Photosynthesis Research **75:** 211–221, 2003.

Regular paper

Cyanobacterial signature genes

Kirt A. Martin[1], Janet L. Siefert[2], Sailaja Yerrapragada[1], Yue Lu[1], Thomas Z. McNeill[1,3], Pedro A. Moreno[1], George M. Weinstock[3], William R. Widger[1] & George E. Fox[1,*]
[1]*Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA;* [2]*Department of Statistics, Rice University, Houston, TX 77251-1892, USA;* [3]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; *Author for correspondence (e-mail: fox@uh.edu; fax: +1-713-743-8351)*

Received 28 June 2002; accepted in revised form 7 November 2002

Key words: comparative genomics, cyanobacteria, signature genes

Abstract

A comparison of 8 cyanobacterial genomes reveals that there are 181 shared genes that do not have obvious orthologs in other bacteria. These signature genes define aspects of the genotype that are uniquely cyanobacterial. Approximately 25% of these genes have been associated with some function. These signature genes may or may not be involved in photosynthesis but likely they will be in many cases. In addition, several examples of widely conserved gene order involving two or more signature genes were observed. This suggests there may be regulatory processes that have been preserved throughout the long history of the cyanobacterial phenotype. The results presented here will be especially useful because they identify which of the many genes of unassigned function are likely to be of the greatest interest.

Introduction

Although claims for the earliest fossilized cyanobacteria at 3.5 Ga (Schopf and Packer 1987) have been seriously questioned (Brasier et al. 2002) there is strong agreement that membrane biomarkers in well-preserved sediments reveal the presence of cyanobacteria at 2.7 Ga (Brocks et al. 1999). It is therefore of interest to studies of the early Earth to understand what shared properties these early cyanobacteria likely had. The obvious example is the ability to carry out oxygenic photosynthesis, which is widely regarded as the cause the rise of oxygen in the Archaean atmosphere at 2.2 Ga(Knoll 1999; Catling et al. 2001). There may, however, be other shared properties of the cyanobacteria that would have had more subtle impact on the early Earth and which may persist even to this day.

The ability to sequence whole genomes makes it possible to examine the distribution of genes in a very detailed way. The completion of the *Synechocystis* 6803 genome (Kaneko et al. 2001) made the full complement of genes carried by a cyanobacterium available for the first time. In order to determine which of these genes, if any, might contribute to a unique shared cyanobacterial genotype we compared this initial cyanobacterial genome to seven other genomes that are now available. Based on rRNA comparisons, these eight genomes represent five major lineages at the crown of the cyanobacterial tree (Turner et al. 1999) (Figure 1). Their common ancestor would predate the rise of the heterocyst, which has been dated at 2.1 Ga by geologic evidence (Amard and Bertrand-Sarfati 1997). An inter-comparison of the genomic content of these eight genomes, defines a working set of signature genes, Table 1. This signature set contains 181 genes that were initially among the almost 1000 putative open reading frames found in the *Synechocystis* 6803 genome (Kaneko et al. 2001).

Materials and methods

The gene content of eight completely sequenced cyanobacterial genomes was examined. The genomes

Table 1. The 181 cyanobacterial signature gene set derived from the set of core genes. Column headings indicate the cyanobacterial species used in this study and the numbers in each column refer to the gene name of the likely ortholog detected by the analysis reported here. The gene designators are those used by the individual genome projects as of October of 2002. In some cases, e.g. *Nostoc* the name consists of a contig followed by the gene number on that contig. In some cases the designator, e.g. sll0558 indicates a relative direction of transcription too. The signature genes are tabulated in the physical order in which they occur in *Synechocystis* sp. PCC 6803. This allows the identification of putative operons containing two or more signature genes by simply scanning the Table. Signature genes included in these putative operons are tabulated in bold.

