

## Ferroplasma acidarmanus

The genome of *Ferroplasma acidarmanus* strain fer1 was sequenced and published in a draft form in 2002. This sequence is public and is included in NMPDR and SEED as "Ferroplasma acidarmanus 97393.1".

The draft sequence was split among 61 non-overlapping pieces, or contigs. During the past few years, more sequencing and analysis has been done to fill some of the gaps. The new version of the sequence is in 18 contigs. The team who worked on the sequence assigned unique IDs to contigs and used these as labels in the FASTA file. The ID numbers begin with "gi|126007703|ref|" which is a genome ID for the NCBI database. The remainder of the ID numbers are as follows:

contig number	length (bp)
NZ_AABC05000001.1	53233
NZ_AABC05000002.1	1251
NZ_AABC05000003.1	1188
NZ_AABC05000004.1	42314
NZ_AABC05000005.1	137254
NZ_AABC05000006.1	293
NZ_AABC05000007.1	975
NZ_AABC05000008.1	226344
NZ_AABC05000009.1	232613
NZ_AABC05000010.1	177479
NZ_AABC05000011.1	321683
NZ_AABC05000012.1	59696
NZ_AABC05000013.1	89805
NZ_AABC05000014.1	392286
NZ_AABC05000015.1	194171
NZ_AABC05000016.1	1661
NZ_AABC05000017.1	668
NZ_AABC05000018.1	2249
<b>Total</b>	<b>1935163</b>
<b>Total contigs &gt; 20 kbp</b>	<b>1926878</b>

One file containing the FASTA-format sequences of the 18 contigs was submitted for annotation to RAST under the name "Ferroplasma acidarmanus complete" because this meets our definition of essentially complete: more than 70% of data in contigs at least 20 kbp in length. This job was shared with members of the class.

### Browse genome in SEED-Viewer

To start the assignment, login to the RAST server at <http://rast.nmpdr.org>. Your jobs overview will appear, and unless you have run other jobs of your own, the jobs list will list only "Ferroplasma acidarmanus complete." Notice that the table reports that 18 contigs were uploaded for analysis.

⇒ Click the link to "view details." At the top of the job details page, click the link to "Browse genome in SEED-Viewer."

The Organism Overview for your genome shows that on the basis of sequence similarity and genomic context (which proteins are located nearby), the automated analysis assigned 27% of the protein-encoding genes (pegs) names that included them in subsystems. Mouse over the blue bar to find that of the 73% of total genes not included in subsystems, more than half are hypothetical. Not much is known about the proteins in this genome, and your assignment is to discover more information.

### Organism Overview for *Ferroplasma acidarmanus* complete (666666.1681)

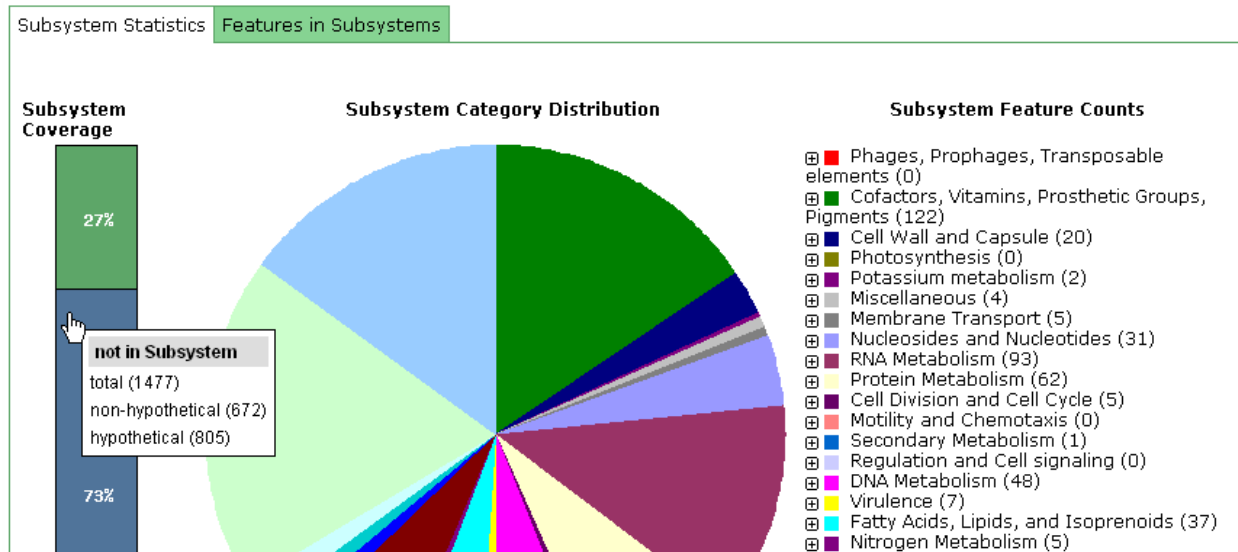
<b>Genome</b>	Ferroplasma acidarmanus complete
<b>Domain</b>	Archaea
<b>Size</b>	1,935,163 bp
<b>Number of Contigs</b>	16
<b>Number of Subsystems</b>	181
<b>Number of Coding Sequences</b>	2004
<b>Number of RNAs</b>	47

