JC

10. juuni 2013

Aare Abroi

Diversity of Virophages in Metagenomic Data Sets

Jinglie Zhou,[a,b,c] Weijia Zhang,[b] Shuling Yan,[d] Jinzhou Xiao,[a,b] Yuanyuan Zhang,[b,c] Bailin Li,[a,b] Yingjie Pan,[a,b] Yongjie Wang[a,b]

Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage & Preservation (Shanghai), Ministry of Agriculture, Shanghai, China[a]; College of Food Science and Technology, Shanghai Ocean University, Shanghai, China[b]; Department of Biological Science, College of Sciences and Mathematics, Auburn University, Auburn, Alabama, USA[c]; Institute of Biochemistry and Molecular Cell Biology, University of Goettingen, Goettingen, Germany[d]

Virophages, e.g., Sputnik, Mavirus, and Organic Lake virophage (OLV), are unusual parasites of giant double-stranded DNA (dsDNA) viruses, yet little is known about their diversity. Here, we describe the global distribution, abundance, and genetic diversity of virophages based on analyzing and mapping comprehensive metagenomic databases. The results reveal a distinct abundance and worldwide distribution of virophages, involving almost all geographical zones and a variety of unique environments. These environments ranged from deep ocean to inland, iced to hydrothermal lakes, and human gut- to animal-associated habitats. Four complete virophage genomic sequences (Yellowstone Lake virophages [YSLVs]) were obtained, as was one nearly complete sequence (Ace Lake Mavirus [ALM]). The genomes obtained were 27,849 bp long with 26 predicted open reading frames (ORFs) (YSLV1), 23,184 bp with 21 ORFs (YSLV2), 27,050 bp with 23 ORFs (YSLV3), 28,306 bp with 34 ORFs (YSLV4), and 17,767 bp with 22 ORFs (ALM). The homologous counterparts of five genes, including putative FtsK-HerA family DNA packaging ATPase and genes encoding DNA helicase/primase, cysteine protease, major capsid protein (MCP), and minor capsid protein (mCP), were present in all virophages studied thus far. They also shared a conserved gene cluster comprising the two core genes of MCP and mCP. Comparative genomic and phylogenetic analyses showed that YSLVs, having a closer relationship to each other than to the other virophages, were more closely related to OLV than to Sputnik but distantly related to Mavirus and ALM. These findings indicate that virophages appear to be widespread and genetically diverse, with at least 3 major lineages.

April 2013 Volume 87 Number 8 Journal of Virology p. 4225–4236

The Evolutionary Landscape of Alternative Splicing in Vertebrate Species

Nuno L. Barbosa-Morais,[1,2] Manuel Irimia,[1]* Qun Pan,[1]* Hui Y. Xiong,[3]* Serge Gueroussov,[1,4]* Leo J. Lee,[3] Valentina Slobodeniuc,[1] Claudia Kutter,[5] Stephen Watt,[5] Recep Çolak,[1,6] TaeHyung Kim,[1,7] Christine M. Misquitta-Ali,[1] Michael D. Wilson,[4,5,7] Philip M. Kim,[1,4,6] Duncan T. Odom,[5,8] Brendan J. Frey,[1,3] Benjamin J. Blencowe[1,4]†

How species with similar repertoires of protein-coding genes differ so markedly at the phenotypic level is poorly understood. By comparing organ transcriptomes from vertebrate species spanning ~350 million years of evolution, we observed significant differences in alternative splicing complexity between vertebrate lineages, with the highest complexity in primates. Within 6 million years, the splicing profiles of physiologically equivalent organs diverged such that they are more strongly related to the identity of a species than they are to organ type. Most vertebrate species-specific splicing patterns are cis-directed. However, a subset of pronounced splicing changes are predicted to remodel protein interactions involving trans-acting regulators. These events likely further contributed to the diversification of splicing and other transcriptomic changes that underlie phenotypic differences among vertebrate species.

Gene Transfer from Bacteria and Archaea Facilitated Evolution of an Extremophilic Eukaryote

Gerald Schönknecht,[1,2]*† Wei-Hua Chen,[3,4]† Chad M. Ternes,[1]† Guillaume G. Barbier,[5]†‡ Roshan P. Shrestha,[5]†§ Mario Stanke,[6] Andrea Bräutigam,[2] Brett J. Baker,[7] Jillian F. Banfield,[8] R. Michael Garavito,[9] Kevin Carr,[10] Curtis Wilkerson,[5,10] Stefan A. Rensing,[11]‖ David Gagneul,[12] Nicholas E. Dickenson,[13] Christine Oesterhelt,[14] Martin J. Lercher,[3,15] Andreas P. M. Weber[2,5,15]*

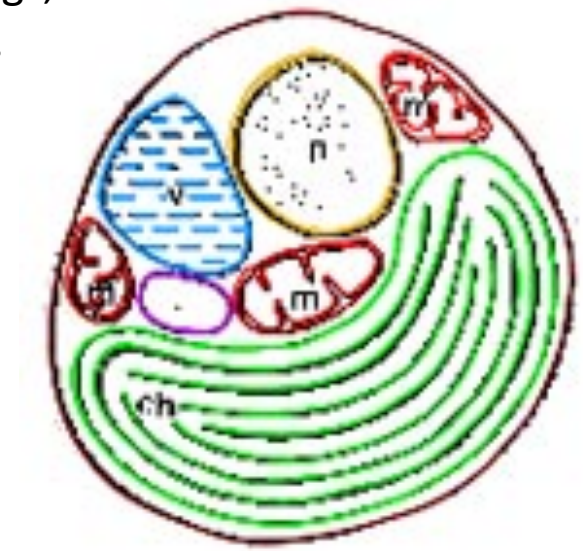
Some microbial eukaryotes, such as the extremophilic red alga *Galdieria sulphuraria*, live in hot, toxic metal-rich, acidic environments. To elucidate the underlying molecular mechanisms of adaptation, we sequenced the 13.7-megabase genome of *G. sulphuraria*. This alga shows an enormous metabolic flexibility, growing either photoautotrophically or heterotrophically on more than 50 carbon sources. Environmental adaptation seems to have been facilitated by horizontal gene transfer from various bacteria and archaea, often followed by gene family expansion. At least 5% of protein-coding genes of *G. sulphuraria* were probably acquired horizontally. These proteins are involved in ecologically important processes ranging from heavy-metal detoxification to glycerol uptake and metabolism. Thus, our findings show that a pan-domain gene pool has facilitated environmental adaptation in this unicellular eukaryote.