Synechosystis sp. PCC6803	*Anabaena* sp. PCC7120	*N. punctiforme*	*P. marinus* MED4	*P. marinus* MIT9313	*Synechococcus* sp. WH8102	*Trichodesmium erythraeum*	*T. elongatus*
sll0558	all1826	458.29	765	2886	82	55.6548	tlr1594
sll1399	alr0301	459.44	1200	1210	28	12.1088	tlr0325
sll1398	all0801	452.29	1406	1052	3501	10.339	tlr0493
slr1495	alr4075	507.14	341	1921	2774	32.4372	tlr1240
slr1122	alr1700	474.20	1139	384	2939	10.349	tll2191
slr0728	alr5367	492.30	1594	1515	2420	70.7692	tll2101
slr0730	**alr4016**	**412.47**	**1670**	**2786**	**2040**	**100.511**	**tll1109**
slr0731	**alr4017**	**412.48**	**1671**	**3564**	**2043**	**100.512**	**tll1108**
slr0732	alr0613	434.12	–	2774	446	94.8798	tlr2434
slr0734	all4252	507.39	1877	2347	1095	46.5725	tll2053
slr0737	all0329	485.8	331	3845	2764	16.1883	tll1724
sll0226	all4289	504.113	600	474	1588	14.1557	tll1388
slr0250	alr4067	477.40	–	1079	121	13.1355	tll1572
ssr0390	asr4775	468.78	1390	980	241	1.151	tsr2273
sll1321	all0011	502.192	412	220	1456	69.7448	tlr0429
slr1780	all0216	504.12	646	–	1752	51.6329	tll1499
ssr2998	asl4482	466.40	1078	1305	570	20.2700	tsr0524
slr1796	all2716	466.40	858	2830	1656	–	tlr1421
slr1800	alr5279	468.70	715	2525	821	10.455	tll0399
sll1656	all3977/all4113	357.8	542	2695	3096	50.6270	tll0396
sll1654	alr4132	501.187	160	2219	2543	2.2628	tlr2269
sll1194	alr1216	441.3	1167	1395	879	8.2226	tll2409
slr1306	alr4172	501.122	1196	1198	3180	40.5311	tll1435
sll0933	all0748	472.34	1418	2919	3527	17.2115	tlr1208
ssl1972	asl4547	–	1961	2117	908	–	tsr0804
ssr1789	asl2354	507.12	398	194	1545	50.6231	tsr1916
slr1596	all1673	454.43	1631	2988	3388	2.2572	tll0748
slr1599	alr3954	429.13	1975	2131	927	38.4925	tll0146
slr1600	alr3603	344.2	2126	2553	1582	29.3820	tll1300
sll1507	alr4178	465.53	1885	2361	1110	1.53	tll1717
sll1752	all4871	623	1530	3330	1814	7.759	tlr0806
slr2032	all3013	439.16	281	2413	1160	15.1731	tlr1249
slr2034	**alr3844**	**493.50**	**2047**	**2045**	**2296**	**8.8173**	**tll1695**
ssr3451	**asr3845**	**493.51**	**2048**	**2046**	**2297**	**8.8174**	**tsr1541**
smr0006	**asr3846**	**493.52**	**2049**	**2047**	**2298**	**8.8175**	**tsr1542**
slr2049	alr0617	623	88	475	2221	1.138	tll1699
sll1934	alr1356	507.250	318	–	1224	57.6669	tlr1140
slr1908	alr2231	–	910	2622	2612	6.6850	tlr2324
slr1915	alr2980	423.49	514	289	1511	3.4146	tll2044
slr1918	all3259	505.36	1241	914	155	2.2455	tlr2348
ssr1698	asl4369	498.7	1467	3805	2716	–	tll2113
slr1034	all4779	501.53	1036	3462	629	22.3033	tlr0289
slr1384	all4892	–	1958	2103	897	13.1241	tll1632
slr1699	alr1085	464.53	977	2718	3536	4.5055	tll2476
ssr2843	asl5079	431.22	925	777	1720	11.782	tlr1507

Table 1. Continued

Synechosystis sp. PCC6803	*Anabaena* sp. PCC7120	*N. punctiforme*	*P. marinus* MED4	*P. marinus* MIT9313	*Synechococcus* sp. WH8102	*Trichodesmium erythraeum*	*T. elongatus*
slr1702	all4574	508.79	779	2959	3357	13.1258	tlr2221
sll1578	**alr0529**	**313.2**	**2053**	**22**	**2746**	**4.5118**	**tlr1958**
sll1577	**alr0528**	**313.1**	**2053**	**21**	**2745**	**4.5134**	**tlr1957**
sll1926	all2971	474.14	–	1141	3531	43.5529	tlr2451
sll1071	all2545	498.172	939	999	259	28.3708	tlr0249
ssl3451	asl2557	494.69	1295	1060	90	1.258	tsl2428
sll1797	all0938	501.62	1857	2289	1011	70.7710	tll1562
slr1896	alr4351	463.41	1555	3392	1960	3.4123	tlr1934
slr1900	alr4979	399.18	592	878	1870	2.2487	tll0916
slr1195	alr2014	458.41	1242	915	156	16.1844	tlr0208
slr1206	all2750	506.114	–	2023	2266	56.6664	tll1929
sll1968	alr3655	430.13	766	2887	83	13.1249	tlr1461
ssl3712	all7022	509.351	1673	3568	2054	10.338	tlr1849
sll1737	all4804	493.57	1231	2351	1098	1.110	tlr1073
slr1834	alr5154	456.7	374	1878	734	12.1150	tlr0731
slr1841	all7614	373.16	910	1722	2609	6.6851	tll1706
slr1535	all0967	463.50	–	3235	373	10.402	tll1668
ssr2595	asl0514	397.9	462	101	1880	–	tsl2208
sll1632	alr3857	502.59	1657	1267	548	16.1948	tlr0136
slr1220	alr1129	479.57	1923	3718	2206	23.3115	tll0853
sll1271	alr2231	–	636	654	2609	6.6850	tlr2324
sll0860	alr2431	489.37	773	2948	3345	3.4005	tll0315
sll1317	all2452	506.87	1061	3500	671	29.3867	tlr0960
ssl2598	asl0846	501.108	118	1823	2948	1.235	tsl1386
smr0009	all4665	501.138	1133	374	2930	62.7131	tsr1387
ssl3379	all0404	505.20	1255	992	254	16.1906	tsr1087
ssl3364	asl2850	372.10	138	3665	2669	2.2613	tsr1820
sll2013	all4333	423.22	387	877	2855	14.1574	tll1711
sll2002	alr4888	507.62	1023	3439	608	12.1104	tlr1986
slr2144	all0476	494.32	154	2216	2541	34.4556	tlr2015
sll1702	alr2762	492.31	993	2752	424	9.8694	tll1092
sll0854	all3378	352.13	378	1885	742	3.4011	tll2274
sll0851	alr4291	496.2	1563	3402	1973	14.1466	tlr1631
sll1162	alr1535	405.5	222	2536	583	13.1245	tlr2008
slr1263	alr1370	453.53	1632	3366	1943	–	tlr2402
slr1990	alr1215	441.1	1946	1394	878	18.2224	tll0200
sll1902	all4902	362.22	1009	2781	460	3.4058	tlr0207
sll1586	all0462	463.65	1192	1179	3173	3.4041	tll2375
ssl0461	asl3112	489.111	–	1855	1303	36.4704	tsl2468
sll0247	all4001	363.3	–	2809	1973	57.6686	tlr1050
sll1979	all4118	506.216	2130	926	896	77.7948	tsl0866
ssl2781	asr2932	409.17	–	365	2922	1.74	tsr0968
slr1506	all0969	445.50	1506	1500	2412	48.5830	tll1717
sll1418	all3076	493.23	1528	3322	1806	15.1653	tlr2075
sll1414	all0646	378.2	881	989	251	26.3452	tlr1134
slr0816	alr3417	507.98	596	3536	703	12.1036	tll0077
slr1342	alr3454	504.121	317	1749	1221	20.2820	tll2430
sll1898	all0949	493.119	1113	343	2904	87.8443	tll1894

