Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA

ALLAN CAMPBELL^{†‡}, JAN MRÁZEK[§], AND SAMUEL KARLIN[§]

Departments of [†]Biological Sciences and [§]Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Allan Campbell, May 26, 1999

ABSTRACT Our basic observation is that each genome has a characteristic “signature” defined as the ratios between the observed dinucleotide frequencies and the frequencies expected if neighbors were chosen at random (dinucleotide relative abundances). The remarkable fact is that the signature is relatively constant throughout the genome; i.e., the patterns and levels of dinucleotide relative abundances of every 50-kb segment of the genome are about the same. Comparison of the signatures of different genomes provides a measure of similarity which has the advantage that it looks at all the DNA of an organism and does not depend on the ability to align homologous sequences of specific genes. Genome signature comparisons show that plasmids, both specialized and broad-range, and their hosts have substantially compatible (similar) genome signatures. Mammalian mitochondrial (Mt) genomes are very similar, and animal and fungal Mt are generally moderately similar, but they diverge significantly from plant and protist Mt sets. Moreover, Mt genome signature differences between species parallel the corresponding nuclear genome signature differences, despite large differences between Mt and host nuclear signatures. In signature terms, we find that the archaea are not a coherent clade. For example, *Sulfolobus* and *Halobacterium* are extremely divergent. There is no consistent pattern of signature differences among thermophiles. More generally, grouping prokaryotes by environmental criteria (e.g., habitat propensities, osmolarity tolerance, chemical conditions) reveals no correlations in genome signature.

Extensive data support the proposal that each living organism possesses a genomic signature consisting of dinucleotide relative abundance values calculated from genomic sequences (1–3). Explicitly, the genomic signature profile consists of the array $\{\rho_{XY}^* = f_{XY}^*/f_X^*f_Y^*\}$, where f_X^* denotes the frequency of the mononucleotide X and f_{XY}^* the frequency of the dinucleotide XY , both computed from the sequence concatenated with its inverted complement. These dinucleotide relative abundance values $\{\rho_{XY}^*\}$ minus 1 (also termed dinucleotide biases) effectively assess differences between the observed dinucleotide frequencies and those expected from random associations of the component mononucleotide frequencies. From data simulations and statistical theory, the estimates $\rho_{XY}^* \leq 0.78$ or $\rho_{XY}^* \geq 1.23$ convey significant underrepresentation or overrepresentation, respectively, for 50-kb random DNA contigs (1–3). We present substantial data showing that the genome signatures of bacterial plasmids are pervasively similar to those of their natural hosts. By contrast, the signatures of animal mitochondrial DNA are not close to those of their hosts but are generally concordant with those of other animal mitochondrial (Mt) DNA.

Justifications for Using Genome Signature. Biochemical experiments in the 1960s and 1970s measuring nearest-

neighbor frequencies (4, 5) established that the set of dinucleotide relative abundance values $\{\rho_{XY}^*\}$ is a remarkably stable property of the DNA of an organism. From this perspective, the set of dinucleotide relative abundance values constitutes a *genomic signature* that is diagnostic and can discriminate sequences from different organisms (3, 6). What causes the uniformity of signature throughout the genome? It pervades both noncoding and coding DNA (7), and hence cannot be explained by preferential codon usage. A reasonable explanation postulates differences in the replication and repair machinery of different species, which either preferentially generate or preferentially select specific dinucleotides in the DNA. These effects might operate through local DNA structures (base step conformational tendencies), context-dependent mutation rates, methylation, and/or other DNA modifications (1–3, 6).

A measure of genomic signature difference between two sequences f and g (from different organisms or from different regions of the same genome) is the average absolute dinucleotide relative abundance difference calculated as

$$\delta^*(f, g) = 1/16 \sum |\rho_{XY}^*(f) - \rho_{XY}^*(g)| \quad (\delta^* \text{ difference}),$$

where the sum extends over all dinucleotides. Levels of δ^* differences of 50-kb contigs for some reference examples are described in the Introduction of the accompanying paper (8).

Genome Signature Comparisons. Fig. 1 displays the genome signature profiles for all currently available extensive genome sequence sets. These include 22 eubacterial genome ensembles (mostly complete genomes), 5 archaeal chromosomes, the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* complete genomes, and extensive nonredundant sequence collections from 5 other eukaryotic species. For each genome, the dinucleotide biases $\{\rho_{XY}^*\}$ are substantially invariant (9).

The dinucleotide TA is broadly underrepresented or low normal in prokaryotic sequences about the level $0.50 \leq \rho_{TA}^* \leq 0.82$ (exceptions include *Rickettsia prowazekii*, *Clostridium acetobutylicum*, and the archaea *P. aerophilum*, *P. horikoshii*, and *Sulfolobus* sp. (full names appear in the legend to Figs. 1 and 2). TA underrepresentation is also pervasive in eukaryotic chromosomes (exception *P. falciparum*) but not in eukaryotic small viral genomes or in animal mitochondrial genomes (10, 11).