Punavetikas *Galdieria sulphuraria*

Eukaryota; Rhodophyta; Bangiophyceae; Cyanidiales; Cyanidiaceae; Galdieria.

The unicellular red micro-alga *Galdieria sulphuraria* (Cyanidiales) is a eukaryote that can represent up to 90% of the biomass in extreme habitats such as hot sulfur springs with pH values of 0 to 4 and temperatures of up to 56°C. This red alga thrives autotropically as well as heterotropically on more that **50** different carbon sources, including a number of rare sugars and sugar alcohols.

G.sulphuraria naturally inhabits volcanic hot sulfur springs, solfatara soils, and anthropogenic hostile environments.



Ultrastructure:
Size: 3-9 µm; highly proteinaceous cell wall; single chloroplast (Kuroiwa et al., 1989) with chlorophyll a, phycocyanin and allophycocyanin; one single mitochondrion (Suzuki et al., 1994) and several peroxisomes; floridean starch outside the chloroplast.

http://genomics.msu.edu/galdieria

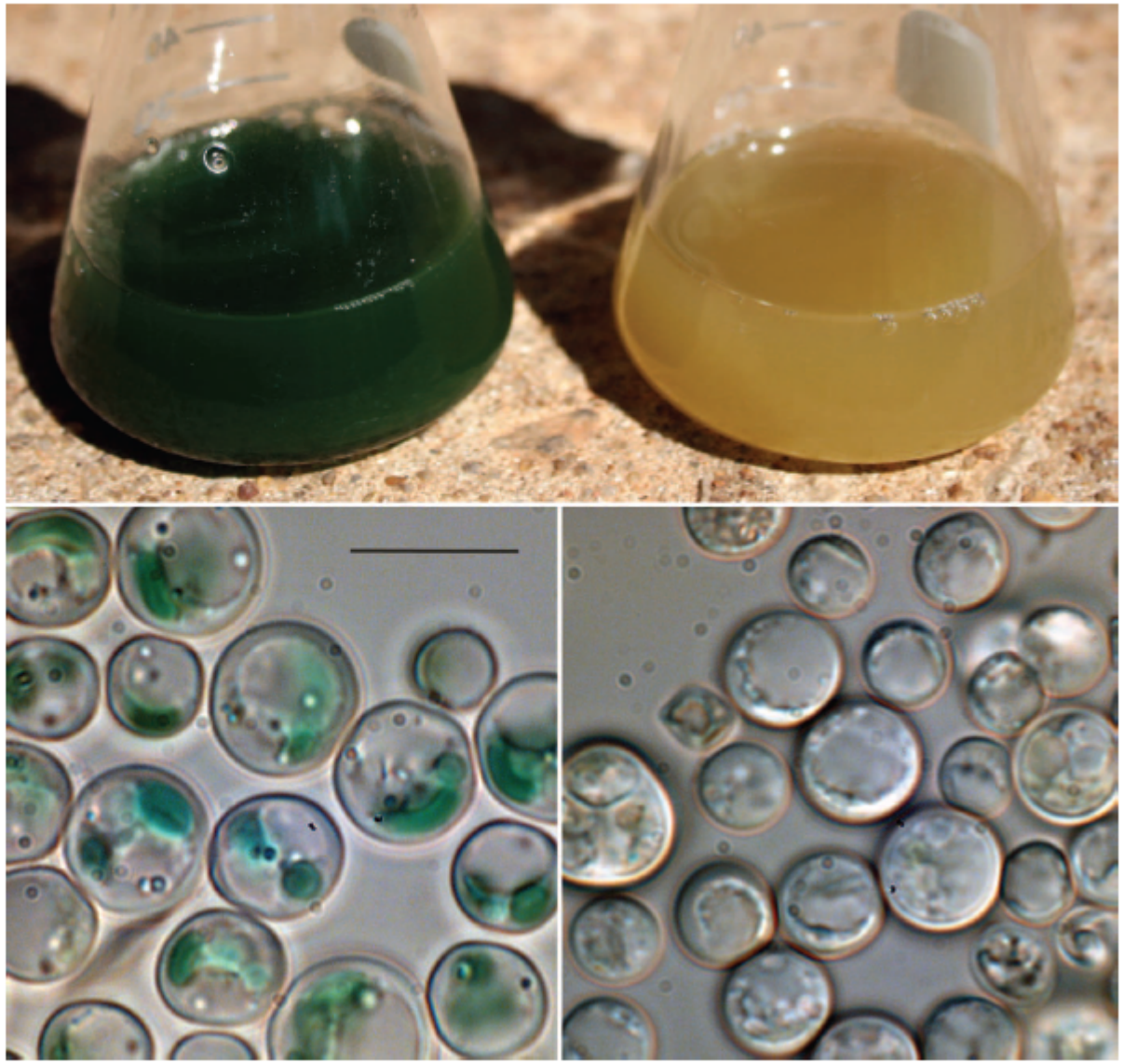
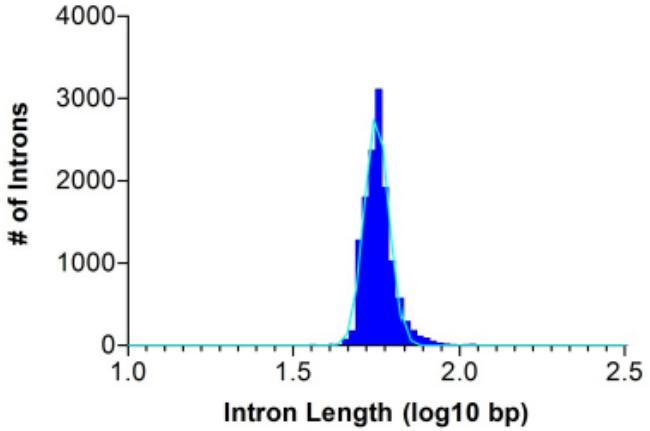


Fig. 1. Photoautotrophic (left) and heterotrophic (right) *G. sulphuraria* cells. Cell cultures (top) and light microscopic images (bottom; bar represents 10 μm) of *G. sulphuraria* cells grown under continuous illumination in the absence of glucose (left) or in darkness in the presence of 200 mM glucose (right).