Table 1. Continued

Synechosystis sp. PCC6803	*Anabaena* sp. PCC7120	*N. punctiforme*	*P. marinus* MED4	*P. marinus* MIT9313	*Synechococcus* sp. WH8102	*Trichodesmium erythraeum*	*T. elongatus*
slr1978	alr0484	468.81	2119	2545	1573	15.1804	tlr0284
sll1390	alr4100	486.28	2118	2987	3387	21.2881	tll0404
slr1470	alr3414	507.95	1584	3534	701	12.1033	tll0601
sll1382	all2919	477.92	1382	2301	226	5.6163	tlr1656
sll1376	alr1909	476.85	1050	3489	658	20.2798	tll1891
sll1372	alr0296	484.94	1109	337	289	23.3109	tlr1402
slr1273	alr2060	454.61	1016	2940	3335	69.7432	tlr0757
slr1287	alr0786	478.98	422	232	1467	1.31	tlr0672
slr1530	alr3399	506.25	1978	2134	932	26.3534	tll1658
slr1160	all4869	443.26	122	1828	1815	5.6050	tll0135
slr1677	alr4005	504.190	225	104	1327	8.8130	tll1550
ssr2831	asr4319	452.35	68	1384	2692	56.6628	tsl1567
slr0941	all2381	506.63	195	1731	2622	68.7417	tll0336
slr0954	alr1121	477.5	753	1007	50	88.8459	tlr0428
slr1623	all1732	434.47	1973	2129	925	1.67	tll0447
slr1636	alr4066	477.39	1310	1078	120	11.756	tll1133
slr1638	all5165	498.72	982	2730	3546	9.8638	tll2024
slr1645	all1258	429.9	1033	345	626	20.2791	tll2464
slr1649	all5339	378.29	–	3810	2747	21.2953	tlr2156
sll0427	all3854	502.159	133	1857	1306	1.186	tll0444
slr0443	alr3855	502.58	816	1985	2233	1.149	tll1059
sll0272	alr3297	428.21	812	3005	3402	29.3858	tlr0472
slr0280	all4343	423.12	1599	1970	2845	–	tll1850
sll0258	all0259	497.55	1510	3579	2069	32.4411	tll1285
sll0359	alr0946	501.182	1072	2777	686	14.1452	tll1150
ssr0657	asl3656	430.31	1404	1047	3498	13.1378	tsl0253
slr1926	all4664	464.2	1134	375	2931	62.7132	tlr0863
slr1946	all4180	366.10	–	1173	3165	37.4809	tll2292
slr1949	all4162	501.118	823	3037	1694	39.4981	tll1052
ssl0563	asr3463	507.20	309	1737	1209	3.4159	tsl1013
sll0295	alr3863	372.20	1808	1899	770	7.7572	tlr1691
slr0169	all3257	464.26	1171	449	2997	2.2608	tlr0311
sll0169	all2707	493.84	533	2677	3082	21.2839	tlr0758
sll0157	all5037	457.38	1459	3551	1020	15.1659	tlr2433
sll0350	alr1278	492.17	2026	2014	2258	22.2992	tlr1075
slr0376	all1871	506.229	402	200	3349	24.3302	tlr1877
ssl1417	asr0062	472.16	1952	2094	888	1.137	tsr1483
slr0013	alr4373	498.167	–	2962	3358	13.1261	tlr0729
sll0208	alr5283	468.71	1162	428	2986	13.1351	tll1313
sll0199	all0258	497.56	1000	2761	433	7.7632	–
slr0438	alr3231	502.3	761	1035	–	10.405	tsr1840
sll0071	all2549	454.38	1609	1755	1226	–	tll0418
ssr0109	asl0272	–	1992	2170	959	44.5634	tsr1584
sll0456	all3908	508.91	999	2760	432	20.2779	tlr0742
slr0630	alr3419	473.96	1834	294	1520	1.206	tlr0872
sll0609	**asl4507**	**455.49**	**1389**	**979**	**240**	**20.2811**	**tsr2284**
sll0608	**all4508**	**455.48**	**1388**	**977**	**239**	**20.2810**	**tlr1437**
slr0418	alr4674	382.16	1248	2927	3320	35.4689	tll0771

