EVOLUTION AND POPULATION GENETICS OF *PROCHLOROCOCCUS MARINUS*

by

Ena Urbach

B.S. Molecular Biophysics and Biochemistry
Yale University, 1980

Submitted in partial fulfillment of the
requirements of the

DEGREE OF DOCTOR OF PHILOSOPHY

in Biology
and Civil and Environmental Engineering

at the
Massachusetts Institute of Technology

February 1995

Signature of Author _____
Departments of Biology and
Civil and Environmental Engineering

Certified by _____
Sallie W. Chisholm, Professor
Biology and Civil and Environmental Engineering

Accepted by _____
Departmental Committee on Graduate Studies
Civil and Environmental Engineering
Joseph M.Sussman

Accepted by _____
Departmental Committee on Graduate Studies
Biology
Jo-Ann Murray

EVOLUTION AND POPULATION GENETICS OF *PROCHLOROCOCCUS MARINUS*

by

Ena Urbach

Submitted to the Departments of Biology and
Civil and Environmental Engineering
February, 1995, in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in
Biology and Civil and Environmental Engineering

ABSTRACT

Prochlorococcus marinus is a tiny, photosynthetic marine prokaryote containing the unusual photosynthetic pigments divinyl chlorophylls *a* and *b*, and lacking phycobilisomes, the light harvesting apparatus found in most cyanobacteria. The unique pigments of this organism suggest a phylogenetic affinity with the Prochlorophyta, a recently described group of (monovinyl) chlorophyll *a* and *b*-containing prokaryotes which also lack phycobilisomes. The Prochlorophyta have been proposed as a monophyletic group sharing a common ancestry with "green" chloroplasts in algae and higher plants.

Phylogenetic analysis of relationships among *P. marinus*, the prochlorophytes *Prochloron sp.* and *Prochlorothrix hollandica*, and other members of the cyanobacterial radiation indicates that the distribution of prochlorophytes within the cyanobacterial radiation is polyphyletic, and that none of the known chlorophyll *b*-containing prokaryotes is specifically related to chloroplasts. This result, inferred from 16S ribosomal RNA (rRNA) sequences, has been confirmed by a parallel, independent study using *rpoC* gene sequences. The polyphyletic distribution of prochlorophytes and green chloroplasts among the cyanobacteria implies that photosynthetic systems employing (monovinyl and divinyl) chlorophyll *b* arose several times during the evolution of photosynthetic organisms. However, since the publication of these results it has been pointed out that they are also consistent with the existence of an ancestral cyanobacterium containing both chlorophyll *b* and phycobilisomes, with subsequent loss of one or the other light harvesting system in all known surviving lineages.

P. marinus is a major constituent of the photosynthetic picoplankton across wide regions of the tropical and subtropical open oceans, with cell densities of 10^5 cells/ml frequently encountered in both the Sargasso Sea and the Pacific. Cultured strains of this ecologically important organism isolated from Sargasso Sea, north Atlantic, Mediterranean and Pacific waters were found by phylogenetic analysis of 16S rRNA, *psbB* and *petB* and *D* sequences to belong to a single lineage within the cyanobacteria. *P. marinus* shares this lineage with strains of marine A *Synechococcus*, a planktonic cyanobacterium which also shares its open ocean habitat.

An investigation into the genetic structure of *P. marinus* field populations in depth profiles from the Sargasso Sea and the Gulf Stream revealed a high degree of genetic heterogeneity within water samples, detected by partial sequencing of cloned PCR products containing portions of the single-copy genes *petB* and *D* and their intergenic region, amplified from flow cytometrically sorted cells. Populations within water samples contained a minumum of six alleles recovered from a sample of 23 clones.

The presence of numerous additional alleles were implied by rarefaction analyses, which indicated that diversity was incompletely sampled in datasets containing 19 to 28 clones per water sample. Overlapping sets of alleles were recovered from the two water columns, from different depths within each water column and from flow cytometrically distinguishable subpopulations within water samples, suggesting that each of these populations drew its membership from a single gene pool.

Thesis Subervisor: Prof. Sallie W. Chisholm

Title: Professor of Biology and Civil and Environmental Engineering

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Penny Chisholm, for allowing me an unusual degree of intellectual freedom in pursuing what I wanted to study. Penny's intellectual courage has allowed her to develop a laboratory with an extraordinary range of skills and interests. I would also like to thank the members of my thesis committee, past and present, who have helped me with sharp critical eyes and large amount of patience: Alan Grossman and Ethan Signer of the Biology Department, Bill Thilly of the Department of Civil and Environmental Engineering, Bruce Levin of the Zoology Departement at University of Massachusetts at Amherst (and currently at Emory University), John Waterbury of the Woods Hole Oceanographic Institution and Mitchel Sogin of the Center for Molecular Evolution at the Marine Biological Laboratories. Special thanks to Bill Thilly for his generous contributions of laboratory space, equipment and supplies and for his intellectual input and to Mitchel Sogin for generous contributions of his phylogenetic experitise and computer time.

Over the years at MIT I have been fortunate to learn from a number of talented individuals who have been generous in sharing their technical expertise. Brian Binder, Phouthone Keohavong, Jasper Reese, Jason Seaman, Michael Way and Sheila Frankel have all served as both teacher and friend, and without them I would never have been able to complete this thesis. Samantha Roberts has been a great companion and source of information as the other "molecular person" in the Parsons lab. I would like to thank Cremilda Dias for her excellent technical help in obtaining four of the 16S ribosomal gene seuqneces used in Chapter Two.

I have also had the good fortune to interact with the unflaggingly friendly and stimulating members of the Chisholm laboratory: Ginger Armbrust, Lana Arras, Kent Barres, Brian Binder, Jim Bowen, Jeff Dusenberry, Sheila Frankel, Mark Gerath, Karina Gin, Liz Mann and Lisa Moore.

Finally, I would like to acknowledge the financial support provided by NSF (BSR-9020254) and by NIH through a training grant to the Department of Biology.

TABLE OF CONTENTS

LIST OF FIGURES

Page

APPENDIX IV

LIST OF TABLES

Page

APPENDIX III

INTRODUCTION

G. Ledyard Stebbins has pointed out that "the ultimate aim of scientific endeavor in any discipline is to obtain facts by reductionist methods and use them to synthesize broadly integrated theories that provide new insights into the world of nature" (Stebbins 1982). The evolutionary and population genetic studies that comprise this thesis have been designed to obtain, by the reductionist methods of DNA sequencing and phylogenetic analysis, information leading to new theories about the evolution of light harvesting pigments in oxygenic photosynthetic organisms, and about the population structure of a prokaryotic phytoplankter which inhabits the largest and possibly best-mixed environment on earth, the tropical and subtropical open oceans.

Procholorococcus marinus is a eubacterial photosynthetic plankter less than 0.8 μm in diameter which contains the unusual photosynthetic pigments divinyl chlorophylls *a* and *b* (chl a_2 and b_2), and lacks phycobilisomes, multicomponent protein and pigment complexes used for light harvesting by most cyanobacteria (Chisholm et al 1988, 1992, Goericke and Repeta 1992). Upon its discovery in 1988, the unusual pigments found in *P. marinus* suggested a taxonomic affinity with the newly recognized *Prochlorophyta* (Lewin 1977), photosynthetic prokaryotes containing "normal" chlorophylls *a* and *b* (chl a_1 and b_1) and lacking phycobilisomes (Chisholm et al. 1988).

According to the endosymbiotic hypothesis proposed by Mereschkowsky (1905, 1910), plastids of higher plants and green algae originated in a free living prokaryote containing green pigments, and which came to live endosymbiotically within the cytoplasm of a eukaryote. Over the course of the symbiosis, the chl *a* and *b*-containing prokaryote degenerated into an organelle incapable of independent survival (Margulis 1970, 1981, Raven 1970). At the time research for this thesis began, the evolutionary origin of photosynthetic organelles within the cyanobacteria had become well established by molecular phylogenetic analyses (Fox et al. 1980, Woese 1987, Giovannoni et al.

1988, Van den Eynde et al. 1988, Witt and Stackebrandt 1988, Ludwig et al. 1990, Valentin and Zeitsche 1990a, b), but whether the chlorophyll *b* photosynthetic antenna had been "invented" by an intracellular prokaryote containing phycobilisomes or arose by endosymbiosis of a free-living, chl *b*-containing prokaryote remained to be established (Cavalier-Smith 1982). The recently discovered chlorophyll *b*-containing prokaryotes *Prochloron didemni* (Lewin 1977, Lewin and Cheng 1989), and *Prochlorothrix hollandica* (Burger-Wiersma et al. 1986) had been proposed as descendants of the putative free-living, chlorophyll *b*-containing ancestor to green chloroplasts, but *P. hollandica* had subsequently been shown to branch separately from cholorplasts in 16S rRNA and *psbA* phylogenetic trees (Turner et al. 1989, Morden and Golden 1989a, b).

Chapter One of this thesis contributed to the formulation of new hypotheses about the origin of chlorophyll *b*-containing photosynthetic antennae by providing a 16S rRNA phylogenetic analysis of relationships among the prochlorophytes *P. marinus*, *Prochloron sp., P. hollandica*, chloroplasts and other cyanobacteria (Urbach et al. 1992). The results were congruent with those of an independent study using *rpoC* sequences (Palenik and Haselkorn 1992). Chapter Four is an Epilogue discussing responses in the literature to this publication.

Chapter Two of this thesis is a phylogenetic study which infers evolutionary relationships among *P. marinus* cultures isolated from diverse ocean provinces, expanding on the previous work. Phylogenetic patterns inferred from three sequences, 16S rRNA, *psbB* and *petB/D*, are used to address questions of whether oceanic prochlorophytes derive from multiple lineages dispersed among the cyanobacteria, whether phylogenetic patterns for the three genes are consistent, and whether they reflect geographic relationships among sites of *P. marinus* culture isolation. In addition, this study provides sequence data which can be used to link the properties of *P. marinus* in

culture to those of populations in the sea and a phylogenetic framework for interpretation of the evolution of characteristics among *P. marinus* cultures.

P. marinus is a major constituent of the picophytoplankton in tropical and subtropical regions of the open ocean, with depth integrated cell abundances in the Sargasso Sea and north Pacific of 5×10^8 and 2×10^9 cells cm^{-2}, respectively (Olson et al. 1990, Cambell and Vaulot 1993) and a global population of approximately 10^{26} cells (according to a back of the envelope calculation which assumes a tropical and subtropical open ocean area of 2×10^{14} m^2, an average *P. marinus* cell density of 10^4 cells/ml and a euphotic zone depth of 150 m). Experiments with cloned, cultured isolates have established that individual *P. marinus* genotypes can grow over wide ranges of light intensity and temperature, but differ in their growth potential at environmentally relevant extremes of high and low light (Partensky et al. 1993, Moore et al. 1994). In order to enable future workers to interpret the patterns of distribution and abundance in *P. marinus* field populations and to use the properties of *P. marinus* in culture to predict the growth and succession of populations in the wild, it is necessary to assess the patterns of distribution of *P. marinus* genotypes in natural populations (Wood 1988, Wood and Leatham 1992, Weisse 1993). Specifically, it is necessary to answer the questions, are *P. marinus* natural populations genetically heterogeneous, and what is the temporal and spatial scale of genetic variability among populations?

Chapter Three is a direct analysis of *P. marinus* populations in which the questions of genetic heterogeneity and differences among populations at different depths and in different water masses are addressed by sequencing cloned *petB/D* PCR products amplified from natural populations sorted by flow cytometry. With the basic outline of the population structure of *P. marinus* known, it becomes possible to formulate improved

hypotheses about properties and processes involving this important component of the marine food web.

REFERENCES

Bryant, D.A. (1992). Puzzles of chloroplast ancestry. Curr. Biol. 2:240-242.

Burger-Wiersma, T., Veenhuis, M., Korthals, J., Van der Weil, C.C.M. and Mur, L (1986). A new prokaryote containing chlorophylls a and b. Nature 320:262-294.

Campbell, L., and Vaulot, D. (1993). Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawii (station ALOHA). Deep-Sea Res. 40:2043-2060.

Chisholm, S.W., R.J. Olson, E.R. Zettler, R. Goericke, J. Waterbury, and N. Welschmeyer. (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. Nature, 334(6180):340-343.

Chisholm, S.W., Frankel, S.L., Goericke, R., Olson, R.J., Palenik, B., Waterbury, J.B., West-Johnsrud, L. and Zettler, E.R. (1992). Prochlorococcus marinus nov. gen nov. sp.: a marine prokaryote containing divinyl chlorophyll a and b. Arch. Microbiol. 157:297-300.

Cavalier-Smith, T. (1982). The origins of plastids. Biol. J. Linn. Soc. 17:289-306>

Fox, G., Stackebrandt, E., Hespell, R., Gibson, J., Maniloff, J., Dyer, T., Wolf, R., Balch, W., Tanner, R., Magrum, L., Zablen, L., Blekemore, R., Gupta, R., Bonen, L., Lewis, B., Stahl, D., Luehrsen, K., Chen, K. and Woese, C. (1980). The phylogeny of prokaryotes. Science 209:457-463.

Giovannoni, S.J., Turner, S., Olsen, G.J., Barnes, S., Lane, D.J. and Pace, N.R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. J. Bacteriol. 170:3584-3592.

Goericke, R. and Repeta, D.J. (1992). The pigments of Procholorcoccus marinus: the presence of divinyl chlorophyll a and b in a marine prokaryote. Limnol. Oceanogr. 37:425-433.

Lewin, R.A. (1977). Prochloron, type genus of the Prochlorophyta. Phycologia 16:217.

Lewin, R.A. (1981). Prochloron and the theory of symbiogenesis. Ann. N.Y. Acad. Sci. 361:325-329

Lewin, R.A. and Cheng, L. (1989). Introduction. In Prochloron: a microbial enigma. Lewin, R.A., and L. Cheng, eds. Chapman and Hall.

Ludwig, W., Weizenegger, D., Betzel, D., Leidel, E., Lenz, T., Ludvigsen, D., Mollenhoff, D., Wenzig, P. and Schleifer, K.H. (1990). Complete nucleotide sequences of seven eubacterial genes coding for the elongation factor Tu: functional, structural and phylogenetic evaluations. Arch. Microbiol. 153:241-247.

Margulis, L. (1970). Origin of eukaryotic cells. New Haven, Yale University Press.

Margulis, L. (1981). Symbiosis in Cell Evolution. Freeman.

Martin, W., Somerville, C.C. and Loiseaux-de Goer, S. (1992). Molecular phylogenies of plastid origins and algal evolution. J. Mol. Evol. 35:385-404.

Mereschkowsky, C. (1905). Uber Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol. Centralbl 25:593-604.

Mereschkowsky, C. (1910). Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. Biologische Centralblatt 30:278-303; 321-347; 353-367.

Moore, L.R., Goericke, R., and Chisholm, S.W. (1994). The comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, flourescence and absorptive properties. Mar. Ecol. Prog. Ser., *in press*.

Morden, C.W. and Golden, S.S. (1989a). *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts. Nature 337:382-385.

Morden, C.W. and Golden, S.S. (1989b). Corrigendum: *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts. Nature 339:400.

Olson, R.J., Chisholm, S.W., Zettler, E.R., Altabet, M.A., and Dusenberry, J.A. (1990) Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. Deep Sea Res. 37, 1033-1051.

Palenik, B. and Haselkorn, R. (1992). Multiple evolutionary origins of prochlorophytes, the chlorophyll b-containing prokaryotes. Nature 355:265-267.

Partensky, R., Hoepffner, N., Li, W.K.W., Ulloa, O. and Vaulot, D. (1993). Photoacclimation of Prochlorococcus sp. (Prochlorophyta) strains isolated from the north Atlantic and the Mediterranean Sea. Plant Physiol. 101:285-296.

Raven, P. (1970). A multiple origin for plastids and mitochondria. Science 169:641-646.

Stebbins, G.L. (1982). Modal themes: a new framework for evolutionary syntheses. In Perspectives on Evolution (R. Milkman, ed.). Sinaur pp. 1-14.

Turner, S., Burger.Wiersma, T., Giovannoni, S.J., Mur, L.R., and N.R. Pace. (1989). The relationship of a prochlorophyte *Prochlorothrix hollandica* to green chloroplasts. Nature 337:380-382.

Urbach, E., Robertson, D. and Chisholm, S.W. (1992). Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355:267-269.

Valentin, K. and Zetsche, K. (1990a). Rubisco genes indicate a close phylogenetic relationship between the plastids of Chromophyta and Rhodophyta. Plant Mol. Biol. 15:575-584.

Valentin, K. and Zetsche, K. (1990b). Structure of the Rubisco operon from the unicellular red alga Cyanidium caldarium: evidence for a polyphyletic origin of plastids. Mol. Gen. Genet. 222:425-430.

Van den Eynde, H., De Baere, R., De Roeck, E., Van de Peer, Y., Vandenberghe, A., Willekens, P. and De Wachter, R. (1988). The 5S ribosomal RNA sequences of a red algal chloroplast and a gymnosperm chloroplast. Implications for the evolution of plastids and cyanobacteria. J. Mol. Evol. 27:126-132.

Witt, D. and Stackebrandt, E. (1988). Disproving the hypothesis of a common ancestry for the *Ochromonas danica* chrysoplast and *Heliobacterium chlorum.* Arch. Microbiol. 150:244-248.

Woese, C.R. (1987). Bacterial Evolution. Microbiol. Rev. 51:221-271.

Weisse, T. (1993). Dynamics of autotrophic picoplankton in marine and freshwater ecosystems. Adv. Microbial. Ecol. 13:327-370.

Wood, A.M. (1988). Molecular biology, single cell analysis and quantitative genetics: new evolutionary genetic approaches in phytoplankton ecology. In *Immunochemical Approaches to Coastal, Estuarine and Oceanographic Questions* (C.M. Yentsch, F.C. Mague and P.K. Horan, eds.) Springer-Verlag pp. 41-73.

Wood, A.M., and Leatham, T. (1992). The species concept in phytoplankton ecology. J. Phycol. 28:723-729.

Chapter One

MULTIPLE EVOLUTIONARY ORIGINS OF PROCHLOROPHYTES WITHIN THE CYANOBACTERIAL RADIATION

Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation

Ena Urbach, Deborah L. Robertson* & Sallie W. Chisholm

Ralph M. Parsons Laboratory,
48-425 Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA
* Department of Molecular Genetics and Cellular Biology,
University of Chicago, Chicago, Illinois 60637, USA

THE taxonomic group Prochlorales (Lewin 1977) Burger-Wiersma, Stal and Mur 1989 was established to accommodate a set of prokaryotic oxygenic phototrophs which, like plant, green algal and euglenoid chloroplasts, contain chlorophyll *b* instead of phycobiliproteins. Prochlorophytes were originally proposed (with concomitant scepticism[1–3]) to be a monophyletic group sharing a common ancestry with these 'green' chloroplasts[4–6]. Results from molecular sequence phylogenies, however, have suggested that *Prochlorothrix hollandica*[7] is not on a lineage that leads to plastids[8–12]. Our results from 16S ribosomal RNA sequence comparisons, which include new sequences from the marine picoplankter *Prochlorococcus marinus*[13,14] and the *Lissoclinum patella* symbiont *Prochloron* sp.[15], indicate that prochlorophytes are polyphyletic within the cyanobacterial radiation, and suggest that none of the known species is specifically related to chloroplasts. This implies that the three prochlorophytes and the green chloroplast ancestor acquired chlorophyll *b* and its associated structural proteins in convergent evolutionary events. We report further that the 16S rRNA gene sequence from *Prochlorococcus* is very similar to those of open ocean *Synechococcus* strains (marine cluster A (ref. 16)), and to a family of 16S rRNA genes shotgun-cloned from plankton in the north Atlantic and Pacific Oceans[17–19].