G. sulphuraria genome

- 13.7 MB
- 6623 proteins
- Coding sequence make up 77.5 % of the genome
- Median intergenic distance 22 bp
- Protein coding genes on average 2 introns, median lengths 55 bp



Intron size distributions of the G. sulphuraria genome. Size distributions of 13630 introns are displayed as frequency against logarithmic length (log10 bp). The distribution of logarithmic intron lengths is described by a simple Gaussian distribution with Ampl = 2796±34, Mean = 1.752±0.0005 (56.5 bp), and SD = 0.0363±0.0005.

Genome sequencing

Three libraries:

Small-insert (~2kbp inserts)

Fosmid library (~40 kbp)

BAC libraries (> 100 kbp)

ABI 3730xl capillary sequencing system

$147*10^6$ Q20 bases ~10-fold coverage

PLUSS

~164 000 reads (8.55x coverage) using GS20 Genomic Sequncer (Roche)

ARACHNE genome assambler

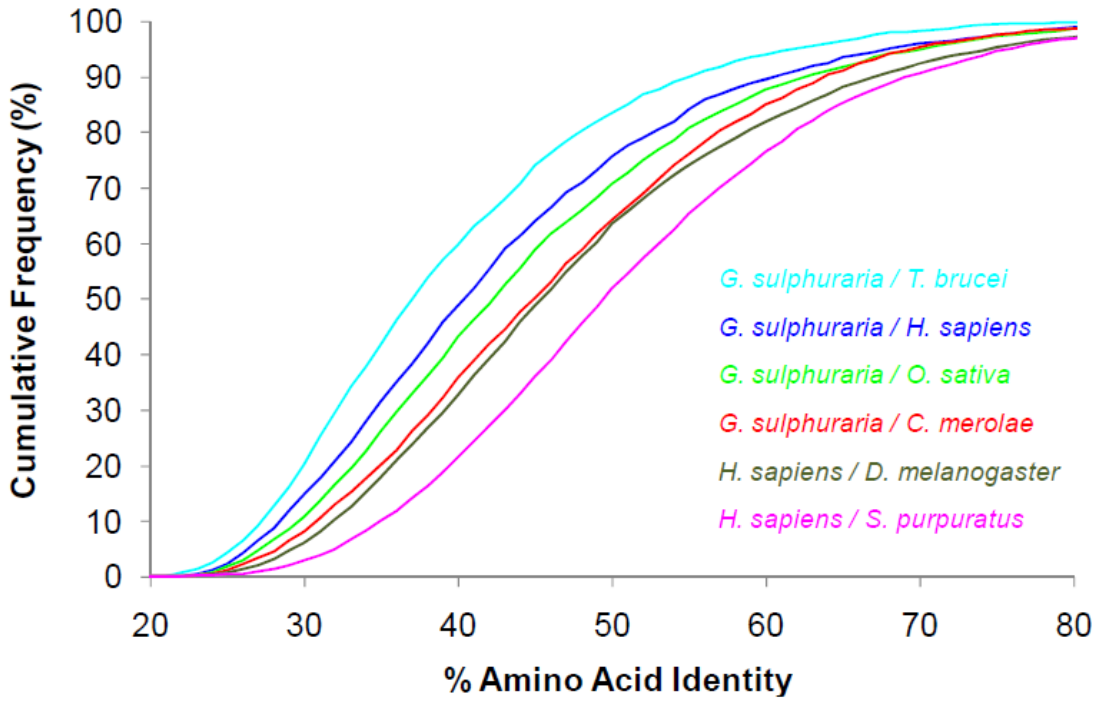
433 scaffolds

N50 172 322 bases

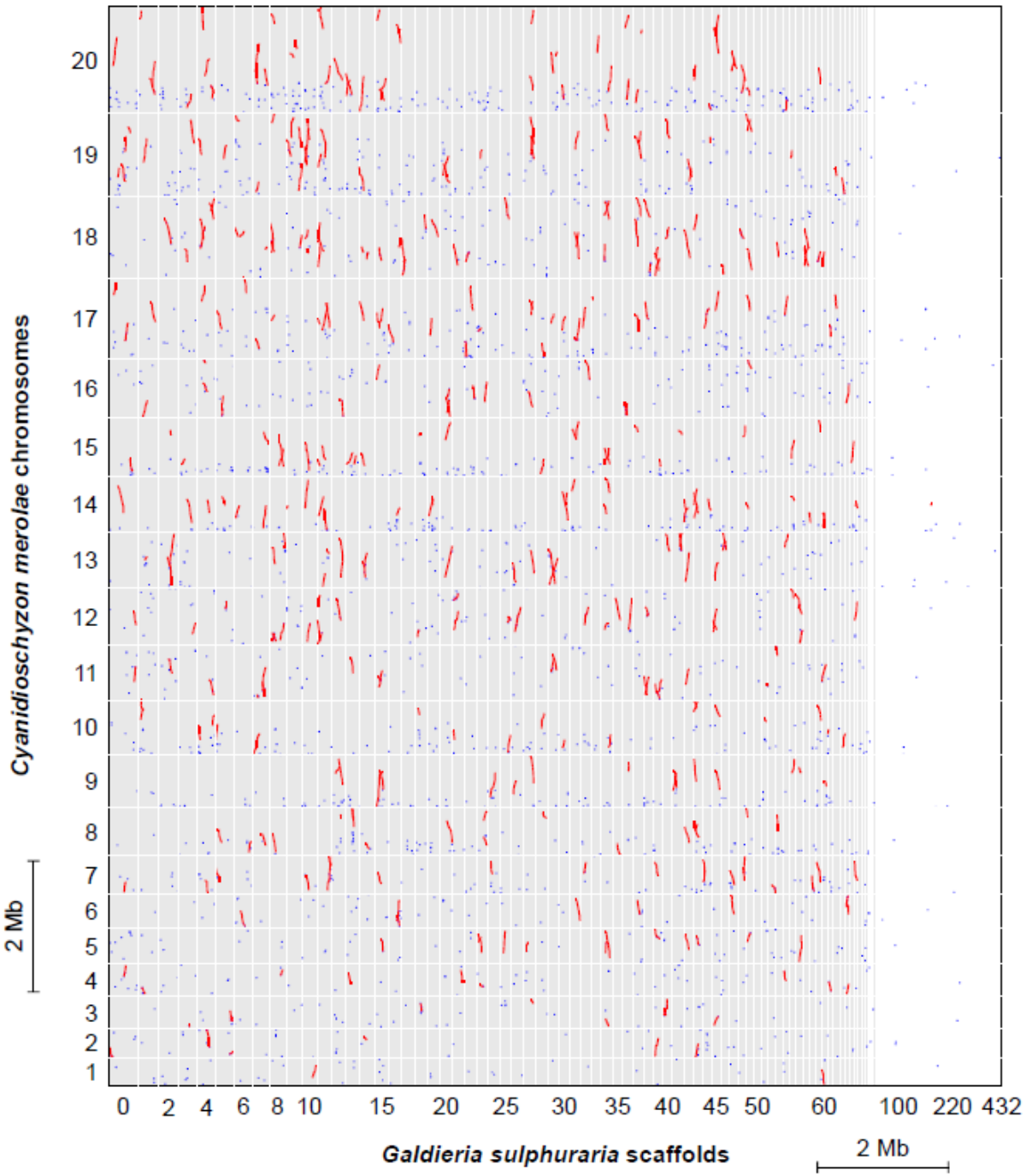
13 712 004 bases (of which 292 650 bases are gaps).

AUGUSTUS for gene prediction (EST sequences by Sanger and mRNA by GS20)

Closest sequenced organism Cyanidioschyzon merolae



Molecular divergence between G. sulphuraria and C. merolae. As a measure of molecular divergence, percent amino acid identity of orthologous gene pairs was compared for different species pairs. Sets of orthologous gene pairs were identified by reciprocal best BLAST hits with BLAST scores > 50. Normalized (to 100%) cumulative frequencies are plotted against % amino acid identity for six different species pairs. The two red algal genomes display a slightly lower amino acid identity (median **44.9%**) compared to Homo sapiens / Drosophila melanogaster (**45.4%**). The vertebrate – insect divergence probably occurred about **910 million** years ago (55), indicating a similar age for the G. sulphuraria – C. merolae split.



Colinear regions between G. sulphuraria and C. merolae genomes.

A colinearity plot of all 20 C. merolae chromosomes (Y axis) against 433 G. sulphuraria scaffolds (X axis) was generated with ColinearScan (http://colinear.cbi.pku.edu.cn/#overview) using a minimum BLAST score of 100. Blue dots indicate orthologous genes identified using InParanoid (41), which are linked by red lines if they are in synteny blocks. Genes in one block are not necessarily next neighbors, but may be separated by other genes, to allow for gene loss, gene creation, and/or minor chromosomal re-arrangements.