Table 1. Continued

Synechosystis sp. PCC6803	*Anabaena* sp. PCC7120	*N. punctiforme*	*P. marinus* MED4	*P. marinus* MIT9313	*Synechococcus* sp. WH8102	*Trichodesmium erythraeum*	*T. elongatus*
sll0372	all2849	472.78	126	1837	1295	2.2614	tlr1952
ssl0353	asl0940	501.60	224	103	1326	66.7305	tlr1577
ssl0352	asr0654	356.8	311	1739	1213	26.3516	tlr0636
slr0204	alr2465	382.21	1989	2165	951	9.8547	tll0488
slr0208	all1363	379.13	1546	1449	1243	84.8351	tlr0320
slr0906	all0138	494.8	80	3796	2708	12.1153	tlr1530
sll0544	alr3444	505.39	1840	1452	2366	34.4609	tlr1573
slr0575	alr3596	479.98	1037	3463	631	8.8215	tll0792
sll0832	alr4394	493.20	2023	2012	2255	54.6491	tlr0651
sll0827	alr3827	479.30	–	3454	623	49.5949	tlr1856
ssr1425	asr3137	475.30	184	1710	2596	24.3272	tsl2457
sll0822	all2080	388.32	1007	3518	2815	37.4783	tll2172
slr0503	alr0942	501.179	1926	3722	2210	66.7316	tlr2014
ssr1041	asr2378	506.212	1565	3407	589	68.7392	tsl1557
sll0584	alr4170	353.4	825	–	1696	15.1780	tll1063
slr0116	alr3707	493.86	1296	1062	91	35.4695	tll2308
ssl0546	**asr3457**	**504.124**	**79**	**3784**	**2705**	**2.2477**	**tlr2018**
sll0288	**alr3455**	**504.122**	**77**	**3782**	**2703**	**2.2475**	**tlr2016**
slr0304	alr3097	478.108	967	832	1716	38.4877	tll1363
sll0286	alr0113	320.6	467	1183	2768	47.5821	tlr1682
sll0662	**alr0045**	**506.101**	**403**	**201**	**1554**	**7.7540**	**tlr2140**
sll0661	**alr0044**	**506.100**	**402**	**200**	**1553**	**7.7539**	**tlr2139**
ssl1263	**asr0043**	**506.99**	**401**	**199**	**1552**	**7.7538**	**tsr2138**
slr0022	alr5129	435.38	807	2999	3396	13.1337	tlr0653
sll0031	alr2308	481.68	1807	1897	767	52.6407	tlr2308
slr0042	alr2231	362.17	910	899	2837	6.6850	tlr2324
sll0509	alr4466	472.37	788	2973	3374	80.8228	tll0991
sll1340	asl4395	493.128	120	1826	1289	1.240	tsl2214
slr1459	all2327	481.91	2053	22	85	40.5258	tlr2034
slr0651	all4101	486.93	201	2512	2631	21.2884	tlr0351
sll0621	alr3122	469.12	1771	3615	2117	19.2279	tlr0052
slr1557	all1338	448.41	586	320	2876	16.1838	tlr0610
slr1579	all4042	479.48	–	1692	2579	3.4044	tll1913
slr1655	all0107	423.1	377	1959	2830	118.973	tlr2404
slr1660	all4118	374.6	1974	2130	926	77.7948	tlr1444
slr1177	alr3362	507.131	–	1550	3563	16.1901	tlr0353
sll1109	alr3101	509.283	862	2834	1660	11.816	tll1867
slr0589	alr3980	357.7	327	3840	2758	1.189	tll0625
slr0590	alr3874	432.18	1844	3313	2372	53.6441	tlr2271
slr0598	alr2454	506.182	1125	382	2914	29.3790	tll0958

were *Synechocystis* sp. PCC 6803 (3.6 MB) (Kaneko et al. 2001), *Anabaena* PCC 7120 (7.2 MB), and *Thermosynechococcus elongates* BP-1 (2.6 MB) available at http://www.kazusa.or.jp/cyano/cyano.html, and *Synechococcus* WH8102 (2.72 MB), *Prochlorococcus marinus* MED4 (1.6 MB), *Prochlorococcus marinus* MIT9313 (2.4 Mb), *Nostoc punctiforme* (9.2 MB), and *Trichodesmium erythraeum* IMS101 (6.5 MB) available at http://jgi.doe.gov/JGI_microbial/html/index.html. Some of these sequences are currently in draft

