Journal of
Molecular Microbiology
and Biotechnology

J Mol Microbiol Biotechnol 2007;12:165–179
DOI: 10.1159/000099639

Large-Scale Comparative Genomic Analyses of Cytoplasmic Membrane Transport Systems in Prokaryotes

Qinghu Ren Ian T. Paulsen

The Institute for Genomic Research, Rockville, Md., USA

Key Words
Membrane transporter · Prokaryotic genomes · Eubacteria · Archaea · Phylogenetic profiling

Abstract
The recent advancements in genome sequencing make it possible for the comparative analyses of essential cellular processes like transport in organisms across the three domains of life. Membrane transporters play crucial roles in fundamental cellular processes and functions in prokaryotic systems. Between 3 and 16% of open reading frames in prokaryotic genomes were predicted to encode membrane transport proteins, emphasizing the importance of transporters in their lifestyles. Hierarchical clustering of phylogenetic profiles of transporter families, which are derived from the presence or absence of a certain transporter family, showed distinct clustering patterns for obligate intracellular organisms, plant/soil-associated microbes and autotrophs. Obligate intracellular organisms possess the fewest types and number of transporters presumably due to their relatively stable living environment, while plant/soil-associated organisms generally encode the largest variety and number of transporters. A group of autotrophs are clustered together largely due to their absence of transporters for carbohydrate and organic nutrients and the presence of transporters for inorganic nutrients. Inside of each group, organisms are further clustered by their phylogenetic properties. These findings strongly suggest the correlation of transporter profiles to both evolutionary history and the overall physiology and lifestyles of the organisms.

Introduction

Membrane transport systems are vital to every living organism. Transporters function in the acquisition of organic nutrients, maintenance of ion homeostasis, extrusion of toxic and waste compounds, environmental sensing and cell communication, and other important cellular functions [Saier, 1999], therefore playing essential roles in life-endowing processes like metabolism, communication, and reproduction. There has also been increasing evidence suggesting the relevance of the composition of membrane transport systems to the general physiology and lifestyles of the organisms [Paulsen et al., 1998, 2000; Ren and Paulsen, 2005].

Various transport systems differ in their putative membrane topology, energy coupling mechanisms and substrate specificities [Saier, 2000]. The most commonly utilized energy sources to drive transport are adenosine triphosphate (ATP), phosphoenolpyruvate, or chemiosmotic energy in the form of sodium ion or proton electrochemical gradients. Primary active transporters couple the transport process to a primary source of energy

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2007 S. Karger AG, Basel
1464–1801/07/0124–0165$23.50/0

Accessible online at:
www.karger.com/mmb

Ian T. Paulsen
The Institute for Genomic Research
9712 Medical Center Dr.
Rockville, MD 20850 (USA)
Tel. +1 301 795 7531, Fax +1 301 838 0208, E-Mail ipaulsen@tigr.org

(ATP hydrolysis), for example, the MalKGFE maltose transporter from *Escherichia coli* [Bohm et al., 2002; Boos and Shuman, 1998]. Secondary transporters utilize an ion or solute electrochemical gradient, such as the proton/sodium motive force, to drive the transport process, e.g. *E. coli* LacY lactose permease [Abramson et al., 2003; Newman et al., 1981; Viitanen et al., 1986]. Group translocators transport and phosphorylate their substrates. *E. coli* MtlA mannitol PTS transporter phosphorylates exogenous mannitol using phosphoenolpyruvate as the phosphoryl donor and energy source and releases the phosphate ester, mannitol-1-P, into the cell cytoplasm [Elferink et al., 1990; Postma et al., 1993]. Compared to other transporter types, channels are unique in that they are energy-independent transporters that transport water, specific types of ions or hydrophilic small molecules down a concentration or electrical gradient with higher rates of transport and lower stereospecificity, e.g. *E. coli* GlpF glycerol channel [Sweet et al., 1990].

Cytoplasmic membrane transporters typically consist of at least one membrane-localized protein component with multiple transmembrane-spanning α-helical segments. This has led to membrane transport systems being difficult to study experimentally. The genomic/bioinformatic analyses provide an attractive alternative to study membrane transporters [Ren et al., 2004]. As of today, over 400 prokaryotic genomes have been sequenced and deposited in the public databases (Gold Genomes Online Database, http://www.genomesonline.org/) [Bernal et al., 2001; Janssen et al., 2005]. These genomes cover a broad range of microbial organisms from different phylogenetic groupings, allowing comparative genomic analyses across a diverse range of organisms and lifestyles. The functional prediction and classification of complete membrane transport systems in these sequenced genomes, as well as the comparative analyses of transporter profiles from related organisms are of great value in understanding organisms' physiology and lifestyles.

In this review, we present a comparative genomic study of prokaryotic membrane transport systems from 201 sequenced genomes, with the focus on their relationship to their overall physiology and lifestyles.

Comparative Genomic Analysis of Membrane Transport Systems

Bioinformatic analyses of 201 species, including 178 eubacteria and 23 archaea (table 1) enabled us to identify a total of 53,669 transport proteins. Based on sequence similarities and phylogenetic analyses, these transport proteins could be categorized into 94 families, including 5 families of primary transporters, 70 families of secondary transporters, 11 channel protein families, 2 phosphotransferase systems, and 6 unclassified families. Some of these families are very large superfamilies with numerous members, such as the ATP-binding cassette superfamily (ABC) and the major facilitator superfamily (MFS), both of which are widely distributed across the eubacterial and archaeal species. Some families, on the contrary, only exist in a very limited phylogenetic spectrum and/or are present in only limited numbers.

The total number of predicted cytoplasmic membrane transport proteins (fig. 1a) and the percentage of transport proteins relative to the total number of open reading frames (ORFs) (fig. 1b) were compared for the 201 prokaryotes (listed by their phylogenetic groupings). Between 3 and 16% of ORFs in prokaryotic genomes were predicted to encode membrane transport proteins, emphasizing the importance of transporters in the lifestyles of all eubacterial and archaeal species. There is considerable variation on the quantity of transport proteins, even for species within the same phylogenetic group. For example, organisms within the α-Proteobacteria exhibit distinct lifestyles and corresponding differences in transporter contents. They include the rhizosphere-dwelling organisms *Mesorhizobium loti* (884 transport proteins, 12.2% of ORFs), *Bradyrhizobium japonicum* (987, 11.9%) and *Sinorhizobium meliloti* (827, 13.3%); the plant pathogen *Agrobacterium tumefaciens* (824, 15.3%); the human pathogens *Brucella* spp. (360–379, 11.0–11.9%); marine Roseobacters, like *Silicibacter pomeroyi* (571, 13.4%) and *Jannaschia* sp. (507, 12.0%), and obligate intracellular pathogens or endosymbionts such as *Rickettsia* spp., *Wolbachia* spp., *Anaplasma* spp., and *Ehrlichia* spp. (53–59, 4.5–7.0%). Across all phyla, obligate endosymbionts and intracellular pathogens generally seem to possess the most limited repertoire of membrane transporters.

Organisms with the lowest percent of ORFs encoding transport proteins include *Pirellula* sp. (225, 3.1%), a marine aerobic heterotrophic planctomycete; *Leptospira interrogans* (147, 3.1%), a parasitic pathogenic spirochaete, and several archaeal species, such as *Methanococcus jannaschii* (68, 3.9%), *Methanopyrus kandleri* (54, 3.2%), and *Nanoarchaeum equitans* (17, 3.0%). One of the contributing factors could be the very limited experimental characterization of species in these phylogenetic groupings, which serves the base for bioinformatic predictions. Pre-

Table 1. Organisms used in this study and their transport proteins

Taxonomy	Organism name	Organism ID	Total transport proteins	Percent of ORFs (%)
Archaea-Crenarchaeota	*Aeropyrum pernix* K1	1	158	8.6
	Pyrobaculum aerophilum IM2	2	146	5.6
	Sulfolobus solfataricus P2	3	191	6.4
	Sulfolobus tokodaii strain7	4	166	5.9
	Sulfolobus acidocaldarius DSM639	5	152	6.8
Archaea-Euryarchaeota	*Archaeoglobus fulgidus* DSM4304	6	184	7.6
	Halobacterium sp. NRC-1	7	160	6.1
	Methanosarcina acetivorans C2A	8	394	8.7
	Methanococcus jannaschii DSM	9	68	3.9
	Methanopyrus kandleri AV19	10	54	3.2
	Methanococcus maripaludis S2	11	138	8.0
	Methanosarcina mazei Goe1	12	248	7.4
	Methanobacterium thermoautotrophicum ΔH	13	102	5.4
	Pyrococcus abyssi GE5	14	177	10.0
	Pyrococcus furiosus DSM3638	15	195	9.4
	Pyrococcus horikoshii OT3	16	159	8.8
	Picrophilus torridus DSM9790	17	171	11.1
	Thermoplasma acidophilum DSM1728	18	145	9.8
	Thermoplasma volcanium GSS1	19	143	4.7
	Haloarcula marismortui ATCC43049	20	330	7.8
	Natronomonas pharaonis DSM2160	21	216	7.7
	Thermococcus kodakaraensis KOD1	22	198	8.6
Archaea-Nanoarchaea	*Nanoarchaeum equitans* Kin4-M	23	17	3.0
Actinobacteria	*Bifidobacterium longum* NCC2705	24	233	13.5
	Corynebacterium diphtheriae NCTC13129	25	251	11.0
	Corynebacterium efficiens YS-314	26	303	10.3
	Corynebacterium glutamicum ATCC13032	27	355	11.9
	Leifsonia xyli CTCB07	28	179	8.8
	Mycobacterium avium K-10	29	293	6.7
	Mycobacterium bovis AF2122/97	30	240	6.1
	Mycobacterium leprae TN	31	100	6.2
	Mycobacterium tuberculosis H37Rv	32	238	6.1
	Nocardia farcinica IFM10152	33	443	7.5
	Propionibacterium acnes KPA171202	34	297	12.9
	Streptomyces avermitilis MA-4680	35	705	9.3
	Streptomyces coelicolor A3(2)	36	702	8.9
	Tropheryma whippelii TW08/27	37	64	8.2
	Tropheryma whipplei Twist	38	25	8.7
Aquificae	*Aquifex aeolicus* VF5	39	89	5.8
Bacteroidetes	*Bacteroides fragilis* YCH46	40	256	5.5
	Bacteroides fragilis NCTC9343	41	256	6.0
	Bacteroides thetaiotaomicron VPI-5482	42	253	5.3
Chlamydia	*Chlamydophila caviae* GPIC	44	76	7.6
	Chlamydia muridarum Nigg	45	66	7.2
	Chlamydia pneumoniae AR39	46	74	6.7
	Chlamydophila pneumoniae TW-183	47	73	6.6
	Chlamydia trachomatis serovar D	48	69	7.7
	Parachlamydia sp. UWE25	49	121	6.0
Chlorobi	*Chlorobium chlorochromatii* CaD3	50	114	5.7
	Chlorobium tepidum TLS	43	122	5.4
	Pelodictyon luteolum DSM273	51	164	7.9

Table 1 (continued)

Taxonomy	Organism name	Organism ID	Total transport proteins	Percent of ORFs (%)
Chloroflexi	*Dehalococcoides ethenogenes* 195	52	101	6.4
	Dehalococcoides sp. CBDB1	53	102	7.0
Cyanobacteria	*Gloeobacter violaceus* PCC7421	54	246	5.6
	Nostoc sp. PCC7120	55	382	6.2
	Prochlorococcus marinus MIT9313	56	142	6.3
	Prochlorococcus marinus SS120(CCMP1375)	57	88	4.7
	Prochlorococcus marinus MED4(CCMP1378)	58	94	5.5
	Synechococcus elongatus PCC6301	59	185	7.3
	Synechocystis sp. PCC6803	60	220	6.9
	Synechococcus sp. WH8102	61	148	5.9
	Thermosynechococcus elongatus BP-1	62	166	6.7
Deinococcus-Thermus	*Deinococcus radiodurans* R1	63	262	8.2
	Thermus thermophilus HB27	64	212	9.6
Firmicutes	*Bacillus anthracis* Ames	65	564	10.6
	Bacillus anthracis A2012	66	682	12.3
	Bacillus cereus ATCC14579	67	571	10.9
	Bacillus halodurans C-125	68	510	12.5
	Bacillus licheniformis ATCC14580	69	503	12.1
	Bacillus subtilis 168	70	423	10.3
	Bacillus thuringiensis konkukian 97-27	71	626	12.2
	Clostridium acetobutylicum ATCC824	72	371	9.6
	Carboxydothermus hydrogenoformans Z-2901	73	166	6.3
	Clostridium perfringens 13	74	311	11.4
	Clostridium tetani E88	75	266	11.2
	Enterococcus faecalis V583	76	393	12.6
	Geobacillus kaustophilus HTA426	77	319	9.0
	Lactobacillus acidophilus NCFM	78	268	14.4
	Listeria innocua Clip11262 (rhamnose-negative)	79	380	12.5
	Lactobacillus johnsonii NCC533	80	286	15.7
	Lactococcus lactis IL1403	81	245	10.8
	Listeria monocytogenes EGD-e	82	387	13.6
	Listeria monocytogenes 4b	83	370	13.1
	Lactobacillus plantarum WCFS1	84	401	13.3
	Mesoplasma florum L1	85	74	10.8
	Mycoplasma gallisepticum R	86	76	10.5
	Mycoplasma genitalium G-37	87	55	11.4
	Mycoplasma hyopneumoniae 232	88	94	13.6
	Mycoplasma mobile 163K	89	69	10.9
	Mycoplasma mycoides PG1T	90	103	10.1
	Mycoplasma penetrans HF-2	91	94	9.1
	Mycoplasma pneumoniae M129	92	47	6.9
	Mycoplasma pulmonis UAB CTIP	93	89	11.4
	Oceanobacillus iheyensis HTE831	94	439	12.6
	Phytoplasma asteris OY-M	95	55	7.3
	Streptococcus agalactiae 2603V/R	96	274	12.9
	Streptococcus agalactiae NEM316	97	278	13.3
	Staphylococcus aureus N315	98	324	12.3
	Staphylococcus aureus COL	99	276	10.5
	Staphylococcus epidermidis ATCC12228	100	274	11.3
	Staphylococcus epidermidis RP62a	101	269	10.6
	Streptococcus mutans UAB159	102	240	12.2
	Streptococcus pneumoniae TIGR4	103	261	12.5

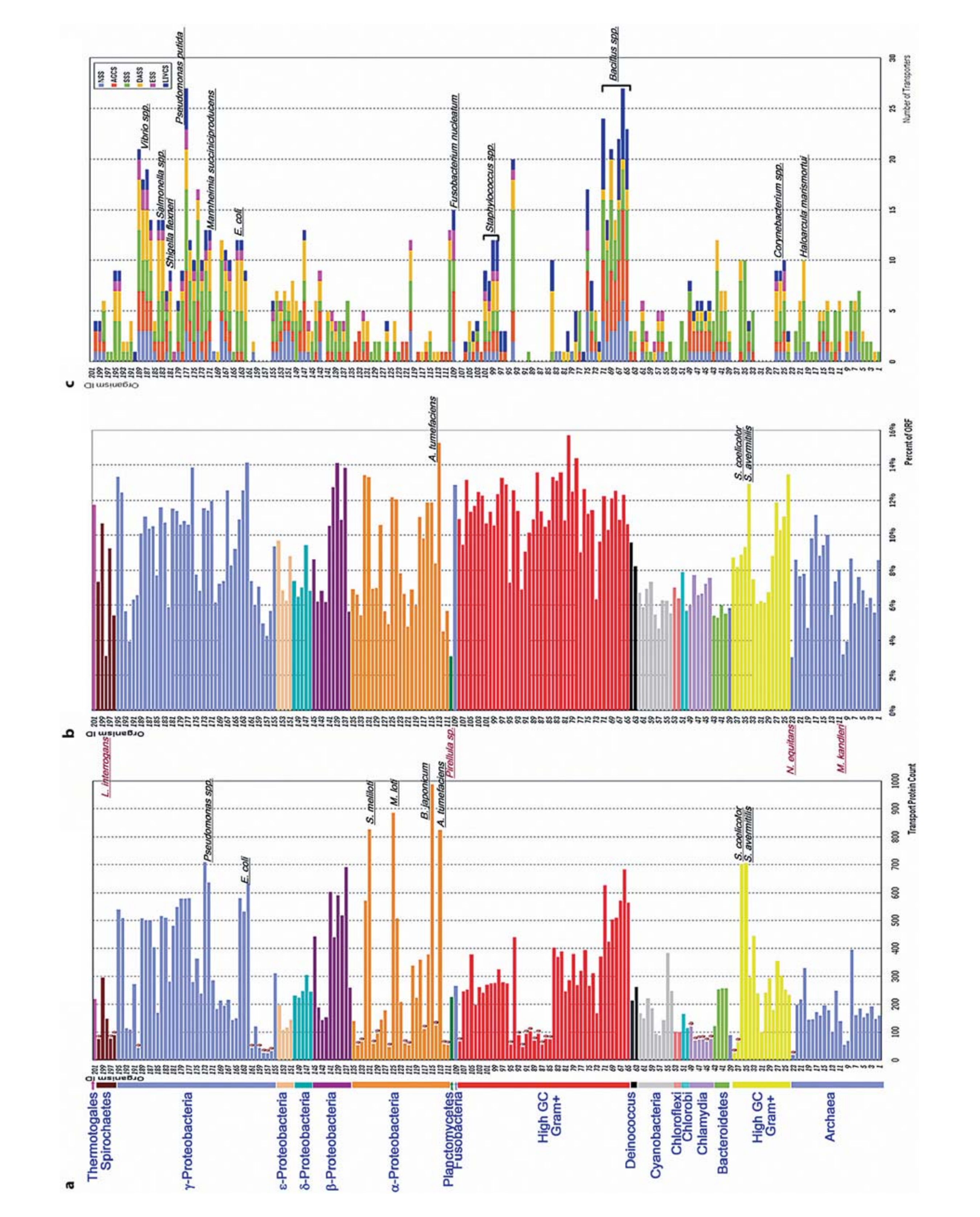
Table 1 (continued)

Taxonomy	Organism name	Organism ID	Total transport proteins	Percent of ORFs (%)
	Streptococcus pyogenes M1	104	198	11.7
	Symbiobacterium thermophilum IAM14863	105	378	11.3
	Streptococcus thermophilus CNRZ1066	106	252	13.2
	Thermoanaerobacter tengcongensis MB4	107	245	9.5
	Ureaplasma urealyticum serovar 3	108	67	10.9
Fusobacteria	*Fusobacterium nucleatum* ATCC25586	109	266	12.9
Planctomycetes	*Pirellula* sp. 1	110	225	3.1
α-Proteobacteria	*Anaplasma marginale* St. Maries	111	54	5.7
	Anaplasma phagocytophilum HZ	112	57	4.5
	Agrobacterium tumefaciens C58	113	824	15.3
	Bartonella henselae Houston-1	114	125	8.4
	Bradyrhizobium japonicum USDA110	115	987	11.9
	Brucella melitensis 16M	116	379	11.9
	Bartonella quintana Toulose	117	112	9.8
	Brucella suis 1330	118	360	11.0
	Caulobacter crescentus CB15	119	223	6.0
	Colwellia psychroerythraea 34H	120	338	6.9
	Ehrlichia chaffeensis Arkansas	121	53	4.8
	Ehrlichia ruminantium Welgevonden	122	59	6.6
	Gluconobacter oxydans 621H	123	208	7.8
	Jannaschia sp. CCS1	124	507	12.0
	Mesorhizobium loti MAFF303099	125	884	12.2
	Neorickettsia sennetsu Miyayama	126	46	4.9
	Nitrobacter winogradskyi Nb-255	127	177	5.7
	Candidatus Pelagibacter ubique HTCC1062	128	143	10.6
	Rickettsia conorii Malish7	129	96	7.0
	Rickettsia prowazekii MadridE	130	58	6.9
	Sinorhizobium meliloti 1021	131	827	13.3
	Silicibacter pomeroyi DSS-3	132	571	13.4
	Wolbachia pipientis wMel	133	65	5.4
	Wolbachia sp. TRS *(Brugia malayi)*	134	53	6.6
	Zymomonas mobilis ZM4	135	138	6.9
β-Proteobacteria	*Azoarcus* sp. EbN1	136	259	5.6
	Bordetella bronchiseptica RB50 NCTC-13252	137	691	13.8
	Burkholderia mallei ATCC23344	138	518	10.9
	Bordetella parapertussis 12822 NCTC-13253	139	590	14.1
	Bordetella pertussis Tohama I NCTC-13251	140	439	12.7
	Burkholderia pseudomallei K96243	141	603	10.5
	Nitrosomonas europaea ATCC19718	142	152	6.2
	Neisseria meningitidis MC58	143	142	6.8
	Nitrosococcus oceani ATCC19707	144	187	6.2
	Ralstonia solanacearum GMI1000	145	441	8.6
δ-Proteobacteria	*Bdellovibrio bacteriovorus* HD100	146	244	6.8
	Desulfotalea psychrophila LSv54	147	305	9.4
	Desulfovibrio vulgaris Hildenborough	148	247	7.0
	Geobacter sulfurreducens PCA	149	223	6.5
	Pelobacter carbinolicus DSM2380	150	230	7.4
ε-Proteobacteria	*Campylobacter jejuni* NCTC11168	151	144	8.8
	Helicobacter hepaticus ATCC51449	152	117	6.2
	Helicobacter pylori 26695	153	108	6.9
	Wolinella succinogenes	154	198	9.7

Table 1 (continued)

Taxonomy	Organism name	Organism ID	Total transport proteins	Percent of ORFs (%)
γ-Proteobacteria	*Acinetobacter* sp. ADP1	155	311	9.4
	Buchnera aphidicola Sg	156	31	5.7
	Buchnera aphidicola APS	157	24	4.3
	Buchnera aphidicola Bp	158	25	5.0
	Candidatus Blochmannia pennsylvanicus BPEN	159	43	7.0
	Coxiella burnetii RSA493	160	121	6.0
	Candidatus Blochmannia floridanus	161	43	7.4
	Erwinia carotovora SCRI1043	162	632	14.1
	Escherichia coli K12-MG1655	163	532	12.6
	Escherichia coli O157:H7 EDL933	164	580	10.9
	Francisella tularensis Schu4	165	148	9.2
	Haemophilus ducreyi 35000HP	166	142	8.3
	Haemophilus influenzae KW20	167	215	12.5
	Idiomarina loihiensis L2TR	168	194	7.4
	Legionella pneumophila Philadelphia 1	169	212	7.2
	Methylococcus capsulatus Bath	170	182	6.2
	Mannheimia succiniciproducens MBEL55E	171	285	12.0
	Pseudomonas aeruginosa PAO1	172	635	11.4
	Pseudomonas fluorescens Pf-5	173	708	11.5
	Pseudoalteromonas haloplanktis TAC125	174	238	6.8
	Photorhabdus lumines laumondii TTO1	175	363	7.8
	Photobacterium profundum SS9	177	580	10.6
	Pseudomonas syringae pv. tomato DC3000	179	579	10.6
	Rhodopseudomonas palustris CGA009	180	548	11.4
	Shigella flexneri 2a 301	181	481	11.5
	Shewanella oneidensis MR-1	182	281	5.9
	Salmonella typhi CT18	183	510	10.7
	Salmonella typhimurium LT2	184	516	11.6
	Thiomicrospira crunogena XCL-2	185	169	7.7
	Vibrio cholerae El Tor N16961	186	403	10.5
	Vibrio parahaemolyticus RIMD2210633	187	500	10.3
	Vibrio vulnificus CMCP6	188	501	11.0
	Vibrio vulnificus YJ016	189	507	10.1
	Wigglesworthia glossinidia P-endosymbiont	190	43	6.6
	Xanthomonas axonopodis pv. citri 306	191	272	6.3
	Xylella fastidiosa 9a5c	192	109	3.9
	Xylella fastidiosa Temecula1	193	115	5.7
	Yersinia pestis CO-92	194	508	12.4
	Yersinia pseudotuberculosis IP32953	195	539	13.3
Spirochaetes	*Borrelia burgdorferi* B31	196	89	5.4
	Borrelia garinii PB1	197	77	9.3
	Leptospira interrogans serovar lai56601	198	147	3.1
	Treponema denticola ATCC35405	199	295	10.7
	Treponema pallidum Nichols	200	76	7.3
Thermotogales	*Thermotoga maritima* MSB8	201	218	11.7

Fig. 1. The overall numbers of recognized transport proteins. Organisms from distinct phylogenetic groups are labeled with different colors. The obligate intracellular parasites/pathogens are marked with red stars. **a** Total number of transport proteins in 201 prokaryotes. **b** Transport proteins as the percentage of total ORFs. **c** Distribution of sodium-dependent amino acid/solute symporters across six families: NSS = light blue; AGCS = orange; SSS = salmon; DASS = lime; glutamate:sodium symporter family (ESS) = pink; LIVCS = dark blue.


a
b
c
Thermotogales
Spirochaetes
γ-Proteobacteria
ε-Proteobacteria
δ-Proteobacteria
β-Proteobacteria
α-Proteobacteria
Planctomycetes
Fusobacteria
High GC Gram+
Deinococcus
Cyanobacteria
Chloroflexi
Chlorobi
Chlamydia
Bacteroidetes
High GC Gram+
Archaea
Organism ID
L. interrogans
Pseudomonas spp.
E. coli
S. meliloti
M. loti
B. japonicum
A. tumefaciens
Pirellula sp.
S. coelicolor
S. avermitilis
N. equitans
M. kandleri
0
100
200
300
400
500
600
700
800
900
1000
Transport Protein Count
0%
2%
4%
6%
8%
10%
12%
14%
16%
Percent of ORF
NSS
ACCS
SSS
DASS
ESS
LIVCS
Vibrio spp.
Salmonella spp.
Shigella flexneri
Pseudomonas putida
Mannheimia succiniciproducens
E. coli
Fusobacterium nucleatum
Staphylococcus spp.
Bacillus spp.
Corynebacterium spp.
Haloarcula marismortui
0
5
10
15
20
25
30
Number of Transporters

vious studies show that archaeal species have much higher percentages of membrane proteins assigned to the role category of 'hypothetical proteins' than eubacterial species [Ren and Paulsen, 2005]. Some of these 'hypothetical proteins' could function in novel transport processes.

Although obligate intracellular organisms and small free-living parasites overall present the fewest transport proteins, they still devote a relatively high percentage of ORFs towards transport functions. For example, *Mycoplasma* spp. have over 10% of their ORFs encoding transport proteins. Transport proteins consist of average 7.7% of ORFs in 38 obligate intracellular and small free-living parasites, compared to an average of 8.9% in all organisms. Although most of these organisms appear to have undergone substantial reductive evolution, it seems that they have not preferentially lost or retained transporter genes as a consequence of their adaptation to intracellular lifestyles. Most of these organisms have various defective biosynthesis pathways, and have to uptake essential nutrients and intermediate metabolites from their host. Detailed examinations of transporter profiles show that these organisms have different degrees of reduction as to the types of transporters and categories of substrate. Compared to other prokaryotes, obligate intracellular organisms exhibit greater degree of reduction in efflux pumps and transporters for ions and small inorganic compounds. However, they appear to have retained a significant percentage of importers in the genome for essential nutrients and intermediate metabolites.

Phylogenetic Profiling as a Tool for Investigating Membrane Transport Content

The phylogenetic profile of a gene is a pattern representing the presence or absence of homologues in a set of fully sequenced genomes. Genes with similar phylogenetic profiles, as assessed by Pearson correlation coefficient, likely could function together in a pathway or are part of a complex because they are likely to evolve in a correlated fashion and tend to be either preserved or eliminated during evolution [Pellegrini et al., 1999; Pellegrini, 2001]. Phylogenetic profiling has many applications in genomics studies, such as detection of conserved core genes, lineage-specific gene family expansions [Vandepoele and Van de Peer, 2005], subcellular localization of proteins [Marcotte et al., 2000], prediction of physical and functional interactions and deduction of the functions of genes that have no well-characterized homologues [Levesque et al., 2003; Wu et al., 2005].

Previously we employed a novel application of phylogenetic profiling to investigate the presence or absence of transporter protein families across 141 sequenced prokaryotes and eukaryotes [Ren and Paulsen, 2005]. Compared to other studies, we used protein families rather than individual proteins as the unit of comparison. The phylogenetic profiling of transporter families provided interesting insights into the distribution of transporters across a broad range of organisms. Organisms from various phylogenetic groups which are adapted to similar environmental niches were often found in clusters. Inside each cluster, organisms were further grouped together by their phylogenetic history. Given that the profiling approach solely utilizes presence/absence of a transporter family and does not use sequence similarity directly, these findings suggest that the types of transporters utilized by an organism are related both to their physiology and to their evolutionary history.

The fast growing number of completely sequenced genomes enabled us to enhance the resolution of this phylogenetic profiling analysis. With the data on membrane transport systems from 201 fully sequenced prokaryotes, we were able to construct more detailed phylogenetic profiles for each transporter family (fig. 2). In line with our previous observations, hierarchical clustering of phylogenetic profiles showed a strong correlation between the observed clustering pattern and phylogeny, with distinct phylogenetic groupings of eubacteria and archaea clearly separated into different clusters, such as high GC Gram+, low GC Gram+, Proteobacteria, Chlamydia, Cyanobacteria, etc. (fig. 2, 3). Additionally, the clustering patterns are influenced by the lifestyle of organisms. The obligate intracellular pathogens/symbionts, the soil/plant-associated microbes and a collection of autotrophs are separated into distinct super-clusters (fig. 2, 3). These findings demonstrate that phylogenetic profiling is a viable and potent approach to the bioinformatic study of membrane transporters.

The obligate intracellular pathogens/symbionts cluster includes a group of phylogenetically diverse organisms (fig. 3b), including Chlamydia (pathogens); γ-Proteobacteria like *Buchnera* spp., *Wigglesworthia glossinidia* and *Candidatus Blochmannia* spp. (endosymbionts); α-Proteobacteria such as *Wolbachia* spp. (endosymbionts) and *Anaplasma* spp., *Ehrlichia* spp., *Rickettsia* spp., *Neorickettsia sennetsu*, *Bartonella* spp. (pathogens); low GC Gram+-like organisms *Mycoplasma* spp., *Ureaplasma urealyticum, Phytoplasma asteris* and *Tropheryma whipplei* (pathogens); Spirochetes like *Treponema pallidum*, *Borrelia* spp. (pathogens); and an archaeal endo-

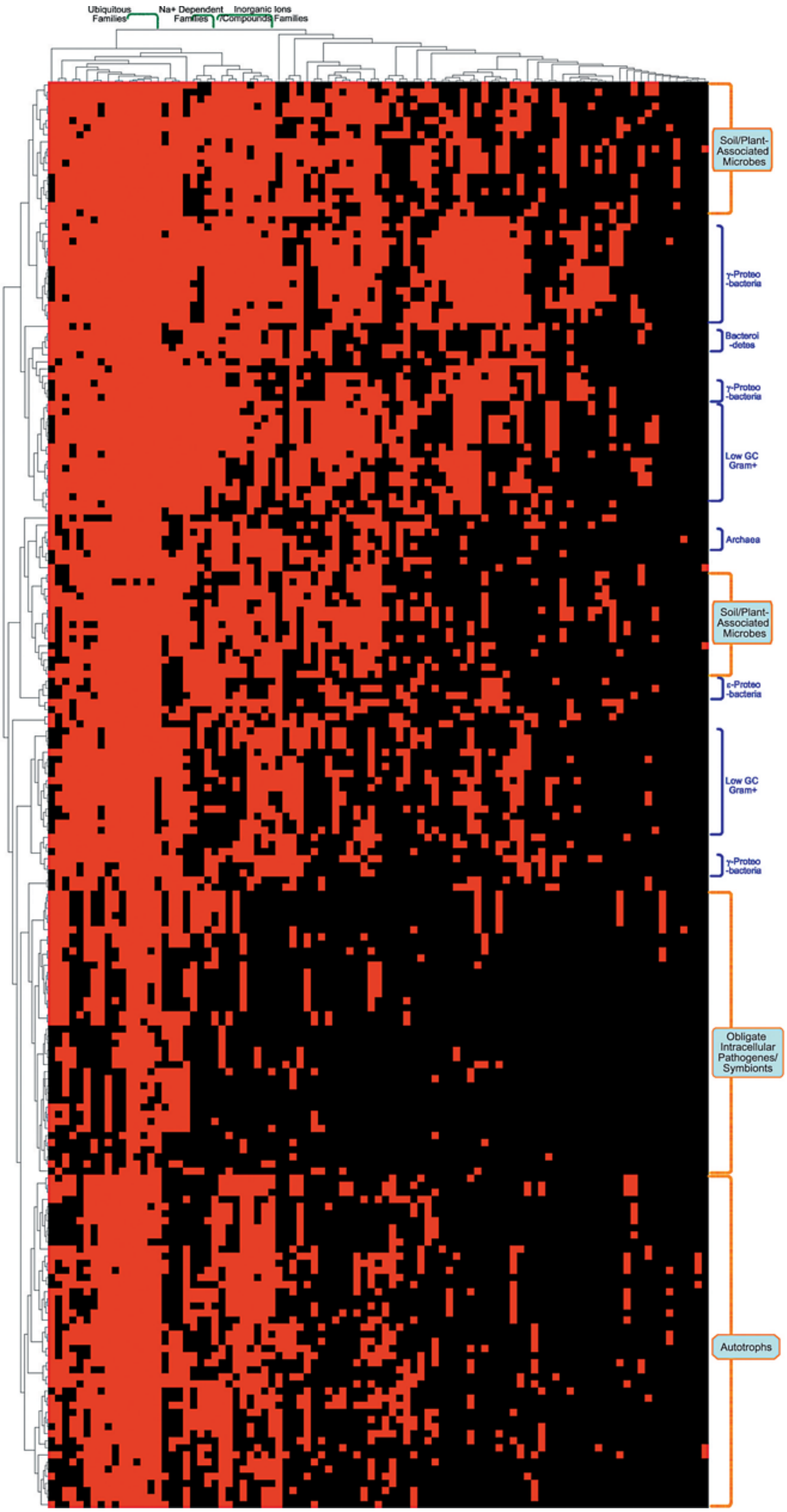


Fig. 2. Phylogenetic profiling of transporter families. Phylogenetic profiles were created for each transporter family. Each profile is a string with 201 entries (number of organisms analyzed). If a given family is present in an organism, the value '1' is assigned at this position (red color). If not, '0' is assigned (black color). Organisms and transporter families were clustered according to the similarity of their phylogenetic profiles.

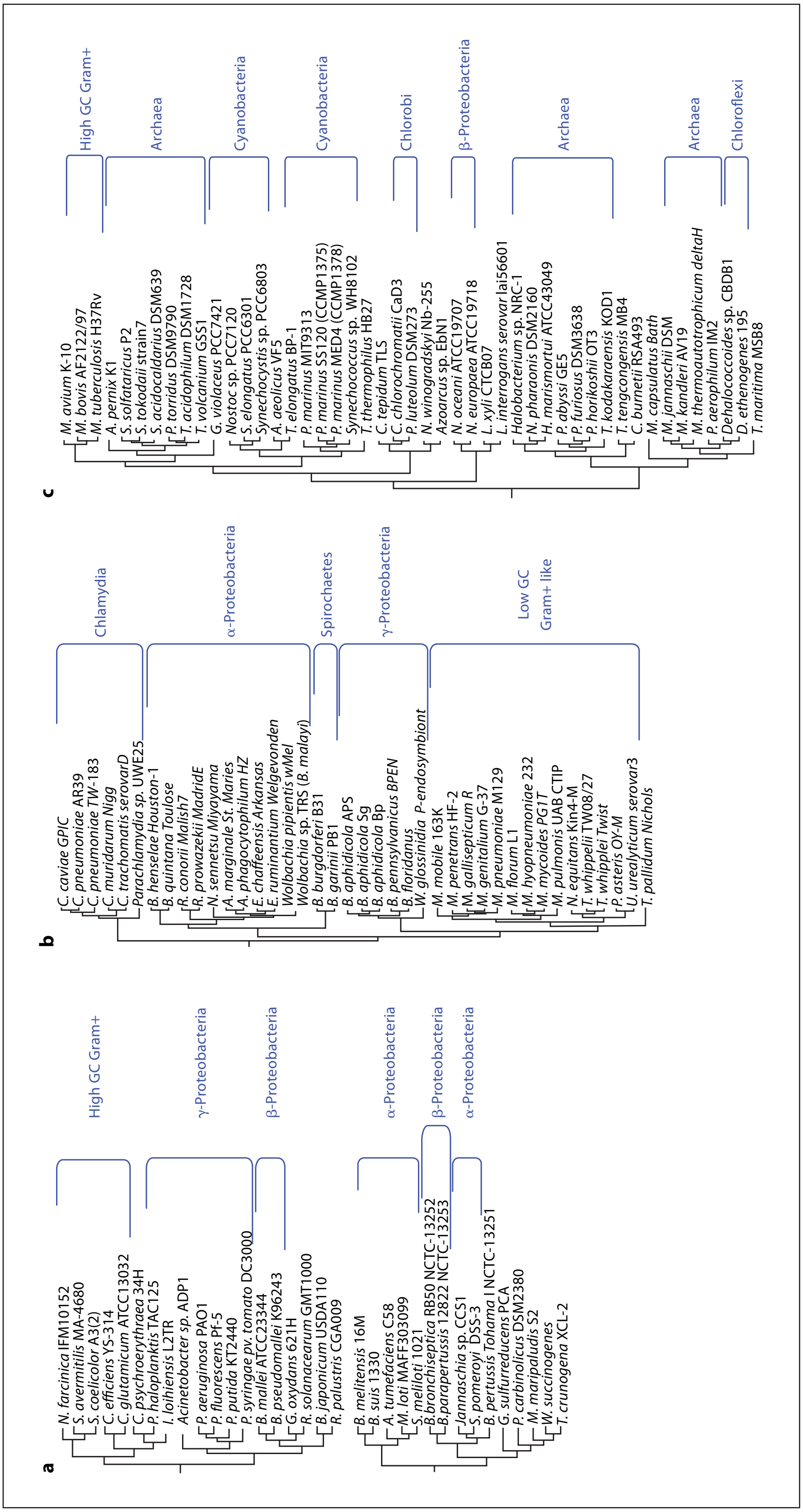


Fig. 3. Detailed view of three clusters of organisms generated by hierarchical clustering of their phylogenetic profiles of transporter families: soil/plant-associated microbes (**a**), obligate intracellular pathogens/symbionts (**b**) and autotrophs (**c**).

symbiont, *N. equitans*. Organisms in this cluster are mostly obligate intracellular organisms, with one or two exceptions, e.g. *Bartonella* spp. that are facultative intracellular pathogens. The transport needs for these obligate intracellular organisms are probably more specialized than those of environmental organisms due to the less dynamic nature of their intracellular environments. This may have allowed them to shed, for example, transporters for alternative nitrogen/carbon sources, drug/toxic metabolite efflux, osmoregulation, and ion homeostasis. The residual transport systems conserved in these obligate intracellular organisms probably belong to the core essential genes required for the acquisition of key nutrients and metabolic intermediates. For example, in *Rickettsia* species, genes coding for proteins functioning in glycolysis and the biosynthesis of S-adenosylmethionine and nucleotides are absent [Andersson and Andersson, 1999; Andersson et al., 1998; Dunning Hotopp et al., 2006; Ogata et al., 2001]. They completely rely on the hosts for these small molecules. As expected, transporter systems for the uptake of nucleoside monophosphates (ATP:ADP antiporter family), S-adenosylmethionine (drug/metabolite transporter superfamily) [Tucker et al., 2003], and glycerol-3-phosphate (MFS family) have been identified [Dunning Hotopp et al., 2006]. The essential glutamate transporters in two obligate endosymbionts *Candidatus Blochmannia floridanus* and *W. glossinidia* provides another example: The GltP glutamate:proton symporter (DAACS family) is encoded in *B. floridanus* [Tolner et al., 1995], while the GltJKL ABC transporter is expressed in *W. glossinidia* [Linton and Higgins, 1998]. Both of these organisms have a truncated TCA cycle which begins with α-ketoglutarate and ends with oxaloacetate [Zientz et al., 2004]. Their TCA cycle could be closed by the transamination of the imported glutamate to aspartate, catalyzed by an aspartate aminotransferase which uses oxaloacetate as a cosubstrate and produces α-ketoglutarate. Compared to the plant/soil-associated microbes, obligate intracellular organisms show a higher degree of variation in terms of energy coupling mechanism and transport mode. These variations may reflect the unique internal environment inside the host cells. All these observations illustrate how adaptation of an organism to certain living conditions leads to changes in its transporter repertoire and at the same time determine the set of transporters that the organism cannot afford to lose.

The soil/plant-associated microbes form two clusters, including organisms from various phylogenetic groups (fig. 3a). The first cluster includes Actinobacteria (*Corynebacterium* spp., *Nocardia farcinica* and *Streptomyces* spp.), γ-Proteobacteria *(Actinobacter* sp., *Idiomarina loihiensis*, *Pseudomonas* spp., *Pseudoalteromonas haloplanktis* and *Rhodopseudomonas palustris)*, and β-Proteobacteria *(Burkholderia* and *Ralstonia)*. The second one includes mainly α-Proteobacteria *(A. tumefaciens, Brucella* spp., *Jannaschia*sp., *M. loti, S. pomeroyi* and *S. meliloti)*, β-Proteobacteria (*Bordetella* spp.), δ-Proteobacteria *(Geobacter sulfurreducens* and *Pelobacter carbinolicus)*, and ε-Proteobacteria *(Wolinella succinogenes)*. This is in contrast to our previous analysis [Ren and Paulsen, 2005] in which these organisms formed one coherent cluster with two major branches comparable to the two clusters shown here. The first cluster is close to other γ-Proteobacteria like *E. coli* which has the highest diversity of transporter families among all prokaryotic organisms, partly due to the extensive experimental studies carried on this model organism. The second cluster is close to other δ-Proteobacteria and ε-Proteobacteria. Therefore, the respective phylogenetic relationships of these two clusters override the linkage by the influence of living environment on transporter contents as observed previously. One of the possible reasons causing the disparity could be the great expansion of γ-Proteobacteria species used in this study which may have exerted a stronger influence on the clustering. All of the organisms in these two clusters possess a robust collection of transporter systems. The similarity of phylogenetic profiles of organisms in these clusters probably reflects the versatility of these organisms and their exposure to a wide range of different substrates in their natural environment. The majority of species in this cluster can be free-living in the soil and some are capable of living in a diverse range of environments. They generally share a broad range of transport capabilities for plant-derived compounds specifically and for organic nutrients in general. Interestingly, some of the human facultative pathogens, such as *Bordetella* and *Brucella*, are also grouped in this cluster. These pathogens have close relatives that are soil or plant-associated environmental organisms [Parkhill et al., 2003; Paulsen et al., 2002], so their transport capabilities probably reflect a combination of their evolutionary heritage, original environmental niche and their current transport needs.

The third significant cluster of phylogenetic profiling of transporter families consists primarily of autotrophs (fig. 3c). This cluster was not found in our previous analysis [Ren and Paulsen, 2005] because of the lack of data on autotrophs. Obligate autotrophs obtain energy exclusively by the oxidation of inorganic substrates and use

CO_2 as the only resource of carbon [Kowalchuk and Stephen, 2001], such as the nitrifying bacteria *Nitrobacter winogradskyi* (oxidizing nitrite ion); *Nitrosomonas europaea* and *Nitrosococcus oceani* (oxidizing ammonium ion). Facultative autotrophs obtain some part of their energy from oxidation of iron, sulfur, hydrogen, nitrogen, and carbon monoxide. These include green sulfur bacteria *(Chlorobium* spp. and *Pelodictyon luteolum)*, green nonsulfur bacteria (*Dehalococcoides* spp.), both of which are anaerobic photosynthetic bacteria; Cyanobacteria *(Prochlorococcus* spp., *Synechococcus* spp., *Synechocystis* sp., *Nostoc* sp., *Gloeobacter violaceus* and *Thermosynechococcus elongates)* which are aerobic photosynthetic bacteria; a hydrogen-oxidizing, microaerophilic, obligate chemolithoautotrophs *(Aquifex aeolicus)*; an obligate methanotroph, *Methylococcus capsulatus*; and a group of autotrophic archaeal species (*Aeropyrum pernix, Sulfolobus* spp., *Picrophilus torridus, Thermoplasma* spp., *Methanobacterium thermoautotrophicum, M. kandleri, M. jannaschii, Pyrobaculum aerophilum, Pyrococcus* spp., *Thermococcus kodakaraensis, Natronomonas pharaonis, Haloarcula marismortui,* and *Halobacterium* sp.). In line with their metabolism features, organisms in this cluster generally lack transporters for carbohydrates, amino acids, carboxylates and nucleosides, etc. Instead, they encode a full array of transporters for various cations and anions, ammonium, inorganic phosphate, and sulfate which feed into their autotrophic metabolism. These features distinguish this group of autotrophs from organisms in the plant/soil-associated and intracellular pathogen/endosymbiont clusters. Interestingly, some heterotrophic bacteria were included in this cluster. They generally fall into several categories: Pathogens that are evolved from environmental organisms, like *Leifsonia xyli* and *L. interrogans*; organisms with extensive ion transport systems and/or few organic nutrient transporters, like *Thermoanaerobacter tengcongensis, Coxiella burnetii* and *Mycobacterium* spp., and a Thermotogales *(Thermotoga maritima)* with extensive array of archaeal-lineage genes [Nelson et al., 1999], and was found to cluster with the archaeal species in this super-cluster.

Comparison of the transporter profiles of marine microbes shows a close relationship between their transporter profiles and their physiology and ecological niches. The sequenced marine microbes to date can be categorized into three groups according to their metabolism and ecological niche: Cyanobacteria clade (photosynthetic autotrophs); Roseobacter clade (such as *Jannaschia* sp. and *S. pomeroyi*) that are metabolically versatile and capable of utilizing diverse organic and inorganic nutrients in the coastal and oceanic planktonic environment [Moran et al., 2004]; and a group of oligotrophic bacteria that are metabolically conservative and more specialized in scavenging organic nutrients in seawater [Button, 1991], such as *Oceanobacillus iheyensis, Vibrio vulnificus, I. loihiensis, Pelagibacter ubique* and *Photobacterium profundum.*

Cyanobacteria, which feature few importers for organic nutrients and a more substantial array of transporters for ion and inorganic compounds, belong to the autotroph cluster (fig. 3). Detailed examination of two Cyanobacteria species with different ocean environmental niches shows quite different transporter profiles [Palenik et al., 2006]: the coastal cyanobacterium, *Synechococcus* sp. strain CC9311, has a much larger capacity to transport, store, utilize or export metals, especially iron and copper than an open ocean oligotrophic strain, *Synechococcus* sp. strain WH8102, which could be related to its greater capacity to sense and respond to changes in its (coastal) environment. In contrast, WH8102 has systems predicted for the efflux of arsenite and chromate [Palenik et al., 2003] that are not found in CC9311. The Roseobacter clade, however, was clustered with plant/soil-associated clusters (fig. 3) due to their abundant and diverse transporters for both organic nutrients (peptides, amino acids, sugars, putrescine and spermidine, taurine, glycine betaine and dimethylsulphoniopropionate, etc.) and inorganic compounds (urea, phosphate, inorganic ions, sulfate, etc.) which enable them to take advantage of transient occurrences of high-nutrient niches within a bulk low-nutrient environment. One of the distinguishing features of Roseobacters are their uncommonly high number of TRAP transporter systems (26 systems for *S. pomeroyi* and 28 for *Jannaschia* sp., no other sequenced genome has more), probably reflecting their capability to import carboxylic acids produced in surface waters during photo-oxidation of dissolved organic matters, like glyoxylate and acetate [Moran et al., 2004]. The metabolically conservative marine heterotrophs did not form any distinct grouping and were clustered primarily by their phylogenetic traits. For example, *O. iheyensis* was clustered with other *Bacillus* spp.; and *V. vulnificus* and *P. profundum* were clustered with other *Vibrio* spp. *P. ubique* represents one of the smallest free-living nonparasitic microorganism [Giovannoni et al., 2005] with 1,354 ORFs, of which 143 encode transport proteins (10.6%). Compared to obligate intracellular organisms, it encodes a large number of transporters for diverse nitrogenous compounds, such as ammonium, urea, basic amino acids, spermidine, and putrescine. These features clearly exclude it from the obligate intracellular organism cluster.

The clustering of transporter families also show features related to the lifestyles of organisms. The ubiquitous families, like ABC, MFS, P(F)-type ATPase, which are present in virtually every organism we analyzed, are clustered together. A group of sodium ion-dependent transporter families, the neurotransmitter:sodium symporter (NSS), alanine/glycine:cation symporter (AGCS), solute:sodium symporter (SSS), and divalent anion:sodium symporter (DASS) are clustered together. Transporters in these families are all symporters which utilize the sodium ion gradient to transport amino acid, solute, and/or divalent ions to cytoplasm. This clustering may suggest that these families co-occur in a specific set of organisms, presumably those most reliant on sodium ion-driven transport. Figure 1d shows the detailed distribution of six sodium ion-dependent amino acid/solute transporter families in the 201 prokaryotic organisms we analyzed. We see considerable variation in the distribution of these families among phylogenetically related species. For example, *Mycobacterium* spp. and *Corynebacterium* spp. are closely related Actinobacteria. *M. tuberculosis* and *C. diphtheriae* are both pathogens of human respiratory systems. *Corynebacterium* spp. encode members of all six sodium-dependent transporter families, while *Mycobacterium* spp. have none. In fact, *Mycobacterium* spp. were clustered with the autotrophic bacteria as an artifact on our phylogenetic profiling studies (fig. 3c) at least in part due to their lack of sodium-dependent transporters.

In general, environmental organisms such as *Bacillus* spp. (including *Oceanbacillus* and *Lactobacillus* spp.), *Colwellia psychroerythraea*, *Pirellula* sp. and *Pseudomonas* spp. present the highest number of sodium-dependent pumps, while organisms with autotrophic lifestyles encode very few sodium ion-driven transporters, and those they do possess are more likely involved in the uptake of simple compounds such as sulfate rather than amino acids or carboxylates. Some of these autotrophs completely lack this type of transporters, such as *Dehalococcoides* spp., *M. kandleri, N. winogradskyi,* and *N. europaea*. There are a couple of interesting exceptions: *H. marismortui*, a halophilic microorganism that thrives in extreme saline environments, encodes 10 members of sodium-dependent transporters in NSS, SSS and DASS families, the highest number among all archaeal species studied. *Halobacterium* sp., another archaeal organism, which, like *H. marismortui*, proliferates in saturating salt solutions, also has 6 members of such transporters. These probably reflect their adaptation to a high-salt environment. A group of human pathogens encode relatively large numbers of sodium-dependent pumps, including Enterobacteriaceae (such as *E. coli*, *Salmonella*, *Shigella*, *Yersinia* and *Vibrio* spp.), *Staphylococcus* spp., *Corynebacterium* spp. and *Fusobacterium nucleatum*, etc. These could also be related to the high-salt environment in human GI tract, oral cavity and respiratory tract. Actually human epithelial cells utilize the same mechanism to uptake nutrients from the GI tract and to regulate the internal homeostasis. Among those organisms with obligate intracellular lifestyles, which need to obtain nutrients like amino acids from their host, the majority do not encode sodium-dependent amino acid transporters. Instead, they typically encode ABC family amino acid transporter and/or APC family amino acid:proton symporters or amino acid:amino acid antiporters. *Chlamydia* spp. are the only group of obligate intracellular organisms that show homologues in each of the six sodium-dependent amino acid/solute families.

There is another cluster of families for inorganic ions and small compounds, including potassium and chloride ion channels, ammonium transporter, inorganic phosphate transporter, sulfate permease, and calcium: cation antiporter. Autotrophic eubacterial and archaeal organisms generally utilize transporters in these families for the uptake of inorganic compounds, as well as soil/plant-associated microbes and other environmental organisms. As expected, obligate intracellular pathogens and endosymbionts generally lack this type of transporters.

Conclusion

The era of complete genome sequencing has opened new horizons in our understanding of complex biological questions. Comparative genomic approaches for the analysis of membrane transport systems have provided us invaluable insights on how microbes adapt to their environment. The observations that organisms with similar lifestyles and/or ecologic niches (obligate intracellular, soil/plant-associated, or autotrophic) display similar phylogenetic profiles despite their phylogenetic differences strongly suggest the influence of their environments on their membrane transport gene complement.

References

Abramson J, Smirnova I, Kasho V, Verner G, Iwata S, Kaback HR: The lactose permease of *Escherichia coli*: overall structure, the sugar-binding site and the alternating access model for transport. FEBS Lett 2003;555:96–101.

Andersson JO, Andersson SG: Genome degradation is an ongoing process in *Rickettsia*. Mol Biol Evol 1999;16:1178–1191.

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 1998;396:133–140.

Bernal A, Ear U, Kyrpides N: Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucl Acids Res 2001;29:126–127.

Bohm A, Diez J, Diederichs K, Welte W, Boos W: Structural model of MalK, the ABC subunit of the maltose transporter of *Escherichia coli*: implications for *mal* gene regulation, inducer exclusion, and subunit assembly. J Biol Chem 2002;277:3708–3717.

Boos W, Shuman H: Maltose/maltodextrin system of *Escherichia coli*: transport, metabolism, and regulation. Microbiol Mol Biol Rev 1998;62:204–229.

Button DK: Biochemical basis for whole-cell uptake kinetics: specific affinity, oligotrophic capacity, and the meaning of the Michaelis constant. Appl Environ Microbiol 1991;57:2033–2038.

Dunning Hotopp JC, Lin M, Madupu R, Crabtree J, Angiuoli SV, Eisen J, Seshadri R, Ren Q, Wu M, Utterback TR, Smith S, Lewis M, Khouri H, Zhang C, Niu H, Lin Q, Ohashi N, Zhi N, Nelson W, Brinkac LM, Dodson RJ, Rosovitz MJ, Sundaram J, Daugherty SC, Davidsen T, Durkin AS, Gwinn M, Haft DH, Selengut JD, Sullivan SA, Zafar N, Zhou L, Benahmed F, Forberger H, Halpin R, Mulligan S, Robinson J, White O, Rikihisa Y, Tettelin H: Comparative genomics of emerging human *Ehrlichiosis* agents. PLoS Genet 2006;2:e21.

Elferink MG, Driessen AJ, Robillard GT: Functional reconstitution of the purified phosphoenolpyruvate-dependent mannitol-specific transport system of *Escherichia coli* in phospholipid vesicles: coupling between transport and phosphorylation. J Bacteriol 1990;172:7119–7125.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ: Genome streamlining in a cosmopolitan oceanic bacterium. Science 2005;309:1242–1245.

Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA: Genome coverage, literally speaking: the challenge of annotating 200 genomes with 4 million publications. EMBO Reports 2005;6:397–399.

Kowalchuk GA, Stephen JR: Ammonia-oxidizing bacteria: a model for molecular microbial ecology. Annu Rev Microbiol 2001;55:485–529.

Levesque M, Shasha D, Kim W, Surette MG, Benfey PN: Trait-to-gene: a computational method for predicting the function of uncharacterized genes. Curr Biol 2003;13:129–133.

Linton KJ, Higgins CF: The *Escherichia coli* ATP-binding cassette (ABC) proteins. Mol Microbiol 1998;28:5–13.

Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D: Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci USA 2000;97:12115–12120.

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye W, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren Q, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J, Ward N: Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 2004;432:910–913.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM: Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 1999;399:323–329.

Newman MJ, Foster DL, Wilson TH, Kaback HR: Purification and reconstitution of functional lactose carrier from *Escherichia coli*. J Biol Chem 1981;256:11804–11808.

Ogata H, Audic S, Renesto-Audiffren P, Fournier PE, Barbe V, Samson D, Roux V, Cossart P, Weissenbach J, Claverie JM, Raoult D: Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 2001;293:2093–2098.

Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, Paulsen I, Dufresne A, Partensky F, Webb EA, Waterbury J: The genome of a motile marine *Synechococcus*. Nature 2003;424:1037–1042.

Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, Badger JH, Madupu R, Nelson WC, Brinkac LM, Dodson RJ, Durkin AS, Daugherty SC, Sullivan SA, Khouri H, Mohamoud Y, Halpin R, Paulsen IT: Genome sequence of *Synechococcus* CC9311:Insights into adaptation to a coastal environment. Proc Natl Acad Sci USA 2006;103:13555–13559.

Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MTG, Churcher CM, Bentley SD, Mungall KL, Cerdeno-Tarraga AM, Temple L, James K, Harris B, Quail MA, Achtman M, Atkin R, Baker S, Basham D, Bason N, Cherevach I, Chillingworth T, Collins M, Cronin A, Davis P, Doggett J, Feltwell T, Goble A, Hamlin N, Hauser H, Holroyd S, Jagels K, Leather S, Moule S, Norberczak H, O'Neil S, Ormond D, Price C, Rabbinowitsch E, Rutter S, Sanders M, Saunders D, Seeger K, Sharp S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Unwin L, Whitehead S, Barrell BG, Maskell DJ: Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 2003;35:32–40.

Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr: Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J Mol Biol 2000;301:75–100.

Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Umayam L, Brinkac LM, Beanan MJ, Daugherty SC, Deboy RT, Durkin AS, Kolonay JF, Madupu R, Nelson WC, Ayodeji B, Kraul M, Shetty J, Malek J, Van Aken SE, Riedmuller S, Tettelin H, Gill SR, White O, Salzberg SL, Hoover DL, Lindler LE, Halling SM, Boyle SM, Fraser CM: The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci USA 2002;99:13148–13153.

Paulsen IT, Sliwinski MK, Saier MH Jr: Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol 1998;277:573–592.

Pellegrini M: Computational methods for protein function analysis. Curr Opin Chem Biol 2001;5:46–50.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 1999;96:4285–4288.

Postma PW, Lengeler JW, Jacobson GR: Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol Rev 1993;57:543–594.

▶Ren Q, Kang KH, Paulsen IT: TransportDB: a relational database of cellular membrane transport systems. Nucleic Acids Res 2004; 32:D284–D288.

Ren Q, Paulsen IT: Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. PLoS Comput Biol 2005;1:190–201.

Saier MH Jr: Classification of Transmembrane Transport Systems in Living Organisms. San Diego, Academic Press, 1999.

▶Saier MH Jr: A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 2000; 64:354–411.

▶Sweet G, Gandor C, Voegele R, Wittekindt N, Beuerle J, Truniger V, Lin EC, Boos W: Glycerol facilitator of *Escherichia coli*: cloning of *glpF* and identification of the *glpF* product. J Bacteriol 1990;172:424–430.

▶Tolner B, Ubbink-Kok T, Poolman B, Konings WN: Cation-selectivity of the L-glutamate transporters of *Escherichia coli*, *Bacillus stearothermophilus* and *Bacillus caldotenax*: dependence on the environment in which the proteins are expressed. Mol Microbiol 1995;18:123–133.

▶Tucker AM, Winkler HH, Driskell LO, Wood DO: S-adenosylmethionine transport in *Rickettsia prowazekii*. J Bacteriol 2003;185: 3031–3035.

▶Vandepoele K, Van de Peer Y: Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol 2005;137:31–42.

▶Viitanen P, Newman MJ, Foster DL, Wilson TH, Kaback HR: Purification, reconstitution, and characterization of the *lac* permease of *Escherichia coli*. Methods Enzymol 1986; 125:429–452.

Wu M, Ren Q, Durkin AS, Daugherty SC, Brinkac LM, Dodson RJ, Madupu R, Sullivan SA, Kolonay JF, Haft DH, Nelson WC, Tallon LJ, Jones KM, Ulrich LE, Gonzalez JM, Zhulin IB, Robb FT, Eisen JA: Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. PLoS Genet 2005;1:e65.

▶Zientz E, Dandekar T, Gross R: Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev 2004;68:745–770.

Editorial

Journal of
Molecular Microbiology
and Biotechnology

J Mol Microbiol Biotechnol 2006;11:289–290
DOI: 10.1159/000095630

The Bioinformatic Study of Transmembrane Molecular Transport

Membrane transporters are the cell's equivalent of delivery vehicles, garbage disposals, and communication systems – proteins that span the cell membrane and form an intricate system of pumps and channels through which they deliver essential nutrients, eject waste products, and assist the cell to sense environmental conditions. Membrane transport systems play indispensable roles in the fundamental cellular processes of all organisms. Knowledge of the suite of transporters present in an organism sheds light on its lifestyle and physiological adaptations. Until recently, analyses of membrane transporters have been limited primarily to the examination of transporter genes in individual organisms. However, with the advent of the genomics era, comprehensive bioinformatic comparisons of predicted membrane transporters across a range of organisms in all three domains of life have become possible.

New computational application of the phylogenetic profiling approach to cluster organisms together that appear to have similar suites of transporters provides an even more recent advance. For example, obligate intracellular pathogens and endosymbionts possess limited numbers of transport systems in spite of the massive metabolite fluxes one would expect between the pathogens or symbionts and their hosts. This is believed to be due to the relatively static nature of their intracellular environments which require minimal degrees of adaptation, particularly to stress conditions. Limited types of nutrients need to be taken up, which are provided by the host in nearly constant amounts. Because the host provides a homeostatic home, the bacteria can streamline their genomes and eliminate excess baggage. The consequence is a bacterium that can live only in one or a few host organisms.

To provide the research community as well as the general public with easy access to the extensive information about transporters, web-based databases have been created. These include the Transporter Classification Database (TCDB) (http://www.tcdb.org) which classifies transport proteins based on both function and phylogeny, and the TransportDB Database (http://www.membranetransport.org/) which surveys fully sequenced genomes for genes encoding transport proteins. In the latter database, the transporter profiles of each sequenced organism are available to view, search, compare, and download in an easy-to-navigate format. Extensive links and references are also provided on the site. This is the only active database dedicated to the comprehensive, comparative study of membrane transporters in different organisms with fully sequenced genomes. In the former database (TCDB), over 400 families of transporters are described, and functional data for representative well-characterized members of these families are provided. When possible, these families have been grouped into superfamilies that define the evolutionary relationships between individual families. Web-based search tools and useful bioinformatic software packages render TCDB user-friendly. There is also a major section devoted to poorly characterized families of transport or putative transport proteins where more studies are needed. TCDB and TransportDB are tremendous resources both for research scientists and for students and their teachers. Thousands of researchers visit these databases on the web regularly and have applied the information provided to their research efforts. The availability of these databases has inspired numerous laboratory studies on membrane transporters, promoting the rapid expansion of this area of research.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1464–1801/06/0116–0289$23.50/0

Accessible online at:
www.karger.com/mmb

Out of the hundreds of thousands of proteins encoded in bacterial, archaeal and eukaryotic genomes, over 10% function in transport. Transport systems employ a large variety of mechanisms to import various ions and nutrients and to prevent the excessive build-up of other ions, end products of metabolism, toxins and drugs. It is hoped that someday, the identification of the complete complement of transporters in sequenced genomes will become possible.

Currently over 300 fully sequenced genomes are available for analysis. Many of the encoded transporters have been classified into different families, and their functions have been predicted. Interdisciplinary expertise in genomics and bioinformatics allows the development of computational tools and models that should expedite, improve and advance traditional biological studies of transporters.

Genome sequencing projects have had a tremendous impact on medical and environmental microbiology. For example, the genome sequencing of a methanotrophic bacterium, *Methylococcus capsulatus*, which feeds off of methane, one of the major greenhouse gases, is of importance because of its potential to mitigate global warming, a problem that we will face on an ever increasingly intense scale in the future. The genome sequence of this bacterium has allowed biologists and environmental bioengineers to focus on an understanding of the parameters of this organism in an effort to utilize it and other bacteria for solutions to our real-life environmental problems.

M. capsulatus relies heavily on its large repertoire of metal cation pumps (12 P-type cation ATPases, 4 of which have a copper-binding P-ATPase motif) to take up copper ions for the regulation of methane oxidation, a critical step in methane metabolism. In contrast, this organism possesses very few transport systems for organic carbon compounds, which emphasizes the importance of methane as its sole carbon source for energy production and growth. This is yet another example illustrating the relevance of transport protein profiles to the overall physiology of organisms. These findings, combined with bioinformatic analyses of metabolic features, deepen our understanding of methanotrophic lifestyles and emphasize this bacterium's potential for biotechnological applications that could lead to environmental improvements.

In conclusion, studies at all levels, including genetic, biochemical, biophysical and physiological, allow computational microbiologists to view a living organism as a complete system communicating with its environment. The implications with respect to improvement of the environment and global biosphere healthcare systems are staggering. Novel techniques will allow the more rapid advance of scientific discovery leading to solutions to our immense environmental and health problems.

Milton H. Saier, Jr., Editor-in-Chief
Qinghu Ren, Staff Scientist
The Institute for Genomic Research (TIGR)

References

Busch, W. and Saier, M.H., Jr. (2002). The Transporter Classification (TC) System, 2002. CRC Crit Rev Biochem Mol Biol 37:287–337.

Chang, A.B., Lin, R., Studley, W.K., Tran, C.V., and Saier, M.H., Jr. (2004). Phylogeny as a guide to structure and function of membrane transport proteins. Mol Membr Biol 21:171–181.

Paulsen, I.T., Nguyen, L., Sliwinski, M.K., Rabus, R., and Saier, M.H., Jr. (2000a). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J Mol Biol 301:75–100.

Ren, Q., Kang, K.H., and Paulsen, I.T. (2004). TransportDB: a relational database of cellular membrane transport systems. Nucleic Acids Res 32:D284–D288 (database issue).

Ren, Q. and Paulsen, I.T. (2005). Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. PLoS Comput Biol 1:190–201.

Saier, M.H., Jr. (2000). A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64:354–411.

Saier, M.H., Jr., Tran, C.V., and Barabote, R.D. (2006). TCDB: The transporter classification database for membrane transport protein analyses and information. Nucl Acids Res 34:D181–D186 (database issue).

Ward, N., Larsen, O., Sakwa, J., Bruseth, L., Khouri, H., Durkin, A.S., Dimitrov, G., Jiang, L., Scanlan, D., Kang, K.H., Lewis, M., Nelson, K.E., Methe, B., Wu, M., Heidelberg, J.F., Paulsen, I.T., Fouts, D., Ravel, J., Tettelin, H., Ren, Q., Read, T., DeBoy, R.T., Seshadri, R., Salzberg, S.L., Jensen, H.B., Birkeland, N.K., Nelson, W.C., Dodson, R.J., Grindhaug, S.H., Holt, I., Eidhammer, I., Jonasen, I., Vanaken, S., Utterback, T., Feldblyum, T.V., Fraser, C.M., Lillehaug, J.R., and Eisen, J.A. (2004). Genomic insights into methanotrophy: the complete genome sequence of *Methylococcus capsulatus* (Bath). PLoS Biol 2:e303 (Epub).

D274–D279 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 28 November 2006*
doi:10.1093/nar/gkl925

TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels

Qinghu Ren, Kaixi Chen and Ian T. Paulsen*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 8, 2006; Revised October 16, 2006; Accepted October 17, 2006

ABSTRACT

TransportDB (http://www.membranetransport.org/) is a comprehensive database resource of information on cytoplasmic membrane transporters and outer membrane channels in organisms whose complete genome sequences are available. The complete set of membrane transport systems and outer membrane channels of each organism are annotated based on a series of experimental and bioinformatic evidence and classified into different types and families according to their mode of transport, bioenergetics, molecular phylogeny and substrate specificities. User-friendly web interfaces are designed for easy access, query and download of the data. Features of the TransportDB website include text-based and BLAST search tools against known transporter and outer membrane channel proteins; comparison of transporter and outer membrane channel contents from different organisms; known 3D structures of transporters, and phylogenetic trees of transporter families. On individual protein pages, users can find detailed functional annotation, supporting bioinformatic evidence, protein/DNA sequences, publications and cross-referenced external online resource links. TransportDB has now been in existence for over 10 years and continues to be regularly updated with new evidence and data from newly sequenced genomes, as well as having new features added periodically.