Table S1.

Orthologous relationships between *G. sulphuraria* and *C. merolae* genomes. For intron and exon size distribution in *G. sulphuraria* see fig. S5, for a comparison of intergenic distances see fig. S3.

	G. sulphuraria	***C. merolae***
Genome size (Mb)	13.7	16.5
GC (%)	37.7	55.0
GC (%) coding sequence (excluding introns)	38.6	56.7
CpG occurrence (Obs./Exp.)	0.715	1.151
CpG islands	27	2
Repeat content	713	313
Number of rRNA units (18S/5.8S/28S + 5S)	4 + 10	3 + 3
Number of tRNAs	85	30
Predicted proteins	6623	5771
Gene density (kb per gene)	2.07	2.86
Average gene length (bp)	1601	1553
Average transcript length (bp)	1388	1552
Average number of amino acids per polypeptide	421	518
Average number of exons per gene	3.16	1.005
Average exon length (bp)	417.3	1527.6
Introns	13630	26
Genes with introns (%)	72.4	0.5
Average intron length (bp)	56.5	248
Median intergenic distance (bp)	20.0	1404.5
Coding sequence (%)	77.5	44.9

- 42% of G. sulphuraria proteins have orthologs in C. merolae
- 25% of both genomes constitute syntenic blocks
- Only 1,259 annotated proteins (19%) from G. sulphuraria have proteins from C. merolae as best BLAST hits, and only slightly more than 60% (4,017 proteins) give a significant BLAST hit (score >50) with C. merolae. The latter percentage is comparable to that of BLAST hits with proteins from Arabidopsis thaliana (4,032 proteins).

All HGT candidates was identified by phylogenetic analyses

Blast against 208 species database (InParanoid), if blast hits only in A or B to next round. If best hit in A or B and hits to E as well - to phylogenetic analysis.

Short sequences removed; trees with less than ten species where excluded; HGT candidate was submitted to NCBI BLAST service (nr), tree of best blast hits was generated, HGT candidates that were not confirmed by a tree of best BLAST hits were not accepted; if tree not informative about origin of G. sulphuraria sequence then gene was removed.