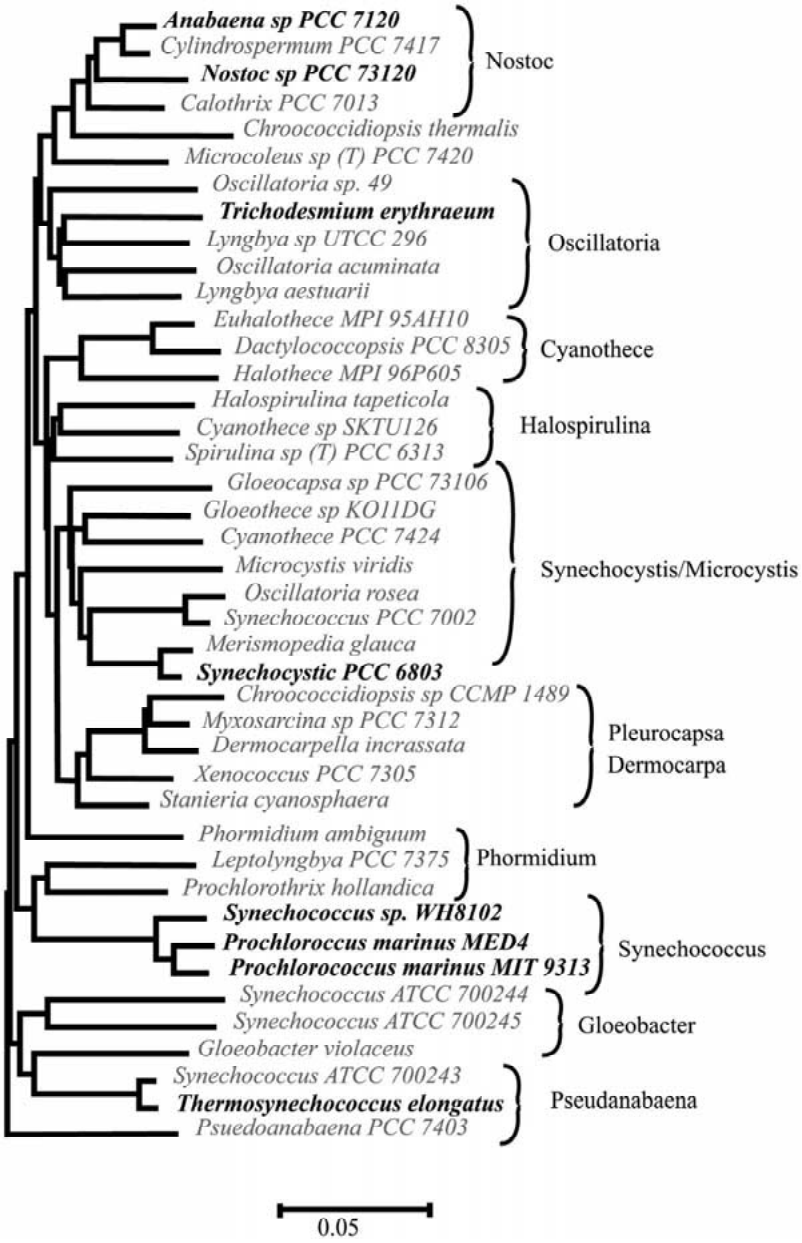

Figure 1. A representative phylogenetic tree of cyanobacteria showing the position of strains whose complete genome has been sequenced and used in this study. The cyanobacterial 16S rRNA tree was constructed from 1063 unambiguously aligned nucleotides under the Kimura 2-parameter using the Neighbor Joining tree making algorithm in Bioedit (Hall 1999). The branches are designated with order-level nomenclature (Turner et al. 1999 with modifications). Bold names indicate the position of the strains whose genomic sequences were publicly available in October of 2002.

form. At the time this work was undertaken, the *Synechocystis* sp. PCC 6803 was by far the best annotated genome among these and, as such, was used as a reference sequence for much of the work reported here. Except for *Anabaena* PCC 7120, each genome had previously been inter-compared with the other seven genomes and the results have been posted on the respective genome sites. The genes found to be common to at least seven cyanobacterial genomes were extracted and assembled into individual sequence files using the Bioedit platform (Hall 1999). Multigene sequence alignment was performed using CLUSTALW

in Bioedit (Higgins et al. 1994) and the results examined to verify that the genes extracted from the various genomes were likely homologs or orthologs.

Each of the conserved genes was next compared against the NCBI protein database by use of BLASTP. BLASTP tables were individually examined for score and corresponding organism. Proteins with E-values $<10^{-10}$ to species other than chloroplasts or chloroplast-containing eukaryotes were culled from the list. Because the NCBI site does not include data from the genomes of several photosynthetic bacteria, we separately examined the residual genes for affinity with genes in the genomes of *Chlorobium tepidum*, *Rhodobacter sphaeroides*, *Rhodopseudomonas palustris*, and *Rhodospirilum rubrum* and *Chloroflexus auranticus* with an expectation cutoff value $E < 10^{-6}$.

We also briefly examined the 181-gene signature set for putative cyanobacterial specific operons. The signature genes in Table 1 were arranged according to their position in the *Synechocystis* PCC 6803 genome. Because, the gene names in the other genomes typically relate to location in the genome we were able to quickly screen the table for sets of signature genes that were in close proximity to one another in all the genomes. The likely operons detected are shown as bold entries in Table 1.

