Supplementary Data

WASPS: Web Assisted Symbolic Plasmid Synteny Server

Catherine BADEL[1], Violette DA CUNHA[1], Ryan CATCHPOLE1[1], Patrick FORTERRE[1,2] & Jacques OBERTO[1,3]

[1]Institute for Integrative Biology of the Cell (I2BC), Microbiology Department, CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

[2]Institut Pasteur, Unité de Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, 25 rue du Docteur Roux, 75015 Paris, France

Contents

1. WASPS Database structure, generation and update pipeline

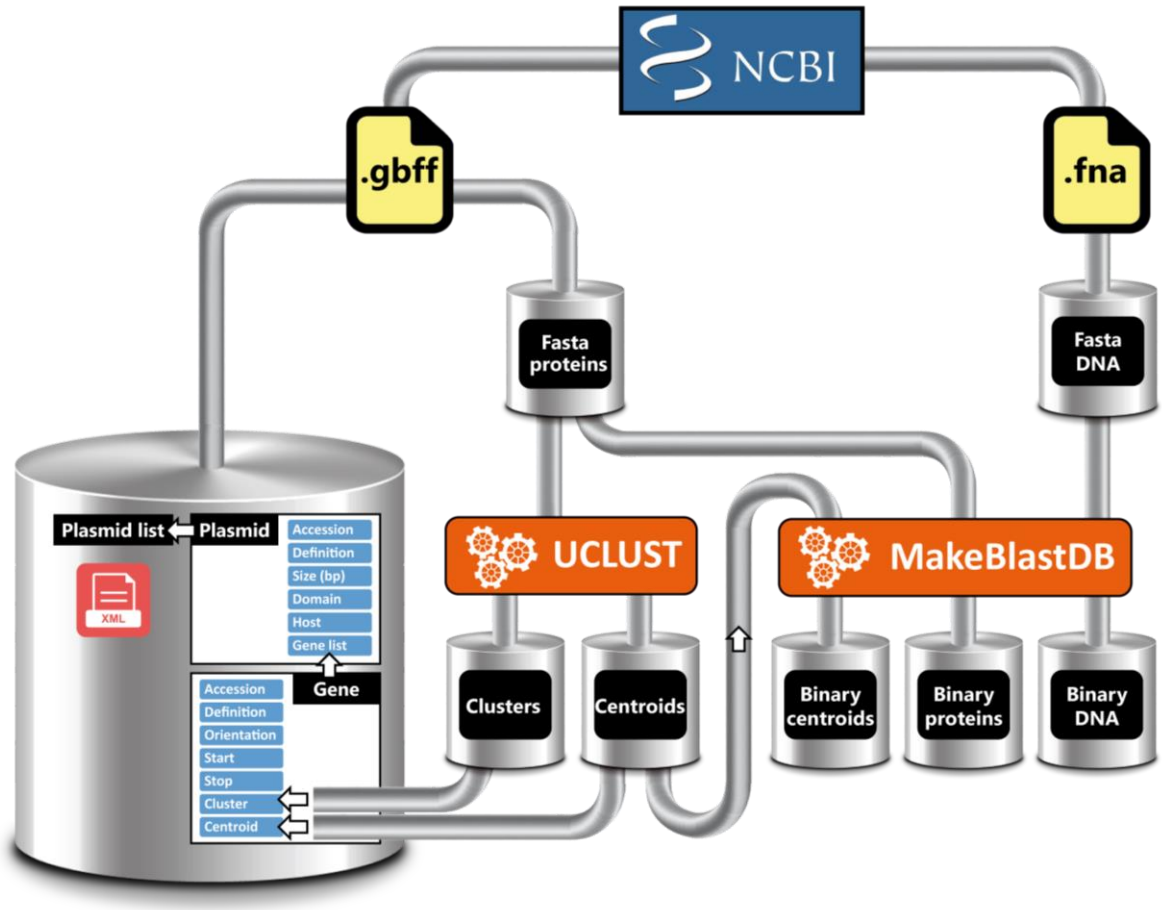
Natural plasmid data (RefSeq) is collected from the FTP site of the NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid) in binary form and processed locally on the server to generate the WASPS Database. At this stage, only the RefSeq plasmid releases are included in the WASPS Database for the reason that they constitute a non-redundant and well-annotated set of sequences, updated at regular intervals. The WASPS Database is updated by the WASPS Updater application in order to follow NCBI RefSeq refreshes. This process involves the addition of new entries which receive a two part GenBank identifier (accession.version). Updates of pre-existing entries keep the same accession number and will increment the version number by one unit. These 'accession.version' identifiers univocally describe both plasmid and gene entries. These identifiers are used to link the different data bins composing the WASPS relational database. The central part of the WASPS database consists of a single XML file containing a sequential list of plasmids exposing their relevant fields. Each plasmid contains a gene list field to store relevant genetic data (Suppl. Fig.1). Each protein in the WASPS database is therefore identified by double 'accession.version' under the format 'gene_accession.version=plasmid_accession.version'. Plasmid DNA sequences and protein sequences are stored in separate bins but intimately linked to the central XML using the 'accession.version' identifiers. All fields and DNA or protein sequences are extracted or