INTRODUCTION

Membrane transporters are a large group of proteins that span the cell membrane and form an intricate system of pumps and channels through which they deliver essential nutrients, eject waste products and assist the cell to sense environmental conditions. Transporters represent a large and diverse group of proteins that differ in membrane topology, energy coupling mechanism and substrate specificities. They play indispensable roles in the fundamental cellular processes of all organisms (1).

With the advent of the genomics era, comprehensive genome-wide bioinformatic comparisons of predicted membrane transporters across a range of organisms in all three domains of life have become possible. Previously, we have reported a series of comparative analyses of transport systems in a collection of prokaryotic and eukaryotic organisms (2–4). We started a web portal showing our bioinformatic prediction of transporters in sequenced genomes back in 1996, and have had a continual web presence since then. The current incarnation of TransportDB dates back to 2002, when we moved to a relational database structure and greatly enhanced the available features (5). The aim of TransportDB is to present the comprehensive transporter profiles of each sequenced prokaryotic and eukaryotic organisms, as well as to provide comparative and phylogenetic tools to view, search, compare and download the transporter data in an easy-to-navigate format. We describe in this paper the data content and web features of TransportDB, with a focus on the recent additions and improvements.

DATABASE STRUCTURE AND CONTENT

TransportDB uses a relational database to store all the data associated with membrane transporters and outer membrane channels. It was built specifically to hold many different genomes and to allow cross-genome queries and comparisons. TransportDB database consists of 20 tables and is implemented in MySQL (http://www.mysql.com/). Data stored in TransportDB database can be accessed using Structured Query Language (SQL). Users can search TransportDB via a web interface which facilitates building a custom query without interacting directly with the database. Examples include searches for transporter class, family, protein and substrate. The relational database format also allows users

*To whom correspondence should be addressed. Tel: +1 301 795 7531; Fax: +1 301 838 0208; Email: ipaulsen@tigr.org
Present address:
Kaixi Chen, Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA

to select a subset of organisms of their interest and compare the overall transporter features as well as each individual transporter family.

TransportDB adopts a '3-tier' database architecture. The top tier of TransportDB is the user web interface. It sends the requests for data, formats the output of the query and displays it on the web. The Application Programmer Interface (API), or middle tier, connects the database and retrieves datasets by explicit keys (e.g. the name of a transporter gene) using a single query. We adopt PHP (http://www.php.net/), a widely used server-side scripting language, as our API. The database itself is the bottom tier. This architecture enables the web application to utilize SQL to query the database, while not limiting the top tier to any specific database system.

TransportDB stores the complete array of predicted transporters and outer membrane channels from various prokaryotic and eukaryotic organisms, with detailed information and supporting evidence for each protein. Figure 1 shows the evolution of data collection in TransportDB. When we moved to a relational database format in 2002, we had bioinformatic analyses of 54 organisms and over 10 000 transporter genes (5). During the past 4 years, a very considerable number of organisms have been added to the database and the total number of annotated transporter genes has increased by a factor of seven. Currently, TransportDB contains data from 248 organisms, including 197 bacteria, 24 archaea and 27 eukaryota. This collection of organisms represents a broad phylogenetic diversity. A total of 71 659 transport proteins and 4790 outer membrane channels have been annotated and assigned to 183 families according to the TC classification system (1,6). These families are further classified into different types according to their transport mode and energy coupling mechanism: seven families of primary active transporter that couple the transport process with a primary energy source (e.g. ATP hydrolysis); 83 families of secondary transporter that utilize an ion or solute electrochemical gradient; 33 families of energy-independent channels; two families of group translocators which modify their substrates during transport; 38 families of porins/outer membrane channels that are prevalent in the outer membrane of Gram-negative bacteria and certain eukaryotic organelles; and 11 families with an unknown transport mechanism. Transporters are unevenly distributed among these families: some are very large superfamilies with thousands of members, such as the ABC superfamily (7) (32 099 proteins annotated) and the MFS superfamily (8) (7942 proteins), both of which are widely distributed across prokaryotic and eukaryotic species; some families, however, only exist in a very limited phylogenetic spectrum and/or are present in only limited numbers.

DATABASE ACCESS AND WEB FEATURES

TransportDB is accessible online at http://www.membranetransport.org/.

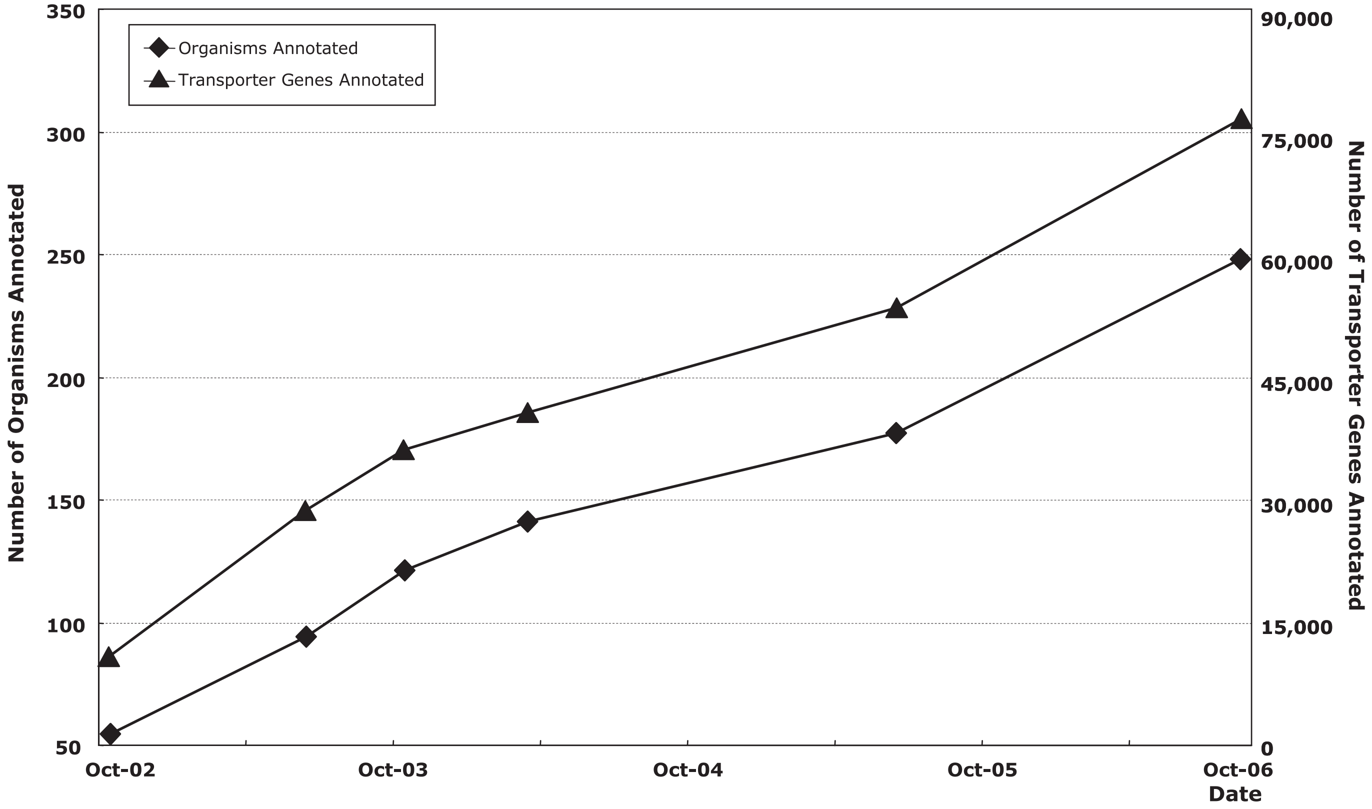
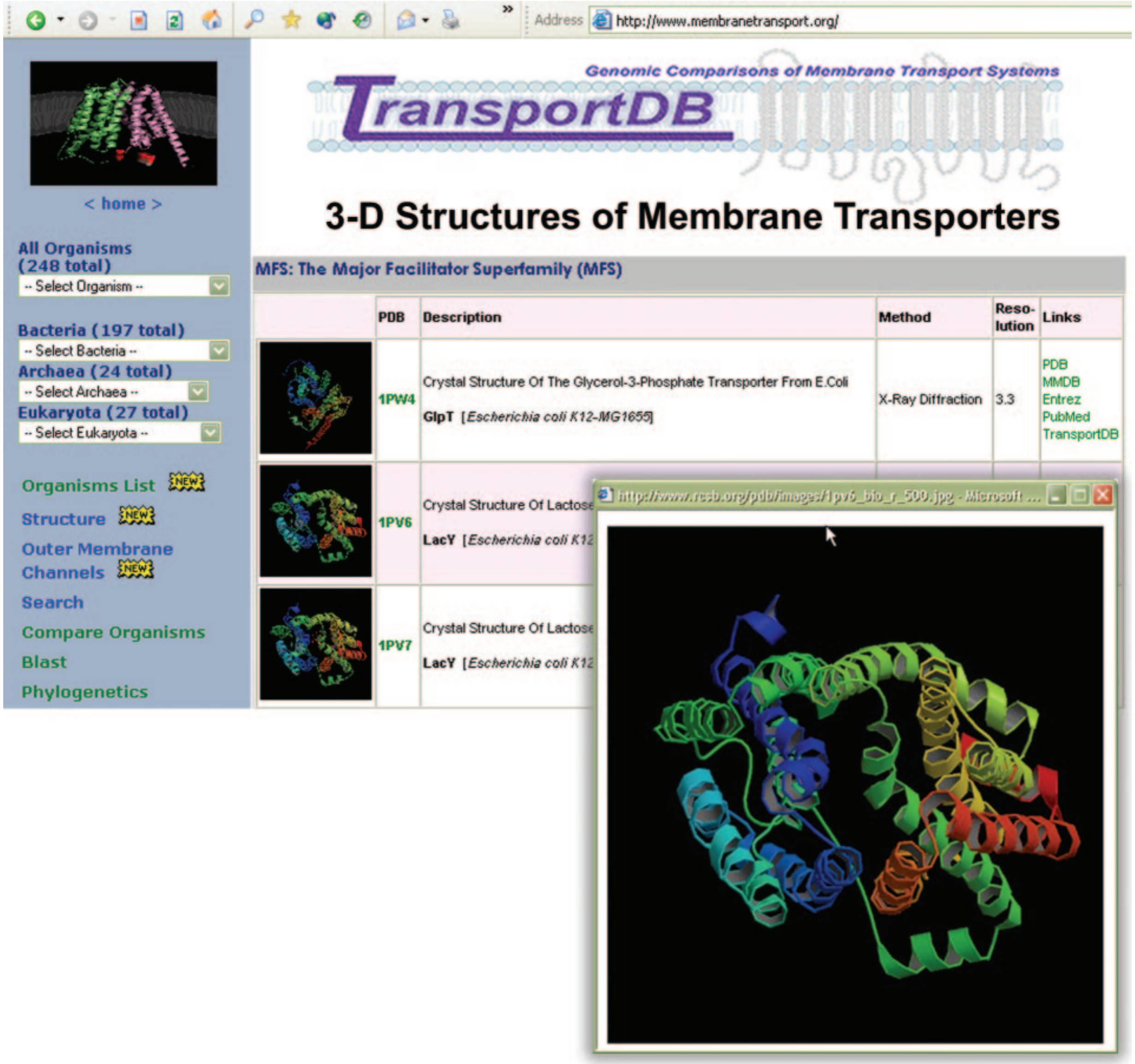


Figure 1. The evolution of TransportDB data storage. The rhomboid points represent number of organisms annotated over 4 years. The triangle points show the number of transporter genes annotated.

There are several ways for users to access data stored in TransportDB. Users can browse the database by selecting the organism from drop-down boxes on the left of the web page, or by clicking the links from the 'Organism List' page (Figure 2). All the transport proteins are listed in a tabular format with predicted substrate or function. A hierarchical top-down structure was deployed for easy data access, which arranges transporters in the orders of kingdom (bacteria, archaea or eukaryota), organism, transporter type, transporter family/subfamily and transport protein. Each transport protein is presented in separate web pages where users can find detailed information such as transporter substrate/function annotation, TC classification, transmembrane segment prediction by TMHMM (9), genomic locus information, protein/DNA sequence, etc. Evidence types associated with functional annotation are included, such as hidden Markov models [Pfam (10) and TIGRfam (11)], BLAST (12) and COG (13) data. Cross-referenced links to external database are also provided, including Entrez Gene (14), TIGR's Comprehensive Microbial Resource (CMR) (15), MIPS (16), EcoCyc (17) and PubMed. A keyword-based text search is available for users to search by criteria such as transporter type, transporter family, transporter protein name or substrate. The results are grouped by transporter family and organism. Each result contains links to individual family and protein pages. The protein and DNA sequences in TransportDB are readily available for BLAST search. Users can submit a single peptide or nucleotide sequence in the 'Blast' section. The output of the BLAST search includes transporter family information (TC number, family name) in addition to the standard features.

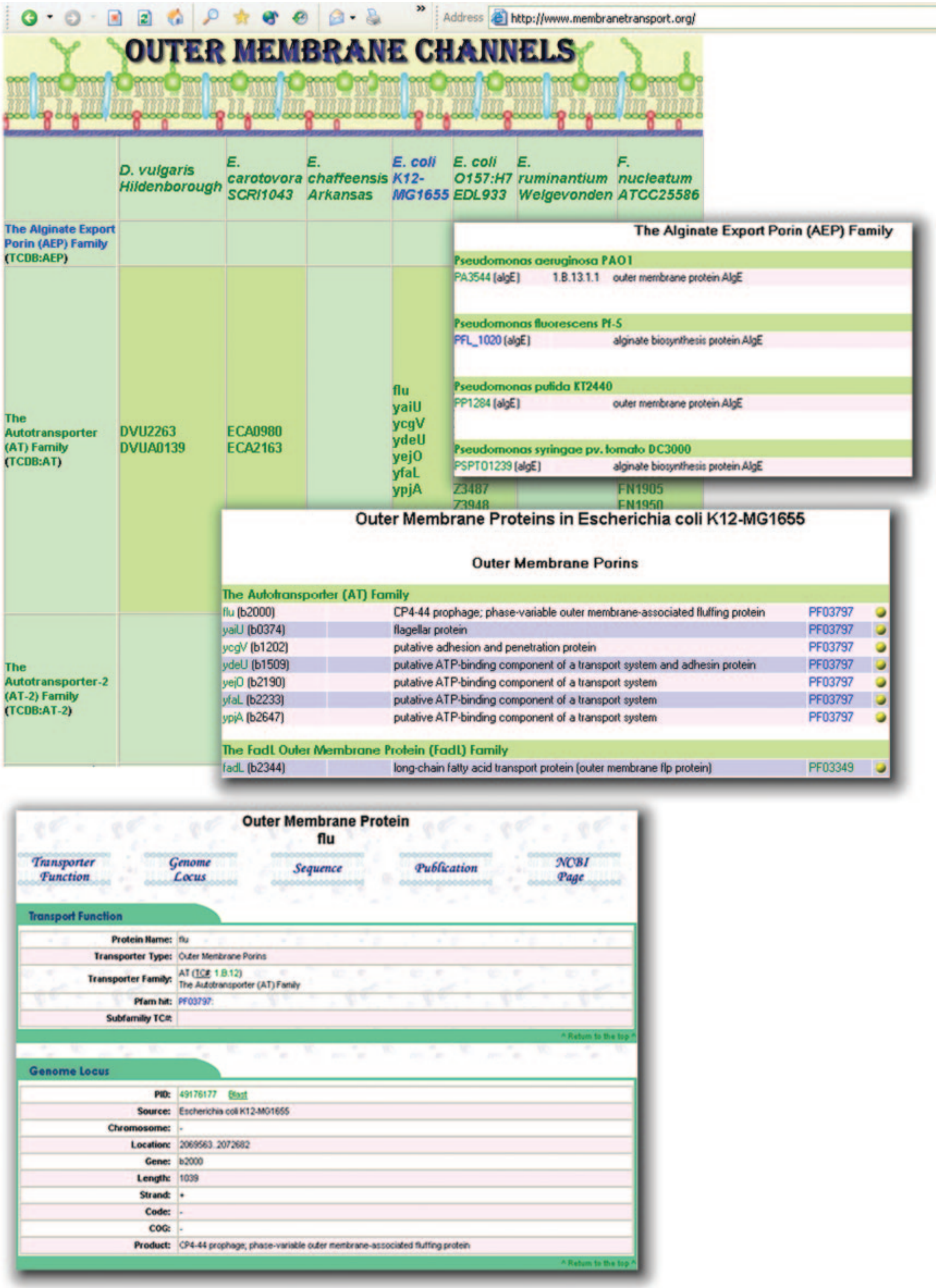


3-D Structures of Membrane Transporters

MFS: The Major Facilitator Superfamily (MFS)

	PDB	Description	Method	Resolution	Links
	1PW4	Crystal Structure Of The Glycerol-3-Phosphate Transporter From E.Coli GlpT [*Escherichia coli K12-MG1655*]	X-Ray Diffraction	3.3	PDB MMDB Entrez PubMed TransportDB
	1PV6	Crystal Structure Of Lactose LacY [*Escherichia coli K12*			
	1PV7	Crystal Structure Of Lactose LacY [*Escherichia coli K12*			

Figure 2. Graphic illustrations of the TransportDB web interface describing 3D structures of membrane transporters. These structures are listed in a tabular format and arranged by transporter families. Information on structure description, method and resolution are included. Cross-referenced links to PDB, PDB_TM, MMDB, Entrez Gene and PubMed are also provided.



	D. vulgaris Hildenborough	E. carotovora SCRI1043	E. chaffeensis Arkansas	E. coli K12-MG1655	E. coli O157:H7 EDL933	E. ruminantium Welgevonden	F. nucleatum ATCC25586
The Alginate Export Porin (AEP) Family (TCDB:AEP)							
The Autotransporter (AT) Family (TCDB:AT)	DVU2263 DVUA0139	ECA0980 ECA2163		flu yaiU ycgV ydeU yejO yfaL ypjA	Z3487 Z3948		FN1905 FN1950
The Autotransporter-2 (AT-2) Family (TCDB:AT-2)							

The Alginate Export Porin (AEP) Family

Pseudomonas aeruginosa PAO1		
PA3544 (algE)	1.B.13.1.1	outer membrane protein AlgE
Pseudomonas fluorescens Pf-5		
PFL_1020 (algE)		alginate biosynthesis protein AlgE
Pseudomonas putida KT2440		
PP1284 (algE)		outer membrane protein AlgE
Pseudomonas syringae pv. tomato DC3000		
PSPTO1239 (algE)		alginate biosynthesis protein AlgE

Outer Membrane Proteins in Escherichia coli K12-MG1655

Outer Membrane Porins

The Autotransporter (AT) Family			
flu (b2000)	CP4-44 prophage; phase-variable outer membrane-associated fluffing protein	PF03797	●
yaiU (b0374)	flagellar protein	PF03797	●
ycgV (b1202)	putative adhesion and penetration protein	PF03797	●
ydeU (b1509)	putative ATP-binding component of a transport system and adhesin protein	PF03797	●
yejO (b2190)	putative ATP-binding component of a transport system	PF03797	●
yfaL (b2233)	putative ATP-binding component of a transport system	PF03797	●
ypjA (b2647)	putative ATP-binding component of a transport system	PF03797	●
The FadL Outer Membrane Protein (FadL) Family			
fadL (b2344)	long-chain fatty acid transport protein (outer membrane flp protein)	PF03349	●

Outer Membrane Protein
flu

Transport Function

Protein Name:	flu
Transporter Type:	Outer Membrane Porins
Transporter Family:	AT (TC# 1.B.12) The Autotransporter (AT) Family
Pfam hit:	PF03797;
Subfamily TC#:	

Genome Locus

PID:	49176177 Blast
Source:	Escherichia coli K12-MG1655
Chromosome:	-
Location:	2069563..2072682
Gene:	b2000
Length:	1039
Strand:	+
Code:	-
COG:	-
Product:	CP4-44 prophage; phase-variable outer membrane-associated fluffing protein

Figure 3. Graphic illustrations of the TransportDB web interface describing outer membrane channels. Proteins are presented in a tabular format. Each outer membrane channel has individual pages showing supporting bioinformatic evidence, protein/DNA sequence, publications and cross-referenced external links, etc. Users can also pull out a list of outer membrane channels from a specific organism, or a list of proteins from a specific family in all organisms.

The relational database format allows easy manipulation of the data stored in TransportDB. An overview page is accessible for each organism, summarizing its complete transporter content, including transporter types and individual transporter families, and their statistics. Users can choose any two or more organisms from the 'Compare Organisms' section to compare their transport gene complement. All these results are generated on the fly to reflect the most recent updates.

In the ‘Phylogenetics’ section, users can view the pre-computed neighbor-joining trees for each of the transporter families through an ATV java applet (18). This enables users to access the up-to-date phylogenetic trees of every transporter family, and to manipulate the trees to display subtrees, zoom in and out, or collapse subtrees to single nodes, etc. Transport proteins in each family are also available for download in FASTA or multiple sequence alignment formats.

RECENT FEATURE ENHANCEMENT

In addition to bioinformatic predictions, we have recently begun to comprehensively track experimental evidence for transporter gene function based on the primary literature. We retrieved from Entrez (19) all related publications on transport proteins in TransportDB by e-utilities (20), which submitted queries containing gene name and organism to NCBI server and returned the related literature. The publications on genomic sequencing and massive gene expression studies were manually excluded. A total of 13 936 PubMed entries were retrieved which covers 3862 transporters and outer membrane channels. The abstracts of all these literature as well as links to PubMed are accessible at the individual transporter protein pages.

A new section has been added to TransportDB describing experimentally determined 3D structures of membrane transporters from crystallization or NMR-based studies. The data to populate this new section were derived by searching our entire collection of transport proteins against the protein data bank (PDB) (21). A total of 273 structures were retrieved, representing 98 transporters (multiple structures are available for some transporters). On the TransportDB website, these structures are listed in a tabular format and arranged by transporter families (Figure 2). Information on structure description, method and resolution are included. Cross-referenced links to PDB, PDB_TM (22), MMDB (23), Entrez Gene and PubMed are also provided. Membrane transporters represent 3–12% of total proteins of various organisms (2–4). Currently there are more than 39 000 structures deposited in the PDB. Membrane transporters are highly underrepresented and consist of <1% of all structures. This lack of representation reflects the difficulties in the purification and crystallization of transporter proteins due to their hydrophobic nature and solubility only in the presence of detergents.

Another recently added section to TransportDB describes outer membrane channels. Gram-negative bacteria and certain eukaryotic organelles, such as mitochondria and peroxisomes, are characteristically surrounded by an outer membrane that shows little permeability for hydrophilic solutes. Outer membrane proteins form nonspecific diffusion channels across the outer membrane to allow the influx of nutrients as well as the extrusion of wastes (24). A total of 4664 outer membrane channels from 142 organisms are currently annotated in TransportDB and classified into 38 families according to the TC classification. These proteins are listed in a tabular format in the ‘Outer Membrane Channels’ section (Figure 3). Each outer membrane channel has an individual page showing supporting bioinformatic evidence, protein/DNA sequences, publications and cross-referenced external links. Users can also pull out a list of outer membrane channels from a specific organism, or a list of proteins from a specific family in all the organisms.

FUTURE PERSPECTIVES

In summary, TransportDB was developed as a relational database for the comprehensive representation of cytoplasmic membrane transport systems. This is the only active database in the field dedicated to the comprehensive and comparative study of membrane transporters and outer membrane channels in different organisms with fully sequenced genomes. We are continuing to expand the TransportDB database to incorporate data from newly published genomes. TransportDB will be routinely updated with new annotation information and with data from newly sequenced organisms.

New enhancements that we are focusing on for the short- to medium-term future include the following: (i) Make our transporter annotation pipeline available to the public through our web portal so that they may customize it for their genome annotation efforts. Over the past 4 years, we have undertaken transporter annotation of over 60 unpublished genomes from custom requests from a broad range of different research groups (these data are not released to public until the publication of the relevant genome paper or until the genome data has been deposited into the public databases). We believe that providing access to our annotation pipeline through the web will serve to fulfill the increasing demands of such efforts. (ii) Add additional bioinformatic analyses to the transporter annotation pipeline, such as examining the genomic context of candidate genes, and increased use of phylogenetic approaches. (iii) Move to using a controlled vocabulary, a carefully selected list of words and phrases, for membrane transporter substrate prediction, so that they may be retrieved by searches more efficiently. The controlled vocabulary for substrate prediction can also facilitate substrate specificity comparisons and aid the automatic derivation of transport reaction equations for metabolic modeling and flux balance analysis for different sequenced genomes. As a starting point, we plan to use the hierarchical compound lists that are defined in the MetaCyc database (25).

ACKNOWLEDGEMENTS

We would like to thank all the past contributors to TransportDB. We would also like to thank Dr Martin Wu and Dr Dongying Wu for their assistance on the construction of the phylogenetic trees and for their helpful feedback on other aspects of TransportDB website. Funding to pay the Open Access publication charges for this article was provided by TIGR.

Conflict of interest statement. None declared.

REFERENCES

1. Saier,M.H.,Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
2. Paulsen,I.T., Sliwinski,M.K. and Saier,M.H.,Jr (1998) Microbial genome analyses: global comparisons of transport capabilities based on

phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.*, **277**, 573–592.

3. Paulsen,I.T., Nguyen,L., Sliwinski,M.K., Rabus,R. and Saier,M.H.,Jr (2000) Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.*, **301**, 75–100.
4. Ren,Q. and Paulsen,I.T. (2005) Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. *PLoS Comput. Biol.*, **1**, 190–201.
5. Ren,Q., Kang,K.H. and Paulsen,I.T. (2004) TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res.*, **32**, D284–D288.
6. Saier,M.H.,Jr, Tran,C.V. and Barabote,R.D. (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D181–D186.
7. Tomii,K. and Kanehisa,M. (1998) A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.*, **8**, 1048–1059.
8. Pao,S.S., Paulsen,I.T. and Saier,M.H.,Jr (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.*, **62**, 1–34.
9. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
10. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
11. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
12. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
13. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
14. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
15. Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.
16. Mewes,H.W., Frishman,D., Mayer,K.F., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stumpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
17. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
18. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
19. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
20. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
21. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
22. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.
23. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
24. Nikaido,H. (2003) Molecular basis of bacterial outer membrane permeability revisited. *Microbiol. Mol. Biol. Rev.*, **67**, 593–656.
25. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J., Rhee,S.Y. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.

Comparative Analyses of Fundamental Differences in Membrane Transport Capabilities in Prokaryotes and Eukaryotes

Qinghu Ren, Ian T. Paulsen[*]

The Institute for Genomic Research, Rockville, Maryland, United States of America

Whole-genome transporter analyses have been conducted on 141 organisms whose complete genome sequences are available. For each organism, the complete set of membrane transport systems was identified with predicted functions, and classified into protein families based on the transporter classification system. Organisms with larger genome sizes generally possessed a relatively greater number of transport systems. In prokaryotes and unicellular eukaryotes, the significant factor in the increase in transporter content with genome size was a greater diversity of transporter types. In contrast, in multicellular eukaryotes, greater number of paralogs in specific transporter families was the more important factor in the increase in transporter content with genome size. Both eukaryotic and prokaryotic intracellular pathogens and endosymbionts exhibited markedly limited transport capabilities. Hierarchical clustering of phylogenetic profiles of transporter families, derived from the presence or absence of a certain transporter family, showed that clustering patterns of organisms were correlated to both their evolutionary history and their overall physiology and lifestyles.

Citation: Ren Q, Paulsen IT (2005) Comparative analyses of fundamental differences in membrane transport capabilities in prokaryotes and eukaryotes. PLoS Comp Biol 1(3): e27.

Introduction

Membrane transport systems play essential roles in cellular metabolism and activities. Transporters function in the acquisition of organic nutrients, maintenance of ion homeostasis, extrusion of toxic and waste compounds, environmental sensing and cell communication, and other important cellular functions [1]. Various transport systems differ in their putative membrane topology, energy coupling mechanisms, and substrate specificities [2]. Among the prevailing energy sources are adenosine triphosphate (ATP), phosphoenolpyruvate, and chemiosmotic energy in the form of sodium ion or proton electrochemical gradients.

The transporter classification system (http://www.tcdb.org/) represents a systematic approach to classify transport systems according to their mode of transport, energy coupling mechanism, molecular phylogeny, and substrate specificity [2–5]. Transport mode and energy coupling mechanism serve as the primary basis for classification because of their relatively stable characteristics. There are four major classes of solute transporters in the transporter classification system: channels, primary (active) transporters, secondary transporters, and group translocators. Transporters of unknown mechanism or function are included as a distinct class. Channels are energy-independent transporters that transport water, specific types of ions, or hydrophilic small molecules down a concentration or electrical gradient; they have higher rates of transport and lower stereospecificity than the other transporter classes (e.g., *Escherichia coli* GlpF glycerol channel [6]). Primary active transporters (e.g., *Lactococcus lactis* LmrP multidrug efflux pump [7]) couple the transport process to a primary source of energy (ATP hydrolysis). Secondary transporters utilize an ion or solute electrochemical gradient, e.g., proton/sodium motive force, to drive the transport process. *E. coli* LacY lactose permease [8,9] is probably one of the best characterized secondary transporters [10]. Group translocators modify their substrates during the transport process. For example, *E. coli* MtlA mannitol PTS transporter phosphorylates exogenous mannitol using phosphoenolpyruvate as the phosphoryl donor and energy source and releases the phosphate ester, mannitol-1-P, into the cell cytoplasm [11,12]. Each transporter class is further classified into individual families and subfamilies according to their function, phylogeny, and/or substrate specificity [3].

Since the advent of genomic sequencing technologies, the complete sequences of over 200 prokaryotic and eukaryotic genomes have been published to date, representing a wide range of species from archaea to human. There are also more than 1,100 additional genome sequencing projects currently underway around the world (Gold Genomes Online Database, http://www.genomesonline.org/) [13,14]. Convenient and effective computational methods are required to handle and analyze the immense amount of data generated by the whole-genome sequencing projects. An in-depth look at transport proteins is vital to the understanding of the metabolic capability of sequenced organisms. However, it is often problematic to annotate these transport proteins by current primary annotation methods because of the occurrence of large and complex transporter gene families, such as the ATP-binding cassette (ABC) superfamily [15,16] and the major facilitator superfamily (MFS) [17,18], and the presence of multiple transporter gene paralogs in many organisms. We have been working on a systematic genome-wide analysis of cellular membrane transport systems. Previously, we reported

Received March 24, 2005; Accepted July 8, 2005; Published August 19, 2005
DOI: 10.1371/journal.pcbi.0010027

Abbreviations: ABC, adenosine triphosphate–binding cassette; ATP, adenosine triphosphate; GIC, glutamate-gated ion channel; MFS, major facilitator superfamily; ORF, open reading frame; PTS, phosphotransferase system

Editor: Peer Bork, EMBL Heidelberg, Germany

[*] To whom correspondence should be addressed. E-mail: ipaulsen@tigr.org

Synopsis

Membrane transporters are the cell's equivalent of delivery vehicles, garbage disposals, and communication systems—proteins that negotiate through cell membranes to deliver essential nutrients, eject waste products, and help the cell sense environmental conditions around it. Membrane transport systems play crucial roles in fundamental cellular processes of all organisms. The suite of transporters in any one organism also sheds light on its lifestyle and physiology. Up to now, analysis of membrane transporters has been limited mainly to the examination of transporter genes of individual organisms. But advances in genome sequencing have now made it possible for scientists to compare transport and other essential cellular processes across a range of organisms in all three domains of life.

Ren and Paulsen present the first comprehensive bioinformatic analysis of the predicted membrane transporter content of 141 different prokaryotic and eukaryotic organisms. The scientists developed a new computational application of the phylogenetic profiling approach to cluster together organisms that appear to have similar suites of transporters. For example, a group of obligate intracellular pathogens and endosymbionts possess only limited transporter systems in spite of the massive metabolite fluxes one would expect between the symbionts and their host. This is likely due to the relatively static nature of their intracellular environment. In contrast, a cluster of plant/soil-associated microbes encode a robust array of transporters, reflecting the organisms' versatility as well as their exposure to a wide range of different substrates in their natural environment.

a comprehensive analysis of the transport systems in 18 prokaryotic organisms [19,20] and in yeast [21]. Here we expand our analyses to 141 species and compare the fundamental differences in membrane transport systems in prokaryotes and eukaryotes. Phylogenetic profiling of transporter families and predicted substrates was utilized to investigate the relevance of transport capabilities to the overall physiology of prokaryotes and eukaryotes.

Results/Discussion

Numbers of Recognized Transporter Families and Proteins

A total of 40,678 transport proteins from 141 species (Table S1), including 115 Eubacteria, 17 Archaea, and 9 Eukaryota, were predicted by our analysis pipeline. They were classified into 134 families, including 7 families of primary transporters, 80 families of secondary transporters, 32 channel protein families, 2 phosphotransferase systems (PTSs), and 13 unclassified families. Some of these families are very large superfamilies with numerous members, such as the ABC superfamily and MFS, both of which are widely distributed in Eubacteria, Archaea, and Eukaryota. Some are small families with only a single or a few members. The distribution of transporter families varies significantly across the three domains of life (Figure 1). There are 42 eukaryotic-specific families, mostly ion channel families that exist exclusively in multicellular eukaryotic organisms like *Drosophila melanogaster, Arabidopsis thaliana,* and humans. These channels are involved in processes like cell communication, signal transduction, and maintenance of internal homeostasis in a multicellular environment. Most of these families are restricted to a single organismal type. Many of them may have arisen later during evolution, after the separation of the three domains. Alternatively, some families may have diverged too extensively from their prokaryotic counterparts to be recognized as homologs. Interestingly, a bacterial homolog to the previously described "eukaryotic-specific" glutamate-gated ion channel (GIC) family of neurotransmitter receptors has now been characterized in *Synechocystis* [22,23], and its orthologs have been identified in other sequenced Cyanobacteria. The *Synechocystis* transporter binds glutamate and forms a K^+-selective ion channel. These observations suggest that eukaryotic GIC family transporters arose from a primordial prokaryotic counterpart.

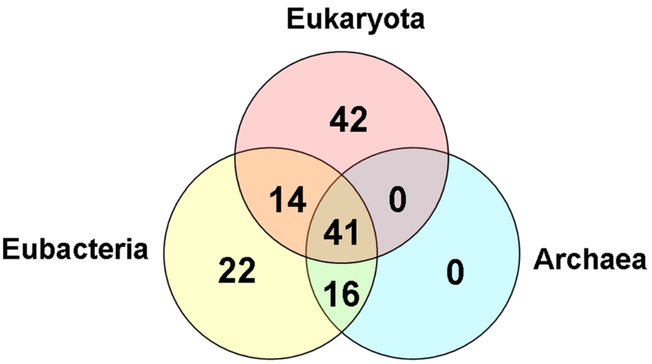


Figure 1. Venn Diagram Showing the Distribution of Transporter Families across the Three Domains of Life

DOI: 10.1371/journal.pcbi.0010027.g001

There are 38 prokaryotic-specific transporter families, of which 22 families exist exclusively in Eubacteria, such as the bacterial sugar PTS systems (see below), and 16 are shared by Eubacteria and Archaea. In contrast to eukaryotic-specific families, which are usually limited to single species, the majority of prokaryotic-specific ones are broadly distributed among prokaryotes. There are no Archaea-specific transporter families currently known. Due to the very limited experimental characterization of Archaea species relative to Eubacteria and Eukaryota, many aspects of the physiology and biochemistry of Archaea are poorly understood [24]. We compared the annotation of membrane proteins in selected species of Archaea and Eubacteria in The Institute for Genomic Research's Comprehensive Microbial Resource database [25]. The percentage of the membrane proteins assigned to the role category of "hypothetical proteins" is significantly greater in Archaea than in Eubacteria (Figure S1). These observations suggest that the sparse functional characterization could be the primary reason for the lack of any known Archaea-specific transporter families.

There are 41 transporter families represented in all three domains of life, highlighting the fundamental importance of these families. These are presumably very ancient families shared by the last common ancestor of Archaea, Eukaryota, and Eubacteria. Most of them were found within the secondary transporter class. These ubiquitous transporter families function in the transport of a diverse spectrum of substrates, including sugars, amino acids, carboxylates, nucleosides, and various cations and anions. There are 14 families shared by Eubacteria and Eukaryota and 16 shared by Eubacteria and Archaea. Some of these families shared only in two domains may ultimately be discovered in all three domains once a greater diversity of organisms is sequenced.

The overall quantity of recognized transport proteins (Figure 2A) and the percentage relative to the total number of open reading frames (ORFs) (Figure 2B) were compared for the organisms analyzed. Between 2% and 16% of ORFs in prokaryotic and eukaryotic genomes were predicted to encode membrane transport proteins, emphasizing the importance of transporters in the lifestyles of all species. In general, eukaryotic species, especially multicellular eukaryotic organisms, exhibit the largest total number of transport proteins, e.g., *Drosophila* (682 transport proteins, 3.7% of ORFs), *Arabidopsis* (882, 3.5%), *Caenorhabditis elegans* (669, 4.1%), and humans (841, 3.0%). However, the transport proteins of eukaryotic species account for a relatively smaller percentage of total ORFs than in Eubacteria (average 9.3% ± 2.9%) and Archaea (average 6.7% ± 2.3%) species. Considerable variations in the quantity of transport proteins have been observed among species belonging to the same phylogenetic group. For example, α-Proteobacteria species exhibit a wide variety of lifestyles and corresponding differences in transporter content; they range from rhizosphere-dwelling organisms such as *Mesorhizobium loti* and *Sinorhizobium meliloti* [26] with 883 (12.1%) and 826 (13.3%) transport proteins each, to obligate intracellular pathogens or symbionts such as *Rickettsia prowazekii* and *Wolbachia* sp. with 57 (6.8%) and 65 (5.4%) transport proteins, respectively. Overall, prokaryotic obligate endosymbionts and intracellular pathogens, as well as the eukaryotic intracellular parasites (*Plasmodium falciparum* [27] and *Encephalitozoon cuniculi* [28]), possess the most limited repertoire of membrane transporters.

Genome Size versus Diversity of Transporter Families and Numbers of Paralogs

Organisms with a larger genome size and therefore more ORFs generally encode a greater number of transporters [19,29]. In addition to transporters, regulatory genes, secondary metabolism genes, and transcription factors, also appear to increase with genome size [29–31]. Two major factors could

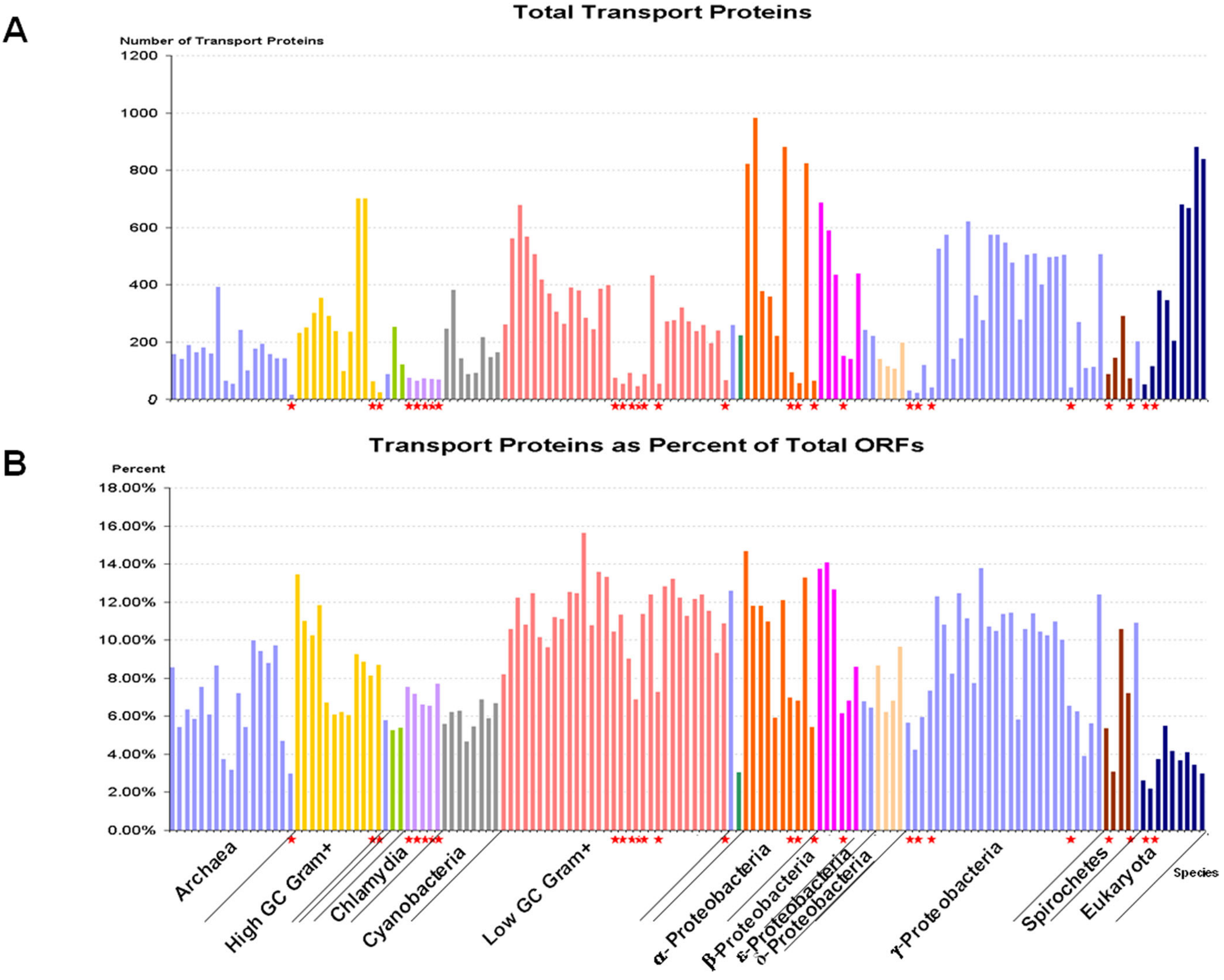


Figure 2. Numbers of Recognized Transport Proteins and Percentage of Total ORFs

The overall numbers of recognized transport proteins (A) and percentage of total ORFs encoding transport proteins (B) were compared for the 141 organisms analyzed. Species from distinct phylogenetic groups are labeled with different colors. The prokaryotic and eukaryotic obligate intracellular parasites/pathogens are marked with red stars.

DOI: 10.1371/journal.pcbi.0010027.g002

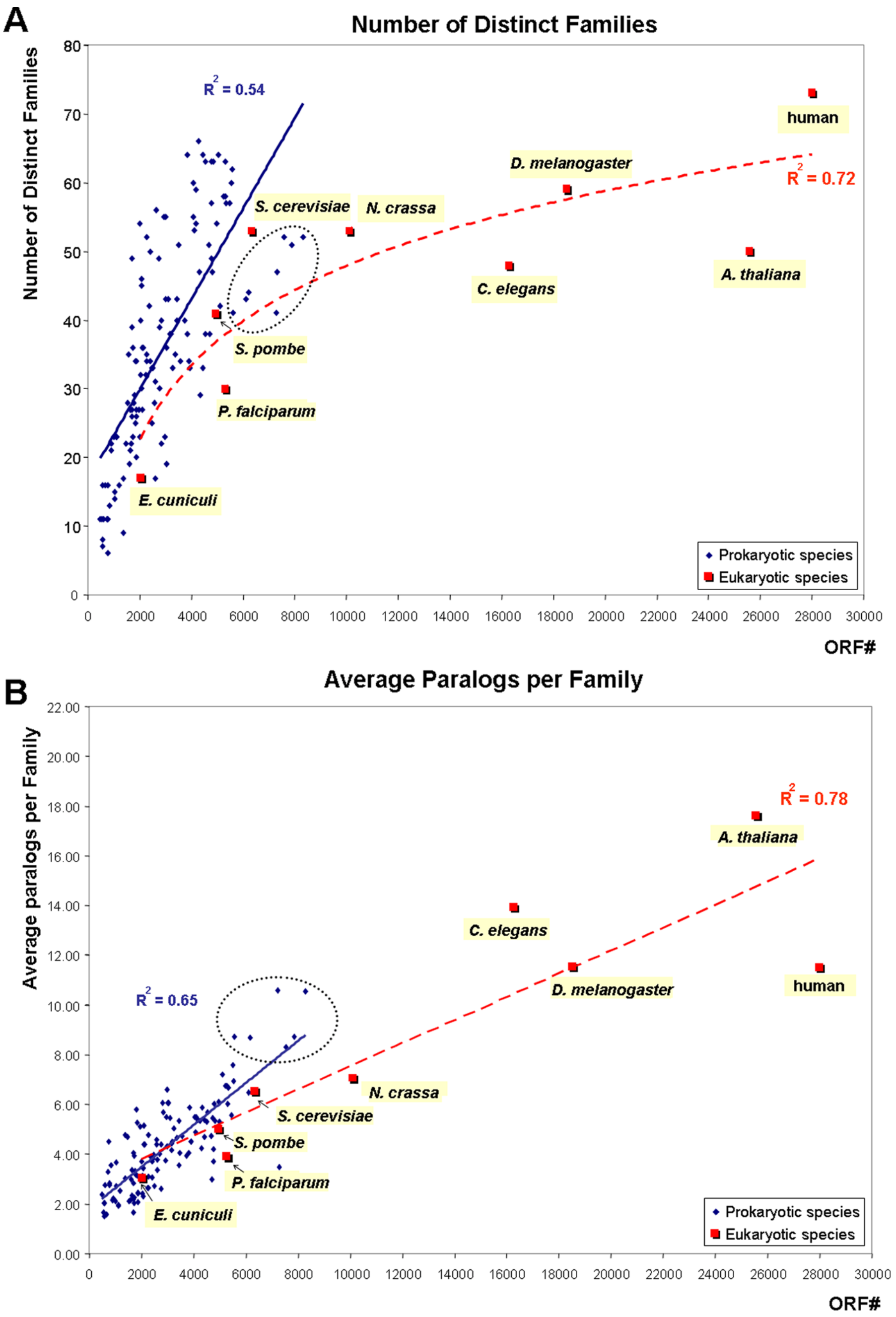


Figure 3. Number of Total ORFs versus Number of Distinct Transporter Families or Average Number of Paralogs per Family

The number of total ORFs in the genome for each of the 141 sequenced prokaryotic and eukaryotic organisms (*x*-axis) was plotted as a function of either the number of distinct transporter families (A) or the average number of paralogs per family (B) (*y*-axis). Blue diamonds represent prokaryotic organisms and red squares represent eukaryotic organisms. Trend line and power correlation R^2 value are shown for prokaryotes and eukaryotes, respectively. A group of α-Proteobacteria are enclosed by a circle (see text for discussion).

DOI: 10.1371/journal.pcbi.0010027.g003

contribute to the expansion of transporters in organisms with large genome sizes: (1) an increased number of distinct transporter families, and (2) a higher degree of gene duplication or expansion, leading to a greater number of paralogs in certain transporter families. To investigate the relationship between genome size and these two factors, we plotted the total number of ORFs from 141 organisms as a function of either the number of distinct transporter families (Figure 3A), or the average number of paralogs per family (Figure 3B). Prokaryotes and eukaryotes exhibit distinct differences. For prokaryotic species, there is a relatively linear relationship between the genome size and the number of transporter families ($R^2 = 0.54$) or average number of paralogs ($R^2 = 0.65$). As genome size increases, the rate of increase in the number of families per organism is approximately eight times greater than that of the average number of paralogs per family. The increase in genome size can only partially explain the expansion of transporter families and

Table 1. The Relative Percentage of Each Transporter Type within Major Phylogenetic Groups

Phylogenetic Group[a]	Primary Transporters	Ion Channels	Secondary Transporters	PTSs	Unclassified
Archaea (17)	39.1% ± 10.4%	6.8% ± 4.2%	52.0% ± 12.1%	NF	2.3% ± 2.0%
Actinobacteria (12)	46.2% ± 10.7%	3.4% ± 1.4%	48.0% ± 9.3%	1.1% ± 1.2%	1.4% ± 1.0%
Chlamydia (5)	39.9% ± 3.8%	NF	49.6% ± 3.1%	8.3% ± 1.1%	2.2% ± 0.1%
Cyanobacteria (8)	51.7% ± 8.0%	8.4% ± 2.5%	38.3% ± 7.6%	NF	1.6% ± 0.5%
Firmicutes (30)	45.6% ± 11.9%	4.2% ± 1.6%	38.7% ± 15.2%	10.9% ± 6.1%	1.1% ± 1.0%
Proteobacteria-alpha (10)	43.9% ± 12.4%	2.9% ± 2.0%	50.0% ± 13.1%	1.5% ± 1.6%	1.6% ± 1.3%
Proteobacteria-beta (6)	41.0% ± 6.8%	3.2% ± 0.7%	51.4% ± 7.5%	2.5% ± 1.3%	2.0% ± 1.8%
Proteobacteria-gamma (27)	30.8% ± 5.6%	5.1% ± 2.0%	55.6% ± 9.3%	7.2% ± 7.2%	1.4% ± 1.0%
Proteobacteria-delta (2)	48.5% ± 8.3%	6.0% ± 0.6%	42.1% ± 5.7%	1.8% ± 2.5%	1.7% ± 0.4%
Proteobacteria-epsilon (3)	34.7% ± 3.4%	4.7% ± 2.5%	57.6% ± 2.5%	NF	3.0% ± 0.8%
Spirochetes (4)	45.3% ± 16.1%	3.1% ± 2.3%	44.2% ± 8.6%	5.9% ± 8.4%	1.6% ± 1.2%
Fungi (3)	15.3% ± 2.5%	4.2% ± 0.3%	77.9% ± 2.2%	NF	1.8% ± 0.2%
Protozoa (1)	48.4%	1.6%	50.0%	NF	0.0%
Microsporidia (1)	41.9%	11.6%	46.5%	NF	0.0%
Nematodes (1)	11.7%	31.1%	56.5%	NF	0.6%
Insects (1)	13.7%	27.9%	56.6%	NF	1.4%
Plants (1)	20.2%	11.9%	65.5%	NF	2.4%
Primates (1)	14.9%	43.3%	38.9%	NF	1.6%

[a]Number of organisms analyzed is indicated in the parenthesis.

NF, not found.

DOI: 10.1371/journal.pcbi.0010027.t001

paralogs (as indicated by the correlation R^2 value). The strain-specific properties and lifestyles could also have an impact. For example, a group of α-Proteobacteria exhibit the most paralogs per family but have relatively lower diversity of transporter families. These organisms include rhizobial microsymbionts *M. loti, S. meliloti,* and *Bradyrhizobium japonicum* [26], and a closely related plant pathogen, *Agrobacterium tumefaciens* (enclosed by a circle on Figure 3). All of these organisms have more ABC transporters than any other sequenced organisms [29]. ABC family transporters mediate the uptake of a variety of nutrients and the extrusion of drugs and metabolite wastes. Having a large complement of high-affinity ABC uptake systems may be an advantage for organisms in the competition among microbes for nutrients. Two *Streptomyces* species, *St. avermitilis* and *St. coelicolor,* also exhibit a similar trend, with a significant expansion of the ABC and MFS family transporters.

The number of eukaryotic species analyzed is smaller, so it is more difficult to draw robust conclusions. The single-celled eukaryotes such as the yeasts appear to display characteristics similar to those of the prokaryotes, showing expansions in both transporter families and paralogs as genome size increases, with the former being a more important factor. However, in multicellular eukaryotic organisms such as animals and plants, the tremendous number of paralogs in certain transporter families accounts for a significant portion of the increase of transporters. Although multicellular eukaryotes exhibit fewer transporter families than some of the prokaryotic species, they have generated an extraordinary number of paralogs by gene duplication or expansion within certain families, like the ABC superfamily, MFS, and the voltage-gated ion channel superfamily. For example, the *Arabidopsis* genome encodes 110 paralogs of the ABC superfamily [32,33] and 92 paralogs of the MFS.

These differences in the relative abundances of transporter paralogs and distinct transporter families probably represent fundamental differences in transporter needs or priorities of these organisms. Multicellular organisms with many apparently redundant transporter paralogs appear to be utilizing a strategy of specialization. Many of their closely related paralogous transporters are presumably expressed only in specific tissues or subcellular localizations, or at specific developmental time points. Many appear to be involved in cell–cell communication and signal transduction processes, emphasizing the importance of intercellular communication in complex multicellular organisms. In contrast, the single-celled prokaryotes and eukaryotes, with relatively fewer paralogs but a greater emphasis on numbers of different families of transporters, appear to be utilizing a strategy of diversification. This probably reflects that one of the primary roles of membrane transport systems in these organisms is nutrient acquisition. A greater diversity of transporter types presumably allows for a broader range of substrate utilization.

Distribution of Transporter Types According to Energy Coupling Mechanism

A wide range of variations were observed in the relative usage of energy coupling mechanisms to drive transport processes among the prokaryotes and eukaryotes analyzed. Table 1 shows the relative percentage of each transporter type in organisms from major phylogenetic groups. Transporters were categorized into five major types according to transport mode and energy coupling mechanism: primary transporters, secondary transporters, ion channels, group translocators, and unclassified. Primary and secondary carriers are ubiquitous, being present in all organisms analyzed. However, their percentage among the total transporters varies greatly (12%–78% for primary carriers and 17%–80% for secondary carriers). In prokaryotic and unicellular eukaryotic systems, primary and secondary carriers are the predominant types of transporters, together contributing more than 90% of the total transporters.

Channel proteins make up a greater percentage (12%–43%) in higher eukaryotic organisms.

Compared to eukaryotes, prokaryotic organisms rely heavily on primary active transporters, largely because of the usage of ABC uptake systems that are absent in eukaryotes [34]. Organisms with the highest percentage of primary transporters generally belong to one of the three groups. (1) The first group includes organisms that lack a citrate cycle and an electron transfer chain, and therefore can only generate a proton motive force by indirect methods such as substrate-level phosphorylation followed by ATP hydrolysis. These organisms include *Mycoplasma* spp., spirochetes, *Streptococcus* spp., *Tropheryma whipplei, Mycobacterium leprae, Thermoanaerobacter tengcongensis,* and *Thermotoga maritime*. ATP is their primary source of energy, and therefore is most frequently used to drive nutrient uptake and maintain ion homeostasis. (2) The second group includes photosynthetic organisms with the ability to synthesize an ATP pool via photosynthesis, including *Synechocystis* sp., *Nostoc* sp., and *Thermosynechococcus elongates*. (3) The third group is a group of α-Proteobacteria that possess a significant expansion of the ABC superfamily [29], including soil/plant-associated bacteria, such as *M. loti* [26], *S. meliloti* [26], *A. tumefaciens,* and related human/animal pathogens such as *Brucella suis.* Unlike the first two groups, in which the usage of primary transporters seems to be predicated on bioenergetic constraints, the expansion of the ABC transporter family in these α-Proteobacteria does not have any obvious energetic explanations. Instead, it may reflect an organismal requirement for high-affinity transport since ABC transporters typically show higher substrate affinities than most secondary transporters.

The PTS is only present in a subset of Eubacteria, while completely lacking in Archaea and Eukaryota. Gram-negative enteric bacteria, such as *E. coli, Shigella flexneri,* and *Salmonella typhimurium*, as well as Gram-positive species associated with the human gastrointestinal tract, like *Listeria monocytogenes* and *Lactobacillus plantarum,* encode the most abundant PTS systems. Owing to the absorption capacity and efficiency of the intestine, these species have to compete with hundreds of other types of bacteria in an environment containing only small amounts of free carbohydrates or other easily absorbable forms of nutrients. The enrichment of sugar PTS systems in these species could be an advantage to thrive in their ecological niches.

Channel proteins contribute a relatively smaller percentage of transporters in the prokaryotic species we analyzed, and their functions in vivo are largely unknown. Nine organisms lack recognizable channels, including *Chlamydia* spp., *T. whipplei, Treponema pallidum, Wolbachia* sp., and *R. prowazekii,* all of which are obligate intracellular pathogens/symboints. All other prokaryotic species, including all extremophiles sequenced to date, encode channel proteins, suggesting these channels could function in responding promptly to osmotic and other environmental stresses [35]. Intracellular pathogens and endosymbionts may not need water or ion channels because of their relatively static intracellular environment and may largely depend on their host organisms for maintenance of ion homeostasis.

The percentage of channel proteins increases significantly in multicellular eukaryotes. In animals, these consist largely of ion channels with communication roles, such as in signal transduction, or roles as sensors for external stimuli. For example, members in the ligand-gated ion channel family [36] and the GIC family [37] are activated by major excitatory (glutamate) and inhibitory neurotransmitters (GABA) and participate in neuronal communication in the brain [38]. Recent studies show that some subunits of ligand-gated ion channels and GIC-type channels are expressed prominently during embryonic and postnatal brain development, while others are expressed mainly in the adult brain, suggesting that a switch in subunit composition may be required for normal brain development [38]. In plants, approximately one-third of the channel proteins are aquaporins (water channels) [39], many of which show a cell-specific expression pattern in the root, emphasizing the importance of regulating and maintaining turgor pressure through the plant [40].

Three fungal species, *Saccharomyces cerevisiae, Schizosaccharomyces pombe,* and *Neurospora crassa,* possess the largest portion of secondary transporters (76%–80%), mainly because of the prominent gene expansion of two types of functionally diverse MFS family transporters: (1) drug efflux pumps, which could play roles in the secretion of secondary metabolites, toxic compounds, and signaling molecules, and (2) sugar symporters, which could allow a broader range of sugar utilization [41,42].

Phylogenetic Profiling of Transporter Family and Substrate Shows Strong Correlations to Organisms' Overall Physiology

The phylogenetic profile of a given protein is a string that encodes the presence or absence of that protein in every fully sequenced genome. Proteins that function together in a pathway or a common structural complex are likely to evolve in a correlated fashion, and therefore tend to be either preserved or eliminated together in a new species during evolution [43,44]. Phylogenetic profiling has been an effective way to detect conserved core genes, species-specific gene families, lineage-specific gene family expansions [45], and subcellular localization of proteins [46]. It can also facilitate the prediction of physical and functional interactions and assist in the deduction of the functions of genes that have no well-characterized homologs [47,48].

We have undertaken a novel application of phylogenetic profiling to investigate the presence or absence of transporter protein families across sequenced genomes. To our knowledge this represents the first application of a phylogenetic profiling approach using protein families rather than individual proteins as the unit of comparison. With the data on membrane transport systems from 141 fully sequenced organisms, we were able to construct the phylogenetic profiles for each transporter family (Figures 4 and S2). Hierarchical clustering of phylogenetic profiles showed a strong correlation between the observed clustering pattern and phylogeny, with Eubacteria, Archaea, and Eukaryota clearly separated into different clusters. Inside the bacterial cluster, Gram-positive bacteria, Proteobacteria, *Chlamydia,* and Cyanobacteria are also clearly defined into different groups. Given that the profiling approach solely utilizes presence or absence of a transporter family and does not use sequence similarity directly, this indicates that the types of transporters utilized by organisms are related to their evolutionary history. Additionally, the clustering appears to be influenced by habitat or lifestyle of organisms. For example, the obligate intracellular pathogens/symbionts and

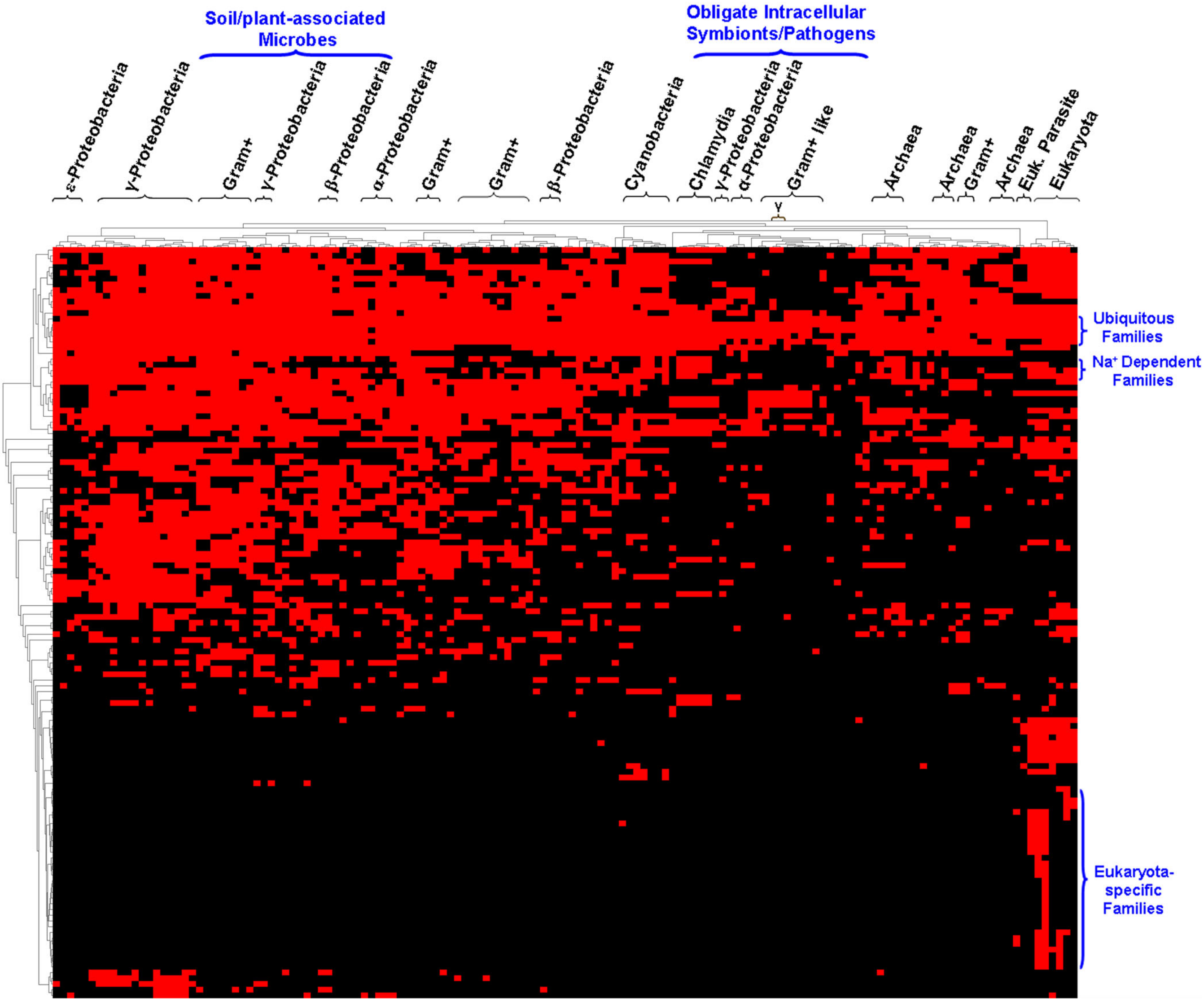


Figure 4. Phylogenetic Profiling of Transporter Families

Phylogenetic profiles were created for each transporter family. Each profile is a string with 141 entries (number of organisms analyzed). If a given family is present in an organism, the value one is assigned at this position (red). If not, zero is assigned (black). Organisms and transporter families were clustered according to the similarity of their phylogenetic profiles.

DOI: 10.1371/journal.pcbi.0010027.g004

a collection of soil/plant-associated microbes are separated into two distinct superclusters (Figure 5).

The obligate intracellular pathogens/symbionts cluster includes a group of phylogenetically diverse organisms, including *Chlamydia* spp. (pathogens); γ-Proteobacteria such as *Buchnera* spp., *Wigglesworthia glossinidia brevipalpis,* and *Candidatus Blochmannia floridanus* (endosymbionts); α-Proteobacteria such as *Wolbachia* sp. (endosymbiont) and *R. prowazekii* (pathogen); Gram-positive-like organisms *Mycoplasma* spp. and *T. whipplei* (pathogens); Spirochetes such as *Tr. pallidum* and *Borrelia burgdorferi* (pathogens); and an archaeal symbiont, *Nanoarchaeum equitans*. Organisms in this cluster share an obligate intracellular lifestyle as well as reduced genome size. The clustering does not appear to be due to genome size alone as nonobligate intracellular organisms with small genome sizes do not fall into this cluster. One possibility is that the transport needs of these obligate intracellular organisms are more specialized than those of environmental organisms because of the much more static nature of their intracellular environments. This may have allowed them to shed, for example, transporters for alternative nitrogen/carbon sources, osmoregulatory functions, and ion homeostasis. Similar to their prokaryotic counterparts, two eukaryotic intracellular parasites, *P. falciparum* and *En. cuniculi,* form a distinct cluster separate from the other eukaryotes.

The soil/plant-associated microbe cluster also contains species from various phylogenetic groups, such as Actinobacteria *(Corynebacterium* and *Streptomyces),* Firmicutes *(Bacillus* and *Oceanobacillus),* α-Proteobacteria *(Brucella, Agrobacterium, Mesorhizobium, Sinorhizobium,* and *Bradyrhizobium),* β-Proteobacteria *(Bordetella* and *Ralstonia),* γ-Proteobacteria *(Pseudomonas* and *Rhodopseudomonas),* δ-Proteobacteria *(Geobacter),* Deinococcus *(Deinococcus radiodurans),* Planctomycetes (*Pirellula* sp.),

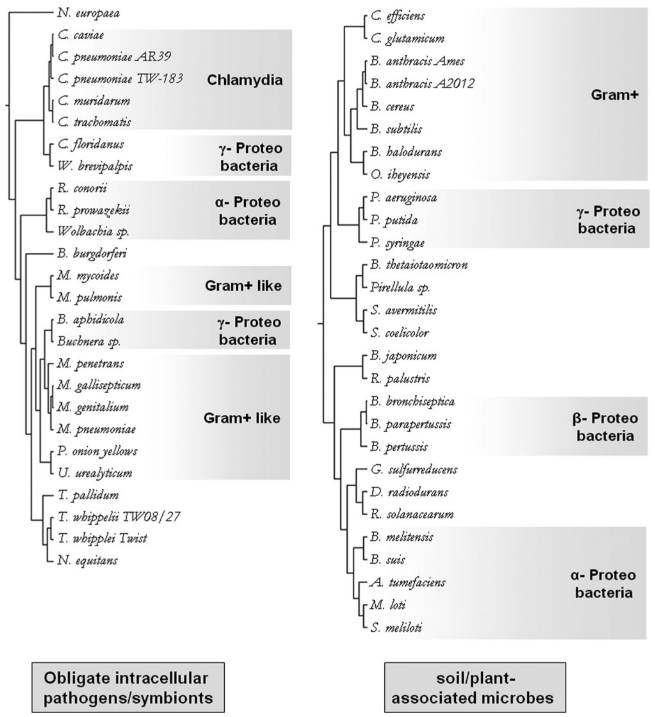


Figure 5. Hierarchial Clustering of Phylogenetic Profiles of Obligate Intracellular Pathogens/Symbionts versus Soil/Plant-Associated Microbes
Detailed view of two clusters of organisms generated by hierarchical clustering of their phylogenetic profiles of transporter families: obligate intracellular pathogens/symbionts and soil/plant-associated microbes.
DOI: 10.1371/journal.pcbi.0010027.g005

and Bacteroidetes *(Bacteroides thetaiotaomicron).* All of these organisms possess a robust collection of transporter systems. It is unlikely that these species are merely clustered by their genome sizes because some species in this cluster have relatively smaller genome sizes, like *Corynebacterium efficiens* (3.0 Mb), *D. radiodurans* (3.2 Mb), and *Brucella melitensis* (3.2 Mb). In addition, hierarchical clustering of organisms exclusively by genome size generates clusters with no apparent phylogenetic relationship (data not shown). The similarity of phylogenetic profiles of organisms in this cluster probably reflects the versatility of these organisms and their exposure to a wide range of different substrates in their natural environment. The majority of species in this cluster can be free-living in the soil, and some are capable of living in a diverse range of environments. They generally share a broad range of transport capabilities for plant-derived compounds specifically and for organic nutrients in general. Interestingly, some of the human pathogens, e.g., *Bordetella, Brucella, Bacillus anthracis* [26], and *Bacteroides thetaiotaomicron,* are also grouped in this cluster. All of these pathogens have close relatives that are soil- or plant-associated environmental organisms [49–52], so their transport capabilities probably reflect a combination of their evolutionary heritage, original environmental niche, and current transport needs.