Results

We herein report the results of the genomic comparisons for eight cyanobacteria. The inter-comparison allowed us to identify hundreds of genes that are shared by at least seven of the genomes and therefore comprise the core (Makarova et al. 1999) of the cyanobacterial genome. Of these core genes, it is found that 181 have not been found to have an obvious homolog or ortholog in non-cyanobacterial bacterial genomes (Table 1). Only 43 of these 181 signature genes have been associated with any specific functional role according to the recently revised annotation of the *Synechocystis* PCC 6803 genome at the Cyanobase web site (http://www.kazusa.or.jp/cyano/cyano.html). For the reader's convenience, these genes are separately listed with their genetic nomenclature and an indication of the function they are associated with in Table 2. Not surprisingly, 34 of these, including many of the Photosystem I and Photosystem II subunits, are directly or indirectly involved in photosynthesis. The remaining 9 known genes are involved in other functions that may not be directly related to photosynthesis. The overwhelming majority of the signature genes (138, or 76.2%) remain annotated as hypothetical genes in *Synechocystis* PCC 6803. Since equivalent genes are found in at least seven out of eight of the organisms it is clear that these are actual genes of unassigned function. These hypothetical genes include 16 genes that are designated as 'ycf' that are frequently found in chloroplasts. These are ycf 21, 23, 33, 34, 35, 36, 41, 49, 51, 52, 53, 54, 58, 60, 66, and 83.

Our screen of the signature set for genes of conserved proximity revealed six putative operons, Table 1. In two cases, nothing was known regarding the function of the genes. The other putative operons are: (a) a cluster of three Photosytem II genes consisting of *ycf*48, *psb*E and *psb*F; (b) a phycocyanin cluster containing *cpc*A and *cpc*B; (c) a cluster of three cell division associated proteins including the two septum site determining proteins *min*C, *min*E and the non-signature gene *mind*; and (d) two genes sll0608 and sll0609 that includes a putative homolog of transcription factor *dev*T.

Discussion

The comparison of eight cyanobacterial genomes allowed us to identify 181 genes that are found in all the cyanobacterial genomes. These genes do not have obvious homologs or orthologs in other bacterial genomes, whether photosynthetic or not. Together, these synapomorphic genes likely account for the unique shared characteristics of the cyanobacterial phenotype and are therefore a characteristic signature (Graham et al. 2000) of the group. The relative portion of the genes in the cyanobacterial signature set ranges from 2.6% of the total number of coding regions in the case of the large *Nostoc punctiforme* genome to 11.4% for the much smaller *Prochlorococcus marinus* MED4 genome. Since the list contains genes conserved primarily between cyanobacteria, and chloroplasts, it would not be expected to include genes acquired by lateral transfer, unless such events occurred before the branching of cyanobacteria.

This first approximation of a cyanobacterial signature set will likely be subject to modification as further data emerges. The addition of cyanobacterial genomes from currently unrepresented branches may on the one hand cause some genes to be relegated to being signatures of subgroups of the cyanobacteria. An example might be genes associated with thylakoid membranes,

Table 2. Listing of 43 signature genes that have been associated with some function. This subset of signature genes is loosely grouped according to function. The table indicates the usual genetic nomenclature for each gene. The brief annotation comments were taken from the 2002 annotation of the *Synechocystis* sp. PCC 6803 genome which is available at the cyanobase web site http://www.kazusa.or.jp/cyano/cyano.html.

PCC 6803 Gene name	Locus	Genetic comment
slr1834	*psa*A	P700 apoprotein subunit Ia
ssl0563	*psa*C	Photosystem I subunit VII
slr0737	*psa*D	Photosystem I subunit II
ssr2831	*psa*E	Photosystem I subunit IV
ssr0390	*psa*K	Photosystem I subunit X
slr1655	*psa*L	Photosystem I subunit XI
sll0226	*ycf*4	Photosystem I assembly related protein
slr0906	*psb*B	Photosystem II core light harvesting protein
sll0851	*psb*C	Photosystem II CP43 protein
ssr3451	*psb*E	Cytochrome b559 alpha subunit
smr0006	*psb*F	Cytochrome b559 b subunit
ssl2598	*psb*H	Photosystem II PsbH protein
smr0009	*psb*N	Photosystem II PsbN protein
sll0427	*psb*O	Photosystem II manganese-stabilizing polypeptide
sll1194	*psb*U	Photosystem II 12 kDa extrinsic protein
sll0258	*psb*V	Cytochrome c550
sll1398	*psb*W	Photosystem II reaction center W protein (*psb*13, *ycf*79)
slr1645	*psb*Z	Photosystem II 11 kD protein
slr2034	*ycf*48	Photosystem II stability/assembly factor
sll1418		similar to II oxygen-evolving complex 23K protein *psb*P
sll1317	*pet*A	ApoCytochrome *f*, component of cytochrome b6/f complex
sll0199	*pet*E	plastocyanin
sll0621	*ccd*A	putative c-type cytochrome biogenesis protein CcdA
sll1578	cpcA	Phycocyanin alpha subunit
sll1577	cpcB	Phycocyanin beta subunit
slr0116		Phycocyanobilin:ferredoxin oxidoreductase
sll1382:		Ferredoxin, petF-like protein
slr1459	*apc*F	Phycobilisome core component
ssr2595	*hli*B	High light-inducible polypeptide HliB
ssr1789	*hli*D	CAB/ELIP/HLIP-related protein HliD
slr1596	*pxc*A	Cytoplasmic membrane protein-light-induced proton extrusion.
sll1968	*pmg*A	Photomixotrophic growth related protein, PmgA
sll0247	*isi*A	Iron-stress chlorophyll-binding protein
ssl3364	*cp*12	CP12 polypeptide
slr1841		Probable porin; major outer membrane protein
sll1271		Probable porin; major outer membrane protein
slr0042		Probable porin; major outer membrane protein
sll1321	*atp*1	ATP synthase protein I
sll1908	*ser*A	D-3-phosphoglycerate dehydrogenase
slr0418		Putative transcripton factor DevT homolog
sll0169		Cell division protein Ftn2 homolog
ssl0546	*min*E	Septum site-determining protein MinE
sll0288	*min*C	Septum site-determining protein MinC