[3] Corresponding author : jacques.oberto@i2bc.paris-saclay.fr

parsed from the downloaded NCBI GenBank and DNA Fasta files. Protein orthology relationships are determined using UCLUST and injected appropriately in the XML file. The UCLUST orthology parameter used by WASPS amounts to 0.35 and is slightly lower than the recommended values (Edgar, 2010). This particular value was chosen since it empirically corresponds to the orthology threshold of 30% similarity proposed by (Lerat, et al., 2003)(see also Section 3, below) which is calculated as follows for proteins A and B:

$$BLAST_{bits(A\times B)} \times \frac{100}{BLAST_{bits(A\times A)}} \geq 30 \qquad \text{(Equation 1)}$$

The cluster centroids calculated with UCLUST are collected separately into an additional bin. The text bins containing DNA, total proteins and centroid proteins sequences are then converted to binary format in order to be efficiently queried by DIAMOND, BlastN, BlastP, TBlastN or PsiBlast. The database compilation is optimized and fully automated to allow frequent updates and ensure exhaustiveness of the analyses. The WASPS Updater pipeline is shown in Supplemental Figure 1.

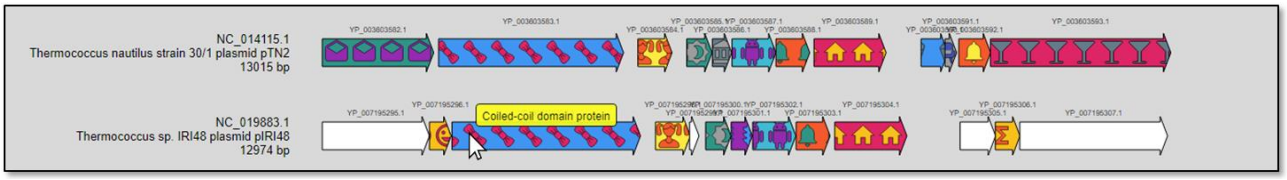


Supplemental figure 1. WASPS database structure and update pipeline. *The WASPS update process is fully automated and optimized for low CPU cycles. Due to its structure, the database is completely regenerated at each update to allow increased robustness and accuracy. Flux directionality is top-down except where noted. All metallic bins are qweryable in WASPS. Binary bins are used specifically by BLAST and DIAMOND.*

2. WASPS client-side user interface

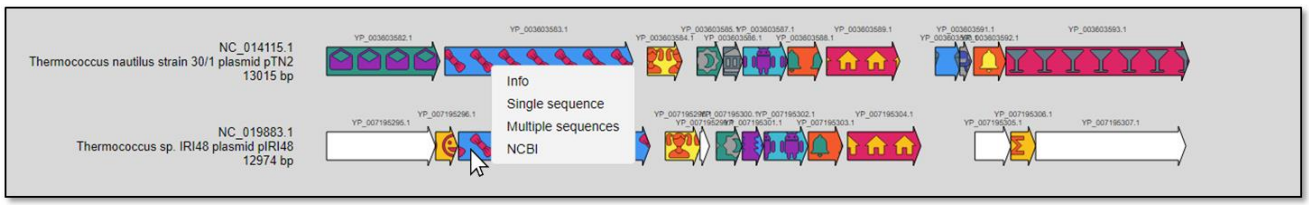
WASPS Synteny Map Interface has been developed to allow maximal user-interactivity. User interactivity is achieved by the means of a 'three button wheel mouse', a standard equipment for most modern desktop computers. Equivalent gestures are available for laptops, trackpads or touch screen devices and are provided by the respective operating systems. 2D synteny maps can be smoothly panned and zoomed directly in the web browser without requiring data transfer from or to the server.