To compare the transport capabilities of organisms in the intracellular pathogen/symbiont cluster and the soil/plant-associated microbe cluster, we carried out statistical analysis on their number of transporters, percentage of ORFs encoding transport proteins, and compositions in each transporter type (data not shown). Organisms in the soil/plant-associated microbe cluster on average have about eight times as many transporters as those in the intracellular organism cluster ($p < 0.0001$; p-value denotes the confidence level that the correlation observed is significantly different from the null hypothesis). The difference in the relative percentage of ORFs that are transporters is smaller but still significant (1.5-fold increase, $p < 0.0001$), suggesting that systematic gene loss and genome compaction is one of the important factors in reducing the number of transport proteins in intracellular organisms. The residual transport systems conserved in these obligate intracellular organisms probably belong to the core essential genes required for the acquisition of key nutrients and metabolic intermediates. For example, a glutamate transporter is encoded in two obligate endosymboints: the GltP glutamate:proton symporter (DAACS family) [53] in *Candidatus Blochmannia floridanus,* and GltJKL ABC transporter [54] in *Wigglesworthia glossinidia brevipalpis*. These organisms have a truncated citrate cycle that begins with α-ketoglutarate and ends with oxaloactetate [55]. Their citrate cycle could be closed by the transamination of the imported glutamate to aspartate, catalyzed by an aspartate aminotransferase (AspC) that uses oxaloactetate as a cosubstrate and produces α-ketoglutarate. As to the distribution of transporter types, there is no significant difference between these two clusters although intracellular organisms show a higher degree of variation in each transporter type than the plant/soil-associated microbes. These variations may reflect the unique internal environment inside the host cells. All these observations illustrate how adaptation of an organism to certain living conditions leads to changes in its transporter repertoire and at the same time determines the set of transporters that the organism cannot afford to lose.

In addition to investigating the relationship between organisms based on their transporter profiles, we also examined the clustering of transporter families. The essentially ubiquitous families, like ABC, MFS, P(F)-type ATPase, that are present in virtually every organism we analyzed, are clustered together. Eukaryotic-specific families, most of which are single-organism-specific ion channels, are grouped together. Interestingly, the sodium-ion-dependent families, like neurotransmitter:sodium symporter, alanine/glycine:cation symporter, solute:sodium symporter, and divalent anion:sodium symporter [56–58], are clustered together. Transporters in these families are all symporters that utilize the sodium ion gradient to transport amino acid, solute, and/or divalent ions into cytoplasm. This clustering may suggest that these families co-occur in a specific set of organisms, presumably those reliant on sodium-ion-driven transport.

Previous studies have shown that transporters with similar functions characteristically cluster together in phylogenetic analyses; hence, substrate specificity appears to be a conserved evolutionary trait in transporters [19,20,59,60]. The phylogenetic profiles of predicted substrates for all 141 organisms were generated and clustered by MeV (see Figure S3). Overall, similar patterns were observed as with the clustering by families. Organisms were grouped together either by their phylogenetic history or by their physiology or living habits. Ubiquitous substrates (e.g., cation, amino acid, sugar, and phosphate) and eukaryotic-specific substrates (e.g., cholesterol, UDP-sugars, and phosphoenolpyruvate) each form distinct clusters.

Distribution of Transporter Families among Species in the Same Genus

With the transporter data from a great diversity of sequenced organisms, we were able to compare the distribution of transporter families in closely related species (i.e., from the same genus) (Figures 6 and S4). In most of the cases we studied, species from the same genus share highly parallel distributions of transporter families. For example, three *Pseudomonas* species, *Ps. aeruginosa* [61], *Ps. putida* [62] and *Ps. syringae* [63], all of which are metabolically versatile soil/plant-associated bacteria, show highly similar patterns of transporter family distribution. Among the 66 transporter families present in this genus, 47 are shared by all three species and 14 are shared by two species (Figure 6A). All three species encode transporters for a diverse spectrum of substrates, including sugars, amino acids, peptides, carboxylates, and various cations and anions.

The distribution of transporter families in three *Corynebacterium* species represents an exception. *Co. glutamicum* [64] and *Co. efficiens* [65] are widely used in the industrial production of amino acids like glutamic acid and lysine by fermentation. The closely related *Co. diphtheriae* [66], however, is a human pathogen causing the respiratory illness diphtheria and lacks amino acid productivity. Compared to the other two species, *Co. diphtheriae* shows a dramatically different transporter family profile (Figure 6B). There are eight families specific to *Co. diphtheriae,* while only one for *Co. glutamicum* and three for *Co. efficiens*. More importantly, *Co. diphtheriae* uses totally different mechanisms to transport potassium ion and C4-dicarboxylates than the other two species. In *Co. diphtheriae,* potassium ions are transported into cytoplasm via a Trk family K^+:H^+ symporter [67], while both *Co. glutamicum* and *Co. efficiens* encode a KUP family potassium ion uptake permease [68]. *Co. diphtheriae* utilizes the DcuABC antiporter system [69] for the uptake of C4-dicarboxylate, while the other species use the ATP-independent tripartite periplasmic symporter systems (TRAP-T family) [70]. The common orthologs of transporters in families specific to one or two *Corynebacterium* species were identified in sequenced high-GC Gram-positive bacteria, and the phylogenetic trees were constructed by the neighbor-joining method (data not shown). For those families with orthologs in *Co. glutamicum* and *Co. efficiens* but not in *Co. diphtheriae,* orthologs were also identified in the majority of high-GC Gram-positive species. The trees of transport protein are similar to the 16S rRNA tree, suggesting certain transporter families in *Co. efficiens* are missing because of specific gene losses. By contrast, *Co. diphtheriae*–specific transporter families, like Dcu, DcuC, and Trk families, tend to have either no apparent orthologs or only distantly related homologs in other sequenced high-GC Gram-positive species, suggesting possible evolutionary gene acquisition events in *Co. diphtheriae*. The recent finding that both gene loss and horizontal gene transfer are responsible for the functional differentiation in amino acid biosynthesis of the three *Corynebacterium* species [71] further supports this conclusion.

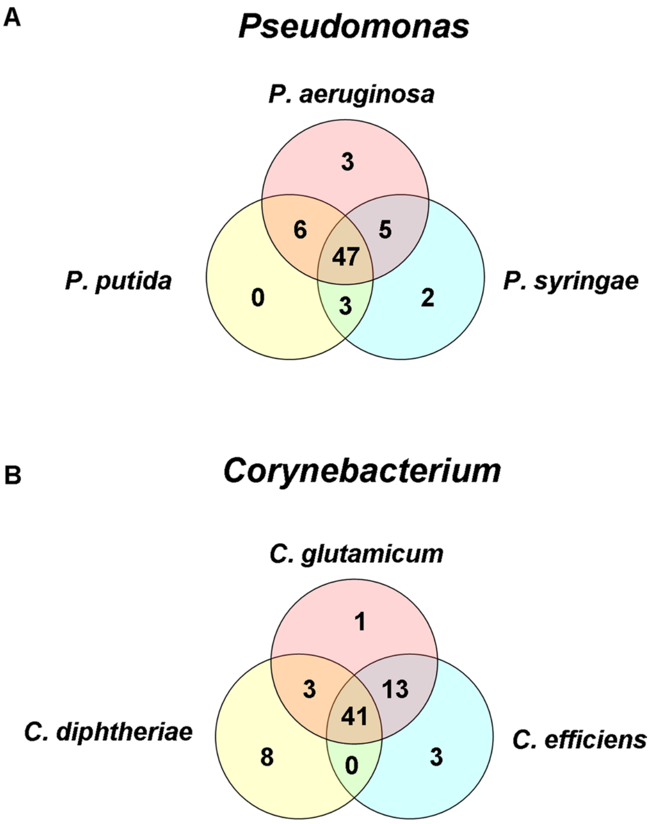


Figure 6. Venn Diagrams Showing the Distribution of Transporter Families among Species Belonging to the Same Genus

(A) Transporter family distribution among three *Pseudomonas* species.
(B) Transporter family distribution among three *Corynebacterium* species.
DOI: 10.1371/journal.pcbi.0010027.g006

All three *Corynebacterium* species share 41 transporter families. Interestingly, although *Co. diphtheriae* shows no amino acid productivity and has a reduced genome size [71], all the major types of amino acid exporters in *Co. glutamicum* [72] are conserved in *Co. diphtheriae*, e.g., the LysE family transporter for the export of basic amino acids, the RhtB family transporter for threonine efflux, the ThrE family transporter for threonine and serine export, and the LIV-E family transporter (BrnFE in *Co. glutamicum*), which is a two-component efflux pump exporting branched-chain amino acids [73]. The only difference observed among these organisms is the number of paralogs in the RhtB family: three in *Co. glutamicum,* two in *Co. efficiens,* and only one in *Co. diphtheriae*. The phylogenetic tree of the RhtB family suggests that gene duplication took place in the common ancestor of *Corynebacterium,* and that specific gene loss was responsible for the single RhtB transporter in *Co. diphtheriae*.

Conclusion

The rapid expansion of complete genome sequencing enabled us to conduct analyses of transporter capabilities on the whole-genome level. By comparing the membrane transport systems in Eubacteria, Archaea, and Eukaryota, we could draw conclusions as follows. (1) Eukaryotic species generally encode a larger number of transporters, but transporters account for a smaller percentage of total ORFs in eukaryotic than in prokaryotic species. Prokaryotic obligate intracellular pathogens and endosymbionts, as well as the eukaryotic parasites, possess the most limited repertoire of membrane transporters. (2) Organisms with a larger genome size tend to have a higher number of transporters. In prokaryotes and unicellular eukaryotes, this increase is primarily due to increased diversity of types of transporter. In multicellular eukaryotes, this increase is largely due to the greater number of paralogs by gene duplication or expansion in certain transporter families. (3) The distribution of different transporter types according to transport mode and energy coupling mechanism generally correlates with organisms' primary mechanism of energy generation. Compared to eukaryotes, prokaryotic species rely heavily on primary (active) transporters. Primary type transporters in Eubacteria and Archaea account for a much larger percentage of total transporters than any other transporter type. This phenomenon may be related to the absence of ABC-type uptake permeases in eukaryotes and, in some cases, the bioenergetic requirements and environmental constraints of prokaryotic organisms. (4) Energy-independent channel proteins are far more numerous in multicellular organisms and are often involved in cell–cell communication and signal transduction processes. Many channels are restricted to a single organismal type. The expression of different subunits of a channel in a timely fashion may be an essential step during embryonic development in mammals. (5) The PTS is only present in a subset of Eubacteria, and is completely absent in Archaea and Eukaryota. The expansion of sugar PTS systems in species dwelling in the gastrointestinal tract could provide the advantage to thrive in their ecological niches. (6) Hierarchical clustering of the phylogenetic profiles of transporter families showed that the distribution of transporter families appears to reflect a combination of evolutionary history and environment and lifestyle factors. (7) The distribution pattern of transporter families in species belonging to the same genus is usually parallel, with some notable exceptions that may reflect specific environmental differences.

Materials and Methods

We developed a semi-automated pipeline to annotate transport systems genome-wide, input the data into TransportDB database, and visualize the result through a Web interface [74]. The complete protein sequences from specific organisms were first searched against our curated database of transport proteins for similarity to known or putative transport proteins using BLAST [75,76]. All of the query proteins with significant hits (*E*-value < 0.001) were collected and searched against the NCBI nonredundant protein database and Pfam database [77]. Transmembrane protein topology was predicted by TMHMM [78]. A Web-based interface was created to facilitate the annotation processes, which incorporates number of hits to the transporter database, BLAST and HMM search *E*-value and score, number of predicted transmembrane segments, and the description of top hits to the nonredundant protein database. We also set up direct links between transporter classification family and COG classification [79] so that COG-based searches can inform the transporter annotation. The results can be viewed at the TransportDB Web site (http://www.membranetransport.org/).

To analyze the phylogenetic profiles of transporter families and predicted substrates, we assigned a profile to each transporter family or substrate. Each profile is a string with 141 entries (number of species analyzed). If a given family is present or a given substrate is transported in certain species, the value one was assigned at these positions (red for transporter families/purple for predicted substrates). If not, zero was assigned (black). Transporter families or substrates were clustered according to the similarity of their phylogenetic profiles using The Institute for Genomic Research's microarray multi-experiment viewer (MeV) [80] with two-dimensional hierarchical clustering as described by Eisen et al. [81].

Supporting Information

Figure S1. Comparison of the Percentage of Membrane Proteins with Six or More Transmembrane Segments That Were Annotated as "Hypothetical Protein" in Selected Archaea and Eubacteria

Found at DOI: 10.1371/journal.pcbi.0010027.sg001 (37 KB PDF).

Figure S2. Detailed View of the Hierarchical Clustering of Phylogenetic Profiles of Transporter Families

(A) Clustering of species.
(B) Clustering of transporter families.

Found at DOI: 10.1371/journal.pcbi.0010027.sg002 (722 KB PPT).

Figure S3. Phylogenetic Profiling of Predicted Transporter Substrates

Phylogenetic profiles were created for each predicted substrate. Each profile is a string with 141 entries (number of organisms analyzed). If a specific substrate is transported in a given organism, the value one is assigned at this position (purple). If not, zero is assigned (black). Organisms and substrates were clustered according to the similarity of their phylogenetic profiles.

Found at DOI: 10.1371/journal.pcbi.0010027.sg003 (671 KB PDF).

Figure S4. Venn Diagrams Showing the Distribution of Transporter Families among Species Belonging to the Same Genus

(A) Transporter family distribution among three *Bordetella* species.
(B) Transporter family distribution among three *Chlamydia* species.
(C) Transporter family distribution among three *Mycobacterium* species.
(D) Transporter family distribution among three *Pyrococcus* species.
(E) Transporter family distribution among three *Streptococcus* species.
(F) Transporter family distribution among three *Vibrio* species.

Found at DOI: 10.1371/journal.pcbi.0010027.sg004 (947 KB PDF).

Table S1. List of 141 Organisms Analyzed in This Study

Found at DOI: 10.1371/journal.pcbi.0010027.st001 (194 KB DOC).

Acknowledgments

We would like to thank Dr. Jonathan A. Eisen for comments and suggestions relating to the phylogenetic profiling analyses and Dr. Julie C. Dunning Hotopp for assistance with the MeV software and for critical reading of the manuscript. We also thank Robert L. Koenig for his contribution to the synopsis.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. QR and ITP conceived and designed the experiments and wrote the paper. QR analyzed the data and contributed materials/analysis tools. ■

References

1. Saier MH Jr (1999) Classification of transmembrane transport systems in living organisms. In: VanWinkle L, editor. Biomembrane transport. San Diego: Academic Press. pp. 265–276
2. Saier MH Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64: 354–411.
3. Saier MH Jr (1999) A functional-phylogenetic system for the classification of transport proteins. J Cell Biochem 75 (Suppl 32): 84–94.
4. Busch W, Saier MH Jr (2002) The transporter classification (TC) system, 2002. Crit Rev Biochem Mol Biol 37: 287–337.
5. Busch W, Saier MH (2004) The IUBMB-endorsed transporter classification system. Mol Biotechnol 27: 253–262.
6. Sweet G, Gandor C, Voegele R, Wittekindt N, Beuerle J, et al. (1990) Glycerol facilitator of *Escherichia coli:* Cloning of glpF and identification of the glpF product. J Bacteriol 172: 424–430.
7. Bolhuis H, Poelarends G, van Veen HW, Poolman B, Driessen AJ, et al. (1995) The lactococcal lmrP gene encodes a proton motive force-dependent drug transporter. J Biol Chem 270: 26092–26098.
8. Newman MJ, Foster DL, Wilson TH, Kaback HR (1981) Purification and reconstitution of functional lactose carrier from *Escherichia coli.* J Biol Chem 256: 11804–11808.
9. Viitanen P, Newman MJ, Foster DL, Wilson TH, Kaback HR (1986) Purification, reconstitution, and characterization of the lac permease of *Escherichia coli.* Methods Enzymol 125: 429–452.
10. Abramson J, Smirnova I, Kasho V, Verner G, Iwata S, et al. (2003) The lactose permease of *Escherichia coli:* Overall structure, the sugar-binding site and the alternating access model for transport. FEBS Lett 555: 96–101.

11. Elferink MG, Driessen AJ, Robillard GT (1990) Functional reconstitution of the purified phosphoenolpyruvate-dependent mannitol-specific transport system of *Escherichia coli* in phospholipid vesicles: Coupling between transport and phosphorylation. J Bacteriol 172: 7119–7125.
12. Postma PW, Lengeler JW, Jacobson GR (1993) Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol Rev 57: 543–594.
13. Bernal A, Ear U, Kyrpides N (2001) Genomes OnLine Database (GOLD): A monitor of genome projects world-wide. Nucleic Acids Res 29: 126–127.
14. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA (2005) Genome coverage, literally speaking: The challenge of annotating 200 genomes with 4 million publications. EMBO Rep 6: 397–399.
15. Davidson AL, Chen J (2004) ATP-binding cassette transporters in bacteria. Annu Rev Biochem 73: 241–268.
16. Schneider E, Hunke S (1998) ATP-binding cassette (ABC) transport systems: Functional and structural aspects of the ATP-hydrolyzing subunits/domains. FEMS Microbiol Lett 22: 1–20.
17. Marger MD, Saier MH Jr (1993) A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. Trends Biochem Sci 18: 13–20.
18. Saier MH Jr, Beatty JT, Goffeau A, Harley KT, Heijne WH, et al. (1999) The major facilitator superfamily. J Mol Microbiol Biotechnol 1: 257–279.
19. Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH Jr (2000) Microbial genome analyses: Comparative transport capabilities in eighteen prokaryotes. J Mol Biol 301: 75–100.
20. Paulsen IT, Sliwinski MK, Saier MH Jr (1998) Microbial genome analyses: Global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol 277: 573–592.
21. Paulsen IT, Sliwinski MK, Nelissen B, Goffeau A, Saier MH Jr (1998) Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. FEBS Lett 430: 116–125.
22. Chen GQ, Cui C, Mayer ML, Gouaux E (1999) Functional characterization of a potassium-selective prokaryotic glutamate receptor. Nature 402: 817–821.
23. Mayer ML, Olson R, Gouaux E (2001) Mechanisms for ligand binding to GluR0 ion channels: Crystal structures of the glutamate and serine complexes and a closed apo state. J Mol Biol 311: 815–836.
24. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273: 1058–1073.
25. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. Nucleic Acids Res 29: 123–125.
26. Ruder K, Winstead ER, Gibbs MS (2004) A quick guide to sequenced genomes. Rockville (Maryland): Genome News Network. Available: http://www.genomenewsnetwork.org/resources/sequenced__genomes/genome__guide__p1.shtml. Accessed 13 July 2005.
27. Gardner MJ, Hall N, Fung E, White O, Berriman M, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419: 498–511.
28. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, et al. (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414: 450453.
29. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101: 3160–3165.
30. Cases I, de Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. Trends Microbiol 11: 248.
31. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res 11: 555–565.
32. Sanchez-Fernandez R, Davies TG, Coleman JO, Rea PA (2001) The *Arabidopsis thaliana* ABC protein superfamily, a complete inventory. J Biol Chem 276: 30231–30244.
33. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.
34. Dean M, Allikmets R (2001) Complete characterization of the human ABC gene family. J Bioenerg Biomembr 33: 475–479.
35. Kung C, Blount P (2004) Channels in microbes: So many holes to fill. Mol Microbiol 53: 373–380.
36. Hong X (1998) Identification of major phylogenetic branches of inhibitory ligand-gated channel receptors. J Mol Evol 47: 323.
37. Nakanishi S, Masu M (1994) Molecular diversity and functions of glutamate receptors. Annu Rev Biophys Biomol Struct 23: 319–348.
38. Lujan R, Shigemoto R, Lopez-Bendito G (2005) Glutamate and GABA receptor signalling in the developing brain. Neuroscience 130: 567–580.
39. Johanson U, Karlsson M, Johansson I, Gustavsson S, Sjovall S, et al. (2001) The complete set of genes encoding major intrinsic proteins in *Arabidopsis* provides a framework for a new nomenclature for major intrinsic proteins in plants. Plant Physiol 126: 1358–1369.
40. Javot H, Maurel C (2002) The role of aquaporins in root water uptake. Ann Bot 90: 301–313.
41. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422: 859–868.
42. Borkovich KA, Alex LA, Yarden O, Freitag M, Turner GE, et al. (2004) Lessons from the genome sequence of *Neurospora crassa*: Tracing the path from genomic blueprint to multicellular organism. Microbiol Mol Biol Rev 68: 1–108.
43. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
44. Pellegrini M (2001) Computational methods for protein function analysis. Curr Opin Chem Biol 5: 46–50.
45. Vandepoele K, Van de Peer Y (2005) Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol 137: 31–42.
46. Marcotte EM, Xenarios I, van der Bliek AM, Eisenberg D (2000) Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci U S A 97: 12115–12120.
47. Levesque M, Shasha D, Kim W, Surette MG, Benfey PN (2003) Trait-to-gene: A computational method for predicting the function of uncharacterized genes. Curr Biol 13: 129–133.
48. Kriventseva EV, Biswas M, Apweiler R (2001) Clustering and analysis of protein families. Curr Opin Struct Biol 11: 334–339.
49. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, et al. (2002) The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci U S A 99: 13148–13153.
50. Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, et al. (2003) A genomic view of the human–*Bacteroides thetaiotaomicron* symbiosis. Science 299: 2074–2076.
51. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, et al. (2003) Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat Genet 35: 32–40.
52. Gerlach G, von Wintzingerode F, Middendorf B, Gross R (2001) Evolutionary trends in the genus *Bordetella*. Microbes Infect 3: 61–72.
53. Tolner B, Ubbink-Kok T, Poolman B, Konings WN (1995) Cation-selectivity of the L-glutamate transporters of *Escherichia coli, Bacillus stearothermophilus* and *Bacillus caldotenax:* Dependence on the environment in which the proteins are expressed. Mol Microbiol 18: 123–133.
54. Linton KJ, Higgins CF (1998) The *Escherichia coli* ATP-binding cassette (ABC) proteins. Mol Microbiol 28: 5–13.
55. Zientz E, Dandekar T, Gross R (2004) Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev 68: 745–770.
56. Reizer J, Reizer A, Saier MH Jr (1994) A functional superfamily of sodium/solute symporters. Biochim Biophys Acta 1197: 133–166.
57. Saier MH Jr, Eng BH, Fard S, Garg J, Haggerty DA, et al. (1999) Phylogenetic characterization of novel transport protein families revealed by genome analyses. Biochim Biophys Acta 1422: 1–56.
58. Pajor AM (2000) Molecular properties of sodium/dicarboxylate cotransporters. J Membr Biol 175: 1–8.
59. Pao SS, Paulsen IT, Saier MH Jr (1998) Major facilitator superfamily. Microbiol Mol Biol Rev 62: 1–34.
60. Jack DL, Paulsen IT, Saier MH (2000) The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. Microbiology 146: 1797–1814.
61. Stover CK, Pham XQ, Erwin AL, Mizoguchi SD, Warrener P, et al. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature 406: 959–964.
62. Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, et al. (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 4: 799–808.
63. Buell CR, Joardar V, Lindeberg M, Selengut J, Paulsen IT, et al. (2003) The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. Proc Natl Acad Sci U S A 100: 10181–10186.
64. Kalinowski J, Bathe B, Bartels D, Bischoff N, Bott M, et al. (2003) The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of aspartate-derived amino acids and vitamins. J Biotechnol 104: 5–25.
65. Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, et al. (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. Genome Res 13: 1572–1579.
66. Cerdeno-Tarraga AM, Efstratiou A, Dover LG, Holden MTG, Pallen M, et al. (2003) The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res 31: 6516–6523.
67. Schlosser A, Meldorf M, Stumpe S, Bakker EP, Epstein W (1995) TrkH and its homolog, TrkG, determine the specificity and kinetics of cation transport by the Trk system of *Escherichia coli*. J Bacteriol 177: 1908–1910.
68. Trchounian A, Kobayashi H (1999) Kup is the major K+ uptake system in *Escherichia coli* upon hyper-osmotic stress at a low pH. FEBS Lett 447: 144–148.
69. Engel P, Kramer R, Unden G (1994) Transport of C4-dicarboxylates by anaerobically grown *Escherichia coli*. Energetics and mechanism of exchange, uptake and efflux. Eur J Biochem 222: 605–614.
70. Kelly DJ, Thomas GH (2001) The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. FEMS Microbiol Rev 25: 405–424.
71. Nishio Y, Nakamura Y, Usuda Y, Sugimoto S, Matsui K, et al. (2004)

Evolutionary process of amino acid biosynthesis in *Corynebacterium* at the whole genome level. Mol Biol Evol 21: 1683–1691.

72. Eggeling L, Sahm H (2003) New ubiquitous translocators: Amino acid export by *Corynebacterium glutamicum* and *Escherichia coli*. Arch Microbiol 180: 155–160.
73. Kennerknecht N, Sahm H, Yen MR, Patek M, Saier MH Jr, et al. (2002) Export of L-isoleucine from *Corynebacterium glutamicum:* A two-gene-encoded member of a new translocator family. J Bacteriol 184: 3947–3956.
74. Ren Q, Kang KH, Paulsen IT (2004) TransportDB: A relational database of cellular membrane transport systems. Nucleic Acids Res 32: D284–D288.
75. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
76. Altschul SF, Gish W (1996) Local alignment statistics. Methods Enzymol 266: 460–480.
77. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26: 320–322.
78. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol 305: 567–580.
79. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: New developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22–28.
80. Saeed AI, Sharov V, White J, Li J, Liang W, et al. (2003) TM4: A free, open-source system for microarray data management and analysis. Biotechniques 34: 374–378.
81. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.

D284–D288 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh016

TransportDB: a relational database of cellular membrane transport systems

Qinghu Ren, Katherine H. Kang and Ian T. Paulsen*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received August 5, 2003; Revised and Accepted September 3, 2003

ABSTRACT

TransportDB (http://www.membranetransport.org) is a relational database designed for describing the predicted cellular membrane transport proteins in organisms whose complete genome sequences are available. For each organism, the complete set of membrane transport systems was identified and classified into different types and families according to putative membrane topology, protein family, bioenergetics and substrate specificities. Web pages were created to provide user-friendly interfaces to easily access, query and download the data. Additional features, such as a BLAST search tool against known transporter protein sequences, comparison of transport systems from different organisms and phylogenetic trees of individual transporter families are also provided. TransportDB will be regularly updated with data obtained from newly sequenced genomes.

INTRODUCTION

Transport systems, which function in the translocation of solutes, play essential roles in cellular metabolism and activities. They mediate the entry of nutrients into cytoplasm and the extrusion of metabolite wastes, maintain a stable internal environment inside the cell by regulating the uptake and efflux of ions, protect cells from environmental insults, and enhance communications between cells through the secretion of proteins, carbohydrates and lipids. Different transport systems differ in their putative membrane topology, energy coupling mechanism and substrate specificities (1). The most common energy coupling mechanisms are the utilization of adenosine triphosphate (ATP), phosphoenolpyruvate (PEP), or chemiosmotic energy in the form of sodium ion or proton electrochemical gradients. The Transporter Classification (TC) system (http://www-biology.ucsd.edu/~msaier/transport/) represents a systematic approach to classify transport systems according to the mode of transport, energy coupling mechanism, molecular phylogeny and substrate specificity (2–4). The transport mode and the energy coupling mechanism serve as the primary base for the classification due to their relatively stable characteristics. There are four characterized classes of solute transporters in the TC system: channels, secondary transporters, primary active transporters and group translocators. Transporters of unknown mechanism or function are included as a distinct class. Channels are energy-independent transporters that exhibit higher rates of transport and lower stereospecificity compared with other transporter classes. Primary active transporters couple the transport process to a primary source of energy, such as a chemical reaction (e.g. ATP hydrolysis). Secondary transporters utilize an ion or solute electrochemical gradient, e.g. proton/sodium motive force, to drive the transport process. Group translocators modify their substrates during the transport process. For example, the bacterial phsphotransferase system (PTS) phosphorylates its sugar substrates using PEP as the phosphoryl donor and energy source and releases them into cytoplasm as sugar–phosphates. Each transporter class is further classified into individual families or superfamilies according to their function, phylogeny and/or substrate specificity (1).

Since the advent of genomic sequencing technologies, such as whole-genome shotgun sequencing, the complete sequences of 135 prokaryotic and eukaryotic genomes have been published to date, with more than 500 additional genome sequencing projects currently underway around the world (Gold Genomes Online Database, http://ergo.integratedgenomics.com/GOLD/). Convenient and effective methods have to be developed to handle and analyze the immense amount of data generated by whole-genome sequencing projects. It has been found that 5–12% of the complete bacterial genome is often dedicated to transport proteins and associated factors (4,5). An in-depth look at transport proteins is vital to the understanding of the metabolic capability of organisms. However, due to the occurrence of large complex transporter gene families, such as the ATP-binding cassette (ABC) and major facilitator superfamily (MFS), and the presence of multiple transporter gene paralogs in many organisms, it is often problematic to annotate these transport proteins by current primary annotation methods. We have been working on a systematic genome-wide analysis of cellular transport systems. Previously, we reported a comprehensive analysis of the transport systems in 18 prokaryotic organisms (4,5) and in yeast (6) based on the TC system. Here we have expanded our analyses to 121 prokaryotic and eukaryotic systems. TransportDB (http://www.membranetransport.org/), a web-integrated database, was built up to store the results of our analyses and to provide user-friendly interfaces to access the data. A semi-automated pipeline was also set up to facilitate the efficient analyses of transport systems.

*To whom correspondence should be addressed. Tel: +1 301 838 3531; Fax: +1 301 838 0200; Email: ipaulsen@tigr.org

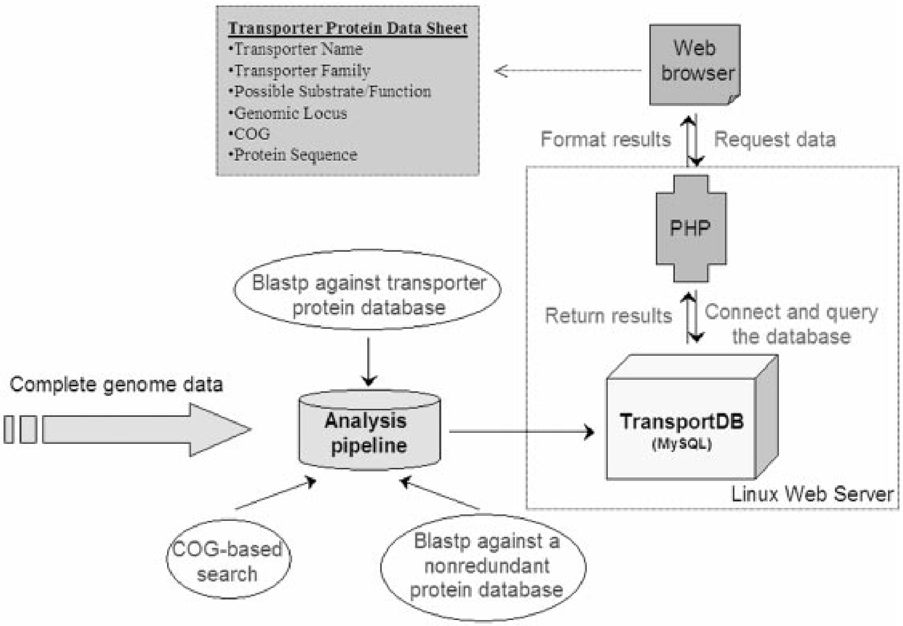


Figure 1. Overview of TransportDB and transport system analysis pipeline. Transporter protein data sets are stored in a MySQL relational database. Users can search the database through a web interface. All web pages are generated dynamically using PHP, which connects, queries the database and formats the results to generate a final data sheet. A pipeline was set up to use the complete genomic protein sequence as input, retrieve transporter proteins and assign them to specific transporter families and substrate/function information. The output can be loaded directly into the database and visualized on the web pages.

DATABASE CONTENT AND STRUCTURE

TransportDB is a MySQL database (http://www.mysql.com/) that is queried using PHP (http://www.php.net/) (see Fig. 1). PHP, a server-side scripting language, mediates the interaction with the user, the database and the computational tools. The database and PHP pages are stored on a Linux web server. TransportDB contains the complete predicted transport profile for each organism, including information on transporter family, TC classification, transporter name, possible substrate/function, genomic locus, COG classification and protein sequence. Where appropriate, links are provided to other databases, such as Entrez (7), COG (8), PubMed (9), TCDB (1) and individual organism genome sequence databases.

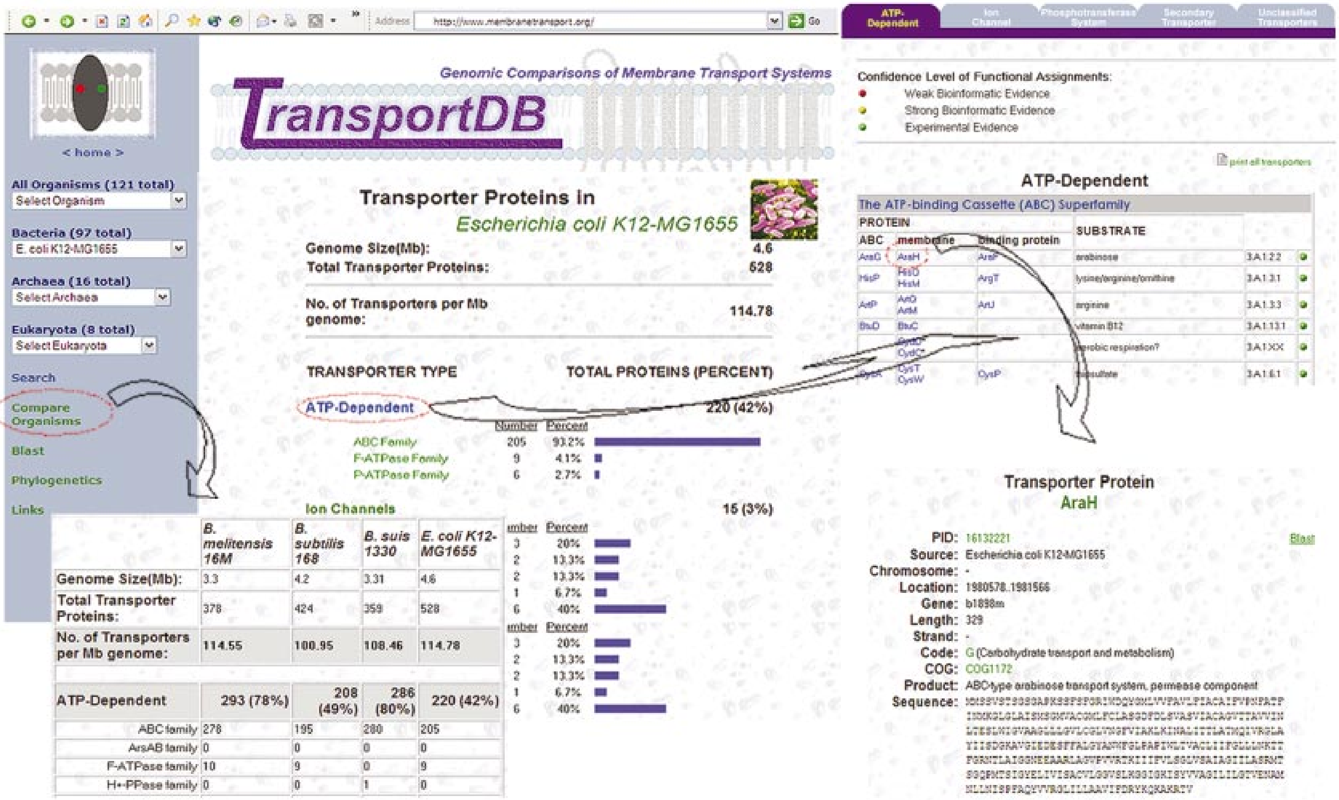
Currently, TransportDB contains data from 121 organisms, including 97 bacteria, 16 archaea and eight eukaryota. This collection of organisms represents a broad phylogenetic diversity. A total of 36 137 transporter proteins was assigned to 136 families. Some of these families are very large superfamilies with over a thousand members, such as the ABC superfamily (17 209 total) (10,11), the MFS superfamily (3635 total) (12,13) and the bacterial sugar-specific PTS superfamily (1341 total) (14,15). The transporter profiles from other organisms whose genomic sequencing are underway will be added to the database once their genome sequences are published.

DATABASE ACCESS

TransportDB is available on the web at http://www.membranetransport.org/ (Fig. 2).

The database can be browsed by organism name using the drop-down boxes on the left of the web page. For each organism, its complete membrane transport complement was classified into different families according to the TC classification system. These families were grouped into five distinct types based on mode of transport and energy-coupling mechanisms: ion channels, secondary transporters, ATP-dependent (primary active) transporters, PTSs and unclassified transporters, which have unknown mechanisms of action. Individual transporter types can be accessed by clicking the tabs at the top or the links on the summary page (Fig. 2). For each transporter family, a detailed list of transporters with their predicted substrates is shown with links to the individual protein page which contains genomic locus, COG, protein sequence and annotation information. A summary page is also available for each organism, summarizing the whole transporter system, including transporter types and individual transporter families, and their statistics. TransportDB is searchable by transporter type, transporter family, transporter protein name or substrate. The results are grouped by transporter family and organism, with links to individual family and protein pages.

Comparisons of the transporter contents of different organisms can provide insight into their physiology and life-style. To view the transport profiles across species, we created a ‘Compare Organisms’ section. Users can choose any two or more organisms to compare their overall numbers of recognized transporters, numbers of transporters relative to genome size and the constituents of each transporter type and family. Previous studies have shown that transporters of similar function characteristically cluster together in


Address http://www.membranetransport.org/
Go
ATP-Dependent
Ion Channel
Phosphotransferase System
Secondary Transporter
Unclassified Transporters
< home >
TransportDB
Genomic Comparisons of Membrane Transport Systems
Search
Compare Organisms
Blast
Phylogenetics
Links
print all transporters
Blast

All Organisms (121 total)
Select Organism

Bacteria (97 total)
E. coli K12-MG1655

Archaea (16 total)
Select Archaea

Eukaryota (8 total)
Select Eukaryota

Transporter Proteins in
Escherichia coli K12-MG1655

Genome Size(Mb):	4.6
Total Transporter Proteins:	528
No. of Transporters per Mb genome:	114.78

TRANSPORTER TYPE	Number	Percent	TOTAL PROTEINS (PERCENT)
ATP-Dependent			220 (42%)
ABC Family	205	93.2%	
F-ATPase Family	9	4.1%	
P-ATPase Family	6	2.7%	
Ion Channels			15 (3%)

umber	Percent
3	20%
2	13.3%
2	13.3%
1	6.7%
6	40%

umber	Percent
3	20%
2	13.3%
2	13.3%
1	6.7%
6	40%

	B. melitensis 16M	*B. subtilis 168*	*B. suis 1330*	*E. coli K12-MG1655*
Genome Size(Mb):	3.3	4.2	3.31	4.6
Total Transporter Proteins:	378	424	359	528
No. of Transporters per Mb genome:	114.55	100.95	108.46	114.78
ATP-Dependent	293 (78%)	208 (49%)	286 (80%)	220 (42%)
ABC family	278	195	280	205
ArsAB family	0	0	0	0
F-ATPase family	10	9	0	9
H+-PPase family	0	0	1	0

Confidence Level of Functional Assignments:

- ● Weak Bioinformatic Evidence
- ● Strong Bioinformatic Evidence
- ● Experimental Evidence

ATP-Dependent

The ATP-binding Cassette (ABC) Superfamily					
PROTEIN			SUBSTRATE		
ABC	membrane	binding protein			
AraG	AraH	AraF	arabinose	3.A.1.2.2	●
HisP	HisQ HisM	ArgT	lysine/arginine/ornithine	3.A.1.3.1	●
ArtP	ArtQ ArtM	ArtJ	arginine	3.A.1.3.3	●
BtuD	BtuC		vitamin B12	3.A.1.13.1	●
	CydD CydC		aerobic respiration?	3.A.1.XX	●
CysA	CysT CysW	CysP	thiosulfate	3.A.1.6.1	●

Transporter Protein
AraH

PID: 16132221
Source: Escherichia coli K12-MG1655
Chromosome: -
Location: 1980578..1981566
Gene: b1898m
Length: 329
Strand: -
Code: G (Carbohydrate transport and metabolism)
COG: COG1172
Product: ABC-type arabinose transport system, permease component
Sequence: MMSSVSTSGSGAPKSSFSFGRIWDQYGMLVVFAVLFIACAIFVPNFATF
INMKGLGLAISMSGMVACGMLFCLASGDFDLSVASVIACAGVTTAVVIN
LTESLWIGVAAGLLLGVLCGLVNGFVIAKLKINALITTLATMQIVRGLA
YIISDGKAVGIEDESFFALGYANWFGLPAPIWLTVACLIIFGLLLNKTT
FGRNTLAIGGNEEAARLAGVPVVRTKIIIFVLSGLVSAIAGIILASRMT
SGQPMTSIGYELIVISACVLGGVSLKGGIGKISYVVAGILILGTVENAM
NLLNISPFAQYVVRGLILLAAVIFDRYKQKAKRTV

Figure 2. Graphic illustration of the structure of TransportDB. Transport proteins are grouped by organism, transporter type and transporter family. Users can choose an organism from the drop-down boxes at the left. Each organism has a summary page to overview the whole transport system. Individual transporter type or family can be viewed by clicking the links or tabs at the top. Each transporter protein also has an individual page to show the genomic locus, COG information and protein sequence. Links to GenBank, Entrez and COG are also provided. The whole transport profile can also be compared with any other organisms stored in the database.

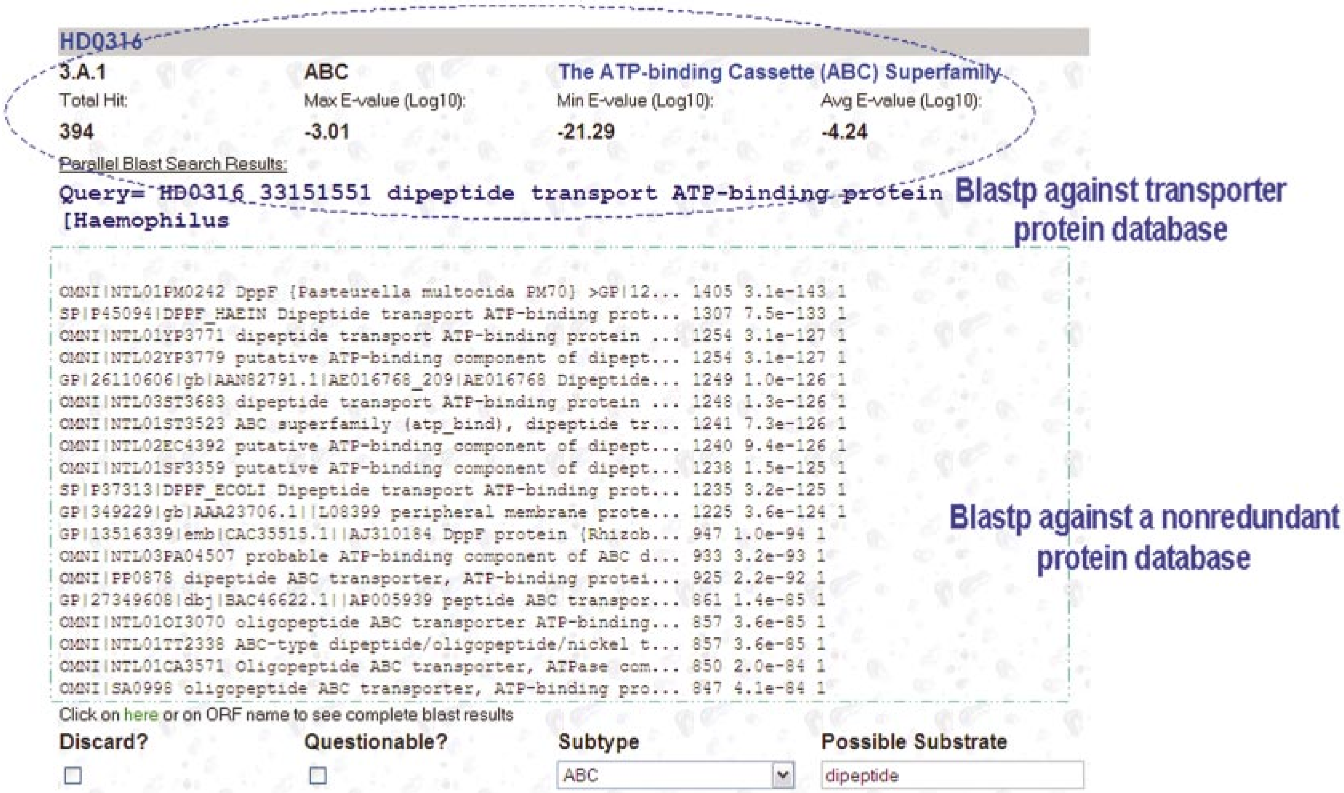
phylogenetic analyses, hence substrate specificity appears to be a conserved evolutionary trait in transporters (4,5,12,16). In the 'Phylogenetics' section of the TransportDB website, pre-computed neighbor-joining trees for each of the transporter families are available to view. All members of each family in the current database are also available to view or download in FASTA or multiple sequence alignment formats. In addition, the whole transporter database is also available for BLAST search. Users can submit the unknown protein sequence in the 'Blast' section. The output of the BLAST search includes transporter family information in addition to the standard features (17).

It should be noted that TransportDB focuses on solute and ion transport across the cytoplasmic membrane, and hence does not include some types of transporters that are shown in TC classification: outer membrane transporter proteins (18); proteins in the *Escherichia coli* TonB/ExbB/ExbD complex that transduce energy to drive outer membrane transport processes (19); proteins involved in the protein secretory pathways (20); proteins involved in proton and sodium ion-translocating electron transfer processes (21); sodium ion-transporting carboxylic acid decarboxylases (22); flagellar motor proteins (23); proteins involved in DNA uptake (24). Auxiliary transport proteins (such as the MFP family) (25) or membrane–periplasmic auxiliary proteins of the MPA1 and MPA2 families (26) were treated as components of the transporters with which they function, rather than separate transport systems.

ANALYSIS PIPELINE

With the rapid increase in the number of published genomes, efficient and effective approaches are required to speed up transport system analysis processes. Previous methods used by us for transporter analysis (4,5) required intensive personal involvement and manual curation. Recently we have developed a new semi-automated pipeline to analyze a genome-wide transport system, input the data into TransportDB and visualize it on the web page (Fig. 1).

The methodology we have developed is as follows: the complete protein sequences from specific organisms were first searched against the curated set of proteins with family assignment in our transporter protein database for similarity to known or putative transporter proteins using BLAST (27,28). All the proteins with an e-value of <0.001 were collected and searched against a non-redundant general protein database. A web-based interface was created to incorporate the output of two BLAST searches (Fig. 3) and to help a human annotator make a decision and assign possible substrates or functions. The useful information includes: number of hits to the



HD0316

3.A.1	ABC	The ATP-binding Cassette (ABC) Superfamily	
Total Hit:	Max E-value (Log10):	Min E-value (Log10):	Avg E-value (Log10):
394	-3.01	-21.29	-4.24

Parallel Blast Search Results:

```
Query= HD0316_33151551 dipeptide transport ATP-binding protein
[Haemophilus

OMNI|NTL01PM0242 DppF {Pasteurella multocida PM70} >GP|12... 1405 3.1e-143 1
SP|P45094|DPPF_HAEIN Dipeptide transport ATP-binding prot... 1307 7.5e-133 1
OMNI|NTL01YP3771 dipeptide transport ATP-binding protein ... 1254 3.1e-127 1
OMNI|NTL02YP3779 putative ATP-binding component of dipept... 1254 3.1e-127 1
GP|26110606|gb|AAN82791.1|AE016768_209|AE016768 Dipeptide... 1249 1.0e-126 1
OMNI|NTL03ST3683 dipeptide transport ATP-binding protein ... 1248 1.3e-126 1
OMNI|NTL01ST3523 ABC superfamily (atp_bind), dipeptide tr... 1241 7.3e-126 1
OMNI|NTL02EC4392 putative ATP-binding component of dipept... 1240 9.4e-126 1
OMNI|NTL01SF3359 putative ATP-binding component of dipept... 1238 1.5e-125 1
SP|P37313|DPPF_ECOLI Dipeptide transport ATP-binding prot... 1235 3.2e-125 1
GP|349229|gb|AAA23706.1||L08399 peripheral membrane prote... 1225 3.6e-124 1
GP|13516339|emb|CAC35515.1||AJ310184 DppF protein (Rhizob... 947 1.0e-94 1
OMNI|NTL03PA04507 probable ATP-binding component of ABC d... 933 3.2e-93 1
OMNI|PP0878 dipeptide ABC transporter, ATP-binding protei... 925 2.2e-92 1
GP|27349608|dbj|BAC46622.1||AP005939 peptide ABC transpor... 861 1.4e-85 1
OMNI|NTL01OI3070 oligopeptide ABC transporter ATP-binding... 857 3.6e-85 1
OMNI|NTL01TT2338 ABC-type dipeptide/oligopeptide/nickel t... 857 3.6e-85 1
OMNI|NTL01CA3571 Oligopeptide ABC transporter, ATPase com... 850 2.0e-84 1
OMNI|SA0998 oligopeptide ABC transporter, ATP-binding pro... 847 4.1e-84 1
```

Click on here or on ORF name to see complete blast results

Discard?	Questionable?	Subtype	Possible Substrate
□	□	ABC	dipeptide

Blastp against transporter protein database
Blastp against a nonredundant protein database

Figure 3. Transporter annotation page. The complete genomic protein sequences were searched against our transporter database and a non-redundant general protein database using BLAST. The results were incorporated into a web-based interface to help the annotator to make a decision on family and substrate properties. Useful information includes: number of hits to the transporter database; maximum, minimum and average e-values; and the description of top hits to the general protein database. Links to TCDB, Entrez and COG are also provided.

transporter database; maximum, minimum and average e-values; and the description of top hits to the general protein database. We also set up direct links between TC family and the COG classification (8) so that COG-based search can also help the annotation processes. The output of the analysis process is in a tab-delimited format, which can be loaded directly into TransportDB and shown on the web pages.

To test the new analysis pipeline, we compared the analysis process on several test genomes by the new pipeline to that by the approaches we had used earlier (4,5). The new pipeline greatly reduced the time annotators spent on the analysis process. In addition, the new pipeline has shown improved sensitivity and selectivity over other approaches. This pipeline has been used in the analysis of over 40 prokaryotic and eukaryotic genomes since its inception.

FUTURE PERSPECTIVES

In summary, we developed a relational database and an analysis pipeline for the comprehensive representation of cellular membrane transport systems in various prokaryotic and eukaryotic organisms. User-friendly web interfaces were designed to easily query the database and access the various features. To our knowledge, this is the only database devoted to the identification and classification of transporter homologs in complete genomes, as well as providing comparative and phylogenetic tools for analyzing the data.

We are continuing to expand the TransportDB database to incorporate data from newly published genomes. TransportDB will be routinely updated at least once per month to ensure the timely report of data. Future planned improvements will include the prediction of transmembrane segments (TMSs) in each transporter protein, prediction of orthologs from different organisms, automated pictorial representation of transport system, links to Swiss-Prot (29) and other online resources, and web-based data submission for TransportDB users.

ACKNOWLEDGEMENTS

We would like to thank Dr Jean-Francois Tomb for comments and suggestions to our web site and Dr Martin Wu for help with the construction of phylogenetic trees. The Linux web server and technical support were kindly provided by Dr Jon Oliver.

REFERENCES

1. Saier,M.H.,Jr (2000) A functional–phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 354–411.
2. Saier,M.H.,Jr (1999) Genome archeology leading to the characterization and classification of transport proteins. *Curr. Opin. Microbiol.*, **2**, 555–561.
3. Saier,M.H.,Jr (1999) Classification of transmembrane transport systems in living organisms. In VanWinkle,L. (ed.), *Biomembrane Transport*. Academic Press, San Diego, CA, pp. 265–276.

4. Paulsen,I.T., Sliwinski,M.K. and Saier,M.H.,Jr (1998) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.*, **277**, 573–592.
5. Paulsen,I.T., Nguyen,L., Sliwinski,M.K., Rabus,R. and Saier,M.H.,Jr (2000) Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.*, **301**, 75–100.
6. Paulsen,I.T., Sliwinski,M.K., Nelissen,B., Goffeau,A. and Saier,M.H.,Jr (1998) Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett.*, **430**, 116–125.
7. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
8. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
9. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
10. Tomii,K. and Kanehisa,M. (1998) A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.*, **8**, 1048–1059.
11. Saurin,W., Hofnung,M. and Dassa,E. (1999) Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J. Mol. Evol.*, **48**, 22–41.
12. Pao,S.S., Paulsen,I.T. and Saier,M.H.,Jr (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.*, **62**, 1–34.
13. Saier,M.H.,Jr, Beatty,J.T., Goffeau,A., Harley,K.T., Heijne,W.H., Huang,S.C., Jack,D.L., Jahn,P.S., Lew,K., Liu,J. *et al.* (1999) The major facilitator superfamily. *J. Mol. Microbiol. Biotechnol.*, **1**, 257–279.
14. Hu,K.Y. and Saier,M.H.,Jr (2002) Phylogeny of phosphoryl transfer proteins of the phosphoenolpyruvate-dependent sugar-transporting phosphotransferase system. *Res. Microbiol.*, **153**, 405–415.
15. Reizer,J., Bachem,S., Reizer,A., Arnaud,M., Saier,M.H.,Jr and Stulke,J. (1999) Novel phosphotransferase system genes revealed by genome analysis—the complete complement of PTS proteins encoded within the genome of *Bacillus subtilis*. *Microbiology*, **145**, 3419–3429.
16. Jack,D.L., Paulsen,I.T. and Saier,M.H. (2000) The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology*, **146**, 1797–1814.
17. Yuan,Y.P., Eulenstein,O., Vingron,M. and Bork,P. (1998) Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–289.
18. Jeanteur,D., Lakey,J.H. and Pattus,F. (1991) The bacterial porin superfamily: sequence alignment and structure prediction. *Mol. Microbiol.*, **5**, 2153–2164.
19. Braun,V., Pilsl,H. and Gross,P. (1994) Colicins: structures, modes of action, transfer through membranes and evolution. *Arch. Microbiol.*, **161**, 199–206.
20. Saier,M.H.,Jr, Werner,P.K. and Muller,M. (1989) Insertion of proteins into bacterial membranes: mechanism, characteristics and comparisons with the eucaryotic process. *Microbiol. Rev.*, **53**, 333–366.
21. Dimroth,P. (1997) Primary sodium ion translocating enzymes. *Biochim. Biophys. Acta*, **1318**, 11–51.
22. Buckel,W. (2001) Sodium ion-translocating decarboxylases. *Biochim. Biophys. Acta*, **1505**, 15–27.
23. Nguyen,C.C. and Saier,M.H.,Jr (1996) Structural and phylogenetic analysis of the MotA and MotB families of bacterial flagellar motor proteins. *Res. Microbiol.*, **147**, 317–332.
24. Macfadyen,L.P., Dorocicz,I.R., Reizer,J., Saier,M.H.,Jr and Redfield,R.J. (1996) Regulation of competence development and sugar utilization in *Haemophilus influenzae* Rd by a phosphoenolpyruvate:fructose phosphotransferase system. *Mol. Microbiol.*, **21**, 941–952.
25. Dinh,T., Paulsen,I.T. and Saier,M.H.,Jr (1994) A family of extracytoplasmic proteins that allow transport of large molecules across the outer membranes of Gram-negative bacteria. *J. Bacteriol.*, **176**, 3825–3831.
26. Paulsen,I.T., Beness,A.M. and Saier,M.H.,Jr (1997) Computer-based analyses of the protein constituents of transport systems catalysing export of complex carbohydrates in bacteria. *Microbiology*, **143**, 2685–2699.
27. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
28. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
29. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

MOLECULAR AND CELLULAR BIOLOGY, Apr. 2007, p. 2648–2660 Vol. 27, No. 7
0270-7306/07/$08.00+0 doi:10.1128/MCB.01910-06

Phosphorylation of the SQ H2A.X Motif Is Required for Proper Meiosis and Mitosis in *Tetrahymena thermophila*▽

Xiaoyuan Song,[1]†‡ Elizabeta Gjoneska,[2]† Qinghu Ren,[1]†§ Sean D. Taverna,[2]
C. David Allis,[2] and Martin A. Gorovsky[1]*

Department of Biology, University of Rochester, Rochester, New York,[1] and Laboratory of Chromatin Biology, Rockefeller University, New York, New York[2]

Received 9 October 2006/Returned for modification 30 November 2006/Accepted 8 January 2007

Phosphorylation of the C terminus SQ motif that defines H2A.X variants is required for efficient DNA double-strand break (DSB) repair in diverse organisms but has not been studied in ciliated protozoa. *Tetrahymena* H2A.X is one of two similarly expressed major H2As, thereby differing both from mammals, where H2A.X is a quantitatively minor component, and from *Saccharomyces cerevisiae* where it is the only type of major H2A. *Tetrahymena* H2A.X is phosphorylated in the SQ motif in both the mitotic micronucleus and the amitotic macronucleus in response to DSBs induced by chemical agents and in the micronucleus during prophase of meiosis, which occurs in the absence of a synaptonemal complex. H2A.X is phosphorylated when programmed DNA rearrangements occur in developing macronuclei, as for immunoglobulin gene rearrangements in mammals, but not during the DNA fragmentation that accompanies breakdown of the parental macronucleus during conjugation, correcting the previous interpretation that this process is apoptosis-like. Using strains containing a mutated (S134A) SQ motif, we demonstrate that phosphorylation of this motif is important for *Tetrahymena* cells to recover from exogenous DNA damage and is required for normal micronuclear meiosis and mitosis and, to a lesser extent, for normal amitotic macronuclear division; its absence, while not lethal, leads to the accumulation of DSBs in both micro- and macronuclei. These results demonstrate multiple roles of H2A.X phosphorylation in maintaining genomic integrity in different phases of the *Tetrahymena* life cycle.

Histone H2A.X is defined by the presence of a conserved SQ motif at the C terminus of a histone H2A, regardless of whether this H2A is a minor variant that is distinct and longer than the major H2A, as in mammals (49, 55), or whether it is the "major" H2A, as in *Saccharomyces cerevisiae* (16, 61), or is on another conserved variant (H2A.Z), as in *Drosophila* (83). The SQ motif is invariant and, in all cases but one (a predicted H2A.X in *Gallus*, accession no. XP_416906, derived from an annotated genomic sequence NW_060235), it localizes 3 to 4 residues from the C terminus, followed by a penultimate acidic residue (E/D) and then a terminal hydrophobic residue (Y/F/I/L) (61). Defined in this way, the H2A.X motif is found in most, if not all, eukaryotes from primitive ones, like *Giardia* (89), to higher organisms, like humans (49), although the histone it is on may not always have been named H2A.X.

DNA double strand breaks (DSBs), whether induced by external sources (ionizing radiation or drugs), by endogenous damage (free-radicals or replication fork collapse), or by developmentally programmed events [V(D)J joining, mating type switching, meiotic recombination, or apoptosis], invariably cause the serine in the SQ motif to become phosphorylated within minutes to produce an isoform commonly referred to as γ-H2A.X (28, 29, 44). The enzymes that phosphorylate the SQ motif are phosphatidylinositol 3-kinase-like kinase family members, Mec1 and Tel1 in *S. cerevisiae* (23, 74) and ATM, ATR, and DNA-dependent protein kinase (DNA-PK) in higher eukaryotes (9, 58, 77, 84).

Formation of γ-H2A.X is an evolutionarily conserved response to DSBs, as indicated by the fact that an anti-γ-H2A.X antibody raised against a synthetic phosphorylated peptide containing the mammalian γ-H2A.X sequence can recognize DSB-induced γ-H2A.X from diverse species (67). H2A.X phosphorylation in response to DSBs extends for megabases in sequences flanking the DSB sites in mammalian cells (67) and for 50 to 100 kb surrounding a single induced DSB in budding yeast (81). Thus, H2A.X phosphorylation is a highly sensitive DNA damage sensor and is required for efficient DSB repair (12, 23). In addition, γ-H2A.X is required to maintain genome stability (12) and has a role in condensing and inactivating sex chromosomes in male meiosis in mice (27).

Many proteins interact, either directly or indirectly, with γ-H2A.X and appear at the break sites or in broader areas surrounding DNA breaks after the appearance of γ-H2A.X (reviewed in references 28 and 29). These proteins include histone modifiers like the histone acetyltransferase NuA4 (22), the histone deacetylase Sin3 (34), chromatin remodelers like Ino80 (22, 53, 82) and SwrC (22), the Tip60 complex which has both histone acetyltransferase and ATP-dependent chromatin remodeling activities (43), DNA repair complexes (Mre11/Rad50/Nbs1 in mammals, and Mre11/Rad50/Xrs2 in budding yeast) (12, 36, 58), checkpoint proteins 53BP1 (85) and Crb2

* Corresponding author. Mailing address: Department of Biology, University of Rochester, Rochester, NY 14627. Phone: (585) 275-6988. Fax: (585) 275-2070. E-mail: goro@mail.rochester.edu.

† X.S., E.G., and Q.R. were equal contributors to this paper.

‡ Present address: School of Medicine, University of California—San Diego, La Jolla, CA 92037.

§ Present address: The Institute for Genomic Research, Rockville, MD 20850.

▽ Published ahead of print on 22 January 2007.

(56), and cohesin (81). Despite initial speculation that it recruits repair proteins to DNA breaks, γ-H2A.X appears to function in DSB repair by concentrating or retaining the modifying, remodeling, and repair proteins (11, 53), which may be recruited to the damage sites by redundant or alternative mechanisms (53). Removing γ-H2A.X after DNA repair also is required for cells to recover from the DNA damage checkpoint and resume their normal functions (35). By recruitment or retention of the SwrC complex (22) and Tip60 (43), γ-H2A.X may mediate its own removal from the altered chromatin structure produced by these recruited remodeling or modifying complexes. The proteasome also localizes at DNA damage sites and is required for proper DNA-damage responses (41), providing another possible mechanism for turning off the pathway initiated by γ-H2A.X. Recently, protein phosphatase 2A was shown to dephosphorylate γ-H2A.X in mammalian cells (17), and phosphatase Pph3 in *S. cerevisiae* was identified in a phosphatase complex responsible for γ-H2A.X dephosphorylation in yeast (35).

Repairing DSBs is crucial to maintain genome integrity in eukaryotes, as a failure to do so will result in acentric chromosome fragments that will be lost during mitosis. DSBs are mainly repaired either by homologous recombination (HR), in which two broken DNA ends join together based on homologous DNA pairing and strand exchange (79), or by nonhomologous end-joining (NHEJ), in which the broken DNA ends are joined together without using long homologous regions (for a review, see reference 70). Absence of γ-H2A.X or loss of H2A.X results in inefficient NHEJ (23) and HR (12). Defects in NHEJ or HR activities result in sensitivity to genotoxic agents, mitotic and meiotic chromosome aberrations, and destabilization of the genome (39, 64, 73, 80).

We have been studying the function of histones and their modifications in the ciliated protozoan, *Tetrahymena thermophila*. As in most ciliates, cells in this organism contain two highly dimorphic nuclei: a germ line micronucleus (MIC) and a somatic macronucleus (MAC). The diploid MIC contains five pairs of chromosomes, divides mitotically, and is transcriptionally inactive during vegetative growth. In contrast, the MAC is transcriptionally active, contains ~225 acentric chromosomes (13, 19, 24), each in ~45 copies, which are derived by fragmentation, telomere addition, and endoreplication from the MIC chromosomes during the sexual process of conjugation. MACs divide by amitosis, a process in which chromosomes assort randomly without condensing or attaching to a mitotic spindle. Because amitosis routinely assorts previously fragmented chromosomes without deleterious consequences, it is not clear whether MACs require mechanisms to efficiently repair DSBs.

The currently held models for meiotic recombination assume that a DSB is an essential recombinogenic substrate in DNA (42, 57, 76), and γ-H2A.X is associated with meiotic DSBs (47) which precede synapsis in mouse germ cell development and which are thought to initiate meiotic recombination, as in yeast (65, 95). During *Tetrahymena* conjugation, micronuclei undergo meiosis, adopting a highly elongate crescent shape (60). During crescent formation the round MIC elongates gradually to the crescent form, which, when maximally extended, can be up to twice as long as the cell, and then shortens and condenses at metaphase I. Because it precedes the meiotic divisions, the crescent stage is thought to be analogous to most of the prophase of meiosis I. However, while crescents in *Tetrahymena* exhibit some features of meiotic prophase found in other organisms, such as bouquet-like clustering of both telomeres (45) and centromeres (19), synaptonemal complexes (SCs) have not been detected (45, 88), making it difficult to correlate the different stages of micronuclear meiotic prophase to the key events in meiosis such as homologous chromosome pairing and recombination. Thus, different mechanisms could be used in *Tetrahymena*. Also, during conjugation, the parental MAC is destroyed by a process that has been suggested to be related to apoptosis in higher eukaryotes (20, 26), and phosphorylation of H2A.X accompanies formation of the DSBs associated with DNA fragmentation during apoptosis in mammals (46, 54, 68).

Based on the above considerations, we sought to investigate the role of H2A.X phosphorylation in amitosis, in the unusual meiosis, in chromosome rearrangement, and in MAC degeneration in *Tetrahymena.* We also sought to utilize the timing of H2A.X phosphorylation during meiosis to help determine when meiotic recombination occurred. In *T. thermophila* there are four H2As, and only one of the two major H2As has the SQ motif (44; X. Song and M. A. Gorovsky, unpublished data). Using a mammalian γ-H2A.X-specific monoclonal antibody (MAb) that recognizes phosphorylated *T. thermophila* H2A.X (formerly H2A.1), we determined that serine 134 in the SQ motif of *T. thermophila* H2A.X is phosphorylated in both MACs and MICs in response to DSBs induced by chemical agents. We show that *Tetrahymena* γ-H2A.X appears in meiotic MICs at early stage II (when the MICs just start to elongate) (18, 45, 78), indicating that DNA DSBs occur before the MICs acquire Rad51 (45, 72) and during MAC development when chromosome rearrangements occur. Surprisingly, H2A.X is not phosphorylated when parental MACs are being degraded. We also provide evidence that the *HTAX S134A* mutation abolishes SQ motif phosphorylation and causes accumulation of DNA DSBs in both meiotic and mitotic MICs and in amitotic MACs. This mutation makes cells sensitive to chemical agents causing DNA DSBs, causes mitotic delays and DNA loss, and produces meiotic defects, including chromosome loss at metaphase I and lagging chromosomes in anaphase I and II, leading to premature cessation of conjugation. These studies argue that H2A.X SQ motif phosphorylation functions in DSB repair in mitosis, meiosis, and amitosis but not during programmed nuclear death in *Tetrahymena*.

MATERIALS AND METHODS

Strains, culture, and conjugation. Table 1 lists the *T. thermophila* strains used in this study. Strains CU428, CU427, and B2086 were provided by P. J. Bruns (Cornell University). Major histone H2A (H2A.X and H2A.1) germ line double knockout heterokaryon strains G4A1F14A and G4B1G6A and all mutant strains were generated as previously described (62). For studies of vegetative growth, *Tetrahymena* cells were grown in super proteose peptone (SPP) medium (31) containing 1% proteose peptone (1× SPP). For conjugation, two strains of different mating types were washed, starved (15 to 24 h, without shaking at 30°C), and mated in 10 mM Tris-HCl (pH 7.5) as previously described (3). Major H2A genes (*HTAX* and *HTA1*) germ line double knockout heterokaryons and site-directed mutagenesis were generated and performed as described (62).