Browse **Compare** **Download**

Browse through the features of [Ferroplasma acidarmanus complete](#) both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click [here](#) to get to the Genome Browser

### Subsystem Information



⇒ Browse this genome either by selecting genome browser from the organism menu, or by clicking the link in the small tabbed box that floats in the upper right.

The genome browser has a tabbed box of controls in the top left, a graphical view of the genome in the top right, and a table that shows 15 pegs at a time at the bottom. The tabbed box has various controls to redraw the graphic. One set of controls is in the "location" tab, which allows you to choose which contig to view and the size of the region to view. In this case, the contig menu shows that the contig shown, number 18, is only 2,249 bp in length. The window size of the graphic is much larger; 16,000 bp by default.

⇒ Use the menu to select a smaller window, then click the draw button.

You may also color the proteins in the graphic according to focus (default), which is the peg most recently clicked on, or by other characteristics.

⇒ Use the drop-down menu to select "subsystems" then click the draw button.

Location **Focus** Upload List

contig: gi|126007513|ref|NZ\_AABC05000018.1| (2,249 bp)

start base: 0

window: 4,000 bp

Color features: by focus, by subsystem, by table filter options, by list, do not color

export table clear all filters

display 15 items per page  
displaying 1 - 15 of 2051

[next»](#) [last»](#)

Feature ID	Type	Contig	Start	Stop	Length (bp)	Function	Subsystems	Region
<a href="#">fig 666666.1681.peg.1</a>	CDS	gi 126007513 ref NZ_AABC05000018.1	402	1	402	glycosyl transferase	- none -	<a href="#">show</a>
<a href="#">fig 666666.1681.peg.2</a>	CDS	gi 126007513 ref NZ_AABC05000018.1	1393	455	939	Integrase, catalytic domain	- none -	<a href="#">show</a>
<a href="#">fig 666666.1681.peg.3</a>	CDS	gi 126007513 ref NZ_AABC05000018.1	1668	1393	276	Transposase IS3/IS911	- none -	<a href="#">show</a>
<a href="#">fig 666666.1681.peg.4</a>	CDS	gi 126007519 ref NZ_AABC05000016.1	1286	546	741	dolichol-phosphate mannosyltransferase	- none -	<a href="#">show</a>
<a href="#">fig 666666.1681.peg.5</a>	CDS	gi 126007521 ref NZ_AABC05000015.1	80	853	774	glucose/galactose transporter	- none -	<a href="#">show</a>
<a href="#">fig 666666.1681.peg.6</a>	CDS	gi 126007521 ref NZ_AABC05000015.1	1697	912	786	Glucose 1-dehydrogenase (EC 1.1.1.41)	D-gluconate and lactulose	<a href="#">show</a>

The table has column headers that are searchable and sortable. Notice that the first graphical view of contig 18 has only three proteins, and the table also lists only the first three proteins as located on that contig. This makes sense due to the short length of the contig. The table lists no entries for contig 17.