- ***Pan.*** The Synteny Map Interface can be panned by holding down the left mouse button.
- ***Zoom.*** The Synteny Map Interface can be zoomed by rotating the mouse wheel.
- ***Hovering.*** Context-sensitive information is available for each displayed gene in the synteny maps. Mouse hovering on a specific gene will present its definition in a yellow tooltip (Suppl. Fig. 2).



Supplemental figure 2. *The hovering tooltip appears in yellow color.*

- ***Context menu.*** Right clicking on a specific gene will open a context menu with four options (Suppl. Fig. 3):
  i) Info : protein gene accession, plasmid accession, protein definition and protein cluster.
  ii) Single sequence : protein sequence of the highlighted gene in Fasta format.
  iii) Multiple sequences : all protein sequences of the WASPS cluster related to the highlighted gene.
  iv) NCBI : external link to the protein (in GenBank format) at the NCBI.



Supplemental figure 3. *The context menu appears in white background upon right mouse keypress.*

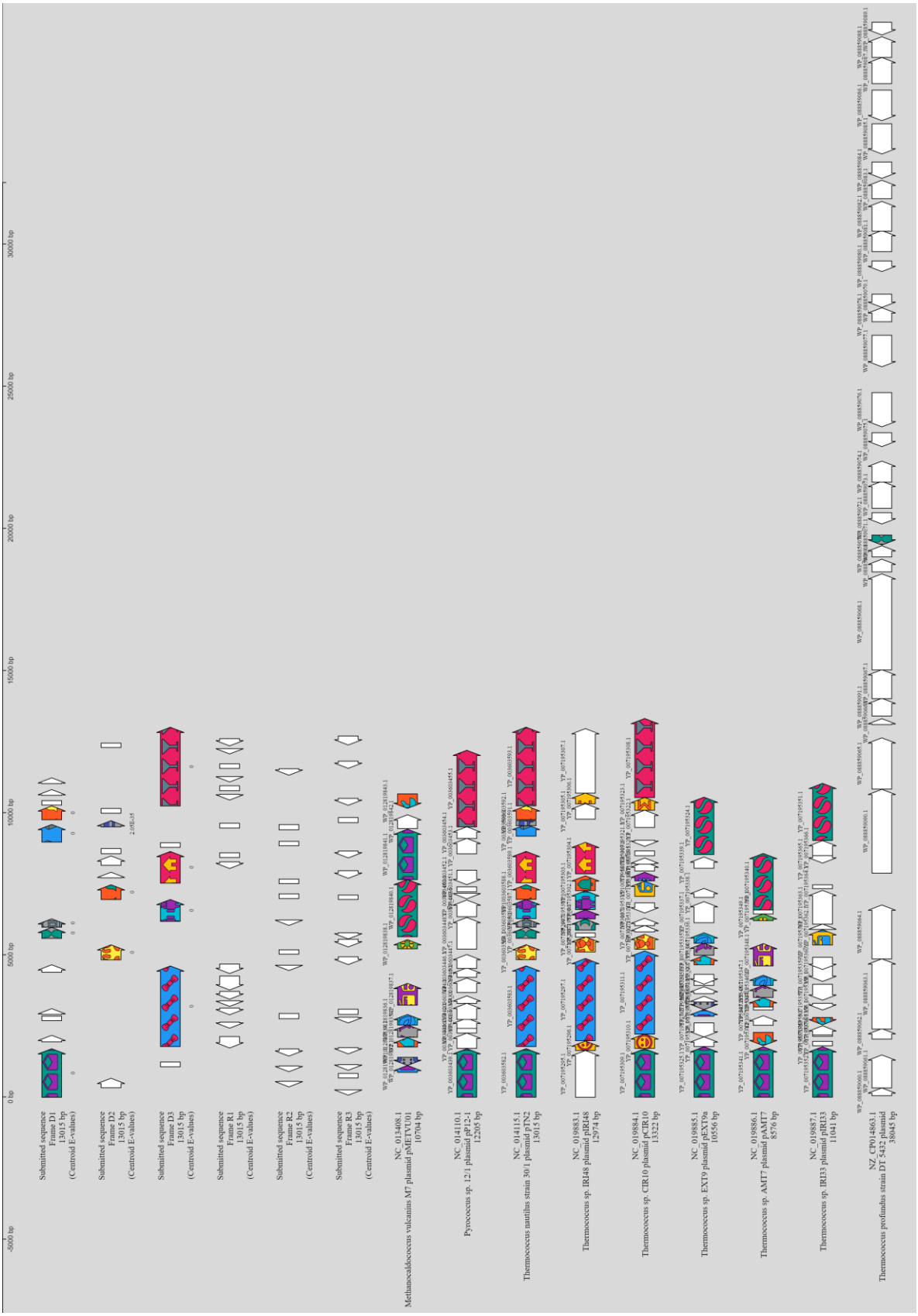
3. Plasmid pTN2 synteny map and assessment of the prediction quality