as these membranes are not present in *Gloeobacter*. On the other hand, some of the genomes used in this study are still undergoing analysis by their annotators, and it is possible that equivalent genes have been overlooked in some cases, resulting in an incomplete signature set. Finally, it should be appreciated that what constitutes the presence of an equivalent gene in other organisms is somewhat subjective. Not only will workers disagree on appropriate choices for BLASTP cutoff values, the program itself may give different values depending on the size of the database and type of filtering used. In addition, in some cases only a portion of a gene, e.g., a domain may be shared.

Regardless of uncertainties in precisely defining the signature set, it is clear that one can expect such sets to exist for at least some other groups or subgroups of related organisms as well. This is especially true for groupings that have characteristic properties such as the production of an endospore that involve multiple genes. The large numbers of genes in the cyanobacterial signature set thus provides further evidence that it may eventually be possible to determine phylogeny by gene content (Ochman and Bergthorsson 1995; Fitz-Gibbon and House 1999; Snel et al. 1999) for at least some groupings in the tree of life. This is important because bacterial genomes are dynamic, and are subject to repeated events of gene acquisition and deletion (Doolittle 1999; Jain 1999). In order to unravel these events, one needs to know what defines the essence of any genome. It may also be possible to use signature sets to construct an internal history of the group under study, if lateral transfer of the signature genes is uncommon within the group.

Of the 181 signature genes, 46 are included in a set of 434 genes that have been proposed as likely candidates for interdomain horizontal gene transfer (Koonin et al. 2001). Forty-five of the forty-six have blast-derived best hits to either *Arabidopsis* genes, or various chloroplast or cyanelle genomes. The remaining gene, sll0031, had an Archaeal best hit (Koonin 2001). This putative homology is to a region internal to sll0031 that contains a ferredoxin type motif rather than the whole gene. The inter-domain horizontal transfer that is being detected (Koonin 2001) in the 45 genes is interpreted to reflect an endosymbiotic event between cyanobacteria and eukaryotic organisms that led to the formation of the chloroplast.

The fact that the phylogenetic signal from these genes is still sufficient to define them as orthologs in extant cyanobacteria testifies to a continuing important biochemical role. This result further demonstrates that these forty-six genes were in fact widely distributed among cyanobacteria at the time when chloroplasts came into existence. The usual assumption would be that the remaining 135 were lost following the endosymbiotic event but they may have been transferred to the nucleus and not yet detected except in *Arabidopsis*. Alternatively, if the original endosymbiotic events occurred before 2.1 Ga it is not impossible that some of these genes were simply not yet present at this earlier stage.

Although gene order has been shown to be relatively unstable (Mushegian and Koonin 1996; Siefert et al. 1997; Itoh et al. 1999), proteins that function together in a pathway or structural complex are nevertheless likely to evolve in a correlated fashion (Pellegrini et al. 1999; Huynen et al. 2000). Several of the signature genes were found to be in close proximity to another signature gene in at least seven of the eight genomes. There were six of these putative operons containing at least two signature genes. We also observed cases in which a single signature gene was repeatedly associated with genes that are found in some other photosynthetic bacteria but not non-photosynthetic bacteria. At this stage, there is an unknown number of examples of conserved gene order involving one signature gene and one or more non-signature genes. The analysis performed here would not have detected putative operons of this type. Regardless of the numbers of these, the existence of several examples of a conserved operon-like architecture among the signature genes suggests the possible existence of regulatory systems that have been shared by all cyanobacteria for over two billion years.

Two cyanobacterial signature genes, *min*C and *min*E, that are septum-site determining proteins could impact the coordination of nitrogen fixation and an oxygen evolving complex. Although *Synechocystis* PCC 6803 does not fix nitrogen, at least two of the cyanobacterial species in this study are known to do so. In the case of sheathless, unicellular cyanobacteria, there is evidence for multiple gains and or losses of nitrogen-fixing ability (Turner 1997). It may well be that the cyanobacterial core includes much of the underlying machinery needed for nitrogen fixation allowing relatively rapid evolution of that capability. Thus, some groups of ancient cyanobacteria might have been able to evolve nitrogen fixation and coordinate it with photosynthesis in response to ecosystem challenges during the Archaean.

The most interesting aspect of the signature set is of course the large number of genes that have not

been assigned a function. This may simply reflect incompleteness in efforts to correlate functional studies with the genomic results. However, if one accepts the finding at face value, it clearly suggests these organisms have more shared characteristics than has been appreciated to date. In particular, there are either far more genes associated with the cyanobacterial photosynthetic processes than previously thought, or that cyanobacteria possess pathways and/or other biochemical activities that are largely unknown. Given the amount of effort that has been focused on understanding photosynthesis in cyanobacteria, it is unlikely that many of these unassigned genes are directly involved in that process. It is far more likely that they are carrying out key supporting roles, such as coordinating the various activities associated with photosynthesis.