Among prokaryotes, CG is suppressed (underrepresented) in *M. genitalium* (but not in *M. pneumoniae*), in *R. prowazekii*, in *B. burgdorferi*, in *C. jejuni*, in the low-G+C Gram-positive sequences of *Streptococcus* and *Clostridium*, and in several thermophiles, including *M. jannaschii*, *M. thermoautotrophicum*, *Sulfolobus* spp., but not in *P. aerophilum* or *P. horikoshii*. At the other extreme, CG is overrepresented in *B. stearothermophilus*, in halobacteria, and also in several β - and α -proteobacterial genomes (e.g., *Neisseria* spp. and *Rhizobium* spp.). Among eukaryotes, CG shows potent suppression in vertebrates (even in deuterostomes). Overall, ρ_{CG}^* values in verte-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Abbreviation: Mt, mitochondria(1).

[†]To whom reprint requests should be addressed. E-mail: fa.amc@forsythe.stanford.edu.

genome (available DNA)	CG	GC	TA	AT	CC GG	TT AA	TG CA	AG CT	AC GT	GA TC	G+C
<i>Escherichia coli</i> (4.6Mb)*	1.16	1.28	0.75	1.10	0.91	1.21	1.12	0.82	0.88	0.92	51%
<i>Haemophilus influenzae</i> (1.8Mb)*	1.09	1.43	0.75	0.95	1.01	1.25	1.12	0.82	0.85	0.87	38%
<i>Neisseria gonorrhoeae</i> (877kb)	1.32	1.26	0.63	1.05	0.99	1.50	0.99	0.67	0.83	0.89	53%
<i>Neisseria meningitidis</i> (2.2Mb)	1.31	1.27	0.64	1.05	0.96	1.44	1.01	0.70	0.84	0.91	52%
<i>Rhodobacter capsulatus</i> (1.4Mb)	1.19	1.19	0.33	1.61	0.88	1.30	1.03	0.84	0.71	1.16	67%
<i>Rickettsia prowazekii</i> (1.1Mb)*	0.77	1.53	0.98	0.98	1.03	1.05	1.02	1.06	0.86	0.91	29%
<i>Helicobacter pylori</i> (1.7Mb)*	0.93	1.56	0.73	0.86	1.17	1.37	0.97	0.97	0.67	0.87	39%
<i>Campylobacter jejuni</i> (1.6Mb)*	0.62	1.75	0.77	0.83	1.11	1.25	1.03	1.09	0.71	0.92	31%
<i>Bacillus subtilis</i> (4.2Mb)*	1.04	1.27	0.65	1.02	0.97	1.24	1.08	0.91	0.75	1.06	44%
<i>Streptococcus pyogenes</i> (985kb)	0.71	1.19	0.76	0.89	1.04	1.17	1.12	1.04	0.86	0.99	39%
<i>Clostridium acetobutylicum</i> (4.0Mb)	0.45	1.28	0.93	0.95	1.22	1.08	1.02	1.12	0.81	0.97	31%
<i>Streptomyces coelicolor</i> (2.4Mb)	1.14	0.97	0.51	0.93	0.88	0.82	1.00	0.95	1.14	1.25	72%
<i>Mycobacterium leprae</i> (1.7Mb)	1.13	1.07	0.75	1.10	0.88	1.04	1.14	0.86	1.05	1.02	58%
<i>Mycobacterium tuberculosis</i> (4.4Mb)*	1.18	1.07	0.58	1.24	0.86	1.05	1.11	0.80	1.05	1.08	65%
<i>Mycoplasma genitalium</i> (580kb)*	0.39	1.19	0.75	0.77	1.13	1.23	1.16	1.06	0.96	0.89	32%
<i>Mycoplasma pneumoniae</i> (816kb)*	0.82	1.14	0.77	0.71	1.12	1.30	1.08	0.96	1.02	0.81	40%
<i>Synechocystis</i> sp. (3.6Mb)*	0.75	1.02	0.75	1.00	1.36	1.32	1.05	0.85	0.79	0.86	48%
<i>Deinococcus radiodurans</i> (3.0Mb)	1.07	1.16	0.49	0.89	0.87	1.24	1.12	1.00	0.93	1.01	67%
<i>Treponema pallidum</i> (1.1Mb)*	1.08	1.22	0.74	0.93	0.86	1.18	1.13	0.94	0.96	0.95	53%
<i>Borrelia burgdorferi</i> (911kb)*	0.48	1.47	0.77	0.88	1.29	1.22	1.02	1.07	0.69	1.01	29%
<i>Chlamydia trachomatis</i> (1.0Mb)*	0.79	1.12	0.77	0.89	1.01	1.16	0.96	1.18	0.76	1.15	41%
<i>Aquifex aeolicus</i> (1.6Mb)*	0.87	0.75	0.82	0.66	1.24	1.29	0.74	1.18	0.89	1.12	43%
<i>Methanococcus jannaschii</i> (1.7Mb)*	0.32	1.12	0.83	0.94	1.38	1.14	1.03	1.11	0.72	1.05	31%
<i>Methanobacterium thermoautotrophicum</i> (1.8Mb)	0.51	0.76	0.74	1.13	1.25	0.95	1.17	1.07	0.85	1.14	50%
<i>Archaeoglobus fulgidus</i> (2.2Mb)*	0.78	1.02	0.61	0.86	1.04	1.21	1.01	1.17	0.77	1.19	49%
<i>Pyrococcus horikoshii</i> (1.7Mb)*	0.61	0.89	0.90	0.92	1.30	1.11	0.85	1.22	0.73	1.13	42%
<i>Pyrobaculum aerophilum</i> (2.2Mb)*	0.97	1.15	1.07	0.93	1.10	1.18	0.86	1.06	0.83	0.90	51%
human (5.8Mb)	0.25	1.00	0.74	0.88	1.25	1.12	1.20	1.17	0.83	0.99	43%
mouse (1.1Mb)	0.22	0.95	0.72	0.80	1.19	1.08	1.24	1.26	0.88	1.01	46%
<i>Drosophila melanogaster</i> (4.3Mb)	0.94	1.29	0.75	0.97	1.08	1.23	1.12	0.87	0.84	0.90	41%
<i>Caenorhabditis elegans</i> (74Mb)	0.97	1.04	0.62	0.86	1.05	1.28	1.09	0.90	0.86	1.09	36%
yeast (12Mb)*	0.80	1.02	0.77	0.94	1.06	1.14	1.10	0.99	0.89	1.06	38%
<i>Arabidopsis thaliana</i> (2.0Mb)	0.72	0.93	0.74	0.90	1.03	1.13	1.11	1.04	0.91	1.11	36%
<i>Plasmodium falciparum</i> (947kb)	0.74	0.93	0.99	1.07	1.54	1.00	1.10	0.83	0.92	0.97	20%