- ⇒ Use the contig menu in the location controls to select the next contig, gi|126007703|ref|NZ\_AABC05000017.1, then click the draw button

Notice that there are no proteins shown because none are encoded on this very short, 668-bp piece of DNA. This agrees with the information in the table. The table also lists only one peg on the next contig, which is also very small. The controls for drawing the graphic do not govern what is shown in the table.

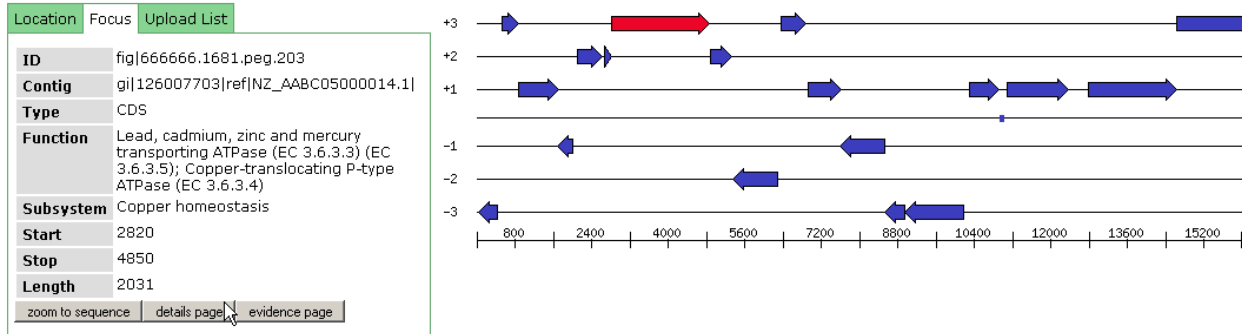
- ⇒ Use the contig menu to select number 14, the longest contig, then click the draw button.
- ⇒ Use the window menu to select a different view, then click the draw button.
- ⇒ Use the right arrow next to the draw button to advance the graphic along the contig. Mouse over any of the arrows to see which peg is displayed.

Notice that the pegs displayed in the table have not changed in response to the controls that redraw the graphic.

- ⇒ In the table header select the same contig shown in the graphic, number 14 (gi|126007703|ref|NZ\_AABC05000014.1).
- ⇒ Click the "show" button to focus the graphic on fig|666666.1681.peg.203

The selected peg is included in a subsystem, but is surrounded by pegs with poorly or totally undefined function. Notice that the controls for the graphic have switched from the "location" tab to the "focus" tab, and that the protein whose button you clicked in the table is shown red in the graphic. In other words, the graphic controls do not redraw the table, but the show button in the table will redraw the graphic focused on the selected peg.

## Browse Genome: *Ferroplasma acidarmanus* complete (666666.1681)



There are now two avenues for finding more information about the focus protein. In the focus controls box next to the graphic, you can use the "details page" button to open the Annotation Overview page for the focus peg. This automatically opens a new tab or window of your browser. The same page will open if you click the Feature ID link for that peg in the table, but the page will open in the same browser window. Also in the graphic controls box, there is a button to open the evidence page. The same page will open from an Annotation Overview by clicking the "feature evidence" link in the table at the top of that page.

### Annotation Overview and Feature Evidence

The Annotation Overview page will present the same type of information for all proteins. The Annotation Overview and Evidence for fig|666666.1681.peg.203 are presented here as an example. This peg is annotated as Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-translocating P-type ATPase (EC 3.6.3.4). This protein belongs to a subsystem, Copper homeostasis, in its role as Copper-translocating P-type ATPase. There seems to be some uncertainty in the role of this protein. Given the physiology and environment of this organism, it might be possible for this one protein to transport all the listed metals. From this page we can look to the EC numbers, the subsystem, the compare regions, and the evidence for more information about these possibilities.

- ⇒ Click on the EC number links at the top of the page, which will open (in a new tab/window) a description, provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG), of the reaction defined by the number.

Do the assigned functions conflict? EC 3.6.3.3 is a cadmium-exporting, "P-type ATPase that undergoes covalent phosphorylation during the transport cycle. This enzyme occurs in protozoa, fungi and plants." We are examining an enzyme from an archaeon. Continue scrolling through the information on this page, and you'll find literature references for articles about this enzyme in several bacteria.