The order Prochlorales subsumes two formally described genera: *Prochloron*, a symbiont of marine ascidians[20], and *Prochlorothrix*, a free-living, filamentous, fresh-water plankter[7]. Like green chloroplasts, these prokaryotes have appressed thylakoid membranes containing chlorophylls *a* and *b* and lacking phycobilisomes. The recently discovered *Prochlorococcus*, a free-living, tiny unicell that is widespread and abundant in the oceans[13], shares these traits with its 'relatives', but differs in that it contains divinyl chlorophylls *a* and *b* (exclusively), and α- rather than β-carotene[14,21].

TABLE 1 Evolutionary distance and fractional similarity matrix for 16S rRNA sequences from *A. tumefaciens* and members of the prokaryotic oxygenic phototroph radiation

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. *Agrobacterium tumefaciens*		813	798	805	810	808	806	811	802	812	794	812	798	802	797	788	805	808
2. SAR100	215		918	978	887	878	892	889	892	891	895	892	889	892	885	844	985	993
3. *Synechococcus* PCC6301	236	87		913	883	873	893	881	887	883	895	889	893	911	886	856	919	911
4. *Prochlorococcus marinus*	226	23	93		886	874	882	887	897	892	892	897	888	898	886	852	975	975
5. *Prochloron* sp.	219	122	128	123		920	906	885	921	941	903	920	903	885	910	879	883	885
6. *Synechocystis* PCC6308	221	134	139	138	85		900	854	892	908	885	896	889	866	888	848	868	875
7. *Synechocystis* PCC6906	225	117	115	129	101	108		863	923	914	899	892	895	883	890	845	885	890
8. *Gloeobacter* PCC7421	218	120	129	122	125	163	151		883	878	898	905	883	889	884	855	881	885
9. *Gloeothece* PCC6501	229	117	122	110	83	117	81	127		914	903	907	895	890	896	862	891	886
10. *Myxosarcina* PCC7312	216	118	127	117	62	98	91	134	91		906	913	896	892	893	866	886	889
11. *Lyngbya* PCC7419	242	113	113	117	104	125	109	110	104	101		923	908	892	910	868	892	896
12. *Oscillatoria* PCC6304	217	117	120	110	85	111	117	102	99	93	81		901	896	897	867	888	893
13. *Anabaena* PCC7122	236	120	115	121	104	120	113	127	113	112	98	106		876	905	872	891	886
14. *Prochlorothrix hollandica*	231	116	95	110	125	147	127	121	119	117	117	112	136		883	863	895	889
15. *Cyanophora* cyanelle	237	125	123	124	96	121	119	126	112	115	96	110	102	128		883	878	881
16. *Marchantia* chloroplast	249	175	160	165	131	169	173	161	152	147	146	146	141	151	127		846	840
17. *Synechococcus* WH7805	226	15	86	25	127	145	125	130	118	124	117	121	118	113	133	172		986
18. *Synechococcus* WH8103	222	7	95	25	125	136	119	124	123	120	112	116	123	120	130	180	14	

Estimated evolutionary distances[29] (×1,000, below the diagonal) and fractional similarity values (×1,000, above) for pairs of 16S rRNA sequences for *Prochlorothrix hollandica*[8], other prochlorophytes (this work), cyanobacteria[22,30] (D. L. Distel and J. B. Waterbury, manuscript in preparation), photosynthetic organelles[22,31], a full-length, shotgun-cloned sequence from the Sargasso Sea[19], and *A. tumefaciens*[32], used in the construction of the phylogenetic tree in Fig. 1*a*. Sequences were aligned according to their conserved primary and secondary structures and compared by the computer program of Olsen[23], using 788 nucleotides of sequence unambiguously aligned for all organisms considered. Gap values were set at 0.5 nucleotide substitution. *Marchantia*, *Marchantia polymorpha*; *Cyanophora*, *Cyanophora paradoxa*. Genomic DNA from *Prochlorococcus* (isolated from the Sargasso Sea, 30 May 1988, 28° 58.9' N, 64° 21.5' W, at a depth of 120 m; ref. 14) was extracted using standard techniques[33] from a dilution culture inoculated with an average cell density of one cell per culture (isolate SSW5, 58% probability of clonality). *Prochloron* cells were collected in February 1990 from reef flats in the Kamori Channel, Koror, Palau, West Caroline Islands (7° 25' N, 134° 30' E). Collection and method of isolating the symbiont from the host, *Lissoclinum patella*, are described in ref. 34. DNA was extracted from frozen cells using standard protocols[35], except that the extraction buffer was adjusted to 0.25 M EDTA. 16S rRNA genes were amplified from genomic DNA of *Prochlorococcus* using the polymerase chain reaction (PCR) with the primers PLG1.1 (ACGGGTGAGTAACGCGTRA) and PLG2.1 (CTTATGCAGGCGAGTTGCAGC), designed to be specific for members of the oxygenic phototroph lineage and thus select against sequences from heterotrophic contaminants in the culture. *Prochloron* sequences were amplified using the primers P1 (AGAGTTTGATCCTGGCTCAG) and P2 (CTTGTTACGACTTCACCCC), which amplify 16S rRNA sequences from all eubacteria, as there was no significant contamination from heterotrophs in this preparation. PCR reactions contained 10 ng genomic DNA, 100 nM each primer, 200 μM each dNTP, 2.25 mM $MgCl_2$, 50 mM KCl, 10 mM Tris (pH 8.4), and 5 units *Taq* polymerase (Perkin Elmer Cetus). Reaction volumes were 200 μl, covered with 100 μl sterile mineral oil. Cycle parameters were: 94 °C for 1 min, T_d for 1 min, and 72 °C for 2 min. Samples were processed for 40 to 50 cycles. T_d was 66 °C for the PLG1.1/PLG2.1 primer pair, and 58 °C for the P1/P2 pair. PCR products were purified from preparative polyacrylamide gels by electro-elution (D-gel, EpiGene), and reamplified asymmetrically using 2 nM of one limiting primer. Asymmetric PCR products from both coding and complementary strands were ethanol-precipitated twice with ammonium acetate and used for dideoxy-nucleotide sequencing with Sequenase version 2 (US Biochemical Corporation) in parallel reactions containing dGTP and dITP, plus and minus single-strand binding protein. 1,147 bases of continuous sequence (*Escherichia coli* 16S rRNA nucleotide positions 128 to 1,312) were determined for *Prochlorococcus* and 1,377 bases (*E. coli* positions 28 to 1,452) for *Prochloron*. Sequences are available through GenBank, accession numbers X63140 and X63141.

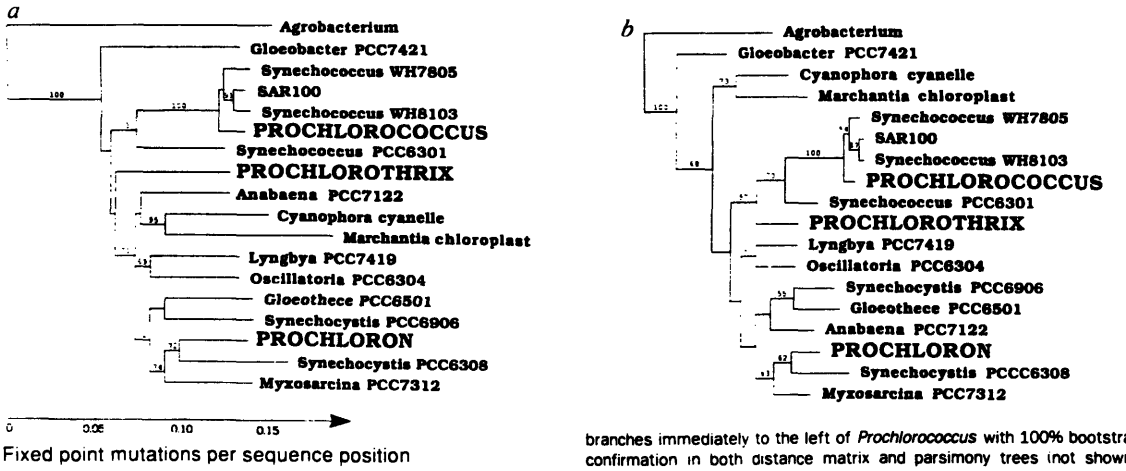


FIG. 1 Nucleic-acid sequence phylogenies, inferred from 16S rRNA data, illustrating the apparent independent appearance of chlorophyll *b*-containing organisms in the cyanobacterial lineage. The phylogenetic position of a full-length, shotgun-cloned 16S rRNA gene from Sargasso Sea plankton[19] (SAR100) is also shown. A sequence for the freshwater *Synechococcus* strain PCC6307 (D. L. Distel and J. B. Waterbury, manuscript in preparation) branches immediately to the left of *Prochlorococcus* with 100% bootstrap confirmation in both distance matrix and parsimony trees (not shown). Numbers indicate per cent confirmation by bootstrap analysis, summarizing the tree output from 100 resampled datasets, of the phylogenetic group to the right. Only values greater than 50% are labelled. *a*, Distance matrix analysis. Horizontal distances are proportional to evolutionary distances (Table 1). *b*, Parsimony analysis of the same data using the heuristic tree search option of PAUP[24]. An equally parsimonious tree reverses the relative positions of *Prochlorococcus* and *Synechococcus* WH7805.

TABLE 2 Evolutionary distance and fractional similarity matrix for 16S rRNA sequences from *A. tumefaciens*, members of the prokaryotic oxygenic phototroph radiation and shotgun clones from marine plankton communities

	1	2	3	4	5	6	7	8	9	10
1. *Agrobacterium tumefaciens*		782	808	821	797	812	805	799	812	805
2. SAR6	258		967	954	918	964	912	951	951	971
3. SAR7	222	33		987	925	977	919	964	984	971
4. SAR100	205	47	13		931	977	932	964	984	971
5. *Synechococcus* PCC6301	236	87	79	72		935	974	922	915	935
6. *Prochlorococcus marinus*	217	37	23	23	68		942	987	981	994
7. *Prochlorothrix hollandica*	226	94	86	72	27	61		929	922	942
8. *Synechococcus* WH7805	234	51	37	37	83	13	75		981	981
9. *Synechococcus* WH8103	217	51	16	16	90	20	82	20		974
10. ALO37	226	30	30	30	68	7	61	20	26	

Estimated evolutionary distances (×1,000, below the diagonal) and fractional similarity values (×1,000, above) for pairs of 16S rRNA sequences for *Prochlorothrix*[8], other prochlorophytes (this work), cyanobacteria[3C] (D. L. Distel and J. B. Waterbury, manuscript in preparation), shotgun-cloned 16S rRNA genes[17-19], and *Agrobacterium*[32], used in the construction of the phylogenetic tree in Fig. 2. The 186 bases of aligned sequence corresponding to position 306 to 514 in the *E. coli* 16S rRNA sequence were analysed as described in the legend to Table 1.

We compared 16S rRNA gene sequences for both *Prochlorococcus* and *Prochloron* with sequences from the oxygenic phototroph database[8,22] (D. L. Distel and J. B. Waterbury, manuscript in preparation) using both distance matrix[23] (Fig. 1*a*; Table 1) and parsimony[24] (Fig. 1*b*) algorithms. Bootstrap resampling[25] was used to estimate reliability of the inferred trees (Fig. 1*a*, *b*). For simplicity we refer only to the results of the distance matrix analysis in our discussion, although results from parsimony analysis are in essential agreement.

The 16S rRNA tree strongly supports the derivation of both *Prochlorococcus* and green chloroplasts (represented in our analysis by the *Marchantia polymorpha* chloroplast) from different, phycobilisome-containing ancestors among the cyanobacteria (Fig. 1*a*). *Prochlorococcus* forms a shallowly branching cluster with marine A *Synechococcus* strains WH7805 and WH8103 (ref. 16) (confirmed in 100% of bootstrap resamplings), before which branch the cyanobacteria *Synechococcus* PCC6307 (D. L. Distel and J. B. Waterbury, manuscript in preparation) (not shown; confirmed in 100% of bootstrap resamplings) and *Synechococcus* PCC6301 (confirmed in 73% of bootstrap resamplings). The *Marchantia* chloroplast branches with the *Cyanophora paradoxa* cyanelle (95% bootstrap confirmation), in accord with evidence that green chloroplasts and the cyanobacterium-like cyanelle descend from a common ancestor[22,26]. Branching patterns for *Prochloron* and *Prochlorothrix* are less than 95% confirmed, thus their descent from separate, phycobilisome-containing ancestors, though likely, is less certain than for *Prochlorococcus* and the chloroplasts.

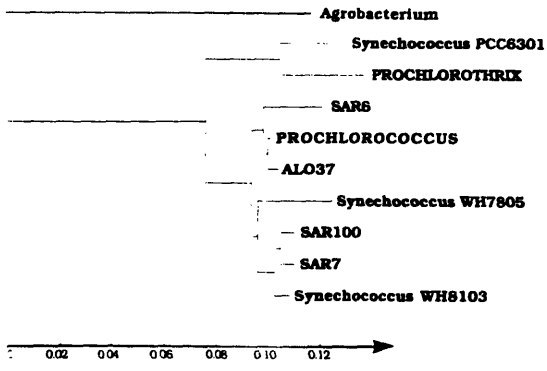


FIG. 2 Phylogenetic tree illustrating relationships among prochlorophytes, members of the *Synechococcus* group and shotgun-cloned sequences from Sargasso Sea plankton[17,19] (SAR) and Pacific Ocean picoplankton[18] (ALO). The tree was constructed using distance matrix analysis and 16S rRNA sequence data (Table 2). Differences in the branching order between this tree and the one shown in Fig. 1*a* are presumably due to the short length of the shotgun-cloned sequences, which limits the resolution of this analysis. Bootstrap analysis indicates that *Prochlorococcus*, *Synechococcus* strains WH8101, WH7805, and the cloned sequences form a coherent phylogenetic cluster distinct from *Prochlorothrix* and *Synechococcus* PCC6301, but the branching order in this cluster is indeterminate.

Prochlorococcus is closely related not only to the open-ocean strains, *Synechococcus* WH7805 and WH8103, but also to a diversity of shotgun-cloned 16S rRNA gene sequences from the Sargasso Sea[17,19] and the north Pacific Ocean[18] (Figs 1 and 2; Tables 1 and 2). Individual samplings from these sites, where both *Prochlorococcus* and *Synechococcus* are common, have routinely netted a puzzling multiplicity of sequences, exhibiting greater than 95% similarity, interpreted as belonging to open-ocean *Synechococcus*[17-19]. This diversity can now be partially attributed to sympatric populations of *Synechococcus* and *Prochlorococcus*, which, despite their sequence similarity, are separate species by ecological as well as phenotypic criteria[27].

We conclude that the phycobilisome$^-$/chlorophyll b^+ (green chloroplast) phenotype seems to have evolved *de novo* at least three, and possibly four times during the evolution of cyanobacteria: once in the ancestry of the green chloroplast lineage, again in that of *Prochlorococcus*, and once or twice in the ancestries of *Prochloron* and *Prochlorothrix*, the mutual independence of which has not positively been demonstrated. The shallowness of the *Prochlorococcus* cluster suggests a particularly recent origin for this organism's pigment phenotype. In the light of these results, and those of Palenik and Haselkorn[28], the order Prochlorales can no longer be justified as a natural grouping. The taxon should therefore be abandoned, and prochlorophytes reclassified in the cyanobacteria. □

Received 27 August; accepted 18 October 1991.

1. Cavalier-Smith, T. *Biol. J. Linn. Soc. Lond.* **17**, 289-306 (1982).
2. Whitton, B. A. & Carr, N. G. in *The Biology of the Cyanobacteria* (eds Carr, N. G. & Whitton, B. A.) 1-8 (University of California, Berkely, 1982).
3. Lewin, R. A. & Cheng, L. in *Prochloron a Microbial Enigma* (eds Lewin, R. A. & Cheng, L.) 1-7 (Chapman and Hall, New York, 1989).
4. Van Valen, L. M. & Maiorana, V. C. *Nature* **287**, 248-250 (1980).
5. Lewin, R. A. *Ann. N. Y. Acad. Sci.* **361**, 325-329 (1981).
6. Margulis, L. *Symbiosis in Cell Evolution* (Freeman, San Francisco, 1981).
7. Burger-Wiersma, T. *et al. Int. J. Syst. Bacteriol.* **39**, 250-257 (1989).
8. Turner, S. *et al. Nature* **337**, 380-382 (1989).
9. Morden, C. W. & Golden, S. S. *Nature* **337**, 382-385 (1989).
10. Morden, C. W. & Golden, S. S. *Nature* **339**, 400 (1989).
11. Morden, C. W. & Golden, S. S. *J. molec. Evol.* **32**, 379-395 (1991).
12. Kishino, H., Miyata, T. & Hasegawa, M. *J. molec. Evol.* **31**, 151-160 (1990).
13. Chisholm, S. W. *et al. Nature* **334**, 340-343 (1988).
14. Chisholm, S. W. *et al. Archiv. Microbiol.* (in the press).
15. Lewin, R. A. & Withers, N. W. *Nature* **256**, 735-737 (1975).
16. Waterbury, J. B. & Rippka, R. in *Bergey's Manual of Systematic Bacteriology* Vol. 3 (eds Stanley, J. T. *et al.*) 1728-1746 (Williams and Wilkins, Baltimore, 1989).
17. Giovannoni, S. J. *et al. Nature* **345**, 60-62 (1990).
18. Schmidt, T. M., DeLong, E. F. & Pace, N. R. *J. Bact.* **173**, 4371-4378 (1991).
19. Britschgi, T. B. & Giovannoni, S. J. *Appl. envir. Microbiol.* **57**, 1707-1713 (1991).
20. Lewin, R. A. *Phycologia* **1**, 217 (1977).

21. Goericke, R. & Repeta, D. *Limnol. Oceanogr.* (in the press).
22. Giovannoni, S. J. *et al. J. Bact.* **170,** 3584–3592 (1988).
23. Olsen, G. J. *Meth. Enzym.* **164,** 793–838 (1988).
24. Swofford, D. L. *Paup* Version 3.0 (Illinois Natural History Survey, Champaign, Illinois, 1989).
25. Felsenstein, J. *Evolution* **39,** 783–791 (1985).
26. Douglas, S. E. & Turner, S. *J. molec. Evol* **33,** 267–273 (1991).
27. Olson, R. J., Chisholm, S. W., Zettler, E. R., Altabet, M. A. & Dusenberry, J. A. *Deep Sea Res.* **37,** 1033–1051 (1990).
28. Palenik, B. & Haselkorn, R. *Nature* **355,** 265–267 (1992).
29. Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (ed. Munro, H. N.) 21 (Academic, New York, 1969).
30. Tomioka, N. & Sugiura, M. *Molec. gen. Genet.* **191,** 45–50 (1983).
31. Ohyama, K. *et al. Pl. molec. Biol. Reporter* **4,** 148–175 (1986).
32. Yang, D. *et al. Proc. natn. Acad. Sci. U.S.A.* **82,** 4443–4447 (1985).
33. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: a Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989).
34. Swift, H. & Robertson, D. L. *Symbiosis* **10,** 95–113 (1991).
35. Dzelzkalns, V. A., Szekeres, M. & Mulligan, B. J. in *Plant Molecular Biology* (ed. Shaw, C. H.) 277–299 (IRL, Washington DC, 1988).

ACKNOWLEDGEMENTS. We thank M. L. Sogin for assistance with phylogenetic analyses, D. L. Distel and J. B. Waterbury for unpublished sequences, and W. Thilly and P. Keohavong for technical support. This work was supported by grants from the NSF, EPA and Office of Naval Research. D.L.R. was supported by a training grant from the NIH.