* indicates complete genome

<0.50 0.50–0.70 0.70–0.78 0.78–1.23 1.23–1.30 1.30–1.50 >1.50

FIG. 1. Genome signature (dinucleotide relative abundances) of complete genomes and large DNA sequence samples (>500 kb).

brates range from 0.23 to 0.40, whereas they are in the normal range for insects, worms, and most fungi. It has been shown (12) that unmethylated CG dinucleotides of normal frequency in most enteroproteobacteria can induce an immune response in mammalian genomes, where CG is very low. Is this a concomitant of genomic signature biases? The reverse dinucleotide, GC, is predominantly overrepresented in many β - and γ -proteobacterial sequences and in several low-G+C Gram-positive bacterial genomes (e.g., *B. subtilis* and *C. acetobutylicum*; Fig. 1). In eukaryotes, moderate overrepresentation of GC occurs in *Drosophila* species but apparently not in other metazoa.

TT/AA is overrepresented in several proteobacteria, in Mycoplasmas, in *Synechocystis* sp., in *Deinococcus radiodurans*, in *A. aeolicus* among prokaryotes, and in insects and worms among eukaryotes. There are no underrepresentations of TT/AA. Overrepresentations of CC/GG include *Synechocystis*, *B. burgdorferi*, *A. aeolicus*, *M. jannaschii*, *M. thermoautotrophicum*, and *P. horikoshii*. There are no underrepresentations of CC/GG.

δ^* Differences Among Prokaryotic Sequences. Fig. 2 reports average δ^* differences based on multiple disjoint 50-kb contigs among prokaryotic genomic sets, each at least 100 kb in aggregate length and most genome sequence collections exceeding 800 kb. The δ^* differences for all pairs of contigs for

every pair of genomes are highly stable with limited variation (cf. refs. 1–6).

(i) Within-species δ^* differences (diagonal elements of Fig. 2) range from 12 to 52 (all δ^* differences are multiplied by 1000), and average between species δ^* differences range from 34 to 340.

(ii) The rickettsial sequences are grouped with α -proteobacteria, apparently on the basis of rRNA gene comparisons. Is this consistent? The classical α -proteobacterial types are divided into two major subgroups: $\mathcal{A}_1$, including *Rhizobium* spp., that function importantly in nitrogen fixation, and $\mathcal{A}_2$, including *Rhodobacter* spp. and *P. denitrificans*, found predominantly in soil and marine habitats and doing anoxygenic photosynthesis. A tentative third group, $\mathcal{A}_3$, includes the *Rickettsia* and *Ehrlichia* clades (obligate intracellular parasites). Genome signature comparisons indicate drastic discrepancies between the combined groups $\{\mathcal{A}_1, \mathcal{A}_2\}$ and the group $\mathcal{A}_3$. Moreover, the $\mathcal{A}_1$ and $\mathcal{A}_2$ genomes are pervasively of high G+C content ($\geq 60\%$), whereas $\mathcal{A}_3$ genomes are of low G+C content ($< 32\%$). The δ^* differences between *Rickettsia* and classical α -proteobacterial sequences are generally > 200 , very distant (see Fig. 2).