Transformation and gene replacement. Constructs containing the wild-type (WT) or mutated *HTAX* gene were digested with XhoI and BamH and transformed into 24-h conjugating *HTA* double knockout heterokaryons (for germ line rescue) or 15- to 17-h conjugating CU428 and B2086 cells (for somatic

TABLE 1. Strains used in this study

Strain	Genotype (micronuclei)	Phenotype (macronuclei)
CU428	*CHX1/CHX1 mpr1-1/mpr1-1 HTAX/HTAX HTA1/HTA1*	WT; Cy^{s} Mp^{s} Pm^{s}; VII
CU427	*chx1-1/chx1-1 MPR1/MPR1 HTAX/HTAX HTA1/HTA1*	WT; Cy^{s} Mp^{s} Pm^{s}; VI
B2086	*CHX1/CHX1 MPR1/MPR1 HTAX/HTAX HTA1/HTA1*	WT; Cy^{s} Mp^{s} Pm^{s}; II
HTAX S1P+5R	*Δhtax/Δhtax Δhta1/Δhta1*[a] *CHX1/CHX1 mpr1?/mpr1?*[b]	H2A.X PRRRRR, ΔH2A.1; Pm^{r} Cy^{s} Mp?[cd]
HTAX S1P+5R+(AAAAS)$_{C}$	*Δhtax/Δhtax Δhta1/Δhta1*[a] *CHX1/CHX1 mpr1?/mpr1?*	H2A.X PRRRRR+(AAAAS)$_{C}$, ΔH2A.1; Pm^{r} Cy^{s} Mp?[d]
HTAX S134A rescued	*Δhtax/Δhtax Δhta1/Δhta1*[a] *CHX1/CHX1 mpr1?/mpr1?*	H2A.X S134A, ΔH2A.1; Pm^{r} Cy^{s} Mp?[d]
Rejuvenated *HTAX S134A* rescued	*chx1-1/chx1-1 MPR1/MPR1 HTAX/HTAX HTA1/HTA1*	H2A.X S134A, ΔH2A.1; Pm^{r} Cy^{s} Mp?[d]
HTAX rescued	*Δhtax/Δhtax Δhta1/Δhta1*[a] *CHX1/CHX1 mpr1?/mpr1?*	H2A.X, ΔH2A.1; Pm^{r} Cy^{s} Mp?[d]
Somatic *HTAX S134A*	*CHX1/CHX1 mpr1-1/mpr1-1 HTAX/HTAX HTA1/HTA1* or *CHX1/CHX1 MPR1/MPR1 HTAX/HTAX HTA1/HTA1*	H2A.X S134A, H2A.1; Pm^{r} Cy^{s} Mp^{s}; VII or II
Somatic *HTAX*	*CHX1/CHX1 mpr1-1/mpr1-1 HTAX/HTAX HTA1/HTA1* or *CHX1/CHX1 MPR1/MPR1 HTAX/HTAX HTA1/HTA1*	H2A.X, H2A.1; Pm^{r} Cy^{s} Mp^{s}; VII or II
HTAX-neo3 somatic replacing *S134A* in *S134A* rescued	*Δhtax/Δhtax Δhta1/Δhta1*[a] *CHX1/CHX1 mpr1?/mpr1?*	H2A.X, ΔH2A.1; Pm^{r} Cy^{s} Mp?[d]

[a] The full genetic nomenclature for *Δhtax*/*Δhtax Δhta1*/*Δhta1* is htax-1::*neo2*/*htax-1*::*neo2 hta1-1*::*neo2*/*hta1-1*::*neo2* (2); the abbreviation has been used to conserve space.

[b] Question mark indicates undetermined genotype.

[c] Mp?, 6-methyl-purine sensitivity not determined.

[d] Mating type not determined.

transformation) as previously described (10). Germ line rescued progeny or somatic transformants were initially selected with paramomycin sulfate (Sigma) at 60 μg/ml and serially transferred every 2 to 3 days to fresh medium with increasing concentrations of paromomycin. The genotypes of all transformed cells were confirmed by sequencing the PCR products using *HTAX*-specific primers from genomic DNA of the transformants.

To replace the mutated *HTAX S134A* gene in the MACs of the S134A rescued strain, the *HTAX* gene with a selectable marker inserted in the 5′ flanking region was somatically transformed into the paromomycin-sensitive (Pm^{s}) S134A rescued cells.

Rejuvenation of the double H2A knockout MICS in the S134A rescued strain with a WT MIC through round I genomic exclusion. Round I genomic exclusion (1, 8, 21) is a special type of abortive mating between WT and star (*) strains which have defective, hypodiploid MICs. Star strains can form pairs with WT but are not able to produce pronuclei during nuclear exchange and fertilization stages of conjugation (18, 50, 78). As a result, both partners of the pairs have only haploid MICs received from the WT cell, which are then endoreplicated to form homozygous diploid MICs in both cells. After this step, conjugation is aborted, and the pairs separate as two round I exconjugants, with each cell retaining its original MAC but obtaining a new homozygous MIC, whose genotype depends on which meiotic product was provided from the normal parent. Cells with defective (star) MICs that have received a WT MIC produced by this process are often referred to as rejuvenated because they obtain a new MIC that should be competent for conjugation. Note also that, because these cells do not form a new MAC, they retain their original mating type and, unlike cells that complete normal conjugation, which are immature for ~65 fissions (69), can mate immediately.

S134A or WT rescued cells were mated with CU427 cells, and single pairs were picked into drops of 1× SPP medium at 5 h postmixing. Individual, separated round I exconjugants were picked from each drop into fresh drops of 1× SPP medium at about 11 h postmixing. After cells grew up in the drops, they were transferred to 1× SPP medium in 96-well plates and tested for sensitivity at 120 μg/ml paromomycin. Pm^{r} round I exconjugants from the S134A rescued cells were starved and mated with CU428. The progeny were then tested for cycloheximide resistance (Cy^{r}) as well as paromomycin sensitivity. The cells whose progeny are Cy^{r} Pm^{s} are the rejuvenated cells, which have the WT *HTAX* and *HTA1* genes (instead of the disrupted versions of those genes) in their MICs but retain the *H2AX* S134A mutation or *H2AX* WT copy in their MACs.

Short-circuit genomic exclusion. Short-circuit genomic exclusion (7) occurs in a small fraction of cells during the same type of matings described above for round I genomic exclusion when, instead of simply aborting conjugation, a small percentage of cells form a new MAC. The genotype of this new MIC (and the phenotype of the cell) is determined by the genetic makeup of the cell with the functional MIC.

Indirect immunofluorescence microscopy. γ-H2A.X was immunostained with anti-phospho H2A.X MAb (Upstate), a monoclonal antibody raised against a phosphorylated peptide corresponding to residues 134 to 142 (KATQA[pS]QE Y) of human H2A.X. Growing or mating cells were fixed as previously described (87) with some modifications. Briefly, 5 μl of partial Schaudin's fixative (two parts saturated HgCl$_2$ to one part 100% ethanol) was added directly to 1.5 ml of cells (2 × 10^{5} cells/ml of growing cells or the same cell density of mating cells in 10 mM Tris, pH 7.5), hand mixed, and incubated for 5 min at room temperature (RT). Cells were gently pelleted (130 × *g* for 30 sec), resuspended in 3 ml of RT methanol, repelleted, and resuspended in 1 ml of RT methanol. A total of 30 to 50 μl of cells was spread onto a coverslip and air dried for 30 min. Cells were stained with anti-γH2A.X (1:100) followed by incubation with AlexaFluor 568 goat anti-mouse immunoglobulin G (IgG; 1:500) (Invitrogen). Nuclei were stained with the DNA-specific dye 4′,6′-diamidino-2-phenylindole (DAPI; Roche) at 10 ng/ml for 10 min. Images were obtained with an Olympus BH-2 fluorescence microscope equipped with filters specific for AlexaFluor and DAPI using a 100× or 40× lens and the Scion VisiCapture and Scion TWAIN 1394 camera import programs (Scion Corporation). Adobe Photoshop software was used for coloring images.

Nucleus isolation, histone extraction, and phosphatase treatment. Log-phase (2 × 10^{5} cells/ml) or 3.5-h conjugating cells were used to isolate MACs and/or MICs as described previously (31). Histones were extracted from MACs with 0.4 N H$_2$SO$_4$ (4) and precipitated with 20% trichloroacetic acid. Aliquots of 25 μg of histones were treated with λ protein phosphatase (New England Biolabs, Inc.) at 10 U/μl for 5 h at 30°C and precipitated with 20% trichloroacetic acid.

AU-PAGE. Acid-urea polyacrylamide gel electrophoresis (AU-PAGE) was performed as described previously (5, 62).

Immunoblotting. Histones from mutated and WT *HTAX* rescued strains, with or without pretreatment with λ protein phosphatase, were separated on long AU-PAGE. Macronuclear and micronuclear extracts from WT cells from log phase and 3.5-h conjugation were separated on 12% sodium dodecyl sulfate-PAGE. Separated histones or nuclear extracts were transferred to Immobilon-P membranes. After being blocked in 5% nonfat milk, the blot was incubated with anti-H2A (1:5,000) or anti-γ-H2AX (1:1,000, Upstate) overnight at 4°C. A 1:10,000 dilution of horseradish peroxidase-conjugated goat anti-rabbit IgG (Jackson ImmunoResearch) or goat anti-mouse IgG-IgA-IgM (Zymed Labs Inc.) was used as secondary antibody. Blots were developed using an ECL Western blotting detection kit (Perkin-Elmer) according to the manufacturer's instructions.

DAPI stain for photomicroscopy. One microliter of 0.1 mg/ml DAPI was added to 1 ml of log-phase cells fixed with 50 μl of formaldehyde (37% solution; J. T. Baker). Cells were stained for 5 min. Images were obtained using an Olympus BH-2 fluorescence microscope with the Scion VisiCapture and Scion TWAIN 1394 camera import programs (Scion Corporation).

Comet assay. A neutral comet assay was performed as previously described (75, 92) with minor modifications. Briefly, 40 μl of 0.3% low-melting-point agarose (Sigma) in phosphate-buffered saline was coated onto the frosted area (18 by 18 mm) of a double-frosted slide (Fisher) and air dried for several days before being used. Five microliters of 2 ×10^{5} to 5 ×10^{5} cells/ml was mixed with 35 μl of 1% low-melting-point agarose in phosphate-buffered saline and spread onto the precoated area, solidified by being immediately put onto a metal tray on ice for 3 min, to form microgels. From that point, all steps were performed in dimmed light or in the dark to reduce DNA damage, and slides were set horizontally during lysis and the following treatments. Microgels were lysed with freshly made, cold (at 4°C for 0.5 h) lysis buffer (2.5 M NaCl, 100 mM EDTA, 10 mM Tris, pH 10) with 1% Triton for 2 h at 4°C. They were then treated with 10

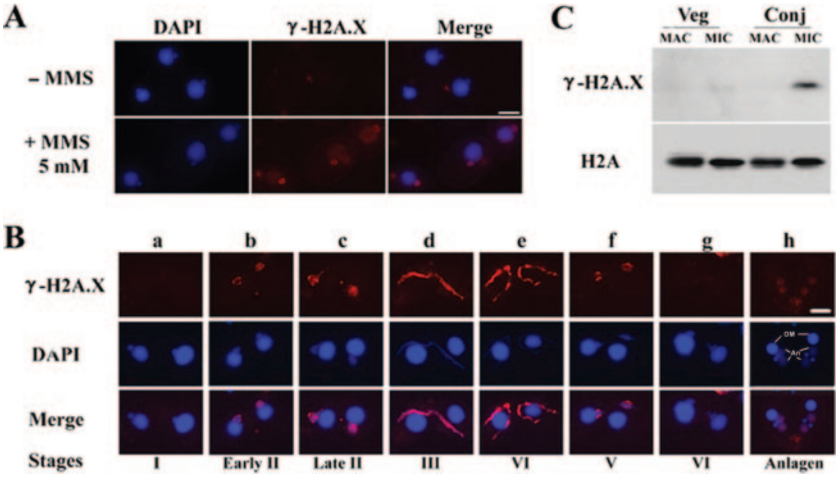


FIG. 1. Human anti-γ-H2A.X antibody detects chemical induced and meiotic DSBs in *T. thermophila*. (A) IF analysis of WT *Tetrahymena* cells with a specific MAb raised against the phosphorylated C-terminal region of human H2A.X (γ-H2A.X) shows that MMS treatment induces macro- and micronuclear staining. Scale bar, 10 μm. (B) IF analysis of different stages during conjugation in *Tetrahymena*. DAPI stain shows the nuclei, and the anti-γ-H2A.X MAb staining indicates that DSBs, indicative of meiotic recombination, appear as the MICs begin to elongate in early stage II (b) and disappear at stage VI (diakinasis/metaphase, g). γ-H2A.X also appears in the newly developing MACs (anlagen) undergoing DNA fragmentation and rearrangement (An, h;), but not in the parental MAC (OM, h) undergoing programmed nuclear death. Scale bar, 10 μm. (C) Western blot of a sodium dodecyl sulfate-PAGE gel of macro- and micronuclear extracts, probed with anti-γ-H2A.X or anti-H2A antibodies. Only MICs from early conjugating cells contain a significant amount of γ-H2A.X.

μg/ml RNase A in lysis buffer without Triton for 2 h at 37°C, followed by 2 h at 37°C in 1 mg/ml proteinase K in lysis buffer without Triton, and equilibrated in 1 liter of freshly made electrophoresis buffer (300 mM sodium acetate, 100 mM Tris, pH 9.0) in an electrophoresis apparatus for 20 min at RT (slides were put side by side tightly and at one end of the apparatus). Electrophoresis was carried out in the same buffer at 12 V (0.6 V/cm) and 100 mA at RT for 1 h. After electrophoresis, microgels were neutralized with 0.4 M Tris, pH 7.4 (drop-wise added on top of the microgels and drained; the procedure was repeated three times), dehydrated with 100% ethanol, and air dried. DNA in microgels was stained with DAPI (20 ng/ml) in 1× Tris-acetate-EDTA buffer for 5 min, followed by destaining for 5 min in 1× Tris-acetate-EDTA buffer. Images were obtained with an Olympus BH-2 microscope equipped for fluorescence, with the Scion VisiCapture and Scion TWAIN 1394 camera import programs (Scion Corp.). The Image J program was used to measure the tail lengths and total DNA content.

RESULTS

H2A.X in *T. thermophila* is phosphorylated in response to induced DSBs. To determine whether phosphorylation of H2A.X in *Tetrahymena* occurs in response to induced DSB formation, we carried out immunofluorescence (IF) analyses using a specific MAb to human γ-H2A.X that reacts across species to examine *Tetrahymena* cells treated with methyl methanesulfonate (MMS), an alkylating agent that causes base alkylations and DNA lesions that are converted to DSBs similar to those produced by ionizing radiation (40, 71, 90, 91). No signal was detected in nuclei of untreated cells (Fig. 1A, −MMS), demonstrating that this antibody does not recognize an epitope in untreated vegetative cells. In contrast, MICs and, to a lesser extent, MACs were stained with the anti-γ-H2A.X MAb after MMS treatment (Fig. 1A, +MMS), indicating that γ-H2A.X is produced in nuclei from vegetative cells when DSBs are induced by exogenous DNA-damaging agents. The difference in γ-H2A.X staining intensity in MICs and MACs from MMS-treated cells could be due to the differential activities of the phosphorylation machinery in the different nuclei or it could reflect the previously observed higher levels of H2A.X (formerly H2A.1) as a fraction of total H2A in MICs than in MACs (5).

H2A.X in *T. thermophila* is phosphorylated during meiosis. As noted above, *Tetrahymena* meiosis is unusual in lacking synaptonemal complexes. In addition, previously observed histone steady-state levels in *Tetrahymena* show enrichment of histone H2A.X in the MICs for reasons that were unclear (5). We reasoned that these differences in relative levels of H2A forms between nuclei could suggest distinct functional roles for different *Tetrahymena* H2As. In particular, higher micronuclear levels of H2A.X might correlate with the specialized function of MICs during conjugation, i.e., its potential to undergo homologous recombination during meiosis. These considerations led us to examine the role of H2A.X phosphorylation during meiosis in *Tetrahymena*.

Based on morphological changes, *Tetrahymena* meiotic prophase has been divided into six stages, with stages II to V being the crescent stages (18, 45, 78). Soon after conjugation starts, the MIC moves away from the macronuclear pocket where it resides during interphase and begins to elongate into a teardrop shape (early stage II), then a spindle-shape (late stage II) (45), and then a head-neck-trunk crescent (stage III), followed by bidirectional elongation to become fully elongated in stage IV, where the five bivalent chromosomes are in parallel arrangement, with telomeres at one end of the crescent (45) and the centromeres at the other end (19). MICs stain with the anti-γ-H2A.X MAb beginning at early stage II, when they have just started to elongate (Fig. 1B, frame b). The signal continues through stage V (Fig. 1B, frames b to f) and disappears or is only barely visible at stage VI (diakinesis/metaphase I) (Fig. 1B, frame g). Immunoblotting analysis of extracts from purified MACs and MICs isolated from either vegetative cells or early conjugation also confirmed the IF analyses showing

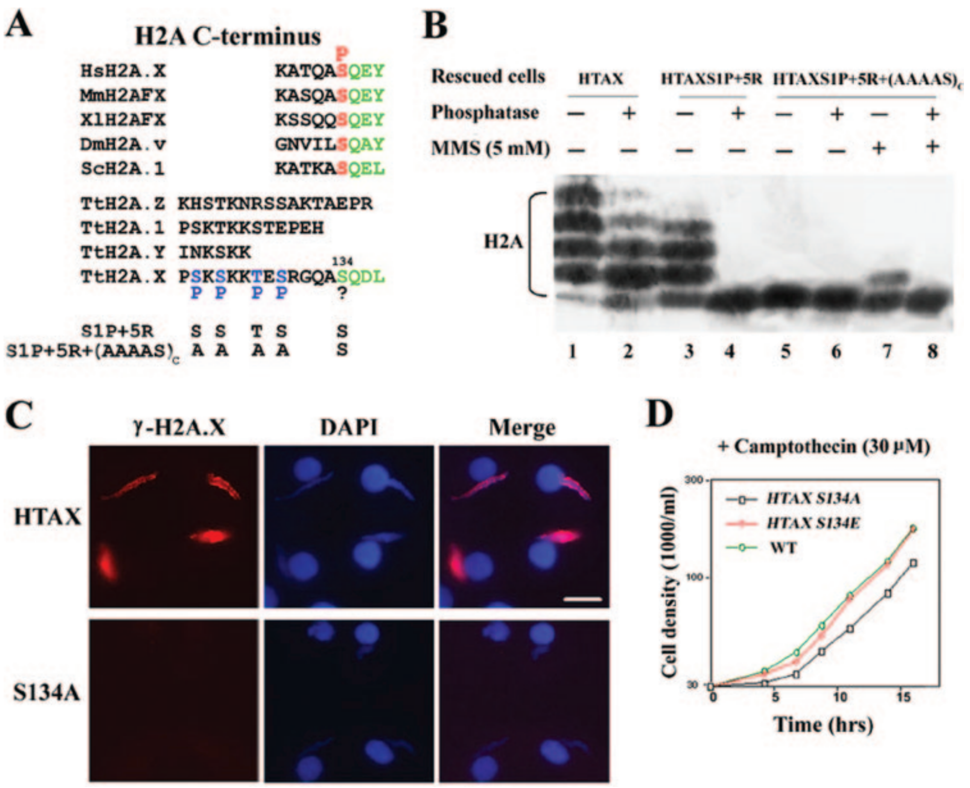


FIG. 2. H2A.X S134 residue in the SQ motif is responsible for the γ-H2A.X signal induced by DSBs. (A) Alignment of the C-terminal tails of H2A.Xs in different organisms and the four H2As in *T. thermophila* using Clustal X. The sequences were obtained from GenBank. The accession numbers are as follows: HsH2A.X, NP_002096; MmH2AFX, NP_034566; XlH2AFX, Q6GM86; DmH2A.v, NP_524519; ScH2A.1, CAA24611; TtH2A.Z (hv1), CAA33554; TtH2A.1 (previously H2A.2), AAC37292; TtH2A.Y, AAU87547; TtH2A.X (previously H2A.1), AAC37291. The serine in the SQ motif that is known to be phosphorylated upon γ irradiation is labeled in red. The conserved SQ motif is labeled in green. TtH2A.X is the only H2A in *T. thermophila* that has an SQ motif, and it has four serine/threonine residues upstream of the SQ motif (labeled in blue) which are the sites for normal phosphorylation of the protein in vegetative growth (see panel B). (B) Western blot of an AU-PAGE gel separating nuclear histones from WT or mutated *HTAX* rescues of HTA double knockout heterokaryons, stained by anti-H2A polyclonal antibody. Lanes 1 and 2 show that WT H2A.X is phosphorylated but that it is impossible to determine the phosphorylation status in detail due to the presence of acetylation on the protein. Lanes 3 and 4 show that there are three phosphatase-sensitive isoforms of the H2A.X in the N-terminal mutation strain (HTAX S1P+5R) that abolished the acetylation sites. Lanes 5 and 6 show that changing the four serine/threonine residues upstream of the SQ motif to alanines eliminates the three phosphatase-sensitive isoforms. Lanes 7 and 8 show that an additional phosphatase-sensitive isoform is detected upon MMS treatment in the mutant *Tetrahymena* strain [HTAX S1P+5R+(AAAAS)$_C$] lacking all other known acetylation and phosphorylation sites. (C) IF analysis of the HTAX or S134A rescued cells during conjugation (4 h). γ-H2A.X staining is not detectable in S134A rescued cells. Scale bar, 10 μm. (D) Growth curve of WT, S134A, and S134E rescued cells in 1× SPP medium with 30 μM camptothecin.

that only MICs from early conjugation stages corresponding to meiotic prophase I contain substantial amount of γ-H2A.X (Fig. 1C).

γ-H2A.X is detectable in developing MACs undergoing DNA rearrangement but not in MACs undergoing programmed nuclear death. H2A.X phosphorylation accompanies V(D)J rejoining (15), and, in mouse cells, the formation of DNA ladders that accompanies apoptosis requires H2A.X phosphorylation (46). During conjugation in *Tetrahymena*, DNA breakage is known to occur at two stages: when chromosome fragmentation and DNA elimination are occurring in developing MACs (94) and when parental MACs undergo programmed nuclear death, which has been viewed as an apoptotic-like process during which DNA is degraded to produce oligonucleosome-sized DNA ladders (20). No staining of γ-H2A.X in the parental MAC was detected at any stage of conjugation, including late stages when the parental MAC was undergoing programmed nuclear death (Fig. 1B, frame h, OM). In contrast, γ-H2A.X staining was easily detectable in late-stage developing MACs when DNA rearrangement events were occurring (Fig. 1B, frame h, An).

H2A.X phosphorylation occurs on S134. To determine if phosphorylation at serine 134 in the SQ motif in *Tetrahymena* H2A.X (Fig. 2A) is associated with anti-γ-H2A.X staining, we performed site-directed mutagenesis together with gene replacement. When a WT *HTAX* gene was used to rescue *HTAX* and *HTA1* (formerly *HTA2*) germ line double knockout heterokaryons (32, 62), the histones isolated from viable progeny (*HTAX* rescued cells) exhibited phosphatase-resistant H2A isoforms when analyzed on acid-urea acrylamide gels (Fig. 2B, lanes 1 and 2), due to charge-altering acetylations in the N-terminal tail (62). When histones isolated from vegetatively growing cells that were rescued with a gene encoding H2A.X that was mutated to eliminate all acetylation sites were analyzed, the viable progeny (HTAX S1P+5R rescued) exhibited three phosphatase-sensitive isoforms in addition to unmodified

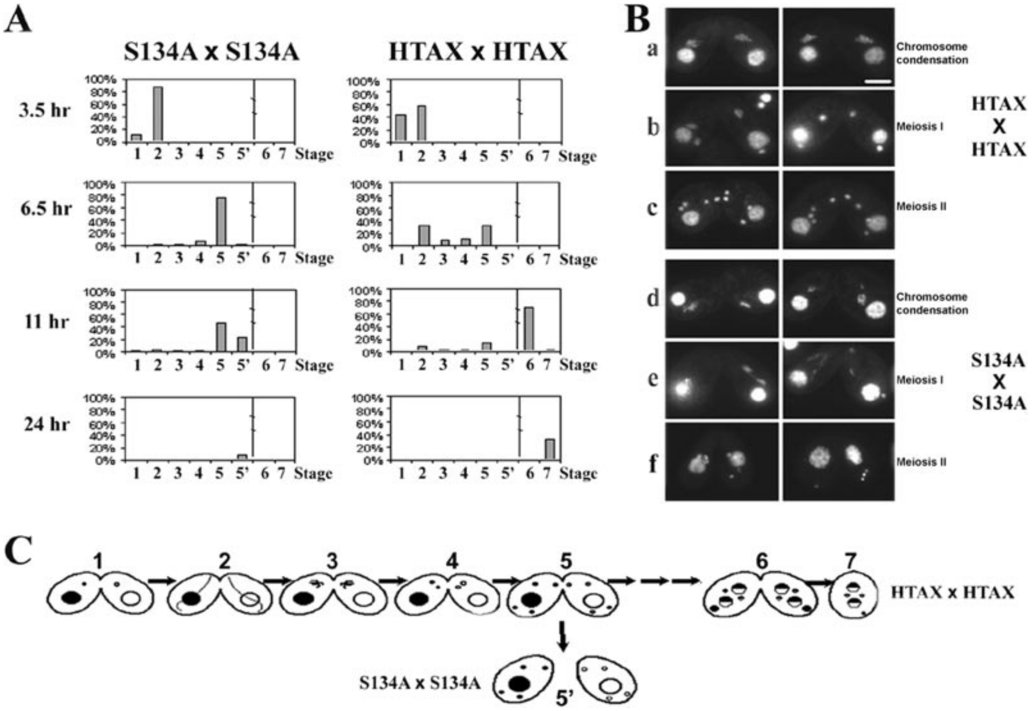


FIG. 3. Absence of S134 phosphorylation leads to meiosis defects and premature termination of conjugation. (A) Conjugation profiles of matings between two S134A rescued strains or two HTAX rescued strains. Different stages during conjugation (indicated in panel C) were scored in samples removed at different times. The vertical line between stage 5′ (see panel C) and 6 indicates that there are several noninformative stages in normal conjugation that were not precisely staged in this study but were counted in the total number of cells ($n \geq 346$) analyzed. (B) DAPI staining of the nuclei from the HTAX or S134A rescued strains at different stages of conjugation times as indicated. Scale bar, 10 μm. (C) Different stages of conjugation scored in panel A. S134A rescued matings prematurely terminated mating after meiosis II (stage 5), which was not seen in HTAX rescued matings and was denoted as stage 5′. There are several stages, indicated by several horizontal arrows, between stage 5 and 6 that were not plotted. Black and white nuclei indicate two mating types of the same genetic background cells.

H2A.X (Fig. 2B, lanes 3 and 4). Upon mutating four S/T residues (S122, S124, T127, and S129) in the H2A.X C terminus (Fig. 2A) to alanines [HTAX S1P+5R+(AAAAS)c rescued] (63), the three phosphorylated isoforms disappeared (Fig. 2B, lanes 5 and 6). These cells produce viable progeny and grow normally. These results indicate that three of the four mutated sites (S122, S124, T127, and S129) in H2A.X are phosphorylated in untreated vegetatively growing cells, yet they are dispensable (63).

HTAX S1P+5R+(AAAAS)c rescued cells, lacking any acetylation sites and in which the only remaining phosphorylatable serine is S134, were then treated with MMS to induce DSBs. A single, phosphatase-sensitive isoform (Fig. 2B, lanes 7 and 8) (63) was detected. These observations suggest that S134 in the SQ motif is responsible for this phosphorylation event in response to DSBs. To test this, we made an HTAX S134A rescued strain, in which an *HTAX* gene bearing a mutation of S to A at residue 134 was used to rescue major *HTA* double knockout heterokaryons (32, 62). The viable progeny will be referred to as S134A rescued cells; they have a MIC with both the *HTAX* and *HTA1* genes knocked out and a MAC containing the double knockout genes plus a mutated *HTAX S134A* gene. To examine whether the S134A mutation abolished DSB-induced phosphorylation, S134A rescued cells of different mating types were mated, fixed, and stained with the anti-γ-H2A.X MAb. No γ-H2A.X signal was detected in crescent MICs of the S134A rescued cells (Fig. 2C). Control HTAX rescued cells, in which a WT *HTAX* gene was used to rescue major *HTA* double knockout heterokaryons, stained like WT cells (Fig. 2C). Thus, the S134A mutation in H2A.X can specifically abolish γ-H2A.X staining in meiotic prophase crescent MICs, confirming that S134 in the H2A.X SQ motif is the site for phosphorylation in response to meiotic DSBs. This S134A mutation also was more sensitive to the DSB-producing chemicals camptothecin (Fig. 2D) (63) and MMS (data not shown) than WT cells or than an S134E mutant which is otherwise isogenic with S134A mutant cells, arguing that phosphorylation on S134 is important for the signaling and/or repair of DNA damage in *Tetrahymena*.

Absence of S134 phosphorylation leads to meiotic defects. To investigate the function(s) of SQ phosphorylation in meiosis in *Tetrahymena*, we examined whether the absence of S134 phosphorylation during homologous recombination in prophase of meiosis I leads to any conjugation defects. We monitored the conjugation process between two different mating types of HTAX or S134A rescued strains and scored the percentages of different stages at various time points after cells were mixed. Aliquots of the mating cells also were fixed and stained with the DNA-specific dye DAPI to examine nuclear morphology. HTAX rescued cells could go through conjugation to pair separation and formed 34% exconjugants after 24 h postmixing (Fig. 3A and C, stage 7). This result showed

TABLE 2. *HTAX-neo3* somatic replacing *HTAX S134A* mutant could not rescue its phenotype during conjugation[a]

Mating	Percentage of exconjugants by conjugation type					Pm^r/Cy^r (%)
	6-h pairing	10-h pairing	10-h anlagen	10-h exconjugants	29.5-h exconjugants	
CU427 × CU428	81.5	70.0	100.0	15.4	68.4	0
CU427 × S134A	52.0	23.0	3.0	0	6.7	16
CU427 × rejuvenated S134A (50-3)	23.2	18.2	9.5	0	5.9	16
CU427 × rejuvenated S134A (62-5)	48.6	10.6	5.9	0	2.5	10.4

[a] To determine whether the S134A mutation only causes the meiosis defect, the mutant *HTAX S134A* gene was replaced with the WT *HTAX-neo3* gene in MACs of the S134A rescued cells (Fig. 5A). These cells now contain WT *HTAX* genes in MACs but retain MICs that could have kept any DSBs accumulated earlier. Two such strains were made (50-3 and 62-5). The conjugation results showed that the strains that replaced the S134A mutation in MACs of S134A rescued cells did not reduce the meiotic defects, as indicated by the low percentage of exconjugants and low Pm^r/Cy^r ratio. Experiments were repeated twice with similar results.

that, although they do so less efficiently than WT cells, the H2A double knockout heterokaryon cells are able to complete conjugation with a transformed WT *HTAX* gene only in the MACs. Matings between S134A rescued cells appeared cytologically normal during prophase of meiosis I (data not shown), but at metaphase I, there is a decrease in micronuclear DAPI staining compared to HTAX rescued mating cells, and DAPI-stained fragments were seen near the condensed chromosomes (Fig. 3B, compare frames d and a), suggesting that there was (partial or whole) chromosome loss in S134A rescued cells. During anaphase I, lagging chromosomes were often observed in S134A mating cells (Fig. 3B, frames e). Although they could still enter meiosis II, the four meiotic MICs of S134A rescued cells were much smaller than those of HTAX rescued cells (Fig. 3B, compare frames f and c), and some of the meiotic MICs were missing. Conjugation of the S134A rescued cells was then aborted, and the two partners separated as single cells with four or fewer defective meiotic MICs (Fig. 3A and C, stage 5′). Occasionally, pairs separated after meiosis I, leaving single cells with MICs undergoing meiosis II (data not shown).

When S134A rescued cells were mated with WT CU427 cells, the WT cells could not fully rescue the *HTAX S134A* mutation. Since the S134A rescued strain has a homozygous double *HTA* knockout (both major H2A genes replaced by *neo2* cassettes) in its MIC (62) and since the CU427 MIC is homozygous for the *chx1-1* gene, which confers dominant Cy^r when these two cells mate, true progeny that have completed conjugation should be both Cy^r (since they received one of the *chx1-1* genes from CU427) and Pm^r (since they also obtained the *htax::neo2* and *hta1::neo2* genes, which confer Pm^r, from the S134A rescued parent). Note that progeny cells also can be obtained at low frequency from a process known as short-circuit genomic exclusion (7) which occurs when cells containing a functional MIC are mated to cells with defective MICs (see Materials and Methods for details). In the studies performed here, these cells will be Cy^r. Only a few exconjugants were observed after 30 h postmixing, and the Pm^r/Cy^r ratio was far below 100% (Table 2). Most of the pairs contained four MICs, indicating that they had stopped conjugation after meiosis was completed. In the following experiments, a low percentage of exconjugants and a low Pm^r/Cy^r ratio when cells were mated to CU427 cells were used as indicators to identify the S134A mutant phenotype in conjugation.

Because MICs develop into new MACs and MICs during conjugation and, in the rescued H2A double knockout strains, the MICs contain only knockout copies of the major H2A genes, there was a concern whether the conjugation defects observed for S134A rescued cells were caused by the absence of major H2A genes in the conjugating knockout heterokaryon cells before the mutated *HTAX S134A* gene was transformed into the cells. To clarify this issue, a WT MIC was reintroduced into the S134A rescued strain to create a "rejuvenated" strain. This is possible in *Tetrahymena* through round I genomic exclusion (1, 8, 21), a special type of abortive mating between WT and star (*) strains which have defective, hypodiploid MICs (see Materials and Methods for details). Since the S134A rescued cells stopped mating at meiosis II and did not produce pronuclei, they behaved like star cells, suggesting that they had defective MICs. S134A rescued cells were mated with WT CU427 cells, and two mating types of rejuvenated S134A rescued cells were obtained through round I genomic exclusion (Fig. 4A). These rejuvenated S134A rescued cells have WT *HTAX* and *HTA1* genes (instead of the double knockouts) in their MICs but retain the mutated *HTAX S134A* genes in their MACs. Rejuvenated S134A rescued cells of different mating types were then mated with each other or mated with CU427, and the conjugation process was monitored. Matings with rejuvenated S134A rescued cells showed the same conjugation phenotype as the original S134A rescued cells (data not shown), indicating that the phenotypes observed are due to the expression of the *HTAX S134A* gene in the parental MACs, not to defects in the MICs or the newly formed zygotic MAC.

As an alternative approach to analyze the phenotype produced by *HTAX S134A* in an otherwise normal background, the macronuclear *HTAX* genes were somatically replaced with *HTAX S134A-neo2* or *HTAX-neo2* (the selectable marker *neo2* cassette was inserted in *HTAX* 3′ flanking region) genes in WT cells of different mating types (Fig. 4B). We found that these somatic *HTAX S134A* mutants had the same conjugation phenotype as the S134A rescued cells (data not shown). These experiments demonstrate that expression of the H2A.X S134A mutation in MACs of conjugating cells produces meiotic defects and premature termination of conjugation.

Absence of S134 phosphorylation causes mitotic defects. To determine if the S134A mutant phenotype described above is the result of events that occur during meiosis or is the result of accumulated defects in the MICs during vegetative growth in

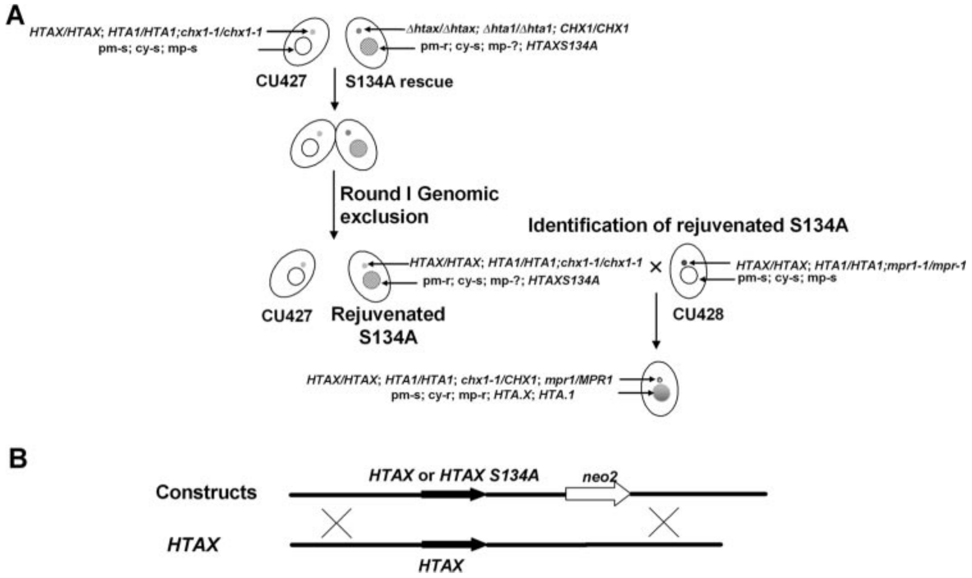

A
HTAX/HTAX; HTA1/HTA1;chx1-1/chx1-1
pm-s; cy-s; mp-s
Δhtax/Δhtax; Δhta1/Δhta1; CHX1/CHX1
pm-r; cy-s; mp-?; HTAXS134A
CU427
S134A rescue
Round I Genomic exclusion
Identification of rejuvenated S134A
HTAX/HTAX; HTA1/HTA1;chx1-1/chx1-1
pm-r; cy-s; mp-?; HTAXS134A
×
HTAX/HTAX; HTA1/HTA1;mpr1-1/mpr-1
pm-s; cy-s; mp-s
CU427
Rejuvenated S134A
CU428
HTAX/HTAX; HTA1/HTA1; chx1-1/CHX1; mpr1/MPR1
pm-s; cy-r; mp-r; HTA.X; HTA.1
B
Constructs
HTAX or HTAX S134A
neo2
HTAX
HTAX

FIG. 4. Strategy to determine if *HTAX S134A* mutation phenotype is truly meiotic. (A) Genetic manipulations used to replace the *HTAX* and *HTA1* double knockout MIC in the S134A rescued strain with a WT MIC to generate the rejuvenated S134A strain. This was used to determine if the phenotype of the *HTAX S134A* mutation was truly meiotic (see text for description). (B) Diagram of somatic replacement of the *HTAX* gene in MACs of WT cells of different mating types with the *HTAX-neo2* or *HTAX S134A-neo2*.

the absence of phosphorylatable H2A.X, the mutated *HTAX S134A* genes in S134A rescued MACs were replaced with WT *HTAX* genes (Fig. 5A). If the S134A mutation causes only meiosis-specific defects in the MIC of a conjugating cell and if these cells now go through conjugation with a WT MAC, the conjugation phenotype should be rescued. On the other hand, if the S134A mutation also causes irreversible damage to mitotic MICs during vegetative growth, this damage should accumulate before the cells are somatically rescued by replacing the *HTAX S134A* gene with a WT *HTAX* gene. Therefore, although the reintroduced *HTAX* completely replaced the *HTAX S134A* genes (data not shown), these cells would retain damaged MICs that might not be able to complete conjugation. Table 2 shows that the S134A mutation phenotype in conjugation (determined as a low percentage of exconjugants and low Pm^r/Cy^r ratio when cells were mated with CU427 cells) is still present in S134A rescued cells in which the mutated genes in MACs had been replaced with WT *HTAX* genes, arguing that, in addition to affecting meiosis, the *HTAX S134A* mutation causes micronuclear defects during vegetative

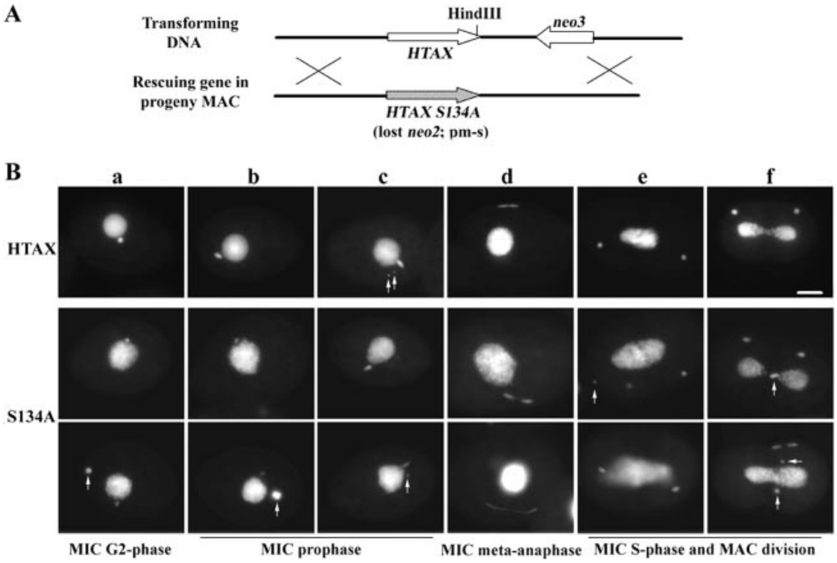


FIG. 5. Eliminating H2A.X SQ motif phosphorylation causes micronuclear DNA loss and abnormal mitosis in *T. thermophila*. (A) Strategy for replacing the mutant *HTAX S134A* gene with WT *HTAX-neo3* gene in MACs of S134A rescued cells. (B) Vegetatively growing HTAX rescued cells (top row) and S134A rescued cells (lower rows) were fixed and stained with DAPI. Shown are MICs and MACs at different stages of the vegetative cell cycle as indicated. Arrows indicate CEBs (see text for description). Scale bar, 10 μm.

growth. Therefore, H2A.X phosphorylation on the SQ motif is likely required for both normal meiosis and mitosis.

Because cells must be grown vegetatively before they can be conjugated, we cannot rule out the possibility that the meiotic defects we observed in both the rejuvenated S134A cells or somatic S134A cells described above were due to the accumulation of mitotic defects during the period of vegetative growth before the cells were conjugated. However, the behavior of other mitotic defect mutants or knockouts suggests that the mitotic defects of S134A cannot account for all of the meiotic defects. (i) *HTAY* conditional knockout cells in nonpermissive conditions have mitotic defects (X. Song and M. A. Gorovsky, submitted for publication) but can mate with WT cells and finish conjugation, producing 50% exconjugants (Song and Gorovsky, unpublished observations), while in the mating between S134A and WT CU427, only a low percentage (<7%) of exconjugants were produced and most of the paired cells stopped mating after meiosis II. (ii) *DCL1* knockout cells, which have severe defects in mitotic chromosome segregation, could finish mating at a low percentage (52), while in matings between two different mating types of S134A rescued cells stopped mating after meiosis II. For these reasons, we believe that, besides the mitotic defects, S134A has meiotic defects.

To confirm that the *HTAX S134A* mutation causes micronuclear damage during mitosis in vegetative growth, we examined micronuclear morphology by DAPI staining of log-phase S134A or HTAX rescued cells. MICs in more than 70% of the S134A rescued cells (Fig. 5B, lower rows) were smaller, more irregular, and less strongly stained than the bright and distinct MICs of HTAX rescued cells in both mitotically dividing and nondividing stages (Fig. 5B, top row), suggesting there was DNA loss during vegetative growth in S134A rescued cells. In HTAX rescued cells, as in WT cells (30), when macronuclear division and cytokinesis start, the MICs have already finished their mitotic division and are visualized as two round dots near the ends of the cells, each of which will enter one of the daughter cells (Fig. 5B, top row, frames e and f). In mitotic S134A rescued cells, lagging chromosomes (Fig. 5B, lower rows) were often observed along with delayed mitotic divisions in which MICs were still in different stages of mitosis when the MACs and cells were dividing (Fig. 5B, lower rows, frames e and f).

MACs in S134A cells also showed defects. There are more cells with chromatin exclusion bodies (CEBs) in the S134A rescued strain (Fig. 5B, lower rows, arrows in frames a, b, c, e and f) than in HTAX rescued cells. DNA loss by elimination of CEBs occurs normally in *Tetrahymena* and is thought to be a mechanism to maintain the level of macronuclear ploidy (6). Thus, H2A.X phosphorylation is required for normal micronuclear mitosis and probably also for normal macronuclear division.

***The HTAX S134A* mutation in the MAC causes DNA damage accumulation in both MACs and MICs.** Next, we studied whether the DSB repair machinery is affected in S134A rescued cells. It has been reported that foci of Rad51, a recombinase required for HR in both meiosis and mitosis (72), are present in vegetative as well as conjugating cell MACs (45). IF analyses showed that the Rad51 signal in MACs in vegetatively growing S134A or HTAX rescued cells was similar as was the appearance of Rad51 in meiotic MICs (data not shown). Thus, in *Tetrahymena*, Rad51 accumulation is not dependent on H2A.X phosphorylation. We then tested whether the S134A mutation causes less efficient repair and leads to accumulation of DNA damage. Since γ-H2AX is one of the earliest cellular responses to DNA DSBs, we first checked γ-H2AX expression in S134A rescued cells. To overcome the fact that, in S134A rescued cells, the mutated H2A.X S134A cannot be phosphorylated on S134 and thus cannot be stained by the anti-γ-H2AX MAb, we utilized the well-established phenomenon of conjugation-mediated transfer of protein and/or mRNA between two mating cells in *Tetrahymena* (51). A WT cell was mated with an S134A rescued cell whose nuclei, lacking WT H2A.X, cannot be stained with the anti-γ-H2A.X MAb. When the mutant cell receives WT H2A.X (or *HTAX* mRNA) from the WT cell by conjugation-mediated transfer (Fig. 6A), both cells in the pair should be stained with the anti-γ-H2AX if H2A.X is phosphorylated on SQ. The staining observed in the mutant cell then becomes an assay for the presence of DSBs in that cell.

IF staining by anti-γ-H2A.X of matings between WT and S134A or HTAX rescued cells are shown in Fig. 6B. In matings between WT and HTAX rescued cells, γ-H2A.X staining is observed initially in early stage II meiotic prophase cells (Fig. 6B, row c), as in WT mating cells (Fig. 1B). However, in matings between WT and S134A rescued cells, γ-H2A.X signal is detectable in the S134A cell in stage I, even before the MICs elongate (Fig. 6B, row b), indicating that DNA DSBs had accumulated in MICs before conjugation. In the S134A rescued cell, the MAC also shows a strong signal (Fig. 6B, row b) that is not observed in the WT partner. These differences between the cells containing WT H2A.X and the S134A mutant H2A.X distinguish the WT from S134A rescued cells in a pair. When MICs elongate, the γ-H2A.X staining appears in MICs of both WT and S134A rescued cells, and the macronuclear staining in S134A cells persists (Fig. 6B, rows d, f, and h). γ-H2A.X staining in parental MACs of conjugating WT cells is never observed. From the DAPI staining, as observed in vegetative cells, the MICs, and probably also the MACs, of S134A rescued cells contain less DNA than those in WT cells. As seen in matings of two S134A rescued cells (Fig. 3B), meiotic defects in matings between WT and S134A cells were also observed. γ-H2A.X signal was observed in the hypodiploid MICs as well as lagging chromosomes in metaphase I (Fig. 6C, rows a and b), anaphase I (Fig. 6C, row c), and anaphase II (Fig. 6C, rows d and e). These defects were observed only in the S134A cell in the pair but not in the WT partner, indicating that they reflected intrinsic defects in the S134A MICs. These abnormalities were not observed in matings between WT and HTAX rescued cells (data not shown). These results argue that DSBs exist in meiotic MICs before and after the period of HR as well as in MACs of the S134A cells. Although the major function of γ-H2AX appears to be associated with DSBs, it has also been suggested to have other functions (reviewed in reference 59). Note, however, that one commonly cited case where H2A.X phosphorylation appears to function independently of DSBs, its association with the inactive X chromosome, has recently been shown to occur during late replication, where it could also be associated with transient, replication-associated DSBs (14). To confirm that S134A cells accumulated DSBs, we used single-cell, neutral gel electrophoresis

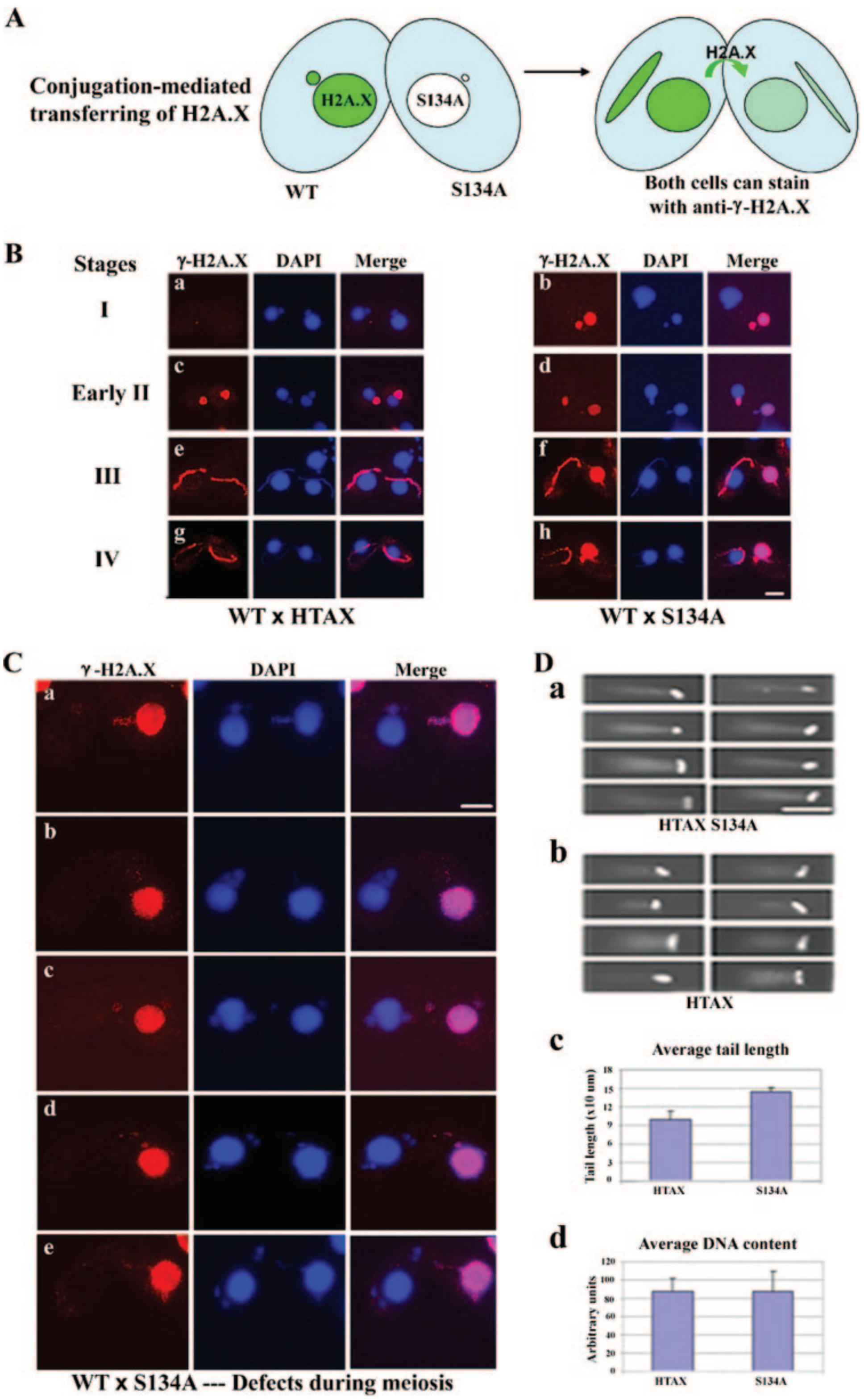


FIG. 6. *HTAX S134A* mutation in MACs causes DNA damage accumulation in both MICs and MACs. (A) Diagram of conjugation-mediated transfer of protein (or mRNA) between the two partners of a pair. (B) IF analysis of conjugation between WT and HTAX rescued cells (rows a, c, e, and g) or WT and S134A rescued cells (rows b, d, f, and h), stained with anti-γ-H2A.X and DAPI. WT cell is on the left in each pair of matings between WT and S134A rescued cells. WT cells and HTAX rescued cells are indistinguishable. (C) IF analysis of conjugation between WT and S134A rescued cells, stained with anti-γ-H2A.X and DAPI, showing the meiotic defects of S134A MICs (see text for details). WT cell is on the left in each pair. (D) Neutral comet assay showing the DAPI-stained nuclei (a and b) after neutral single-cell gel electrophoresis. Graphs show quantification of the average tail lengths (c) and total DNA contents (including tail and chromosomal DNA; d) from about 50 nuclei of S134A or HTAX rescued cells using the Image J program. Scale bars, 10 μm (B and C) and 100 μm (D).

(comet assay) (75). As shown in Fig. 6D, S134A rescued cells exhibited longer comet tails (average tail length, 145 μm; n = 51; standard deviation [SD], 6.7) than HTAX rescued cells (average tail length, 99 μm; n = 50; SD, 13.6) (Fig. 6D, a to c), while they have similar total (mostly macronuclear) DNA contents (Fig. 6D, d), indicating that they contain more DSBs in their macronuclear DNA.

DISCUSSION

We show in this study that one of the major histone H2As in *Tetrahymena* is a typical H2A.X. It can be phosphorylated at the serine 134 residue in its SQ motif in response to DSBs induced by chemical agents and during meiosis. Using a MAb specific to a phosphopeptide (KATQA[pS]QEY) corresponding to residues 134 to 142 of human H2A.X, together with a mutant strain (S134A) that abolished the phosphorylation site, we demonstrate that the SQ motif phosphorylation is important for cells to recover from exogenous DNA damage and to repair breaks associated with normal micronuclear meiosis and mitosis and macronuclear amitosis. The inability to phosphorylate this site leads to meiotic, mitotic, and amitotic defects and accumulation of DSBs in both MICs and MACs of *Tetrahymena* cells.

Although the H2A.X S134A mutation causes visible defects in mitosis of MICs, and appears to affect amitosis of MACs, the mutation is not lethal for vegetative growth. Most previously described mutations that affect *Tetrahymena* MICs are not lethal (52, 86), which is likely due to the transcriptional inertness of the MIC in vegetative growth. Mutations that abolish the SQ phosphorylation site or knock out H2AX in other organisms also are not lethal (12, 23), suggesting that either there are alternative, less efficient, pathways to repair DSBs in the absence of γ-H2A.X or that cells can tolerate DSBs. While our studies do not allow us to determine whether *Tetrahymena* has such alternative pathways, it is clear that this organism can tolerate unrepaired DSBs.

Since the S134A rescued strain grew but had severe meiosis defects and stopped conjugation prematurely, it appears that, in *Tetrahymena*, either the role of γ-H2A.X is more important in repair of meiotic DSBs than in repair of breaks created during mitosis or that meiosis is more sensitive to the existence of DSBs. During vegetative growth, the MIC is not transcribed, and the damage accumulated in vegetative MICs can cause mitotic defects but will not have a major phenotypic effect until the next round of conjugation. Since *Tetrahymena* cells can replicate indefinitely in the absence of functional MICs, it may have been evolutionarily advantageous to eliminate mitotic DNA damage checkpoints. In MACs, which are transcriptionally active and control the phenotype of vegetative cells, there are ~45 copies of each chromosome, so breaks in some genes could be compensated by other copies that are still intact, enabling cells to survive with unrepaired DSBs.

Given that the MIC is the germ line nucleus and will give rise to the new MIC and MAC in the process of conjugation, it seems reasonable that *Tetrahymena* cells would have meiotic checkpoints that protect the long-term survival of the cells by stopping the process of conjugation to prevent production of progeny with aneuploid macronuclei. However, although H2A.X phosphorylation at the SQ motif is essential for proper meiosis and leads to premature termination of conjugation after meiosis II, matings between the S134A rescued cells showed no evidence of a block in prophase I even though they show chromosome loss at metaphase I and chromosome segregation defects in anaphase I. These results suggest that the recombination (or pachytene) checkpoint observed in many organisms, which monitors DSBs in meiotic recombination and delays cell cycle progression in prophase I until all the DSBs have been repaired (33, 66), is either weak or nonexistent in the absence of γ-H2A.X in *Tetrahymena* cells. It is possible that an adaptation process, which renders cells capable of overcoming the checkpoint-dependent block and permits meiotic progression with unrepaired DSBs, could be operating in the S134A rescued cells. Such adaptations have been observed in the budding yeast recombination checkpoint (33, 48, 93). However, this seems unlikely since there is no detectable delay in the progress of S134A cells through meiosis until after meiosis II (Fig. 3). The arrest observed in later stages of conjugation, after meiosis II and before the third prezygotic mitosis in matings between S134A rescued cells, could be explained by the activity of a mitotic DNA damage checkpoint activated by the unrepaired DSBs that persist past meiosis II in S134A rescued cells. If this is the case, there is no specific meiotic checkpoint in *Tetrahymena* in the absence of γ-H2A.X. However, if such a mitotic checkpoint exists in *Tetrahymena*, it must be subject to adaptation in vegetative S134A cells, which continue to grow, but not in conjugating cells.

Meiotic recombination events in mouse and yeast are well studied, and γ-H2A.X appearance in these organisms precedes and is spatially distinct from synapsis (47, 65, 95). SCs have not been identified in *Tetrahymena*. Our results show that *Tetrahymena* γ-H2A.X follows a similar time line, even though it lacks SCs; i.e., γ-H2A.X appears in early stage II cells, before the reported appearance of Rad51 in MICs in late stage II and before the close pairing of homologous chromosomes in stage IV (45), and disappears when the micronuclear chromosomes become condensed in metaphase I. These studies provide the first evidence for the timing of the appearance of meiotic DSBs in *Tetrahymena*, demonstrating that they occur in the very early crescent stage of *Tetrahymena* conjugation at the beginning of prophase of meiosis I and that they likely persist until the end of the crescent stage, when meiotic crossing over is probably completed. Finally, our results shed new light on the nature of the programmed degradation of the parental MAC during conjugation. Previous studies showed that this process is accompanied by the production of oligonucleosome-sized DNA fragments (20) and the appearance of caspase 1-, 8-, and 9-like (25, 37) and endonuclease G-like activities (38), suggesting that it occurred by a mechanism resembling apoptosis in higher organisms. However, apoptosis in higher organisms is also accompanied by phosphorylation of H2A.X (46, 54, 68), and we were unable to demonstrate any γ-H2A.X staining in degenerating MACs. We also failed to detect any open reading frames with significant homology to caspases 1, 8, and 9 or to endonuclease G in a BLASTP search of the TIGR gene predictions based on the *Tetrahymena* macronuclear genome sequence (http://seq.ciliate.org/cgi-bin/BLAST-tgd.pl), making it highly unlikely that an apoptotic-like mechanism is involved.

ACKNOWLEDGMENTS

We thank Josephine Bowen for critical reading of the manuscript. E.G., S.D.T., and C.D.A. also thank Linda Langman for technical support.

This work was supported by NIH grants to M.A.G. (GM 21793), S.D.T. (GM53512), and C.D.A. (GM63959).