EC 3.6.3.5 is a zinc-exporting, "P-type ATPase that undergoes covalent phosphorylation during the transport cycle. This enzyme also exports Cd<sup>2+</sup> and Pb<sup>2+</sup>." Again, scrolling down to the bottom of this description provides links to the literature.

EC 3.6.3.4 is a copper-exporting, "P-type ATPase that undergoes covalent phosphorylation during the transport cycle. This bacterial and mammalian enzyme exports Cu<sup>2+</sup> from cells. In humans, it is involved in Menkes disease and Wilson's disease." Although only bacterial and mammalian versions are claimed, there are several archaea listed toward the bottom of the genes list: AFU stands for *Archeoglobus fulgidus*. Notice as you scroll down the list that while these genes all catalyze the same reaction, there are several different gene names used: ATP7A, copA, ybaR, zntA, copB. Some of these may be paralogs rather than orthologs; that is, the progenitor gene duplicated within an ancestral organism, and the copies continued to evolve somewhat differently, preferentially exporting one metal ion over another. Duplication and divergence is a classic mechanism for the evolution of new enzymatic functions. It is often difficult to figure out, based on sequence similarity alone, whether two enzymes from two different organisms are orthologs, evolved from the same ancestral version; or paralogs, evolved from slightly

different ancestral versions that duplicated in the common ancestral organism. Try opening a few of the genes listed with different names. You'll see that there is a variety of amino acid sequence lengths.

Examining the reactions described the three EC numbers does not clarify whether the *Ferroplasma* enzyme can perform all three functions, or if it performs only one or two of them. Sequence homology in context with nearby conserved clusters may provide evidence of function.

⇒ Look at the Compare Regions graphic.

The top line (no matter which gene you are focused on) is the newly annotated version of *Ferroplasma*. The second line will show a genomic region surrounding the most similar protein to the focus peg, so in this case, it will usually be the 2002 version of the genome, simply labeled "*Ferroplasma acidarmanus*." There are no features depicted to the far left of the focus protein in *Ferroplasma* because it is located close to the start of this contig—there are only 2680 bp of DNA upstream of this peg.

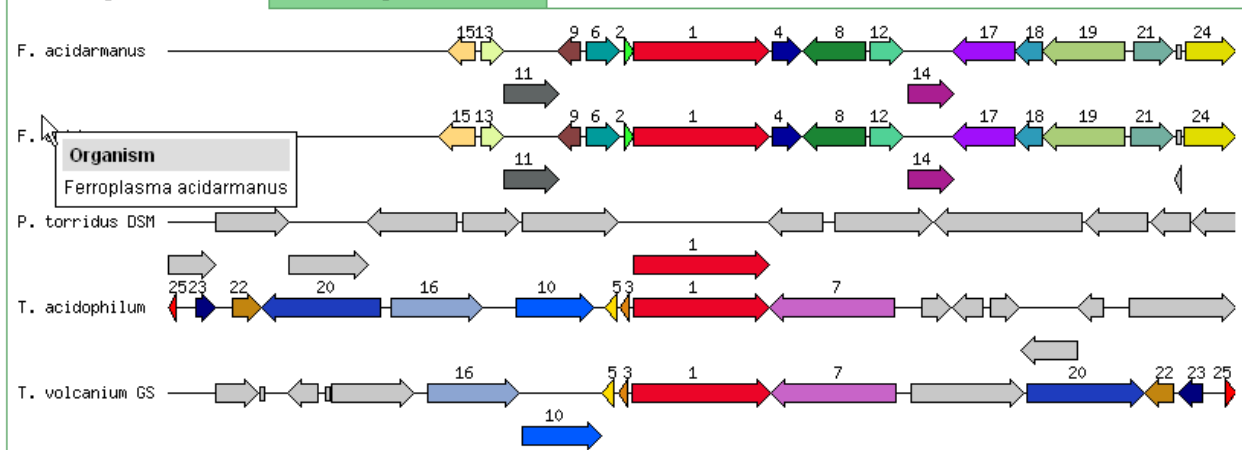
### Compare Regions

Display options

Region Size (bp)