(iii) *Sulfolobus* spp. δ^* differences from other prokaryotes. Fig. 2 includes seven archaeal sequence collections consisting of one halobacterial genus (HALSP) (see legends to Figs. 1

FIG. 2. Average δ^* differences within (diagonal entries) and between (nondiagonal entries) prokaryotic DNA sequence samples based on pairwise comparisons of all disjoint nonredundant 50-kb samples available. See also Fig. 5 for 70 prokaryotic species [published as supplemental data on the PNAS web site (www.pnas.org)].

and 2 for full names, and we use the Swiss-Prot abbreviations), two methanogens (METJA, METTH), one "derived" methanogen (ARCFU), and three archaeal thermophiles (PYRHO, PYRAE, SULSP). Thermophiles (exception, PYRHO) tend to be relatively closer to vertebrate eukaryotes than to eubacterial sequences, with δ^* differences in the range 100–150 (3, 9), whereas HALSP have generally $\delta^* > 200$. Thus, the highly diverse δ^* differences indicate that a consistent description for the archaeal sequences is problematical. We summarize the genomic signature contrasts for *Sulfolobus* in the following array.

δ^* (*Sulfolobus*, *Clostridium*) ≈ 85 , moderately similar; δ^* (*Sulfolobus*, *Rickettsia*, and *Buchnera*) ≈ 125 –130, distantly similar; δ^* (*Sulfolobus*, other thermophilic archaea) ≈ 71 –114, weakly to distantly similar; δ^* (*Sulfolobus*, purple proteobacteria and high G+C Gram-positive) ≈ 190 –270, very distant; δ^* (*Sulfolobus*, cyanobacteria) ≈ 145 –177, distant.

(iv) Halobacterial genome sequences are outliers. Intriguingly, with respect to genome signature comparisons, the closest to *Halobacterium* spp. are the *Streptomyces* sequences, weakly similar. The δ^* differences of *Halobacterium* spp. from other prokaryotes mostly exceed the extreme level of 250.

Genome Compatibility Between Plasmids and Host. Plasmids are genetic mobile elements among bacterial cells, generally laterally transferred by conjugation (on occasion by transduction or transformation). Plasmids carry restriction systems, antibiotic resistance genes, heavy-metal cofactors, Nif (nitrogen fixation) genes, and other contingency functions. Replication of plasmids is largely governed by host machinery. How the genome signature of a plasmid sequence compares

with that of its host sequence is assessed for available prokaryotic genomes with completely sequenced plasmids. Fig. 3 reports average δ^* differences for each plasmid compared with representatives of 52 prokaryotic genomic collections. Importantly, the δ^* differences between each plasmid and its natural host rank among the closest. Thus, the examples of Fig. 3 are consistent with the proposition that plasmids are viable in bacterial hosts only when their genome signatures are sufficiently compatible (moderately similar, $55 \leq \delta^* \leq 85$, or at least weakly similar, $90 < \delta^* < 115$) with the host genome signature or they can rapidly be made similar.

B. burgdorferi. The complete genome extends 0.911 Mb and contains 17 plasmids, of which 11 have been sequenced from 10 kb to 54 kb in size labeled A to K. The average δ^* difference within the *B. burgdorferi* (BORBU) genome is 25 with range 3–66. The δ^* differences among all plasmids and compared with a complete set of 50-kb contigs of the host genome range from 42 to 84, clearly moderately similar (data not shown). Mutual plasmid sequence δ^* differences show moderate to weak similarity.

Specific plasmids. *M. jannaschii*, mja. This archaeal genome contains two plasmids of 58 kb (called large) and 16.5 kb (small). Again, we find that the two plasmids are mutually close ($\delta^* = 35$) and each moderately similar to the host genome. *Rhizobium* spp., rhi. We have available about 150 kb of nonredundant aggregate sequences from *R. leguminosarum* and about 250 kb of sequence from *R. meliloti*. There is a mammoth plasmid of about 550 kb. Partitioning these sequences into 50-kb contigs yields the average δ^* differences given in Fig. 3 of close similarity. *Halobacterium* spp., hal. An