REFERENCES

1. **Allen, S. L.** 1967. Genomic exclusion: a rapid means for inducing homozygous diploid lines in *Tetrahymena pyriformis*, Syngen I. Science **155:**575–578.
2. **Allen, S. L., M. I. Altschuler, P. J. Bruns, J. Cohen, F. P. Doerder, J. Gaertig, M. Gorovsky, E. Orias, and A. Turkewitz.** 1998. Proposed genetic nomenclature rules for *Tetrahymena thermophila*, *Paramecium primaurelia* and *Paramecium tetraurelia*. Genetics **149:**459–462.
3. **Allis, C. D., and D. K. Dennison.** 1982. Identification and purification of young macronuclear anlagen from conjugating cells of *Tetrahymena thermophila*. Dev. Biol. **93:**519–533.
4. **Allis, C. D., C. V. Glover, and M. A. Gorovsky.** 1979. Micronuclei of *Tetrahymena* contain two types of histone H3. Proc. Natl. Acad. Sci. USA **76:** 4857–4861.
5. **Allis, C. D., C. V. Glover, J. K. Bowen, and M. A. Gorovsky.** 1980. Histone variants specific to the transcriptionally active, amitotically dividing macronucleus of the unicellular eucaryote, *Tetrahymena thermophila*. Cell **20:**609–617.
6. **Bodenbender, J., A. Prohaska, F. Jauker, H. Hipke, and G. Cleffmann.** 1992. DNA elimination and its relation to quantities in the macronucleus of *Tetrahymena*. Dev. Genet. **13:**103–110.
7. **Bruns, P. J., T. B. Brussard, and A. B. Kavka.** 1976. Isolation of homozygous mutants after induced self-fertilization in *Tetrahymena*. Proc. Natl. Acad. Sci. USA **73:**3243–3247.
8. **Bruns, P. J.** 1986. Genetic organization of *Tetrahymena*, p. 27–44. *In* J. G. Gall (ed.), The molecular biology of ciliated protozoa. Academic Press, Inc., Orlando, FL.
9. **Burma, S., B. P. Chen, M. Murphy, A. Kurimasa, and D. J. Chen.** 2001. ATM phosphorylates histone H2AX in response to DNA double-strand breaks. J. Biol. Chem. **276:**42462–42467.
10. **Cassidy-Hanley, D., J. Bowen, J. Lee, E. S. Cole, L. A. VerPlank, J. Gaertig, M. A. Gorovsky, and P. J. Bruns.** 1997. Germline and somatic transformation of mating *Tetrahymena thermophila* by particle bombardment. Genetics **146:**135–147.
11. **Celeste, A., O. Fernandez-Capetillo, M. J. Kruhlak, D. R. Pilch, D. W. Staudt, A. Lee, R. F. Bonner, W. M. Bonner, and A. Nussenzweig.** 2003. Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks. Nat. Cell Biol. **5:**675–679.
12. **Celeste, A., S. Petersen, P. J. Romanienko, O. Fernandez-Capetillo, H. T. Chen, O. A. Sedelnikova, B. Reina-San-Martin, V. Coppola, E. Meffre, M. J. Difilippantonio, C. Redon, D. R. Pilch, A. Olaru, M. Eckhaus, R. D. Camerini-Otero, L. Tessarollo, F. Livak, K. Manova, W. M. Bonner, M. C. Nussenzweig, and A. Nussenzweig.** 2002. Genomic instability in mice lacking histone H2AX. Science **296:**922–927.
13. **Cervantes, M. D., R. S. Coyne, X. Xi, and M. C. Yao.** 2006. The condensin complex is essential for amitotic segregation of bulk chromosomes, but not nucleoli, in the ciliate *Tetrahymena thermophila*. Mol. Cell. Biol. **26:**4690–4700.
14. **Chadwick, B. P., and T. F. Lane.** 2005. BRCA1 associates with the inactive X chromosome in late S-phase, coupled with transient H2AX phosphorylation. Chromosoma **114:**432–439.
15. **Chen, H. T., A. Bhandoola, M. J. Difilippantonio, J. Zhu, M. J. Brown, X. G. Tai, E. P. Rogakou, T. M. Brotz, W. M. Bonner, T. Ried, and A. Nussenzweig.** 2000. Response to RAG-mediated V(D)J cleavage by NBS1 and gamma-H2AX. Science **290:**1962–1964.
16. **Choe, J., D. Kolodrubetz, and M. Grunstein.** 1982. The two yeast histone H2A genes encode similar protein subtypes. Proc. Natl. Acad. Sci. USA **79:**1484–1487.
17. **Chowdhury, D., M. C. Keogh, H. Ishii, C. L. Peterson, S. Buratowski, and J. Lieberman.** 2005. γ-H2AX dephosphorylation by protein phosphatase 2A facilitates DNA double-strand break repair. Mol. Cell **20:**801–809.
18. **Cole, E. S., D. Cassidy-Hanley, J. Hemish, J. Tuan, and P. J. Bruns.** 1997. A mutational analysis of conjugation in *Tetrahymena thermophila.* 1. Phenotypes affecting early development: meiosis to nuclear selection. Dev. Biol. **189:**215–232.
19. **Cui, B., and M. A. Gorovsky.** 2006. Centromeric histone H3 is essential for vegetative cell division and for DNA elimination during conjugation in *Tetrahymena thermophila*. Mol. Cell. Biol. **26:**4499–4510.
20. **Davis, M. C., J. G. Ward, G. Herrick, and C. D. Allis.** 1992. Programmed nuclear death: apoptotic-like degradation of specific nuclei in conjugating *Tetrahymena*. Dev. Biol. **154:**419–432.
21. **Doerder, F. P., and S. K. Shabatura.** 1980. Genomic exclusion in *Tetrahymena thermophila*: a cytogenetic and cytofluorometric study. Dev. Genet. **1:**205–218.
22. **Downs, J. A., S. Allard, O. Jobin-Robitaille, A. Javaheri, A. Auger, N. Bouchard, S. J. Kron, S. P. Jackson, and J. Cote.** 2004. Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. Mol. Cell **16:**979–990.
23. **Downs, J. A., N. F. Lowndes, and S. P. Jackson.** 2000. A role for *Saccharomyces cerevisiae* histone H2A in DNA repair. Nature **408:**1001–1004.
24. **Eisen, J. A., R. S. Coyne, M. Wu, D. Wu, M. Thiagarajan, J. R. Wortman, J. H. Badger, Q. Ren, P. Amedeo, K. M. Jones, L. J. Tallon, A. L. Delcher, S. L. Salzberg, J. C. Silva, B. J. Haas, W. H. Majoros, M. Farzad, J. M. Carlton, R. K. Smith, J. Garg, R. E. Pearlman, K. M. Karrer, L. Sun, G. Manning, N. C. Elde, A. P. Turkewitz, D. J. Asai, D. E. Wilkes, Y. Wang, H. Cai, K. Collins, B. A. Stewart, S. R. Lee, K. Wilamowska, Z. Weinberg, W. L. Ruzzo, D. Wloga, J. Gaertig, J. Frankel, C. C. Tsao, M. A. Gorovsky, P. J. Keeling, R. F. Waller, N. J. Patron, J. M. Cherry, N. A. Stover, C. J. Krieger, C. Del Toro, H. F. Ryder, S. C. Williamson, R. A. Barbeau, E. P. Hamilton, and E. Orias.** 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLOS Biol. **4:**e286.
25. **Ejercito, M., and J. Wolfe.** 2003. Caspase-like activity is required for programmed nuclear elimination during conjugation in *Tetrahymena*. J. Eukaryot. Microbiol. **50:**427–429.
26. **Endoh, H., and T. Kobayashi.** 2006. Death harmony played by nucleus and mitochondria: nuclear apoptosis during conjugation of tetrahymena. Autophagy **2:**129–131.
27. **Fernandez-Capetillo, O., S. K. Mahadevaiah, A. Celeste, P. J. Romanienko, R. D. Camerini-Otero, W. M. Bonner, K. Manova, P. Burgoyne, and A. Nussenzweig.** 2003. H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. Dev. Cell. **4:**497–508.
28. **Fillingham, J., M. C. Keogh, and N. J. Krogan.** 2006. γH2AX and its role in DNA double-strand break repair. Biochem. Cell Biol. **84:**568–577.
29. **Foster, E. R., and J. A. Downs.** 2005. Histone H2A phosphorylation in DNA double-strand break repair. FEBS J. **272:**3231–3240.
30. **Gavin, R. H.** 1965. The effects of heat and cold on cellular development in synchronized *Tetrahymena pyriformis* WH-6. J. Protozool. **12:**307–318.
31. **Gorovsky, M. A., M.-C. Yao, J. B. Keevert, and G. L. Pleger.** 1975. Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods Cell Biol. **9:**311–327.
32. **Hai, B., J. Gaertig, and M. A. Gorovsky.** 2000. Knockout heterokaryons enable facile mutagenic analysis of essential genes in *Tetrahymena*. Methods Cell Biol. **62:**513–531.
33. **Hochwagen, A., and A. Amon.** 2006. Checking your breaks: surveillance mechanisms of meiotic recombination. Curr. Biol. **16:**R217—R228.
34. **Jazayeri, A., A. D. McAinsh, and S. P. Jackson.** 2004. *Saccharomyces cerevisiae* Sin3p facilitates DNA double-strand break repair. Proc. Natl. Acad. Sci. USA **101:**1644–1649.
35. **Keogh, M. C., J. A. Kim, M. Downey, J. Fillingham, D. Chowdhury, J. C. Harrison, M. Onishi, N. Datta, S. Galicia, A. Emili, J. Lieberman, X. Shen, S. Buratowski, J. E. Haber, D. Durocher, J. F. Greenblatt, and N. J. Krogan.** 2006. A phosphatase complex that dephosphorylates γH2AX regulates DNA damage checkpoint recovery. Nature **439:**497–501.
36. **Kobayashi, J., H. Tauchi, S. Sakamoto, A. Nakamura, K. Morishima, S. Matsuura, T. Kobayashi, K. Tamai, K. Tanimoto, and K. Komatsu.** 2002. NBS1 localizes to gamma-H2AX foci through interaction with the FHA/BRCT domain. Curr. Biol. **12:**1846–1851.
37. **Kobayashi, T., and H. Endoh.** 2003. Caspase-like activity in programmed nuclear death during conjugation of *Tetrahymena thermophila*. Cell Death Differ. **10:**634–640.
38. **Kobayashi, T., and H. Endoh.** 2005. A possible role of mitochondria in the apoptotic-like programmed nuclear death of *Tetrahymena thermophila*. FEBS J. **272:**5378–5387.
39. **Kolodner, R. D.** 2000. Guarding against mutation. Nature **407:**687–689.
40. **Korch, R. S. a. C. T.** 1970. Alkylation induced gene conversion in yeast: use in fine structure mapping. Mol. Genet. Genomics **107:**201–208.
41. **Krogan, N. J., M. H. Lam, J. Fillingham, M. C. Keogh, M. Gebbia, J. Li, N. Datta, G. Cagney, S. Buratowski, A. Emili, and J. F. Greenblatt.** 2004. Proteasome involvement in the repair of DNA double-strand breaks. Mol. Cell **16:**1027–1034.
42. **Kupiec, M.** 2000. Damage-induced recombination in the yeast Saccharomyces cerevisiae. Mutat. Res. **451:**91–105.
43. **Kusch, T., L. Florens, W. H. Macdonald, S. K. Swanson, R. L. Glaser, J. R. Yates III, S. M. Abmayr, M. P. Washburn, and J. L. Workman.** 2004. Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. Science **306:**2084–2087.
44. **Li, A., J. M. Eirin-Lopez, and J. Ausio.** 2005. H2AX: tailoring histone H2A for chromatin-dependent genomic integrity. Biochem. Cell Biol. **83:**505–515.
45. **Loidl, J., and H. Scherthan.** 2004. Organization and pairing of meiotic chromosomes in the ciliate *Tetrahymena thermophila*. J. Cell Sci. **117:**5791–5801.
46. **Lu, C., F. Zhu, Y. Y. Cho, F. Tang, T. Zykova, W. Y. Ma, A. M. Bode, and Z. Dong.** 2006. Cell apoptosis: requirement of H2AX in DNA ladder formation, but not for the activation of caspase-3. Mol. Cell **23:**121–132.
47. **Mahadevaiah, S. K., J. M. Turner, F. Baudat, E. P. Rogakou, P. de Boer, J. Blanco-Rodriguez, M. Jasin, S. Keeney, W. M. Bonner, and P. S. Burgoyne.**

2001. Recombinational DNA double-strand breaks in mice precede synapsis. Nat. Genet. **27:**271–276.

48. **Malkova, A., L. Ross, D. Dawson, M. F. Hoekstra, and J. E. Haber.** 1996. Meiotic recombination initiated by a double-strand break in rad50 delta yeast cells otherwise unable to initiate meiotic recombination. Genetics **143:**741–754.
49. **Manniron, C., W. M. Bonner, and C. L. Hatch.** 1989. H2A.X. a histone isoprotein with a conserved C-terminal sequence, is encoded by a novel mRNA with both DNA replication type and poly(A) 3′ processing signals. Nucleic Acids Res. **17:**9113–9126.
50. **Martindale, D. W., C. D. Allis, and P. J. Bruns.** 1982. Conjugation in *Tetrahymena thermophila.* A temporal analysis of cytological stages. Exp. Cell Res. **140:**227–236.
51. **McDonald, B. B.** 1966. The exchange of RNA and protein during conjugation in *Tetrahymena*. J. Protozool. **13:**277–285.
52. **Mochizuki, K., and M. A. Gorovsky.** 2005. A Dicer-like protein in *Tetrahymena* has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase. Genes Dev. **19:**77–89.
53. **Morrison, A. J., J. Highland, N. J. Krogan, A. Arbel-Eden, J. F. Greenblatt, J. E. Haber, and X. Shen.** 2004. INO80 and gamma-H2AX interaction links ATP-dependent chromatin remodeling to DNA damage repair. Cell **119:** 767–775.
54. **Mukherjee, B., C. Kessinger, J. Kobayashi, B. P. Chen, D. J. Chen, A. Chatterjee, and S. Burma.** 2006. DNA-PK phosphorylates histone H2AX during apoptotic DNA fragmentation in mammalian cells. DNA Repair **5:**575–590.
55. **Nagata, T., T. Kato, T. Morita, M. Nozaki, H. Kubota, H. Yagi, and A. Matsushiro.** 1991. Polyadenylated and 3′ processed mRNAs are transcribed from the mouse histone H2A.X gene. Nucleic Acids Res. **19:**2441–2447.
56. **Nakamura, T. M., L. L. Du, C. Redon, and P. Russell.** 2004. Histone H2A phosphorylation controls Crb2 recruitment at DNA breaks, maintains checkpoint arrest, and influences DNA repair in fission yeast. Mol. Cell. Biol. **24:**6215–6230.
57. **Paques, F., and J. E. Haber.** 1999. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol. Mol. Biol. Rev. **63:**349–404.
58. **Paull, T. T., E. P. Rogakou, V. Yamazaki, C. U. Kirchgessner, M. Gellert, and W. M. Bonner.** 2000. A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage. Curr. Biol. **10:**886–895.
59. **Petrini, J. H., and T. H. Stracker.** 2003. The cellular response to DNA double-strand breaks: defining the sensors and mediators. Trends Cell Biol. **13:**458–462.
60. **Ray, C.** 1956. Meiosis and nuclear behavior in *Tetrahymena pyriformis*. J. Protozool. **3:**88–96.
61. **Redon, C., D. Pilch, E. Rogakou, O. Sedelnikova, K. Newrock, and W. Bonner.** 2002. Histone H2A variants H2AX and H2AZ. Curr. Opin. Genet. Dev. **12:**162–169.
62. **Ren, Q., and M. A. Gorovsky.** 2003. The nonessential H2A N-terminal tail can function as an essential charge patch on the H2A.Z variant N-terminal tail. Mol. Cell. Biol. **23:**2778–2789.
63. **Ren, Q. H.** 2002. In vivo functions of the post-translational modifications of histone H2A and H2A.Z in *Tetrahymena thermophila.* Ph.D. thesis. University of Rochester, Rochester, NY.
64. **Richardson, C., and M. Jasin.** 2000. Recombination between two chromosomes: implications for genomic integrity in mammalian cells. Cold Spring Harb. Symp. Quant. Biol. **65:**553–560.
65. **Roeder, G. S.** 1997. Meiotic chromosomes: it takes two to tango. Genes Dev. **11:**2600–2621.
66. **Roeder, G. S., and J. M. Bailis.** 2000. The pachytene checkpoint. Trends Genet. **16:**395–403.
67. **Rogakou, E. P., C. Boon, C. Redon, and W. M. Bonner.** 1999. Megabase chromatin domains involved in DNA double-strand breaks in vivo. J. Cell Biol. **146:**905–915.
68. **Rogakou, E. P., W. Nieves-Neira, C. Boon, Y. Pommier, and W. M. Bonner.** 2000. Initiation of DNA fragmentation during apoptosis induces phosphorylation of H2AX histone at serine 139. J. Biol. Chem. **275:**9390–9395.
69. **Rogers, M. B., and K. M. Karrer.** 1985. Adolescence in *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA **82:**436–439.
70. **Sasaki, S., M. Sato, Y. Katsura, A. Kurimasa, D. J. Chen, S. Takeda, H. Kuwano, J. Yokota, and T. Kohno.** 2006. Rapid assessment of two major repair activities against DNA double-strand breaks in vertebrate cells. Biochem. Biophys. Res. Commun. **339:**583–590.
71. **Schwartz, J. L.** 1989. Monofunctional alkylating agent-induced S-phase-dependent DNA damage. Mutat. Res. **216:**111–118.
72. **Sehorn, M. G., and P. Sung.** 2004. Meiotic recombination: an affair of two recombinases. Cell Cycle **3:**1375–1377.
73. **Sharpless, N. E., D. O. Ferguson, R. C. O'Hagan, D. H. Castrillon, C. Lee, P. A. Farazi, S. Alson, J. Fleming, C. C. Morton, K. Frank, L. Chin, F. W. Alt, and R. A. DePinho.** 2001. Impaired non-homologous end-joining provokes soft tissue sarcomas harboring chromosomal translocations, amplifications, and deletions. Mol. Cell **8:**1187–1196.
74. **Shroff, R., A. Arbel-Eden, D. Pilch, G. Ira, W. M. Bonner, J. H. Petrini, J. E. Haber, and M. Lichten.** 2004. Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. Curr. Biol. **14:** 1703–1711.
75. **Singh, N. P., and R. E. Stephens.** 1997. Microgel electrophoresis: sensitivity, mechanisms, and DNA electrostretching. Mutat. Res. **383:**167–175.
76. **Stahl, F.** 1996. Meiotic recombination in yeast: coronation of the double-strand-break repair model. Cell **87:**965–968.
77. **Stiff, T., M. O'Driscoll, N. Rief, K. Iwabuchi, M. Lobrich, and P. A. Jeggo.** 2004. ATM and DNA-PK function redundantly to phosphorylate H2AX after exposure to ionizing radiation. Cancer Res. **64:**2390–2396.
78. **Sugai, T., and K. Hiwatashi.** 1974. Cytologic and autoradiographic studies of the micronucleus at meiotic prophase in *Tetrahymena pyriformis*. J. Protozool. **21:**542–548.
79. **Sung, P., L. Krejci, S. Van Komen, and M. G. Sehorn.** 2003. Rad51 recombinase and recombination mediators. J. Biol. Chem. **278:**42729–42732.
80. **Takata, M., M. S. Sasaki, E. Sonoda, C. Morrison, M. Hashimoto, H. Utsumi, Y. Yamaguchi-Iwai, A. Shinohara, and S. Takeda.** 1998. Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. EMBO J. **17:**5497–5508.
81. **Unal, E., A. Arbel-Eden, U. Sattler, R. Shroff, M. Lichten, J. E. Haber, and D. Koshland.** 2004. DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. Mol. Cell **16:**991–1002.
82. **van Attikum, H., O. Fritsch, B. Hohn, and S. M. Gasser.** 2004. Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. Cell **119:**777–788.
83. **van Daal, A., E. M. White, M. A. Gorovsky, and S. C. Elgin.** 1988. *Drosophila* has a single copy of the gene encoding a highly conserved histone H2A variant of the H2A.F/Z type. Nucleic Acids Res. **16:**7487–7497.
84. **Ward, I. M., and J. Chen.** 2001. Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. J. Biol. Chem. **276:**47759–47762.
85. **Ward, I. M., K. Minn, K. G. Jorda, and J. Chen.** 2003. Accumulation of checkpoint protein 53BP1 at DNA breaks involves its binding to phosphorylated histone H2AX. J. Biol. Chem. **278:**19579–19582.
86. **Wei, Y., L. Yu, J. Bowen, M. A. Gorovsky, and C. D. Allis.** 1999. Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. Cell **97:**99–109.
87. **Wenkert, D., and C. D. Allis.** 1984. Timing of the appearance of macronuclear-specific histone variant hv1 and gene expression in developing new macronuclei of *Tetrahymena thermophila*. J. Cell Biol. **98:**2107–2117.
88. **Wolfe, J., B. Hunter, and W. S. Adair.** 1976. A cytological study of micronuclear elongation during conjugation in *Tetrahymena*. Chromosoma **55:** 289–308.
89. **Wu, G., A. G. McArthur, A. Fiser, A. Sali, M. L. Sogin, and M. Müller.** 2000. Core histones of the amitochondriate protist, *Giardia lamblia*. Mol. Biol. Evol. **17:**1156–1163.
90. **Xiao, W., and B. L. Chow.** 1998. Synergism between yeast nucleotide and base excision repair pathways in the protection against DNA methylation damage. Curr. Genet. **33:**92–99.
91. **Xiao, W., B. L. Chow, and L. Rathgeber.** 1996. The repair of DNA methylation damage in *Saccharomyces cerevisiae*. Curr. Genet. **30:**461–468.
92. **Xie, H., S. S. Wise, A. L. Holmes, B. Xu, T. P. Wakeman, S. C. Pelsue, N. P. Singh, and J. P. Wise, Sr.** 2005. Carcinogenic lead chromate induces DNA double-strand breaks in human lung cells. Mutat. Res. **586:**160–172.
93. **Xie, S., B. Xie, M. Y. Lee, and W. Dai.** 2005. Regulation of cell cycle checkpoints by polo-like kinases. Oncogene **24:**277–286.
94. **Yao, M. C., and J. L. Chao.** 2005. RNA-guided DNA deletion in *Tetrahymena*: an RNAi-based mechanism for programmed genome rearrangements. Annu. Rev. Genet. **39:**537–559.
95. **Zenvirth, D., T. Arbel, A. Sherman, M. Goldway, S. Klein, and G. Simchen.** 1992. Multiple sites for double-strand breaks in whole meiotic chromosomes of *Saccharomyces cerevisiae*. EMBO J. **11:**3441–3447.

MOLECULAR AND CELLULAR BIOLOGY, Apr. 2003, p. 2778–2789 Vol. 23, No. 8
0270-7306/03/$08.00+0 DOI: 10.1128/MCB.23.8.2778–2789.2003

The Nonessential H2A N-Terminal Tail Can Function as an Essential Charge Patch on the H2A.Z Variant N-Terminal Tail

Qinghu Ren† and Martin A. Gorovsky*

Department of Biology, University of Rochester, Rochester, New York 14627

Received 2 October 2002/Returned for modification 19 November 2002/Accepted 21 January 2003

***Tetrahymena thermophila* cells contain three forms of H2A: major H2A.1 and H2A.2, which make up ~80% of total H2A, and a conserved variant, H2A.Z. We showed previously that acetylation of H2A.Z was essential (Q. Ren and M. A. Gorovsky, Mol. Cell 7:1329–1335, 2001). Here we used in vitro mutagenesis of lysine residues, coupled with gene replacement, to identify the sites of acetylation of the N-terminal tail of the major H2A and to analyze its function in vivo. *Tetrahymena* cells survived with all five acetylatable lysines replaced by arginines plus a mutation that abolished acetylation of the N-terminal serine normally found in the wild-type protein. Thus, neither posttranslational nor cotranslational acetylation of major H2A is essential. Surprisingly, the nonacetylatable N-terminal tail of the major H2A was able to replace the essential function of the acetylation of the H2A.Z N-terminal tail. Tail-swapping experiments between H2A.1 and H2A.Z revealed that the nonessential acetylation of the major H2A N-terminal tail can be made to function as an essential charge patch in place of the H2A.Z N-terminal tail and that while the pattern of acetylation of an H2A N-terminal tail is determined by the tail sequence, the effects of acetylation on viability are determined by properties of the H2A core and not those of the N-terminal tail itself.**

In eukaryotic cell nuclei, DNA associates with histones to form chromatin. The basic unit of chromatin is the nucleosome core particle consisting of ~146 bp of DNA wrapped around an octamer of two of each of the four conserved core histones, H2A, H2B, H3, and H4 (85). All core histones contain a highly structured C-terminal histone-fold domain and a highly charged, structurally undefined, N-terminal tail domain that emerges from the histone core. These tails are thought to be important for DNA-histone and histone-histone interactions within and/or between nucleosomes and for interactions with nonhistone proteins (52, 89). In addition, H2A has an extended C-terminal tail contacting DNA near the dyad axis at the center of the nucleosome core (77, 87).

The seemingly simple, repetitive nature and highly condensed state of chromatin in the nucleus provide apparent limitations to chromatin functions, especially to transcription regulation in different cell types and physiological states. Several mechanisms have evolved to produce heterogeneity in the chromatin, including complex patterns of histone modifications and sequence variation of histones. Most histone modifications, including acetylation, phosphorylation, methylation, and ubiquitination, occur on the histone amino- and carboxyl-terminal tails (34). Of all the histone modifications, acetylation of the ε-amino group of lysine, which occurs after histone deposition and is restricted to the N-terminal tails, is probably the most abundant and is the best characterized. This acetylation has been closely linked to transcriptional activation (12, 63) by findings that many transcriptional activators or coactivators possess histone acetyltransferase (HAT) activity (13, 64), while corepressors containing histone deacetylase activity can repress transcription (54, 72). Recent studies also have shown that acetylation patterns of chromatin domains are important for establishing stable patterns of gene expression (28).

Besides the relationship between histone acetylation and transcription regulation, histone acetylation is also involved in processes such as DNA replication (80), nucleosome assembly, and chromosome condensation (74).

Two mechanisms have been proposed for how acetylation and other histone posttranslational modifications might act. Acetylation might modify chromatin structure and function by affecting histone-DNA interactions or histone-histone interactions. Alternatively or in addition, acetylation might act by altering the ability of nonhistone transcription or replication factors to bind to the N-terminal tails (44, 86).

It has recently been suggested that histone modifications, acting at specific sites either alone, in combination, or in sequential fashion on one or multiple histone tails, can form a complex histone code that can either enhance or reduce the interaction affinities with chromatin-associated proteins and thereby specify unique chromatin functions (42, 69, 76). The hallmark of this histone code mechanism is that the posttranslational modification provides a site-specific signal (either a structural motif, a structural change, or a specific charge) that affects recognition of the site by another molecule. There is ample evidence to support this kind of mechanism. Catalytic HATs and histone deacetylases do not acetylate and deacetylate histones nonspecifically (23, 84). Bromodomains found in HATs and in some transcription factors (PCAF, $TAF_{II}250$) can selectively interact with acetylated lysines either individually or in specific combination in the histone N-terminal tails (20, 41). The chromodomains of some heterochromatin proteins and histone methyltransferases are highly selective for methylated H3 at K9 but not for methylated H3 at K4 (11, 55). A lot of evidence also shows the interplay between different modifications on single or multiple histone tails (91). H3 phos-

* Corresponding author. Mailing address: Department of Biology, University of Rochester, Rochester, NY 14627. Phone: (585) 275-6988. Fax: (585) 275-2070. E-mail: goro@mail.rochester.edu.

† Present address: The Institute for Genomic Research, Rockville, MD 20850.

phorylation at serine 10 can enhance acetylation on lysine 14 and affect transcription at specific genes (17, 51). H4-R3 methylation mediated by PRMT1 (70, 81) facilitates p300-mediated acetylation on H4-K8 and H4-K12. In the reverse direction, histone H4 acetylation on any of the four lysines (K5, K8, K12, or K16) also inhibits the subsequent methylation at H4-R3 by PRMT1 (81).

Charge-altering modifications also can affect chromatin function by a second mechanism in which they alter the charge of a protein domain rather than affect a specific site. This "charge patch" mechanism has been shown to apply to regulation of the expression of specific genes by phosphorylation of linker histone H1 in *Tetrahymena* (21, 22) and to modulation of the essential function of histone H2A.Z by acetylation in *Tetrahymena* (62). In these cases, the function of the modification is to alter the charge of the domain in which it resides. Unlike the histone code, these changes need not be site specific. Modulation of the charge at any one of a number of clustered sites can have the same effect. The ability of acetylation to inhibit the salt-induced condensation of nucleosome oligomers in vitro (75) could be such an effect, and if acetylation were to inhibit nucleosome condensation in vivo, it could facilitate transcription.

In addition to posttranslational modifications of histones, another factor that contributes to chromatin functional heterogeneity is the existence of nonallelic histone variants (37). The demonstration that some histone mutations have highly specific effects on transcription, coupled with the observation that expression of some histone variants is temporally, developmentally, and spatially regulated, suggests that variant nucleosomes perform distinct functions (88).

The best-studied core histone variant is H2A.Z (61). H2A.Z has been found in diverse organisms, including *Tetrahymena* (83), *Saccharomyces cerevisiae* (39, 65), *Schizosaccharomyces pombe* (15), *Drosophila* (79), *Arabidopsis thaliana* (14), sea urchins (24), chickens (19, 36), and mammals (10). Phylogenetic analysis of H2A protein sequences (73) demonstrated that the major H2As and the H2A.Z variants diverged early in eukaryotic evolution and that the H2A.Zs show even less evolutionary divergence than the major H2As. Therefore, there were two types of H2A genes in primitive eukaryotes before the divergence of ciliates, fungi, animals, and plants, and they have been under different selective pressures since that time.

The evolutionary implication that the major H2As and H2A.Z variants have distinct and important functions has been confirmed experimentally. In *Drosophila* and in *S. cerevisiae*, the distribution of the two types of H2A in chromatin differs (47, 65). Deletions of genes encoding H2A.Z are lethal in *Tetrahymena* (50), *Drosophila* (78), and mice (25) and cause slow growth and/or conditional lethality in yeasts (2, 15, 40, 65). In both *S. cerevisiae* (43, 50, 66) and *Tetrahymena* (48), mutants without at least one of the two major histone genes (*HTA1* or *HTA2*) cannot survive. In *S. cerevisiae*, expression of the gene encoding H2A.Z (*HTZ1*) cannot rescue disruptions of both genes encoding the major H2As even when overexpressed or placed under control of the *HTA1* promoter (40, 65). Chimeric genes with different domains on the H2A.Z replaced with those of major H2A were injected into *Drosophila* H2A.Z null embryos to investigate which domain(s) is essential for H2A.Z function (18). Surprisingly, the essential portions of H2A.Z are the αC helix and H3/H4-binding domains. Thus, it is clear that the H2A.Z variants and the major H2As have distinct functions, both of which are either essential or required for normal growth in all organisms tested.

Little is known about the specific functions that are distinct to either the major H2As or to the H2A.Z variants. Pinto and Winston (58) argued that the major H2A of *S. cerevisiae* was required for normal centromere function because two cold-sensitive H2A mutations showed chromosome segregation phenotypes and interacted genetically with mutations in known centromere components. While this study did not specifically test the same mutations in H2A.Z, the fact that the H2A mutations had centromere-specific effects in cells containing a wild-type *HTZ1* gene suggests that this centromere function is specific for the major H2A. Considerable circumstantial evidence suggests that H2A.Z has a transcription-related function. In *Tetrahymena*, H2A.Z is present only in the transcriptionally active macronuclei and not in the transcriptionally inactive micronuclei of vegetative cells but appears in premeiotic micronuclei of conjugating cells when they become transcriptionally active (68). In *S. cerevisiae*, mutations in *HTZ1* are synthetically lethal with deletion of *SNF2* (65), a component of the SWI/SNF remodeling complex required for transcription of many genes, while mutations in the major histones suppress *SWI2* deletion (45, 46, 60, 82).

The likely function of H2A.Z in transcription and the well-documented relationship between acetylation and transcription led us to analyze the function of acetylation of H2A.Z. We showed that acetylation of *Tetrahymena* H2A.Z is essential and that it acts to modulate a charge patch on its N-terminal tail (62). Because the function of the major H2A of *Tetrahymena* is essential but distinct from that of H2A.Z, in this study we sought to identify the sites of acetylation of the major H2A in this organism and analyze their function. We changed the acetylation sites on major histone H2A.1 either from lysine to arginine, which conserves the net positive charge of lysine but cannot be neutralized by acetylation, or from lysine to glutamine, which resembles acetylated lysine in charge and structure. We showed that *Tetrahymena* cells survive plus five lysines of their major H2A replaced by arginines plus a mutation that abolishes the N-terminal acetylation of serine normally found in the wild-type protein (29, 30). Thus, acetylation of the major histone H2A is quite different from acetylation of H2A.Z: it is not essential, even though the protein itself is essential and constitutes ~80% of the total H2A. We also found that the N-terminal tail of H2A can replace the H2A.Z N-terminal tail and that the nonessential acetylation of the major H2A N-terminal tail can provide modulation of the charge patch on the H2A.Z N-terminal tail, which is essential for H2A.Z function. We conclude that when H2A.Z has a highly positively charged tail, it is essential that at least one of the positive charges of the tail can be neutralized in vivo by acetylation. However, this essential function of acetylation depends on properties of the H2A molecule that are independent of those of the tail itself.

MATERIALS AND METHODS

Strains, culture, and conjugation. Table 1 lists the *Tetrahymena thermophila* strains used in this study. Strains CU428, CU427, and B2086 were kindly provided by P. J. Bruns (Cornell University). Histone H2A knockout heterokaryon

TABLE 1. Strains used in this study

Strain	Genotype (micronucleus)	Phenotype (macronucleus)
CU428.2	*HTA1/HTA1 CHX1/CHX1 mpr1-1/mpr1-1*	wt,[c] pm-s cy-s mp-s VII
CU427.4	*HTA1/HTA1 chx1-1/chx1-1 MPR1/MPR1*	wt, pm-s cy-s mp-s VI
B2086.1	*HTA1/HTA1 CHX1/CHX1 MPR1/MPR1*	wt, pm-s cy-s mp-s II
G115B5	*ΔHTA1/ΔHTA1*[a] *HTA2/HTA2 CHX1/CHX1 mpr1?/mpr1?*[b]	wt, pm-s cy-s mp-r ?
G114B11	*ΔHTA1/ΔHTA1 HTA2/HTA2 CHX1/CHX1 mpr1?/mpr1?*	wt, pm-s cy-s mp-r ?
G209C4	*HTA1/HTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	wt, pm-s cy-s mp-r ?
G204F2	*HTA1/HTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	wt, pm-s cy-s mp-r ?
G4A1F14A	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	wt, pm-s cy-s mp-r ?
G4B1G6A	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	wt, pm-s cy-s mp-r ?
HTA1-D1A01	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	wt H2A.1 ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D9B6	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 RRRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D11B1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 ARRRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D18C1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 PRRRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D19B1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VRRRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D26A	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VKRRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D25C	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VRKRRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D23A1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VRRKRR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D27A1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VRRRKR ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D24A1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1 VRRRRK ΔH2A.2 pm-r cy-s mp-? ?
HTA3-T57E1	*ΔHTA3/ΔHTA3 CHX1/CHX1 mpr1?/mpr1?*	H2A.Z(H2A.1)$_{N}$ pm-r cy-s mp-? ?
HTA3-T65E2	*ΔHTA3/ΔHTA3 CHX1/CHX1 mpr1?/mpr1?*	H2A.Z(H2A.1RRRRR)$_{N}$ pm-r cy-s mp-? ?
HTA3-T90N	*ΔHTA3/ΔHTA3 CHX1/CHX1 mpr1?/mpr1?*	H2A.Z(H2A.1S1V+5R)$_{N}$ pm-r cy-s mp-? ?
HTA3-T89K	*ΔHTA3/ΔHTA3 CHX1/CHX1 mpr1?/mpr1?*	H2A.Z(H2A.1S1V+7R)$_{N}$ pm-r cy-s mp-? ?
HTA1-D35C	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1(H2A.Z)$_{N}$ ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D33A1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1(H2A.Z 6K)$_{N}$ ΔH2A.2 pm-r cy-s mp-? ?
HTA1-D34B1	*ΔHTA1/ΔHTA1 ΔHTA2/ΔHTA2 CHX1/CHX1 mpr1?/mpr1?*	H2A.1(H2A.Z 5K)$_{N}$ ΔH2A.2 pm-r cy-s mp-? ?

[a] The correct genetic nomenclature for Δ*HTA1*/Δ*HTA1* is *hta1-1*::*neo2*/*hta1-1*::*neo2* (3), but we use the abbreviation to conserve space.

[b] ?, genotype or mating type not determined.

[c] wt, wild type.

strains G4A1F14A and G4B1G6A and all mutant strains were generated as described below. For studies of vegetative growth, *Tetrahymena* cells were grown in SPP medium containing 1% proteose peptone (1× SPP) (32). For conjugation, two strains of different mating types were washed, starved (16 to 24 h, 30°C), and mated in 10 mM Tris-HCl (pH 7.5) as described previously (4).

Plasmid construction. The *HTA1* gene knockout construct (pQR10B) is based on plasmid pXL53, a pBluescript KS(+) derivative, which contains a copy of the *HTA1* coding sequence (49), 3.5 kb of 5′ flanking sequence, and 1.8 kb of 3′ flanking sequence. A 0.5-kb *Hin*dIII-*Hin*cII fragment, which included the whole coding sequence of the *HTA1* gene, was removed and replaced with a 1.5-kb *Hin*dIII-*Sma*I fragment from p4T2-1, a pBluescript KS(+) derivative, which contains a copy of the *neo2* gene cassette (31). The *neo2* gene, controlled by a histone H4 gene (*HHF1*) promoter, is transcribed in the direction opposite to that of the wild-type *HTA1* gene (Fig. 1A). The final fragment, *HTA1*::*neo2*, was released from pQR10B by digestion with *Kpn*I and *Sac*I.

For the *HTA2* gene knockout construct (pQR17), a 1.5-kb fragment of the *HTA2* 5′ flanking sequence (from an *Nsi*I site to the ATG start codon) was PCR amplified from *Tetrahymena* genomic DNA and inserted into the *Sma*I site of the 3′ polylinker region of p4T2-1. A 1.5-kb fragment of the *HTA2* 3′ flanking sequence (from the TGA stop codon to a *Bgl*II site) was PCR amplified from pXL46, a pBluescript KS(+) derivative that contains a copy of the *HTA2* gene, and inserted into the 5′ polylinker region (between *Kpn*I and *Eco*RV) of p4T2-1. The *neo2* gene is transcribed in the direction opposite to that of the wild-type *HTA2* gene (Fig. 1A). The final fragment, *HTA2*::*neo2*, was released from pQR17 by digestion with *Kpn*I and *Sac*I.

Site-directed mutagenesis. Oligonucleotide-directed, double-strand mutagenesis was performed as described previously (9) on pXL53, which contains a copy of the wild-type *HTA1* gene. In some cases, a silent mutation was introduced to generate a restriction enzyme site used to monitor transformation. All mutated genes were sequenced with an automatic sequencing system (ABI Prism) and released by digestion with *Kpn*I and *Sac*I before being introduced into knockout heterokaryons.

Construction of *HTA1* and *HTA2* double-knockout heterokaryons and transformation of mutated genes. Using the DuPont Biolistic PDS-1000/He particle delivery system (Bio-Rad Laboratories) as described previously (16), the *HTA1* and *HTA2* genes encoding H2A.1 and H2A.2 were disrupted individually with *HTA1*::*neo2* or *HTA2*::*neo2* by biolistic transformation into early stage (2.5 h) conjugating CU428 and B2086 cells. Homozygous knockout heterokaryon strains of *HTA1* (G115B5 and G114B11) and *HTA2* (G209C4 and G204F2) with different mating types were created as described previously (35).

To create major histone H2A double germ line knockout heterokaryons, the homozygous *HTA1*::*neo2* strain, G115B5, was mated with a homozygous *HTA2*::*neo2* strain, i.e., G209C4 or G204F2. The heterozygous progeny were then mated to a B*VI strain (56) as described previously (90) to obtain strains with homozygous *HTA1*::*neo2* and *HTA2*::*neo2* in the micronucleus. Two strains (G4A1F14A and G4B1G6A) with different mating types were created. These strains contain disrupted *HTA1* and *HTA2* genes in their micronuclei and wild-type genes in their macronuclei. When these paromomycin-sensitive heterokaryons conjugate, the old paromomycin-sensitive macronuclei are replaced by new ones produced by meiosis, fertilization, and mitotic division of the micronuclei of the cells. Consequently, the drug resistance genes that disrupt the *HTA1* and *HTA2* genes are expressed in the new macronuclei, allowing simple drug selection for successful mating. However, because major histone H2A.1 and H2A.2 together are essential in *Tetrahymena* (50) and the new macronucleus contains only disrupted copies of both genes, the progeny from this mating die unless they are transformed during mating with an *HTA1* gene that functions well enough to support growth.

Successful creation of germ line knockout heterokaryons of the *HTA1* and *HTA2* genes was demonstrated by showing that no viable progeny were obtained when double-knockout heterokaryons of two different mating types were mated and that progeny were able to be rescued by transformation with a wild-type copy of *HTA1*. In addition, the physical structure of the disrupted *HTA* gene in the micronucleus of the heterokaryons was examined by mating knockout heterokaryons with wild-type CU427 cells and selecting for retention of *HTA1*::*neo2* and *HTA2*::*neo2* by increasing the paromomycin concentration to 2.0 mg/ml. When genomic DNA was analyzed by PCR using primers specific for the 5′ and 3′ flanking sequences of *HTA1* or *HTA2*, the presence of disrupted genes (indicated by the presence of the 1.5-kb *neo2* cassette) was demonstrable in progeny cell macronuclei, indicating that the parental heterokaryons have the disrupted gene in their micronuclei. As expected, the heterozygous macronuclei of these progeny cells also have wild-type copies of the *HTA1* and/or *HTA2* gene as required to provide the essential major H2A functions (Fig. 1B).

Because matings between two knockout heterokaryons fail to produce viable offspring and their progeny can be rescued by either a wild-type or a nonlethal mutated version of the *HTA1* gene, these strains greatly facilitate systematic mutagenesis studies on major histone H2A modification sites, as illustrated in

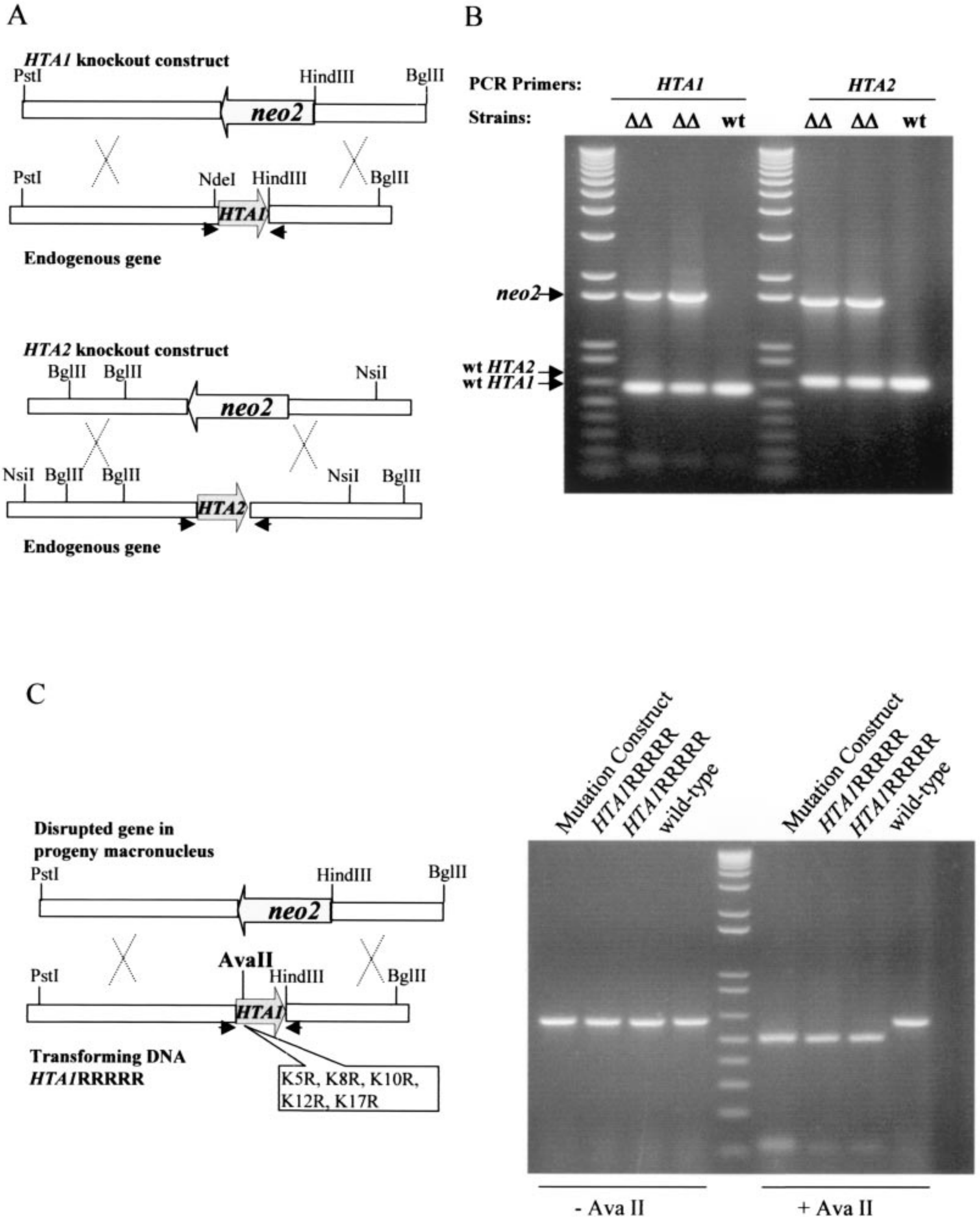


FIG. 1. Creation of major H2A knockout heterokaryons. (A) Knockout constructs for *HTA1* and *HTA2*. The entire coding region of either *HTA1* or *HTA2* was replaced by a *neo2* cassette that confers resistance to paromomycin when expressed in macronuclei. The macronuclear genomic *HTA1* gene is shown as a 5-kb *Pst*I-*Bgl*II fragment containing the *HTA1* coding region. The *HTA1*::*neo2* knockout construct is shown as a *neo2* cassette with 3.2-kb 5′ and 1.8-kb 3′ *HTA1* flanking sequences. The macronuclear genomic *HTA2* gene is shown as a 3.5-kb *Nsi*I-*Bgl*II fragment containing the *HTA2* coding region. The *HTA2*::*neo2* knockout construct is shown as a *neo2* cassette with 1.5-kb 5′ and 1.5-kb 3′ *HTA2* flanking sequences. In both knockout constructs, the *neo2* gene was transcribed in the direction opposite to that of the *HTA* gene. (B) PCR analysis of double-knockout heterokaryon progeny. The physical structure of the disrupted *HTA1* and *HTA2* genes in the micronuclei of the heterokaryons was examined by mating knockout heterokaryons with wild-type CU427 cells and selecting for retention of *HTA1*::*neo2* and *HTA2*::*neo2* by gradually increasing the paromomycin concentration to 2.0 mg/ml. When genomic DNA was analyzed by PCR using primers specific for the 5′ and 3′ flanking sequences of *HTA1* or *HTA2* (indicated by arrows in panel A), the presence of the disrupted *HTA1* and *HTA2* genes was demonstrable in progeny cell macronuclei *(neo2* band), indicating that the parental heterokaryons had the disrupted gene in their micronuclei. As expected, the heterozygous macronuclei of these progeny cells also had wild-type copies of the *HTA1* and *HTA2* genes required to provide the essential H2A functions. ΔΔ, progeny of mating CU427 × double germ line knockout heterokaryon; wt, progeny of mating CU427 × CU428 wild-type strain. (C) Example of an experiment in which knockout heterokaryon progeny were rescued by transformation with a mutated *HTA1* gene. The mutated *HTA1*RRRRR construct is shown as a 5.0-kb *Pst*I-*Bgl*II fragment containing specific mutations that also introduce a new *Ava*II restriction enzyme site. After this mutated gene was transformed into knockout heterokaryons, PCR analysis was done to test for the presence of the mutated gene in the transformants. The *HTA1* coding regions from wild-type control cells and *HTA1*RRRRR transformants were PCR amplified using the primers indicated by arrows in panel C. The PCR products, with or without *Ava*II digestion, were run on an agarose gel. The 600-bp PCR products from *HTA1*RRRRR mutant transformants contained the *Ava*II site and were cleaved into two shorter fragments, while the products from the wild-type strain were not cleaved.

Fig. 1C. A mutated form of the *HTA1* gene containing five arginine replacements at its N-terminal tail (see Results for details) was introduced into mating G4A1F14A and G4B1G6A knockout heterokaryons at late stages of conjugation (24 h) by biolistic transformation (16), and progeny were selected with paromomycin at 120 μg/ml. Viable progeny were obtained, indicating that the mutation is nonlethal. When the *HTA1* coding region of the progeny is PCR amplified using primers specific for the *HTA1* gene, the mutated and newly introduced *HTA1* gene is easily differentiated from the wild type because the mutated gene contains an *Ava*II restriction enzyme site such that only the PCR product from the desired mutants is cleaved by this enzyme. Finally, the genotypes of all mutants were confirmed by sequencing the PCR products from genomic DNA of the transformed progeny.

Growth analysis. Specific mutant strains, along with a strain rescued with the wild-type *HTA1*, were used in vegetative growth assays as described previously (67). Cells from each strain were inoculated into 50 ml of 1× SPP medium at starting densities of 1×10^4 cells/ml. Cultures were grown at 30°C with vigorous shaking, and samples (100 μl) were counted at frequent intervals with a ZB1 Coulter counter (Coulter Electronics, Inc.). Growth data were plotted using Cricket Graph III software (Computer Associates). Doubling times were calculated using the linear portion of the logarithmic growth curves.

Nuclear isolation and histone extraction. Rescued strains were grown to log phase (cell density, 2×10^5 cells/ml), and macronuclei were isolated by the method of Gorovsky et al. (32). Histones were extracted from macronuclei with 0.4 N H_2SO_4 (5) and precipitated with 20% trichloroacetic acid.

Acid-urea polyacrylamide gel electrophoresis and immunoblotting. Nuclear histones (25 μg) from mutants and wild-type *HTA1* rescued strains, with or without pretreatment with λ protein phosphatase (New England Biolabs, Inc.) at 10 U/μl for 5 h at 30°C, were separated on long acid-urea polyacrylamide slab gels (15% acrylamide, 6 M urea, 5% acetic acid) as described previously (6) and transferred onto an Immobilon-P membrane (Millipore). After blocking in 5% nonfat milk, anti-H2A (1:5,000) or anti-hv1 (1:10,000) (68) was added and the blot was incubated overnight at 4°C. A 1:100,000 dilution of horseradish peroxidase-conjugated goat anti-rabbit immunoglobulin G (Sigma) was used as secondary antibody. Blots were developed using the ECL Western blotting detection kit (NEN) according to the manufacturer's instructions.

RESULTS

Acetylation occurs on at least three of the lysines in the N-terminal tail. There are five lysine residues (5, 8, 10, 12, and 17) in the *T. thermophila* H2A.1 N-terminal tail (49). To identify which ones were acetylated, we mutated them, singly or in combination, to arginine. This conserves the net positive charge of lysine, but arginine cannot be neutralized by acetylation (53). We then used the mutated genes to rescue the progeny of a mating between two H2A knockout heterokaryon strains.

To determine the acetylation status of H2A.1, nuclear histones from strains rescued with wild-type or mutated *HTA1* genes were separated on acid-urea gels, which separate histones by both molecular weight and charge. Gels were then immunoblotted and stained with a highly specific antibody for *Tetrahymena* H2A to differentiate H2A.1 from any other comigrating *Tetrahymena* histones. *Tetrahymena* H2A is modified by both acetylation and phosphorylation (7), both of which alter the charge and therefore produce differences in mobility in this gel system. To eliminate the effects of phosphorylation on heterogeneity, histones were pretreated with λ protein phosphatase (33). This assay was used to characterize the acetylation status of all viable H2A.1 mutants.

If all lysines in the wild-type H2A.1 N-terminal tail can be modified by acetylation in vivo, up to six separable, phosphatase-resistant isotypes might be expected in wild-type cells. Strains rescued by the wild-type gene yielded five isoforms after treatment with phosphatase (Fig. 2B, lane 2), likely representing unmodified H2A.1 (bottom band) and isoforms containing one to four acetyl groups.

Since the exact acetylation sites were unknown, we began by changing all five available lysine residues to arginines (referred to as the RRRRR mutation). *Tetrahymena* survives with this mutated form of H2A.1 as its only source for major histone H2A, but the mutant strain grew slowly at 30°C (Fig. 2A). Surprisingly, histones extracted from mutant cells with all five lysines changed to arginines still have two phosphatase-resistant isoforms (Fig. 2B, lane 4). These results indicated that while at least three of the five internal lysines were acetylated, there was an additional, charge-altering modification in addition to phosphorylation and acetylation of lysine.

Neither N-terminal nor lysine acetylation of the H2A N-terminal tail is essential. The initial residue of the histone H2A N termini of many organisms (10, 38), including *Tetrahymena* (29, 30), can be blocked by N-terminal acetylation, a conserved process that adds an acetyl group to the first amino acid of many histone and nonhistone proteins (57). Since N-terminal acetylation abolishes one positive charge at the major H2A N terminus, it affects the mobility of major H2A.1 in the mutant, and if this process occurs for some but not all H2A.1 molecules, it might account for the electrophoretic heterogeneity observed in the RRRRR mutation. Because not all N-terminal residues can be acetylated, we attempted to remove any effect of N-terminal acetylation in the RRRRR mutation by further mutating the initial serine residue of H2A.1 to alanine, proline, or valine, residues which still allow removal of the initiator methionine (59). These mutation constructs (H2A.1ARRRRR, H2A.1PRRRRR, and H2A.1VRRRRR) gave viable transformants, all of which had slow-growth phenotypes (Fig. 2A). Nuclear histones from these three transformants were then extracted and separated on an acid-urea gel. Consistent with the observation that alanines following the initiator methionine are frequently acetylated (59), H2A.1ARRRRR still had two phosphatase-resistant isoforms, although the amount of the slower-migrating isoform was greatly reduced. Mutants H2A.1PRRRRR and H2A.1VRRRRR each contained only a single isoform after phosphatase treatment (Fig. 2B, lanes 6, 8, and 10), consistent with observations that these residues are less likely to be acetylated after the initiator methionine is removed. These data argue that the N-terminal serine of H2A.1 is normally acetylated and that acetylation of the major H2A, including its N-terminal acetylation, is not essential for viability in *Tetrahymena*.

All lysines in the H2A N-terminal tail can be acetylated. To map the exact acetylation sites of *Tetrahymena* histone H2A, a series of mutation constructs were generated from the H2A.1VRRRRR mutation, in which a single lysine residue replaced an arginine at each of the five positions. All of these mutations produced viable transformants with a slow-growth phenotype (Fig. 3). Acid-urea gel analyses showed that all of these mutants had two phosphatase-resistant isoforms: an unmodified H2A.1 and a mono-acetylated H2A.1 (data not shown). These results indicate that each of the five lysines in the N-terminal tail of *Tetrahymena* H2A can be acetylated.

Given that there are six acetylatable sites (those of the N-terminal residue and five lysines) in the N-terminal tail of wild-type H2A, it may seem surprising that we have only observed 5 H2A isoforms on acid-urea gels after dephosphory-

A

Major H2A N-terminal Tail	Rescue	Pase-resistant Isoforms	Acetylation
Ac STTGKGGKAKGKTASSKQ 5 8 10 12 17	+	5	+
R R R R R	+*	2	+
A R R R R R	+*	2	+
P R R R R R	+*	1	–
V R R R R R	+*	1	–

B

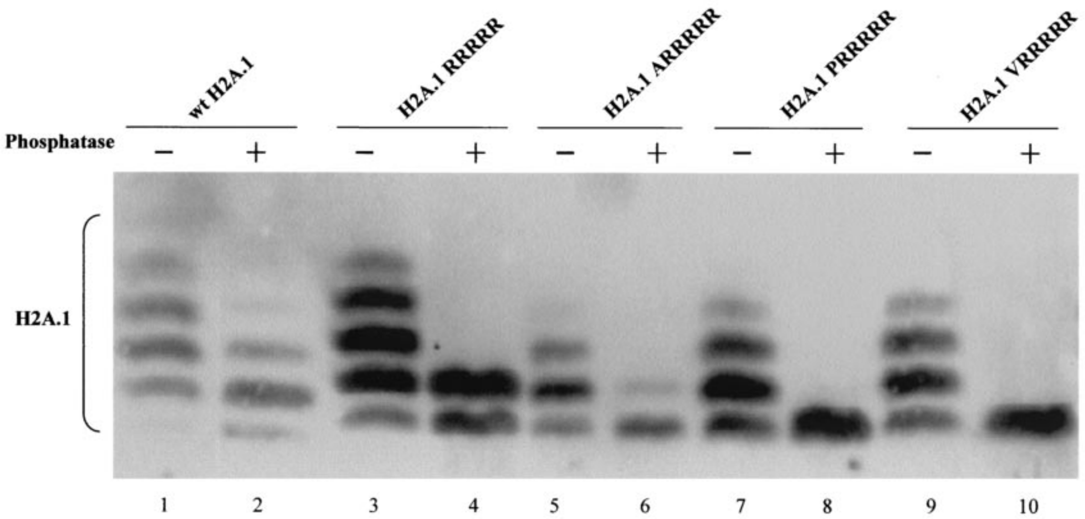


FIG. 2. Major H2A acetylation is not essential. (A) *Tetrahymena* can survive with all five lysines changed to arginines at the major H2A N-terminal tail. To abolish N-terminal acetylation, the first residue (serine) in the 5R mutated gene (RRRRR) was changed to alanine (ARRRRR), proline (PRRRRR), or valine (VRRRRR). All mutations generated viable transformants. *, mutants with severe phenotypes, including slow growth, variable size, and irregular surfaces. (B) Macronuclear histones from 5R transformants containing different N-terminal residues were analyzed on acid-urea gels. Histones from viable transformants were separated, blotted, and detected with an antiserum specific for H2A. H2A.1RRRRR and H2A.1ARRRRR show two phosphatase-resistant isoforms, indicating that N-terminal acetylation still occurs in these mutants (see text for details). H2A.1PRRRRR and H2A.1VRRRRR both show only one phosphatase-resistant isoform, indicating that these mutations abolish all acetylation at the H2A N-terminal tail.

lation (Fig. 2B). However, prediction of the number of observed isoforms from the number of acetylatable sites is not simple. First, the most highly acetylated isoforms are invariably faint and they are slightly variable in appearance. This is likely due to the fact that even small amounts of deacetylation during cell pelleting, nuclear isolation, or histone extraction can cause these isoforms to disappear and contribute to slower-migrating, less-acetylated isoforms. Another possibility is that while six acetylation sites can be identified by mutational analysis, few (if any) molecules in vivo are simultaneously acetylated at all six sites.

The major H2A N-terminal tail can replace the function of the H2A.Z variant N-terminal tail. The region encoding the entire N-terminal tail of *HTA1*, including all of the acetylatable lysines, was used to replace the region encoding the N-terminal tail of H2A.Z in the *HTA3* gene. This chimeric gene, encoding an H2A.Z variant core with an H2A.1 N-terminal tail, was used to rescue the progeny of mating H2A.Z germ line knockout heterokaryons (62). Viable transformants were obtained (Fig. 4A). Growth of these rescued strains was indistinguishable from that of wild-type cells. Using hv1, a specific antibody to H2A.Z, the modification state of this chimeric protein was then determined by immunoblotting the nuclear histones. In wild-type *Tetrahymena* cells, H2A.Z shows five or six phosphatase-resistant isoforms (Fig. 4B, lane 2). The chimeric protein H2A.Z(H2A.1)$_N$ shows four or five phosphatase-resistant isoforms (Fig. 4B, lane 4), a pattern similar to that of wild-type H2A.1. In addition, the mobility of H2A.Z(H2A.1)$_N$ isoforms on acid-urea gels is intermediate between that of wild-type H2A.1 and that of wild-type H2A.Z. The fastest migrating isoform, presumably the unmodified H2A.Z(H2A.1)$_N$, which has six positive charges in the N-terminal tail (including one positive charge of the N-terminal residue), has mobility similar to that of wild-type H2A.Z with three acetyl groups, whose tail

Major H2A N-terminal Tail	Rescue	Pase-resistant Isoforms	Acetylation
Ac STTGKGGKAKGKTASSKQ 5 8 10 12 17	+	5	+
V R R R R R	+*	1	-
V K R R R R	+*	2	+
V R K R R R	+*	2	+
V R R K R R	+*	2	+
V R R R K R	+*	2	+
V R R R R K	+*	2	+

FIG. 3. All lysines in the N-terminal tail of H2A can be acetylated. A series of mutation constructs were made starting from the H2A.1 VRRRRR mutation, leaving a single lysine at different positions. All mutations containing a single lysine generated viable transformants with slow-growth phenotypes. *, mutants with severe phenotypes, including slow growth, variable size, and irregular surfaces.

also has a net positive charge of +6. These data demonstrate that the N-terminal tail of H2A.1 can provide the function of the H2A.Z variant N-terminal tail.

The H2A.Z(H2A.1)$_N$ chimeric protein exhibits the essential charge patch properties of wild-type H2A.Z. Our previous studies had demonstrated that the N-terminal tail of H2A.Z in which all of the acetylatable lysines had been replaced by arginines cannot support growth in *Tetrahymena*. However, mutants with even a single lysine, or with lysines replaced by glutamine or with the N-terminal tail deleted, were viable (62). These observations were used to argue that the H2A.Z tail functioned as a charge patch in which neutralization or removal of at least one of the lysine positive charges was required for the creation of a viable gene. The net positive charge of the unmodified H2A.Z N-terminal tail is +9. Since a tail containing five arginines and one acetylatable lysine is viable, this suggests that an H2A.Z N terminus that can be modified to a net charge of +8 in vivo is viable.

Tetrahymena can survive with all five lysines at the N-terminal tail of the chimeric gene H2A.Z(H2A.1)$_N$ changed to arginines (Fig. 4A). The acid-urea gel analysis shows that H2A.Z(H2A.1RRRRR)$_N$ has two phosphatase-resistant isoforms (Fig. 4B, lane 6), indicating that the chimeric protein is modified by N-terminal acetylation to produce a blocked N-terminal tail with a net positive charge of +5 and a unblocked, completely unacetylated N-terminal tail with the maximum possible positive charge of +6. In similarity to the result seen with H2A.1, changing the first serine residue in this chimeric protein to valine eliminates N-terminal acetylation (Fig. 4B, lane 8), creating a cell that is viable and whose N-terminal tail has a total charge of +6. To determine whether the H2A.1 N-terminal tail attached to H2A.Z also functions as a charge patch, we introduced additional arginine residues into the chimeric gene by changing the last two or three amino acids of the H2A.1 N-terminal tail to arginines to increase its total of positive charges. *Tetrahymena* cells can survive with a total of eight positive charges in the tail of the chimeric protein but not with nine (Fig. 4A). Thus, just as in the case of the H2A.Z N-terminal tail itself, the H2A.1 N-terminal tail attached to H2A.Z cannot function with nine positive charges but produces viable progeny with eight. This argues that the H2A.1 N-terminal tail also can function as an essential charge patch when it replaces the H2A.Z N-terminal tail.

The H2A.Z N-terminal tail can provide the function of the major histone H2A N-terminal tail. The entire N-terminal tail of H2A.Z, including six acetylation sites and two nonacetylatable lysines, was used to replace the H2A.1 N-terminal tail. This chimeric gene, H2A.1(H2A.Z)$_N$, was introduced into the progeny of major H2A double germ line knockout heterokaryons. Viable H2A.1(H2A.Z)$_N$ transformants were obtained but grew slowly, as seen in a comparison of the growth curve with that of a strain rescued with the wild-type *HTA1* gene (Fig. 5B). Cells rescued with a gene [H2A.1(H2A.Z 6K)$_N$] in which the two nonacetylatable lysines in the H2A.Z N-terminal tail were changed to glutamines (reducing the maximum positive charge in vivo to +7) still grew slowly (Fig. 5B). When a third, acetylatable residue was changed to glutamine, the H2A.1 (H2A.Z 5K)$_N$ transformants, in which the N-terminal tail of the chimeric protein had a maximum total of six positive charges, exhibited growth rates that were indistinguishable from those of the wild-type cells (Fig. 5B). Because the maximum positive charge of the H2A.1 tail itself is +6, these data suggest that H2A N-terminal tail has an optimum maximum number of positive charges that can be well tolerated, even if most molecules can be acetylated and have a lower total number of positive charges. This suggests that the H2A tail is carrying out two distinct nonessential roles, one requiring an unacetylated tail with a charge of +6 and the other requiring acetylation.

The acid-urea gel analysis shows that the chimeric proteins H2A.1(H2A.Z)$_N$, H2A.1(H2A.Z 6K)$_N$, and H2A.1(H2A.Z 5K)$_N$ have five or six phosphatase-resistant isoforms, a pattern similar to that of wild-type H2A.Z (Fig. 5C, lanes 4, 6, and 8), indicating that the acetylation state is determined mainly by the N-terminal tail of the histones. The lack of apparent difference in the number of phosphatase-resistant isoforms be-

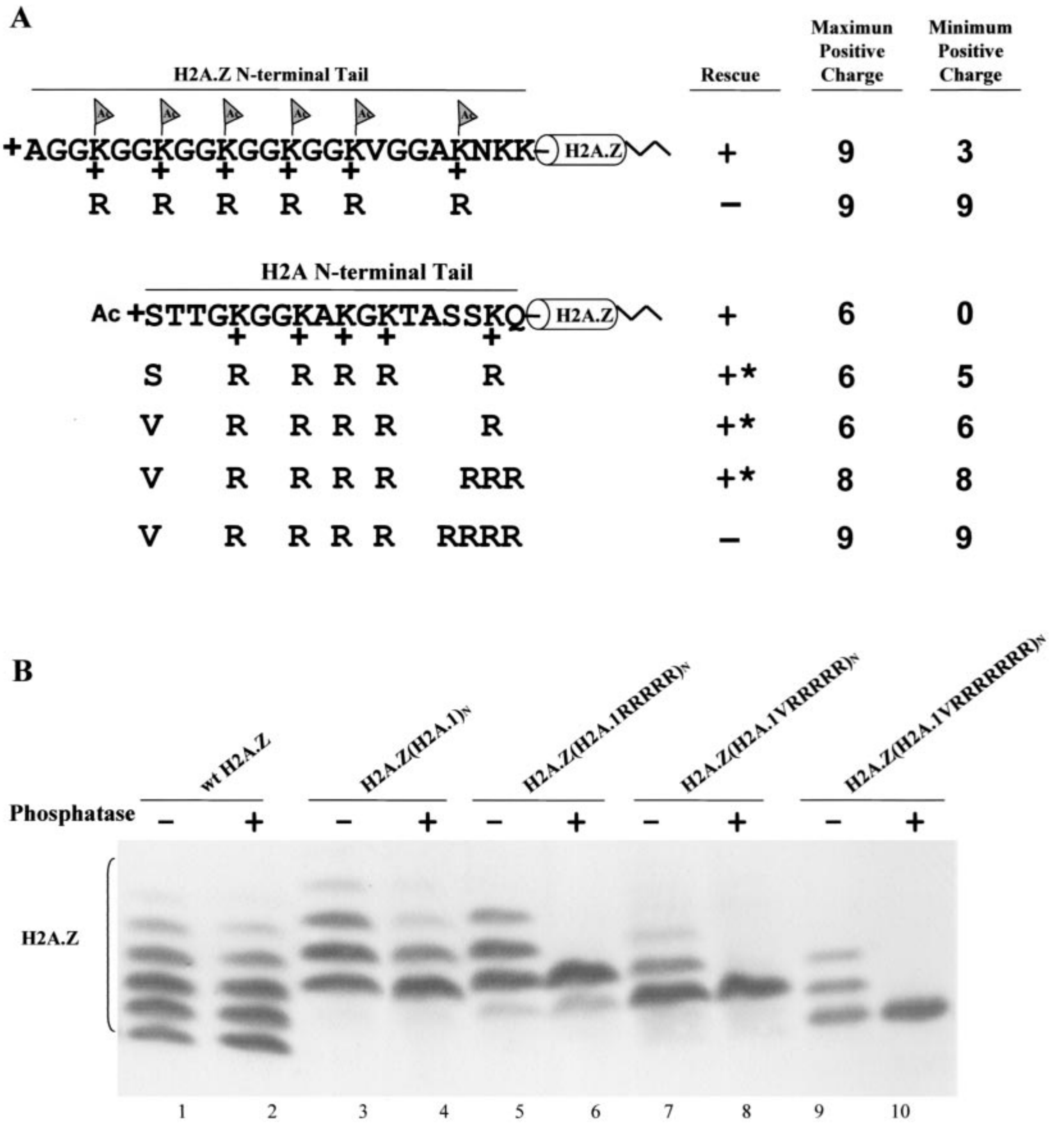


FIG. 4. The major H2A N-terminal tail can replace the essential function of the H2A.Z variant N-terminal tail. (A) Constructs in which the N-terminal tail of H2A.Z, whose acetylation has an essential function, was replaced either by a wild-type H2A.1 tail or by mutated tails containing increasing numbers of arginine replacements were transformed into H2A.Z knockout heterokaryons. Transformants with the chimeric gene containing a wild-type H2A.1 N-terminal tail on H2A.Z grow normally. *Tetrahymena* cells can also survive with arginine replacement mutations which increase the total positive charge of the N-terminal tail to +8 but cannot survive when the total positive charge is increased to +9. *, mutants with severe phenotypes, including slow growth, variable size, and irregular surfaces. (B) Macronuclear histones from viable transformants were extracted, blotted, and detected with a specific antibody to H2A.Z. The banding patterns of the chimeric proteins (H2A.1 tails on H2A.Z) caused by acetylation are similar to those of the major H2A.1, with three or four phosphatase-resistant isoforms for H2A.Z(H2A.1)$_N$ and two for H2A.Z(H2A.1RRRRR)$_N$. H2A.Z(H2A.1VRRRRR)$_N$ and H2A.Z(H2A.1VRRRRRRR)$_N$ both show a single phosphatase-resistant isoform.

tween mutants H2A.1(H2A.Z 6K)$_N$ andH2A.1(H2A.Z 5K)$_N$, although the latter has one acetylatable lysine changed into glutamine, can be explained by the observation that despite the possibility that all six lysine residues serve as acetylation sites, H2A.Z isolated from the wild-type strain does not show an isoform with six acetyl groups, likely reflecting the dynamic balance between the acetylation and deacetylation processes. The differences in the mobility of the chimeric protein on the acid-urea gel among the mutants are caused by the molecular difference between lysine and glutamine.

DISCUSSION

A number of important conclusions can be drawn from this work. (i) Although the major histone H2A is essential and constitutes ~80% of the total H2A, its acetylation is not essential. This conclusion applies to both cotranslational N-terminal acetylation and internal acetylation of lysine residues, as shown by the observation that both the VRRRRR and the PRRRRR versions of the major H2A showed no detectable electrophoretic heterogeneity and were viable. (ii) The N-terminal tail of histone H2A can replace the function of the

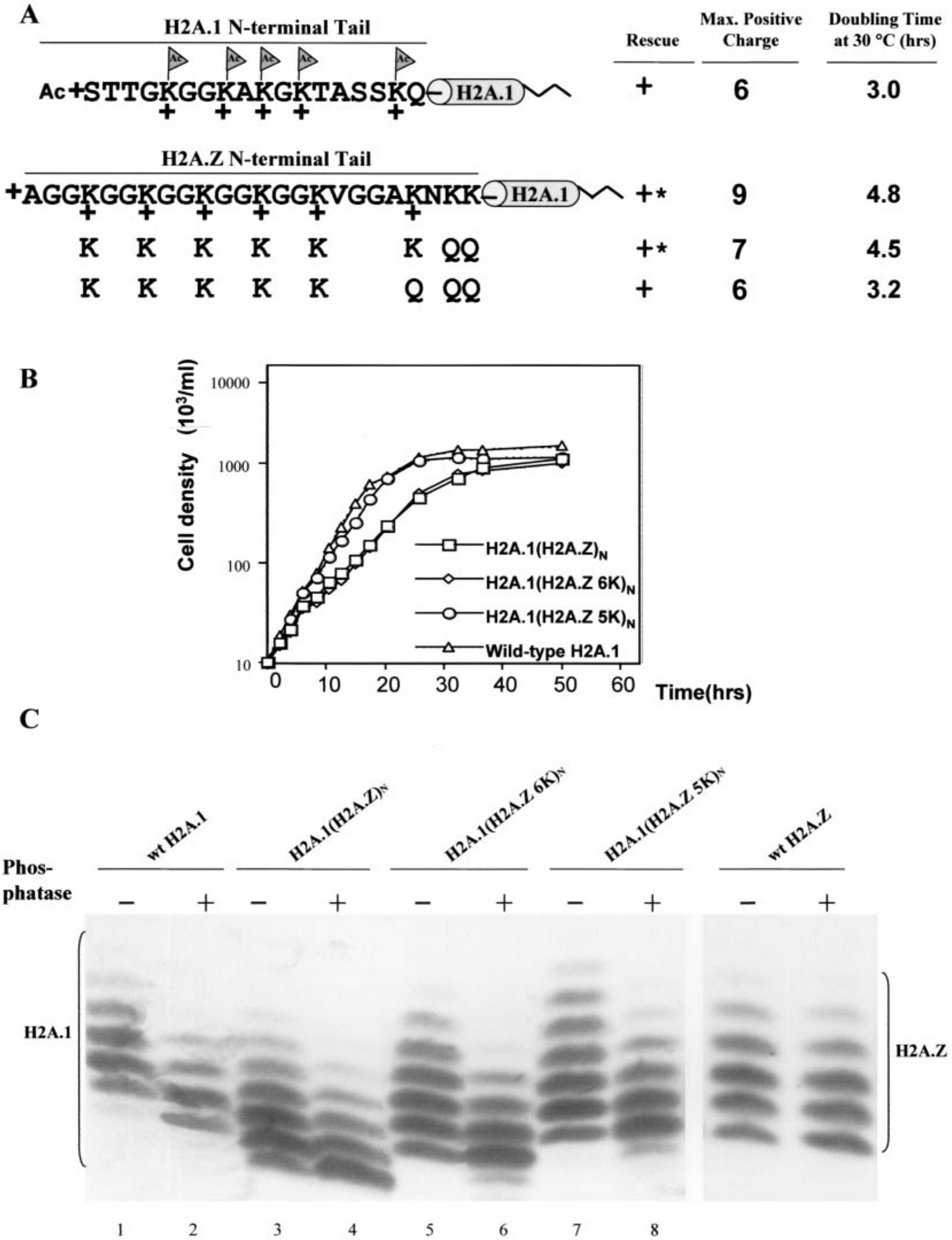


FIG. 5. The H2A.Z N-terminal tail can also replace the function of the major H2A N-terminal tail. (A) The N-terminal tail of H2A.1 was replaced either by the wild-type H2A.Z N-terminal tail or by tails with different numbers of glutamine replacements at nonacetylatable residues that reduce the positive charges of the N-terminal tail without eliminating acetylation sites. The wild-type H2A.Z N-terminal tail on H2A.1 yields transformants with a slow-growth phenotype, while glutamine replacement mutations, which decrease the maximum positive charge of the chimeric protein's N-terminal tail to +6, generate transformants whose doubling time at 30°C is indistinguishable from that of wild-type cells. *, mutants with severe phenotypes, including slow growth, variable size, and irregular surfaces. (B) Mutant H2A.1(H2A.Z)$_N$, H2A.1(H2A.Z 6K)$_N$, and H2A.1(H2A.Z 5K)$_N$, as well as strains rescued with the wild-type gene, were grown in 1× SPP medium at 30°C. Cell densities were measured for up to 50 h and plotted on a log scale. Doubling times in hours are listed in Fig. 5A. (C) Macronuclear histones from the mutants were extracted, blotted, and detected with a specific antibody to major H2A. While wild-type H2A.1 has four or five phosphatase-resistant isoforms, the chimeric protein, H2A.1(H2A.Z)$_N$, shows five or six phosphatase-resistant isoforms, a pattern similar to that of wild-type H2A.Z. H2A.1(H2A.Z 6K)$_N$ and H2A.1(H2A.Z 5K)$_N$ show patterns similar to that of H2A.1(H2A.Z)$_N$, except for small mobility differences likely caused by the extra glutamine mutations that abolish two or three positive charges on the N-terminal tail.