Number of Regions

Visual Region Information



The focus protein and its homologs are colored red and numbered 1. The set of homologous proteins most frequently located close to the focus are colored green and numbered 2. Homologous proteins, those with similar amino acid sequence, share the same color and numerical label, ordered by number of displayed genomes in which similarities occur. Non-homologous proteins and non-protein features such as RNAs are shown in gray. Is there a cluster of homologs surrounding the focus peg in the genomes displayed? If so, they may function together. In the case of this heavy metal transporter, there is no obvious functional clustering even in the closely related genomes shown. (How do you know which organisms are closely related? Go to [www.nmpdr.org](http://www.nmpdr.org), select Organisms from the Search menu in the grey bar, click the link under Taxonomy, and use your browser to find *Ferroplasma*.) The most closely related organisms are *Picrophilus* and the *Thermoplasmas*.

⇒ Use the text boxes above the graphic to expand the number and/or size of regions, then click the button to redraw the display.

This focus gene (set 1, red) is not universally paired with the next-most frequently found set of homologs, set 2, green. Conserved pairs of closely spaced homologs (PCH) frequently provide some clue to the mutual function of the paired genes.

- ⇒ Click on the Advanced button. Increase the number of genomic regions shown to 20 or 30, select collapse close genomes, then click the button to redraw the graphic.

This reveals that the transporter is found in a wide variety of genomes, but it is relatively isolated on the genomes. To detect similarities among very small proteins, you can use the advanced controls to increase the expectation value used to define similar groups to 1e-10 or 1e-5. If you are trying to disambiguate duplicates, lower the cut-off to 1e-50. If sets 2 or 3 are consistently shown close to set 1 when few genomes are shown, select the PCH pin (pairs of close homologs) to see the extent of conserved co-location. If there is no obvious clustering, leave the pin set at similarity, which pulls in genomes based on sequence similarity of the focus gene alone, rather than similarity of the focus gene and its neighbors. From the results you may discover information that adds more meaning to the functional annotation. In this case, the transporter is isolated in the genomes in which it appears, so genomic context does not add functional information.

All the information depicted in the graphic is contained in a table in a tab behind the graphic, with the addition of one more tool.

- ⇒ Click "Tabular Region Information" to view the table.

The focus gene is again highlighted red. We have determined from the graphic that it is not clustered on the genome of *Ferroplasma acidarmanus* or the most closely related organisms. The gene may be clustered in distantly related organisms.

- ⇒ Click the cluster button in the row corresponding to the focus protein (the red row).

Visual Region Information    Tabular Region Information

export table

Genome ▲▼	ID ▲▼	Start ▲▼	Stop ▲▼	Size (nt)	Strand	Function ▲▼	FC ▲▼	SS ▲▼	Set ▲▼	CL
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.197</a>	431	33	399	-	Ubiquinol-cytochrome C reductase iron-sulfur subunit (EC 1.10.2.2)			15	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.198</a>	540	878	339	+	Twin-arginine translocation protein TatA		1	13	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.199</a>	880	1704	825	+	Twin-arginine translocation protein TatC		1	11	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.200</a>	2007	1684	324	-	hypothetical protein			9	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.201</a>	2114	2620	507	+	transcriptional regulatory protein			6	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.202</a>	2681	2815	135	+	conserved hypothetical protein			2	cluster
Ferroplasma acidarmanus complete	<a href="#">figl666666.1681.peg.203</a>	2820	4850	2031	+	Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-		10	1	cluster

The table lists which organisms have a conserved cluster containing the focus peg, as well as which other functions are in the cluster. None of the listed functions contradicts or is inconsistent with the annotation of the version in *Ferroplasma*.

Sequence similarity analysis is also available by running BLAST live from this page, or by opening the Evidence page to look at pre-computed results of BLASTP. At the top of the Annotation Overview page, there is a drop-down menu of sequence analysis tools. Psi BLAST is listed first.