individual plasmids															Broad host range plasmids				
esc	ype	ype	ype	agr	rhl	hpy	lac	chl	aqu	hal	mja	mja	rsf	inc	esc	salty	yeren	yerpe	serma
pl	_1	_2	_3	Ti	pl	pl	pl	pl	pl	pl	_1	_2	10	Pa					
74	66	75	84	94	106	118	144	176	253	215	220	215	83	66	esc	coo			
92	86	84	100	82	95	118	141	168	240	211	213	201	96	70	salty				
44	81	38	67	121	132	97	114	149	229	240	172	166	86	100	yeren				
55	78	55	76	106	119	108	126	159	241	237	191	186	76	91	yerpe				
98	79	96	105	81	86	131	162	193	261	212	238	233	83	48	serma				
50	55	32	44	110	118	79	95	127	211	209	167	163	64	82	vibch				
89	102	84	93	135	149	87	135	166	235	253	211	207	111	100	haein				
67	42	72	68	79	78	151	133	166	250	187	196	192	66	53	psepu				
91	45	82	71	66	61	140	112	135	219	166	177	176	84	52	pseae				
108	66	107	91	68	57	179	146	156	244	153	207	203	102	69	azovi				
95	74	104	107	89	86	182	179	212	296	196	240	235	87	75	borpe				
100	70	106	99	78	72	182	168	201	285	192	231	227	84	62	alceu				
161	162	152	160	149	161	176	201	216	284	262	277	259	163	130	neigo				
147	141	138	145	132	145	163	188	208	276	246	264	250	145	111	neime				
125	85	123	106	33	32	168	147	148	234	167	220	211	101	62	agrtu				
151	108	147	132	40	36	193	173	162	255	170	246	233	111	80	rhile				
171	124	165	148	61	52	213	189	160	256	147	261	244	130	101	rhime				
191	150	186	172	93	89	222	212	200	280	242	286	274	144	105	rhoca				
193	150	187	171	111	99	237	207	187	282	149	264	245	154	141	rhosp				
113	139	90	121	163	189	70	105	136	205	248	153	143	145	166	ricpr				
126	88	129	103	124	115	178	149	152	239	149	201	207	137	119	myxxa				
167	187	150	165	205	212	104	138	155	196	301	209	191	158	177	helpy				
171	195	153	173	217	224	68	105	137	169	325	167	165	165	188	camje				
95	72	81	67	85	92	105	108	122	198	212	180	176	73	61	bacsu				
82	100	78	70	146	153	126	73	118	159	183	106	99	146	140	bacth				
45	67	40	53	131	138	101	91	130	206	199	147	142	101	109	staa				
85	91	78	63	122	129	69	42	81	142	217	118	122	109	108	lacia				
74	93	64	72	143	151	53	43	92	154	238	118	121	102	118	stpy				
141	170	121	151	205	210	103	91	131	148	285	82	81	168	193	cloac				
150	123	142	135	100	110	172	175	183	251	218	249	230	133	85	bacst				
199	150	195	163	154	148	250	212	206	281	92	259	248	214	168	stroc				
218	167	215	163	160	150	271	234	231	311	90	289	279	236	181	sacer				
80	65	99	81	111	121	181	157	187	268	144	212	207	122	95	mycle				
122	90	137	110	97	96	218	186	208	289	150	242	239	133	98	myctu				
131	161	132	141	221	228	94	102	159	176	291	134	158	159	183	mycge				
124	150	120	125	202	210	109	109	159	183	257	186	184	169	161	mycpn				
81	68	68	56	92	95	93	82	104	180	205	154	150	70	67	synsp				
59	88	50	62	142	150	60	53	100	168	225	125	123	104	120	anasp				
134	174	138	138	186	201	126	120	173	173	297	124	114	172	167	synsq				
105	63	99	85	107	105	124	118	149	218	191	199	198	98	61	deira				
79	51	72	72	109	112	114	122	154	237	176	198	196	84	77	trepa				
178	191	160	168	212	219	85	94	138	148	319	102	104	173	191	borbu				
134	135	120	108	138	141	102	80	31	103	215	146	140	143	159	chltr				
224	228	217	197	214	214	175	133	104	59	254	155	151	229	224	thesp				
211	223	205	188	243	248	190	149	116	61	267	176	176	258	233	aquae				
220	171	219	187	166	161	286	248	221	287	28	290	277	243	192	halps				
173	188	159	160	214	221	143	99	132	150	306	49	65	195	214	metja				
156	182	166	157	209	211	181	128	171	165	235	111	119	193	213	metth				
152	152	141	120	136	138	105	89	48	110	217	149	146	149	152	arclu				
205	220	193	187	232	234	178	126	103	70	261	102	95	233	251	pyrho				
123	147	105	123	174	182	102	98	97	147	244	158	147	149	163	pyrae				
173	195	162	162	227	233	162	111	118	102	240	99	88	222	224	sulsp				

FIG. 3. Plasmids are abbreviated as follows: rsf10 (broad-host-range plasmid RSF1010), incPa (Broad host range plasmid Birmingham IncP α), escpl (enterohemorrhagic *E. coli* plasmid pO157), ype_1 (*Yersinia pestis* plasmid pMT-1), ype_2 (*Y. pestis* plasmid pCD1), ype_3 (*Y. pestis* plasmid pPCP1), agrTi (*A. tumefaciens* plasmid Ti), rhipl (*Rhizobium* plasmid pNGR234a), hpypl (*H. pylori* plasmid pHPM186); lacpl (*L. lactis* DPC3147 plasmid pMRC01), chlpl (virulence cluster from *C. trachomatis* plasmid pCHL1), aqupl (*A. aeolicus* plasmid ece1), halpl (*Halobacterium* sp. NRC-1 plasmid), mja_1 (*M. jannaschii* large plasmid), mja_2 (*M. jannaschii* small plasmid). See Fig. 2 for prokaryotic species abbreviations. Additional prokaryotes not included in Fig. 2 are salty (*Salmonella typhimurium*), yeren (*Yersinia enterocolitica*), yerpe (*Yersinia pestis*), serma (*Serratia marcescens*), vibch (*Vibrio cholerae*), and rhile (*Rhizobium leguminosarum*). Yellow background indicates δ^* differences of a plasmid from its confirmed host. See also Fig. 6, which is published as supplemental data on the PNAS web site (www.pnas.org).

