Biochimie xxx (2015) 1–9

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 xxx

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.

Please cite this article in press as: M. Cossu, et al., Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes, Biochimie (2015), http://dx.doi.org/10.1016/j.biochi.2015.07.008