M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, H. Kikuchi, Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement), DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes 5 (1998) 147–155.
[62] H.S. Lee, S.S. Bae, M.S. Kim, K.K. Kwon, S.G. Kang, J.H. Lee, Complete genome sequence of hyperthermophilic Pyrococcus sp. strain NA2, isolated from a deep-sea hydrothermal vent area, J. Bacteriol. 193 (2011) 3666–3667.
[63] J.H. Jung, J.H. Lee, J.F. Holden, D.H. Seo, H. Shin, H.Y. Kim, W. Kim, S. Ryu, C.S. Park, Complete genome sequence of the hyperthermophilic archaeon Pyrococcus sp. strain ST04, isolated from a deep-sea hydrothermal sulfide chimney on the Juan de Fuca Ridge, J. Bacteriol. 194 (2012) 4434–4435.
[64] X. Jun, L. Lupeng, X. Minjuan, P. Oger, W. Fengping, M. Jebbar, X. Xiang, Complete genome sequence of the obligate piezophilic hyperthermophilic archaeon *Pyrococcus yayanosii* CH1, J. Bacteriol. 193 (2011) 4297–4298.
[65] P. Vannier, V.T. Marteinsson, O.H. Fridjonsson, P. Oger, M. Jebbar, Complete genome sequence of the hyperthermophilic, piezophilic, heterotrophic, and carboxydotrophic archaeon *Thermococcus barophilus* MP, J. Bacteriol. 193 (2011) 1481–1482.
[66] W. Zhao, X. Xiao, Complete genome sequence of *Thermococcus eurythermalis* A501, a conditional piezophilic hyperthermophilic archaeon with a wide temperature range, isolated from an oil-immersed deep-sea hydrothermal chimney on Guaymas Basin, J. Biotechnol. 193 (2015) 14–15.
[67] Y. Zivanovic, J. Armengaud, A. Lagorce, C. Leplat, P. Guerin, M. Dutertre, V. Anthouard, P. Forterre, P. Wincker, F. Confalonieri, Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea, Genome Biol. 10 (2009) R70.
[68] T. Fukui, H. Atomi, T. Kanai, R. Matsumi, S. Fujiwara, T. Imanaka, Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with Pyrococcus genomes, Genome Res. 15 (2005) 352–363.
[69] A.F. Gardner, S. Kumar, F.B. Perler, Genome sequence of the model hyperthermophilic archaeon *Thermococcus litoralis* NS-C, J. Bacteriol. 194 (2012) 2375–2376.
[70] J. Oberto, M. Gaudin, M. Cossu, A. Gorlas, A. Slesarev, E. Marguet, P. Forterre, Genome sequence of a hyperthermophilic archaeon, Thermococcus nautili 30–1, that produces viral vesicles, Genome Announc. 2 (2014).
[71] H.S. Lee, S.G. Kang, S.S. Bae, J.K. Lim, Y. Cho, Y.J. Kim, J.H. Jeon, S.S. Cha, K.K. Kwon, H.T. Kim, C.J. Park, H.W. Lee, S.I. Kim, J. Chun, R.R. Colwell, S.J. Kim, J.H. Lee, The complete genome sequence of *Thermococcus onnurineus* NA1 reveals a mixed heterotrophic and carboxydotrophic metabolism, J. Bacteriol. 190 (2008) 7491–7499.
[72] A.V. Mardanov, N.V. Ravin, V.A. Svetlitchnyi, A.V. Beletsky, M.L. Miroshnichenko, E.A. Bonch-Osmolovskaya, K.G. Skryabin, Metabolic versatility and indigenous origin of the archaeon *Thermococcus sibiricus*, isolated from a siberian oil reservoir, as revealed by genome analysis, Appl. Environ. Microbiol. 75 (2009) 4580–4588.
[73] X. Wang, Z. Gao, X. Xu, L. Ruan, Complete genome sequence of Thermococcus sp. strain 4557, a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent area, J. Bacteriol. 193 (2011) 5544–5545.
[74] P. Oger, T.G. Sokolova, D.A. Kozhevnikova, N.A. Chernyh, D.H. Bartlett, E.A. Bonch-Osmolovskaya, A.V. Lebedinsky, Complete genome sequence of the hyperthermophilic archaeon Thermococcus sp. strain AM4, capable of organotrophic growth and growth at the expense of hydrogenogenic or sulfidogenic oxidation of carbon monoxide, J. Bacteriol. 193 (2011) 7019–7020.
[75] J.H. Jung, J.F. Holden, D.H. Seo, K.H. Park, H. Shin, S. Ryu, J.H. Lee, C.S. Park, Complete genome sequence of the hyperthermophilic archaeon Thermococcus sp. strain CL1, isolated from a Paralvinella sp. polychaete worm collected from a hydrothermal vent, J. Bacteriol. 194 (2012) 4769–4770.
[76] S.A. Hensley, J.H. Jung, C.S. Park, J.F. Holden, Thermococcus paralvinellae sp. nov. and Thermococcus cleftensis sp. nov. of hyperthermophilic heterotrophs from deep-sea hydrothermal vents, Int. J. Syst. Evol. Microbiol. 64 (2014) 3655–3659.