H2A.Z N-terminal tail and vice versa. (iii) Although it differs in sequence from the H2A.Z tail, the H2A.1 N-terminal tail can be made to function as an essential charge patch when linked to H2A.Z. (iv) Acetylation of the N-terminal residue has an effect similar to that of acetylation of internal lysine residues in modulating the function of the H2A tail as a charge patch. (v) The acetylation patterns of H2A tails are largely determined by the tails themselves, but the effects of acetylation are determined by the nature of the remainder of the histone. (vi) There appears to be a specific maximum charge that can be tolerated by each type of H2A. (vii) The function of acetylation in regulating the charge of the N-terminal tails of H2A is highly sensitive to changes of even a single charge.

Tetrahymena histone H2As are blocked at the N-terminal serine residue by an α-N-acetyl (29, 30). N-terminal blocking of histones and other proteins through acetylation is common in various organisms (57), and N-terminal acetylation of actin has been shown to strengthen the weak interaction between actin and myosin (1). To our knowledge, the finding that cotranslational N-terminal acetylation can affect histone function in a manner similar to that of internal, posttranslational acetylation has not been previously described. Since one of the important mechanisms by which acetylation affects histone function is that of modulating a charge patch (62; this study), the effect of N-terminal acetylation on the positive charge of the histone H2A N terminus must be considered a possible regulatory mechanism, especially in cases (as in that of *Tetrahymena* H2A) in which not all of the molecules contain this modification.

Interestingly, the extent of acetylation of the N terminus of H2A in *Tetrahymena* depends not only on the terminal residue but also on the sequence to which it is attached. The occurrence of N-terminal acetylation is sequence restricted and can often be correctly predicted by protein primary sequence (27). Three N-terminal acetyltransferases have been cloned in *S. cerevisiae*, and the conserved recognition sequences for each enzyme were reported (59). We initially changed the first serine of histone H2A.1 to alanine because alanine is structurally similar to serine and because the N-terminal alanine residue on H2A.Z is not blocked by acetylation (8, 62). Cells containing this mutation, H2A.1ARRRRR, still contained two phosphatase-resistant isoforms, although the amount of the slower-migrating, acetylated form is greatly reduced compared to that of H2A.1SRRRRR. However, when we changed the first serine to proline (P) or valine (V), neither of which is found in the conserved recognition sequence for N-terminal acetyltransferases, H2A.1PRRRRR and H2A.1VRRRRR mutations produced viable transformants in which the mutated H2A.1 showed only one phosphatase-resistant isoform representing unmodified H2A.1. These results demonstrate that major histone H2A acetylation, including acetylation of the N-terminal residue, is not essential in *Tetrahymena*.

We mapped the acetylation sites on H2A.1 by changing all but one wild-type lysine in the N-terminal tail to arginine and analyzing the H2A modification status in each mutant. We found that all five lysines in the N-terminal tail can be acetylated. However, they are not modified to equal extents. Lysines at the first three positions, K5, K8, and K10, are heavily acetylated, since mutants with a single lysine at these positions contain much more of the mono-acetylated isoform than other mutants (data not shown). This approach cannot rule out the possibility that some of these lysines are only acetylated when other sites are not available and are not normally acetylated in wild-type cells.

The studies described here strongly support the hypothesis that the essential function of acetylation of the H2A.Z tail acts by modulating the charge of the tail. We reported previously that *Tetrahymena* H2A.Z acetylation modulated an essential charge patch (62). *Tetrahymena* cannot survive with all six acetylatable lysines on H2A.Z changed into arginines, which produces a tail whose charge cannot be reduced from +9 (resulting from the presence of the amino-terminal α-amino group, six nonacetylatable arginines at the acetylation sites, and two nonacetylatable lysines). However, viable transformants can be obtained simply by reducing the charge to +8 by replacing a neutral residue with a negatively charged residue at other positions in the tail or by replacing any one of the arginine residues with glutamine (62). Remarkably, this same sensitivity to charge can be demonstrated by replacing the N-terminal tail of H2A.Z with an N-terminal tail of H2A.1 to which positive charges were added. We were able to demonstrate that the maximum number of nonneutralizable positive charges allowed on the N-terminal tail, including the one on the N-terminal residue, is eight; mutants in which the chimeric protein contained nine positive charges in the H2A.1 tail were not viable.

Surprisingly, while the extent of acetylation of the H2A tail was determined largely by the nature of the tail, the effect of acetylation on viability depended on the nature of the rest of the H2A molecule. One possibility is that acetylation of the tail affects the structure of major H2A and H2A.Z differently or acts synergistically with properties that differentiate the two types of H2A. Recent studies comparing the crystal structures of nucleosomes containing the major H2A with those containing the H2A.Z variant (71) provide some basis for this hypothesis. The region of H2A.Z essential for viability in *Drosophila* is at the C-terminal tail that is exposed on the surface of the nucleosome and is part of the docking domain involved in maintaining the interactions between the H3/H4 tetramer and H2A/H2B dimer within the nucleosome (18, 37). The H2A.Z nucleosomes have an altered surface that includes a metal ion, which may lead to changes in higher-order structure or in the association between H2A.Z and other nuclear proteins (71). It also has been shown that the presence of H2A.Z variants and tail acetylation of histones can each affect the hydrodynamic properties of nucleosomal arrays, offering the possibility that these two processes cooperate to establish unique chromatin domains (26). An alternative explanation of how the effects of acetylation can be determined by nonacetylated portions of the H2A molecule is based on the observations that nucleosomes containing the major H2A and those containing H2A.Z associate with different DNA sequences in chromatin (47, 65). In this scenario, acetylation can be viewed as a simple switch that is able to alter the properties of both major H2A and H2A.Z nucleosomes similarly but whose effect depends on the specific sequences with which each type of H2A was associated.

The results described here can reconcile our previous study demonstrating the essential function of a single acetylation site in the H2A.Z N-terminal tail (62) with studies of *Drosophila* showing that the only region of *Drosophila* H2A which cannot provide the essential developmental function of H2A.Z resides in the C-terminal α-helix (18). The *Drosophila* results are completely consistent with our finding that acetylation sites on the major H2A can replace those on H2A.Z when associated with the H2A.Z C-terminal helix in a fusion protein.

ACKNOWLEDGMENTS

This work was supported by grant GM21793 from the National Institutes of Health to M.A.G.

We thank Josephine Bowen for technical assistance and for critically reading the manuscript.

REFERENCES

1. **Abe, A., K. Saeki, T. Yasunaga, and T. Wakabayashi.** 2000. Acetylation at the N-terminus of actin strengthens weak interaction between actin and myosin. Biochem. Biophys. Res. Commun. **268:**14–19.
2. **Adam, M., F. Robert, M. Larochelle, and L. Gaudreau.** 2001. H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions. Mol. Cell. Biol. **21:**6270–6279.
3. **Allen, S. L., M. I. Altschuler, P. J. Bruns, J. Cohen, F. P. Doerder, J. Gaertig, M. Gorovsky, E. Orias, A. Turkewitz, and The Seventh International Meeting on Ciliate Molecular Biology Genetics Nomenclature.** 1998. Proposed genetic nomenclature rules for *Tetrahymena thermophila*, *Paramecium primaurelia* and *Paramecium tetraurelia*. Genetics **149:**459–462.
4. **Allis, C. D., and D. K. Dennison.** 1982. Identification and purification of young macronuclear anlagen from conjugating cells of *Tetrahymena thermophila*. Dev. Biol. **93:**519–533.
5. **Allis, C. D., C. D. C. Glover, and M. A. Gorovsky.** 1979. Micronuclei of *Tetrahymena* contain two types of histone H3. Proc. Natl. Acad. Sci. USA **76:**4857–4861.
6. **Allis, C. D., C. V. D. Glover, J. K. Bowen, and M. A. Gorovsky.** 1980. Histone variants specific to the transcriptionally active, amitotically dividing macronucleus of the unicellular eukaryote, *Tetrahymena thermophila*. Cell **20:**609–617.
7. **Allis, C. D., and M. A. Gorovsky.** 1981. Histone phosphorylation in macro- and micronuclei of *Tetrahymena thermophila*. Biochemistry **20:**3828–3833.
8. **Allis, C. D., R. Richman, M. A. Gorovsky, Y. S. Ziegler, B. Touchstone, W. A. Bradley, and R. G. Cook.** 1986. hv1 is an evolutionarily conserved H2A variant that is preferentially associated with active genes. J. Biol. Chem. **261:**1941–1948.
9. **Andrews, C. A., and S. A. Lesley.** 1998. Selection strategy for site-directed mutagenesis based on altered beta-lactamase specificity. BioTechniques **24:** 972–978.
10. **Ball, D. J., C. A. Slaughter, P. Hensley, and W. T. Garrard.** 1983. Amino acid sequence of the N-terminal domain of calf thymus histone H2A.Z. FEBS Lett. **154:**166–170.
11. **Bannister, A. J., P. Zegerman, J. F. Partridge, E. A. Miska, J. O. Thomas, R. C. Allshire, and T. Kouzarides.** 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature **410:**120–124.
12. **Berger, S.** 2002. Histone modifications in transcriptional regulation. Curr. Opin. Genet. Dev. **12:**142–148.
13. **Brownell, J. E., J. X. Zhou, T. Ranalli, R. Kobayashi, D. G. Edmondson, S. Y. Roth, and C. D. Allis.** 1996. *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell **84:**843–851.
14. **Callard, D., and L. Mazzolini.** 1997. Identification of proliferation-induced genes in *Arabidopsis thaliana*. Characterization of a new member of the highly evolutionarily conserved histone H2A.F/Z variant subfamily. Plant Physiol. **115:**1385–1395.
15. **Carr, A. M., S. M. Dorrington, J. Hindley, G. A. Phear, S. J. Aves, and P. Nurse.** 1994. Analysis of a histone H2A variant from fission yeast: evidence for a role in chromosome stability. Mol. Gen. Genet. **245:**628–635.
16. **Cassidy-Hanley, D., J. Bowen, J. Lee, E. S. Cole, L. A. VerPlank, J. Gaertig, M. A. Gorovsky, and P. J. Bruns.** 1997. Germline and somatic transformation of mating *Tetrahymena thermophila* by particle bombardment. Genetics **146:**135–147.
17. **Cheung, P., K. G. Tanner, W. L. Cheung, P. Sassone-Corsi, J. M. Denu, and C. D. Allis.** 2000. Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation. Mol. Cell **5:**905–915.
18. **Clarkson, M. J., J. R. E. Wells, F. Gibson, R. Saint, and D. J. Tremethick.** 1999. Regions of variant histone His2AvD required for *Drosophila* development. Nature **399:**694–697.
19. **Dalton, S., A. J. Robins, R. P. Harvey, and J. R. Wells.** 1989. Transcription from the intron-containing chicken histone H2A.F gene is not S-phase regulated. Nucleic Acids Res. **17:**1745–1756.
20. **Dhalluin, C., J. E. Carlson, L. Zeng, C. He, A. K. Aggarwal, and M. M. Zhou.** 1999. Structure and ligand of a histone acetyltransferase bromodomain. Nature **399:**491–496.
21. **Dou, Y., and M. A. Gorovsky.** 2002. Regulation of transcription by H1 phosphorylation in *Tetrahymena* is position independent and requires clustered sites. Proc. Natl. Acad. Sci. USA **99:**6142–6146.
22. **Dou, Y., and M. A. Gorovsky.** 2000. Phosphorylation of linker histone H1 regulates gene expression in vivo by creating a charge patch. Mol. Cell **6:**225–231.
23. **Edmondson, D. G., M. M. Smith, and S. Y. Roth.** 1996. Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. Genes Dev. **10:**1247–1259.
24. **Ernst, S. G., H. Miller, C. A. Brenner, C. Nocente-McGrath, S. Francis, and R. McIsaac.** 1987. Characterization of a cDNA clone coding for a sea urchin histone H2A variant related to the H2A.F/Z histone protein in vertebrates. Nucleic Acids Res. **15:**4629–4644.
25. **Faast, R., V. Thonglairoam, T. C. Schulz, J. Beall, J. R. E. Wells, H. Taylor, K. Matthaei, P. D. Rathjen, D. J. Tremethick, and I. Lyons.** 2001. Histone variant H2A.Z is required for early mammalian development. Curr. Biol. **11:**1183–1187.
26. **Fan, J. Y., F. Gordon, K. Luger, J. C. Hansen, and D. J. Tremethick.** 2002. The essential histone variant H2A.Z regulates the equilibrium between different chromatin conformational states. Nat. Struct. Biol. **9:**172–176.
27. **Flinta, C., B. Persson, H. Jornvall, and G. von Heijne.** 1986. Sequence determinants of cytosolic N-terminal protein processing. Eur. J. Biochem. **154:**193–196.
28. **Forsberg, E. C., and E. H. Bresnick.** 2001. Histone acetylation beyond promoters: long-range acetylation patterns in the chromatin world. Bioessays **23:**820–830.
29. **Fusauchi, Y., and K. Iwai.** 1983. *Tetrahymena* histone H2A. Isolation and two variant sequences. J. Biochem. **93:**1487–1497.
30. **Fusauchi, Y., and K. Iwai.** 1984. *Tetrahymena* histone H2A. Acetylation in the N-terminal sequence and phosphorylation in the C-terminal sequence. J. Biochem. (Tokyo) **95:**147–154.
31. **Gaertig, J., L. Gu, B. Hai, and M. A. Gorovsky.** 1994. High frequency vector-mediated transformation and gene replacement in *Tetrahymena*. Nucleic Acids Res. **22:**5391–5398.
32. **Gorovsky, M. A., M. C. Yao, J. B. Keevert, and G. L. Pleger.** 1975. Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods Cell Biol. **9:**311–327.
33. **Goto, H., Y. Tomono, K. Ajiro, H. Kosako, M. Fujita, M. Sakurai, K. Okawa, A. Iwamatsu, T. Okigaki, T. Takahashi, and M. Inagaki.** 1999. Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. J. Biol. Chem. **274:**25543–25549.
34. **Grant, P. A.** 5 April 2001, posting date. A tale of histone modifications. Genome Biol. **2:**0003.1–0003.6. [Online.] http://genomebiology.com.
35. **Hai, B., and M. A. Gorovsky.** 1997. Germ-line knockout heterokaryons of an essential α-tubulin gene enable high-frequency gene replacement and a test of gene transfer from somatic to germ-line in *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA **94:**1310–1315.
36. **Harvey, R. P., J. A. Whiting, L. S. Coles, P. A. Krieg, and J. R. Wells.** 1983. H2A.F: an extremely variant histone H2A sequence expressed in the chicken embryo. Proc. Natl. Acad. Sci. USA **80:**2819–2823.
37. **Hayes, J. J.** 2002. Changing chromatin from the inside. Nat. Struct. Biol. **9:**161–163.
38. **Isenberg, I.** 1979. Histones. Annu. Rev. Biochem. **48:**159–191.
39. **Jackson, J. D., V. T. Falciano, and M. A. Gorovsky.** 1996. A likely histone H2A.F/Z variant in *Saccharomyces cerevisiae*. Trends Biochem. Sci. **21:**466–467.
40. **Jackson, J. D., and M. A. Gorovsky.** 2000. Histone H2A.Z has a conserved function that is distinct from that of the major H2A sequence variants. Nucleic Acids Res. **28:**3811–3816.
41. **Jacobson, R. H., A. G. Ladurner, D. S. King, and R. Tjian.** 2000. Structure and function of a human TAFII250 double bromodomain module. Science **288:**1422–1425.
42. **Jenuwein, T., and C. D. Allis.** 2001. Translating the histone code. Science **293:**1074–1080.
43. **Kolodrubetz, D., M. O. Rykowski, and M. Grunstein.** 1982. Histone H2A subtypes associate interchangeably in vivo with histone H2B subtypes. Proc. Natl. Acad. Sci. USA **79:**7814–7818.
44. **Kornberg, R., and Y. Lorch.** 2002. Chromatin and transcription: where do we go from here? Curr. Opin. Genet. Dev. **12:**249–251.
45. **Kruger, W., C. L. Peterson, A. Sil, C. Coburn, G. Arents, E. N. Moudrianakis, and I. Herskowitz.** 1995. Amino acid substitutions in the structured domains of histones H3 and H4 partially relieve the requirement of the yeast SWI/SNF complex for transcription. Genes Dev. **9:**2770–2779.
46. **Kurumizaka, H., and A. P. Wolffe.** 1997. Sin mutations of histone H3: influence on nucleosome core structure and function. Mol. Cell. Biol. **17:** 6953–6969.
47. **Leach, T. J., M. Mazzeo, H. L. Chotkowski, J. P. Madigan, M. G. Wotring, and R. L. Glaser.** 2000. Histone H2A.Z is widely but nonrandomly distributed in chromosomes of *Drosophila melanogaster*. J. Biol. Chem. **275:**23267–23272.
48. **Liu, X., J. Bowen, and M. A. Gorovsky.** 1996. Either of the major H2A genes but not an evolutionarily conserved H2A.F/Z variant of *Tetrahymena ther-*

mophila can function as the sole H2A gene in the yeast *Saccharomyces cerevisiae*. Mol. Cell. Biol. **16:**2878–2887.

49. **Liu, X., and M. A. Gorovsky.** 1996. Cloning and characterization of the major histone H2A genes completes the cloning and sequencing of known histone genes of *Tetrahymena thermophila*. Nucleic Acids Res. **24:**3023–3030.
50. **Liu, X., B. Li, and M. A. Gorovsky.** 1996. Essential and nonessential histone H2A variants in *Tetrahymena thermophila.* Mol. Cell. Biol. **16:**4305–4311.
51. **Lo, W. S., R. C. Trievel, J. R. Rojas, L. Duggan, J. Y. Hsu, C. D. Allis, R. Marmorstein, and S. L. Berger.** 2000. Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14. Mol. Cell **5:**917–926.
52. **Luger, K., and T. J. Richmond.** 1998. The histone tails of the nucleosome. Curr. Opin. Genet. Dev. **8:**140–146.
53. **Megee, P. C., B. A. Morgan, B. A. Mittman, and M. M. Smith.** 1990. Genetic analysis of histone H4: essential role of lysines subject to reversible acetylation. Science **247:**841–845.
54. **Ng, H. H., and A. Bird.** 2000. Histone deacetylases: silencers for hire. Trends Biochem. Sci. **25:**121–126.
55. **Nielsen, P. R., D. Nietlispach, H. R. Mott, J. Callaghan, A. Bannister, T. Kouzarides, A. G. Murzin, N. V. Murzina, and E. D. Laue.** 2002. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature **416:**103–107.
56. **Orias, E., and P. J. Bruns.** 1976. Induction and isolation of mutants in *Tetrahymena.* Methods Cell Biol. **13:**247–282.
57. **Persson, B., C. Flinta, G. von Heijne, and H. Jornvall.** 1985. Structures of N-terminally acetylated proteins. Eur. J. Biochem. **152:**523–527.
58. **Pinto, I., and F. Winston.** 2000. Histone H2A is required for normal centromere function in *Saccharomyces cerevisiae*. EMBO J. **19:**1598–1612.
59. **Polevoda, B., J. Norbeck, H. Takakura, A. Blomberg, and F. Sherman.** 1999. Identification and specificities of N-terminal acetyltransferases from *Saccharomyces cerevisiae*. EMBO J. **18:**6155–6168.
60. **Recht, J., and M. A. Osley.** 1999. Mutations in both the structured domain and N-terminus of histone H2B bypass the requirement for Swi-Snf in yeast. EMBO J. **18:**229–240.
61. **Redon, C., D. Pilch, E. Rogakou, O. Sedelnikova, K. Newrock, and W. Bonner.** 2002. Histone H2A variants H2AX and H2AZ. Curr. Opin. Genet. Dev. **12:**162–169.
62. **Ren, Q., and M. A. Gorovsky.** 2001. H2A.Z acetylation modulates an essential charge patch. Mol. Cell **7:**1329–1335.
63. **Ren, Q. H., and T. J. Tong.** 1997. Histone acetylation and its roles in transcriptional regulation. Prog. Biochem. Biophys. **24:**309–312.
64. **Roth, S. Y., J. M. Denu, and C. D. Allis.** 2001. Histone acetyltransferases. Annu. Rev. Biochem. **70:**81–120.
65. **Santisteban, M. S., T. Kalashnikova, and M. M. Smith.** 2000. Histone H2A.Z regulates transcription and is partially redundant with nucleosome remodeling complexes. Cell **103:**411–422.
66. **Schuster, T., M. Han, and M. Grunstein.** 1986. Yeast histone H2A and H2B amino termini have interchangeable functions. Cell **45:**445–451.
67. **Shen, X., L. Yu, J. W. Weir, and M. A. Gorovsky.** 1995. Linker histones are not essential and affect chromatin condensation in vivo. Cell **82:**47–56.
68. **Stargell, L. A., J. Bowen, C. A. Dadd, P. C. Dedon, M. Davis, R. G. Cook, C. D. Allis, and M. A. Gorovsky.** 1993. Temporal and spatial association of histone H2A variant hv1 with transcriptionally competent chromatin during nuclear development in *Tetrahymena thermophila*. Genes Dev. **7:**2641–2651.
69. **Strahl, B. D., and C. D. Allis.** 2000. The language of covalent histone modifications. Nature **403:**41–45.
70. **Strahl, B. D., S. D. Briggs, C. J. Brame, J. A. Caldwell, S. S. Koh, H. Ma, R. G. Cook, J. Shabanowitz, D. F. Hunt, M. R. Stallcup, and C. D. Allis.** 2001. Methylation of histone H4 at arginine 3 occurs in vivo and is mediated by the nuclear receptor coactivator PRMT1. Curr. Biol. **11:**996–1000.
71. **Suto, R. K., M. J. Clarkson, D. J. Tremethick, and K. Luger.** 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. Nat. Struct. Biol. **7:**1121–1124.
72. **Taunton, J., C. A. Hassig, and S. L. Schreiber.** 1996. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science **272:**408–411.
73. **Thatcher, T. H., and M. A. Gorovsky.** 1994. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res. **22:**174–179.
74. **Th'ng, J. P.** 2001. Histone modifications and apoptosis: cause or consequence? Biochem. Cell Biol. **79:**305–311.
75. **Tse, C., T. Sera, A. P. Wolffe, and J. C. Hansen.** 1998. Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. Mol. Cell. Biol. **18:**4629–4638.
76. **Turner, B. M.** 2000. Histone acetylation and an epigenetic code. Bioessays **22:**836–845.
77. **Usachenko, S. I., S. G. Bavykin, I. M. Gavin, and E. M. Bradbury.** 1994. Rearrangement of the histone H2A C-terminal domain in the nucleosome. Proc. Natl. Acad. Sci. USA **91:**6845–6849.
78. **Van Daal, A., and S. C. R. Elgin.** 1992. A histone variant, H2AvD, is essential in *Drosophila melanogaster.* Mol. Biol. Cell **3:**593–602.
79. **Van Daal, A., E. M. White, M. A. Gorovsky, and S. C. R. Elgin.** 1988. *Drosophila* has a single copy of the gene encoding a highly conserved histone H2A variant of the H2A.F/Z type. Nucleic Acids Res. **16:**7487–7498.
80. **Verreault, A., P. D. Kaufman, R. Kobayashi, and B. Stillman.** 1996. Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4. Cell **87:**95–104.
81. **Wang, H. B., Z. Q. Huang, L. Xia, Q. Feng, H. Erdjument-Bromage, B. D. Strahl, S. D. Briggs, C. D. Allis, J. M. Wong, P. Tempst, and Y. Zhang.** 2001. Methylation of histone H4 at arginine 3 facilitating transcriptional activation by nuclear hormone receptor. Science **293:**853–857.
82. **Wechser, M. A., M. P. Kladde, J. A. Alfieri, and C. L. Peterson.** 1997. Effects of Sin^- versions of histone H4 on yeast chromatin structure and function. EMBO J. **16:**2086–2095.
83. **White, E. M., D. L. Shapiro, C. D. Allis, and M. A. Gorovsky.** 1988. Sequence and properties of the message encoding *Tetrahymena* hv1, a highly evolutionarily conserved histone H2A variant that is associated with active genes. Nucleic Acids Res. **16:**179–198.
84. **Winston, F., and C. D. Allis.** 1999. The bromodomain: a chromatin-targeting module? Nat. Struct. Biol. **6:**601–604.
85. **Wolffe, A.** 1998. Chromatin, 3rd ed. Academic Press, London, United Kingdom.
86. **Wolffe, A. P., and D. Guschin.** 2000. Chromatin structural features and targets that regulate transcription. J. Struct. Biol. **129:**102–122.
87. **Wolffe, A. P., and J. J. Hayes.** 1999. Chromatin disruption and modification. Nucleic Acids Res. **27:**711–720.
88. **Wolffe, A. P., and D. Pruss.** 1996. Deviant nucleosomes: the functional specialization of chromatin. Trends Genet. **12:**58–62.
89. **Wu, J. S., and M. Grunstein.** 2000. 25 years after the nucleosome model: chromatin modifications. Trends Biochem. Sci. **25:**619–623.
90. **Xia, L., B. Hai, Y. Gao, D. Burnette, R. Thazhath, J. Duan, M. H. Bre, N. Levilliers, M. A. Gorovsky, and J. Gaertig.** 2000. Polyglycylation of tubulin is essential and affects cell motility and division in *Tetrahymena thermophila*. J. Cell Biol. **149:**1097–1106.
91. **Zhang, Y., and D. Reinberg.** 2001. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. Genes Dev. **15:**2343–2360.

Molecular Cell, Vol. 7, 1329–1335, June, 2001, Copyright ©2001 by Cell Press

Histone H2A.Z Acetylation Modulates an Essential Charge Patch

Qinghu Ren and Martin A. Gorovsky[1]
Department of Biology
University of Rochester
Rochester, New York 14627

Summary

Histone H2A.Z is structurally and functionally distinct from the major H2As. To understand the function of H2A.Z acetylation, we performed a mutagenic analysis of the six acetylated lysines in the N-terminal tail of *Tetrahymena* H2A.Z. *Tetrahymena* cannot survive with arginines at all six sites. Retention of one acetylatable lysine is sufficient to provide the essential function of H2A.Z acetylation. This essential function can be mimicked by deleting the region encompassing all six sites, or by mutations that reduce the positive charge of the N terminus at the acetylation sites themselves, or at other sites in the tail. These properties argue that the essential function of H2A.Z acetylation is to modify a "charge patch" by reducing the charge of the tail.

Introduction

The nucleosome core particle, consisting of ~146 bp DNA wrapped around an octamer of two of each of four conserved core histones, H2A, H2B, H3 and H4, is a basic unit of chromatin structure in eukaryotic cells (Wolffe, 1998). Core histones contain a highly structured C-terminal domain, important for histone-histone interactions and nucleosome core formation, and a highly charged, structurally undefined, N-terminal tail domain that extends from the core. Tails appear to be important for intranucleosomal DNA-histone interactions, for internucleosomal histone-DNA and histone-histone interactions, and for interactions with nonhistone proteins (Luger and Richmond, 1998; Hansen et al., 1998).

The conserved seemingly constant structure of nucleosomes is subject to processes that create heterogeneity associated with different structural and functional states of chromatin. Major factors producing this heterogeneity are enzymatic posttranslational modifications of the core histone tails, including acetylation, phosphorylation, and methylation (Wolffe and Hayes, 1999). These modifications can function in histone deposition, nucleosome assembly, chromosome condensation, and transcriptional regulation (Turner, 1995; Roth and Allis, 1996; Ura et al., 1997; Wei et al., 1999; Cheung, P. et al., 2000; Rea et al., 2000).

Acetylation of the ϵ-amino group of lysine is the best-characterized core histone modification. This modification, which eliminates the positive charge on the acetylated lysines, has been linked to transcriptional activation (Cheung, W.L. et al., 2000; Bannister and Miska, 2000; Kouzarides, 2000; Mizzen and Allis, 1998; Kuo and Allis, 1998). The discoveries that transcriptional (co-)activators possess histone acetyltransferase (HAT) activity (Brownell et al., 1996; Wade et al., 1997; Kouzarides, 1999), while (co-)repressors contain histone deacetylase (HDAC) activity (Taunton et al., 1996; Ng and Bird, 2000), provided a molecular basis for a long-standing correlation between histone acetylation and transcriptional activation. Similarly, the finding that a factor (CAF-1) required for chromatin assembly in vitro is associated with acetylated histones and histone binding proteins, like those that are associated with histone acetyltransferase, also provided a molecular underpinning for another long-standing correlation between histone acetylation and chromatin replication (Verreault et al., 1996).

Two mechanisms have been proposed to explain how charge-altering modifications might act to change histone-DNA or histone-protein interactions. These modifications, acting at specific sites, either alone, in combination, or sequentially on one or more histone tails, could form a complex "histone code" which specifies unique chromatin functions (Strahl and Allis, 2000; Turner, 2000; Paro, 2000). The second mechanism by which charge-altering modifications could affect chromatin function is by modifying the charge of a protein domain. In *Tetrahymena*, this "charge patch" mechanism has been shown to apply to regulation of the expression of specific genes by phosphorylation of the H1 histone associated with the DNA that links nucleosomal cores (Dou and Gorovsky, 2000).

Another factor that contributes to chromatin functional heterogeneity is the existence of histone variants. The best-studied core histone variant, H2A.Z, has been found in enough diverse organisms (Santisteban et al., 2000) to suggest that it is a universal component of eukaryotic chromatin. Major H2As and H2A.Z diverged early in eukaryotic evolution and H2A.Z proteins show even less evolutionary divergence than the major H2As (Thatcher et al., 1994). Although it comprises only 5%–10% of total H2A (West and Bonner, 1980; Wu et al., 1982), H2A.Z is essential in *Drosophila* (Van Daal and Elgin, 1992), *Tetrahymena* (Liu et al., 1996), and mouse (see Clarkson et al., 1999), and affects growth in yeasts (Carr et al., 1994; Jackson and Gorovsky, 2000; Santisteban et al., 2000). H2A.Z can function both as a positive regulator of the transcription of yeast genes (Santisteban et al., 2000) and in silencing at HMR locus in yeast (Dhillon and Kamakaka, 2000).

We have developed the ciliated protozoan *Tetrahymena thermophila* as a model system for studying the in vivo functions of H2A.Z acetylation. In *Tetrahymena*, each cell contains a germline micronucleus and a somatic macronucleus that differ in structure and function (Gorovsky, 1980). During vegetative growth, the diploid micronucleus divides mitotically and is transcriptionally inert while the polyploid macronucleus divides amitotically and is transcriptionally active. Macro- and micronuclei of vegetative cells have a common origin during conjugation, the sexual stage of the life cycle (Bruns, 1986). There is circumstantial evidence that *Tetrahymena* H2A.Z functions in establishing a transcriptionally

[1] Correspondence: goro@mail.rochester.edu

competent state of chromatin. It is present in macronuclei but not in micronuclei during growth or starvation (Allis et al., 1980). Strikingly, H2A.Z appears in micronuclei during early stages of conjugation, when micronuclei become transcriptionally active (Stargell et al., 1993). Both macronuclear H2A.Z and the H2A.Z which appears in micronuclei early in conjugation are highly acetylated (Stargell et al., 1993).

In this study, we changed the H2A.Z acetylation sites either from lysine to arginine, which conserves the positive charge of lysine, but cannot be acetylated, or we changed lysine to glutamine, which resembles acetylated lysine in charge and structure. We found that *Tetrahymena* cannot survive with arginine replacements at all acetylation sites of H2A.Z. However, cells containing five arginine replacements and either a single lysine or a single glutamine are viable. We also show that cells containing six arginines plus mutations that reduce the positive charge at (unacetylatable) sites in the N terminus are also viable. These studies indicate that H2A.Z acetylation modulates a charge patch with an essential function.

Results

Distinguishing Between a "Charge Patch" and a "Histone Code"

Our results are best understood in the context of the distinguishing features of a charge patch. If the acetylated region functions as a charge patch, it is the overall charge of the modified region that is important. Therefore, any mutation that reduces the charge, either at the acetylation site or at nearby residues, should phenotypically mimic acetylation. Another feature of the only charge patch characterized to date (Dou and Gorovsky, 2000) is that reducing the positive charge of the highly basic histone H1 had the same phenotypic effects as deleting the histone, suggesting that charge reduction mimicked removal of the histone from the DNA. While such an effect is not a necessary feature of a charge patch, it may be a reasonable outcome after reducing the positive charge of a histone domain that is associated with DNA, if exposure of the DNA is the normal function of the charge-reducing modification. In contrast, a defining feature of a site-specific histone code is that different acetylation sites have qualitatively distinguishable functions specific to each acetylation site. This leads to the prediction that only amino acid replacements that closely resemble acetylated lysines, and that occur at the normally acetylated sites, should mimic the function of acetylation of these residues.

H2A.Z Acetylation Has an Essential Function

Changing the acetylation sites on *Tetrahymena* H2A.Z (Allis et al., 1986) from lysine to arginine conserves the net positive charge of lysine but prevents charge neutralization by acetylation (Megee et al., 1990). Changing all six acetylatable lysines to arginines failed to yield any transformants in three experiments in which parallel transformations with the wild-type *HTA3* gene yielded numerous transformants, suggesting H2A.Z acetylation has an essential function in *Tetrahymena*. Leaving a single lysine at any of the acetylatable sites resulted in viable cells that grew slowly at 30°C, and had variable sizes and irregular surfaces (Figure 2A and data not shown). Small but reproducible differences in growth rates and temperature sensitivities were observed among the strains containing a single acetylation site at different positions (data not shown). Leaving two of the six acetylation sites (changing four lysines to arginines) yielded viable transformants that grew normally at 30°C.

To determine the acetylation status of H2A.Z, histones from strains rescued with wild-type or mutated *HTA3* genes were separated on acid urea gels to separate histones by both molecular weight and charge. To differentiate H2A.Z from other *Tetrahymena* histones that comigrate with it, gels were immunoblotted and stained with an antibody for *Tetrahymena* H2A.Z (Stargell et al., 1993). Because *Tetrahymena* H2A.Z is also phosphorylated (Allis and Gorovsky, 1981), the effects of phosphorylation were eliminated by pretreating proteins with λ protein phosphatase.

With six acetylation sites, up to seven separable, phosphatase-resistant isotypes might be expected in wild-type cells. Strains rescued by the wild-type gene showed five or six isoforms after phosphatase treatment (Figure 2D, lane 2), representing unmodified H2A.Z (bottom band) and isoforms containing one to five acetyl groups. We could not detect a band containing six acetates, suggesting it either comprises only a tiny fraction of the total H2A.Z, or is highly susceptible to deacetylation during purification. As expected, H2A.Z from mutants containing only two acetylation sites showed three phosphatase-resistant isoforms, and those containing only a single acetylation site contained only two isoforms (Figure 2D, lanes 4 and 6). This assay was used to characterize the acetylation status of all viable H2A.Z mutants. In every case, the number of isoforms observed was one more than the number of unmutated lysines at the sites of acetylation. These observations demonstrate that the biochemically identified lysines at positions 4, 7, 10, 13, 16, and 21 (Allis et al., 1986) are the only acetylated residues in H2A.Z and that all of the acetylation site mutations have the expected level of H2A.Z acetylation.

Glutamine Partially Mimics H2A.Z Acetylation

A series of *HTA3* constructs was made changing different numbers (1–6) of arginines on the *HTA3* RRRRRR plasmid to glutamines, a neutral amino acid that resembles acetylated lysine. All of these produced viable progeny (Figure 2B). The H2A.Z modification states of two of the mutants, *HTA3* QQQQQR and *HTA3* QQQQQQ were tested by immunoblotting (data not shown). As expected, neither showed any phosphatase-resistant heterogeneity because they did not contain any acetylatable lysine that could exist in either the acetylated or the unacetylated state. All of the viable mutants in Figure 2B showed slow growth rates, variable sizes, and irregular surfaces (data not shown), indicating that glutamine can mimic some, but not all of the functions of acetylatable lysines on H2A.Z. As with the single lysine mutants, strains having the same number of glutamine replacements, but at different sites (e.g., *HTA3* QRRRRR and *HTA3* RRRRRQ), showed differences in growth rates and/or temperature sensitivities (data not shown).

Mutations at Nonacetylatable Sites that Reduce the Net Charge of the H2A.Z Amino Terminus Are Viable

Glutamine could mimic acetylation either by mimicking the charge neutralization caused by acetylation and/or by structurally mimicking acetylated lysine. To distinguish these possibilities, the N-terminal tail of H2A.Z was altered to reduce its charge by mutating sites that normally are not acetylated and by using residues that do not resemble acetyl-lysine. First, the charge-reducing effect of a single acetylated lysine was mimicked by using a negatively charged aspartic acid (D) or glutamic acid (E) to replace a neutral residue adjacent to one of the arginines of the *HTA3* RRRRRR mutant gene. Addition of either of these two amino acids near the beginning or the end of the tail resulted in viable progeny (Figure 2C). Second, two nonacetylated lysines (K23 and K24) near the end of the tail were changed to glutamines. This too yielded viable progeny (Figure 2C). H2A.Z isolated from the *HTA3* DRRRRRR mutant showed a single phosphatase-resistant isoform at approximately the position of monoacetylated H2A.Z. The slight mobility difference from monoacetylated H2A.Z is likely due to a small difference in molecular weight (Figure 2D, lane 10). Thus, the essential function of H2A.Z acetylation can be mimicked by reducing the charge of the amino terminus at sites other than those normally acetylated and by using residues that do not mimic the structure of acetyl-lysine. Because these are defining properties of modifications that create or modify a charge patch, we conclude that the essential function of acetylation of H2A.Z is to alter the charge of the N terminus.

Deletion of the Acetylated Region Results in Viable Progeny

Deletion of H2A.Z is lethal in *Tetrahymena* (Liu et al., 1996). However, we reasoned that, inasmuch as acetylation of H2A.Z is restricted to the N terminus, and N termini are not required to form nucleosome cores, it might be possible to delete the N terminus of H2A.Z. To test this, residues 4–24 of the N terminus were deleted, removing all of the positive charges in this region. This mutation yielded viable progeny (Figure 2C). As expected, *HTA3* DEL(4–24) shows only a single, smaller, phosphatase-resistant isoform (Figure 2D, lane 8). Thus, with respect to viability, deletion of the majority of the H2A.Z tail has the same phenotype as the presence of acetylatable lysines, or of glutamines that mimic acetylation, or of charge reducing alterations. The simplest explanation for these observations is that the essential function of lysine acetylation in the N terminus is nonspecific charge neutralization, which weakens or abolishes the association of the tail with DNA. The finding that deletion of the tail results in viable cells, while elimination of acetylation by arginine replacements is lethal, also argues that the essential function of acetylation is not the site-specific recognition of acetyl-lysines.

Discussion

The important conclusions to be drawn from this work are: (1) that acetylation of a specific, quantitatively minor histone variant has an essential function that cannot be performed by acetylation of other core histones; (2) that the essential function of H2A.Z acetylation is likely to be the reduction of the highly positive charge of the N-terminal tail of this histone; and (3) that a likely structural consequence of H2A.Z acetylation in vivo is weakening of the association of the N-terminal region of this histone with chromatin.

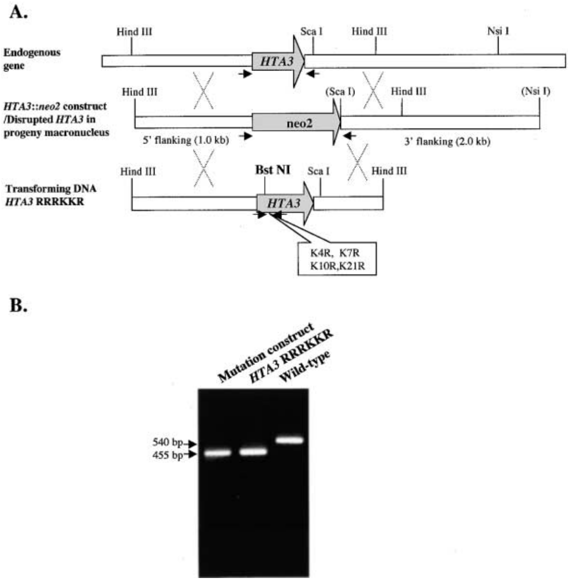
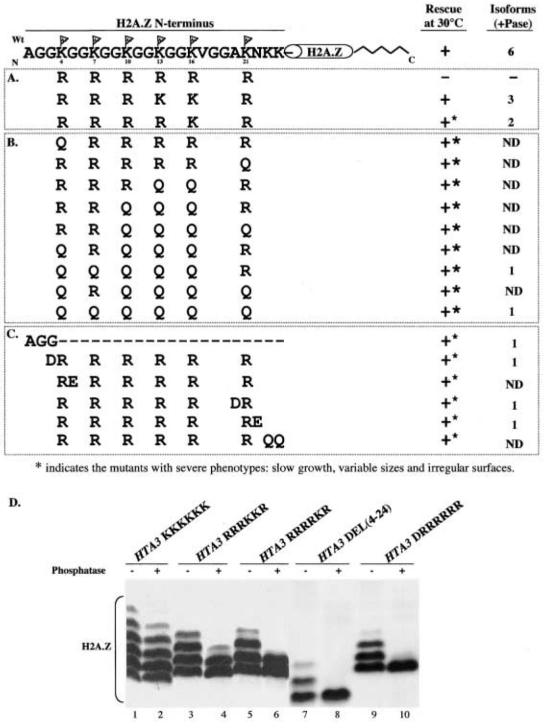


Figure 1. PCR Identification of Knockout Heterokaryons and Rescued Mutant Progeny

(A) The macronuclear genomic *HTA3* gene is shown as a 3.5 kb Hind III-Nsi I fragment containing the *HTA3*-coding region. The *HTA3*::*neo2* KO construct is shown as a *neo2* cassette with 1.0 kb 5′ and 2.0 kb 3′ flanking *HTA3* sequences. The *HTA3* RRRKKR construct is shown as 2.0 kb Hind III-Hind III fragment containing the specific mutations which also introduce a new BstN I restriction site.

(B) The *HTA3*-coding regions from wild-type and *HTA3* RRRKKR transformants were PCR amplified using the primers (arrows) shown in (A). The PCR products were run on an agarose gel. The coding region of the *HTA3* construct is ∼85 bp shorter than the endogenous gene because it lacks the second intron. The *HTA3* RRRKKR transformants give PCR products whose size (455 bp) is the same as that of the transforming mutation construct, while products from the wild-type strain have the size of the endogenous wild-type *HTA3* (540 bp).

The conclusion that acetylation of H2A.Z has an essential function is derived from observations that replacement of all six acetylatable lysines with unacetylatable arginines is lethal, while leaving even a single acetylatable lysine is not. An alternative interpretation that arginine cannot perform some other function of lysine, seems less likely for a number of reasons. First, a single glutamine, whose charge and structure resembles acetyl-lysine more than lysine, also suffices for viability. Second, the existence of six arginines is not itself lethal, because mutants containing six arginines plus charge-reducing mutations at other sites in the tail are viable. Third, the acetylatable lysines themselves perform no essential function in their unmodified form, because they all can be deleted and the cells live. These results argue strongly that it is the ability of these lysines to be acetylated in vivo that is responsible for their essential function.



	H2A.Z N-terminus	Rescue at 30°C	Isoforms (+Pase)
Wt	N AGGKGGKGGKGGKGGKVGGAKNKK– H2A.Z C	+	6
	4 7 10 13 16 21		
A.	R R R R R R	–	–
	R R R K K R	+	3
	R R R R K R	+*	2
B.	Q R R R R R	+*	ND
	R R R R R Q	+*	ND
	R R R Q Q R	+*	ND
	R R Q Q Q R	+*	ND
	R R Q Q Q Q	+*	ND
	Q R Q Q Q R	+*	ND
	Q Q Q Q Q R	+*	1
	Q R Q Q Q Q	+*	ND
	Q Q Q Q Q Q	+*	1
C.	AGG--------------------	+*	1
	DR R R R R R	+*	1
	RE R R R R R	+*	ND
	R R R R R DR	+*	1
	R R R R R RE	+*	1
	R R R R R R QQ	+*	ND

* indicates the mutants with severe phenotypes: slow growth, variable sizes and irregular surfaces.

Figure 2. H2A.Z Acetylation Modulates an Essential "Charge Patch"

The sequence of the N-terminal tail of *Tetrahymena* H2A.Z is shown with the six acetylated lysines indicated by flags.

(A) Changing all six acetylatable lysines to arginines failed to yield any transformants in three independent experiments in which positive controls (the wild-type gene) gave numerous transformants. Cells with two acetylation sites grew normally at 30°C. A single acetylation site was sufficient to maintain viability but produces cells that grew slowly even at 30°C. An asterisk indicates the mutants with severe phenotypes: slow growth, variable sizes, and irregular surfaces.

(B) *HTA3* knockout heterokaryons were rescued by changing one to six arginines to glutamine. All these mutants grew slowly, were variable size, and exhibited an irregular surface.

(C) The lethal phenotype of the RRRRRR mutation was rescued by addition of negatively charged amino acids, either aspartic acid/D or glutamic acid/E adjacent to either the first or the last acetylation site. Changing two nonacetylatable lysines, K23 and K24, into glutamines also rescued *HTA3* RRRRRR. All of these mutants grew slowly at 30°C. A deletion in the N terminus (4–24), which removes the whole acetylation region, also yielded viable progeny that grew slowly at 30°C.

(D) Nuclear histones from strains rescued with wild-type or mutated *HTA3* genes were phosphatase-treated and Western blotted as described in Experimental Procedures. Five or six isoforms can be seen after phosphatase treatment of histones from wild-type (KKKKKK) cells (lane 2). As expected, cells containing *HTA3* genes encoding two (RRRKKR) or one (RRRRKR) acetylation sites show three and two isoforms (lanes 4 and 6) respectively. *HTA3* DEL(4–24) showed only a single phosphatase-resistant isoform that migrated faster than unacetylated H2A.Z (lane 8), while *HTA3* DRRRRRR showed a single isoform which differed only slightly in mobility from monoacetylated H2A.Z (lane 10), probably owing to a small difference in molecular weight.

The conclusion that the essential function of H2A.Z acetylation in *Tetrahymena* is likely to be due to a reduction in the charge of the highly positive N-terminal tail is based on the observations that all mutations examined that reduce the charge of the tail are viable, even if they occur at sites that are not acetylated in vivo. The most telling of these mutations are those in which all the acetylatable lysines were changed to arginine, while glutamate or aspartate residues were used to replace neutral residues at sites adjacent to the first or last positions that are normally acetylated in wild-type cells. These mutant H2A.Zs do not contain any residues that resemble acetyl-lysine and do not have a reduced charge at the normal acetylation sites. These mutations also make it highly unlikely that mutations containing only arginines at the acetylation site are functioning as dominant negatives. The most obvious way in which these mutant H2A.Zs resemble acetylatable H2A.Zs is in the reduced charge of the tail. We conclude, therefore, that charge reduction of the positively charged N-terminal tail is the essential function of H2A.Z acetylation.

It seems rather remarkable that the presence or absence of a single acetylation site in only a small fraction of the nucleosomes can result in the difference between life and death. Assuming that *Tetrahymena* H2A.Z-containing nucleosomes in vivo contain two molecules of H2A.Z and the normal complement of other core histones, there are at least 30 acetylatable lysines in a *Tetrahymena* H2A.Z nucleosome (for a summary of the acetylation sites in the other core histones and references, see Gorovsky, 1986). It is hard to envision how changing only a single positive charge of the H2A.Z tail can have such a dramatic consequence, especially in light of studies of nucleosomal arrays, the form of chromatin in vitro most likely to reflect the chromatin in vivo. In these arrays, no major changes in folding or transcriptional activity were detected when only ~6 of the 26 sites in the core histone tails were acetylated (Tse et al., 1998). However, when 12 of the sites were acetylated, higher-order folding was completely prevented and transcription increased.

The essential function of a single acetylation site in the H2A.Z N-terminal tail is also surprising in light of studies showing that the only region of *Drosophila* H2A that cannot provide the essential developmental function of H2A.Z resides in the C-terminal α-helix (Clarkson et al., 1999). The difference between these studies and ours indicate either that there are fundamental differences between the essential domains of *Drosophila* and *Tetrahymena* H2A.Z, or that acetylation sites on the major H2A can replace those on H2A.Z when associated with the H2A.Z C-terminal helix in a fusion protein. We think the latter alternative is more likely, given the high level of sequence and functional conservation among H2A.Zs of highly divergent organisms (Thatcher and Gorovsky, 1994; Jackson et al., 1996; Jackson and Gorovsky, 2000) and the lack of site-specificity for the essential function of charge neutralization in the H2A.Z N terminus demonstrated in our studies.

The conclusion that the essential function of H2A.Z acetylation acts by dissociating the tail from charge-dependent contacts with DNA or with other proteins is supported by the observation that a complete deletion of the acetylated region is viable. This is similar to the phosphorylation-induced charge patch in which phosphorylation mimics the complete removal of *Tetrahymena* H1 with regard to the expression of (at least) two genes (Dou and Gorovsky, 2000). However, it seems unlikely that neutralizing any one of the six acetylatable lysines, which reduces the positive charges of the H2A.Z N-terminal tail from +8 to +7, can completely dissociate the tail either from DNA or from an acidic patch on another protein. It seems more likely that acetylation lowers the affinity of the tail for the DNA or protein with which it interacts, facilitating access of competing factors that are essential. Such a mechanism is supported by the observation that cells with two acetylation sites grow more normally than those with one, suggesting that additional weakening of this association further increases the ability of the essential factors to interact with chromatin. The observation that extensive acetylation can greatly weaken, but does not abolish, the interaction of an H4 N-terminal peptide with DNA in vitro also supports this model (Norton et. al., 1989).

While our studies demonstrate that reducing the charge of the N-terminal tail plays an essential role in the function of the conserved H2A.Z variant, they do not rule out the possibility that additional, nonessential functions of H2A.Z can be modulated by the existence of a histone code that distinguishes the precise function of the different acetylation sites. This remains one of a number of possible mechanisms that could explain the differences we observed in the growth properties of identical amino acid replacements at different acetylation sites.

Experimental Procedures

Strains, Culture, and Conjugation

Tetrahymena thermophila strains CU428, CU427, and B2086 were kindly provided by P.J. Bruns (Cornell University). *Tetrahymena* cells were grown in SPP medium containing 1% Proteose Peptone (1×SPP) (Gorovsky et al., 1975). For conjugation, two strains of different mating types were washed, starved (16–24 hr, 30°C), and mated in 10mM Tris-HCl (pH 7.5) as described by Allis and Dennison (1982).

Plasmid Construction

Plasmid p4T2-1, a pBluescript KS(+) derivative, contains a copy of the *neo2* gene cassette (Gaertig et al., 1994). To construct the *HTA3* gene-knockout construct, a 1.1 kb fragment of the *HTA3* 5′ flanking sequence (from a Hind III site to the start codon) was PCR amplified from phv1, a pBluescript SK(−) derivative, and inserted into the 5′ polylinker region (between Kpn I and EcoR V) of p4T21. A 2.0 kb Sca I-Nsi I fragment of the *HTA3* 3′ flanking sequence was amplified from *Tetrahymena* genomic DNA and inserted into the Sma I site of the 3′ polylinker region. The final *HTA3*::*neo2* construct, pQR22 (Figure 1A), was released by digestion with Kpn I and Sac I.

Site-Directed Mutagenesis

Oligonucleotide-directed, double strand mutagenesis was performed as described (Andrews and Lesley, 1998) on phv1, which contains a copy of the wild-type *HTA3* gene with a deletion of the second intron (85 bp). All six lysines were changed in a single mutagenesis into arginines to create phv1 RRRRRR. All other mutation constructs were derived from this plasmid. In some cases, a silent mutation was introduced to generate a restriction site to monitor transformation. All mutated genes were sequenced by automatic sequencing system (ABI Prism) and released by digestion with Hind III before being introduced into knockout heterokaryons.

Knockout Heterokaryons and Transformation

The *HTA3* gene encoding H2A.Z was disrupted using pQR22 (Figure 1A) by replacing the entire coding region with a *neo2* cassette, which confers paromomycin (pm) resistance when expressed in macronuclei (Gaertig et al., 1994). A 4.5 kb Hind III-Nsi I fragment containing the disrupted gene was introduced into 2.5 hr conjugating CU428 and B2086 cells using the Biolistic PDS-1000/He Particle delivery system (Bio-Rad), as described (Cassidy-Hanley et al., 1997). Knockout heterokaryons containing disrupted *HTA3* genes in their micronuclei and wild-type *HTA3* genes in their macronuclei (*HTA3*-G311A1 and *HTA3*-G304A1) were created as described (Hai and Gorovsky, 1997). When these heterokaryons conjugate, the old pm-sensitive macronuclei are replaced by new ones produced by meiosis, fertilization, and mitotic division of the cells' micronuclei. Consequently, the *neo2* gene that disrupts the *HTA3* gene is expressed, allowing drug selection for successful mating. However, because H2A.Z is essential in *Tetrahymena* (Liu et al., 1996), and the new macronucleus contains only disrupted *HTA3* genes, the progeny of this mating will die unless they are transformed with an *HTA3* gene that functions well enough to support growth.

Successful creation of *HTA3* germline knockout heterokaryons was demonstrated by the fact that no viable progeny were obtained when *HTA3* heterokaryons of two different mating types were mated and that progeny could be rescued by transforming with a wild-type copy of *HTA3*. In addition, the physical structure of the disrupted *HTA3* in the micronucleus of the heterokaryons was examined by mating knockout heterokaryons with wild-type CU427 cells, and selecting for retention of *HTA3*::*neo2* by increasing the pm concentration to 10 mg/ml. When genomic DNA was analyzed by PCR using primers specific for the 5′ and 3′ flanking sequences of *HTA3*, the disrupted *HTA3* gene was demonstrable in progeny cell macronuclei, indicating that the parental heterokaryons have the disrupted gene in their micronuclei. As expected, the heterozygous macronuclei of these progeny cells also have wild-type copies of *HTA3* gene as required to provide the essential H2A.Z functions (data not shown).

These knockout heterokaryon strains facilitate systematic mutagenesis studies on H2A.Z modification sites as illustrated in Figure 1B. A mutated form of the *HTA3* gene containing only two of the six acetylation sites (see Results for details) was introduced into mating *HTA3*-G311A1 and *HTA3*-G304A1 knockout heterokaryons at late stages (24 hr) by biolistic transformation (Cassidy-Hanley et al., 1997), and progeny were selected with pm at 120 μg/ml. Viable progeny were obtained, indicating it is a non-lethal mutation. When the *HTA3* coding region of the progeny was PCR amplified using the *HTA3* gene primers, the newly introduced, mutated *HTA3* gene is easily differentiated from wild-type because the second intron is missing. Finally, the genotype of all mutant strains was confirmed by sequencing the PCR products from genomic DNA of the transformed progeny.

Growth Analysis

Mutant strains, a control strain rescued with the wild-type *HTA3* and a wild-type CU428 strain were used in growth assays as described (Shen et al., 1995). Cells from each strain were inoculated into 50 ml 1×SPP medium at starting densities of 1×10^4 cells/ml. Cultures were grown at 30°C with vigorous shaking. Samples (100 μl) were counted at frequent intervals using a ZB1 Counter (Coulter Electronics, Incorporated). Doubling times were calculated using the linear portion of the logarithmic growth curves plotted using Cricket GraphIII (Computer Associates). To assay temperature sensitivity, strains were inoculated into 96-well plates at 1×10^4 cells/ml, serially diluted 2-fold, incubated at 40°C for 24–48 hr, and observed under the light microscope.

Histone Extraction, Electrophoresis, and Immunoblotting

Macronuclei were isolated by the method of Gorovsky et al. (1975). Histones were extracted as previously described (Glover et al., 1981). 25 μg nuclear histones were pretreated with λ protein phosphatase (New England Biolabs, Inc.) at 10 units/μl for 5 hr at 30°C. Phosphatase treated and untreated histones were separated on long acid urea polyacrylamide slab gels (15% acrylamide, 6M urea, 5% acetic acid) as described (Allis et al., 1980). The immunoblotting analyses with antibody to H2A.Z (anti-hv1-HPLC; 1:10,000) were performed as described (Stargell et al., 1993).

Acknowledgments

We are grateful to Josephine Bowen for technical assistance, and to Dr. Jeffrey Hayes, Dr. Cheeptip Benyajati, and Josephine Bowen for critically reading the manuscript. This work was supported by grant GM21793 from the NIH. This manuscript is dedicated to the memory of Alan Wolffe, whose many contributions to the chromatin field will be sorely missed.

Received November 20, 2000; revised April 3, 2001.

References

Allis, C.D., Glover, C.V.C., Bowen, J.K., and Gorovsky, M.A. (1980). Histone variants specific to the transcriptionally active, amitotically dividing macronucleus of the unicellular eukaryote, *Tetrahymena thermophila*. Cell *20*, 609–617.

Allis, C.D., and Gorovsky, M.A. (1981). Histone phosphorylation in macro- and micronuclei of *Tetrahymena thermophila*. Biochemistry *20*, 3828–3833.

Allis, C.D., and Dennison, D.K. (1982). Identification and purification of young macronuclear anlagen from conjugating cells of *Tetrahymena thermophila*. Dev. Biol. *93*, 519–533.

Allis, C.D., Richman, R., Gorovsky, M.A., Ziegler, Y.S., Touchstone, B., Bradley, W.A., and Cook, R.G. (1986). hv1 is an evolutionarily conserved H2A variant that is preferentially associated with active genes. J. Biol. Chem. *261*, 1941–1948.

Andrews, C.A., and Lesley, S.A. (1998). Selection strategy for site-directed mutagenesis based on altered beta-lactamase specificity. Biotechniques *24*, 972–974, 976, 978 passim.

Bannister, A.J., and Miska, E.A. (2000). Regulation of gene expression by transcription factor acetylation. Cell Mol. Life Sci. *57*, 1184–1192.

Brownell, J.E., Zhou, J.X., Ranalli, T., Kobayashi, R., Edmondson, D.G., Roth, S.Y., and Allis, C.D. (1996). *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843–851.

Bruns, P. (1986). Genetic organization of *Tetrahymena*. In The Molecular Biology of Ciliated Protozoa. J.G. Gall, ed. (Orlando, FL: Academic Press Inc.), pp. 27–44.

Carr, A.M., Dorrington, S.M., Hindley, J., Phear, G.A., Aves, S.J., and Nurse, P. (1994). Analysis of a histone H2A variant from fission yeast: evidence for a role in chromosome stability. Mol. Gen. Genet. *245*, 628–635.

Cassidy-Hanley, D., Bowen, J., Lee, J., Cole, E.S., VerPlank, L.A., Gaertig, J., Gorovsky, M.A., and Bruns, P.J. (1997). Germline and somatic transformation of mating *Tetrahymena thermophila* by particle bombardment. Genetics *146*, 135–147.

Cheung, P., Allis, C.D., and Sassone-Corsi, P. (2000). Signaling to chromatin through histone modifications. Cell *103*, 263–271.

Cheung, W.L., Briggs, S.D., and Allis, C.D. (2000). Acetylation and chromosomal functions. Curr. Opin. Cell Biol. *12*, 326–333.

Clarkson, M.J., Wells, J.R.E., Gibson, F., Saint, R., and Tremethick, D.J. (1999). Regions of variant histone His2AvD required for *Drosophila* development. Nature *399*, 694–697.

Dhillon, N., and Kamakaka, R.T. (2000). A Histone variant, Htz1p, and a Sir1p-like protein, Esc2p, mediate silencing at HMR. Mol. Cell *6*, 769–780.

Dou, Y., and Gorovsky, M.A. (2000). Phosphorylation of linker histone H1 regulates gene expression in vivo by creating a charge patch. Mol. Cell *6*, 225–231.

Gaertig, J., Gu, L., Hai, B., and Gorovsky, M.A. (1994). High frequency vector-mediated transformation and gene replacement in *Tetrahymena*. Nucleic Acids Res. *22*, 5391–5398.

Glover, C.V.C., Vavra, K.J., Guttman, S.D., and Gorovsky, M.A. (1981). Heat shock and deciliation induce phosphorylation of histone H1 in *Tetrahymena pyriformis*. Cell *23*, 73–77.

Gorovsky, M.A. (1980). Genome organization and reorganization in *Tetrahymena*. Ann. Rev. Genet. *14*, 203–239.

Gorovsky, M.A. (1986). Ciliate chromatin and histones. In The Molecular Biology of Ciliated Protozoa. J.G. Gall, ed. (Orlando: Academic Press, Inc.), pp. 227–261.

Gorovsky, M.A., Yao, M.-C., Keevert, J.B., and Pleger, G.L. (1975). Isolation of micro- and macronuclei of *Tetrahymena pyriformis*. Methods Cell Biol. *9*, 311–327.

Hai, B., and Gorovsky, M.A. (1997). Germ-line knockout heterokaryons of an essential α-tubulin gene enable high-frequency gene replacement and a test of gene transfer from somatic to germ-line nuclei in *Tetrahymena thermophila*. Proc. Natl. Acad. Sci. USA *94*, 1310–1315.

Hansen, J.C., Tse, C., and Wolffe, A.P. (1998). Structure and function of the core histone N-termini: more than meets the eye. Biochemistry *37*, 17637–17641.

Jackson, J.D., Falciano, V.T., and Gorovsky, M.A. (1996). A likely histone H2A.F/Z variant in *Saccharomyces cerevisiae*. Trends Biochem. Sci. *21*, 466–467.

Jackson, J.D., and Gorovsky, M.A. (2000). Histone H2A.Z has a conserved function that is distinct from that of the major H2A sequence variants. Nucleic Acids Res. *28*, 3811–3816.

Kouzarides, T. (1999). Histone acetylases and deacetylases in cell proliferation. Curr. Opin. Genet. Dev. *9*, 40–48.

Kouzarides, T. (2000). Acetylation: a regulatory modification to rival phosphorylation? EMBO J. *19*, 1176–1179.

Kuo, M.H., and Allis, C.D. (1998). Roles of histone acetyltransferases and deacetylases in gene regulation. Bioessays *20*, 615–626.

Liu, X., Li, B., and Gorovsky, M.A. (1996). Essential and nonessential histone H2A variants in *Tetrahymena thermophila*. Mol. Cell. Biol. *16*, 4305–4311.

Luger, K., and Richmond, T.J. (1998). The histone tails of the nucleosome. Curr. Opin. Genet. Dev. *8*, 140–146.

Megee, P.C., Morgan, B.A., Mittman, B.A., and Smith, M.M. (1990). Genetic analysis of histone H4: essential role of lysines subject to reversible acetylation. Science *247*, 841–845.

Mizzen, C.A., and Allis, C.D. (1998). Linking histone acetylation to transcriptional regulation. Cell. Mol. Life Sci. *54*, 6–20.

Ng, H.H., and Bird, A. (2000). Histone deacetylases: silencers for hire. Trends Biochem. Sci. *25*, 121–126.

Norton, V.G., Imai, B.S., Yau, P., and Bradbury, E.M. (1989). Histone acetylation reduces nucleosome core particle linking number change. Cell *57*, 449–457.

Paro, R. (2000). Chromatin regulation. Formatting genetic text. Nature *406*, 579–580.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599.

Roth, S.Y., and Allis, C.D. (1996). Histone acetylation and chromatin assembly: a single escort, multiple dances? Cell *87*, 5–8.

Santisteban, M.S., Kalashnikova, T., and Smith, M.M. (2000). Histone H2A.Z regulates transcription and is functionally redundant with nucleosome remodeling complexes. Cell *103*, 411–422.

Shen, X., Yu, L., Weir, J.W., and Gorovsky, M.A. (1995). Linker histones are not essential and affect chromatin condensation in vivo. Cell *82*, 47–56.

Stargell, L.A., Bowen, J., Dadd, C.A., Dedon, P.C., Davis, M., Cook, R.G., Allis, C.D., and Gorovsky, M.A. (1993). Temporal and spatial association of histone H2A variant hv1 with transcriptionally competent chromatin during nuclear development in *Tetrahymena thermophila*. Genes Dev. *7*, 2641–2651.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications . Nature *403*, 41–45.

Taunton, J., Hassig, C.A., and Schreiber, S.L. (1996). A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. Science *272*, 408–411.

Thatcher, T.H., and Gorovsky, M.A. (1994). Phylogenetic analysis of core histones H2A, H2B, H3 and H4. Nuclei Acids Res. *22*, 174–179.

Tse, C., Sera, T., Wolffe, A.P., and Hansen, J.C. (1998). Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. Mol. Cell. Biol. *18*, 4629–4638.

Turner, B.M. (1995). Histone H4, the cell cycle and a question of integrity. Bioessays *17*, 1013–1015.

Turner, B.M. (2000). Histone acetylation and an epigenetic code. Bioessays *22*, 836–845.

Ura, K., Kurumizaka, H., Dimitrov, S., Almouzni, G., and Wolffe, A.P. (1997). Histone acetylation: influence on transcription, nucleosome mobility and positioning, and linker histone-dependent transcriptional repression. EMBO J. *16*, 2096–2107.

Van Daal, A., and Elgin, S.C.R. (1992). A histone variant, H2AvD, is essential in *Drosophila melanogaster*. Mol. Biol. Cell *3*, 593–602.

Verreault, A., Kaufman, P.D., Kobayashi, R., and Stillman, B. (1996). Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4. Cell *87*, 95–104.

Wade, P.A., Pruss, D., and Wolffe, A.P. (1997). Histone acetylation: chromatin in action. Trends Biochem. Sci. *22*, 128–132.

Wei, Y., Yu, L., Bowen, J., Gorovsky, M.A., and Allis, C.D. (1999). Phosphorylation of histone H3 is required for proper chromosome condensation and segregation. Cell *97*, 99–109.

West, M.H., and Bonner, W.M. (1980). Histone 2A, a heteromorphous family of eight protein species. Biochemistry *19*, 3238–3245.

Wolffe, A.P. (1998). Chromatin: structure and function. Third edition. (San Diego, CA: Academic Press).

Wolffe, A.P., and Hayes, J.J. (1999). Chromatin disruption and modification. Nucleic Acids Res. *27*, 711–720.

Wu, R.S., Tsai, S., and Bonner, W.M. (1982). Patterns of histone variant synthesis can distinguish G0 from G1 cells. Cell *31*, 367–374.

Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment

Brian Palenik*, Qinghu Ren†, Chris L. Dupont*, Garry S. Myers†, John F. Heidelberg†, Jonathan H. Badger†, Ramana Madupu†, William C. Nelson†, Lauren M. Brinkac†, Robert J. Dodson†, A. Scott Durkin†, Sean C. Daugherty†, Stephen A. Sullivan†, Hoda Khouri†, Yasmin Mohamoud†, Rebecca Halpin†, and Ian T. Paulsen†‡

*Scripps Institution of Oceanography, University of California at San Diego, La Jolla, CA 92093; and †The Institute for Genomic Research, Rockville, MD 20850

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved July 10, 2006 (received for review April 11, 2006)

Coastal aquatic environments are typically more highly productive and dynamic than open ocean ones. Despite these differences, cyanobacteria from the genus *Synechococcus* are important primary producers in both types of ecosystems. We have found that the genome of a coastal cyanobacterium, *Synechococcus* sp. strain CC9311, has significant differences from an open ocean strain, *Synechococcus* sp. strain WH8102, and these are consistent with the differences between their respective environments. CC9311 has a greater capacity to sense and respond to changes in its (coastal) environment. It has a much larger capacity to transport, store, use, or export metals, especially iron and copper. In contrast, phosphate acquisition seems less important, consistent with the higher concentration of phosphate in coastal environments. CC9311 is predicted to have differences in its outer membrane lipopolysaccharide, and this may be characteristic of the speciation of some cyanobacterial groups. In addition, the types of potentially horizontally transferred genes are markedly different between the coastal and open ocean genomes and suggest a more prominent role for phages in horizontal gene transfer in oligotrophic environments.

cyanobacteria | genomics | marine

Coastal waters typically have higher nutrient concentrations than open ocean waters because of wind-driven upwelling of nutrients from deeper depths and inputs from land and sediments. The higher nutrient concentrations lead to higher primary productivity. The spectral quality of light is typically different because of the presence of terrestrial material and algal biomass. These conditions contrast strongly with the low-nutrient blue-light-dominated ecosystems of the open ocean. Although each coastal environment has unique elements, these generalizations help us understand the adaptations likely to be found in coastal compared to open ocean microorganisms.