Chapter Two

PHYLOGENETIC RELATIONSHIPS AMONG CULTURED STRAINS OF *PROCHLOROCOCCUS MARINUS* ISOLATED FROM DIVERSE OCEANIC PROVINCES

ABSTRACT

Based on relationships inferred from 16S rRNA, *psbB* and *pet B* and *D* sequences, cultured strains of *Prochlorococcus marinus* were found to belong to a single lineage within the cyanobacteria, which they share with strains of marine A *Synechococcus*. Branching patterns inferred for *P. marinus* strains using 16S rRNA, *psbB* and *pet B* and *D* sequences were consistent, and revealed no correlation between genetic distance and the geographic distance between sites of culture isolation, with the most closely related strains originating in surface waters of the Mediterranean and the Pacific Oceans. The branching order of deeply diverging lineages within the *P. marinus*/marine A *Synechococcus* cluster, including one strain of *P. marinus* and two strains of marine A *Synechococcus*, appear inconsistent in 16S rRNA gene trees inferred by neighbor-joining, parsimony and maximum likelihood methods of analysis. This result suggests a near-simultaneous radiation of *P. marinus* and marine A *Synechococcus* lineages.

INTRODUCTION

Prochlorococcus marinus is a tiny, photosynthetic marine prokaryote containing divinyl chlorophylls *a* and *b* (chl a_2 and chl b_2) and which recently has been recognized as a major constitutent of the photosynthetic picoplankton in wide regions of the Atlantic, Pacific, Mediterranean and Red Seas and in the Banda Sea of Indonesia (Chisholm et al. 1988, Gieskes et al. 1988, Li and Wood 1988, Olson et al. 1990, Vaulot et al. 1990, Li et al. 1992, Vaulot and Partensky 1992, Cambell and Vaulot 1993, Goericke and Welschmeyer 1993, Goericke and Repeta 1993, Lantoine and Neveux 1993, Veldhuis and Kraay 1993, Vaulot et al. 1994). *P. marinus* comprises 35% of the seasonally averaged chl *a* at north Pacific station ALOHA (Letelier et al. 1993) and 30% at Sargasso Sea station OFP (Goericke and Welschmeyer 1993). The ecological importance of this

ubiquitous and unusual organism has sparked efforts to characterize its physiological capabilities in pure culture (Partensky et al. 1993, Morel et al. 1993, Moore et al. 1994).

Phylogenetic analysis using 16S ribosomal RNA gene sequences has established that a strain of *P. marinus* isolated from the Sargasso Sea (strain SSW5, descended from SARG, the original *P. marinus* isolate) is closely related to the marine A *Synechococcus* (Waterbury and Rippka 1989), a group of cyanobacterial strains including isolates from open ocean habitats in which *P. marinus* also is found (Urbach et al. 1992). Cyanobacterial 16S rRNA gene sequences cloned from natural populations in the Sargasso Sea and north Pacific (Giovannoni et al. 1990, Britschgi and Giovannoni 1991, Schmidt et al. 1991) also fall into this cluster (Urbach et al. 1992). An independent study using *rpoC* gene sequences found that the genetic distance between SARG and a second *P. marinus* isolate originating in the Mediterranean Sea (MED), was greater than the distance between two heterocyst-forming cyanobacteria assigned to different genera (Palenik and Haselkorn 1992), suggesting that differences between *P. marinus* isolates may be significant. *rpoC* analyses also cluster *P. marinus* with *Synechococcus* WH8103 (Swift and Palenik 1992). The 16S rRNA and *rpoC* studies each reported, in addition, that *P. marinus* is unrelated to the "normal" chlorophyll *a* and *b* (chl a_1 and b_1) - containing prokaryotes *Prochloron sp.* and *Prochlorothrix hollandica*, and to green plant and algal chloroplasts. These conclusions have been cast into some doubt, however, by the discovery that base substitution biases at rapidly evolving nucleotide sites are correlated with branching patterns in these studies' phylogenetic trees (Lockhart and Penny 1992).

To help resolve these controversial relationships and examine isolates of *P. marinus* from additional oceanographic provinces, we present here the results of an investigation into the phylogenetic relationships among cultured isolates of *P. marinus,*

addressed by comparisons of three gene sequences and including marine A *Synechococcus* and cloned 16S rRNA gene sequences from the Sargasso Sea. Questions to be addressed are (1) whether oceanic prochlorophytes represent more than one independent lineage of chlorophyll b_2-containing prokaryotes dispersed among the cyanobacteria, (2) whether the inferred phylogenetic patterns for each of the three genes are consistent, (3) whether a different phylogenetic pattern is inferred by methods suggested as insensitive to nucleotide substitution bias, and (4) whether evolutionary relationships reflect geographic relationships among sites of *P. marinus* culture isolation. Detailed knowledge of the evolutionary relationships among *P. marinus* cultured strains is, in addition, useful for organizing information on the physiological differences among these strains. Also, if gene transfer does not play a major role in the evolution of *P. marinus,* then knowledge of the phylogenetic relationships will provide a link between phenotypes of cultured cells and sequences recovered by molecular cloning experiments exploring genetic diversity in field populations (Giovannoni et al. 1990, Schmidt et al. 1991, Britschgi and Giovannoni 1991, DeLong et al 1993, Fuhrman et al. 1993, this thesis Chapter Three).

BACKGROUND

Characteristics of *P. marinus* strains. Cultures have been established for *P. marinus* isolated from a variety of geographic and hydrographic regimes (Chisholm et al. 1992, Partensky et al. 1993, Moore et al. 1994, L. Moore, pers. comm.). For this study we selected cultures drawn from widely separated locales within *P. marinus'* geographic range: the Sargasso Sea, Mediterranean Ocean, north Atlantic and south Pacific Oceans. The *P. marinus* type strain, Sargasso Sea clone SS120 (designated CCMP-1375 at the Center for the Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences, West Boothbay Harbort, ME) originated in the deep euphotic zone during spring

stratification and is descended from SARG, the original *P. marinus* isolate (Chisholm et al. 1992). The Sargasso Sea isolate used for 16S rRNA analysis, SSW5, was isolated from SARG by dilution to a mean cell density of one cell per culture tube, and is assumed to be identical to SS120[1]. Mediterranean clone Med4 (CCMP1378) originated in mixed surface waters of the Mediterranean in winter and is descended from strain MED (Chisholm et al 1992, Partensky et al. 1993). North Atlantic isolate FP5 is an uncloned culture collected from 30 m in the North Atlantic by F. Partensky, and Pacific strain MIT9107, also not cloned, came from 25m in the mixed surface layer of the oligotrophic south Pacific, and was isolated by J. Dusenberry and M. DuRand. In addition to these strains, Sargasso Sea culture MIT9303, isolated from 100 m in the Sargasso Sea during summer stratification by L. Moore, was included in some of the analyses. *Synechococcus* clone WH8103, a high phycourobilin (PUB) strain isolated from the surface of the Sargasso Sea, and the reference culture for marine A *Synechococcus* (Waterbury and Rippka 1989), was selected to represent marine A *Synechococcus* (Table 1).

Physiological studies have characterized *P. marinus* clones SS120 and Med4 (or their parent cultures) according to pigment content and respose to variation in temperature and illumination (Partensky et al. 1993, Morel et al. 1993, Moore et al. 1994). These studies confirm the presence in both strains of divinyl chlorophylls *a* and *b* (chl a_2 and b_2) as well as accessory pigments zeaxanthin, α-carotene and an unknown pigment which elutes with chlorophyll c_1 by HPLC (Chisholm et al. 1988, Chisholm et al. 1992, Goericke and Repeta 1992). In addition, it is now known that SS120 (and

[1]In light of the results of Chapter Three, which indicate that *P. marinus* field populations are genetically heterogeneous, the possibility exists that the primary isolate SARG could have contained more than one genetic variant. Since SSW5 and SS120 were independently isolated from SARG, it therefore is possible that they are not genetically identical. Omitting the SSW5 sequence from the dataset does not change the conclusions of this study.

Table 1. *Prochlorococcus marinus* strains used in this study.

Strain	Isolation history					Culture history	Chl b/a_2 Ratio	Ref.
	region	site	depth	date	researcher			
SS120 (CCMP1375)	Sargasso	28°59'N, 64°21'W	120m	5/30/88	B. Palenik	cloned from SARG	0.4 - 2.4	a
SSW5	Sargasso	28°59'N, 64°21'W	120m	5/30/88	B. Palenik	isolated from SARG by dilution		b
Med4 (CCMP1378)	Mediterranean	43°12'N 61°52'E	5m	1/89	D. Vaulot/ F. Partensky	cloned from MED	0.05 - 0.15	a, c
FP5	north Atlantic	38°59'N 49°33'W	30m	4/90	F. Partensky	primary culture		
MIT9107	south Pacific	14°60'S 134°60'W	25m	8/8/91	J. Dusenbery/ M. DuRand	primary culture		
MIT9303	Sargasso	34°45'N 66°11'W	100m	7/15/93	L. Moore	primary culture		

[a]Moore et al. 1994.
[b]Urbach et al. 1992
[c]Chisholm et al. 1992.

SARG) grown at high light (>20 μE m^{-2} s^{-1}) contains chl b_1, which may comprise up to 55% of total chlorophyll *b* (chl *b*, equal to chl b_1 + chl b_2) (Partensky et al. 1993, Morel et al. 1993, Moore et al. 1994). In contrast, clone Med4 (and MED) contains no detectable chl b_1 under any growth conditions tested. Both strains are capable of photoadaptation by changing the ratio of their chl *b* to chl a_2, but chl b/a_2 ratios for Med 4 are one tenth those of SS120 at all growth irradiances (Partensky et al. 1993, Morel et al. 1993, Moore et al. 1994).

It has been hypothesized that the different chl b/a_2 ratios for SS120 and Med4 are genetic adaptations to the low and high light environments from which the two clones were collected, respectively. The low light SS120 clone is capable of acclimating to conditions at the bottom of the euphotic zone in oligotrophic waters by elaborating extensive, chl *b*-containing photosynthetic antennae capable of absorbing blue light, while the high light-adapted Med4 fails to elaborate an extensive antenna (Partensky et al. 1993, Moore at al. 1994). This hypothesis is consistent with the results of growth experiments which have shown that SS120 is capable of growth at low light intensities (2 to 6 μE m^{-2} s^{-1}) at which no growth was observed for Med4, and that Med4 is capable of growth at high light intensities (>100 μE m^{-2} s^{-1}) at which SS120 fails to grow. The compensation light intensity (I_{comp}), which predicts the minimum illumination for cell survival, was significantly lower for SS120 than for Med4 (Moore et al. 1994), again consistent with the hypothesis. Other cultures included in the phylogenetic study are much less well characterized than SS120 and Med4.

Genetic loci. For this study, evolutionary relationships among the four *P. marinus* cultures SS120, Med4, FP5 and MIT9107 and *Synechococcus* WH8103 were examined using three sequences exhibiting varying degrees of evolutionary conservation: the 16S ribosomal RNA genes, *psbB*, and *petB* and *D*. Strain MIT9303 was evaluated

using only 16S rRNA, as *psbB* and *petB/D* sequences were unavailable for this culture due to time limitations. Future work after the completion of this thesis may include *psbB* and *petB/D* sequences for this strain.

16S ribosomal RNA (rRNA) genes are the standard tool for phylogenetic reconstructions (Hillis and Dixon 1991, Olsen and Woese 1993) and have provided phylogenetic information for many diverse taxa (e.g. Woese 1987, Sogin et al. 1989, Medlin et al. 1993). 16S rRNA analyses have the advantage of compatibility with a large and well-characterized database, containing sequences from hundreds of organisms and organelles, and a dedicated computerized clearinghouse which provides automated sequence alignment and similarity searching (Larsen et al. 1993). Several cloning experiments exploring the diversity of microbial communities in the sea have provided 16S rRNA gene sequences, some of which are likely to derive from wild *P. marinus* (Giovannoni et al. 1990, Schmidt et al. 1991, Britschgi and Giovannoni 1991, DeLong et al. 1993, Fuhrman et al., 1993).

psbB and *petB* and *D* encode proteins of the photosynthetic apparatus and are much less constrained evolutionarily than 16S rRNA. These loci were selected to infer phylogenies sensitive to evolutionary differences on a finer scale. Each has the additional advantage of being a single-copy sequence unique to photosynthetic organisms (Vermaas and Ikeuchi 1991), and so is easily amplified from *P. marinus* cultures, which contain heterotrophic bacteria. *psbB* encodes the chlorophyll *a*-binding antenna protein CP47 and *petB* and *D* encode subunits of the photosynthetic b_6/f complex responsible for transferring electrons between Photosystem II and Photosystem I (Widger and Cramer 1991). *petB* and *D* are neighboring cistrons with a consistent tandem orientation in oxygenic photosynthetic prokaryotes and organelles (Vermaas and Ikeuchi 1991, Greer and Golden 1992) and provide PCR priming sites which permit amplification of a

fragment containing the 3' end of *petB* and the 5'end of *petD*, plus the highly variable intergenic region between them (the amplified locus will be referred to as "*petB/D*"). The *psbB* and *petB/D* loci are not close to each other on cyanobacterial chromosomes (Vermaas and Ikeuchi 1991).

Problems posed for molecular phylogenetic analysis by nucleotide substitution biases. An unusal feature of cyanobacterial and chloroplast sequences is the wide variation in their DNA base compositions (Herdman et al. 1979, Lockhart et al. 1992a), which is especially evident at rapidly evolving sites such as noncoding regions and silent third codon positions (Lockhart and Penny 1992, Lockhart et al. 1992a, b). An additional property of the cyanobacterial phylogeny is the presence of many deeply branching lineages which apparently originated within a short period of evolutionary history (Giovannoni et al. 1988). The juxtaposition of these two unusual features creates a situation in which the inferred branching pattern is susceptible to the artefactual clustering of sequences converging towards similar G+C content (Lockhart et al 1992a, b, c, Lockhart and Penny 1992, Lockhart et al. 1993), and there is as yet no accepted method for assessing the relative contributions of nucleotide substitution bias "noise" and true phylogenetic "signal" to trees inferred from such data (Lockhart et al. 1993). Bootstrap resampling analyses, which are frequently used to assess the sensitivity of phylogenetic inferences to sampling error in the nucleotide positions chosen for the analysis (Felsenstein 1988), cannot detect systematic distortions in branching patterns due to nucleotide substitution biases (Lockhart et al. 1992a).

Lockhart et al. (1992a, b) used G+C content at third codon positions to characterize nucleotide substitution biases in the evolution of different taxa, noting that branching patterns inferred from protein-encoding sequences grouped organisms having similar nucleotide substitution biases. They suggest that in instances when nucleotide

substitution biases differ among lineages, standard phylogenetic methods will infer incorrect phylogenetic relationships. Lockhart et al. (1992a) further assert that the influence of nucleotide substitution bias detected at third codon positions in protein encoding genes may determine branching patterns inferred for the same organisms using 16S rRNA genes, even though the 16S rRNA sequences may exhibit little variation in G+C content (c.f. Table 3).

Several molecular phylogenetic methods have been suggested as being relatively immune to nucleotide substitution biases, although none are yet generally accepted as such (Sogin et al. 1989, 1993). Woese et al. (1991) suggested that distance and parsimony methods applied to transversion mismatches would have generally reduced sensitivity, and so would be less susceptible to these artefacts. Hasegawa and Hashimoto (1993) suggested that phylgenetic analysis of amino acid sequences would have this property as well (see also Hashimoto et al. 1994), and amino acid parsimony has been applied to plastid and cyanobacterial data (Morden et al. 1992). Most recently, two similar, new algorithms have been devised which are reported to be insensitive to variation in G+C bias among lineages, LogDet (Lockhart et al. 1994) and paralinear distances (Lake 1994). However, computer programs which impliment these algorithms are not yet publicly available.

For our analysis of *P. marinus* phylogenetic relationships we have employed standard methods of phylogenetic inference: neighbor-joining, parsimony and maximum likelihood. In addition, we explored branching patterns inferred by transversion distance analysis and amino acid parsimony to assess whether these methods, suggested as relatively insensitive to nucleotide substitution biases, infer a phylogeneic pattern consistently different from those inferred by standard analyses.

METHODS

Cell culture and DNA isolation. *P. marinus* clones and isolates were grown in modified K/10 - Cu medium as previously reported (Chisholm et al. 1992). DNA was prepared from dense cultures as described (Urbach et al. 1992), or by a modification of the alkaline lysis protocol of Li et al. (1991). For this protocol, 1 ml of culture containing 10^6 to 10^8 *P. marinus* was concentrated by spin filtration at 2,000 x g using 0.2 μm pore Ultrafree-MC filter units (Millipore) and washed twice with cell suspension buffer (0.5 M NaCl, 10 mM Tris [pH 8.0] 10 mM EDTA). Cells were resuspended in a final volume of 100 μl cell suspension buffer and lysed by addition of 12 μl 0.5 M DTT and 6 μl 10 M NaOH followed by a 10 minute incubation at 65°C. After addition of 120 μl neutralization buffer (90 mM Tris [pH 8.0] plus 0.5 mol/l HCl), DNA was precipitated by addition of 1 μl 20 mg/ml glycogen (Boehringer) and 0.6 ml cold 100% ethanol and storage overnight at -20°C. Precipitates were collected by centrifugation at 16,000 x g for 30 min at 4°C, washed with 70% ethanol and resuspended in TE (10 mM Tris [pH 8.0] 1 mM EDTA). DNA was stored at -20°C.

Synechococcus WH8103 was grown in S/N Medium according to Waterbury et al. (1986). DNA was prepared by CsCl density gradient centrifugation (Sambrook et al. 1989).

PCR and DNA sequencing. 16S rRNA genes from *P. marinus* Med4, FP5, MIT9107 and MIT9303 were amplified using oxygenic phototroph-specific primers and protocols as described (Urbach et al. 1992), except that one primer in each PCR reaction was labelled with biotin (Hultman et al. 1989). *psbB* and *petB/D* were amplified using biotinlyated/unbiotinylated primer pairs containing inosine (Knoth et al. 1988) under the

following conditions: *psbB* primers PPSBB1353 (5'GTIGCIGGIACIATGTGGTA) and PPSBB1928R (5'GCRTGICCRAAIGTRAACCA), 2.25 mM $MgCl_2$, 2 min. 94°C, followed by 1 min 94°C, 1 min 51°C, 1 min 72°C (5 cycles), 1 min 94°C, 1 min 56°C, 1 min 72°C (10 cycles), 1 min 94°C, 1 min 62°C, 1 min 72°C (25 cycles), followed by 10 min 72°C; *petB/D* primers PPETBD314 (5'PPETBD314 (5'ATGATGGTIYTIATGATGAT) and PPPETBD1160R (5'CCRTARTARTTRTGICCCAT), 1.5 mM $MgCl_2$, 2 min 94°C followed by 1 min 94°C, 1 min 45°C, 1 min 72°C (5 cycles), 1 min 94°C, 1 min 50°C, 1 min 72°C (10 cycles), 1 min 94°C, 1 min 57°C, 1 min 72°C (25 cycles), followed by 10 min 72°C. 100 µl PCR reactions contained $MgCl_2$ concentrations as indicated, plus 1x Mg-free Taq polymerase buffer (Promega), 200 µM each dNTP, 100 nM each primer and 5 units Taq polymerase (Promega). PCR products from three or more replicate reactions were pooled and separated by electrophoresis in 1% agarose gels. Bands were purified using GeneClean (Bio 101) and single stranded sequencing templates prepared using Dynabeads (Dynal) according to the manufacturers instructions.