approximately 200-kb plasmid (pNRC100) from *H. halobium* was sequenced. About 100 kb of aggregate sequences from *H. halobium* and *H. salinarum* genomes are also accessible. Fig. 3 reports δ^* differences entailing close to moderate similarity. *A. tumefaciens*, agr-Ti. Compared to nonplasmid sequences,

the classic crown gall tumor plasmid Ti (133 kb available) produces an average δ^* difference of 33 (closely similar). *E. coli*, esc. The plasmid O157 of *E. coli*, of length 93 kb, gives δ^* difference of 74 (moderate similarity) compared with *E. coli*. *L. lactis*, lac. This bacterium has available about 250 kb of aggregate genomic sequence and a complete plasmid (MRC01) of length about 60 kb. The δ^* difference of MRC01 to the average 50-kb genome contig is 42 (closely similar). *A. aeolicus*, aqu. The complete genome sequence (1.6 Mb) is available. A plasmid contig labeled ece1 has also been sequenced. The δ^* difference is 61 (moderately similar). *H. pylori*, hpy. The single plasmid (13 kb) of strain 26695 has δ^* difference 104 (weakly similar) to the genome. *C. trachomatis*, chl. The plasmid (7.5 kb) compared to the complete genome has a close signature difference ($\delta^* = 31$).

Broad-host-range plasmids. Two plasmids that stably replicate in many hosts are RSF1010 and RP4 (13, 14); see legend to Fig. 3. We list next those host prokaryotic species with at least 80 kb of available genome sequence which accept stably these plasmids. RSF1010: *Escherichia coli*, *Pseudomonas aeruginosa*, *Pseudomonas putida*, *Azotobacter vinelandii*, *Rhizobium meliloti*, *Agrobacterium tumefaciens*, and *Alcaligenes eutrophus*. RP4: *Serratia marcescens*, *Azotobacter vinelandii*, *Pseudomonas aeruginosa*, *Pseudomonas fluorescens*, *Escherichia coli*, *Pseudomonas putida*, *Rhodospirillum rubrum*, *Shigella boydii*, *Salmonella typhimurium*, *Vibrio cholerae*, and *Rhizobium meliloti*.

The broad-range plasmids are generally moderately or weakly similar to the other plasmids from proteobacterial genomes. The plasmid from *L. lactis* is distantly similar to the broad-range plasmids. The plasmids from *B. burgdorferi*, *Halobacterium* spp., and *M. jannaschii* are very distant from all proteobacterial plasmids. The δ^* differences among *B. burgdorferi* plasmids are mutually closely or moderately similar, except plasmid G, which is only weakly similar. Plasmid G is close ($\delta^* = 49$) to the *L. lactis* plasmid, perhaps implying that a low G+C Gram-positive bacterium may be the source of plasmid G. The two broad-range plasmids are mutually moderately similar ($\delta^* = 83$). This similarity might mean either that relatively close genome signatures promote plasmid establishment or that the plasmids have acquired their hosts' signatures during long-term residence. Experiments on conjugation can address issues such as specificity vs. wide host range and relevance of size and signature for plasmid compatibility. We interpret the similarities in signature between plasmids and their bacterial hosts as implying that they share much replication and repair machinery, perhaps because the prokaryotic cell is not compartmentalized to the degree that the eukaryotic cell is.

Genomic Signature δ^* Differences Among Mitochondrial Mt Genomes. Fig. 4 details δ^* differences among a broad spectrum of Mt genomes. A succinct summary is given in Table 1.

Mammalian Mt genome signatures appear almost as random samples of each other. Why are the genome signatures among vertebrate Mt sequences significantly more similar than those among the corresponding host nuclear genomic sequences (15)? δ^* differences among organisms putatively reflect variations in replication and/or repair (1–3, 6). The replication machinery for animal Mt DNA apparently varies less than that for host DNA, perhaps because Mt replication is less affected by changes in external environment or developmental programs than is host DNA replication. δ^* differences of Mt from protostomes (e.g., insect, mollusk) versus deuterostomes show δ^* differences mostly in the range 90–115, weakly similar. The worm (LUMTE and *C. elegans*) genomes compared with deuterostome Mt are closely to weakly similar, $\delta^* \approx 41$ –101. The fungal Mt sets (excluding *S. cerevisiae*) compared with deuterostome Mt also yield δ^* differences in the range 43–114. The *S. cerevisiae* mtDNA (≈ 78 kb) composition is an extreme anomaly attested to by the inordinate δ^* differences from all other Mt or nuclear genomes, contributed to by about 100

FIG. 4. Average δ^* differences among complete Mt genomes based on a single sequence sample for genomes of <60 kb and multiple samples of about 50 kb for larger genomes.