Some adaptations of photosynthetic microorganisms to the open ocean vs. coastal environment have included adaptations to nutrient levels and light. Differences in the pigments of coastal vs. open ocean *Synechococcus* have been well documented (1–4). In terms of nutrients, Carpenter (5) noted that coastal phytoplankton (diatom) species had a higher K_s (half-saturation constant for transport) for nitrate, whereas related open ocean diatom species had a lower K_s. The minimum amount of iron and other metals for growth of open ocean phytoplankton is less than that needed for coastal species, suggesting that adaptation to *in situ* metal levels is a significant factor in phytoplankton speciation (6–8). Recently, it has been shown, again in diatoms, that adaptation to low iron in the open ocean involves changes in the cellular concentration of the iron-rich photosynthetic reaction center proteins of photosystem I (9) and the use of plastocyanin, a copper containing protein, instead of iron (10).

We report here the genome sequence of *Synechococcus* sp. strain CC9311. This organism was isolated from the edge of California Current after nitrate enrichment and low light incubation (11). Strains related to CC9311 have been isolated from coastal environments such as Vineyard Sound (12, 13) and have been highly represented in *rpoC* gene sequence libraries of Southern California coastal waters and in the water column of the California Current when it displayed a coastal type chlorophyll profile (ref. 14; B.P., unpublished work). CC9311 possesses an ability to adapt to light quality (blue to green light ratios) not seen in open ocean *Synechococcus* strains such as WH8102, further indicating a coastal ecosystem niche for this strain (12). The availability of the genome sequence of CC9311 (Fig. 1) allows us to compare it to the genome sequence of *Synechococcus* sp. strain WH8102 (15), an open ocean strain, and to begin to understand the adaptation of bacterial genomes to the coastal vs. open ocean environments.

Results and Discussion

Gene Regulation and Two-Component Regulatory Systems. One of the insights from the genome of the open ocean *Synechococcus* WH8102 was that it and other open ocean cyanobacteria have minimal regulatory systems, particularly two-component regulatory systems consisting of a sensor and response regulator pair (15–17). There are only five histidine kinase sensors and nine response regulators in WH8102, and it was suggested that this was due to adaptation to a relatively constant ecosystem. As one would predict from adaptation to the more variable coastal environment, CC9311 has nearly double this number, with 11 histidine kinase sensors and 17 response regulators (Fig. 2). Interestingly, these additional systems occur in pairs in the genome, which is not always the case in WH8102. The function of these sensors is not predictable from their sequences at this time but may regulate the more complex metal metabolism in CC9311.

Despite the presence of additional sensor kinases, based on BLAST and phylogenetic analyses, CC9311 apparently lacks a phosphate sensor-response regulator system seen in other cyanobacteria and bacteria in general (18). Consistent with this, several alkaline phosphatases present in WH8102 are absent, and CC9311 has fewer periplasmic phosphate-binding proteins used in ABC transporter systems. These differences between the open ocean and coastal *Synechococcus* types likely reflect the higher phosphate concentrations in coastal environments compared to some surface ocean environments where phosphate can become limiting.

Metals and CC9311. CC9311 has a number of metal enzymes or cofactors not found in WH8102, suggesting that it has a greater

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. CP000435). The *Synechococcus* CC9311 strain has been deposited in the Provasoli–Guillard National Center for Culture of Marine Phytoplankton (http://ccmp.bigelow.org) under catalog no. CCMP2515.

‡To whom correspondence should be addressed. E-mail: ipaulsen@tigr.org.

PLANT BIOLOGY

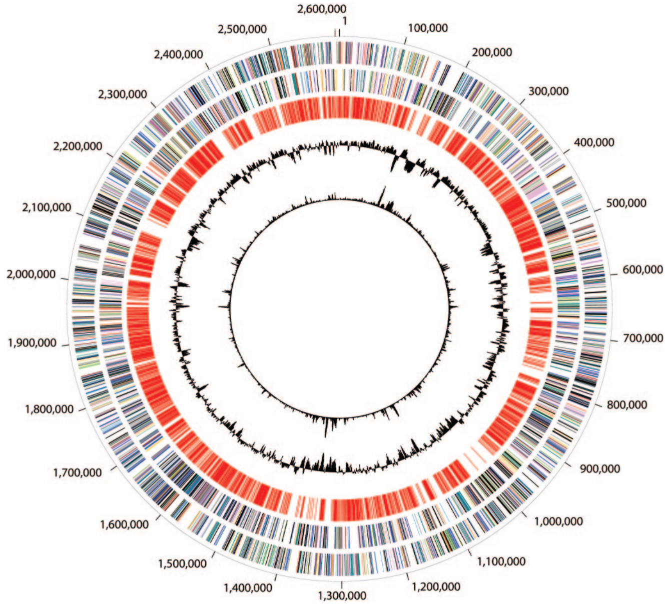


Fig. 1. Circular representation of the *Synechococcus* CC9311 overall genome structure. The outer scale designates coordinates in base pairs. The first circle shows predicted coding regions on the plus strand, color coded by role categories: violet, amino acid biosynthesis; light blue, biosynthesis of cofactors, prosthetic groups, and carriers; light green, cell envelope; red, cellular processes; brown, central intermediary metabolism; yellow, DNA metabolism; light gray, energy metabolism; magenta, fatty acid, and phospholipid metabolism; pink, protein synthesis and fate; orange, purines, pyrimidines, nucleosides, and nucleotides; olive, regulatory functions and signal transduction; dark green, transcription; teal, transport, and binding proteins; gray, unknown function; salmon, other categories; and blue, hypothetical proteins. The second circle shows predicted coding regions on the minus strand color coded by role categories. The third circle shows in red the set of 1,730 genes conserved between *Synechococcus* CC9311 and WH8102, the fourth circle shows percentage G+C in relation to the mean G+C in a 2,000-bp window in black, and the fifth circle shows the trinucleotide composition in black.

use for iron (Fig. 3). This is consistent with higher metal quotas for iron of coastal vs. open ocean phytoplankton, such as diatoms (6), and adds a mechanistic basis to these previous studies. Iron-dependent metalloenzymes unique to CC9311 include a cytochrome P450-like encoding ORF (sync_2424), two additional cytochrome *c* molecules (sync_1753 and sync_1742), and one or two additional ferrodoxins (sync_1953 and sync_0980, the latter truncated). It also has a putative iron-dependent alcohol dehydrogenase (sync_2669).

CC9311 appears to have a greater use for copper than WH8102, because it has a copper zinc superoxide dismutase not seen in marine cyanobacteria (sync_1771) until this work and the recent availability of two marine cyanobacterial genomes (CC9902 and CC9605; http://genome.jgi-psf.org/mic_home.html). It has a putative multicopper oxidase (sync_1489), which could be involved in oxidation of organic compounds or detoxifying high levels of reduced copper (19). Interestingly, it has been shown that *Synechococcus* sp. strain WH8016, a strain in the same clade as CC9311, was more resistant to copper than oligotrophic strains (20) .

For other metal usage, there appears to be a putative vanadium-dependent bromoperoxidase (sync_2681). The latter gene is very interesting, because it is highly similar to one in marine red algae. In red algae, this enzyme generates brominated compounds using hydrogen peroxide (21, 22). Cyanobacteria have been shown to produce brominated compounds such as bromodiphenyl ethers through an unknown mechanism, with the best-studied case being a filamentous cyanobacterial symbiont of a sponge (23). These brominated compounds have been found recently to cause leakage of fungal cell membranes (24), but the role of brominated compounds, if any, in CC9311 is open to speculation.

Possibly because of its more intensive use of metals, CC9311 has some metal transporters not seen in WH8102, including an FeoA/B transporter for iron(II) (sync_0681-0682). Total iron concentrations are higher in coastal environments, and reduced iron(II) may be more abundant as well, because it is likely produced from photochemical reactions of iron and organic matter (25, 26). CC9311 also has three cation-dependent efflux transporters (sync_0686, sync_1861, and sync_1510) compared to two in WH8102, suggesting that it may have an increased capacity to export toxic metal levels if needed.

In contrast, the oligotrophic ocean strain WH8102 has systems predicted for the efflux of arsenite (preceded by its reduction) and chromate (15) that are not found in CC9311. It has been suggested that high arsenate to phosphate ratios in oligotrophic regions result in the need of microorganisms to deal with excess arsenate (27).

Coastal *Synechococcus* strain CC9311 has a greatly enhanced capacity for metal storage. This is seen in the four copies of *smtA*, a gene for bacterial metallothionein (sync_1081, sync_2426, sync_0853, and sync_2379) compared to one in *Synechococcus* WH8102 and none in some *Prochlorococcus* strains. Gene amplification of *smtA* has been found in freshwater *Synechococcus* PCC6301 in response to higher trace metal levels such as cadmium (28). However, in this case, *smtA* copies occur in tandem, not disbursed throughout the genome as seen in CC9311.

CC9311 also has a greatly enhanced capacity specifically for iron storage. It has five copies of bacterial ferritin (sync_0854, sync_0687, sync_1077, sync_1539, and sync_0680) compared to one in most cyanobacterial genomes including *Synechococcus* WH8102. It also has a ferritin-related protein DpsA (DNA-protecting protein under starved conditions) that binds iron. The later is not found in WH8102 but is found in some *Prochlorococcus* strains (PMT2218 in MIT9313).

It is unclear whether the greatly enhanced transport and metal storage capacity for iron and other metals in CC9311 is due to a greater need for metals, the need to respond to excess metal levels, or the possibility that the cells see episodic metal concentrations. Iron concentrations in California coastal environments can vary from limiting to replete with rapid fluctuations (29), thus the ability to store iron may be advantageous. Taken together, these results suggest a much more metal-dependent ecological strategy for CC9311 (Fig. 3 and Table 1, which is published as supporting information on the PNAS web site).

Organic Nitrogen and Other Transporters. CC9311 and WH8102 also differ in other aspects of their membrane transporter complement that may reflect differences in the nutrients they are exposed to in their different environments. Interestingly, CC9311 has multiple AMT family ammonia transporters and based on this, ammonia is arguably its most important nitrogen source, but determining this will require *in situ* gene expression studies. CC9311 encodes a TRAP family dicarboxylate transporter as well as a DASS family transporter that may also be specific for carboxylates and a formate/nitrite transporter that is not present in WH8102. CC9311 also encodes a second type of predicted urea transporter and two APC-type amino acid transporters that are not present in WH8102. These capabilities are consistent with the coastal isolate CC9311 being exposed to more organic matter than its oligotrophic ocean relative WH8102. There is a significant expansion of mechanosensitive ion channels in CC9311, which has five MscS and two MscL members compared with only two MScS channels in WH8102. Mechanosensitive ion channels can function as "emergency

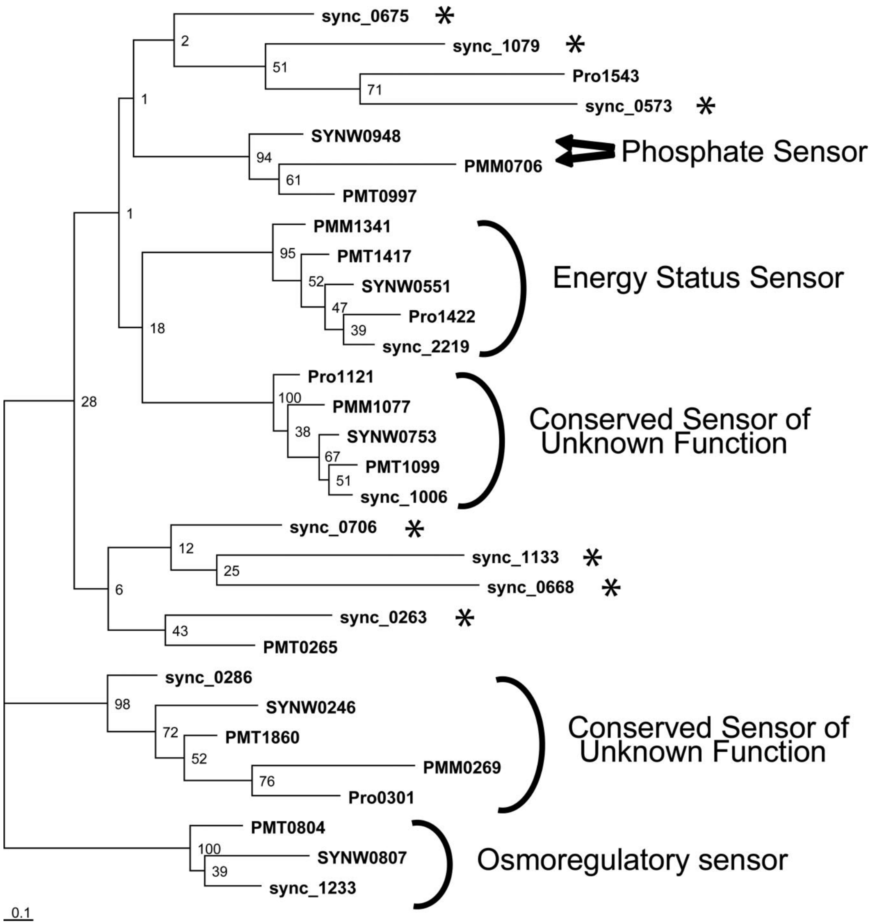


Fig. 2. Phylogenetic tree of sensor kinases from *Synechococcus* CC9311 (sync_xxxx), and WH8102 (SYNWxxxx), *Prochlorococcus marinus* MI9313 (PMTxxxx), MED4 (PMMxxxx), and SS120 (Proxxxx). This maximum-likelihood phylogenetic tree was generated by using PHYLIP, and bootstrap values are indicated next to the branch nodes. Orthologous clusters conserved in all of the cyanobacteria shown are highlighted by lines on the side, the phosphate sensor is labeled, and the divergent sensors unique to CC9311 are highlighted with asterisks.

relief valves" during conditions of osmotic shock, implying that the coastal isolate CC9311 may be subject to a more osmotically challenging environment (30).

Light and CC9311. The predicted ORFs associated with photosynthesis and light harvesting are relatively similar to WH8102. One exception is the much greater number of high light-inducible protein (HLIP) gene family members in CC9311 (with 14) compared to WH8102 (with eight). Increased HLIP content has been associated with cyanobacteria found in high light environments (16), thus these results predict that CC9311 would have the capacity to live in high light surface waters or under changing light conditions found during mixing of the water column.

Some differences in the ORFs clustered in the phycobilisome-encoding region were found between WH8102 and CC9311, and these may play a role in the type IV chromatic light adaptation discovered in CC9311 (12). The genome sequence identifies two ORFs (sync_0485 and sync_0486) as phycobiliprotein lyases not found in WH8102; such proteins were predicted to be involved in chromatic adaptation in a recent biochemical study (31). These ORFs clearly merit further attention.

Horizontal Gene Transfer. Strains WH8102 and CC9311 share 1,730 ORFs. Mapping these on the CC9311 genome indicated they were unevenly distributed, with a number of intervening regions that essentially lacked any genes conserved with WH8102 (Fig. 1). Analysis of these regions indicated that some (≈116 ORFs with 19 regions of >3 kb) displayed an atypical trinucleotide composition and GC percentage, suggesting they may be novel genomic "islands" relatively recently acquired by CC9311 (Table 2, which is published as supporting information on the PNAS web site). Previous analysis of the WH8102 genome (15) had also identified putative similar islands based on their atypical nucleotide content. These WH8102 putative islands also essentially lacked any of the 1,730 conserved *Synechococcus* genes.

The putative genomic islands with atypical nucleotide content from CC9311 and WH8102 appear to differ significantly in terms of gene function. The majority of the WH8102 islands consist largely of hypothetical genes, often flanked by phage integrase genes, suggesting they may be of phage origin. In contrast, none of the CC9311 islands contain phage integrase genes or other identifiable phage genes. It has been hypothesized that lysogenic phages would be more common in nutrient-poor environments such as the open ocean (discussed in ref. 32). The residual phage-related genes in open ocean WH8102 but not CC9311 are some of the first data consistent with this hypothesis.

Both genomes have unique islands consisting of different polysaccharide biosynthesis genes that may be important in changing cell surface characteristics, perhaps in response to phage or grazing selection pressure. Other islands unique to CC9311 encode an ABC secretion system and an RTX family toxin homologue, a predicted secreted nuclease and protease, and some two-component regulatory system genes. Some of the

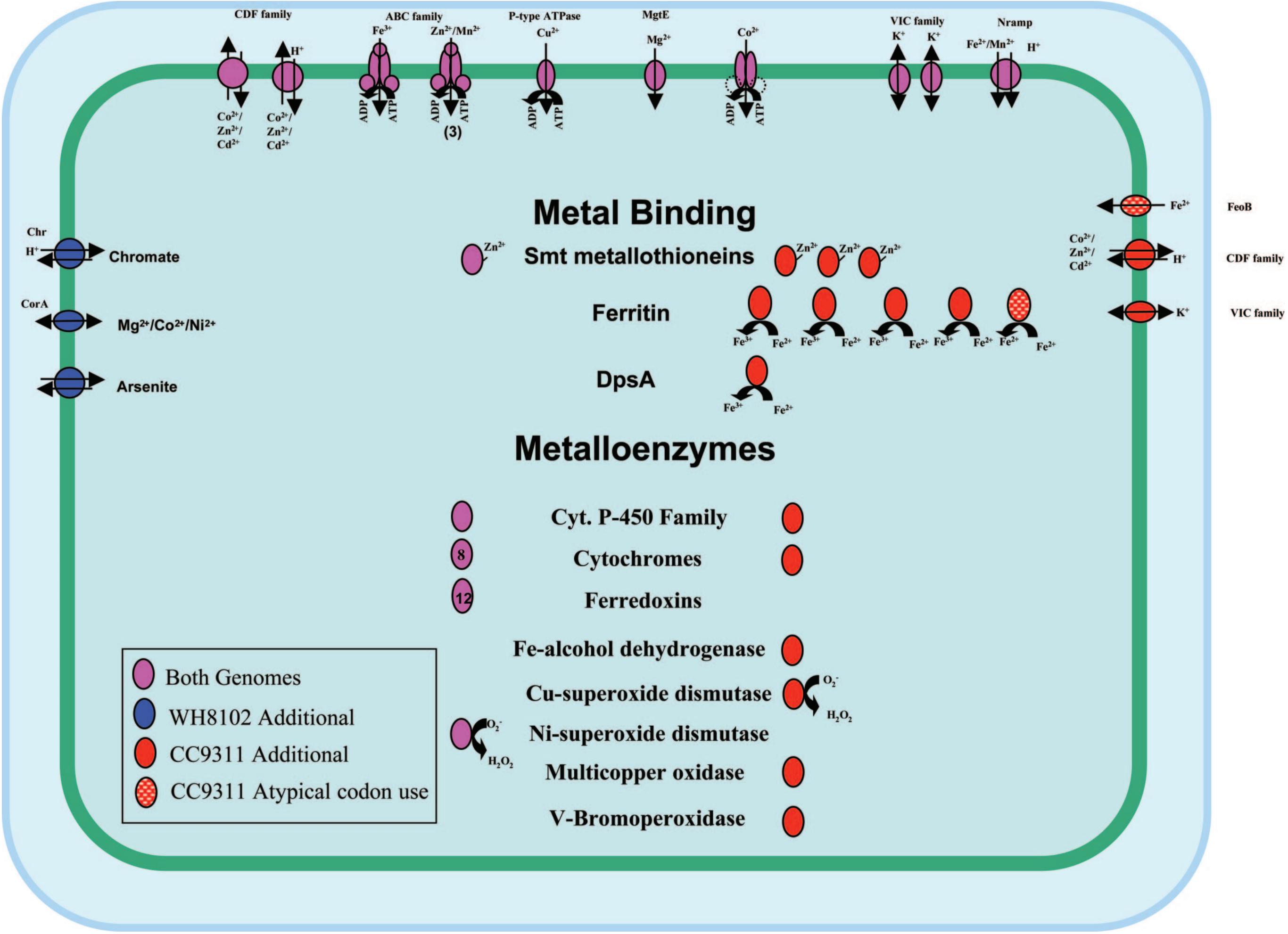


Fig. 3. Overview of metal transport and metabolism in *Synechococcus* CC9311 and WH8102. Metal ion transporters are shown in the membrane, with the arrows indicating the direction of transport. Metal-binding proteins and metalloenzymes are shown inside the cell, and the number of copies of each system is shown in parentheses or within the protein. The color shading of the proteins indicates their distribution: magenta, present in both WH8102 and CC9311; red, present only in CC9311; and blue, present only in WH8102. Hatching indicates that the gene is located in a region with atypical trinucleotide content.

previously mentioned metal metabolism genes, including a ferritin and ferrous iron transport genes, are also found in these islands. The presence of metal-related (especially iron) genes in these islands with atypical codon usage is interesting, because it suggests that metal usage may also be under strong selection. Genes with new physiological capabilities for metal use may be highly favored and maintained in CC9311, if acquired through horizontal gene transfer.

Cell Surfaces: LPS and Pili. CC9311 (relative to WH8102) is missing the genes for the synthesis of KDO, a molecule necessary for the biosynthesis of a typical LPS, and is missing genes for one pathway for the biosynthesis of the sugar rhamnose, a potential component of LPS. The genes for the synthesis of lipid A, the lipid part of LPS, were found. At its simplest level, this suggests that CC9311 has differences in its LPS compared to WH8102. Preliminary LPS analyses suggest this to be the case (B.P., B. Brahamsha, P. Azadi, and S. Snyder, unpublished work). A greatly altered LPS could drastically change the sensitivity of CC9311 to particular phages; because of its abundance at the cell surface, LPS is often a phage receptor (33).

CC9311 has seven putative genes for pili and pilin biosynthesis. Thus it may have pili that would be available for twitching motility or DNA uptake. Both of these could be potentially useful in coastal ecosystems where CC9311 is more likely to encounter surfaces or DNA than in the open ocean. In contrast, CC9311 is missing two major cell surface proteins (*SwmA* and *SwmB*) involved in swimming motility in WH8102 (34, 35). The use of the CC9311 genome and other nonmotile *Synechococcus* genomes will help determine genes unique to WH8102 and thus other genes that could be involved in its unique form of swimming motility. However, our examination of these WH8102 "unique" genes so far has not yielded clues, because many of these genes are annotated only as hypothetical or conserved hypothetical.

Summary. The coastal strain CC9311 has dramatic differences in gene complement compared to the open ocean strain WH8102. Many of these differences are consistent with adaptation to a coastal environment. Because the genus marine *Synechococcus* contains multiple clades (potential species), it will be interesting to see which of these coastal/open ocean differences will be conserved across all clades or whether, even within coastal clades, different strategies exist for adapting to this complex environment.

Methods

Genome Sequencing, Annotation, and Characteristics. The complete genome sequence of *Synechococcus* CC9311 was determined by using the whole-genome shotgun method (36). Physical and sequencing gaps were closed by using a combination of primer walking, generation and sequencing of transposon-tagged libraries of large-insert clones, and multiplex PCR (37). Identification of putative protein-encoding genes and annotation of the genome were performed as described (38). An initial set of ORFs predicted to encode proteins was initially identified by using GLIMMER (39). ORFs consisting of <30 codons and those containing overlaps were eliminated. Frame shifts and point mutations were corrected or designated "authentic." Functional assignment, identification of membrane-spanning domains, and determination of paralogous gene families were performed as described (38). Sequence alignments and phylogenetic trees were generated by using the methods described (38). The CC9311 genome was found to be composed of one circular chromosome of 2,606,748 bp (Fig. 1), with an average GC content of 52.5%. A total of 3,065 ORFs, 2 rRNA operons, and 44 tRNAs were identified within the CC9311 genome.

Trinucleotide Composition. Distribution of all 64 trinucleotides (3 mers) was determined, and the 3-mer distribution in 2,000-bp windows that overlapped by half their length (1,000 bp) across the genome was computed. For each window, we computed the χ^2 statistic on the difference between its 3-mer content and that of the whole chromosome. A large value for χ^2 indicates the 3-mer composition in this window is different from the rest of the chromosome. Probability values for this analysis are based on assumptions that the DNA composition is relatively uniform throughout the genome, and that 3-mer composition is independent. Because these assumptions may be incorrect, we prefer to interpret high χ^2 values as indicators of regions on the chromosome that appear unusual and demand further scrutiny.

Comparative Genomics. The *Synechococcus* CC9311 and WH8102 genomes were compared at the nucleotide level by suffix tree analysis by using MUMmer (40), and their ORFs were compared by a reciprocal best BLAST match analysis by using an E-value cutoff of 10^{-5}.

We thank The Institute for Genomic Research faculty, sequencing facility, and informatics group for expert advice and assistance. This work was supported by National Science Foundation Grant EF0333162.

1. Wood, A. M., Phinney, D. A. & Yentsch, C. S. (1998) *Mar. Ecol. Prog. Ser.* **162,** 25–31.
2. Olson, R. J., Chisholm, S. W., Zettler, E. R. & Armbrust, E. V. (1988) *Deep-Sea Res.* **35,** 425–440.
3. Olson, R. J., Chisholm, S. W., Zettler, E. R. & Armbrust, E. V. (1990) *Limnol. Oceanogr.* **35,** 45–58.
4. Wood, A. M., Lipsen, M. & Coble, P. (1999) *Deep-Sea Res. II* **46,** 1769–1790.
5. Carpenter, E. J. & Guillard, R. R. L. (1971) *Ecology* **52,** 183–185.
6. Sunda, W. G., Swift, D. G. & Huntsman, S. A. (1991) *Nature* **351,** 55–57.
7. Brand, L. E., Sunda, W. G. & Guillard, R. R. L. (1983) *Limnol. Oceanogr.* **28,** 1182–1198.
8. Ryther, J. H. & Kramer, D. D. (1961) *Ecology* **42,** 444–446.
9. Strzepek, R. F. & Harrison, P. J. (2004) *Nature* **431,** 689–692.
10. Peers, G. & Price, N. M. (2006) *Nature* **441,** 341–344.
11. Toledo, G. & Palenik, B. (1997) *Appl. Environ. Microbiol.* **63,** 4298–4303.
12. Palenik, B. (2001) *Appl. Environ. Microbiol.* **67,** 991–994.
13. Waterbury, J. B. & Rippka, R. (1989) in *Bergey's Manual of Systematic Bacteriology*, eds. Staley, J. T., Bryant, M. P., Pfennig, N. & Holt, J. B. (Williams & Wilkins, Baltimore), Vol. 3, pp. 1728–1746.
14. Ferris, M. J. & Palenik, B. (1998) *Nature* **396,** 226–228.
15. Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. A., McCarren, J., *et al.* (2003) *Nature* **424,** 1037–1042.
16. Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., *et al.* (2003) *Nature* **424,** 1042–1047.
17. Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10020–10025.
18. Hirani, T., Suzuki, I., Murata, N., Hayashi, H. & Eaton-Rye, J. (2001) *Plant Mol. Biol.* **45,** 133–144.
19. Grass, G., Thakali, K., Klebba, P., Thieme, D., Muller, A., Wildner, G. & Rensing, C. (2004) *J. Bacteriol.* **186,** 5826–5833.
20. Brand, L. E., Sunda, W. G. & Guillard, R. R. L. (1986) *J. Exp. Mar. Biol. Ecol.* **96,** 225–250.
21. Pedersén, M. (1976) *Physiol. Plant* **37,** 6–11.
22. Carter, J. N., Beatty, K. E., Simpson, M. T. & Butler, A. (2002) *J. Inorg. Biochem.* **91,** 59–69.
23. Unson, M. D., Holland, N. D. & Faulkner, D. J. (1994) *Mar. Biol.* **119,** 1–11.
24. Sionov, E., Roth, D., Sandovsky-Losica, H., Kashman, Y., Rudi, A., Chill, L., Berdicevsky, I., Segal, E., *et al*. (2005) *J. Infect.* **50,** 453–460.
25. Kuma, K., Nakabayashi, S., Suzuki, Y., Kudo, I. & Matsunaga, K. (1992) *Marine Chemistry* **37,** 15–27.
26. Barbeau, K., Rue, E. L., Trick, C. G., Bruland, K. W. & Butler, A. (2003) *Limnol. Oceanogr.* **48,** 1069–1078.
27. Cutter, G., Cutter, L., Featherstone, A. & Lohrenz, S. E. (2001) *Deep-Sea Res. II* **48,** 2895–2915.
28. Gupta, A., Whitton, B. A., Morby, A. P., Huckle, J. W. & Robinson, N. J. (1992) *Proc. R. Soc. London Ser. B* **248,** 273–281.
29. Bruland, K. W., Rue, E. L. & Smith, G. J. (2001) *Limnol. Oceanogr.* **46,** 1661–1674.
30. Booth, I. R. & Louis, P. (1999) *Curr. Opin. Microbiol.* **2,** 166–169.
31. Everroad, C., Six, C., Partensky, F., Thomas, J. C., Holtzendorff, J. & Wood, A. M. (2006) *J. Bacteriol.* **188,** 3345–3356.
32. Ortmann, A. C., Lawrence, J. E. & Suttle, C. A. (2002) *Microb. Ecol.* **43,** 225–231.
33. Traurig, M. & Misra, R. (1999) *FEMS Microbiol. Lett.* **181,** 101–108.
34. McCarren, J., Heuser, J., Roth, R., Yamada, N., Martone, M. & Brahamsha, B. (2005) *J. Bacteriol.* **187,** 224–230.
35. McCarren, J. & Brahamsha, B. (2005) *J. Bacteriol.* **187,** 4457–4462.
36. Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., *et al.* (1997) *Nature* **390,** 580–586.
37. Tettelin, H., Radune, D., Kasif, S., Khouri, H. & Salzberg, S. L. (1999) *Genomics* **62,** 500–507.
38. Paulsen, I. T., Seshadri, R., Nelson, K. E., Eisen, J. A., Heidelberg, J. F., Read, T. D., Dodson, R. J., Umayam, L., Brinkac, L. M., Beanan, M. J., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 13148–13153.
39. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **26,** 544–548.
40. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. (2002) *Nucleic Acids Res.* **30,** 2478–2483.

Life in Hot Carbon Monoxide: The Complete Genome Sequence of *Carboxydothermus hydrogenoformans* Z-2901

Martin Wu[1], Qinghu Ren[1], A. Scott Durkin[1], Sean C. Daugherty[1], Lauren M. Brinkac[1], Robert J. Dodson[1], Ramana Madupu[1], Steven A. Sullivan[1], James F. Kolonay[1], William C. Nelson[1], Luke J. Tallon[1], Kristine M. Jones[1], Luke E. Ulrich[2], Juan M. Gonzalez[3], Igor B. Zhulin[2], Frank T. Robb[3], Jonathan A. Eisen[1,4,*]

1 The Institute for Genomic Research, Rockville, Maryland, United States of America, **2** Center for Bioinformatics and Computational Biology, School of Biology, Georgia Institute of Technology, Atlanta, Georgia, United States of America, **3** Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland, United States of America, **4** Johns Hopkins University, Baltimore, Maryland, United States of America

We report here the sequencing and analysis of the genome of the thermophilic bacterium *Carboxydothermus hydrogenoformans* Z-2901. This species is a model for studies of hydrogenogens, which are diverse bacteria and archaea that grow anaerobically utilizing carbon monoxide (CO) as their sole carbon source and water as an electron acceptor, producing carbon dioxide and hydrogen as waste products. Organisms that make use of CO do so through carbon monoxide dehydrogenase complexes. Remarkably, analysis of the genome of *C. hydrogenoformans* reveals the presence of at least five highly differentiated anaerobic carbon monoxide dehydrogenase complexes, which may in part explain how this species is able to grow so much more rapidly on CO than many other species. Analysis of the genome also has provided many general insights into the metabolism of this organism which should make it easier to use it as a source of biologically produced hydrogen gas. One surprising finding is the presence of many genes previously found only in sporulating species in the Firmicutes Phylum. Although this species is also a Firmicutes, it was not known to sporulate previously. Here we show that it does sporulate and because it is missing many of the genes involved in sporulation in other species, this organism may serve as a "minimal" model for sporulation studies. In addition, using phylogenetic profile analysis, we have identified many uncharacterized gene families found in all known sporulating Firmicutes, but not in any non-sporulating bacteria, including a sigma factor not known to be involved in sporulation previously.

Citation: Wu M, Ren Q, Durkin AS, Daugherty SC, Brinkac LM, et al. (2005) Life in hot carbon monoxide: The complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. PLoS Genet 1(5): e65.

Introduction

Carbon monoxide (CO) is best known as a potent human poison, binding very strongly and almost irreversibly to the iron core of hemoglobin. Despite its deleterious effects on many species, it is also the basis for many food chains, especially in hydrothermal environments such as the deep sea, hot springs, and volcanoes. In these environments, CO is a common potential carbon source, as it is produced both by partial oxidation of organic matter as well as by multiple microbial strains (e.g., methanogens). It is most readily available in areas in which oxygen concentrations are low, since oxidation of CO will convert it to CO_2. In hydrothermal environments, CO use as a primary carbon source is dominated by the hydrogenogens, which are anaerobic, thermophilic bacteria or archaea that carry out CO oxidation using water as an electron acceptor [1]. This leads to the production of CO_2 and H_2. The H_2 is frequently lost to the environment and the CO_2 is used in carbon fixation pathways for the production of biomass. Hydrogenogens have attracted significant biotechnological interest because of the possibility they could be used in the biological production of hydrogen gas.

Hydrogenogens are found in diverse volcanic environments [2–7]. The phylogenetic types differ somewhat depending on the environments and include representatives of bacteria and archaea. *Carboxydothermus hydrogenoformans* is a hydrogenogen that was isolated from a hot spring in Kunashir Island, Russia [2]. It is a member of the Firmicutes Phylum (also known as low GC Gram-positives) and grows optimally at 78 °C. This species has been considered an unusual hydrogenogen, in part because unlike most of the other hydrogenogens, it was believed to be strictly dependent on CO for growth. The other species were found to grow poorly unless CO was supplemented with organic substrates. Thus it was selected for genome sequencing as a potential model obligate CO autotroph.

Surprisingly, initial analysis of the unpublished genome

Received August 8, 2005; Accepted October 19, 2005; Published November 25, 2005
DOI: 10.1371/journal.pgen.0010065

Abbreviations: CDS, protein coding sequence; CO, carbon monoxide; CODH, carbon monoxide dehydrogenase

Editor: Paul M Richardson, Joint Genome Institute, United States of America

* To whom correspondence should be addressed. E-mail: jeisen@tigr.org

A previous version of this article appeared as an Early Online Release on October 19, 2005 (DOI: 10.1371/journal.pgen.0010065.eor).

Synopsis

Carboxydothermus hydrogenoformans, a bacterium isolated from a Russian hotspring, is studied for three major reasons: it grows at very high temperature, it lives almost entirely on a diet of carbon monoxide (CO), and it converts water to hydrogen gas as part of its metabolism. Understanding this organism's unique biology gets a boost from the decoding of its genome, reported in this issue of *PLoS Genetics*. For example, genome analysis reveals that it encodes five different forms of the protein machine carbon monoxide dehydrogenase (CODH). Most species have no CODH and even species that utilize CO usually have only one or two. The five CODH in *C. hydrogenoformans* likely allow it to both use CO for diverse cellular processes and out-compete for it when it is limiting. The genome sequence also led the researchers to experimentally document new aspects of this species' biology including the ability to form spores. The researchers then used comparative genomic analysis to identify conserved genes found in all spore-forming species, including *Bacillus anthracis,* and not in any other species. Finally, the genome sequence and analysis reported here will aid in those trying to develop this and other species into systems to biologically produce hydrogen gas from water.

sequence data led to the discovery that this species is not an obligate CO autotroph [8]. We report here a detailed analysis of the genome sequence of *C. hydrogenoformans* strain Z-2901, the type strain of the species, hereafter referred to simply as *C. hydrogenoformans*.

Results/Discussion

Genome Structure

The *C. hydrogenoformans* genome is a single circular chromosome of 2,401,892 base pairs (bp) with a G+C content of 42.0% (Figure 1, Table 1). Annotation of the genome reveals 2,646 putative protein coding genes (CDSs), of which 1,512 can be assigned a putative function. The chromosome displays two clear GC skew transitions that likely correspond to the DNA replication origin and terminus (Figure 1). Overall, 3.0 % of the genome is made up of repetitive DNA sequences. Included in this repetitive DNA are two large-clustered, regularly interspaced short palindromic repeats (CRISPR, 3.9 and 5.6 kilobases, respectively). Each cluster contains 59 and 84 partially palindromic repeats of 30 bps, respectively (GTTTCAATCCCAGA[A/T]TGGTTCGATTAAAAC). Most repeats within each cluster are identical but they differ for one nucleotide in the middle between clusters. Repeats at ends of the smaller cluster degenerate to some extent. These types of repeats are widespread in diverse groups of bacteria and archaea [9]. The first one-third of the repeat sequence is generally conserved. Although the precise functions of these repeats are unknown, some evidence suggests they are involved in chromosome partitioning [10,11]. In addition, experiments in the thermophilic archaea *Sulfolobus solfataricus* have identified a genus-specific protein binding specifically to the repeats present in that species' genome [11].

One 35-kilobase lambda-like prophage containing 50 CDSs was identified in the genome. It is flanked on one side by a tRNA suggesting this may have served as a site of insertion. Phylogenetic analysis showed this phage is most closely related to phages found in other Firmicutes, particularly the SPP1 phage infecting *Bacillus subtilis.*

As with other members of the Phylum Firmicutes, the directions of leading strand DNA replication and transcription are highly correlated, with 87% of genes located on the leading strand. This gene distribution bias is also highly correlated with the presence of a Firmicutes-specific DNA polymerase PolC in the genome [12]. In *B. subtilis,* PolC synthesizes the leading strand, and another distinct DNA polymerase, DnaE, replicates the lagging strand [13]. In other non-Firmicutes bacteria, DnaE replicates both strands. The asymmetric replication forks of Firmicutes were proposed to contribute to the asymmetry of their gene distributions [12]. One copy of PolC and two copies of DnaE have been identified in *C. hydrogenoformans* genome. At least some of the gene distribution bias can be caused by selection to avoid collision of the RNA and DNA polymerases as well [14,15]. Despite this apparent selection, the lack of significantly conserved gene order across Firmicutes indicates that genome rearrangements still occur at a reasonably high rate.

Phylogeny and Taxonomy

Analysis of the complete genome of *C. hydrogenoformans* suggests that the taxonomy of this species, as well as some other organisms, needs to be revised. More specifically, phylogenetic analysis based on concatenation of a few dozen markers (Figure 2) reveals a variety of conflicts between the organismal phylogeny and the classification of some of the Firmicutes. For example, *C. hydrogenoformans* is currently considered to be a member of the Family Peptococcaceae in the Order Clostridiales [16]. Thus it should form a clade with the *Clostridium* spp. to the exclusion of other taxa for which genomes are available (e.g., *Thermoanaerobacter tengcongensis,* which is considered to be a member of Thermoanaerobacteriales). The tree, however, indicates that this is not the case and that *T. tengcongensis* and the *Clostridia* spp. are more closely related to each other than either is to *C. hydrogenoformans.* Thus we believe *C. hydrogenoformans* should be placed in a separate Order from Clostridiales.

Perhaps more surprisingly, the concatenated genome tree shows *C. hydrogenoformans* grouping with *Symbiobacterium thermophilum. S. thermophilum* is a strictly symbiotic thermophile isolated from compost and is currently classified in the Actinobacteria (also known as high GC Gram-positives) based on analysis of its 16s rRNA sequence [17]. The grouping with Firmicutes is supported by the overall level of similarity of its proteome to other species [18]. We therefore believe the rRNA-based classification is incorrect and that *S. thermophilum* should be transferred to the Firmicutes. Such inaccuracies of the rRNA trees are relatively uncommon and may in this case be due to the mixing of thermophilic and non-thermophilic species into one group. This can cause artifacts when using rRNA genes for phylogenetic reconstruction since the G+C content of rDNA is strongly correlated to optimal growth temperature.

CO Dehydrogenases and Life in CO

Anaerobic species that make use of CO do so using nickel-iron CO dehydrogenase (CODH) complexes [19,20]. These enzymes all appear to catalyze the anaerobic interconversion of CO and CO_2. However, they vary greatly in the cellular role of this conversion and in the exact structure of the complex [19]. Analysis of the genome reveals the presence of five genes encoding homologs of CooS, the catalytic subunit of

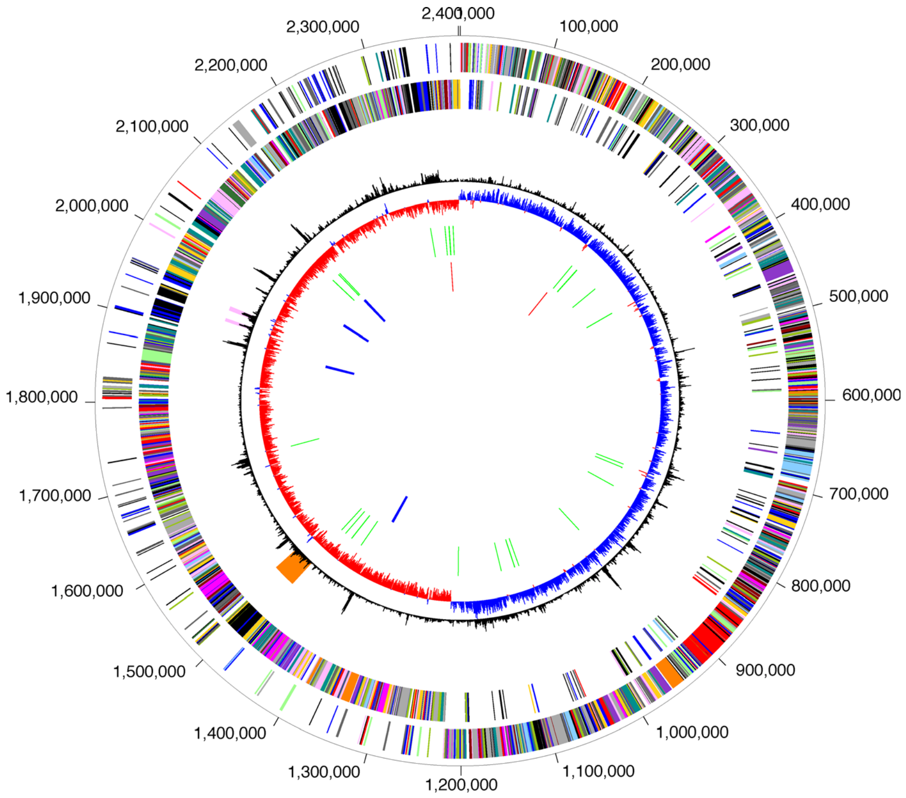


Figure 1. Genomic Organization of *C. hydrogenoformans*

From the outside inward the circles show: (1, 2) predicted protein-coding regions on the plus and minus strands (colors were assigned according to the color code of functional classes; (3) prophage (orange) and CRISPR (pink) regions; (4) χ^2-square score of tri-nucleotide composition; (5) GC skew (blue indicates a positive value and red a negative value); (6) tRNAs (green); (7) rRNAs (blue) and structural RNAs (red).
DOI: 10.1371/journal.pgen.0010065.g001

anaerobic CODHs. These five CooS encoding genes are scattered around the genome, and analysis of genome context, gene phylogeny, and experimental studies in this and other CO-utilizing species suggests they are subunits of five distinct CODH complexes, which we refer to as CODH I-V (Figure 3). The CooS homologs are named accordingly.

Table 1. General Features of the *C. hydrogenoformans* Genome

Feature	Value
Genome size, bp	2,401,892
% G+C	42.0
Predicted protein coding genes (CDSs)	2646
Average CDS length	827
Percent of genome that is coding	91.1
CDSs with assigned function	1512 (57.1%)
Conserved hypothetical CDS[a]	354 (13.4%)
Unknown function CDS[b]	331 (12.5%)
Hypothetical CDS[c]	449 (17.0%)
Transfer RNA	50
Ribosomal RNA	12
Structural RNAs	2
CRISPR regions	2
Prophage	1

[a] Match to genes in other species, but no function known.
[b] Some biochemical function prediction, but cellular role not predictable.
[c] No match to genes in other species.
DOI: 10.1371/journal.pgen.0010065.t001

Specific details about each complex and proposed physiological roles are given in the following paragraphs.

Energy Conservation (CODH-I)

A catalytic subunit (CooS-I, CHY1824) and an electron transfer protein (CooF, CHY1825) of CODH are encoded immediately downstream of a hydrogenase gene cluster (*cooMKLXUH,* CHY1832–27) that is closely related to the one found in *Rhodospirillum rubrum* [21]. These eight proteins form a tight membrane-bound enzyme complex that converts CO to CO_2 and H_2 in vitro [1,22]. In *R. rubrum,* this CODH/hydrogenase complex was proposed to be the site of CO-driven proton respiration where energy is conserved in the form of a proton gradient generated across the cell membrane [21]. Based on the high similarities in protein sequences and their gene organization, this set of genes were suggested to play a similar role in energy conservation in *C. hydrogenoformans* [1]. Consistent with this, this *cooS* gene is in the same subfamily as that from *R. rubrum* (Figure 4).

Carbon Fixation (CODH-III)

Anaerobic bacteria and archaea, such as methanogens and acetogens, can fix CO or CO_2 using the acetyl-CoA pathway (also termed the Wood-Ljungdahl pathway), where two molecules of CO_2, through a few steps, are condensed into one acetyl-CoA, a key building block for cellular biosynthesis and an important source of ATP [23]. The key enzyme of the final step (a CODH/acetyl-CoA synthase complex) has been purified from *C. hydrogenoformans* (strain DSM 6008) cultured

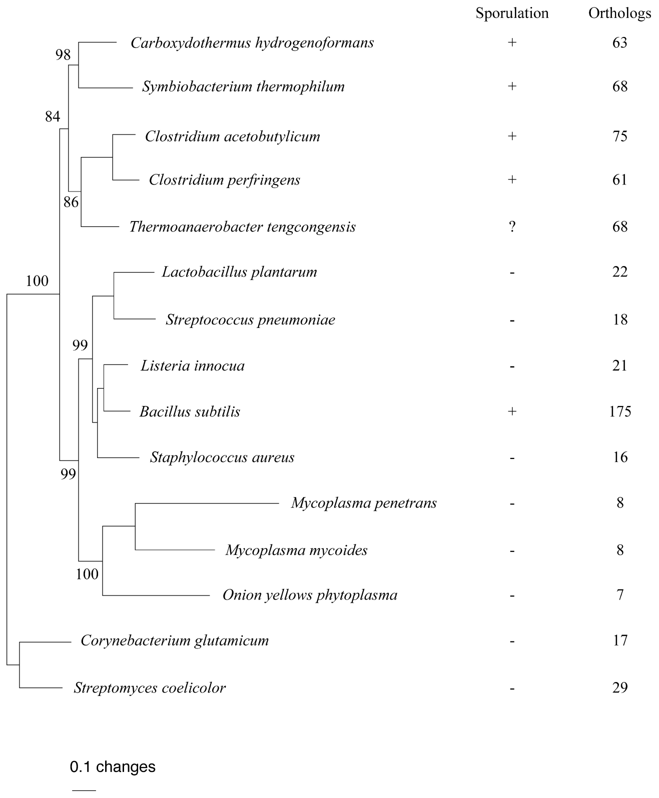


Figure 2. Genome Tree of Representatives of Firmicutes

A maximum likelihood tree was built from concatenated protein sequences of 31 universal housekeeping genes and rooted by two outgroup Actinobacteria (high GC Gram-positives) species: *Corynebacterium glutamicum* and *Streptomyces coelicolor.* Bootstrap support values (out of 100 runs) for branches of interest are shown beside them. Each species' ability to sporulate and its number of putative orthologs of the 175 known *B. subtilis* sporulation genes are also shown.

DOI: 10.1371/journal.pgen.0010065.g002

under limited CO supply and shown to be functional in vitro [24]. Genes encoding this complex and other proteins predicted to be in this pathway are clustered in the genome (CHY1221–7). This cluster is very similar to the *acs* operon from the acetogen *Moorella thermoacetica* which encodes the acetyl-CoA pathway machinery [25]. The phylogenetic tree also shows that CooS-III is in the same subfamily as the corresponding gene in the *M. thermoacetica acs* operon (Figure 4), suggesting they have the same biological functions. In addition, all the genes in the acetyl-CoA pathway have been identified in the *C. hydrogenoformans* genome and activities of some of those gene products have been detected (Figure 5), prompting us to propose that this organism carries out autotrophic fixation of CO through this pathway. This is consistent with the observation that key enzymes for the other known CO_2 fixation pathways, such as the Calvin cycle, the reverse tricarboxylic acid cycle, and 3-hydroxypropionate cycle are apparently not encoded in the genome.

Oxidative Stress Response (CODH-IV)

C. hydrogenoformans, though an anaerobe, has to deal with oxidative challenges present in the environment from time to time. Unlike aerobes, many anaerobes are proposed to use an alternative oxidative stress protection mechanism that depends on proteins such as rubrerythrin [26,27]. With few exceptions, rubrerythrin-like proteins have been found in complete genomes of all anaerobic and microaerophilic microbes but are absent in aerobic microbes [28]. Rubrerythrin is thought to play a role in the detoxification of reactive oxygen species by reducing the intermediate hydrogen peroxide, although the exact details remain elusive [28,29]. *C. hydrogenoformans* encodes three rubrerythrin homologs. One of them forms an operon with genes encoding CooS-IV, a CooF homolog, and a NAD/FAD-dependent oxidoreductase (CHY0735–8, Figure 3), suggesting that their functions are related. Here we speculate that this operon encodes a multisubunit complex where electrons stripped from CO by the CODH are passed to rubrerythrin to reduce hydrogen peroxide to water, with CooF and the NAD/FAD-dependent oxidoreductase acting as the intermediate electron carriers. Therefore, CODH-IV may play an important role in oxidative stress response by providing the ultimate source of reductants.

Others

Two other homologs of CooS are encoded in the genome. The gene encoding CooS-II (CHY0085) was originally cloned with the neighboring *cooF* (CHY0086) [30] and the complex was purified as functional homodimers [1]. This complex (CODH-II) is membrane-associated and an in vitro study showed it might have an anabolic function of generating NADPH [1]. Its structure has been solved [31]. The role of CooS-V (CHY0034) is more intriguing as it is the most deeply branched of the CooSs (Figure 4) and is not flanked by any genes with obvious roles in CO-related processes.

Aerobic bacteria metabolize CO using drastically different CODHs that are unrelated to the anaerobic ones. The CODHs from aerobes are dimers of heterotrimers composed of a molybdoprotein (CoxL), a flavoprotein (CoxM), and an iron-sulfur protein (CoxS) and belong to a large family of molybdenum hydroxylases including aldehyde oxidoreductases and xanthine dehydrogenases [32]. These enzymes characteristically demonstrate high affinity for CO, and the oxidation is typically coupled to CO_2 fixation via the reductive pentose phosphate cycle.

C. hydrogenoformans has one gene cluster (CHY0690–2) homologous to the *coxMSL* cluster in *Oligotropha carboxidovorans,* the most well-studied aerobic CODHs. However, our phylogenetic analysis showed that the *C. hydrogenoformans* homolog of CoxL does not group within the CODH subfamily. Therefore, we conclude that it is unlikely that this gene cluster in *C. hydrogenoformans* encodes a CODH, although that needs to be tested. Of the available published and unpublished genomes, only *R. rubrum* appears to have both an anaerobic CODH and a close relative of the aerobic *O. carboxidovorans* CODH. Accordingly, *R. rubrum,* a photosynthetic bacterium, can grow in the dark both aerobically and anaerobically using CO as an energy source.

Structures of both the Mo- and Ni-containing enzymes have been published recently. The crystal structure of CooS-II from *C. hydrogenoformans* is a dimeric enzyme with dual Ni-containing reaction centers each connected to the enzyme surface by 70-Å hydrophobic channels through which CO transits [31]. This channeling, also confirmed experimentally [33,34], explains the mechanism of CO use as a central metabolic intermediate despite its low solubility and generally low concentration in geothermal environments.

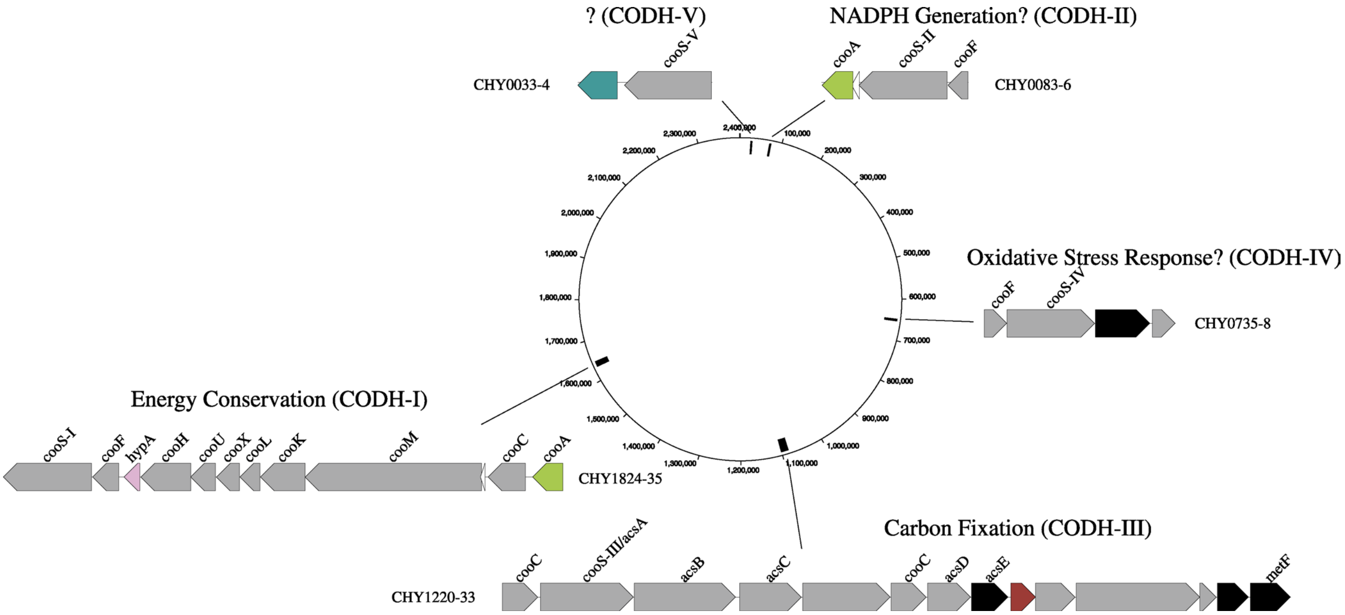


Figure 3. Genome Locations of Genes Predicted to Encode Five CODH Complexes

The genome locations of the genes encoding the five CooS homologs (labelled CooS I-V) are shown. Also shown are neighboring genes that are predicted to encode the five distinct CODH complexes (CODH I-V) with each CooS homolog. Possible cellular roles for four of the five CODH complexes are indicated.

DOI: 10.1371/journal.pgen.0010065.g003

Sporulation

The *C. hydrogenoformans* genome encodes a large number of homologs of genes involved in sporulation in other Firmicutes, spanning all stages of sporulation (Table 2). Among those are the master switch gene *spo0A* and all sporulation-specific sigma factors, σ^H, σ^E, σ^F, σ^G, and σ^K. However, sporulation has not been previously reported for this species. With this in mind, we set out to re-examine the morphology of *C. hydrogenoformans* cells and found endospore-like structures when cultures were stressed (Figure 6).

We then used phylogenetic profile analysis to look for other possible sporulation genes in the genome. Phylogenetic profiling works by grouping genes according to their distribution patterns in different species [35]. Proteins that function in the same pathways or structural complexes frequently have correlated distribution patterns. Phylogenetic profile analysis identified an additional set of 37 potential sporulation-related genes (Figure 7). Those genes are generally *Bacillales*- and *Clostridiales*-specific, consistent with the fact that endospores have so far only been found in these and other closely related Firmicutes. Most of the novel genes are conserved hypothetical proteins, whereas a few are putative membrane proteins. In support, a few of those novel sporulation genes have been shown to be involved in *Bacillus subtilis* sporulation by experimental studies [36,37]. The rest of the genes are thus excellent candidates for encoding known sporulation functions that have not been assigned to genes or previously unknown sporulation activities. Strikingly, within this group of genes, in addition to other known sporulation-specific sigma factors (σ^E, σ^F, σ^G, and σ^K), we identified a sigma factor (CHY1519) that was not known to be associated with sporulation previously. σ^I, its putative ortholog in *B. subtilis,* has shown some association with heat shock [38]. It remains to be determined experimentally whether this sigma factor is involved in sporulation, and if so, the regulatory network it controls.

A search of known sporulation-related genes in *B. subtilis* against *C. hydrogenoformans* revealed that many of them are missing in the genome. Of the 175 *B. subtilis* sporulation-related genes we compiled from the genome annotation and literature [39,40], half have no detectable homologs in *C. hydrogenoformans* using BLASTP with an E-value cutoff of 1e-5. Putative orthologs defined by mutual-best-hit methodology are present for only one third of those genes in *C. hydrogenoformans.* Among those missing genes are *spo0B* and *spo0F,* which encode the key components of the complex phosphorelay pathway in *B. subtilis* that channels various signals such as DNA damage, the ATP level, and cell density to the master switch protein Spo0A and therefore governs the cell's decision to enter sporulation. *C. hydrogenoformans* hence uses either a simplified version of this pathway or an alternative signal transduction pathway to sense the environmental or physiological stimuli. A large number of genes involved in the protective outer layer (cortex, coat, and exosporium) formation, spore germination, and small acid-soluble spore protein synthesis, among a few genes in various stages of spore development, are also missing. A similar, but slightly different, set of genes are missing in the other spore-forming *Clostridia* species as well [41]. Absence of those genes is more pronounced in non-spore-forming Firmicutes such as *Listeria* spp., *Staphylococcus* spp., and *Streptococcus* spp., as they lack all the sporulation-specific genes. When overlaid onto the phylogeny of Firmicutes (Figure 2), this observation can be explained by either multiple independent gene-loss events along branches leading to non-*Bacillus* species or by independent gene-gain events along branches leading to *Bacillus* and *Clostridia,* or by both. Whatever the history is of the sporulation evolution, the core set of sporulation genes

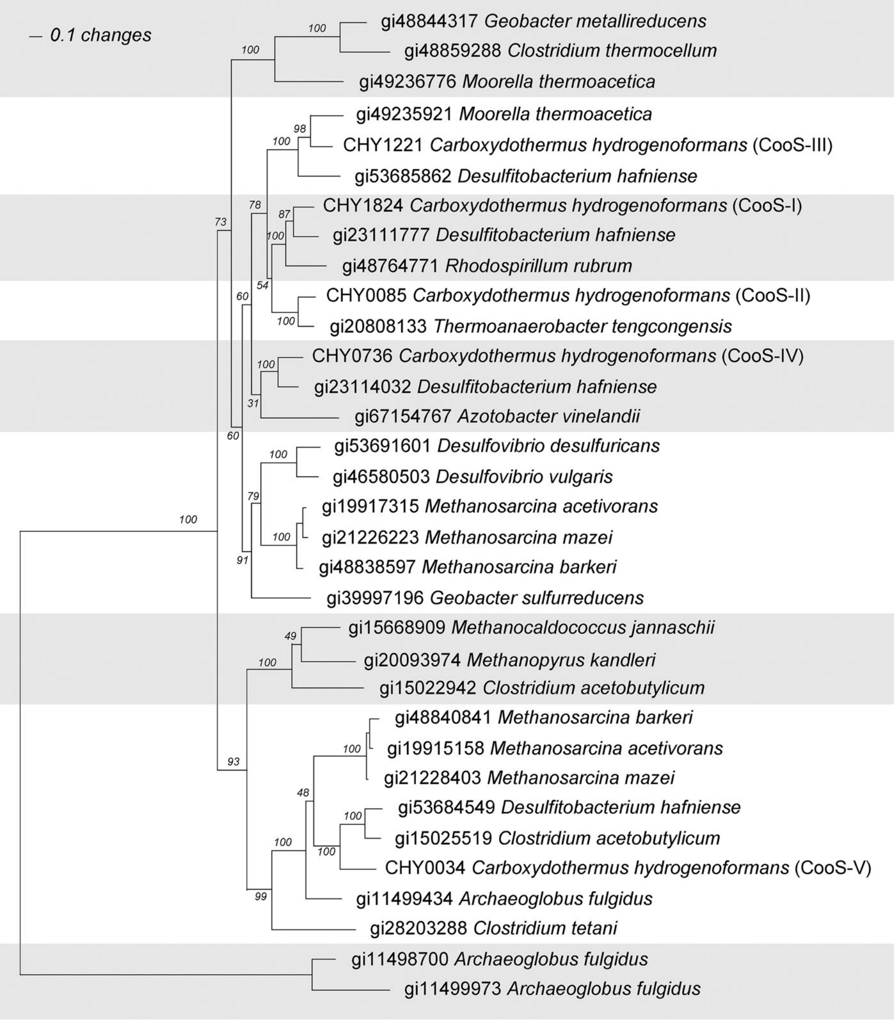

— 0.1 changes
gi48844317 Geobacter metallireducens
gi48859288 Clostridium thermocellum
gi49236776 Moorella thermoacetica
gi49235921 Moorella thermoacetica
CHY1221 Carboxydothermus hydrogenoformans (CooS-III)
gi53685862 Desulfitobacterium hafniense
CHY1824 Carboxydothermus hydrogenoformans (CooS-I)
gi23111777 Desulfitobacterium hafniense
gi48764771 Rhodospirillum rubrum
CHY0085 Carboxydothermus hydrogenoformans (CooS-II)
gi20808133 Thermoanaerobacter tengcongensis
CHY0736 Carboxydothermus hydrogenoformans (CooS-IV)
gi23114032 Desulfitobacterium hafniense
gi67154767 Azotobacter vinelandii
gi53691601 Desulfovibrio desulfuricans
gi46580503 Desulfovibrio vulgaris
gi19917315 Methanosarcina acetivorans
gi21226223 Methanosarcina mazei
gi48838597 Methanosarcina barkeri
gi39997196 Geobacter sulfurreducens
gi15668909 Methanocaldococcus jannaschii
gi20093974 Methanopyrus kandleri
gi15022942 Clostridium acetobutylicum
gi48840841 Methanosarcina barkeri
gi19915158 Methanosarcina acetivorans
gi21228403 Methanosarcina mazei
gi53684549 Desulfitobacterium hafniense
gi15025519 Clostridium acetobutylicum
CHY0034 Carboxydothermus hydrogenoformans (CooS-V)
gi11499434 Archaeoglobus fulgidus
gi28203288 Clostridium tetani
gi11498700 Archaeoglobus fulgidus
gi11499973 Archaeoglobus fulgidus
100
100
98
100
73
78
87
100
60
54
100
100
31
60
100
79
100
100
91
49
100
100
93
48
100
100
100
99

Figure 4. Phylogenetic Tree of CooS Homologs

The figure shows a maximum-likelihood tree of CooS homologs. The tree indicates the five CooS homologs in *C. hydrogenoformans* are not the result of recent duplications but instead are from distinct subfamilies. The other CooS homologs included in the tree were obtained from the NCBI nr database and include some from incomplete genome sequences generated by United States Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/).

DOI: 10.1371/journal.pgen.0010065.g004

shared by *Bacillus* and *Clostridia* might be close to a "minimal" sporulation set, as so far only these two groups have been found to be capable of producing endospores. Alternatively, some spore specific functions may be carried out by non orthologous genes in different species, which would prevent us from identifying them by this type of analysis.

Strictly Dependent on CO?

Until very recently, *C. hydrogenoformans* was thought to be an autotroph strictly depending on CO for growth. An overview of the genome reveals features related to its autotrophic lifestyle. For example, it has lost the entire sugar phosphotransferase system and encodes no complete pathway for sugar compound degradation. However, many aspects of the gene repertoire are suggestive of heterotrophic capabilities. For example, among the transporters encoded in the genome are ones predicted to import diverse carbon compounds including formate, glycerol, lactate, C4-dicarboxylate (malate, fumarate, or succinate; the binding receptor for this has three paralogs in the genome), 2-keto-3-deoxygluconate, 2-oxoglu-

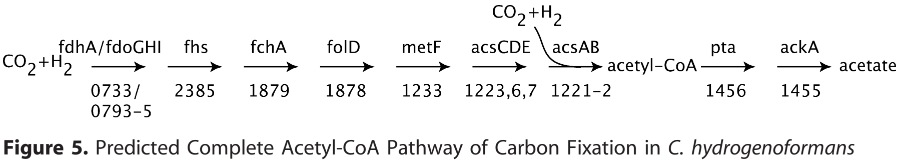


Figure 5. Predicted Complete Acetyl-CoA Pathway of Carbon Fixation in *C. hydrogenoformans*

Genes predicted to encode each step in the acetyl-CoA pathway of carbon fixation were identified in the genome. The locus numbers are indicated on the figure.