DNA sequences were determined using Sequenase (United States Biochemical), according to the manufacturer's protocol. 16S rRNA gene sequences were determined from both strands using amplification primers and internal primers as described (Urbach et al. 1992). 90.5% of 16S rRNA gene sequences were determined on both strands. *psbB* and *petB/D* sequences were determined using amplification primers and internal primers PPSBB1527 (5'ITTYTAYGAYTAYGTIGG), PPSBB 1508R (5'CCIAGRTARTGRTARAAIGC), PPSBB1898R (5'CGAAAIACICCRTC), and PPETBD532R (5'CCIACRCTYTCICCICC). *psbB* sequences were redundantly determined over 92% of their length, with 60% of nucleotides determined on both strands. *petB/D* sequences were determined from one strand only, as the biotinylated

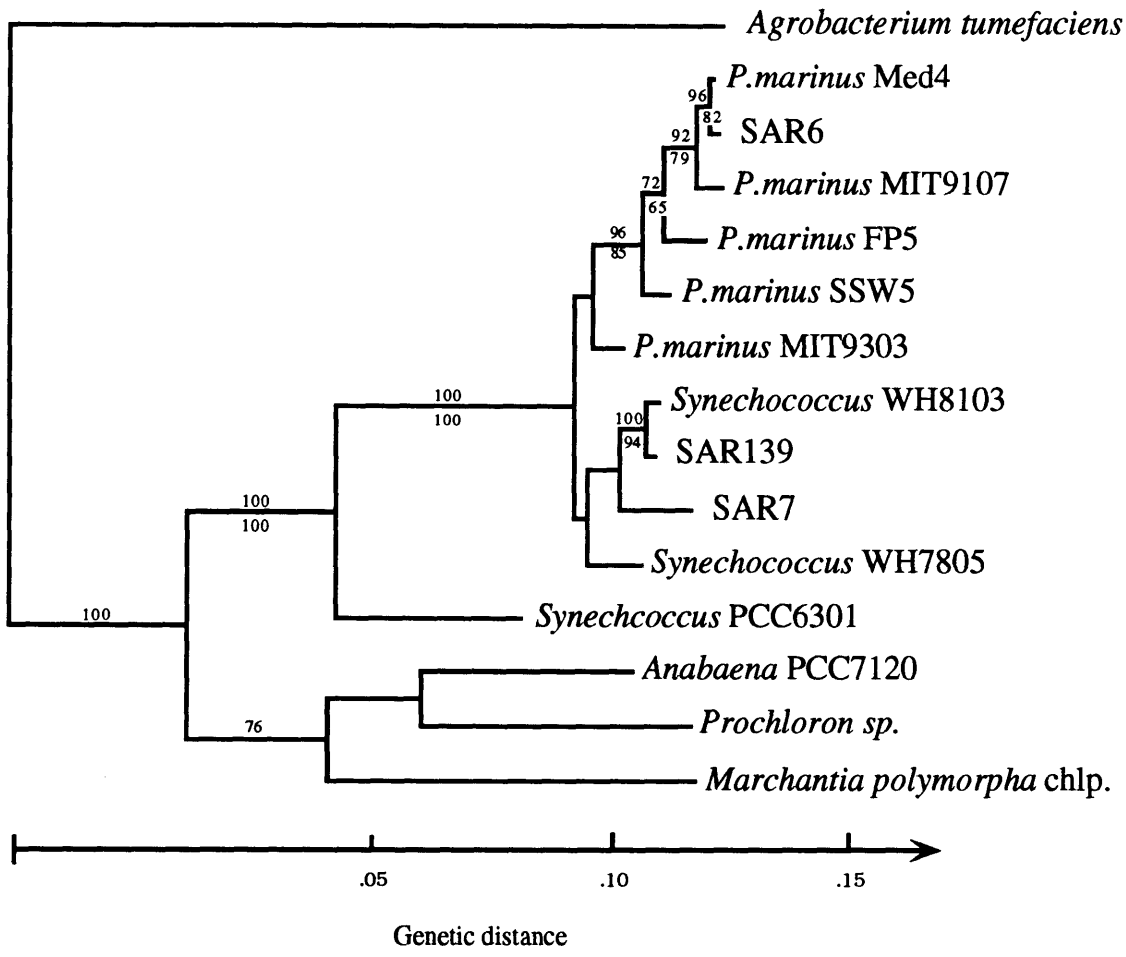
version of PPETBD1160R did not function well as a PCR primer. 94% of *petB/D* sequences were redundantly determined.

Sequence alignment and phylogenetic analysis. Sequences were aligned according to conserved regions of 16S rRNA primary and secondary structure or according to conserved regions in amino acid translations (Lang and Haselkorn 1989, Widger and Cramer 1991) using the Olsen sequence alignment editor (Olsen 1990). Mismatch frequency and base compositions were calculated using the program of Olsen (1988, 1990). Neighbor-joining (Saitou and Nei 1987) and protein parsimony were performed using Kimura two-parameter genetic distance estimates (2:1 transition transversion ratio) and the DNADIST, NEIGHBOR and PROTPARS programs in PHYLIP verision 3.4 (Felsenstein 1991). DNA parsimony was performed with PAUP's branch and bound and MULPARS options (Swofford 1991). DNA maximum likelihood calculations were done using fastDNAml (Olsen et al. 1992), and likelihood comparisons between different phylogenetic trees were performed using PHYLIP's DNAML program. Log likelihood values and standard deviations are from DNAML comparisons.

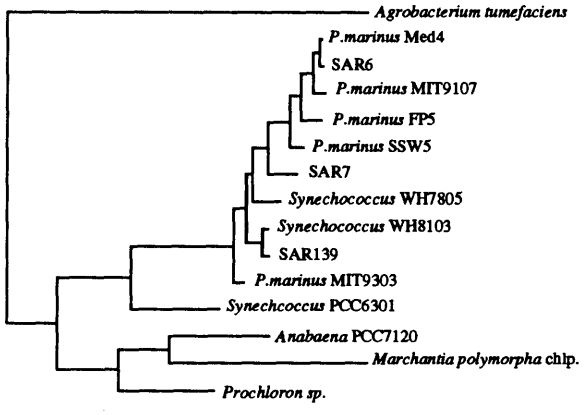
RESULTS

Relationships inferred by standard methods: molecular phylogenetic branching patterns and bootstrap calculations. Relationships among cultured strains of *P. marinus*, marine A *Synechococcus* and other members of the oxygenic photosynthetic radiation were investigated using three standard methods for phylogenetic analysis: neighbor-joining (Saitou and Nei 1987), parsimony, and maximum likelihood (Felsenstein 1988). All strains of *P. marinus* and marine A *Synechococcus* fell into a single lineage in trees inferred by all analyses in this study (Figures 1, 2, 3). This

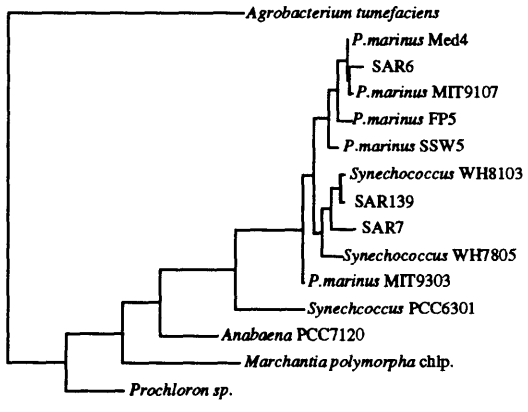
Figure 1. Phylogenetic relationships among cultured strains of *P. marinus*, shotgun cloned sequences from Sargasso Sea picoplankton ("SAR" sequences) and other members of the cyanobacterial lineage, inferred from 16S rRNA gene sequences using *Agrobacterium tumefaciens* as an outgroup. (a), Distance matrix tree inferred by neighbor-joining with Kimura two parameter genetic distance estimates (Kimura 1980) (Table 2a). Numbers indicate the percent of trees inferred from 100 bootstrap datasets which contained the phylogenetic group to the right, with numbers above the lines from neighbor-joining analysis and below from parsimony. Bootstrap values below 50% are omitted. (b), The most parsimonious tree (668 steps) inferred by parsimony. (c), Maximum likelihood tree (ln L = -4767.84277), not significantly better than the tree inferred by neighbor-joining (ln L = -4800.28223, S.D. 19.0879, Δln L = -32.43945). Transversion analysis inferred a branching pattern identical to that in the maximum likelihood tree.


Agrobacterium tumefaciens
P.marinus Med4
SAR6
P.marinus MIT9107
P.marinus FP5
P.marinus SSW5
P.marinus MIT9303
Synechococcus WH8103
SAR139
SAR7
Synechococcus WH7805
Synechcoccus PCC6301
Anabaena PCC7120
Prochloron sp.
Marchantia polymorpha chlp.
96
82
92
79
72
65
96
85
100
100
100
94
100
100
100
76
.05
.10
.15
Genetic distance

a

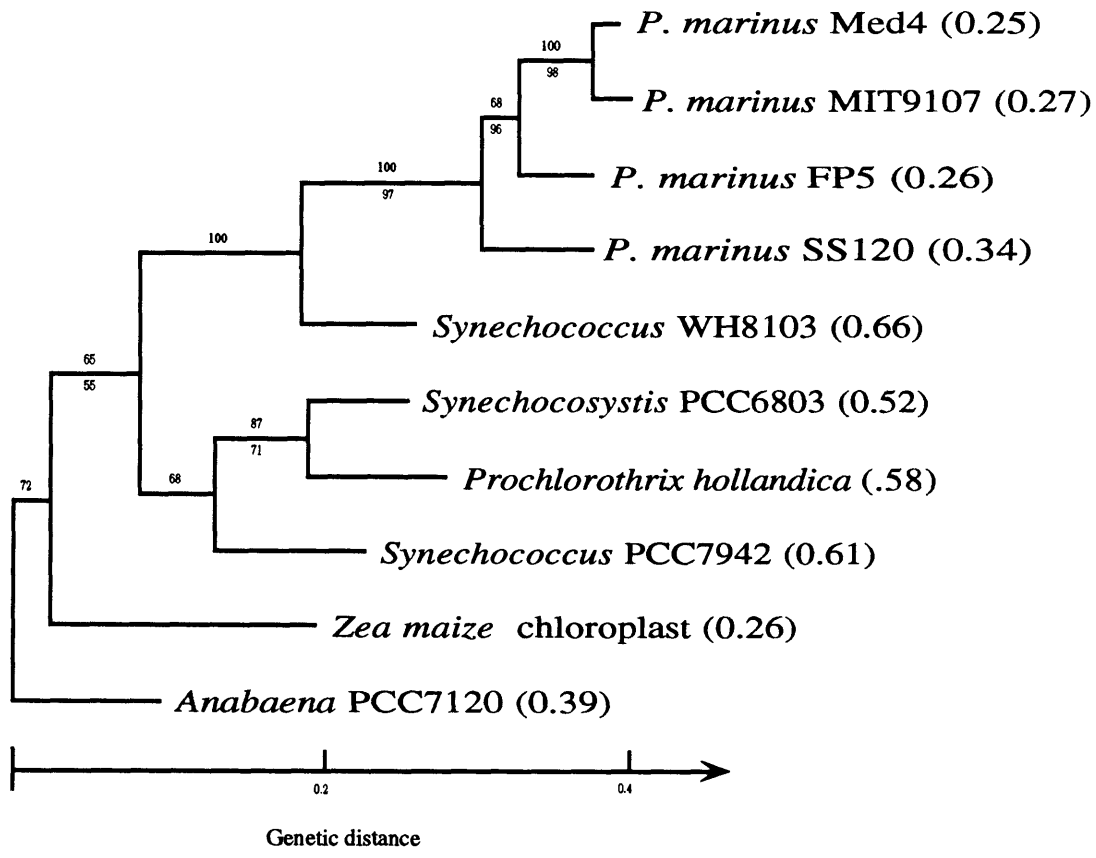

Agrobacterium tumefaciens
P.marinus Med4
SAR6
P.marinus MIT9107
P.marinus FP5
P.marinus SSW5
SAR7
Synechococcus WH7805
Synechococcus WH8103
SAR139
P.marinus MIT9303
Synechcoccus PCC6301
Anabaena PCC7120
Marchantia polymorpha chlp.
Prochloron sp.

b

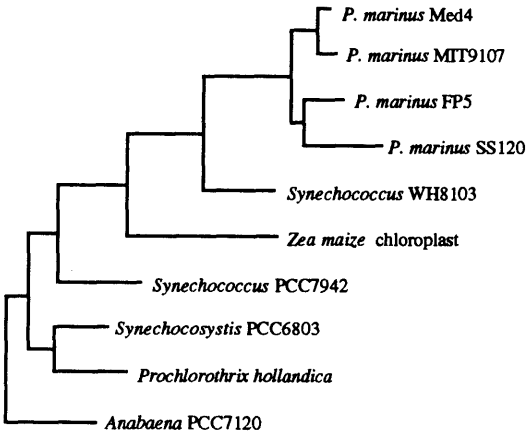

Agrobacterium tumefaciens
P.marinus Med4
SAR6
P.marinus MIT9107
P.marinus FP5
P.marinus SSW5
Synechococcus WH8103
SAR139
SAR7
Synechococcus WH7805
P.marinus MIT9303
Synechcoccus PCC6301
Anabaena PCC7120
Marchantia polymorpha chlp.
Prochloron sp.

c

Figure 2. Phylogenetic relationships among cultured strains of *P. marinus* and other members of the cyanobacterial lineage inferred using *psbB* sequences. Trees are arbitrarily rooted to *Anabaena* PCC7120 to facilitate comparison to those inferred from 16S rRNA gene sequences (Figure 1). (a), Distance matrix tree inferred by neighbor-joining with Kimura two parameter genetic distance estimates (Table 2b). Numbers in parentheses indicate G+C base composition at third codon positions; other numbers are as in Figure 1a. Identical branching patterns were inferred by parsimony, maximum likelihood (ln L = -1925.90128) and transversion neighbor-joining analyses, with some deep branches rearranged in the transversion tree (not shown). (b), Consensus for the four most parsimonious trees (267 steps) inferred by protein parsimony.

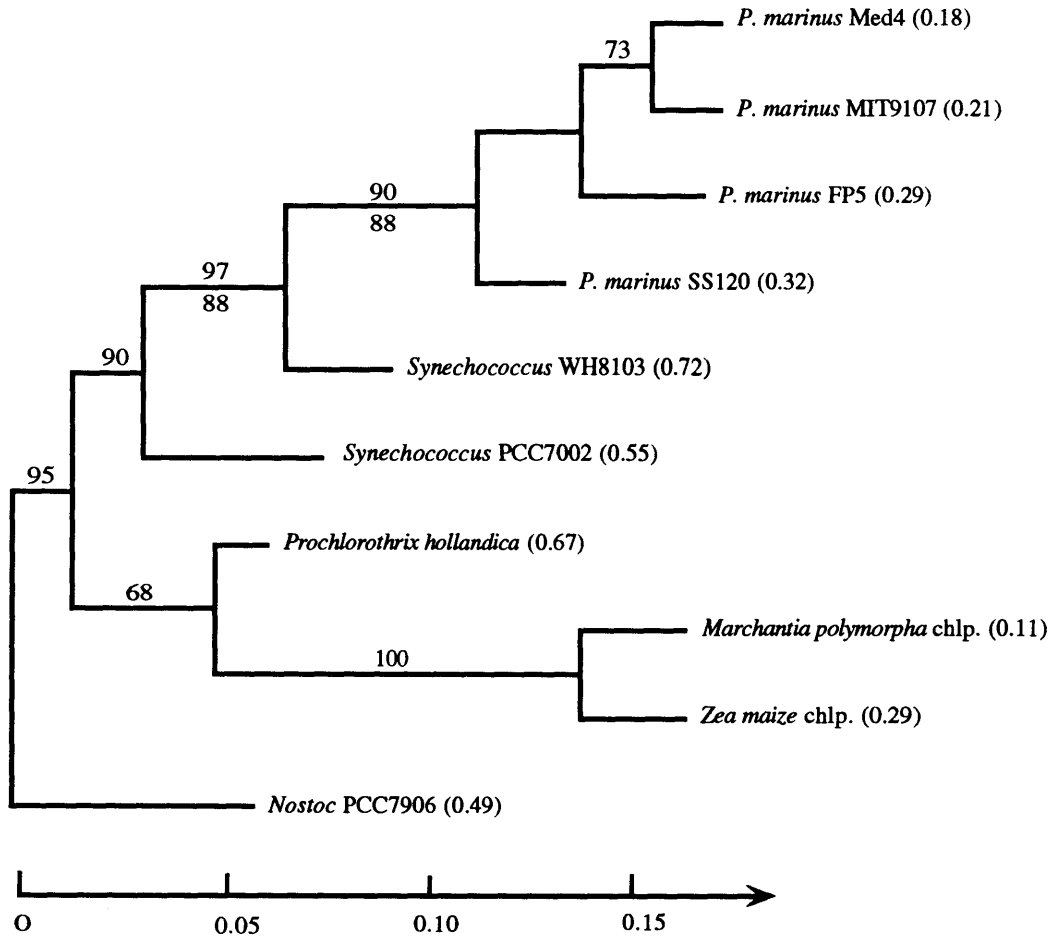

P. marinus Med4 (0.25)
P. marinus MIT9107 (0.27)
P. marinus FP5 (0.26)
P. marinus SS120 (0.34)
Synechococcus WH8103 (0.66)
Synechocosystis PCC6803 (0.52)
Prochlorothrix hollandica (.58)
Synechococcus PCC7942 (0.61)
Zea maize chloroplast (0.26)
Anabaena PCC7120 (0.39)
100
98
68
96
100
97
100
65
55
87
71
68
72
0.2
0.4
Genetic distance

a

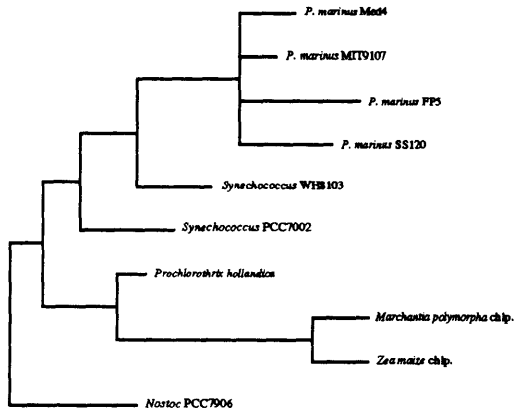

P. marinus Med4
P. marinus MIT9107
P. marinus FP5
P. marinus SS120
Synechococcus WH8103
Zea maize chloroplast
Synechococcus PCC7942
Synechocosystis PCC6803
Prochlorothrix hollandica
Anabaena PCC7120

b

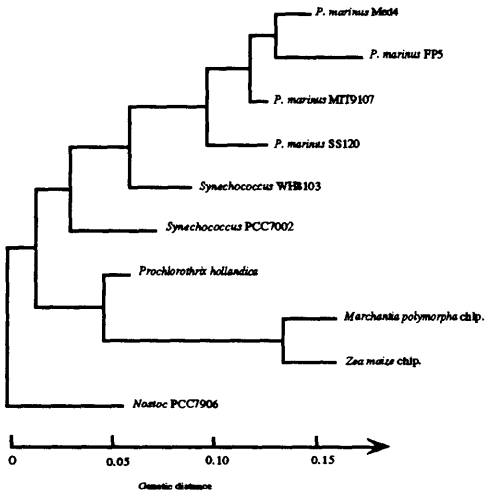
Figure 3. Phylogenetic relationships among cultured strains of *P. marinus* and other members of the cyanobacterial lineage inferred using *petB/D* sequences. Trees are arbitrarily rooted to *Nostoc* PCC7906 to facilitate comparison to those inferred from 16S rRNA gene sequences (Figure 1). (a), Distance matrix tree inferred by neighbor-joining with Kimura two parameter genetic distance estimates (Table 2c). Numbers in parentheses indicate G+C base composition at third codon positions; other numbers are as in Figure 1a. An identical branching pattern was inferred by protein parsimony, and for *P. marinus* and *Synechococcus* WH8103 by neighbor-joining analysis of transversion distances, which inferred a different pattern for more deeply branching lineages (not shown). (b), Consensus of the two best trees inferred by DNA parsimony (129 steps). (c), Tree inferred by maximum likelihood (ln L = -1031.58031), which was not significantly better than the tree inferred by neighbor-joining (ln L = -1038.59363, S.D. = 6.7154, Δln L = -7.01331).