There is little direct evidence at this stage as to what these uncharacterized gene products do. Since operons frequently consist of genes that are functionally related, a more general study of conserved neighboring genes will likely provide clues as to functional roles in some cases. Likewise, examination of the putative proteins from a structural perspective might allow one to recognize characteristic features such as the ability to span membranes. In the end, the assignment of function to the unknown genes will require detailed studies of biochemical function.

One especially relevant biochemical study was a recent DNA microarray analysis of the expression patterns of all PCC-6803 genes during acclimation to high light (Hihara et al., 2001). Although the ability to observe low expression genes was limited to some extent by ribosomal RNA contamination, more than 160 genes were classified as having one of six characteristic response patterns. Three signature genes of known function, *psb*O, *psb*V, and *apc*F were initially repressed and then later increased. It is notable, however, that no other signature gene exhibited an identifiable response to the change in light intensity. Since the characteristic cyanobacterial genes are apparently not involved, one can likely expect that acclimation to light may differ considerably in the various cyanobacterial lineages. Another example of relevant data are the two-dimensional protein gel separations, which are available at the Cyanobase site (http://www.kazusa.or.jp/cyano/cyano.html). Fourteen distinct protein products were found in the thylakoid membrane fraction. Four of these are signature genes of known function, *psa*C, *psa*E, and *psb*O. In addition, three signature genes of unknown function, slr1623, ssl0352 and ssr2998 are found in this fraction.

At this stage, one can ultimately only speculate on the function of the proteins encoded by the unassigned signature genes. However, instead of focusing attention on the almost 1000 hypothetical genes seen in the *Synechocystis* 6803 genome, the results presented here will allow photosynthesis researchers to target efforts to less than 140 genes that are likely to be of considerable interest.

Acknowledgements

This work was supported in part by funding from: the NIH National Human Genome Research Institute to K.M. (F31-HG00186), NSF Postdoctoral Bioinformatics grant (9974214) and an NSF LExEn grant (0085562) to J.L.S., and grants from the National Space and Aeronautics Administration Exobiology Program (NAG5-8140 and NAG5-12366), the Institute of Space Systems Operations and the Shell Scholars Program at the University of Houston to G.E.F.

References

Amard B and Bertrand-Sarfati J (1997) Microfossils in 2000 Ma old cherty stromatolites of the Franceville Group, Gabon. Precambrian Res 81: 197–221.

Brasier MD, Green OR, Jephcoat AP, Kleppe AK, van Kranendonk MJ, Lindsay JF, Steeles A and Grassineau NV (2002) Questioning the evidence for Earth's oldest fossils. Nature 416: 76–81

Brocks JJ, Logan GA, Buick R and Summons RE (1999) Achaean molecular fossils and the early rise of eukaryotes. Science 285: 1033–1036

Catling DC, Zahnle KJ and McKay C (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. Science 293: 839–843

Doolittle WF (1999) Lateral genomics. Trends Cell Biol 9: M5-M8

Fitz-Gibbon ST and House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27: 4218–4222

Gaasterland T and Ragan MA (1998) Constructing multigenome views of whole microbial genomes. Microb Comp Genomics 3: 177–192

Graham DE, Overbeek R, Olsen GJ and Woese CR (2000) An archaeal genomic signature. Proc Natl Acad Sci USA 97: 3304–3308

Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95–98

Hihara Y, Kamel A, Kanehisa M, Kaplan A and Ikeuchi M (2001) DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. Plant Cell 13: 793–806

Higgins D, Thompson J, Gibson T Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: improving the sensitivity

of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Huynen M, Snel B, Lathe W and Bork P (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res 10: 1204–1210

Itoh T, Takemoto K, Mori H and Gojobori T (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol Biol Evol 16: 332–346

Jain KK (1999) Strategies and technologies in functional genomics. Drug Discov Today 4: 50–53

Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M and Tabata S (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120 (supplement). DNA Res 8: 227–253

Kasting JF and Siefert JL (2001) Biogeochemistry. The nitrogen fix. Nature 412: 26–27

Knoll AH (1999) PALEONTOLOGY: enhanced: a new molecular window on early life. Science 285: 1025-1026

Koonin, EV, Makarova, KS and Aravind, L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Ann Rev Microbiol 55: 709–742

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, and Koonin EV (1999) Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9: 608–628

Mushegian AR and Koonin EV (1996) Gene order is not conserved in bacterial evolution. Trends Genet 12: 289–290

Ochman H and Bergthorsson U (1995) Genome evolution in enteric bacteria. Curr Opin Genet Dev 5: 734–738

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

Schopf JW and Packer BM (1987) Early Achaean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. Science 237: 70–73

Siefert JL, Martin KA, Abdi F, Widger WR and Fox GE (1997) Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. J Mol Evol 45: 467–472

Snel B, Bork P and Huynen MA (1999) Genome phylogeny based on gene content. Nat Genet 21: 108–110

Turner S, Pryer KM, Miao VP and Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Euk Microbiol 46: 327–338