The sequence of plasmid pTN2 from *T. nautili* was submitted to the WASPS web service for synteny analysis using the parsimonious option 'Primary Hits Only' to limit the search to the best match for each protein using the fast DIAMOND algorithm. To obtain the results shown in Figure 1 (main article), WASPS internally translated the query plasmid sequence in the six possible reading frames using ATG, GTG or TTG as start codons and TAG, TAA or TGA as stop codons. All predicted open reading frames (ORFs) were retained without size limitation. Internal ORFS in the same reading frame of large predicted genes were not considered. The resulting 51 predicted proteins were then compared to the centroid database using DIAMOND. All matching translated ORFs presenting a DIAMOND E-value hit ≤ 10E-6 were

assigned the corresponding WASPS centroid cluster number and a unique symbolism composed of a SVG icon with foreground and background colors. Translated ORFs above this threshold were non-orthologous singletons and retained the default white color. Three related plasmids were detected as best hits in the WASPS database and drawn similarly using gene cluster numbers pre-calculated at database inception. A combination of graphic symbols and colors would grant the simultaneous display of a quasi-unlimited number or orthologous genes. This 'fully connected' pre-calculated clustering topology allowed for near-instantaneous generation of completely resolved plasmid synteny maps. The consistent ORF symbolism permitted immediate visual synteny analysis.

A second analysis using the same pTN2 plasmid data using the 'Extended cluster hits' option produced a deeper analysis and reported the same 3 plasmids found previously and 6 additional plasmids (Suppl. Fig. 4). In that case, the entire orthologous clusters corresponding to each best DIAMOND centroid hit were considered positive. All plasmids having at least one gene belonging to one of these clusters were retained and drawn according to the same 'fully connected' clustering topology.

We have tested the accuracy of the predictive capabilities of WASPS syntenies using an alternative method. All plasmids depicted in Figure 1 (main article) carry a leftmost gene annotated as 'UvrD'. However in the WASPS synteny map, the corresponding YP_007195295.1 gene (white color) carried by pIRI48 is not considered as part to the same orthologous cluster containing the UvrD genes of the other plasmids. Gene YP_007195295.1 clustered alone in the WASPS database. To investigate the nature of this discrepancy we computed orthologous relationships between gene YP_007195295.1 and the WASPS UvrD cluster composed of genes YP_007195309.1, YP_007195325.1, WP_012819841.1, YP_007195341.1, YP_003603582.1, YP_007195352.1 and YP_003603439.1. We used Equation 1 (see Section 1, above) for the calculations of the similarities between these 8 proteins and the results are tabulated in Supplemental Table 1. The UvrD intra-cluster similarities (orange) were always near or in excess of the 30% orthology threshold recommended by (Lerat, et al., 2003). Interestingly, similarity with YP_007195295.1 (blue) was significantly lower suggesting non-orthology on the basis of protein sequence. These results validated the threshold parameter chosen for UCLUST (see Section 1, above).