P. marinus Med4 (0.18)
73
P. marinus MIT9107 (0.21)
90
88
P. marinus FP5 (0.29)
97
88
P. marinus SS120 (0.32)
90
Synechococcus WH8103 (0.72)
Synechococcus PCC7002 (0.55)
95
Prochlorothrix hollandica (0.67)
68
Marchantia polymorpha chlp. (0.11)
100
Zea maize chlp. (0.29)
Nostoc PCC7906 (0.49)
0
0.05
0.10
0.15
Genetic distance

a


P. marinus Med4
P. marinus MIT9107
P. marinus FP5
P. marinus SS120
Synechococcus WH8103
Synechococcus PCC7002
Prochlorothrix hollandica
Marchantia polymorpha chlp.
Zea maize chlp.
Nostoc PCC7906

b


P. marinus Med4
P. marinus FP5
P. marinus MIT9107
P. marinus SS120
Synechococcus WH8103
Synechococcus PCC7002
Prochlorothrix hollandica
Marchantia polymorpha chlp.
Zea maize chlp.
Nostoc PCC7906
0
0.05
0.10
0.15

c

Table 2a. Genetic distance and fractional similarity for pairwise comparisons of 16S rRNA gene sequences.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. *Agrobacterium tumefaciens*	-	2843	2793	2829	2838	2760	2789	2838	2843	2731	2748	2863	2845	3043	2673
2. *P. marinus* Med4	877	-	0093	0149	0178	0282	0331	0379	0028	0283	0331	1060	1424	1841	1653
3. *P. marinus* MIT9107	881	994	-	0206	0178	0283	0351	0399	0102	0283	0351	1095	1435	1832	1678
4. *P. marinus* FP5	877	994	989	-	0187	0244	0340	0379	0178	0341	0340	1059	1378	1813	1628
5. *P. marinus* SSW5	879	991	991	991	-	0225	0293	0292	0187	0293	0293	1060	1376	1865	1673
6. *P. marinus* MIT9303	878	992	988	994	992	-	0197	0244	0311	0293	0206	0908	1329	1823	1599
7. *Synechococcus* WH8103	877	987	985	989	987	990	-	0225	0351	0234	0046	0986	1374	1909	1614
8. *Synechococcus* WH7805	874	987	985	989	989	992	993	-	0409	0321	0254	0975	1363	1885	1662
9. SAR6	877	998	994	993	991	990	985	985	-	0283	0351	1093	1448	1879	1653
10. SAR7	879	987	987	985	985	986	993	987	987	-	0225	1104	1422	1914	1602
11. SAR139	877	987	985	989	987	990	998	991	985	993	-	0965	1374	1883	1568
12. *Synechococcus* PCC6301	875	960	957	962	960	969	960	962	959	959	960	-	1312	1779	1439
13. *Prochlorothrix hollandica*	871	942	942	944	946	948	947	947	940	944	947	943	-	1453	1079
14. *Anabaena* PCC7120	864	936	935	940	936	941	938	938	935	935	940	939	947	-	1547
15. *Marchantia polymorpha* clp	876	930	928	932	932	936	934	932	930	934	934	938	949	942	-

Genetic distances ($x10^4$) estimated by the Kimura two parameter method (Kimura 1980) with transition/transversion ratio 2.0, above the diagonal, and fractional similarity values ($x10^3$) below, for pairwise comparisons of 16S rRNA gene sequences for *P. marinus* isolates Med4, MIT9107 and FP5 (this work, Genbank accession numbers), the *P. marinus* Sargasso Sea isolate SSW5 (Urbach et al. 1992), cloned sequences from uncharacterized Sargasso Sea picoplankton (Giovannoni et al. 1990, Britschgi and Giovannoni 1991), other members of the cyanobacterial lineage (Tomioka and Sugiura 1983, Ohyama et al. 1986, Giovannoni et al. 1988, Turner et al. 1989, Lingnon et al. 1991, D.L. Distel and J.B. Waterbury, unpublished data) and *Agrobacterium tumefaciens* (Yang et al. 1985). These comparisons and other 16S rRNA phylogenetic analyses in Figure 1 considered 1085 aligned sequence positions corresponding to nucleotides 143 to 1312 in the *Escherichia coli* 16S rRNA sequence, excluding positions where alignment or nucleotide identities were ambiguous. Genetic distance values were used to infer the 16S rRNA neighbor-joining tree (Figure 1a).

Table 2b. Evolutionary distance and fractional similarity at all, first two and third codon positions for pairwise comparisons of *psbB* sequences.

	1	2	3	4	5	6	7	8	9	10
1. *P. marinus* Med4	-	0308	0907	1314	2013	3442	3278	2843	3175	3647
2. *P. marinus* MIT9107	823 **970** *527*	-	0949	1319	2094	3345	3213	2962	3222	3768
3. *P. marinus* FP5	789 **916** *533*	797 **913** *564*	-	1211	2155	3319	3196	3395	3413	3970
4. *P. marinus* SS120	741 **883** *455*	739 **883** *448*	765 **892** *509*	-	2435	3299	3413	3170	3102	3763
5. *Synechococcus* WH8103	663 **832** *321*	653 **826** *303*	667 **823** *352*	659 **805** *364*	-	2805	2584	2704	2453	3432
6. *Synechocystis* PCC6803	588 **745** *273*	608 **751** *321*	612 **748** *339*	616 **751** *345*	659 **781** *412*	-	1286	1920	1784	2684
7. *Prochlorothrix hollandica*	608 **754** *315*	629 **757** *370*	614 **757** *327*	629 **745** *394*	685 **796** *461*	765 **886** *521*	-	2029	1815	2857
8. *Synechococcus* PCC7942	629 **778** *327*	616 **769** *309*	598 **742** *309*	627 **757** *364*	665 **784** *424*	713 **838** *461*	715 **829** *485*	-	2135	3395
9. *Anabaena* PCC7120	627 **760** *358*	639 **757** *400*	645 **745** *442*	639 **763** *388*	681 **802** *436*	729 **850** *485*	733 **847** *503*	695 **823** *436*	-	2423
10. *Zea maize* chloroplast	604 **730** *352*	614 **724** *394*	596 **712** *364*	610 **721** *388*	574 **745** *230*	635 **784** *333*	610 **772** *285*	592 **742** *291*	671 **802** 406	-

(Table 2b, continued)

Genetic distances ($x10^4$) estimated by the Kimura two parameter method with transition/transversion ratio 2.0, for pairwise comparisons of first two codon positions, above the diagonal, and fractional similarity values ($x10^3$) for all codon positions (plain print), first two codon positions (bold print) and third codon positions (italic print), below, for pairwise comparisons of *psbB* sequences for *P. marinus* isolates, *Synechococcus* WH8103 (this work, Genbank accession numbers) and other members of the cyanobacterial lineage (Vermaas et al. 1987, Lang and Haselkorn 1989, Greer and Golden 1991, Rock et al. 1987, Kulkarni and Golden, unpublished). These comparisons and other *psbB* phylogenetic analyses considered 498 aligned sequence positions (333 first and second codon positions) corresponding to nucleotides 803 to 1307 in the *Synechocystis* PCC6803 *psbB* sequence, excluding positions where nucleotide identities were ambiguous for any sequence. Genetic distance values were used to infer the *psbB* neighbor-joining tree (Figure 2a).

TableII-2c. Fractional similarity at all, first two and third codon positions and genetic distance at first two codon positions for pairwise comparisons of *petB/D* sequences.

	1	2	3	4	5	6	7	8	9	10	11
1. *P. marinus* Med4	-	0276	0568	0607	0913	0867	1247	1282	1607	1842	1842
2. *P. marinus* MIT9107	838 **973** *581*	-	0566	0566	1048	0870	1251	1149	1370	1898	1998
3. *P. marinus* FP5	807 **946** *527*	804 **946** *535*	-	0698	0918	0872	1208	1244	1616	2004	1853
4. *P. marinus* SS2	788 **942** *473*	832 **946** *605*	801 **934** *527*	-	0913	0916	1293	1237	1607	2055	2106
5. 85Br100 (*P. marinus* MIT9303)	722 **915** *339*	716 **903** *339*	740 **915** *394*	755 **915** *440*	-	0522	1340	1140	1704	1930	1733
6. *Synechococcus* WH8103	709 **919** *287*	693 **919** *248*	729 **919** *349*	724 **915** *341*	789 **950** *472*	-	1152	1003	1410	1925	1875
7. *Prochlorothrix hollandica*	706 **888** *357*	688 **888** *295*	722 **891** *395*	695 **884** *318*	740 **880** *448*	768 **895** *512*	-	0965	1020	1188	1188
8. *Synechococcus* PCC7002	709 **884** *357*	714 **895** *357*	709 **888** *349*	716 **888** *372*	722 **895** *386*	763 **907** *481*	776 **911** *512*	-	1205	1523	1713
9. *Nostoc* PCC7906	698 **860** *380*	706 **880** *372*	706 **860** *403*	685 **860** *326*	696 **853** *378*	714 **876** *380*	750 **907** *442*	737 **891** *426*	-	1579	1723
10. *Marchantia polymorpha* chlp.	755 **841** *589*	750 **837** *589*	716 **829** *496*	726 **826** *519*	649 **833** *285*	644 **833** *264*	701 **891** *333*	696 **864** *349*	711 **860** *419*	-	0477
11. *Zea maize* chlp.	714 **841** *473*	704 **829** *457*	698 **841** *426*	685 **822** *411*	673 **849** *316*	670 **837** *341*	706 **891** *341*	693 **849** *380*	698 **849** *411*	835 **953** *605*	-

(Table II-2c, continued)

Genetic distances ($x10^4$) estimated by the Kimura two parameter method with transition/transversion ratio 2.0, for pairwise comparisons of first two codon positions, above the diagonal, and fractional similarity values ($x10^3$) for all codon positions (plain print), first two codon positions (bold print) and third codon positions (italic print), below, for pairwise comparisons of coding regions of *petB/D* sequences for *P. marinus* isolates, *Synechococcus* WH8103 (this work, Genbank accession numbers) and other members of the cyanobacterial lineage (Brand et al. 1981, Rock et al. 1987, Kallas et al. 1988, Ohyama et al. 1986, Greer and Golden 1991). These comparisons and other *petB/D* phylogenetic analyses considered 419 aligned sequence positions, (258 first and second codon positions) corresponding to nucleotides 315 to 699 (the 3' end) of *petB* and positions 1 to 38 of *petD* from *Synechococcus* PCC7002, excluding positions where nucleotide identities were ambiguous for any sequence. Genetic distance values were used to infer the *petB/D* neighbor-joining tree (Figure II-4a).

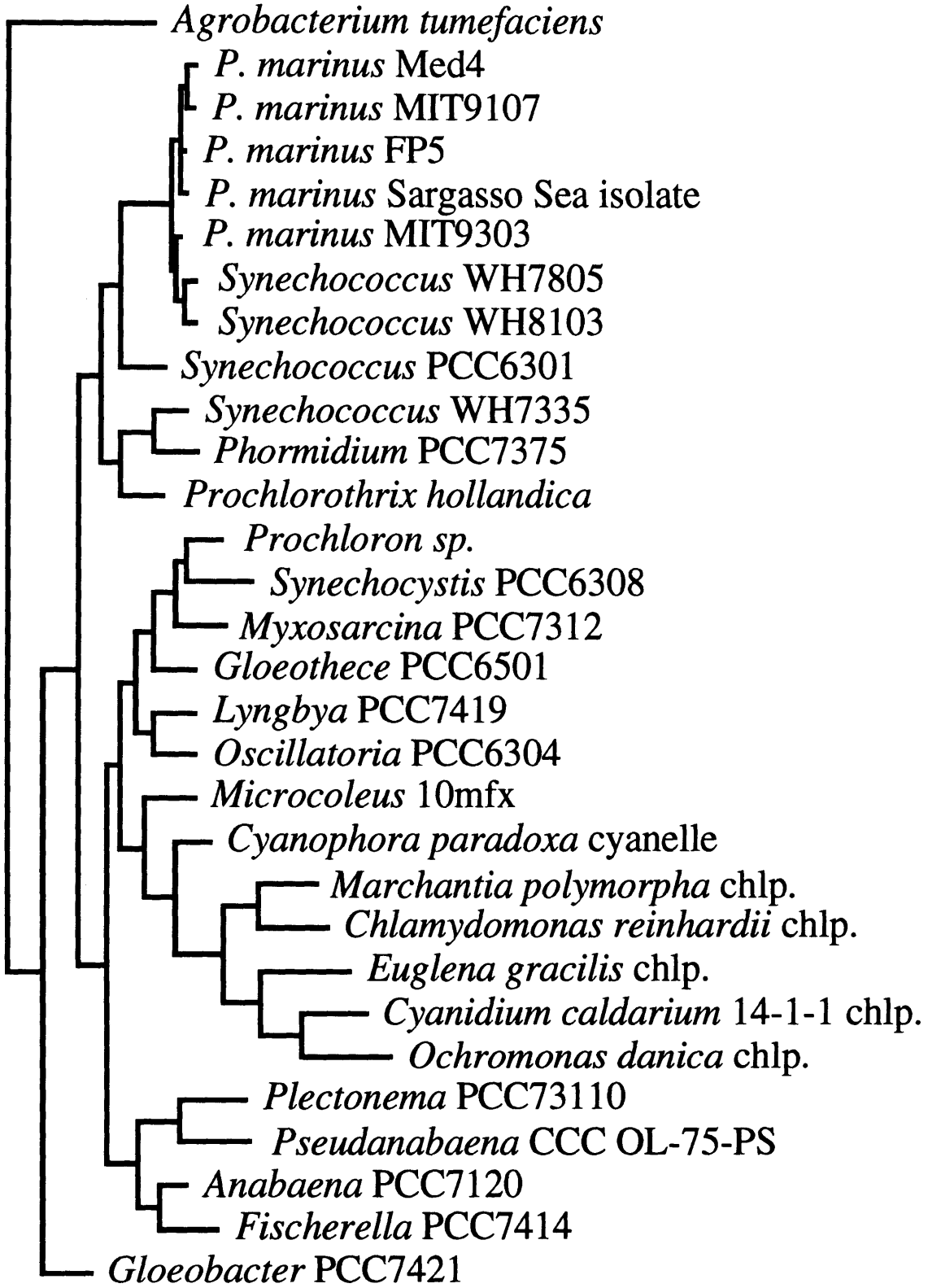
clustering was consistent in neighbor-joining and parsimony analyses, whether or not chloroplast sequences were included in the datasets, and persisted in a 16S rRNA gene tree which included diverse cyanobacterial taxa (Figure 4).

P. marinus strains Med4, MIT9107, FP5 and SS120 (the "*P. marinus* four culture clade") formed a phylogenetic group distinct from *Synecyococcus* WH8103 in all analyses. However, the phylogenetic position of the deeply branching *P. marinus* MIT9303 was different in 16S rRNA trees inferred by different methods.

Identical branching patterns were inferred for relationships among members of the *P. marinus* four culture clade and *Synechococcus* WH8103 for all three genes using the neighbor-joining method, and for 16S rRNA and *psbB* genes using the parsimony and maximum likelihood methods as well (Tables 2a, b, c, Figures 1a, b, c, 2a, 3a). For *petB/D* sequences, however, relationships among the four *P. marinus* cultures could not be resolved using parsimony (Figure 3b), and the maximum likelihood tree, though not significantly more likely than the neighbor-joining tree, exhibited a slightly altered branching order (Figure 3c).

Bootstrap analysis (100 resampled datasets each for neighbor-joining and parsimony calculations) strongly supported the pattern inferred by all three methods for 16S and *psbB* sequences for branching order among members of the *P. marinus* four culture clade and *Synechococcus* WH8103, except that the relative positions of strains FP5 and SS120 received varying levels of support (Figures 1a, 2a). Consistent with the generally lower resolving power in the *petB/D* dataset, bootstrap values were lower in trees inferred from these sequences (Figure 3a). The lower resolution in *petB/D* trees may be due to the shorter length of *petB/D* sequences; after intergenic regions and third

Figure 4. Phylogenetic relationships among *P. marinus* cultured strains and diverse cyanobacterial taxa, inferred by neighbor joining using 16S rRNA sequences. Differences in the branching pattern between this tree and the tree in Figure 1a are presumably due to the smaller number of nucleotides used in the analysis, due to short length and sequence ambiguities of some sequences.


Agrobacterium tumefaciens
P. marinus Med4
P. marinus MIT9107
P. marinus FP5
P. marinus Sargasso Sea isolate
P. marinus MIT9303
Synechococcus WH7805
Synechococcus WH8103
Synechococcus PCC6301
Synechococcus WH7335
Phormidium PCC7375
Prochlorothrix hollandica
Prochloron sp.
Synechocystis PCC6308
Myxosarcina PCC7312
Gloeothece PCC6501
Lyngbya PCC7419
Oscillatoria PCC6304
Microcoleus 10mfx
Cyanophora paradoxa cyanelle
Marchantia polymorpha chlp.
Chlamydomonas reinhardii chlp.
Euglena gracilis chlp.
Cyanidium caldarium 14-1-1 chlp.
Ochromonas danica chlp.
Plectonema PCC73110
Pseudanabaena CCC OL-75-PS
Anabaena PCC7120
Fischerella PCC7414
Gloeobacter PCC7421

codon positions were removed from the analysis, *petB/D* phylogenetic inferences were drawn from 258 nucleotide positions, while *psbB* and 16S rRNA gene analyses considered 333 and 1085 nucleotide postitions, respectively.

The phylogenetic position of *P. marinus* strain MIT9303 was ambiguous. While 16S rRNA gene trees inferred by the three standard methods consistently recovered the *P. marinus* four culture clade as a coherent group, they were inconsistent in their depictions of the relative branching order of five more deeply branching lineages in the *P. marinus*/marine A *Synechococcus* lineage, including *P. marinus* MIT9303, the four culture clade, the two marine A *Synechococcus* strains and SAR7, a cloned sequence from the Sargasso Sea (Giovannoni et al. 1990) (Figure 1a, b, c). In bootstrap calculations, the branching order of these deeply branching lineages received less than 50% support (Figure 1a). The discordance in the branching order of *P. marinus* MIT9303 and marine A *Synechococcus* lineages inferred by different phylogenetic methods and their low bootstrap values suggests a near-simultaneous radiation of *P. marinus* and marine A *Synechococcus* lineages during evolutionary history (Hoelzer and Melnick 1994).

Effects of nucleotide substitution biases. We explored the possible effects of nucleotide substitution bias on our phylogenetic trees by labelling the *psbB* and *petB/D* neighbor-joining trees according to their third codon position G+C content (Lockhart and Penny 1992, Lockhart et al. 1992b) (Figures 2a, 3a). It is immediately apparent from inspection of these labelled trees that *P. marinus* isolates fail to cluster with chloroplasts despite similarities in their third codon position G+C content (Figures 2, 3), and that nucleotide substitution bias is therefore not the sole organizing factor in these trees. 16S rRNA gene trees also fail to group *P. marinus* with chloroplasts (Figure 1). However,

within the *P. marinus*/marine A *Synechococcus* cluster the phylogenetic branching pattern inferred by the majority of analyses *is* correlated with G+C content at third codon positions, with the Med4/MIT9107 phylogenetic pair showing considerably lower third codon position G+C content (0.18 to 0.27) than SS120 (0.32 to 0.34) and *Synechococcus* WH8103 (0.66 to 0.72). Although this clustering may reflect actual phylogenetic relationships in which substitution biases simply differ among lineages, it is also consistent with a systematic error in phylogenetic inference (Lockhart et al. 1992a). Again, there is no accepted molecular criterion for determining which hypothesis is correct.