- ⇒ With Psi-BLAST showing in the menu, click the button to "Run Tool"

The tool will run in a scrolling frame in a new tab or window of your browser. The result shown is the same as running BLASTP at NCBI against the non-redundant (nr) database. To set up a position-specific scoring matrix, you can go through and deselect all but the most closely related similarities, then click the button to run a second iteration. The hits will be scored according to amino acid similarity (the BLOSUM substitution matrix) as well as according to the conserved position of amino acids in the sequences selected. This tool is sometimes helpful in finding proteins with slight functional differences when there is an over-all similarity that masks important differences in the functional site of a transporter or enzyme.

**current assignment** We are uncertain of the precise function of this feature. It is probably one of the following:  
 Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5) **EC Number** [3.6.3.3](#), [3.6.3.5](#)  
 Copper-translocating P-type ATPase (EC 3.6.3.4) **EC Number** [3.6.3.4](#)  
[show encoded function](#)

**taxonomy id** [666666](#) **contig** gi|126007703|ref|NZ\_AABC05000014.1|

**internal links** [genome browser](#) [feature evidence](#) [sequence](#) **ACH** [?] [show essentially identical genes](#)

**annotation history** [show](#) **run tool** Psi-Blast [run tool](#)

Pre-computed BLASTP results are also shown on the Evidence page, along with results of sequence analysis tools designed to predict subcellular location or recognize short regions (domains) with defined functions.

⇒ Click the feature evidence link near the top of the Annotation Overview.

**Sims Filter**

Max Sims:  Max E-val:   
  
 Sort Results By:   
 Group By Genome  
[Resubmit](#)

[Align Selected](#) [Fasta Download Selected](#)

Include query

E-Value Key	Function
Query Fer.aci.	<input type="checkbox"/> Copper-transporting P-type ATPase (EC 3.6.3.4)
Query Pic.tor.DS	<input type="checkbox"/> copper transporting ATPase
Query The.vol.GS	<input type="checkbox"/> Cation transport ATPase
Query The.aci.DS	<input type="checkbox"/> heavy-metal transporting P-type ATPase related pro
Query Nos.sp.PCC	<input type="checkbox"/> Lead, cadmium, zinc and mercury transporting ATPas
Query Nostoc sp. PCC 7120 [103690] Nos.sp.PCC	<input type="checkbox"/> Lead, cadmium, zinc and mercury transporting ATPas
Query Nos.sp.PCC	<input type="checkbox"/> Lead, cadmium, zinc and mercury transporting ATPas
Query Cal.sac.DS	<input type="checkbox"/> Lead, cadmium, zinc and mercury transporting ATPas

**Organism** Nostoc sp. PCC 7120 [103690]

Conserved functional domains and subcellular location can often be predicted from patterns in the amino acid sequence—in this case, the results were inconclusive. The BLASTP results are shown color-coded for e-value, and with a bar that depicts the region of matching amino acid sequence. If there is a possibility of paralogs (closely related duplicate sequences with slightly different functions), you might learn something from checking the box to group genomes, and resubmitting the analysis. Again, in this particular case, the functional annotations of the similar sequences shown are pretty consistent. The reason for this consistency is that this protein has been included in a subsystem. All proteins performing the same function defined in a subsystem are given the same name. If several names are equally likely based on bioinformatic analysis, and there is no experimental evidence in the literature in favor of one function, then the curator assigns all likely functions.

Have we learned anything more about fig|666666.1681.peg.203? No, the annotator has defined the function as well as possible without experimental testing. Given the small size of the genome, the lack of duplicates within this genome, and the lifestyle of the organism, this protein probably transports many different heavy metal ions. Experimental testing would be required before changing the annotation to use "and" rather than "or" between the EC numbers.

You will be assigned a set of proteins with annotations that give no hint of the function of the protein. Your task will be to use the comparative tools as described above to add information to the annotation. You may determine whether the gene is in a conserved cluster, whether it is conserved only in Archaea or does it appear in Bacteria as well? Is it conserved only in organisms that live at an extreme temperatures or pH? Are there any conserved functional domains within the protein? Adding the name of the domain, if there is one, adds meaning to an annotation of "hypothetical protein" because it will differentiate that protein from all the other hypotheticals that do not contain the domain.

You will discover that sequence analysis and comparative bioinformatics can do only so much in isolation from experimental data and biological knowledge. In this case what bioinformatics can do is to identify genes that should be studied experimentally to determine how this organism works.