For this candidate proteins NCBI nr blast and/or KEGG or Pfam to collect sequences. For nonredundand (90%) MSA with T-Coffee (mcoffee or accurate)

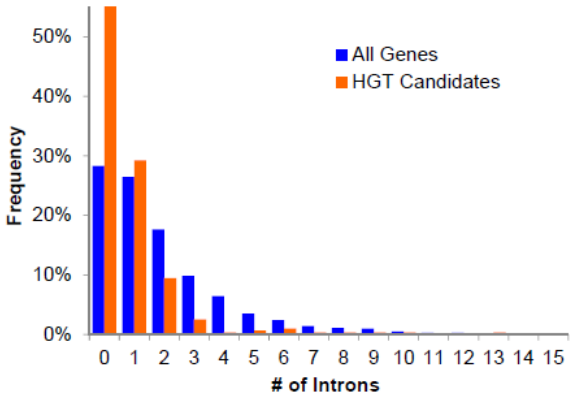
- 178 G. sulphuraria proteins had significant BLAST hits (score > 50) only in bacterial or archaeal sequences.
- Out of those 178 proteins, 110 were accepted as HGT candidates after further inspection using the NCBI Web BLAST service (nr) combined with the Tree View option and criteria as described above. Phylogenetic analyses indicated that those 110 HGT candidates resulted from 25 HGT events.

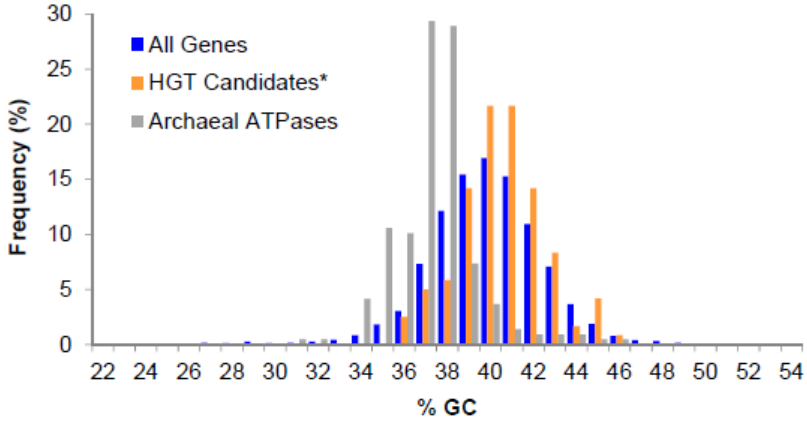
- 618 proteins had best BLAST hits in bacterial or archaeal sequences. Out of those, RELL analyses for maximum likelihood phylogenetic trees confirmed horizontal gene transfer for 163 proteins with statistical significance (5% significance level). From these 163 proteins, 50 were accepted as HGT candidates after further inspection. Phylogenetic analyses indicated that those 50 HGT candidates resulted from 44 HGT events.
- Genes were excluded as HGT candidates during further inspection in case the encoded protein a) was too short (<150 amino acids); b) had too few BLAST hits; c) potentially originated from endosymbiotic gene transfer; d) resulted in a phylogenetic tree that did not allow conclusions; e) had significant sequence similarity with proteins from eukaryotic species not included in the 208 genomes used for the systematic bioinformatics screen, resulting in a phylogenetic tree of best BLAST hits (NCBI nr) that did not confirm HGT. These very stringent criteria for the manual inspection of each HGT candidate were aimed at preventing false positives as far as possible.

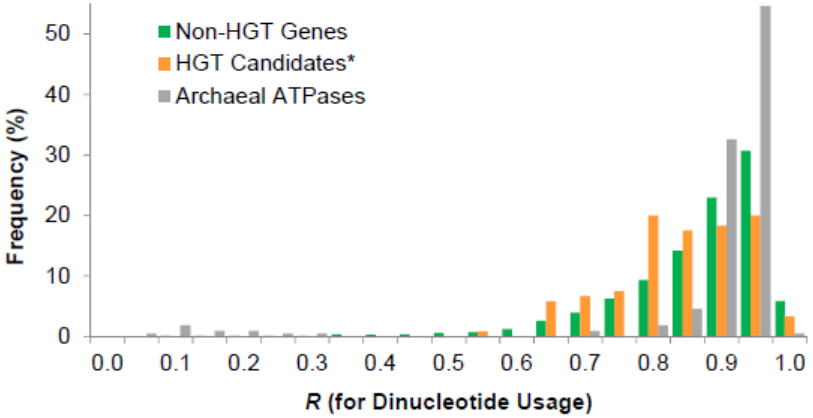
- 337 genes resulting from 75 horizontal gene transfers were detected
- From 337, only 160 HGT candidates were identified by the genome-wide bioinformatics screen (other were identified by manual phylogenetic analyses).