G+C-rich clusters, each about 50 to 100 bp long, separated by A+T-rich spacers and numerous transposable elements.

The two plant Mt genomes (ARATH, MARPO, see Fig. 4), of about 150-kb length, compared with animal Mt are *weakly similar to distant*. The green alga CHLRE Mt is *very distant* from most other Mt sequences ($\delta^* > 180$). The other two algal Mt sequences (PROWI, CHOCHR) are generally *weakly similar* to nonprotist Mt sequences. The protist Mt divide into two subgroups, P1 = (RECAM, ACACA) and P2 = (PARAU, TRYBR, LEITA). The RECAM Mt genome signature compared with deuterostome Mt is *weakly to distantly similar*. ACACA is a bit farther ($\delta^* \approx 94$ –144). The protist group P2 are mostly *distant to very distant* from animal Mt sequences. The plant *A. thaliana* Mt sequence is *distant* from mammalian Mt sequences, possibly suggesting a polyphyletic ancestry separating metazoan, plant, and protist Mt.

Animal Mt sequences show significant underrepresentations of CG dinucleotides, $\rho_{CG}^* \approx 0.40$ –0.60 (10), about to the same extent as occurs in vertebrate nuclear genomic sequences. Virtually all animal Mt maintain normal represen-

tations of TA dinucleotides, whereas the corresponding nuclear DNAs predominantly have TA in low relative abundance, suggesting that mtDNA may be thermodynamically less stable than nuclear DNA because the dinucleotide TA has the lowest stacking energies compared with all other base steps. The fungal Mt of *S. pombe* has $\rho_{CG}^* \approx 0.54$, typical of animal Mt. However, the *Podospora anserina* fungal Mt genome has ρ_{CG}^* in the normal range. The Mt genome of *A. thaliana* has $\rho_{CG}^* = 0.73$, significantly low. The single persistently high ρ^* value occurs for $\rho_{CC/GG}^* \geq 1.30$ in animal and fungal Mt sequences. δ^* differences between species parallel the corresponding nuclear δ^* differences, despite large differences between Mt and corresponding host nuclear signatures (15).

Discussion Prokaryotic molecular taxonomy heretofore has been derived predominantly from sequence comparisons among rRNA genes. There are many uncertainties and controversies regarding divisions among prokaryotes (for recent reviews, see refs. 16 and 17). Protein sequence comparisons and associated phylogenetic tree constructions are even more conflicting relative to evolutionary relationships (18, 19).

Table 1. Summary of δ^* differences of Mt genomes

Comparison	δ^*	Description
Among mammalian Mt	<40	Very close
Among all vertebrates including birds, frogs, fish	<67	Mainly close
Vertebrates vs. nonvertebrate deuterostomes	55–120	Moderate to weak similarity
Deuterostomes vs. protostomes	Mostly 34–152	Moderate to distant similarity
Fungi vs. animals	Mostly 43–128	Moderate to weak similarity
Plants vs. (animals and fungi)	84–179	Weak similarity to distant
<i>C. reinhardtii</i> vs. animals	Mostly 141–212	Distant to very distant

Conventional methods of phylogenetic reconstruction from sequence information employ only similarity or dissimilarity assessments of aligned homologous genes or regions. Some difficulties intrinsic to this approach compared to the use of genome signature include the following: (i) Alignments of distantly related long sequences (e.g., complete genomes) are generally not feasible for various reasons, including chromosomal rearrangements, whereas signature comparisons do not depend on alignments. (ii) Different phylogenetic reconstructions (trees) may result for the same set of organisms based on analysis of different protein, gene, or noncoding sequences. Attempts to overcome these conflicts by "averaging" over many proteins are problematic because of biases in species sampling, effects of lateral transfer, complications of gene duplications, and inadequacies and artifacts of phylogenetic methods. As the signature has remarkably low variance throughout the genome, a tree based on δ^* differences is independent of which genome segments of 50 kb (or longer) is used in its construction. (iii) Chimeric origins and lateral transfer between distantly related organisms complicate alignment-based phylogenies. Signature comparisons are unaffected by these factors. On the other hand, alignment-based comparisons may provide information on individual gene origins that the signature method does not. (iv) Tree construction derived from aligned sequences cannot be applied to organisms for which similar gene sequences are largely unavailable (e.g., for bacteriophages, diverse eukaryotic viruses). (v) That "lateral transfer" is pervasive among prokaryotic genomes is now widely appreciated. Vectors for lateral transfer include exogenous transposons, hitchhiking on plasmids and/or phages, movement via episomes, and cell fusions.