DOI: 10.1371/journal.pgen.0010065.g005

Table 2. Orthologs of Known *Bacillus subtilis* Sporulation Genes in *C. hydrogenoformans*

Locus	Gene	Description
CHY1978	*spo0A*	Stage 0 sporulation protein A
CHY0010	*spo0J*	Stage 0 sporulation protein J
CHY0370	*obg*	spo0B-associated GTP-binding protein
CHY0009	*soj*	Sporulation initiation inhibitor protein soj
CHY1960	*spoIIAB*	Anti-sigma F factor
CHY2541	*spoIID*	Stage II sporulation protein D
CHY1517	*spoIID*	Putative stage II sporulation protein D
CHY0212	*spoIIE*	Putative stage II sporulation protein E
CHY2057	*spoIIGA*	Putative sporulation specific protein SpoIIGA
CHY1965	*spoIIM*	Putative stage II sporulation protein M
CHY1923	*spoIIP*	Putative stage II sporulation protein P
CHY0408	*spoIIP*	Putative sporulation protein
CHY2054	*spoIIR*	Stage II sporulation protein R
CHY0206		Putative stage II sporulation protein D
CHY2007	*spoIIIAA*	Putative sporulation protein
CHY2006	*spoIIIAB*	Putative sporulation protein
CHY2005	*spoIIIAC*	Putative sporulation protein
CHY2004	*spoIIIAD*	Putative sporulation protein
CHY2003	*spoIIIAE*	Putative sporulation protein
CHY2001	*spoIIIAG*	Putative sporulation protein
CHY2534	*spoIIID*	Stage III sporulation protein D
CHY1159	*spoIIIE*	DNA translocase FtsK
CHY0004	*spoIIIJ*	Sporulation associated-membrane protein
CHY1916	*spoIVA*	Stage IV sporulation protein A
CHY1979	*spoIVB*	Putative stage IV sporulation protein B
CHY1957	*spoVAC*	Stage V sporulation protein AC
CHY1956	*spoVAD*	Stage V sporulation protein AD
CHY1955	*spoVAE*	Stage V sporulation protein AE
CHY0960	*spoVB*	Stage V sporulation protein B
CHY1152	*spoVFA*	Dipicolinate synthase, A subunit
CHY1153	*spoVFB*	Dipicolinate synthase, B subunit
CHY1391	*spoVK*	Stage V sporulation protein K
CHY1202	*spoVR*	Stage V sporulation protein R
CHY1171	*spoVS*	Stage V sporulation protein S
CHY0202	*spoVT*	Stage V sporulation protein T
CHY2272	*cotJC*	cotJC protein
CHY0786	*cotJC*	cotJC protein
CHY1463	*sspD*	Small acid-soluble spore protein
CHY1464	*sspD*	Small acid-soluble spore protein
CHY1175	*sspF*	Small acid-soluble spore protein
CHY1465		Putative small acid-soluble spore protein
CHY1941	*spmA*	Spore maturation protein A
CHY1940	*spmB*	Spore maturation protein B
CHY0958		Small acid-soluble spore protein
CHY1160		Putative spore cortex-lytic enzyme
CHY1756	*sleB*	Putative spore cortex-lytic enzyme
CHY0336	*gerKA*	Spore germination protein GerKA
CHY1404	*gerKB*	Spore germination protein
CHY0337	*gerKC*	Spore germination protein
CHY0305	*gerM*	Putative germination protein GerM
CHY1950		Putative spore germination protein
CHY0143		RNA polymerase sigma factor
CHY2056	*sigE*	RNA polymerase sigma-E factor
CHY1959	*sigF*	RNA polymerase sigma-F factor
CHY2055	*sigG*	RNA polymerase sigma-G factor
CHY2333	*sigH*	RNA polymerase sigma-H factor
CHY0617	*sigK*	RNA polymerase sigma-K factor
CHY1462	*gpr*	Spore protease
CHY2672		Sigma-K processing regulatory protein BofA
CHY0424		Putative sporulation protein

DOI: 10.1371/journal.pgen.0010065.t002

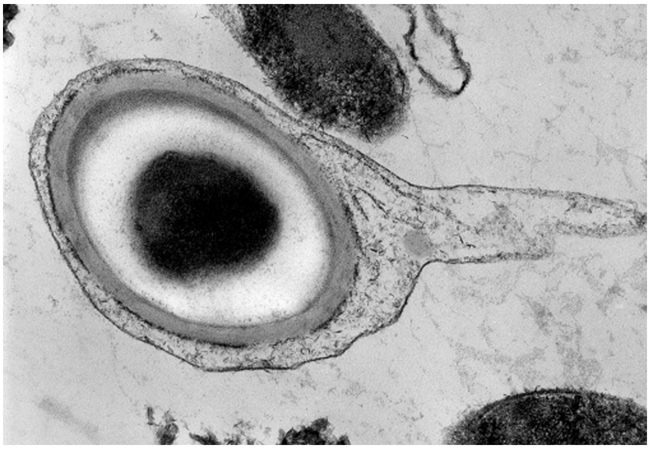


Figure 6. An Electron Micrograph of a *C. hydrogenoformans* Endospore

The finding of homologs of many genes involved in sporulation in other species led us to test whether *C. hydrogenoformans* also could form an endospore. Under stressful growth conditions, endospore-like structures form. We note that even though homologs could not be found in the genome for many genes that in other species are involved in protective outer-layer (cortex, coat, and exosporium) formation, those structures seem to be visible and intact.

DOI: 10.1371/journal.pgen.0010065.g006

tarate, and amino acids. In addition there is a diverse array of signal transduction pathways including chemotaxis not commonly found in the genomes of autotrophs (see below). Consistent with these observations, Henstra et al. recently showed that formate, lactate, and glycerol could be utilized as carbon source provided 9,10-anthraquinone-2,6-disulfonate was used as the electron acceptor [8]. Similarly, sulfite, thiosulfate, sulfur, nitrate, and fumarate were reduced with lactate as electron donor, although heterotrophic growth was relatively slow compared with cultures growing on pure CO [8]. It is not known what electron acceptors are likely to be coupled to these pathways in the isolation locale of *C. hydrogenoformans,* however it is clear that there is a more versatile complement of energy sources than initially concluded by Svetlichny et al. [2].

In terms of autotrophic lifestyle, although *C. hydrogenoformans* and *S. thermophilum* are close phylogenetically, they have gone separate ways in their lifestyles. *S. thermophilum* is an uncultivable thermophilic bacterium growing as part of a microbial consortium [18], while *C. hydrogenoformans* is a hot-spring autotroph that can survive efficiently on CO as its sole carbon and energy source. Accordingly, their metabolic capabilities are very different and only half of their proteomes are homologous. It is not clear why *S. thermophilum* is dependent on other microbes. Unlike other symbiotic microorganisms, no large-scale genome reductions have occurred in *S. thermophilum* [18]. On the other hand, *C. hydrogenoformans* has evolved to live preferably on CO, possibly by acquiring and/or expanding its complement of CODHs. As a result, it has lost many genes associated with a heterotrophic lifestyle, such as the phosphotransferase transporter system, and may be on the verge of becoming an obligate autotroph. Even though *C. hydrogenoformans* is more closely related to *S. thermophilum* than to *T. tengcongensis,* an anaerobic thermophile isolated also from freshwater hot springs [42], *C. hydrogenoformans* actually shares slightly less genes with *S. thermophilum* than with *T. tengcongensis.*

Signal Transduction

C. hydrogenoformans is poised to respond to diverse environmental cues through a suite of signal transduction pathways

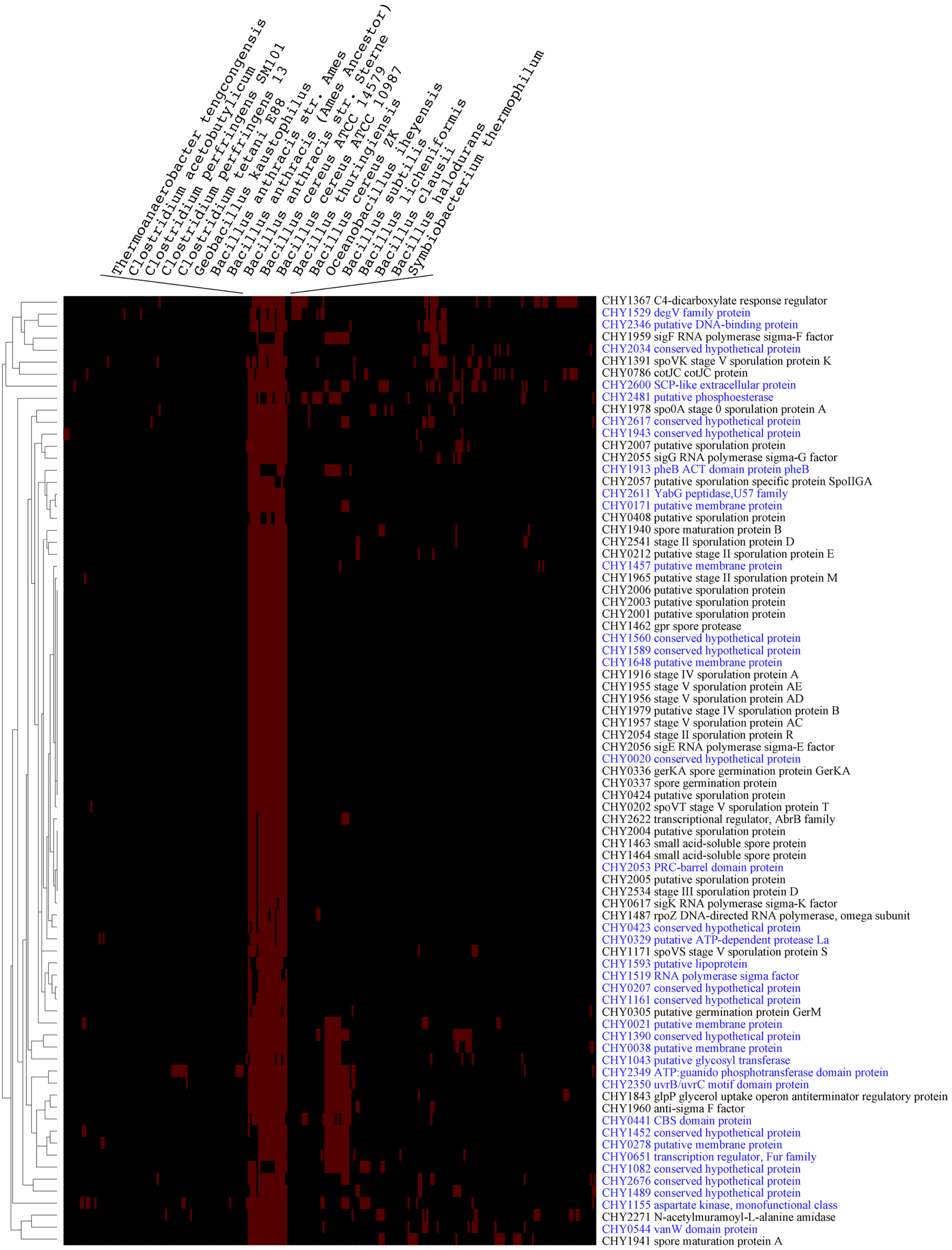

Thermoanaerobacter tengcongensis
Clostridium acetobutylicum
Clostridium perfringens SM101
Clostridium perfringens 13
Clostridium tetani E88
Geobacillus kaustophilus
Bacillus anthracis str. Ames
Bacillus anthracis (Ames Ancestor)
Bacillus anthracis str. Sterne
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Bacillus thuringiensis
Bacillus cereus ZK
Oceanobacillus iheyensis
Bacillus subtilis
Bacillus licheniformis
Bacillus clausii
Bacillus halodurans
Symbiobacterium thermophilum
CHY1367 C4-dicarboxylate response regulator
CHY1529 degV family protein
CHY2346 putative DNA-binding protein
CHY1959 sigF RNA polymerase sigma-F factor
CHY2034 conserved hypothetical protein
CHY1391 spoVK stage V sporulation protein K
CHY0786 cotJC cotJC protein
CHY2600 SCP-like extracellular protein
CHY2481 putative phosphoesterase
CHY1978 spo0A stage 0 sporulation protein A
CHY2617 conserved hypothetical protein
CHY1943 conserved hypothetical protein
CHY2007 putative sporulation protein
CHY2055 sigG RNA polymerase sigma-G factor
CHY1913 pheB ACT domain protein pheB
CHY2057 putative sporulation specific protein SpoIIGA
CHY2611 YabG peptidase,U57 family
CHY0171 putative membrane protein
CHY0408 putative sporulation protein
CHY1940 spore maturation protein B
CHY2541 stage II sporulation protein D
CHY0212 putative stage II sporulation protein E
CHY1457 putative membrane protein
CHY1965 putative stage II sporulation protein M
CHY2006 putative sporulation protein
CHY2003 putative sporulation protein
CHY2001 putative sporulation protein
CHY1462 gpr spore protease
CHY1560 conserved hypothetical protein
CHY1589 conserved hypothetical protein
CHY1648 putative membrane protein
CHY1916 stage IV sporulation protein A
CHY1955 stage V sporulation protein AE
CHY1956 stage V sporulation protein AD
CHY1979 putative stage IV sporulation protein B
CHY1957 stage V sporulation protein AC
CHY2054 stage II sporulation protein R
CHY2056 sigE RNA polymerase sigma-E factor
CHY0020 conserved hypothetical protein
CHY0336 gerKA spore germination protein GerKA
CHY0337 spore germination protein
CHY0424 putative sporulation protein
CHY0202 spoVT stage V sporulation protein T
CHY2622 transcriptional regulator, AbrB family
CHY2004 putative sporulation protein
CHY1463 small acid-soluble spore protein
CHY1464 small acid-soluble spore protein
CHY2053 PRC-barrel domain protein
CHY2005 putative sporulation protein
CHY2534 stage III sporulation protein D
CHY0617 sigK RNA polymerase sigma-K factor
CHY1487 rpoZ DNA-directed RNA polymerase, omega subunit
CHY0423 conserved hypothetical protein
CHY0329 putative ATP-dependent protease La
CHY1171 spoVS stage V sporulation protein S
CHY1593 putative lipoprotein
CHY1519 RNA polymerase sigma factor
CHY0207 conserved hypothetical protein
CHY1161 conserved hypothetical protein
CHY0305 putative germination protein GerM
CHY0021 putative membrane protein
CHY1390 conserved hypothetical protein
CHY0038 putative membrane protein
CHY1043 putative glycosyl transferase
CHY2349 ATP:guanido phosphotransferase domain protein
CHY2350 uvrB/uvrC motif domain protein
CHY1843 glpP glycerol uptake operon antiterminator regulatory protein
CHY1960 anti-sigma F factor
CHY0441 CBS domain protein
CHY1452 conserved hypothetical protein
CHY0278 putative membrane protein
CHY0651 transcription regulator, Fur family
CHY1082 conserved hypothetical protein
CHY2676 conserved hypothetical protein
CHY1489 conserved hypothetical protein
CHY1155 aspartate kinase, monofunctional class
CHY2271 N-acetylmuramoyl-L-alanine amidase
CHY0544 vanW domain protein
CHY1941 spore maturation protein A

Figure 7. Phylogenetic Profile Analysis of Sporulation in *C. hydrogenoformans*

For each protein encoded by the *C. hydrogenoformans* genome, a profile was created of the presence or absence of orthologs of that protein in the predicted proteomes of all other complete genome sequences. Proteins were then clustered by the similarity of their profiles, thus allowing the grouping of proteins by their distribution patterns across species. Examination of the groupings showed one cluster consisting of mostly homologs of sporulation proteins. This cluster is shown with *C. hydrogenoformans* proteins in rows (and the prediced function and protein ID indicated on the right) and other species in columns with presence of a ortholog indicated in red and absence in black. The tree to the left represents the portion of the cluster diagram for these proteins. Note that most of these proteins are found only in a few species represented in red columns near the center of the diagram. The species corresponding to these columns are indicated. We also note that though most of the proteins in this cluster, for which functions can be predicted, are predicted to be involved in sporulation and some have no predictable functions (highlighted in blue). This indicates that functions of these proteins' homologs have not been characterized in other species. Since these proteins show similar distribution patterns to so many proteins with roles in sporulation, we predict that they represent novel sporulation functions.

DOI: 10.1371/journal.pgen.0010065.g007

and processes. The organism has 83 one-component regulators and 13 two-component systems (including two chemotaxis systems), which are average numbers for such a genome size [43] (Table S1). Many of the genes encoding these two-component systems are next to transporters, possibly being involved in regulation of solute uptake, while others are adjacent to oxidoreductases. *C. hydrogenoformans* also possesses an elaborate cascade of chemotaxis genes, including 11 chemoreceptors, and a complete set of flagellar genes, most located within a large cluster of about 70 genes (CHY0963–1033). Chemotaxis allows microbes to respond to environmental stimuli by swimming toward nutrients or away from toxic chemicals. Generally, a heavy commitment to chemotaxis is not a characteristic of autotrophic microorganisms [44], and it is possible that *C. hydrogenoformans* is responding to gradients of inorganic nutrients, or gases such as CO, O_2, H_2, or CO_2.

Critical for sensing CO, two CooA homologs occur in the *C. hydrogenoformans* genome, both of which are encoded within operons containing *cooS* genes. CooA proteins are heme proteins that act as both sensors for CO as well as transcriptional regulators. They belong to the cyclic adenosine monophosphate receptor protein family and induce CO-related genes upon CO binding [45]. CHY1835, encoding CooA1, is at the beginning of the *R. rubrum*-like *coo* operon. CHY0083, encoding CooA2, is at the end of the operon possibly involved in NADPH generation from CO [1] (Figure 3).

C. hydrogenoformans lacks certain subfamilies of transcription factors that are present in its close *Clostridia* relatives, such as those utilizing the following helix-turn-helix domains: iron-dependent repressor DNA-binding domain, LacI, PadR, and DeoR (Pfam nomenclature). The genome does not encode any proteins of the LuxR family, which are usually abundant in both one-component (e.g., quorum-sensing regulators) and two-component systems.

The largest family of transcriptional regulators in *C. hydrogenoformans* is sigma-54- dependent activators. Eight such regulators comprise one-component systems (CHY0581, CHY0788, CHY1254, CHY1318, CHY1359, CHY1376, CHY1547, and CHY2091) and another one is a response regulator of the two-component system (CHY1855). Seven one-component sigma-54-dependent regulators have at least one PAS domain as a sensory module. PAS domains are known to often contain redox-responsive cofactors, such as FAD, FMN, and heme and serve as intracellular oxygen and redox sensors [46]. Overall, there are 18 PAS domains in *C. hydrogenoformans.* It is a very significant number compared to only two PAS domains in *Moorella thermoacetica* (similar genome size) and nine in *Desulfitobacterium hafniense* (a much larger genome). The most abundant sensory domain of bacterial signal transduction, the LysR substrate-binding domain, which binds small molecule ligands, is present only in six copies in *C. hydrogenoformans* (there are 36 copies in *D. hafniense*), re-enforcing the notion that redox sensing via PAS domains might be the most critical signal transduction event for this organism.

The most intriguing signal transduction protein in *C. hydrogenoformans* is the sigma-54-dependent transcriptional regulator that has an iron hydrogenase-like domain as a sensory module (CHY1547). This domain contains 4Fe-4S clusters and is predicted to use molecular hydrogen for the reduction of a variety of substrates. Its fusion with the sigma-54 activator and the DNA-binding HTH__8 domain in the CHY1547 protein strongly suggests that this is a unique regulator that activates gene expression in *C. hydrogenoformans* in response to hydrogen availability. Interestingly, it is located immediately upstream of a ten-gene cluster encoding a Ni/Fe hydrogenase (CHY1537–46). Iron hydrogenases similar to the one in CHY1547 can be identified in several bacterial genomes including *S. thermophilum, Dehalococcoides ethenogenes,* and some *Clostridia;* however, they are not associated with DNA-binding domains. The only organisms where we found a homologous sigma-54 activator are *M. thermoacetica, Geobacter metallireducens, G. sufurreducens,* and *Desulfuromonas acetoxidans.*

Selenocysteine-Containing Proteins

C. hydrogenoformans possesses all known components of the selenocysteine (Sec) insertion machinery (CHY1803:SelA, CHY1802:SelB, CHY2058:SelD) and the Sec tRNA. A total of 12 selenocysteine-containing proteins (selenoproteins) were identified in *C. hydrogenoformans* genome by the Sec/Cys homology method (Table 3). For each of them, an mRNA stem-loop structure, the signature of the so-called Sec Insertion Sequence (SECIS) required for the Sec insertion, is present immediately downstream of the UGA codon. Although most of the identified selenoproteins are redox proteins, as has been shown for other bacteria and archaea [47], three are novel. Two are transporters (CHY0860, CHY0565), while the third is a methylated-DNA-protein-cysteine methyltransferase (CHY0809), a suicidal DNA repair protein that repairs alkylated guanine by transferring the alkyl group to the cysteine residue at its active site. It is striking that although this protein has been found in virtually every studied organism, only the one in *C. hydrogenoformans* has selenocysteine in place of cysteine at its active site. Therefore, this selenoprotein most likely evolved very recently, probably from a cysteine-containing protein. Similar patterns exist for the two selenocysteine-containing transporters, suggesting invention of new selenoproteins is an ongoing process in *C. hydrogenoformans.*

Translational Frameshifts

Analysis of the genome identified many potential cases of frameshifted genes. They are identified by having significant sequence similarity in two reading frames to a single homolog in another species. Examination of sequence traces suggests they are not sequencing errors. Some of these appear to be programmed frameshifts. Programmed frameshifting is a ubiquitous mechanism cells use to regulate translation or generate alternative protein products [48]. The frameshift in the gene *prfB* (CHY0163), encoding the peptide chain release factor 2, is a well-studied example of programmed frameshift that actually regulates its own translation [48].

However, many of the detected frameshifts appear to be the result of mutations from an ancestral un-frameshifted state. This is best exemplified by examination of the frameshift in the *cooS*-III gene (CHY1221), which as described above is predicted to encode one of the key components of the acetyl-CoA carbon fixation pathway. In cultures of another strain of this species (DSM 6008), a functional full-length (i.e., unframeshifted) version of this protein has been purified [24] and sequence comparisons of the gene from that strain with ours revealed many polymorphisms, including a deletion in our strain that gave rise to this frameshift (unpublished data). Studies of DSM 6008 show that in cultures grown in excess CO, the acetyl-CoA synthase (ACS, CHY1222) existed predominantly as monomer and only trace amount of CODH-III/ACS complex could be detected. On the other hand, when the CO supply was limited, CODH-III/ACS complex became the dominant form. It is plausible that CODH-III is not absolutely required for carbon fixation when the CO supply is high. Thus the frameshift and other mutations in *cooS-III* in Z-2901 may reflect the fact that it has been serially cultured in excess CO in the lab for many years. The putative lab-acquired mutations in Z-2901 are yet another reason to sequence type strains of species that have been directly acquired from culture collections and not submitted to extended laboratory culturing [49].

Conclusion

Living solely on CO is not a simple feat and the fact that *C. hydrogenoformans* does it so well makes it a model organism for this unusual metabolism. Our analysis of the genome sequence, and phylogenomic comparisons with other species, provide insights into this species' specialized metabolism. Perhaps most striking is the presence of genes that apparently encode five distinct carbon monoxide dehydrogenase complexes. Analysis of the genome has also revealed many new perspectives on the biology and evolution of this species, for example, leading us to propose its reclassification, providing further evidence that it is not a strict autotroph and revealing a previously unknown ability to sporulate. The analysis reported here and the availibility of the complete genome sequence should catalyze future studies of this organism and the hydrogenogens as a whole.

Table 3. Selenoproteins Identified in *C. hydrogenoformans* Genome

Locus	Description
CHY2058	Selenide, water dikinase
CHY2392	Glycine reductase, selenoprotein A
CHY2393	Glycine reductase, selenoprotein B
CHY0733	NAD-dependent formate dehydrogenase, alpha subunit
CHY0740	Dehydrogenase
CHY0793	Formate dehydrogenase-O, major subunit
CHY0809	Methylated-DNA-protein-cysteine methyltransferase
CHY0860	Cation-transporting ATPase, E1-E2 family
CHY0930	Heterodisulfide reductase, subunit A
CHY0931	Hydrogenase, methyl-viologen-reducing type, delta subunit
CHY1095	Thioredoxin domain selenoprotein/cytochrome C biogenesis family protein
CHY0565	Mercuric transport protein, putative

DOI: 10.1371/journal.pgen.0010065.t003

Materials and Methods

Medium composition and cultivation. *C. hydrogenoformans* Z-2901were cultivated under strictly anaerobic conditions in a basal carbonate-buffered medium composed as described [2]. However, 1.5 g l^{-1} $NaHCO_3$, 0.2 g l^{-1} $Na_2S \cdot 9\ H_2O$, 0.1 g l^{-1} yeast extract, and 2 μmol l^{-1} $NiCl_2$ were used instead of reported concentrations, and the Na_2S concentration was lowered to 0.04 g l^{-1}. Butyl rubber-stoppered bottles of 120 ml contained 50 ml medium. Bottles were autoclaved for 25' at 121 °C. Gas phases were pressurized to 170 kPa and were composed of 20% CO_2 and either 80% of N_2, H_2, or CO. Sporulation was induced by the addition of 0.01 mM $MnCl_2$ to the medium and by a transient heat shock treatment (100 °C for 5 min).

EM of *C. hydrogenoformans* endospore. Samples were fixed with 5% glutaraldehyde for 2 h and 1% OsO_4 for 4 h at 4 °C and then embedded in Epon-812. The thin sections were stained with uranyl acetate and lead citrate according to the method described by Miroshnichenko et al. [50]. The samples were observed and photographed using a JEOL JEM-1210 electron microscope.

Genome sequencing. Genomic DNA was isolated from exponential-phase cultures of *C. hydrogenoformans* Z-2901. This strain was acquired by Frank Robb from Vitali Svetlitchnyi (Bayreuth University, Germany) in 1995 after being serially grown in culture since its original isolation in 1990. Cloning, sequencing, assembly, and closure were performed as described [51,52]. The complete sequence has been assigned GenBank accession number CP000141 and is available at http://www.tigr.org.

Annotation. The gene prediction and annotation of the genome were done as previously described [51,52]. CDSs were identified by Glimmer [53]. Frameshifts or premature stop codons within CDSs were identified by comparison to other species and confirmed to be "authentic" by either their high quality sequencing reads or re-sequencing. Repetitive DNA sequences were identified using the REPUTER program [54].

Comparative genomics. To identify putative orthologs between two species, both of their proteomes were BLASTP searched against a local protein database of all complete genomes with an E-value cutoff of 1e-5. Species-specific duplications were identified and treated as one single gene (super-ortholog) for later comparison. Pair-wise mutual best-hits were then identified as putative orthologs.

Genome tree construction. Protein sequences of 31 housekeeping genes *(dnaG, frr, infC, nusA, pgk, pyrG, rplA, rplB, rplC, rplD, rplE, rplF, rplK, rplL, rplM, rplN, rplP, rplS, rplT, rpmA, rpoB, rpsB, rpsC, rpsE, rpsI, rpsJ, rpsK, rpsM, rpsS, smpB, tsf)* from genomes of interest were aligned to pre-defined HMM models and ambiguous regions were auto-trimmed according to an embedded mask. Concatenated alignments were then used to build a maximum likelihood tree using phyml [55].

Phylogenetic profile analysis. For each protein in *C. hydrogenoformans,* its presence or absence in every complete genome available at the time of this study was determined by asking whether a putative ortholog was present in that species (see above). Proteins were then grouped by their distribution patterns across species (bits of 1 and 0, 1 for presence and 0 for absence) using the CLUSTER program and the clusters were visualized using the TREEVIEW program (http://rana.lbl.gov/EisenSoftware.htm). Species were weighted by their closeness to each other to partially remove the phylogenetic component of the correlation [56].

Identification of selenoproteins. Each CDS of *C. hydrogenoformans* that ends with stop codon TGA was extended to the next stop codon TAA or TAG. It was then searched with BLASTP against the nraa database. A protein with a TGA codon pairing with a conserved Cys site was identified as a putative selenoprotein. The secondary structure of the mRNA immediately downstream of the TGA codon was also checked using MFOLD [57] to look for a possible stem-loop structure.

Supplemental Information

Table S1. Regulatory Genes in *Clostridia* Species

Found at DOI: 10.1371/journal.pgen.0010065.st001 (22 KB DOC).

Acknowledgments

We would like to thank The Institute for Genomic Research's (TIGR) Bioinformatics Department for supporting the infrastructure associated with genome annotation and analysis; the TIGR IT Department for general IT support; Dan Haft for discussions regarding selenoproteins; Anne Ciecko, Kristi Berry, and Chris Larkin for initial work on genome closure; Patrick Eichenberger, Richard Losick, and Jim Brannigan for discussions about sporulation; Vitali Svetlichnyi for general discussions about the organism; Terry Utterback and Tamara Feldblyum for coordinating the shotgun sequencing; William C. Nierman for making the genomic libraries; and Claire M. Fraser for supporting the selection of this genome as part of the DOE project referenced below. The sequencing, annotation, and analysis of the genome were supported by United States Department of Energy, Office of Biological Energy Research, Co-Operative Agreement DE-FC0295ER61962. Support for other work in association with this publication came from the National Science Foundation (NSF/MCB-02383387 for FTR and JMG) and the National Institutes of Health (GM072285 for LEU and IBZ).

Competing interests. The authors have declared that no competing interests exist.

Author contributions. MW, FTR, and JAE conceived and designed the experiments. JMG and FTR performed the physiology experiments. MW, QR, LEU, IBZ, FTR, and JAE analyzed the data. MW, IBZ, FTR, and JAE wrote the paper. LJT and KMJ closed the genome. ASD, SCD, LMB, RJD, RM, SAS, JFK, and WCN annotated the genome. ■

References

1. Svetlitchnyi V, Peschel C, Acker G, Meyer O (2001) Two membrane-associated NiFeS-carbon monoxide dehydrogenases from the anaerobic carbon-monoxide-utilizing eubacterium *Carboxydothermus hydrogenoformans*. J Bacteriol 183: 5134–5144.
2. Svetlichny VA, Sokolova TG, Gerhardt M, Ringpfeil M, Kostrikina NA, et al. (1991) *Carboxydothermus hydrogenoformans* gen. nov., sp. nov., a CO-utilizing thermophilic anaerobic bacterium from hydrothermal environments of Kunashir Island. Syst Appl Microbiol 14: 254–260.
3. Svetlichny V, Sokolova T, Kostrikina N, Lysenko A (1994) A new thermophilic anaerobic carboxydotrophic bacterium *Carboxydothermus restrictus* sp. nov. Mikrobiologiya 63: 294–297.
4. Bonch-Osmolovskaya E, Miroshnichenko M, Slobodkin A, Sokolova T, Karpov G, et al. (1999) Biodiversity of anaerobic lithotrophic prokaryotes in terrestrial hot springs of Kamchatka. Microbiology 68: 343–351.
5. Sokolova TG, Gonzalez JM, Kostrikina NA, Chernyh NA, Tourova TP, et al. (2001) *Carboxydobrachium pacificum* gen. nov., sp. nov., a new anaerobic, thermophilic, CO-utilizing marine bacterium from Okinawa Trough. Int J Syst Evol Microbiol 51: 141–149.
6. Sokolova TG, Gonzalez JM, Kostrikina NA, Chernyh NA, Slepova TV, et al. (2004) *Thermosinus carboxydivorans* gen. nov., sp. nov., a new anaerobic, thermophilic, carbon-monoxide-oxidizing, hydrogenogenic bacterium from a hot pool of Yellowstone National Park. Int J Syst Evol Microbiol 54: 2353–2359.
7. Sokolova TG, Kostrikina NA, Chernyh NA, Tourova TP, Kolganova TV, et al. (2002) *Carboxydocella thermautotrophica* gen. nov., sp. nov., a novel anaerobic, CO-utilizing thermophile from a Kamchatkan hot spring. Int J Syst Evol Microbiol 52: 1961–1967.
8. Henstra AM, Stams AJ (2004) Novel physiological features of *Carboxydothermus hydrogenoformans* and *Thermoterrabacterium ferrireducens*. Appl Environ Microbiol 70: 7236–7240.
9. Mojica FJ, Diez-Villasenor C, Soria E, Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of *Archaea, Bacteria,* and mitochondria. Mol Microbiol 36: 244–246.
10. Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the *Archaea Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. Mol Microbiol 17: 85–93.
11. Peng X, Brugger K, Shen B, Chen L, She Q, et al. (2003) Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. J Bacteriol 185: 2410–2417.
12. Rocha E (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 10: 393–395.
13. Dervyn E, Suski C, Daniel R, Bruand C, Chapuis J, et al. (2001) Two essential DNA polymerases at the bacterial replication fork. Science 294: 1716–1719.
14. Brewer BJ (1988) When polymerases collide: Replication and the transcriptional organization of the *E. coli* chromosome. Cell 53: 679–686.
15. Liu B, Alberts BM (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. Science 267: 1131–1137.
16. Garrity GM, Bell JA, Lilburn TG (2004) Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, Second Edition. New York: Springer-Verlag. 2816 p.
17. Ohno M, Shiratori H, Park MJ, Saitoh Y, Kumon Y, et al. (2000) *Symbiobacterium thermophilum* gen. nov., sp. nov., a symbiotic thermophile that depends on co-culture with a *Bacillus* strain for growth. Int J Syst Evol Microbiol 50 Pt 5: 1829–1832.
18. Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji TO, et al. (2004) Genome sequence of *Symbiobacterium thermophilum,* an uncultivable bacterium that depends on microbial commensalism. Nucleic Acids Res 32: 4937–4944.
19. Ferry JG (1995) CO dehydrogenase. Annu Rev Microbiol 49: 305–333.
20. Ragsdale SW, Kumar M (1996) Nickel-containing carbon monoxide dehydrogenase/acetyl-CoA synthase(,). Chem Rev 96: 2515–2540.
21. Fox JD, He Y, Shelver D, Roberts GP, Ludden PW (1996) Characterization of the region encoding the CO-induced hydrogenase of *Rhodospirillum rubrum*. J Bacteriol 178: 6200–6208.
22. Soboh B, Linder D, Hedderich R (2002) Purification and catalytic properties of a CO-oxidizing:H2-evolving enzyme complex from *Carboxydothermus hydrogenoformans*. Eur J Biochem 269: 5712–5721.
23. Ragsdale SW (2004) Life with carbon monoxide. Crit Rev Biochem Mol Biol 39: 165–195.
24. Svetlitchnyi V, Dobbek H, Meyer-Klaucke W, Meins T, Thiele B, et al. (2004) A functional Ni-Ni-[4Fe-4S] cluster in the monomeric acetyl-CoA synthase from *Carboxydothermus hydrogenoformans*. Proc Natl Acad Sci U S A 101: 446–451.
25. Roberts DL, James-Hagstrom JE, Garvin DK, Gorst CM, Runquist JA, et al. (1989) Cloning and expression of the gene cluster encoding key proteins involved in acetyl-CoA synthesis in *Clostridium thermoaceticum:* CO dehydrogenase, the corrinoid/Fe-S protein, and methyltransferase. Proc Natl Acad Sci U S A 86: 32–36.
26. Adams MW, Jenney FE Jr., Clay MD, Johnson MK (2002) Superoxide reductase: Fact or fiction? J Biol Inorg Chem 7: 647–652.
27. Lynch MC, Kuramitsu HK (1999) Role of superoxide dismutase activity in the physiology of *Porphyromonas gingivalis*. Infect Immun 67: 3367–3375.
28. Weinberg MV, Jenney FE Jr., Cui X, Adams MW (2004) Rubrerythrin from the hyperthermophilic archaeon *Pyrococcus furiosus* is a rubredoxin-dependent, iron-containing peroxidase. J Bacteriol 186: 7888–7895.
29. Sztukowska M, Bugno M, Potempa J, Travis J, Kurtz DM Jr. (2002) Role of rubrerythrin in the oxidative stress response of *Porphyromonas gingivali*s. Mol Microbiol 44: 479–488.
30. Gonzalez JM, Robb FT (2000) Genetic analysis of *Carboxydothermus hydrogenoformans* carbon monoxide dehydrogenase genes cooF and cooS. FEMS Microbiol Lett 191: 243–247.
31. Dobbek H, Svetlitchnyi V, Gremer L, Huber R, Meyer O (2001) Crystal structure of a carbon monoxide dehydrogenase reveals a [Ni-4Fe-5S] cluster. Science 293: 1281–1285.
32. Schubel U, Kraut M, Morsdorf G, Meyer O (1995) Molecular characterization of the gene cluster coxMSL encoding the molybdenum-containing carbon monoxide dehydrogenase of *Oligotropha carboxidovorans*. J Bacteriol 177: 2197–2203.
33. Maynard EL, Lindahl PA (1999) Evidence of a molecular tunnel connecting the active sites for CO2 reduction and acetyl-coA synthesis in acetyl-coA synthase from *Clostridium thermoaceticum*. J Am Chem Soc 121: 9221–9222.
34. Seravalli J, Ragsdale SW (2000) Channeling of carbon monoxide during anaerobic carbon dioxide fixation. Biochemistry 39: 1274–1277.
35. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A 96: 4285–4288.
36. Eichenberger P, Jensen ST, Conlon EM, van Ooij C, Silvaggi J, et al. (2003) The sigmaE regulon and the identification of additional sporulation genes in *Bacillus subtilis*. J Mol Biol 327: 945–972.
37. Molle V, Fujita M, Jensen ST, Eichenberger P, Gonzalez-Pastor JE, et al. (2003) The Spo0A regulon of *Bacillus subtilis*. Mol Microbiol 50: 1683–1701.
38. Zuber U, Drzewiecki K, Hecker M (2001) Putative sigma factor SigI (YkoZ) of *Bacillus subtilis* is induced by heat shock. J Bacteriol 183: 1472–1475.
39. Stragier P, Losick R (1996) Molecular genetics of sporulation in *Bacillus subtilis*. Annu Rev Genet 30: 297–241.
40. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249–256.
41. Nolling J, Breton G, Omelchenko MV, Makarova KS, Zeng Q, et al. (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. J Bacteriol 183: 4823–4838.

42. Bao Q, Tian Y, Li W, Xu Z, Xuan Z, et al. (2002) A complete sequence of the *T. tengcongensis* genome. Genome Res 12: 689–700.
43. Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 13: 52–56.
44. Faguy DM, Jarrell KF (1999) A twisted tale: The origin and evolution of motility and chemotaxis in prokaryotes. Microbiology 145 (Pt 2): 279–281.
45. He Y, Shelver D, Kerby RL, Roberts GP (1996) Characterization of a CO-responsive transcriptional activator from *Rhodospirillum rubrum*. J Biol Chem 271: 120–123.
46. Taylor BL, Zhulin IB (1999) PAS domains: Internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev 63: 479–506.
47. Kryukov GV, Gladyshev VN (2004) The prokaryotic selenoproteome. EMBO Rep 5: 538–543.
48. Farabaugh PJ (1996) Programmed translational frameshifting. Annu Rev Genet 30: 507–528.
49. Ward N, Eisen J, Fraser C, Stackebrandt E (2001) Sequenced strains must be saved from extinction. Nature 414: 148.
50. Miroshnichenko ML, Bonch-Osmolovskaya EA, Neuer A, Kostrikina NA, Chernych NA, et al. (1989) *Thermococcus stetteri* sp. nov., a new extremely thermophilic marine sulfur metabolizing archaebacterium. Syst Appl Microbiol 12: 257–262.
51. Eisen JA, Nelson KE, Paulsen IT, Heidelberg JF, Wu M, et al. (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci U S A 99: 9509–9514.
52. Wu M, Sun LV, Vamathevan J, Riegler M, Deboy R, et al. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: A streamlined genome overrun by mobile genetic elements. PLoS Biol 2: e69. DOI: 10.1371/joournal.pbio.0020069
53. Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res 26: 544–548.
54. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, et al. (2001) REPuter: The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633–4642.
55. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
56. Eisen JA, Wu M (2002) Phylogenetic analysis and gene functional predictions: Phylogenomics in action. Theor Popul Biol 61: 481–487.
57. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406–3415.
58. Onyenwoke RU, Brill JA, Farahi K, Wiegel J (2004) Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). Arch Microbiol 182: 182–192.

Note Added in Proof

It has come to our attention that a complementary comparison of sporulation genes in various Firmicutes was published in 2004 [58]. This study identified homologs of known sporulation genes in Firmicutes by experimental methods and genome analysis. The authors then used these results to study the evolution of sporulation and known sporulation genes.

Genome Sequence of *Theileria parva*, a Bovine Pathogen That Transforms Lymphocytes

Malcolm J. Gardner,[1]* Richard Bishop,[2] Trushar Shah,[2] Etienne P. de Villiers,[2] Jane M. Carlton,[1] Neil Hall,[1] Qinghu Ren,[1] Ian T. Paulsen,[1] Arnab Pain,[3] Matthew Berriman,[3] Robert J. M. Wilson,[4] Shigeharu Sato,[4] Stuart A. Ralph,[5] David J. Mann,[6] Zikai Xiong,[3] Shamira J. Shallom,[1] Janice Weidman,[1] Lingxia Jiang,[1] Jeffery Lynn,[1] Bruce Weaver,[1] Azadeh Shoaibi,[1] Alexander R. Domingo,[1] Delia Wasawo,[2] Jonathan Crabtree,[1] Jennifer R. Wortman,[1] Brian Haas,[1] Samuel V. Angiuoli,[1] Todd H. Creasy,[1] Charles Lu,[1]† Bernard Suh,[1]‡ Joana C. Silva,[1] Teresa R. Utterback,[1] Tamara V. Feldblyum,[1] Mihaela Pertea,[1] Jonathan Allen,[1] William C. Nierman,[1] Evans L. N. Taracha,[2] Steven L. Salzberg,[1] Owen R. White,[1] Henry A. Fitzhugh,[2]§ Subhash Morzaria,[2]‖ J. Craig Venter,[7] Claire M. Fraser,[1] Vishvanath Nene[1]

We report the genome sequence of *Theileria parva*, an apicomplexan pathogen causing economic losses to smallholder farmers in Africa. The parasite chromosomes exhibit limited conservation of gene synteny with *Plasmodium falciparum*, and its plastid-like genome represents the first example where all apicoplast genes are encoded on one DNA strand. We tentatively identify proteins that facilitate parasite segregation during host cell cytokinesis and contribute to persistent infection of transformed host cells. Several biosynthetic pathways are incomplete or absent, suggesting substantial metabolic dependence on the host cell. One protein family that may generate parasite antigenic diversity is not telomere-associated.

Theileria parva is a tick-borne parasite that causes a fatal disease in cattle known as East Coast fever (ECF). This disease, which kills over 1 million cattle each year in sub-Saharan Africa, results in economic losses exceeding $200 million annually (*1*). *Theileria* organisms belong to the phylum Apicomplexa, which is predicted to have originated about 930 million years ago (*2*). Unlike other apicomplexans, penetration of host cells by *T. parva* is not orientation-specific. Rhoptries and microspheres discharge after invasion, coincident with dissolution of the surrounding host cell membrane, leaving the parasite free in the host cell cytoplasm. Morbidity and mortality due to ECF are attributed to the ability of the schizont stage to malignantly transform its host cell, the bovine lymphocyte. Parasitosis increases exponentially because the schizont divides in synchrony with the host cell and infected cells infiltrate all tissues; cattle die of this lymphoproliferative disease 3 to 4 weeks after infection. Little pathology is due to the tick infective piroplasm, the red blood cell stage (*1*).

We sequenced the genome of *T. parva* in order to facilitate research on parasite biology, assist the identification of schizont antigens for vaccine development (*3*), and extend comparative apicomplexan genomics, in particular with *Plasmodium falciparum*, which causes malaria. Comparison with *T. annulata*, which causes tropical bovine theileriosis and mainly transforms macrophages, is described in an accompanying report (*4*). (This whole-genome shotgun project has been deposited at DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank under the project accession AAGK00000000.)

The haploid *T. parva* nuclear genome is 8.3×10^6 base pairs (Mbp) in length and consists of four chromosomes (Table 1). We provide a complete sequence, except for a 1- to 2-kbp gap in chromosome 4 and a gap in chromosome 3 (Tpr locus) that contains a 41-kbp and a 13-kbp set of overlapping sequences (contig) (*5*). The parasite apicoplast and mitochondrial (*6*) genomes have also been sequenced. Like *P. falciparum*, *T. parva* chromosomes contain one extremely A+T-rich region (>97%) about 3 kbp in length that may be the centromere. The regions between the CCCTA$_{3\text{-}4}$ telomeric repeats and the first protein-encoding gene are short, 2.9 kbp on average, and do not contain other repeats. Thus, the structure of the subtelomeric regions in *T. parva* is much less complex than that in *P. falciparum*, where arrays of repeats extend up to 30 kbp (*7*).

The *T. parva* nuclear genome contains about 4035 protein-encoding genes, 20% fewer than *P. falciparum*, but exhibits higher gene density, a greater proportion of genes with introns, and shorter intergenic regions. There are two identical, unlinked 5.8S-18S-28S rRNA units, suggesting that unlike *P. falciparum T. parva* does not possess functionally distinct ribosomes (*8*). Putative functions were assigned to 38% of the predicted proteins (Table 1).

The complexity of the *T. parva* life cycle is not matched by a large number of recognizable cell cycle regulators. Thus, the parasite is more akin to yeasts than higher eukaryotes, lacking discernable components of both the p53-MDM2-p14ARF-p21 and the Ink4-retinoblastoma-E2F pathways (*9*). There are four predicted cyclins and five cyclin-dependent kinases (cdks), most of which have close homologs in *P. falciparum*. However, *T. parva* lacks one cyclin and two cdks found in *P. falciparum*. These parasite cyclins are poorly conserved (∼25% identity), making cross-species comparisons difficult. The reduced recognizable *T. parva* cell cycle machinery suggests that a number of novel regulatory features remain to be discovered.

A unique aspect of *T. parva* biology is that infection of T and B lymphocytes results in a reversible transformed phenotype with uncontrolled proliferation of host cells that remain persistently infected. Parasite proteins that may modulate host cell phenotype are described in an accompanying report (*4*). Host cell microtubules that decorate the surface of schizonts are captured by the host cell spindle during mitosis, favoring infection of both daughter cells (*1*). *T. parva* encodes putative secreted forms of EMAP115- and Tau-like proteins, which are absent from *P. falciparum*; in higher eukaryotes, these proteins interact with microtubules (*10*). In addition, *T. parva* may modulate host cell mitosis by influencing disassembly of the host cell spindle via a secreted cdc48-like AAA–adenosine triphosphatase

[1]Institute for Genomic Research (TIGR), 9712 Medical Center Drive, Rockville, MD 20850, USA. [2]International Livestock Research Institute, Post Office Box 30709, Nairobi, Kenya. [3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]National Institute for Medical Research, Ridgeway, Mill Hill, London NW7 1AA, UK. [5]Institut Pasteur, 25 Rue du Docteur Roux, 75724 Paris Cedex 15, France. [6]Department of Biological Sciences, Imperial College, London SW7 2AZ, UK. [7]Venter Institute, 9708 Medical Center Drive, Rockville, MD, 20850, USA.

*To whom correspondence should be addressed. E-mail: gardner@tigr.org

†Present address: Lewis Thomas Lab, Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA.

‡Present address: Baskin School of Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA.

§Present address: 3709 Summercrest, Ft. Worth, TX 76109, USA.

‖Present address: Food and Agriculture Organization (FAO), 39 Phra Atit Road, Bangkok 10200, Thailand.

(ATPase associated with diverse cellular activities) (*11*). A likely *P. falciparum* homolog of this protein contains an N-terminal signal anchor sequence, whereas the *T. parva* protein contains a signal peptide and lacks a recognizable endoplasmic reticulum retention signal.

We used the Tribe-MCL algorithm (*5*) to identify 384 protein families containing 1063 proteins in the *T. parva* proteome (table S1). The largest family, containing 85 proteins, exists primarily in tandem arrays in the subtelomeric regions of all chromosomes. Many members of the family have a similar architecture, consisting of a secretion signal at the N terminus and a low-complexity glutamine- and proline-rich central domain that may be difficult for vertebrate immune systems to recognize (*12*). These genes are polymorphic between parasite isolates, and specific genes are absent from certain isolates (*13*). Each telomere has a conserved ~140-bp sequence immediately adjacent to the telomeric repeat (*14*), and several subtelomeric regions exhibit 70 to 100% sequence similarity (fig. S1). As in other eukaryotic pathogens, these features may facilitate interchromosomal recombination and the generation of antigenic diversity.

Proteins in the most rapidly evolving *T. parva* protein family, the Tpr (*T. parva* repeat) family, contain complex domain structures reminiscent of a system that has evolved to generate diversity (*15*). Unlike the majority of hypervariable gene families in parasitic protozoa (*16*), Tpr sequences are not telomere-associated. This family comprises a tandem array of highly conserved open reading frames (ORFs) on chromosome 3, located ~570 kbp from a telomere. The locus, estimated to span 100 kbp, contains at least 28 ORFs, of which 18, ranging in length from 192 to 674 amino acids, lack methionine codons in the first 50 amino acids (fig. S2). Eleven additional dispersed copies of Tpr, also of varying length, contain a 268–amino acid membrane-associated helical domain typical of the Tpr family. Massively parallel signature sequencing (*17*) and expressed sequence tags suggest that some genes in the locus are only transcribed in the piroplasm stage, whereas at least two of the dispersed genes are transcribed in the schizont stage. In common with the *var* genes of *P. falciparum* (*18*), domains within the Tpr genes are isolate-specific (*19*), and the 3′ end of Tpr has been used for genotyping of *T. parva* isolates. Tpr proteins have not yet been detected in piroplasms, and the function of these proteins remains unknown.

The genome sequence provides a global view of the metabolic potential of *T. parva* and allows a comparative analysis with *P. falciparum* metabolism. We predict a reduced functional role for the *T. parva* apicoplast and a greater dependence on the host for many substrates (fig. S3). *T. parva* lacks many enzymes in the shikimic acid, porphyrin, polyamine, and type II fatty acid biosynthetic pathways, but it retains the ability to produce isoprenoids via a methyl erythritol phosphate pathway in the apicoplast. *T. parva* cannot salvage purines, its ability to interconvert amino acids is very limited, and it lacks enzymes that permit the alternative nonoxidative production of pentoses and tetroses via the pentose phosphate pathway. Analysis of predicted transporters revealed fewer transporters of organic nutrients and inorganic cations than are present in *P. falciparum*. However, *T. parva* has more adenosine 5′-triphosphate-binding cassette (ABC) transporters of unknown substrate specificity. Another difference is that *T. parva* encodes an amino acid–cation symporter that is not present in *P. falciparum* (*7*) or *C. parvum* (*20*). In contrast to *P. falciparum*, *T. parva* encodes trehalose-6-phosphate synthase and trehalose phosphatase. Trehalose is a disaccharide that plays a role in desiccation and stress tolerance. It may protect the parasite during its long developmental cycle in the tick.

T. parva genes encode all of the enzymes necessary for glycolysis, glycerol catabolism, and the tricarboxylic acid (TCA) cycle. Unlike *P. falciparum*, *T. parva* does not encode malate dehydrogenase, but this could be functionally replaced by malate-quinone oxidoreductase, an activity also predicted to be present in *P. falciparum*. The origin of mitochondrial acetyl-coenzyme A (CoA) in both parasites presents a problem, because *P. falciparum* encodes a single pyruvate dehydrogenase that is targeted to the apicoplast (*21*) and *T. parva* does not encode all the subunits of this enzyme. Both parasites are predicted to contain cytoplasmic acetyl-CoA synthetase and a plasma membrane acetyl-CoA–CoA antiporter, but how mitochondrial oxidation of carbon chains is fueled in these two pathogens remains enigmatic because glycolysis and the tricarboxylic acid cycle do not appear to be linked by a classical route (*22*). Thus, it is not clear whether the complete TCA cycle is functional. Nitrogen metabolism differs from *P. falciparum* because *T. parva* lacks glutamate-ammonia ligase and only contains a nicotinamide adenine dinucleotide (NAD^+)–dependent glutamate dehydrogenase, which is usually associated with glutamate catabolism. This suggests that imported glutamate could play a role in supplementing intermediates in the TCA cycle.

The ionophores valinomycin and gramicidin D kill *T. parva*, suggesting that a mitochondrial electrochemical gradient is essential for parasite survival (*23*), but it is not known whether this is coupled to ATP synthesis. All subunits of the F1 catalytic domain of ATP synthase and subunit c of the F0 domain are present, but genes coding for subunits a and b of F0 were not found. The *T. parva* respiratory complexes are similar to those described in *P. falciparum.* Buparvaquone, a hydroxynapthaquinone drug used in the chemotherapy of ECF, probably inhibits electron transport through complex III (*23*).

The apicoplast is found in most apicomplexans and plays an essential role in parasite metabolism (*24*). An A+T-rich, ~35-kbp apicoplast genome encoding 30 proteins, rRNAs, and tRNAs is present in *Plasmodium*, *Toxoplasma*, and *Eimeria*, but not in *Cryptosporidium* (*20*); the latter lacks an apicoplast. The 39.5-kbp *T. parva* apicoplast genome differs from that of *P. falciparum* in that all of its genes are transcribed in the same direction. In addition, it has one rather than two copies of the rRNA genes, *clp*C is duplicated, the *rpoC2* gene encoding the β'' subunit of RNA polymerase is split into two parts, and it lacks the *sufB* gene (Fig. 1). Twenty-six of the 44 *T. parva* apicoplast genome protein-coding genes share sequence

Table 1. Comparison of *T. parva* nuclear genome coding characteristics with other sequenced apicomplexans. Gene length excludes introns; gene density calculated as genome size/number of protein-encoding genes. Source of data for *P. falciparum* was (*7*), and, for *C. parvum*, (*20*).

Features	Apicomplexan organism		
	T. parva	*P. falciparum*	*C. parvum*
Size (bp)	8,308,027	22,853,764	9,100,000
Number of chromosomes	4	14	8
Total G+C content (%)	34.1	19.4	30
Number of protein encoding genes	4035	5268	3807
Number of hypothetical proteins	2498	3208	925
Mean gene length (bp)	1407	2283	1795
Gene density (gene frequency in bp)	2057	4338	2382
Percent coding	68.4	52.6	75.3
Genes with introns (%)	73.6	53.9	5
Exons per gene (median)	4	2	1
Mean intergenic length (bp)	405	1694	566
G+C content intergenic regions (%)	26.1	13.6	23.9
Number of tRNA genes	47	43	45
Number of 5S rRNA genes	3	3	6
Number of rRNA units	2	7	5

similarity (27 to 61%) with proteins encoded by the *P. falciparum* apicoplast genome.

Most apicoplast proteins are encoded by nuclear genes and imported into the organelle by means of a bipartite targeting presequence (*24*). Comparison of the 345 *T. parva* (*5*) and 551 *P. falciparum* (*7*) predicted apicoplast-targeted (AT) proteins revealed similarities and differences in apicoplast function. The apicoplasts of *Plasmodium* and *Toxoplasma* participate in heme biosynthesis and are the sites of type II fatty acid and isoprenoid biosynthesis. Apicoplast-derived fatty acids in these parasites might contribute to the establishment and modification of the parasitophorous vacuole membrane (*25*). It may be notable that both *T. parva* and *T. annulata*, which have only retained isoprenoid biosynthesis, do not exist within a parasitophorous vacuole. About 100 AT proteins were found in both species, but 40% of these were hypothetical proteins, indicating that many core apicoplast functions have yet to be defined.

Fe-S clusters are required in mitochondria and plastids for the maturation of apoproteins. Fe-S cluster formation in the *T. parva* mitochondrion appears to be similar to that in yeast and *Plasmodium* (*26*) (table S3). However, of the *sufABCDES* genes involved in the assembly of Fe-S clusters in *Arabidopsis thaliana* (*27*) and *P. falciparum* plastids (*26*), only *sufS* was identified in *T. parva*. SufS is a cysteine desulfurase that requires SufE for catalytic activity. The parasite *T. para* genome encodes a plastid-targeted tRNA thiolation enzyme (MnmA) that has an additional domain similar both in sequence and predicted structure to the sulfur-binding domain of SufE. Thus, a previously unknown complex of SufS/MnmA may catalyze thiolation of tRNA in the *T. parva* apicoplast. The *T. parva* nuclear genome also encodes an AT protein with homology to NFU1, a scaffold protein for Fe-S cluster assembly in *A. thaliana* plastids (*28*), suggesting that assembly of Fe-S clusters occurs in the *T. parva* apicoplast despite the absence of most Suf proteins.

T. parva and *T. annulata* exhibit near-complete synteny across all chromosomes (*4*). To examine the extent of conservation of gene synteny between the evolutionarily distant *P. falciparum* and *T. parva*, we applied an iterative syntenic block algorithm and Jaccard-filtered COGs to whole-genome data from *P. falciparum* clone 3D7 (*7*), *P. y. yoelii* (*29*), *C. parvum* (*20*), and *T. parva*. Extensive synteny was found between *P. falciparum* and *P. y. yoelii* but not between *P. falciparum* and *C. parvum* or between *T. parva* and *C. parvum*. A total of 435 microsyntenic regions containing 1279 orthologs were observed between *P. falciparum* and *T. parva*, consisting of groups of 2 to 11 orthologs conserved in position between the two genomes (Fig. 2). This may be an underestimate of the degree of microsynteny as it is possible that, due to its long-term in vitro culture, clone 3D7 may represent an atypical genome. Syntenic clusters were distributed uniformly along each chromosome except for the subtelomeric regions, which contain species-specific gene families.

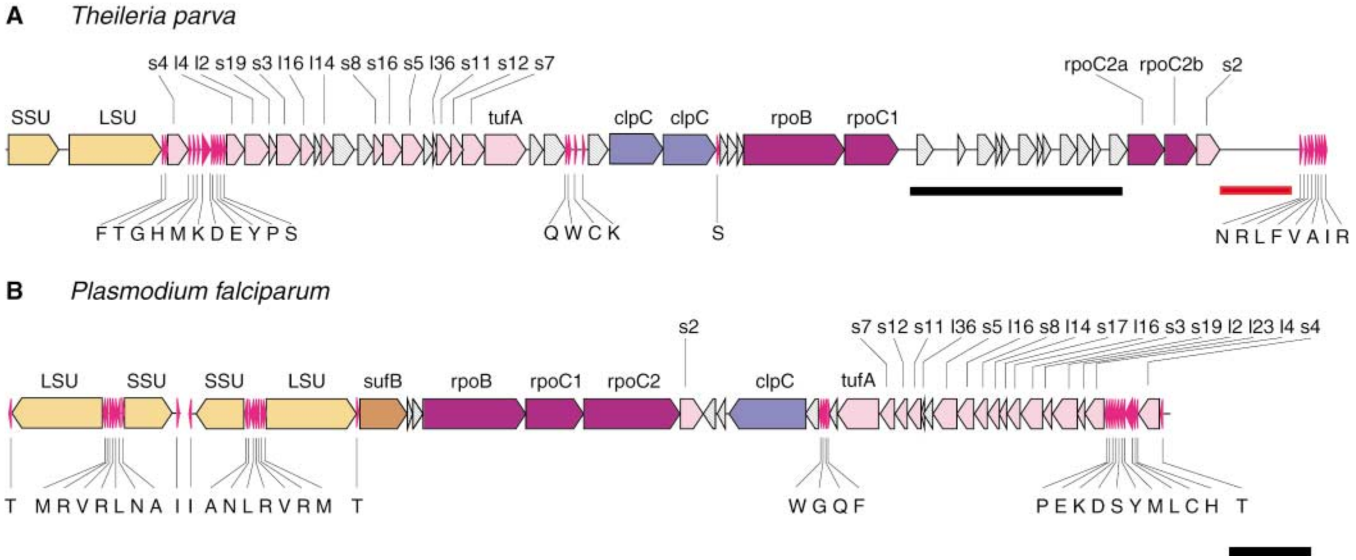


Fig. 1. Comparison of the apicoplast genomes of *T. parva* (**A**) and *P. falciparum* (**B**). A circular contig of the *T. parva* apicoplast genome was obtained after assembly of shotgun sequences, but the in vivo conformation has not been determined. The *P. falciparum* apicoplast genome is circular in vivo (*30*). The genomes are displayed in linear format beginning with the small subunit rRNA genes. Abbreviations and color coding: light orange, small (SSU) and large (LSU) subunit rRNAs; magenta, tRNAs [single-letter amino acid code (*31*)]; pink, ribosomal proteins (s and l for small and large subunit ribosomal proteins, respectively) and elongation factor Tu (tufA); blue, protein import; stippled gray, hypothetical proteins; purple, transcription; brown, SufB subunit of the SufABCDE Fe-S cluster assembly complex. The black and red bars indicate a region containing repeats and short ORFs and another region containing repeats and potential selenocysteine tRNAs, respectively (*5*). Scale bar equals 1 kbp.

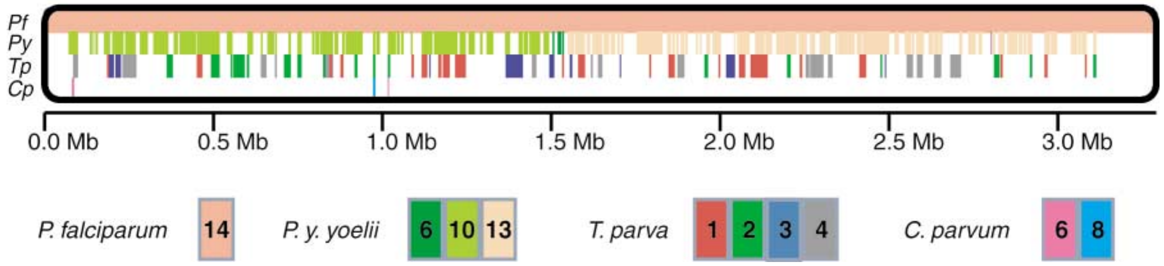


Fig. 2. Regions of microsynteny between *T. parva* and *P. falciparum*. Schematic of a representative *P. falciparum* chromosome showing synteny with three other apicomplexan species. Top row, *P. falciparum* chromosome 14 proteins. Second row, *P. y. yoelii* orthologs from *P. y. yoelii* chromosomes 6, 10, and 13. Third row, *T. parva* orthologs from *T. parva* chromosomes 1, 2, 3, and 4. Fourth row, *C. parvum* orthologs from *C. parvum* chromosomes 6 and 8.

The genome sequence of *T. parva* shows remarkable differences from the other apicomplexan genomes sequenced to date. It provides significant improvements in our understanding of the metabolic capabilities of *T. parva* and a foundation for studying parasite-induced host cell transformation and constitutes a critical knowledge base for a pathogen of significance to agriculture in Africa. Mining of sequence data has already proved useful in the search for candidate vaccine antigens (*3*).

References and Notes

1. R. A. I. Norval, B. D. Perry, A. S. Young, *The Epidemiology of Theileriosis in Africa* (Academic Press, London, 1992), p. 481.
2. A. A. Escalante, F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5793 (1995).
3. S. P. Graham *et al.*, in preparation.
4. A. Pain *et al.*, *Science* **309**, 131 (2005).
5. Materials and methods are available as supporting material on *Science* Online.
6. A. Kairo, A. H. Fairlamb, E. Gobright, V. Nene, *EMBO J.* **13**, 898 (1994).
7. M. J. Gardner *et al.*, *Nature* **419**, 498 (2002).
8. J. H. Gunderson *et al.*, *Science* **238**, 933 (1987).
9. B. Vogelstein, K. W. Kinzler, *Nat. Med.* **10**, 789 (2004).
10. M. Goedert, *Semin. Cell Dev. Biol.* **15**, 45 (2004).
11. I. M. Cheeseman, A. Desai, *Curr. Biol.* **14**, R70 (2004).
12. R. F. Anders, *Parasite Immunol.* **8**, 529 (1986).
13. R. Bishop *et al.*, *Mol. Biochem. Parasitol.* **110**, 359 (2000).
14. B. Sohanpal, D. Wasawo, R. Bishop, *Gene* **255**, 401 (2000).
15. H. A. Baylis, S. K. Sohal, M. Carrington, R. P. Bishop, B. A. Allsopp, *Mol. Biochem. Parasitol.* **49**, 133 (1991).
16. J. D. Barry, M. L. Ginger, P. Burton, R. McCulloch, *Int. J. Parasitol.* **33**, 29 (2003).
17. R. Bishop *et al.*, in preparation.
18. Z. Su *et al.*, *Cell* **82**, 89 (1995).
19. R. Bishop, A. Musoke, S. Morzaria, B. Sohanpal, E. Gobright, *Mol. Cell. Biol.* **17**, 1666 (1997).
20. M. S. Abrahamsen *et al.*, *Science* **304**, 441 (2004); published online 25 March 2004 (10.1126/science.1094786).
21. B. J. Foth *et al.*, *Mol. Microbiol.* **55**, 39 (2005).
22. S. A. Ralph, *Mol. Microbiol.* **55**, 1 (2005).
23. A. A. McColm, N. McHardy, *Ann. Trop. Med. Parasitol.* **78**, 345 (1984).
24. R. F. Waller, G. I. McFadden, *Curr. Issues Mol. Biol.* **7**, 57 (2005).
25. R. F. Waller *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12352 (1998).
26. F. Seeber, *Int. J. Parasitol.* **32**, 1207 (2002).
27. X. M. Xu, S. G. Moller, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9143 (2004).
28. S. Leon, B. Touraine, C. Ribot, J. F. Briat, S. Lobreaux, *Biochem. J.* **371**, 823 (2003).
29. J. M. Carlton *et al.*, *Nature* **419**, 512 (2002).
30. R. J. Wilson *et al.*, *J. Mol. Biol.* **261**, 155 (1996).
31. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

32. We thank T. Irvin, O. Ole-MoiYoi, T. Musoke, C. Sugimoto, H. Leitch, R. von Kaufmann, S. MacMillan, R. Koenig, M. Brown, R. Ndegwa, L. Thairo, B. Anyona, T. Akinyemi, the TIGR conferences staff, the Secretariat of the Consultative Group for International Agricultural Research, and the research staff of the International Livestock Research Institute (ILRI). Supported by the TIGR Board of Trustees, ILRI, J. C. Venter, the Rockefeller Foundation, the U.S. Agency for International Development, and the UK Department for International Development.

Supporting Online Material
www.sciencemag.org/cgi/content/full/309/5731/134/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S3

31 January 2005; accepted 5 May 2005
10.1126/science.1110439

Long-Term Monitoring of Bacteria Undergoing Programmed Population Control in a Microchemostat

Frederick K. Balagaddé,[1*†] Lingchong You,[2†‡] Carl L. Hansen,[1§] Frances H. Arnold,[2] Stephen R. Quake[1*‖]

Using an active approach to preventing biofilm formation, we implemented a microfluidic bioreactor that enables long-term culture and monitoring of extremely small populations of bacteria with single-cell resolution. We used this device to observe the dynamics of *Escherichia coli* carrying a synthetic "population control" circuit that regulates cell density through a feedback mechanism based on quorum sensing. The microfluidic bioreactor enabled long-term monitoring of unnatural behavior programmed by the synthetic circuit, which included sustained oscillations in cell density and associated morphological changes, over hundreds of hours.

By continually substituting a fraction of a bacterial culture with sterile nutrients, the chemostat (*1*, *2*) presents a near-constant environment that is ideal for controlled studies of microbes and microbial communities (*3*–*6*). The considerable challenges of maintaining and operating continuous bioreactors, including the requirement for large quantities of growth media and reagents, have pushed the move toward miniaturization and chip-based control (*7*–*10*), although efforts have been limited to batch-format operation. Microbial biofilms, which exist in virtually all nutrient-sufficient ecosystems (*11*), interfere with continuous bioreactor operation (*12*). Phenotypically distinct from their planktonic counterparts (*11*), biofilm cells shed their progeny into the bulk culture and create mixed cultures. At high dilution rates, the biofilm, which is not subject to wash-out, supplies most of the bulk-culture cells (*13*). The increase in surface area-to-volume ratio as the working volume is decreased aggravates these wall-growth effects (*13*).

We created a chip-based bioreactor that uses microfluidic plumbing networks to actively prevent biofilm formation. This device allows semicontinuous, planktonic growth in six independent 16-nanoliter reactors with no observable wall growth (Fig. 1A). The cultures can be monitored in situ by optical microscopy to provide automated, real-time, noninvasive measurement of cell density and morphology with single-cell resolution.

Each reactor, or "microchemostat," consists of a growth chamber, which is a fluidic loop 10 μm high, 140 μm wide, and 11.5 mm in circumference, with an integrated peristaltic pump and a series of micromechanical valves to add medium, remove waste, and recover cells (Fig. 1B). The growth loop is itself composed of 16 individually addressable segments. The microchemostat is operated in one of two alternating states: (i) continuous circulation, and (ii) cleaning and dilution. During continuous circulation, the peristaltic pump moves the microculture around the growth loop at a linear velocity of $\sim$250 μm s^{-1} (Fig. 1C). During cleaning and dilution, the mixing is halted and a segment is isolated from the rest of the reactor with micromechanical valves. A lysis buffer is flushed through the isolated segment for 50 s to expel the cells it contains, including any wall-adhering cells (Fig. 1D). Next, the segment is flushed with sterile growth medium to completely rinse out the lysis buffer. This segment, filled with sterile medium, is then reunited with the rest of the growth chamber, at which point continuous circulation resumes. This process is repeated sequentially on different growth chamber segments, thus eliminating biofilm formation and enabling pseudocontinuous operation. In comparison, passive treatment of the microfluidic surfaces with nonadhesive surface coatings [such as poly (ethylene glycol), ethylenediaminetetraacetic acid, polyoxyethylene sorbitan monolaurate, and bovine serum albumin] proved ineffective in preventing biofilm forma-

[1]Department of Applied Physics, [2]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA.

*Present address: Department of Bioengineering, Stanford University, Stanford, CA 94305, USA.

†These authors contributed equally to this work.

‡Present address: Department of Biomedical Engineering and Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA.

§Present address: Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

‖To whom correspondence should be addressed. E-mail: quake@stanford.edu