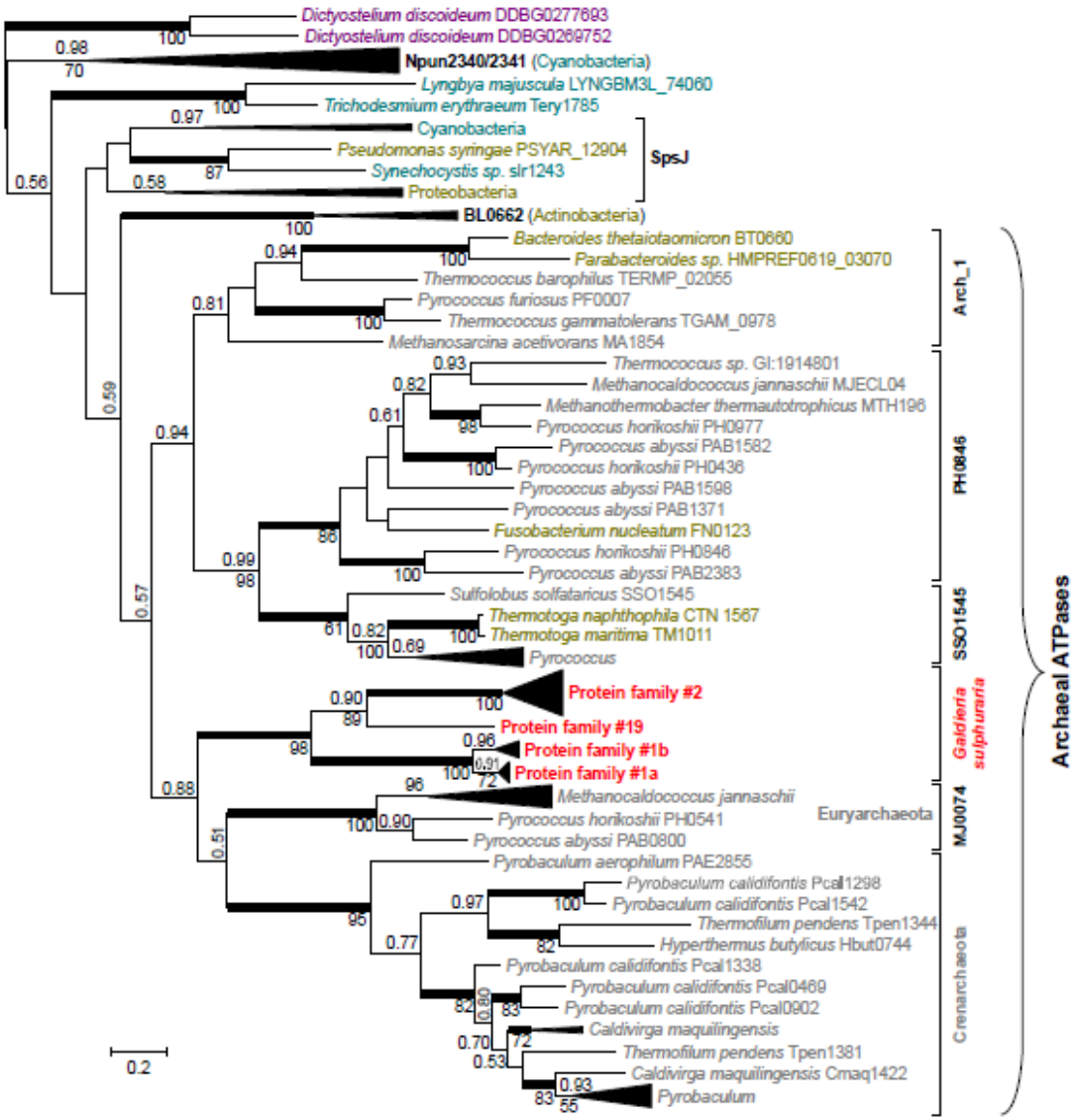
Additional parameters

- Introns per gene 0.8 vs 2.06
- GC content
- Oligonucleotide usage

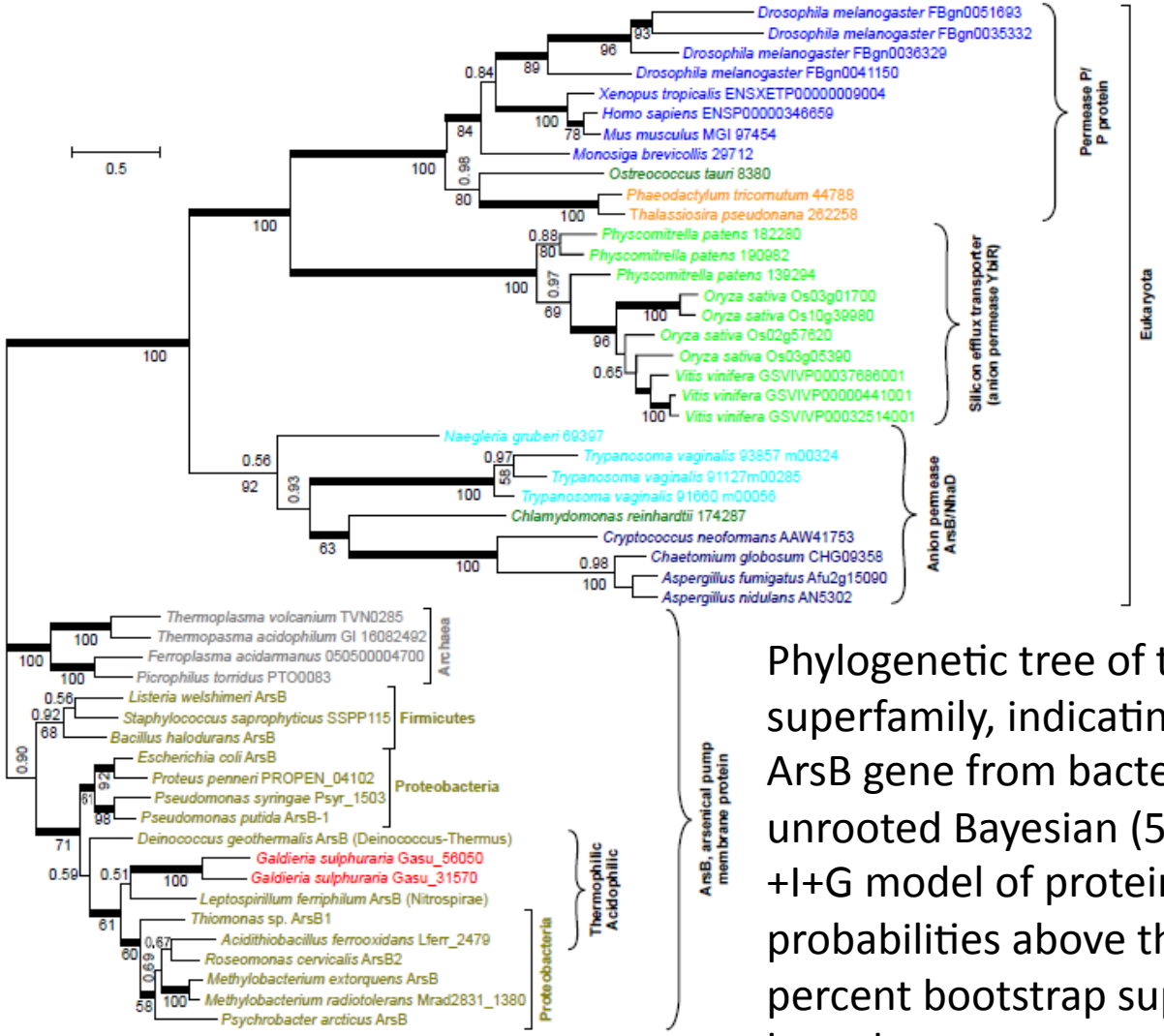








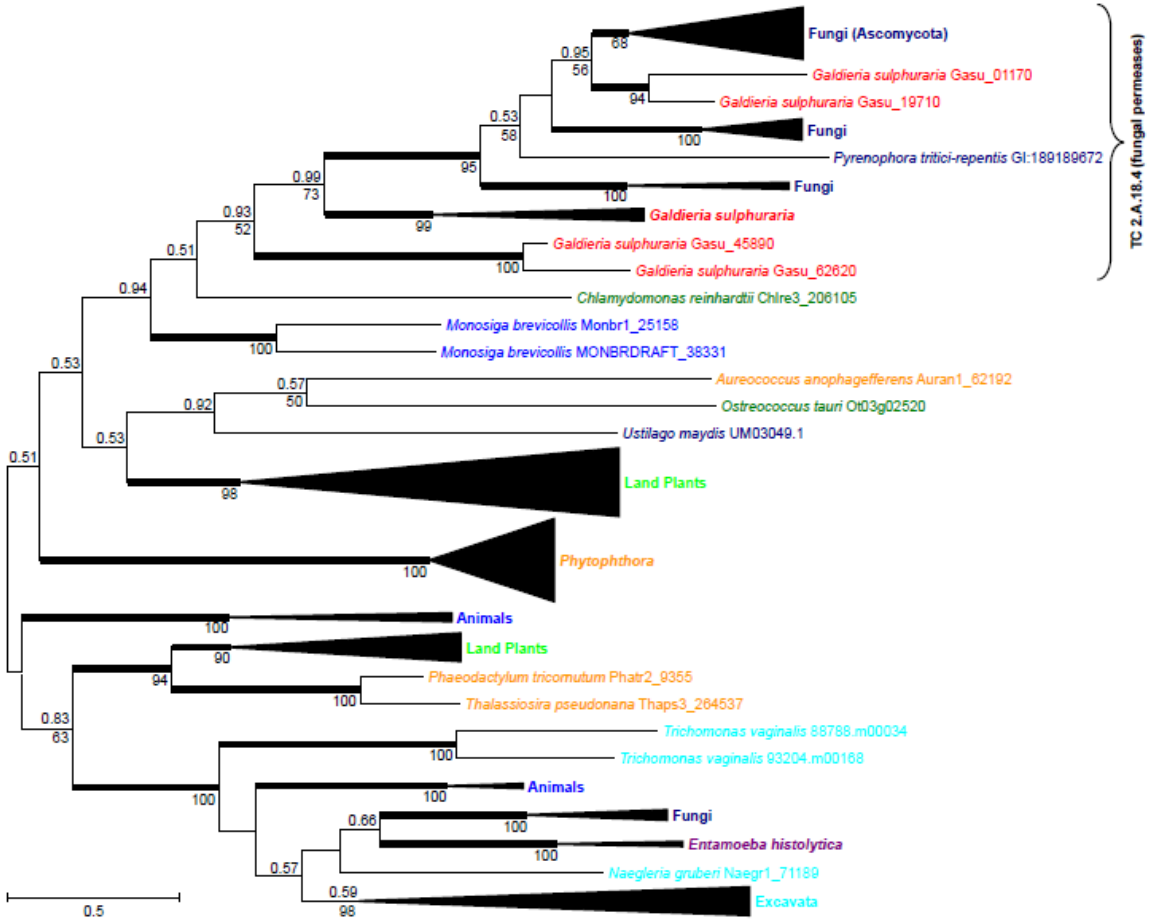
Phylogenetic tree of the MNS clade of the STAND class of P-loop ATPases. The unrooted Bayesian (52) tree calculated with a CpREV+I+G model of protein evolution shows posterior probabilities above the branches and PhyML (75) bootstrap support values (using LG+I+G+F) below the branches. Thick branches indicate 1.0 posterior probability. Constraining all eukaryotic sequences into one monophyletic branch outside the Archaeal ATPases resulted in a tree with pRELL = 0.0. For clarity some sub-branches with sequences from the same clade have been collapsed (elongated triangles) with the height of the triangles reflecting the number of taxa included (3 to 