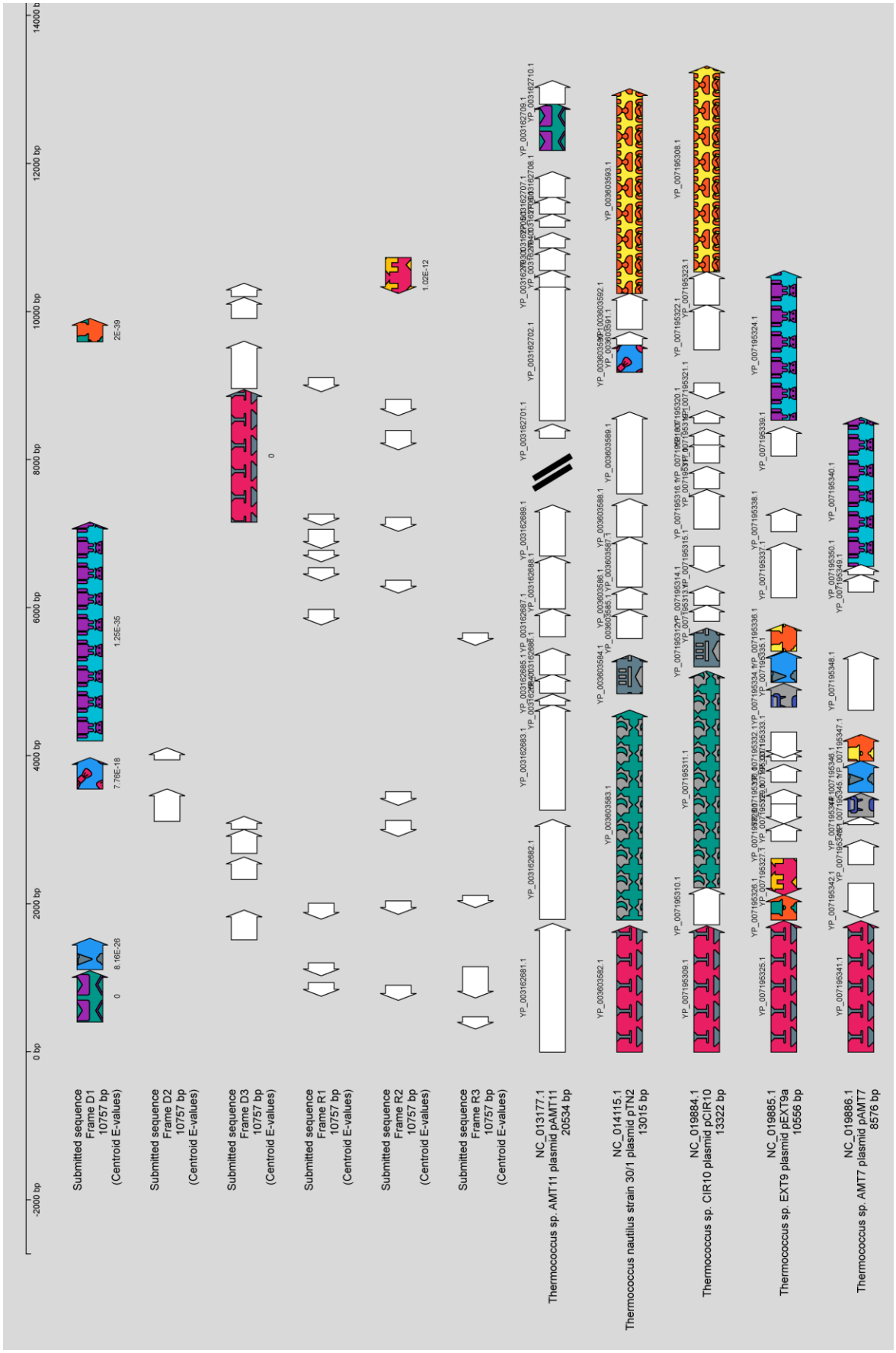
Supplemental figure 4. Extended pTN2 plasmid synteny*. Using the 'Extended cluster hits' option, the synteny analysis produced more extensive results, retrieving all 9 related plasmids from the WASPS database.*

Supplemental Table 1. Inter-cluster (blue) versus intra-cluster orthologous comparisons (orange)