To further explore the possible effects of nucleotide substitution biases, phylogenetic branching patterns were assessed using two methods of analysis thought to be insensitive to them (Woese 1991, Hasegawa and Hashimoto 1993): transversion analysis with neighbor-joining and, for the two protein-encoding sequences, amino acid parsimony. Transversion analysis inferred an identical branching pattern for *Synechococcus* WH8103 and the members of the *P. marinus* four culture clade as was found by standard neighbor-joining for all three gene sequences, although the relative branching order for deep branches in the *P. marinus*/marine A *Synechococcus* clade agreed with the maximum likelihood calculation in the 16S rRNA transversion tree (Figures 1, 2, 3). Relationships inferred by protein parsimony were also largely congruent with trees inferred by standard methods, except that *P. marinus* strains SS120 and FP5 were joined to form a sister clade to *P. marinus* Med4 and MIT9107 in the *psbB* protein parsimony tree (Figure 2b).

In sum, transversion and protein parsimony analyses do not suggest a consistent, alternative phylogenetic branching pattern, but instead reinforce the conclusions of the

standard neighbor-joining, parsimony and maximum likelihood analyses: *P. marinus* Med4 is most closely related to *P. marinus* MIT9107, below which diverge *P. marinus* strains FP5 and SS120 in somewhat indeterminate order, with the most deeply branching culture being *P. marinus* MIT9303. The branching order of five deeply branching lineages in the *P. marinus*/marine A *Synechococcus* cluster is not resolved.

Intergenic region and third codon position mismatches. Intergenic "nonsense" regions and third codon (silent) positions in protein-encoding genes are rapidly evolving, and therefore sensitive indicators of small amounts of evolutionary change. However, when compared to the first two codon positions in protein-encoding genes or to 16S rRNA genes, intergenic regions and third codon positions become rapidly saturated with nucleotide substitutions and (in intergenic regions) insertions and deletions. These tend to obscure evolutionary relationships when homologous nucleotide positions cannot be aligned or when the proportion of superimposed substitutions becomes large. Comparison of aligned sequences of *petB/D* and *psbB* from cultured strains of *P. marinus*, *Synechococcus* WH8103, chloroplasts and other cyanobacteria illustrates this point. In comparisons of *P. marinus petB/D* intergenic regions, which range from 34 to 90 basepairs in length, homologous nucleotide positions could be identified only for a portion of the intergenic regions of strains Med4 and MIT9107 (Figure 5). Third codon position mismatches ranged from 41.1% to 64.3% (similarity 0.589 to 0.357) for *psbB* and from 39.5 to 66.1 percent (similarity 0.65 to 0.339) for *petB/D*, and were considered too large to give reliable phylogenetic information (Tables 2b, c) (for a similar analysis, see Bhattacharya et al. 1991). In addition, third codon positions were highly heterogeneous in G+C content (Table 3), which may bias phylogenetic inferences. Intergenic regions and third codon positions were therefore omitted from phlyogenetic analyses.

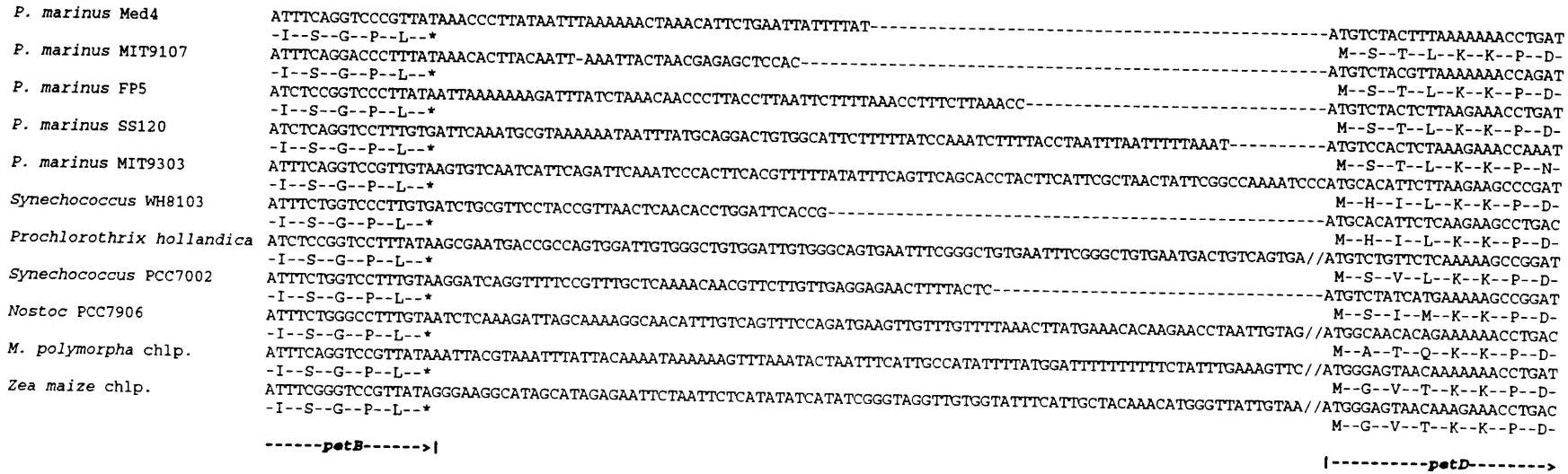
Table 3. G+C Base compositions of sequences used in these analyses.

petB/D	Codon position(s)		
	All	1st2	3rd
P. marinus Med4	0.37	0.47	0.18
P. marinus MIT9107	0.38	0.46	0.21
P. marinus FP5	0.40	0.45	0.29
P. marinus SS120	0.41	0.46	0.32
Synechococcus WH8103	0.57	0.49	0.72
Prochlorothrix hollandica	0.54	0.47	0.67
Synechococcus PCC7002	0.50	0.48	0.55
Nostoc PCC7906	0.49	0.49	0.49
Marchantia chloroplast	0.32	0.43	0.11
Zea maize chloroplast	0.41	0.47	0.29

psbB	Codon position(s)		
	All	1st2	3rd
P. marinus Med4	0.39	0.47	0.25
P. marinus MIT9107	0.40	0.47	0.27
P. marinus FP5	0.41	0.48	0.26
P. marinus SS120	0.45	0.51	0.34
Synechococcus WH8103	0.58	0.54	0.66
Synechocystis PCC6803	0.51	0.50	0.52
Prochlorothrix hollandica	0.56	0.55	0.58
Synechococcus PCC7942	0.55	0.52	0.61
Anabaena PCC7120	0.48	0.53	0.39
Zea maize chloroplast	0.42	0.49	0.26

16S rRNA	G+C composition
Agrobacterium tumefaciens	0.54
P. marinus Med4	0.53
P. marinus Pac7	0.53
P. marinus FP5	0.54
P. marinus SSW5	0.54
P. marinus MIT9303	0.54
Synechococcus WH8103	0.54
Synechococcus WH7805	0.54
SAR6	0.53
SAR7	0.54
SAR139	0.54
Synechococcus PCC6301	0.54
Prochloron sp.	0.52
Anabaena PCC7120	0.53
Marchantia polymorpha chloroplast	0.54

Figure 5. Comparison of *petB/D* intergenic region sequences for cultured strains of *P. marinus* and other members of the oxygenic photosynthetic radiation. Sequences are aligned according to amino acid translations in coding regions, and, with the exception of the sequence for *P. marinus* MIT9107, intergenic regions are arbitrarily represented as continuous with *petB*. An alignment gap has been inserted into the *P. marinus* MIT9107 intergenic region sequence in order to accentuate the similarity between the 5' end of this sequence and that from *P. marinus* Med4. Intergenic region sequences from *Prochlorothrix hollandica*, *Nostoc* PCC7906 and *Marchantia polymorpha* and *Zea maize* chloroplasts have been truncated. Sequence data for *P. marinus* MIT9303 are from this work, other attributions are as in Table 2c.


P. marinus Med4
P. marinus MIT9107
P. marinus FP5
P. marinus SS120
P. marinus MIT9303
Synechococcus WH8103
Prochlorothrix hollandica
Synechococcus PCC7002
Nostoc PCC7906
M. polymorpha chlp.
Zea maize chlp.
ATTTCAGGTCCCGTTATAAACCCTTATAATTTAAAAAACTAAACATTCTGAATTATTTTAT----------ATGTCTACTTTAAAAAAACCTGAT
-I--S--G--P--L--*
M--S--T--L--K--K--P--D-
ATTTCAGGACCCTTTATAAACACTTACAATT-AAATTACTAACGAGAGCTCCAC----------ATGTCTACGTTAAAAAAACCAGAT
-I--S--G--P--L--*
M--S--T--L--K--K--P--D-
ATCTCCGGTCCCTTATAATTAAAAAAAGATTTATCTAAACAACCCTTACCTTAATTCTTTTAAACCTTTCTTAAACC----------ATGTCTACTCTTAAGAAACCTGAT
-I--S--G--P--L--*
M--S--T--L--K--K--P--D-
ATCTCAGGTCCTTTGTGATTCAAATGCGTAAAAAATAATTTATGCAGGACTGTGGCATTCTTTTTATCCAAATCTTTTACCTAATTTAATTTTTAAAT----------ATGTCCACTCTAAAGAAACCAAAT
-I--S--G--P--L--*
M--S--T--L--K--K--P--N-
ATTTCAGGTCCGTTGTAAGTGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTTCAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCCATGCACATTCTTAAGAAGCCCGAT
-I--S--G--P--L--*
M--H--I--L--K--K--P--D-
ATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGGATTCACCG----------ATGCACATTCTCAAGAAGCCTGAC
-I--S--G--P--L--*
M--H--I--L--K--K--P--D-
ATCTCCGGTCCTTTATAAGCGAATGACCGCCAGTGGATTGTGGGCTGTGGATTGTGGGCAGTGAATTTCGGGCTGTGAATTTCGGGCTGTGAATGACTGTCAGTGA//ATGTCTGTTCTCAAAAAGCCGGAT
-I--S--G--P--L--*
M--S--V--L--K--K--P--D-
ATTTCTGGTCCTTTGTAAGGATCAGGTTTTCCGTTTGCTCAAAACAACGTTCTTGTTGAGGAGAACTTTTTACTC----------ATGTCTATCATGAAAAAGCCGGAT
-I--S--G--P--L--*
M--S--I--M--K--K--P--D-
ATTTCTGGGCCTTTGTAATCTCAAAGATTAGCAAAAGGCAACATTTGTCAGTTTCCAGATGAAGTTGTTTGTTTTAAACTTATGAAACACAAGAACCTAATTGTAG//ATGGCAACACAGAAAAAACCTGAC
-I--S--G--P--L--*
M--A--T--Q--K--K--P--D-
ATTTCAGGTCCGTTATAAATTACGTAAATTTATTACAAAATAAAAAAGTTTAAATACTAATTTCATTGCCATATTTTATGGATTTTTTTTTTCTATTTGAAAGTTC//ATGGGAGTAACAAAAAAACCTGAT
-I--S--G--P--L--*
M--G--V--T--K--K--P--D-
ATTTCGGGTCCGTTATAGGGAAGGCATAGCATAGAGAATTCTAATTCTCATATATCATATCGGGTAGGTTGTGGTATTTCATTGCTACAAACATGGGTTATTGTAA//ATGGGAGTAACAAAGAAACCTGAC
-I--S--G--P--L--*
M--G--V--T--K--K--P--D-
------petB------>|
|----------petD-------->
57

DISCUSSION

Chlorophyll a_2 and b_2-containing marine prokaryotes arise from a single lineage within the cyanobacteria. The cultured *P. marinus* strains included in this study are phylogenetically restricted to a single lineage within the cyanobacteria, which they share with marine A *Synechococcus* strains WH8103 and WH7805. Neighbor-joining, parsimony and maximum likelihood analyses of nucleotide sequences, parsimony analysis of amino acid sequences, and analysis of transversions by neighbor-joining, all replicate this result, which links strains of *P. marinus* and marine A *Synechococcus* having very different G+C substitution biases, despite the presence in the dataset of chloroplast sequences having third codon position G+C content similar to those of *P. marinus*. The consistency of these results argues that the *P. marinus*/marine A *Synechococcus* phylogenetic grouping represents a true evolutionary lineage, and not an artefactual cluster.

Branching patterns among four cultured isolates of *P. marinus* and *Syenchococcus* WH8103 are highly consistent. The branching order inferred for *P. marinus* Med4, MIT9107, FP5 and SS120 and *Synechococcus* WH8103 in phylogenies inferred from three gene sequences was consistent. It is therefore likely that gene transfer does not play a major role in the evolution of these organisms and that *P. marinus* exhibits clonal inheritence similar to that in *Escherichia coli* (Selander et al. 1987). This finding buttresses the assumption that prokaryotic organisms exhibiting closely related sequences at a single locus are likely to be genetically similar at most loci, and hence phenotypically similar. This assumption is implicit in most analyses of sequences cloned from natural populations (Young 1989).

Phylogenetic relationships among *P. marinus* isolates do not correlate with geography. *P. marinus* Mediterranean strain Med4 and Pacific strain MIT9107, both isolated from mixed surface waters, were found to be closely related, while north Atlantic strain FP5 and Sargasso Sea deep chlorophyll maximum cultures SS120 and and MIT9303 fell into progressively more deeply branching lineages. There was thus no correlation between genetic distance and the geographic distance between sites of culture isolation, with the two Sargasso Sea isolates exhibiting less sequence similarity than the Med4/MIT9107 pair.

The different possible branching positions for *P. marinus* MIT9303 suggest different evolutionary scenarios for *P. marinus* and marine A *Synechococcus* pigments. The ambiguity in the branching position of *P. marinus* MIT9303 leaves open the question of the origins of *P. marinus* and marine A *Synechococcus* pigments. If the correct placement of *P. marinus* MIT9303 were as the deepest branch of the *P. marinus*/marine A *Synechococcus* cluster, as indicated by the 16S rRNA gene parsimony and maximum likelihood trees (Figures 1b, c), this would suggest that the chl a_2 and b_2 *P. marinus* phenotype was ancestral to the whole *P. marinus*/marine A *Synechococcus* cluster, implying further that cells capable of this phenotype arose from an ancestor shared with *Synechococcus* PCC6301 and gave rise to *Synechococcus* WH8103 (both strains being chl a_1 and phycobilisome-containing cyanobacteria, neither of which contains chl a_2 or chl *b* (Rudiger and Schoch 1988, Waterbury et al. 1979, L. Moore, pers. comm.). If, on the other hand, the correct placement of *P. marinus* MIT9303 were at the base of a unitary *P. marinus* lineage, as inferred by neighbor-joining using 16S rRNA gene sequences (Figure 1a), then the phylogeny would be consistent with a single origin of the chl a_2 and b_2 phenotype without "reversion" to the chl a_1 and phycobilisome-containing phenotype. The characterization of the *P. marinus*

photosynthetic apparatus and its genes, with comparison to those of marine A *Synechococcus* is likely to yield the most conclusive data pertaining to this question.

Ambiguity in the branching order of deeply diverging lineages in the *P. marinus*/marine A *Synechococcus* cluster precludes identification of some cloned sequences from natural populations. Most studies investigating the phylogenetic diversity of uncultured microorganisms in the sea have employed 16S rRNA gene sequences, many of which fall into the *P. marinus*/marine A *Synechococcus* cluster (Giovannoni et al. 1990, Schmidt et al. 1991, Britschgi and Giovannoni 1991, DeLong et al., 1993, Fuhrman et al, 1993). While some of these sequences form close phylogenetic associations with cultured *P. marinus* or marine A *Synechococcus* strains (e.g. SAR6 and SAR139), others (e.g. SAR7) form deep branches in the *P. marinus*/marine A *Synechococcus* cluster and are not specifically related to sequences from characterized cells (Figure 1). Because the branching order of *P. marinus* and marine A *Synechococcus* lineages in the cluster cannot at present be resolved, these cloned sequences cannot confidently be assigned to either taxonomic group. Assignment of phenotypes to these sequences must await the discovery of closely related sequences in phenotypically characterized cells, results of *in situ* hybridization experiments, or future refinements in the methods of phylogenetic inference.

REFERENCES

Bhattacharya, D., Stickel, S.K. and Sogin, M.L. (1991). Molecular phylogenetic analysis of actin genic regions from *Achlya bisexualis* (Oomycota) and *Costaria costata* (Chromophyta). J. Mol. Evol. 33:525-536.

Brand, S.N., Tan, X. and Widger, W.R. (1992). Cloning and sequencing of the *petBD* operon from the cyanobacterium *Synechococcus sp.* PCC7002. Plant Mol. Biol. 20:481-491.

Britschgi, T.B. and Giovannoni, S.J. (1991) Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. 57:1707-1713.

Campbell, L., and Vaulot, D. (1993). Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawii (station ALOHA). Deep-Sea Res. 40:2043-2060.

Chisholm, S.W., R.J. Olson, E.R. Zettler, R. Goericke, J. Waterbury, and N. Welschmeyer. (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature*, 334(6180):340-343.

Chisholm, S.W., Frankel, S.L., Goericke, R., Olson, R.J., Palenik, B., Waterbury, J.B., West-Johnsrud, L. and Zettler, E.R. (1992). *Prochlorococcus marinus nov.* gen *nov.* sp.: a marine prokaryote containing divinyl chlorophyll *a* and *b*. Arch. Microbiol. 157:297-300.

DeLong, E.F., Franks, D.G. and Alldredge, A.L. (1993) Phylogenetic diversity of aggregate-attatched vs. free-living marine bacterial assemblages. Limnol. Oceanogr. 38:924-934.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. 22:521-65.

Felsenstein, J. (1991). Phylip version 3.4 (University of Washington, Seattle, WA).

Fitch. W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. Science 155:279-284.

Fuhrman, J.A., McCallum, K. and Davis, A.A. (1993). Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. Appl. Environ. Microbiol. 59:1294-1302.

Geiskes, W.W.C., G.W. Kraay, A. Nontji, D. Setiapermana, and Sutomo. 1988. Monsoonal alternation of a mixed and a layered structure in the phytoplankton of the euphotic zone of the Banda Sea (Indonesia): A mathematical analysis of algal pigment fingerprints. Neth. J. Sea Res. 22:(2):123-137.

Giovannoni, S.J., Turner, S., Olsen, G.J., Barnes, S., Lane, D.J. and Pace, N.R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. J. Bacteriol. 170:3584-3592.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and K.G. Field. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-62.

Goericke, R. and Repeta, D.J. (1992). The pigments of *Procholorcoccus marinus*: the presence of divinyl chlorophyll *a* and *b* in a marine prokaryote. Limnol. Oceanogr. 37:425-433.

Goericke, R. and Repeta, D.J. (1993). Chromatographic analysis of divinyl-chlorophylls *a* and *b* in samples from the subtropical north Atlantic Ocean. Mar. Ecol. Prog. Ser.

Goericke, R. and Welschmeyer, N.A. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep-Sea Res. 40:2283-2294.

Greer, K.L. and Golden, S.S. (1992). Conserved relationship between *psbH* and *petBD* genes: presence of a shared upstream element in *Prochlorothrix hollandica*. Plant Mol. Biol. 19:355-365.

Hasegawa, M. and Hashimoto, T. (1993). Ribosomal RNA trees misleading? Nature 361:23.

Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T. Adachi, J., Goto, N., Okamoto, K-i. and Hasegawa, M. (1994) Mol. Biol. Evol. 11:65-71.