22). The tree was constructed from 136 sequences: 41 from G. sulphuraria, all 56 seed sequences of the 'Arch_ATPase' family (PF01637) (76), plus sequences mentioned in Leipe et al. (20) plus BLAST (35) hits of these three groups for sequences outside the Archaeal ATPases. Square brackets indicate proteins of monophyletic origin. Six families as established by Leipe et al. (20) are indicated; MJ-type, SSO-type, and PH-type families of the Archaeal ATPases, plus BL0662 family, SpsJ family, and Npun2340/2341 family outside the Archaeal ATPases. Color coding: Archaea, Cyanobacteria, other Bacteria, Rhodophyta, Amoebozoa. Bar represents 0.2 changes per site.



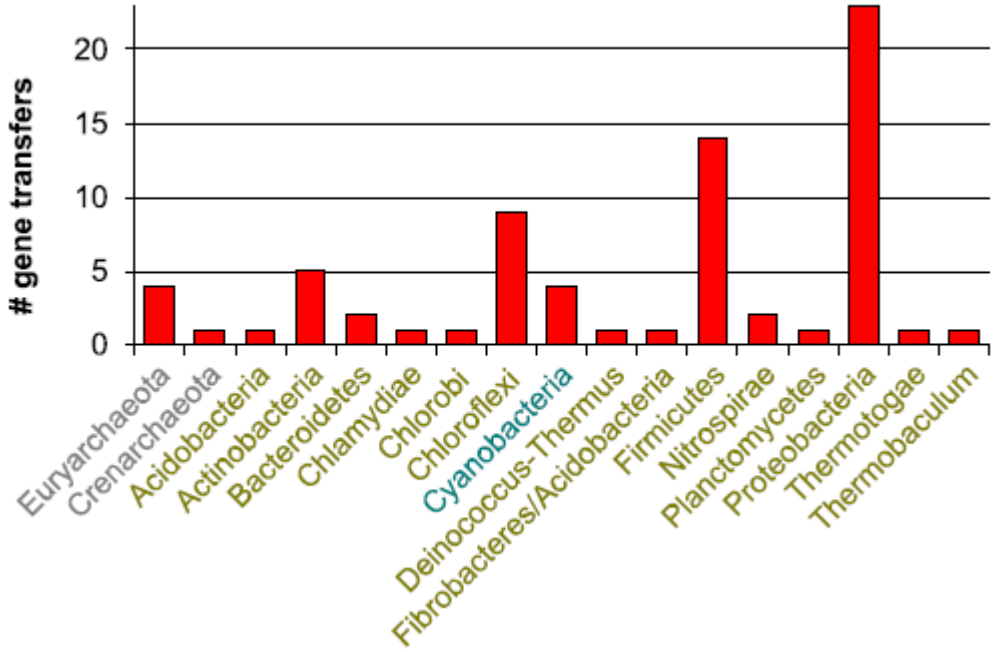
Color coding of major phylogenetic groups: **Bacteria,** Archaea, Streptophyta, Chlorophyta, **Rhodophyta, Stramenopiles,** Excavata, **Fungi, Animals & Choanoflagellates**