What possible reasons and mechanisms can account for the qualitative parallelism between the evolutionary development of host nuclear genomes and the development of Mt organelle genomes despite the pronounced difference between the Mt and host nuclear genome signatures? The Mt and nuclear genomes for animal and fungal organisms use independent DNA polymerase machinery (e.g., γ vs. α , ϵ , and δ subunits in mammals). Also, the methods of replication and the nature of the replication origins are fundamentally different. Explicitly, the animal and fungal Mt transcription-primed replication machinery is distinctive in that most of the heavy strand is synthesized first and the light strand subsequently, whereas the nuclear genomes are replicated bidirectionally from multiple origins. There appears to be no DNA excision repair mechanism to deal with cyclobutane dimers in the Mt, and no repair of bulky lesions (20). Mt DNAs in animals and fungi show elevated levels of single- and double-strand breaks, mismatches, and generally corrupted base pairings, probably due to a paucity of abasic site correction facilities and mismatch repair capacity in Mt genomes (21). Moreover, repair may be less urgent for Mt activity because each cell has many mitochondria (hundreds or thousands) and a modicum of impaired organelles may not significantly curtail energy production. We, nevertheless, propose that Mt genomes retain signatures close to those of their repair-competent prokaryotic ancestor.

The contrast between plasmids (which track host genomic signatures) and mitochondria (which do not) is sharp. The similar signatures of plasmids and hosts may have two bases: (i) As we have postulated for genome fusions (8), perhaps a plasmid whose signature is too different from that of the host

will not be accepted by it. This would prescribe a maximum possible signature deviation. The largest δ^* difference observed in our survey is *weakly similar* (or about the distance from human to sea urchin). (ii) During the plasmid's residence in its current host, the same pressures that homogenize the signature throughout the chromosome will also drive the plasmid's signature towards that of the host. Such amelioration has been postulated for the G+C content of laterally transferred DNA (22). We suspect that the signature should ameliorate even more rapidly, for both plasmids and laterally transferred chromosomal segments. Whereas most successful gene transfer between lineages is very likely intraspecific, an appreciable amount of transfer among distantly related bacteria seems to have accumulated over time (22, 23). Despite such transfer, the signatures currently observed are almost invariant throughout the genome. Some means based on codon usage biases for ascertainties of laterally transferred genes in bacterial organisms are set forth in ref. 23.

We are happy to acknowledge valuable discussions and comments on the manuscript by Drs. B. Edwin Blaisdell, Dale Kaiser, and David Relman. S.K. is supported in part by National Institutes of Health Grants 5R01GM10452-34 and 5R01HG00335-11 and National Science Foundation Grant DMS9704552.

1. Karlin, S. & Burge, C. (1995) *Trends Genet.* **11**, 283–290.
2. Blaisdell, B. E., Campbell, A. M. & Karlin, S. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 5854–5859.
3. Karlin, S., Mrázek, J. & Campbell, A. M. (1997) *J. Bacteriol.* **179**, 3899–3913.
4. Josse, J., Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **263**, 864–875.
5. Russell, G. J., Walker, P. M., Elton, R. A. & Subak-Sharpe, J. H. (1976) *J. Mol. Biol.* **108**, 1–23.
6. Karlin, S. (1998) *Curr. Opin. Microbiol.* **1**, 598–610.
7. Karlin, S. & Mrázek, J. (1996) *J. Mol. Biol.* **262**, 459–472.
8. Karlin, S., Brocchieri, L., Mrázek, J., Campbell, A. M. & Spormann, A. M. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 9190–9195.
9. Karlin, S., Campbell, A. M. & Mrázek, J. (1998) *Annu. Rev. Genet.* **32**, 185–225.
10. Cardon, L. R., Burge, C., Clayton, D. A. & Karlin, S. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3799–3803.
11. Karlin, S., Doerfler, W. & Cardon, L. R. (1994) *J. Virol.* **68**, 2889–2897.
12. Krieg, A. M., Yi, A.-K., Schorr, J. & Davis, H. L. (1998) *Trends Microbiol.* **6**, 23–27.
13. Olsen, R. H. & Shipley, P. (1973) *J. Bacteriol.* **113**, 772–780.
14. Bagdasarian, M., Lurz, R., Ruckert, B., Franklin, F. C., Bagdasarian, M. M., Frey, J. & Timmis, K. N. (1981) *Gene* **16**, 237–247.
15. Karlin, S. & Mrázek, J. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 10227–10232.
16. Brown, J. R. & Doolittle, W. R. (1997) *Microbiol. Rev.* **61**, 456–502.
17. Gupta, R. S. (1998) *Microbiol. Mol. Biol. Rev.* **62**, 1435–1491.
18. Gupta, R. S. & Golding, G. B. (1996) *Trends Biochem. Sci.* **21**, 166–171.
19. Budin, K. & Philippe, H. (1998) *Mol. Biol. Evol.* **15**, 943–956.
20. Shadel, G. S. & Clayton, D. A. (1997) *Annu. Rev. Biochem.* **66**, 409–435.
21. Yakes, F. M. & Van Houten, B. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 514–519.
22. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44**, 383–397.
23. Karlin, S., Mrázek, J. & Campbell, A. M. (1998) *Mol. Microbiol.* **29**, 1341–1355.