Herdman, M., Janvier, M., Waterbury, J.B., Rippka, R. and Stanier, R.Y. (1979) Deoxyribonucleic acid base composition of cyanobacteria. J. Gen Microbiol. 111:63-71.

Hillis, D.M., and Dixon, M.T. (1991). Ribosomal DNA: Molecular evolution and phylogenetic inference. Quarterly Rev. Biol. 66:411-453.

Hoelzer, G.A and Melnick, D.J. (1994) Patterns of speciation and limits to phylogenetic resolution. Trends Ecol. Evol. 9:104-107.

Hultman, T., Stahl, S., Hornes, E. and Uhlen, M. (1989). Nucl. Acids Res. 17:4937-4946.

Kallas, T., Spiller, S. and Malkin, R.C. (1988). Characterization of two operons encoding the cytochrome b6-f complex of the cyanobacterium *Nostoc* PCC7906 and highly conserved sequences but different gene organization than in chloroplasts. J. Biol. Chem. 263:14334-14342.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

Knoth, K., Roberds, S., Poteet, C. and Tamkun, M. (1988). Highly degenerate, inosine-containing primers specifically amplify rare cDNA using the polymerase chain reaction. Nuleic Acids Res. 16:10371.

Lake, J.A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455-1459.

Lang, J.D., and Haselkorn, R. (1989). Isolation, sequence and transcription of the gene encoding the phtotsystem II chlorophyll-binding protein, CP-47, in the cyanobacterium *Anabaena* 7120. Plant Mol. Biol. 13:441-456.

Lantoine, F. and Neveux, J. (1993). Spectrofluorometric analysis of photosynthetic pigments in the Gulf of Lions (Mediterranean Sea) with an emphasis on the autotrophic picoplankton. Ann. Inst. Oceanogr., Paris 69:173-175.

Larsen, N., Olsen, G.J., Maidak, B.L., McCaughey, M.J., Overbeek, R., Macke, T.J., Marsh, T.L., and Woese, C.R. (1993) The ribosomal database project. Nucleic Acids Res. 21 Supplement:3021-3023.

Letelier, R.M., Bidigare, R.R., Hebel, D.V., Ondrusek, M., Winn, C.D. and Karl, D.M. (1993). Temporal variability of phytoplankton community structure based on pigment analysis. Limnol. Oceanogr. 38:1420-1437.

Li, H., Cui, X. and Arnheim, A. (1991). Analysis of DNA sequence variation in single cells. Methods 2:49-59.

Li, W.K.W., Dickie, P.M., Irwin, B.D., Wood, A.M. (1992). Biomass of bacteria, cyanobacteria, prochlorophytes and photosynthetic eukaryotes in the Sargasso Sea. Deep-Sea Res. 39:501-519.

Li, W.K.W., and M. Wood. (1988). Vertical distribution of North Atlantic ultraphytoplankton: analysis by flow cytometry and epifluorescence microscopy. Deep Sea Res. 35:1615-1638.

Lignon, P.J.B., Meyer, K.G., Martin, J.A. and Curtis, S.E. (1991). Nucl. Acids Res. 19:4553.

Lockhart, P.J., Penny, D. (1992). The problem of GC content, evolutionary trees and the origins of chl-*a/b* photosynthetic organelles: are the prochlorophytes a eubacterial model for higher plant photosynthesis? in *Research in Photosynthesis, Vol. III* (N. Murata, ed.). Kluwer Academic, pp. 499-505.

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J. and Larkum, A.W.D. (1992a). Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol 34:153-162.

Lockhart, P.J., Penny, D., Hendy, M.D., Howe, C.J., Beanland, T.J. and Larkum, A.W.D. (1992b). Controversy on chloroplast origins. FEBS Lett. 301:127-131.

Lockhart, P.J., Beanland, Howe, C.J. and Larkum, A.W.D. (1992c). Sequence of Prochloron didemni atpBE and the inference of chloroplast origins. Proc. Natl. Acad. Sci. USA 89:2742-2746.

Lockhart, P.J., Penny, D., Hendy, M.D., and Larkum, A.D. (1993). Is *Prochlorothrix hollandica* the best choice as a prokaryotic model for higher plant chl *a/b* photosynthesis? Photosynthesis Res. 37:61-68.

Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605-612.

Medlin, L.K., Williams, D.M. and P.A. Sims (1993). The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. Eur. J. Phycol. 28:261-275.

Moore, L.R., Goericke, R., and Chisholm, S.W. (1994). The comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, flourescence and absorptive properties. Mar. Ecol. Prog. Ser., *in press*.

Morden, C.W., Delwiche, C.F., Kuhsel, M., and Palmer, J.D. (1992). Gene phylogenies and the endosymbiotic origin of plastids. Biosystems 28:75-90.

Morel, A., Ahn, Y.-H., Partensky, F., Vaulot, D. and Claustre, H. (1993). J. Mar. Res. 51:617-649.

Neveux, J., Vaulot, D., Courties, C., and E. Fukai. (1989). Green photosynthetic bacteria associated with the deep chlorophyll maximum of the Sargasso Sea. C.R. Acad. Sci. Paris 308:9-14.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. and Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572-574.

Olsen, G.J. (1988). Phylogenetic analysis using ribosomal RNA. Methods Enzymol. 164:793-838.

Olsen, G.J. (1990). Sequence editor and analysis package (University of Illinois, Urbana, IL).

Olsen, G.J., Matsuda, H., Hagstrom, R and Overbeek, R. (1992). FastDNAml (Argonne National Laboratory, Argonne, IL).

Olsen, G.J. and Woese, C.R (1993). Ribosomal RNA: a key to phylogeny. FASEB J. 7:113-123.

Olson, R.J., Chisholm, S.W., Zettler, E.R., Altabet, M.A., and Dusenberry, J.A. (1990) Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. Deep Sea Res. **37**, 1033-1051.

Palenik, B. and Haselkorn, R. (1992). Multiple evolutionary origins of prochlorophytes, the chlorophyll *b*-containing prokaryotes. Nature 355:265-267.

Partensky, R., Hoepffner, N., Li, W.K.W., Ulloa, O. and Vaulot, D. (1993). Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains isolated from the north Atlantic and the Mediterranean Sea. Plant Physiol. 101:285-296.

Rock, C.D., Barkan, A. and Taylor, W.C. (1987). The maize plastid *psbB-psbF-petB-petD* RNAs encode alternative products. Curr. Genet 12:69-77.

Rudiger, W., and S. Schoch. (1988). Chlorophylls. In *Plant Pigments*. T.W. Goodwin, ed. Academic Press. 1-60.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning: a laboratory manual.* Cold Spring Harbor Laboratory Press, New York.

Schmidt, T.M., DeLong, E.F., and Pace, N.R. (1991). Analysis of a marine picoplankton community by 16SrRNA gene cloning and sequencing. J. Bacteriol. 173:4371-4378.

Selander, R.K., Caugant, D.A., and Whittam, T.S. (1987). Genetic structure and variation in natural populations of *Escherichia coli.* In Escherichia coli *and* Salmonella typhimunium*: cellular and molecular biology* (F.C. Neidhardt, ed.). Am. Soc. Microbiol. pp. 1625-1648.

Sogin, M.L., Gunderson, J.H., Elwood, H.L., Alonso, R.A. and Peattie, D.A. (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia.* Science 243:75-77.

Sogin, M.L., Hinkle, G., Leipe, D.D. (1993). Universal tree of life. Nature 362:795.

Swift, H. and Palenik, B. (1992). Prochlorophyte evolution and the origin of chloroplasts: morphological and molecular evidence. In *Origins of Plastids* (R.A. Lewin, ed.) Chapman and Hall, pp. 123-139.

Swofford, D.L. (1991). PAUP version 3.0 (Illinous Natural History Survey, Champaign, IL).

Tomioka, N. and Sugiura, M. (1983). The complete nucleotide sequence of a 16S ribosomal RNa gene from a blue-green alga, *Anacystis nidulans.* Mol. Gen. Genet. 191:45-50.

Turner, S., Burger.Wiersma, T., Giovannoni, S.J., Mur, L.R., and N.R. Pace. (1989). The relationship of a prochlorophyte *Prochlorothrix hollandica* to green chloroplasts. Nature 337:380-382.

Urbach, E., Robertson, D. and Chisholm, S.W. (1992). Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355:267-269.

Vaulot, D., Marie, D., Olson, R.J. and Chisholm, S.W. (1994). *Prochlorococcus* is highly synchronized to the diel cycle in the equatorial Pacific and divides up to once a day. submitted.

Vaulot, D., and Partensky, F. (1992). Cell cycle distributions of prochlorophytes in the northwestern Mediterranean Sea. Deep-Sea Res. 39:727-742.

Vaulot, D., Partensky, F., Neveux, J., Mantoura, R.F.C. and C. Llewellyn. (1990). Wintertime presence of prochlorophytes in surface waters of the North-Western Mediterranean Sea. Limnol. Oceanogr. 35:1156-1164.

Veldhuis, M.J.W. and Kraay, G.W. (1993). Cell abundance and Fluorescence of picoplakton in relation to growth irradience and nitrogen availability in the Red Sea. Netherlands J. Sea Res. 31:135-145.

Veldhuis, M.J.W. and Kraay, G.W. (1990). Vertical distribution and pigment composition of a picoplanktonic prochlorophyte in the subtropical N. Atlantic: a study of HPLC-analysis of pigments and flow cytometry. Mar. Ecol. Prog. Seq. 68:121-127.

Vermaas, W.F., Williams, J.G. and Arntzen, C.J. (1978). Sequencing and modification of *psbB*, the gene encoding the CP-47 protein of photosystem II in the cyanobacterium *Synechocustis* 6803. Plant Mol. Biol. 8:317-329.

Vermaas, W.F.J., and Ikeuchi, M. (1991). Photosystem II. in *The Photosynthetic Apparatus: molecular biology and operation* (L. Bogorad and I.K. Vasil, eds.). Academic Press pp. 26-111.

Waterbury, J.B. and Rippka, R. (1991). Subsection 1. Order *Croococcales* Wettstein 1924, emend. Rippka et al., 1979. in *Bergey's Manual of Systematic Bacteriology,* Vol. 3. J.T. Stanley, et al., eds. Williams and Wilkins.

Waterbury, J.B., Valois, F.S., Franks, D.G. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. in *Photosynthetic Picoplankton* (T. Platt and W.K.W. Li, eds.) Can. Bull. Fish Aquatic Sci. 214:71-120.

Waterbury, J.B., Watson, S.W., Guillard, R.R.L. and Brand, L.E. (1979) Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium. Nature 277:293-294.

Widger, W.R. and Cramer, W.A. (1991). The b_6/f complex. in *The Photosynthetic Apparatus: molecular biology and operation* (L. Bogorad and I.K. Vasil, eds.). Academic Press pp. 149-224.

Woese, C.R. 1987. Bacterial Evolution. Microbiol. Rev. 51:221-271.

Woese, C.R., Achenbach, L., Rouviere, P. and Mandelco, L. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeolobus fulgidus* in light of certain composition-induced artifacts. System. Appl. Microbiol. 14:364-371.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G.J., Woese, C.R. (1985). Mitochondrial origins. Proc. Natl. Acad. Sci. USA 82:4443-4447.

Young, J.P.W. (1989). The population genetics of bacteria. in *Genetics of Bacterial Diversity* (D.A. Hopwood and K.F. Chater, eds.). Academic Press pp. 415-438.

Chapter Three

GENETIC DIVERSITY IN NATURAL POPULATIONS OF *PROCHLOROCOCCUS MARINUS* IN THE SARGASSO SEA AND THE GULF STREAM

ABSTRACT

An investigation into the genetic structure of *Prochlorococcus marinus* field populations in depth profiles from the Sargasso Sea and the Gulf Stream revealed a high degree of genetic heterogeneity within water samples, detected by partial sequencing of cloned PCR products amplified from flow cytometrically sorted cells. Overlapping sets of alleles were recovered from the two water columns, from different depths within each water column and from flow cytometrically distinguishable subpopulations within water samples, suggesting that each of these populations drew its membership from a single gene pool.

INTRODUCTION

Prochlorococcus marinus is a unicellular prokaryote which makes up a high proportion of the photosynthetic picoplankton throughout wide regions of the tropical and subtropical marine environment (Chisholm et al 1988, Giskes et al. 1988, Olson et al. 1990, Vaulot et al. 1990, Li et al. 1993, Campbell and Vaulot 1993, Veldhuis and Kraay 1993). Sometimes referred to as a prochlorophyte (Lewin 1981, Chisholm et al. 1988), *P. marinus* is recognized by its tiny size (0.6 to 0.8 μm) and unique set of photosynthetic pigments, which include divinyl chlorophylls *a* and *b* (chl a_2 and b_2) and specifically do not include phycobilipigments, typical of most other cyanobacteria.

In efforts to understand the role of this organism in the marine microbial community, a number of recent studies have characterized the water column distribution of *P. marinus* in terms of cell number, biomass, productivity, chlorophyll fluorescence and DNA content at several locations (Giskes et al. 1988, Li and Wood 1988, Olson et al.

1990, Vaulot et al. 1990, Li et al. 1992, Vaulot and Partensky, 1992, Cambell and Vaulot 1993, Goericke and Repeta 1993, Goericke and Welschmeyer 1993, Vaulot et al. 1994) and over time (Olson et al. 1990, Cambell and Vaulot 1993). These studies have established that *P. marinus* is present throughout the euphotic zone at all times of the year and numerically dominates the picophytoplankton at oligotrophic, deep chlorophyll maxima during thermal stratification. Other, parallel studies have analyzed physiological and pigment differences among cultured *P. marinus* isolated from different depths and locations (Morel et al. 1993, Partensky et al. 1993, Moore et al. 1994), identifying differences in their growth rate response to varying levels of illumination which suggest that some genetic variants may have relatively greater growth rates in brightly lit surface waters, while others may have growth advantages at dimly lit depths (Moore et al. 1994).

Med4 and SS120, a pair of cloned isolates from 5 m and 120 m in the in the Mediterranean and the Sargasso Sea, respectively, are the most extensively characterized *P. marinus* strains (Partensky et al. 1993, Morel et al. 1993, Moore et al. 1994). Each can grow over a wide range of light intensities, but consistent with having been isolated at the surface, Med 4 is able to grow at high light intensities (>100 $\mu E\ m^{-2}\ s^{-1}$) at which SS120 is unable to grow, and SS120, isolated from a deep chlorophyll maximum (DCM) at 120 m, grows at low light intensities (2 - 6 $\mu E\ m^{-2}\ s^{-1}$) at which Med4 shows no growth (Moore et al. 1994). The two isolates differ in their compliment of photosynthetic pigments as well, with SS120 elaborating "normal" chlorophyll *b* (chl b_1) at light intensities above 20 $\mu E\ m^{-2}\ s^{-1}$ and exhibiting a tenfold higher chlorophyl *b* (chl b_1 + chl b_2) to chl a_2 ratio (chl b/a_2 ratio) than Med4 over the entire range of growth light intensities (0.05 to 0.15 for Med4 and 0.4 - 2.4 for SS120) (Morel et al. 1993, Partensky et al. 1993, Moore et al. 1994). Med4 contains no detectable chl b_1 under any conditions tested (Morel et al. 1993, Partensky et al. 1993, Moore et al. 1994).

Phylogenetic analyses using four gene sequences, 16S ribosomal RNA (rRNA), *psbB*, *petB/D* and *rpoC*, have shown that cultured strains of *P. marinus* fall into a single phylogenetic cluster which they share with the marine A *Synechococcus*, phycobilisome-containing cyanobacteria with which they also share their marine habitat (Swift and Palenik 1992, this thesis Chapter Two). *P. marinus* Mediterranean strain (Med4), derived from strain MED (Chisholm et al. 1992, Partensky et al. 1993) and Pacific strain (MIT9107), both isolated from mixed surface waters (Chisholm et al. 1992, J. Dusenberry, pers. comm.), were found to be closely related, while DCM culture SS120 and others fell into more deeply branching lineages (this thesis Chapter Two). There was no correlation between genetic distance and the geographic distance between sites of culture isolation, with two Sargasso Sea isolates exhibiting less sequence similarity than the Mediterranean and Pacific pair. Fractional sequence similarity among four *P. marinus* isolates ranged from 0.994 to 0.989 for 16S rRNA and from 0.973 to 0.934 at the first two codon positions of *petB* and *D* (0.838 to 0.788 at all codon positions). Intergenic region sequences between *petB* and *D* were similar for Med4 and MIT9107, while sequences from other cultures were so divergent that homologous nucleotide positions could not be identified (this thesis Chapter Two).

Data from HPLC and flow cytometric studies suggest that genetic differences may distinguish *P. marinus* populations at different depths in stratified water columns, and that different genetic variants may coexist within water samples. At sites in the Atlantic, Pacific and in the Red Sea, ratios of chl b/a_2 were found to be lower near the surface than at the DCM (Veldhuis and Kraay 1990, 1993, Cambell and Vaulot 1993, Goericke and Repeta 1993), with the range of chl b/a_2 ratios within a water column (0.15 to 2.9 in the Sargasso Sea) being greater than those observed in individual cultures (0.05 to 0.15 for Med4 and 0.4 - 2.4 for SS120) (Goericke and Repeta 1993). HPLC analysis of samples from from DCM's in the subtropical Pacific has shown that cells passing

through a 0.65 μm filter exhibit 1.7-fold lower ratios of chl b_2/a_2 than cells retained by the filter, suggesting that subpopulations with different pigment ratios may coexist in the water sample (Letelier et al. 1993). While the flow cytometric study provided data on fluorescence ratios for individual *P. marinus* cells excited at 457 and 488 nm, exploiting the different fluorescence excitation properties of chl a_2 and *b* at these wavelengths to estimate pigment ratios (Cambell and Vaulot 1993), the HPLC studies examined bulk chlorophyll, and so could have included chl *b* from cells other than *P. marinus*. Differences in chl b/a_2 ratios in these HPLC reports are therefore not direct evidence of genetic heterogeneity between *P. marinus* at different depths.

Flow cytometry of field samples from the Pacific, the Red Sea and the Sargasso Sea have occasionally identified "multiple populations," consisting of discrete distributions of chlorophyll autofluorescence and light scatter from individual cells (Cambell and Vaulot 1993, Veldhuis and Kraay 1993, B. Binder, R. Olson, J. Dusenberry, E. Zettler, unpublished observations). These flow cytometric subpopulations may derive from genetically distinct subpopulations which express different pigment phenotypes when exposed to uniform conditions, but they could also derive from samples containing recently mixed, genetically similar populations acclimated to different conditions of light or nutrient supply (Dusenberry and Chisholm in preparation).

Data consistent with the possibility of genetic heterogeneity within *P. marinus* populations at discrete depths have also come from molecular cloning experiments (Giovannoni et al. 1990, Britchgi and Giovannoni 1991, Schmidt et al. 1991, DeLong et al. 1993, Fuhrman et al. 1993). Investigations in which 16S ribosomal RNA gene sequences were cloned from uncharacterized marine picoplankton have recovered a number of sequences which fall into the *P. marinus*/marine A *Synechococcus* cluster.

While some of these sequences (*e.g.* SAR6 and SAR139) can be phylogenetically linked to cultured *P. marinus* or marine A *Synechococcus*, others (*e.g.* SAR7) form deep branches that are not robustly affiliated with either taxon, and so cannot be identified as *P. marinus* or *Synechococcus* with the data at hand (Urbach et al. 1992, this thesis Chapter Two). In order to assess the the genetic structure of *P. marinus* populations (as defined by their flow cytometric signature), therefore, it is necessary to physically separate *P. marinus* from other members of the community.