Phylogenetic tree of the anion permease ArsB/NhaD superfamily, indicating horizontal gene transfer of an ArsB gene from bacteria to G. sulphuraria. The unrooted Bayesian (52) tree calculated with a CpREV +I+G model of protein evolution shows posterior probabilities above the branches and PhyML (75) percent bootstrap support (using LG+I+G) below the branches.

Thick branches indicate 1.0 posterior probability. Square brackets group proteins according to their origin from Eukaryota, Archaea, or different phyla of Bacteria. Curly brackets indicate different subfamilies of the anion permease ArsB/NhaD superfamily (27). The label 'Thermophilic/Acidophilic' marks G. sulphuraria and four bacteria being thermophilic and/or acidophilic, living in similar habitats as G. sulphuraria does. Bar represents 0.5 changes per site.

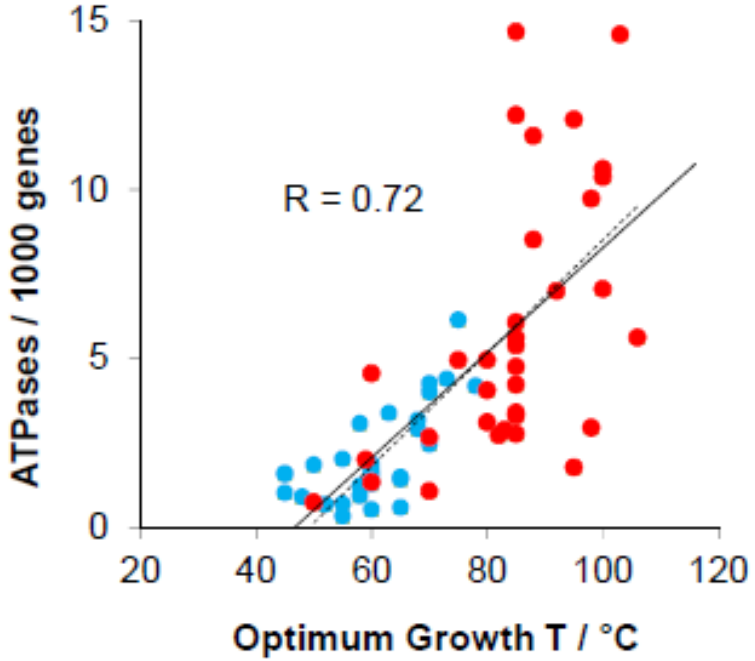


Phylogenetic tree of the amino acid / auxin permease (AAAP) family. The unrooted Bayesian (*52*) tree calculated with a CpREV+I+G model of protein evolution shows posterior probabilities above the branches and PhyML (*75*) percent bootstrap support (using LG+G+F) below the branches. Thick branches indicate 1.0 posterior probability. For clarity most sub-branches have been collapsed (elongated triangles) with the height of the triangles reflecting the number of taxa included (3 to 46). Curly bracket indicates a subfamily of fungal amino acid permeases (TC 2.A.18.4), which includes all *G. sulphuraria* sequences, according to best BLAST hits in the Transporter Classification Database (*38*). Color coding of major phylogenetic groups: Streptophyta, Chlorophyta, Rhodophyta, Stramenopiles, Excavata, Amoebozoa, Fungi, Choanoflagellates & Animals. Bar represents 0.2 changes per site.



Number of horizontal gene transfers from different phyla into the genome of G. sulphuraria. For each horizontal gene transfer the phylum of the 'donor' organism from which the gene might have originated was determined by phylogenetic analysis or best BLAST hits (see table S4 for details).

Tabel ülekaranutest



Number of Archaeal ATPase genes per 1000 coding genes in genomes of thermophilic and hyperthermophilic organisms as function of optimum growth temperature. Genomes from 57 thermophilic and 29 hyperthermophilic archaea and bacteria were downloaded from NCBI via the "microbial genomes properties" portal (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Optimum growth temperatures were obtained from NCBI. If temperatures were given as a range (58-60°C for example), the highest number was chosen (60°C in this case). In total, we obtained valid optimum growth temperatures for 38 thermophilic (including 8 archaea, in red, and 30 bacteria, in cyan) and 26 hyperthermophilic (all archaea) organisms. The HMM model of Archaeal ATPases (accession ID: PF01637.10) was downloaded from the Pfam database (76). HMMER version 3 beta was used to score each protein sequence in each thermophilic and hyperthermophilic species against this HMM model. A protein was considered an Archaeal ATPase if the e-value of the HMM search was less than 10-5. A trend line (solid) plus correlation coefficient are given for both data sets combined. A trend line for data points from archaea only gave a similar result (dashed line; R = 0.56).

HGT from A and B to G. sulphurariae

- At least 75 events, (67 from B, 6 from A and 2 from V)
- 3-fold enriched in membrane transporters
- 14 fold enriched in „extermophilic families“
- Common environment common families??

Kõik.

Suur tänu kuulamast!