Gene_Accession Plasmid_accession	YP_007195295.1 NC_019883.1	YP_007195309.1 NC_019884.1	YP_007195325.1 NC_019885.1	WP_012819841.1 NC_013408.1	YP_007195341.1 NC_019886.1	YP_003603582.1 NC_014115.1	YP_007195352.1 NC_019887.1	YP_003603439.1 NC_014110.1
YP_007195295.1 NC_019883.1	100	14.3	14.6	14.0	15.2	14.1	14.7	15.2
YP_007195309.1 NC_019884.1		100	27.6	26.7	27.2	74.1	26.5	82.7
YP_007195325.1 NC_019885.1			100	94.3	95.0	27.0	93.4	27.0
WP_012819841.1 NC_013408.1				100	97.7	27.3	97.5	26.4
YP_007195341.1 NC_019886.1					100	26.6	95.9	26.6
YP_003603582.1 NC_014115.1						100	27.4	70.1
YP_007195352.1 NC_019887.1							100	26.0
YP_003603439.1 NC_014110.1								100

4. Identification of NCBI contig QMOB01000129.1

The NCBI entry QMOB01000129.1, a 10757bp contig was assigned to a *Chloroflexi* bacterium metagenome (Dombrowski, et al., 2018). We submitted this entry to WASPS as a Fasta file and generated a synteny map using the default E-value of 10E-06, the 'primary hits only' option and BLAST. Very surprisingly, the six-frame translation generated 7 proteins displaying very high similarity (E-values ≤ 1.02E-12) to those encoded by 5 hyperthermophilic archaeal plasmids (Suppl. Fig.4). Since the *Chloroflexi* metagenomic samples originated from Guaymas Basin hydrothermal vents known to host Thermococcales archaea (Canganella, et al., 1998), we surmised that this particular contig in fact corresponded to a low level of archaeal DNA contamination in the metagenomics sample.



Supplemental Figure 4. *Synteny map for NCBI contig QMOB01000129.1*

5. Discussion

The plasticity of natural replicative plasmids contributes to the evolution of their host genome. Robust plasmid comparative genomics is therefore required to accurately assess the evolution of organisms. The conservation of gene order or synteny based on protein sequences has already proven successful to compare cellular genomes and infer evolutionary relationships. In this work, we have developed a novel database-backed natural plasmid synteny web service designed to overcome current limitations of current plasmid databases. The WASPS database is fully relational and carries all natural plasmids from the three domains of life. All plasmid-encoded proteins ranks in the database are pre-calculated according to a 'fully connected' clustering topology. The WASPS Webtool allows rigorous and straightforward analysis of user-submitted plasmid-related protein or DNA sequences. The highly significant and robust WASPS protein clustering allows the software to rapidly assign functions to submitted sequences and to infer their orthologous relationships. WASPS-generated synteny maps are almost identical to their manually computed and hand-drawn counterparts while requiring a fraction of the effort. WASPS' predictive capability allows users to easily identify mobile replicative mobile elements present in metagenomes as well as aid in the detection of potential DNA contaminations. The attractive and intuitive WASPS web interface incorporates the latest web standards and technologies. Among these, a navigable plasmid synteny map boasting striking iconic orthology symbols constitutes the flagship of this web service and should appeal to both wet and dry bench researchers. Important provisions have been adopted to keep the robust and lightweight WASPS database up to date by the means of an optimized and fully automated background task. The quality of the analyses provided by the WASPS Webtool are therefore destined to improve with time following the addition of new plasmid entries. Additional programs could be designed to take full advantage of the standalone WASPS database. The database model developed in this work could be further replicated to study viruses or target specifically integrative and conjugative elements.

6. References

Canganella, F.*, et al.* (1998) Thermococcus guaymasensis sp. nov. and Thermococcus aggregans sp. nov., two novel thermophilic archaea isolated from the Guaymas Basin hydrothermal vent site, *International journal of systematic bacteriology*, **48 Pt 4**, 1181-1185.

Dombrowski, N., Teske, A.P. and Baker, B.J. (2018) Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments, *Nature communications*, **9**, 4999.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics*, **26**, 2460-2461.

Lerat, E., Daubin, V. and Moran, N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria, *PLoS biology*, **1**, E19.