We report the results of a direct analysis of *P. marinus* populations using cells sorted by flow cytometry as starting material. *P. marinus* were functionally defined for this study as cells exhibiting the *P. marinus* flow cytometric "signature:" characteristic values for red fluorescence and forward light scatter (measured relative to standard 0.57 μm beads), in the absence of orange fluorescence, when excited by 488 nm laser light (Chisholm et al. 1988). Genotypes were sampled from sorted *P. marinus* by PCR amplification of the *petB/D* locus[1] with cloning and single run, partial sequencing of the resulting clones (Adams et al. 1991, Adams et al. 1992, Kahn et al 1992). Preliminary analysis of *petB/D* amplification products by direct sequencing revealed the presence of superimposed sequences differing mostly in the intergenic region, suggesting the presence of multiple *petB/D* alleles (Appendix I). 21 to 28 clones were analyzed for each of eight sorted samples from different depths in the Sargasso Sea and the Gulf Stream, and including subpopulations of two flow cytometrically defined "double populations" in the Gulf Stream depth profile. Questions addressed are whether *P. marinus* field populations at discrete depths are genetically homo- or heterogeneous, and whether genetic differences can be detected between populations at different depths, between

[1]The *petB/D* locus includes the 3' end of *petB*, an intergenic region, and the 5' end of *petD*, (this thesis Chapter Two). These genes encode subunits of the photosynthetic b_6/f complex and are single-copy, linked cistrons with a consistent orientation in all photosynthetic prokaryotes and organelles in the literature (Vermaas and Ikeuchi 1991, Greer and Golden 1992).

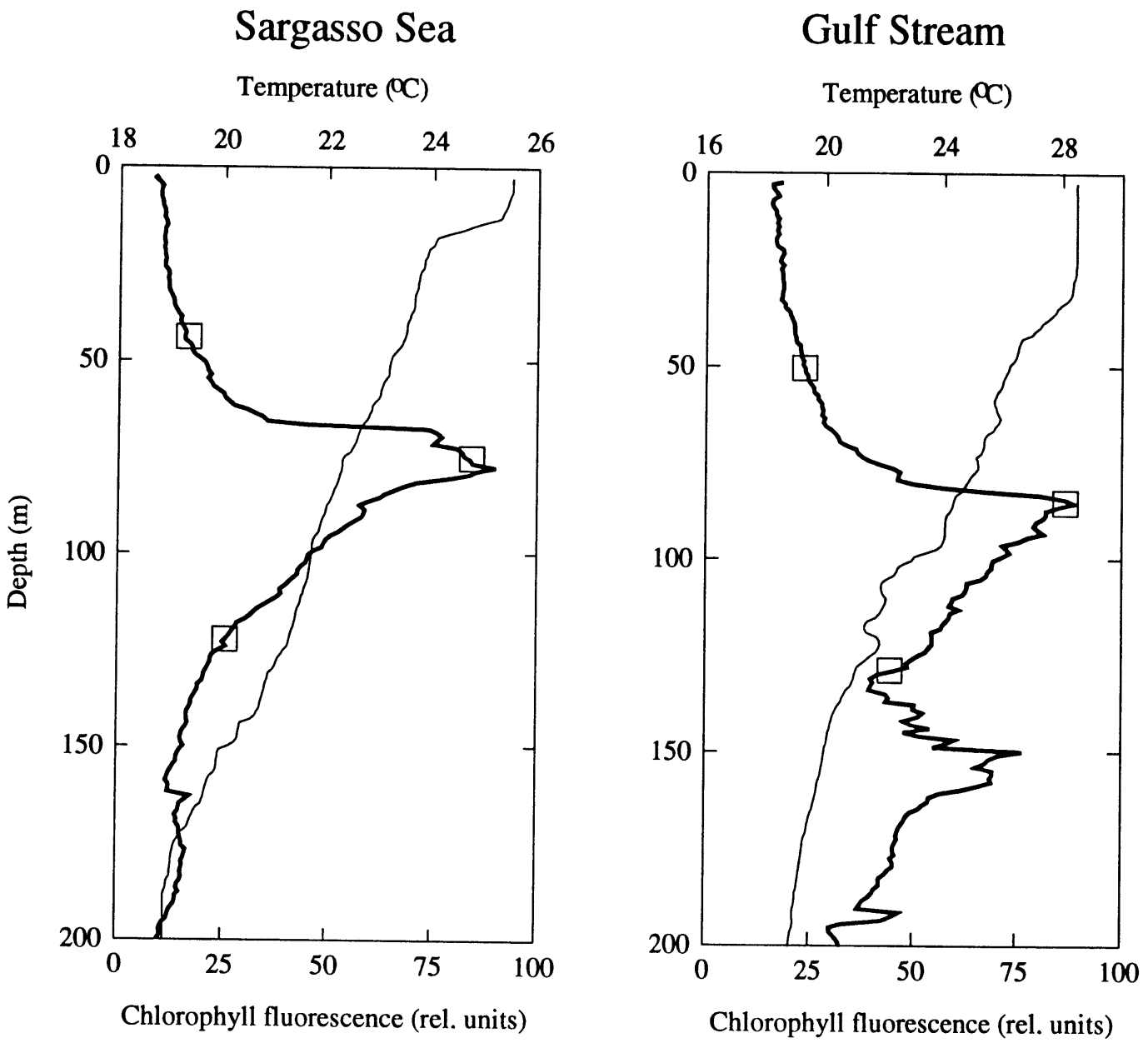
subpopulations in multiple populations (as differentiated by their fluorescence and forward light scatter per cell), and between water columns in adjacent, but hydrographically distinct water masses. Gene sequences obtained from field populations are compared to sequences from cultured cells.

METHODS

Characteristics of sampling sites and flow cytometric signatures. *P. marinus* were collected from depth profiles in the Sargasso Sea (33^{o} 58.65'N, 66^{o} 6.63'W, 10:00 local time, 11 July 1993) and the Gulf Stream (37^{o} 30.68'N, 68^{o} 13.69'W, 08:00 local time, 17 July 1993) during cruise 9306 of RV Columbus Iselin. The Sargasso Sea sampling site exhibited a typical summer hydrographic profile, with a chlorophyll-poor, mixed upper layer overlying thermally stratified waters and a pronounced DCM (as indicated by *in vivo* chlorophyll fluorescence measurements) at 70 m. The Gulf Stream profile was more complex, with relatively low concentrations of chlorophyll at the surface and a pair of subsurface chlorophyll maxima at 85 m and 150 m (Figure 1). Samples were collected at depths predicted to contain the most different *P. marinus* populations, avoiding samples at the immediate surface which are difficult to visualize using the sorting flow cell. Sargasso Sea samples were collected at 40 m, 70 m (the DCM) and below the DCM at 120 m. Gulf Stream samples were collected at 50 m, 85 m (the shallower DCM) and at 135 m, between the two DCM's (Figure 1).

Flow cytometric characteristics of *P. marinus* populations varied with depth according to patterns familiar from previous investigations (Olson et al. 1990): at each depth in the Sargasso Sea *P. marinus* populations formed a single cluster on two dimensional scatter plots of chlorophyll fluorescence versus forward light scatter for

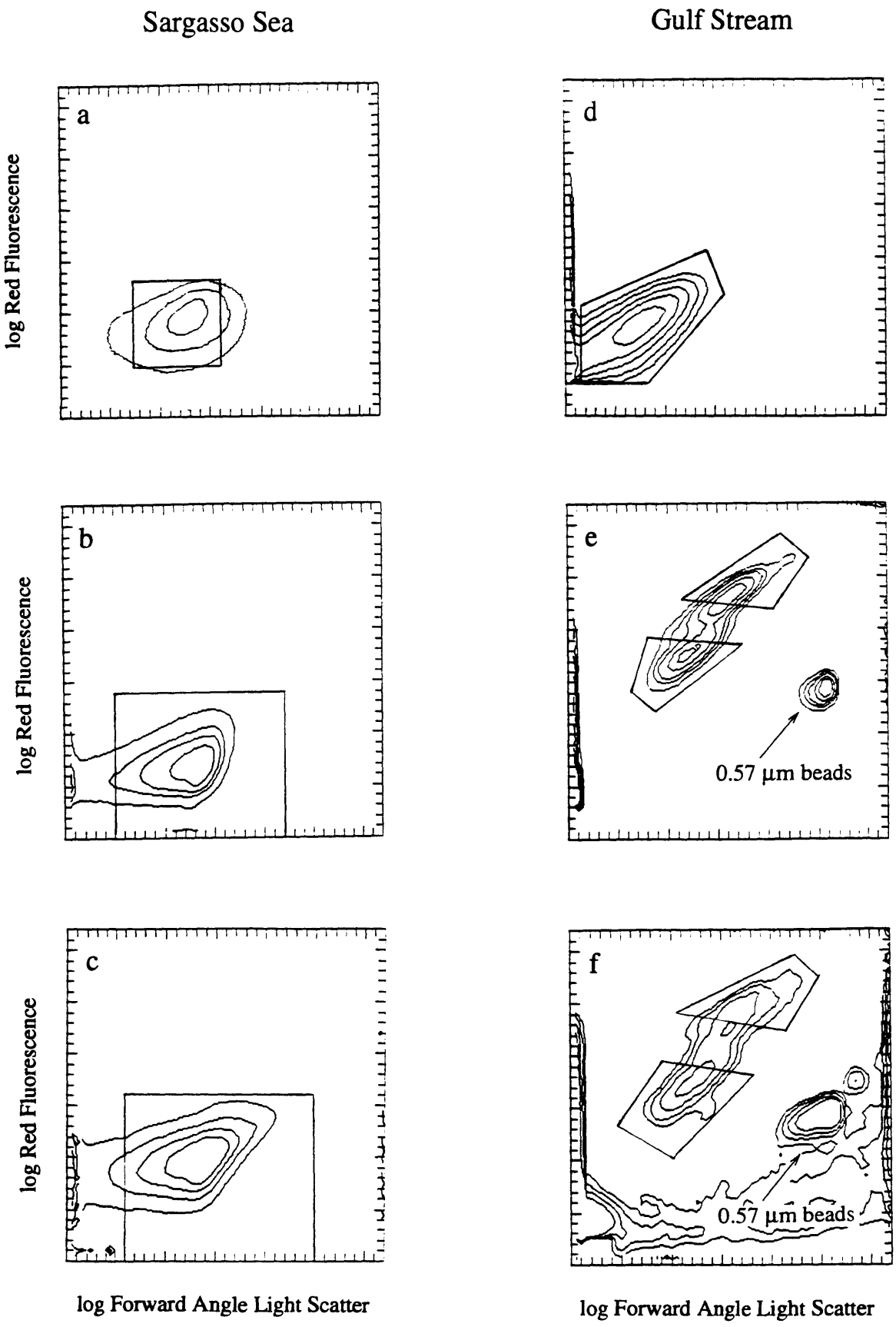
Figure 1. Depth profiles of temperature (thin lines) and *in vivo* chlorophyll fluorescence (heavy lines) at the Sargasso Sea (33^{o} 58.65'N, 66^{o} 6.63'W, 10:00 local time, 11 July 1993) and the Gulf Stream (37^{o} 30.68'N, 68^{o} 13.69'W, 08:00 local time, 17 July 1993) sampling sites. Open squares on the *in vivo* cholrophyll fluorescence plot mark depths at which samples were collected.


Sargasso Sea
Temperature (°C)
18
20
22
24
26
0
50
100
150
200
Depth (m)
0
25
50
75
100
Chlorophyll fluorescence (rel. units)
Gulf Stream
Temperature (°C)
16
20
24
28
0
50
100
150
200
0
25
50
75
100
Chlorophyll fluorescence (rel. units)

individual cells, with the distribution of values for each parameter approximating a lognormal distribution. Modes for chlorophyll fluorescence and forward light scatter showed an increase with depth, reflecting photoacclimative increase in chlorophyll per cell, in addition to possible genetic differences between populations at different depths. In the Gulf Stream, the *P. marinus* surface population again showed a unimodal distribution, but populations at 85m and 135m were double ("multiple") populations (Figure 2).

Sample concentration and flow cytometric sorting. Samples were collected using Niskin (30 l) or Go-Flo (10 or 30 l) bottles. Subsamples (300 to 750 ml) were concentrated to approximately 7 ml over 0.45 μm Durapore filters (Millipore) under light vacuum (5" Hg) and cells adhering to the filter resuspended by vigorous pipetting. *P. marinus* cell recoveries as measured by flow cytometry in similar concentrates were approximately 50%. Concentrates were immediately sorted by flow cytometry aboard ship using an Epics V flow cytometer (Coulter Electronics) or frozen in liquid nitrogen for sorting in the laboratory using an EPICS 753. Frozen cells were sorted over a period of about about 20 minutes immediately after thawing to avoid loss of autofluorescence, which can start to occur at about 30 minutes (Vaulot and Xiuren 1988, J. Dusenberry and R. Olson, unpublished observations). *P. marinus* populations were flow cytometrically defined according to their red fluorescence (660-700 nm) in the absence of orange fluorescence (540-630 nm) and by their characteristic forward light scatter (relative to 0.57 μm beads) when illuminated by blue laser light (488 nm) (Chisholm et al. 1988). Bitmap sort windows were drawn to separate *P. marinus* from other constituents of the photosynthetic planktonic community when *P. marinus* populations were unimodal, and to isolate members of flow cytometric subpopulations when multiple populations were observed. Approximately 10^6 *P. marinus* cells were collected in each sorted sample.

Figure 2. Contour plots for flow cytometric distributions of *P. marinus* sorted from Sargasso Sea (a-c) and Gulf Stream (d-e) field samples. Sorted (sub)populations used for genetic analysis are outlined. Red Fluorescence signals of the dimmly fluorescing cells in the 40 m Sargasso Sea sample (a) (processed in the laboratory) were enhanced by increasing the laser power to 1.2 watts PMT voltage to 1500 volts. Other samples (processed at sea) were processed using laser power of 1.0 watts and a PMT voltage of 1100 volts. To minimize the possibility of contaminaton standard beads were not added to samples used for sorting; distributions (e) and (f) are from reference samples. *P. marinus* cells were identified by their characteristic red vs. FALS signature and absence of orange phycoerythrin fluorescence (orange fluorescence data not shown). Sargasso Sea samples: (a) 40 m, (b) 70 m, (c) 120 m. Gulf Stream samples: (d) 50 m, (e) 85 m, (f) 135 m. Red Fluorescence: *in vivo* fluorescence of chlorophyll excited by 488 nm blue laser light. Forward Angle Light Scatter: forward angle light scatter of 488 nm laser light, a proxy for cell size.


Sargasso Sea
Gulf Stream
a
b
c
d
e
f
log Red Fluorescence
log Red Fluorescence
log Red Fluorescence
0.57 µm beads
0.57 µm beads
log Forward Angle Light Scatter
log Forward Angle Light Scatter

Isolation of genomic DNA. A modification of the protocol of Li et al. (1991) was used to prepare DNA from sorted cells. Cells from the flow cytometric sort were concentrated above 0.2 μm pore size Durapore membranes using Ultrafree-MC filter units (Millipore) spun at 2000 x g and washed twice with cell suspension buffer (0.5 M NaCl, 10 mM Tris [pH 8.0] 10 mM EDTA). Cells were resuspended in a final volume of 100 μl cell suspension buffer and lysed by addition of 12 μl 0.5 M DTT and 6 μl 10 M NaOH followed by a 10 minute incubation at 65°C. After addition of 120 μl neutralization buffer (90 mM Tris [pH 8.0] plus 0.5 mol/l HCl), DNA was precipitated by addition of 1 μl 20 mg/ml glycogen (Boehringer) and 0.6 ml cold 100% ethanol. Precipitates were stored at -20°C for several months. In the laboratory, crude DNA precipitates were collected by centrifugation at 16,000 x g for 30 min at 4°C, washed with 70% ethanol and resuspended in TE (10 mM Tris [pH 8.0] 1 mM EDTA). DNA was stored at -20°C.

Construction of the *petB/D* clone library. Portions of the *petB* and *D* genes and their intergenic region were amplified using degenerate oligonucleotide primers containing inosine (Knoth et al. 1988). Priming sites correspond to highly conserved regions in *petB* and *D* amino acid sequences, with forward primer PPETBD314 (5'ATGATGGTIYTIATGATGAT) corresponding to nucleotide positions 274 to 293 in the *Nostoc* PCC7906 *petB* coding sequence (Kallas et al. 1988), and reverse primer PPPETBD1160R (5'-CCRTARTARTTRTGICCCAT) corresponding to nucleotide positions 64 to 83 in the *Nostoc* PCC7906 *petD* sequence (Kallas et al. 1988). 100 μl PCR reactions contained DNA isolated from approximately 10^5 sorted cells, 5 U of Taq DNA polymerase (Promega) and a PCR reaction cocktail containing 1 x Mg-free PCR reaction buffer (50 mM KCl, 10 mM Tris-HCl [pH 9.0 at 25°C], 0.1% [wt/vol] Triton X-100), 1.5 mM $MgCl_2$, 200 μM (each) dATP, dCTP, dGTP and dTTP, 0.5 μM PPETBD314 and 1 μM PPETBD1160R. After initial denaturation for 2 min at 94°C,

thermal cycle parameters were 1 min 94°C, 1 min 47°C, 1 min 72°C (5 cycles), 1 min 94°C, 1 min 52°C, 1 min 72°C (10 cycles), 1 min 94°C, 1 min 57°C, 1 min 72°C (25 cycles), followed by 10 min 72°C. PCR products from pairs of replicate reactions were pooled and separated by electrophoresis in a 1% agarose gel. Products in the 400 to 800 basepair region of the gel were excised and purified with GeneClean (Bio 101). 1/30th to 1/1000th of each primary amplification product was reamplified for 30 cycles of 1 min 94°C, 1 min 57°C, 1 min 72°C, followed by 10 min 72°C, in reaction cocktail with 5 U Taq polymerase added during the first 57°C incubation. Control reactions lacking template DNA gave no amplication products for either primary amplifications or reamplifications, as judged from agarose gels.

Pooled reamplification products from pairs of replicate reactions were excised from a 1% agarose gel, purified with GeneClean and half to all of each product was inserted into pCRII vector (TA Cloning Kit, Invitrogen) according to the manufacturer's instructions. OneShot Competent cells were transformed using half of each ligation mixture. Both white and blue colonies were picked from LB plates (0.5% tryptone, 1% yeast extract, 0.5% NaCl, 1.5% Bacto-agar) containing 40 µg/ml X-gal and 50 µg/ml of either ampicillin or kanamycin. Plasmid DNA was isolated by InstaPrep (5prime-3prime) or standard alkaline lysis (Sambrook et al. 1989). Plasmids were digested with restriction endonuclease EcoRI and the digests analyzed by agarose gel electrophoresis. Clones containing 400 to 800 basepair inserts were chosen for sequence analysis.

DNA sequencing and database entry. Double-stranded plasmids containing PCR fragment inserts in random orientation were sequenced using Sequenase version 2.0 (USB) and primer PTASP6 (5'GATCCACTAGTAACGGCCG), complimentary to vector sequence flanking the pCRII cloning site. DNA sequences were entered into a sequence alignment editor (Olsen 1990), translated into amino acid sequences and aligned with a

petB/D sequence database according to conserved regions in the amino acid sequence (Widger and Cramer 1991, Greer and Golden 1992). In order to obtain sequences across the intergenic region for all clones recovered, clones giving sequence at the 5' (PPETBD314) end of the amplified *petB/D* fragment, and which did not match sequences already in the database, were sequenced from the opposite side of the pCRII cloning site using PTAT7 (5'GAGCGGCCGCCAGTGTGA), which was closer to the intergenic region. This procedure yielded some sequences spanning the entire length of the cloned fragment, while others covered only portions of the 3' end of *petB*, the intergenic region and the 5' end of *petD*. Sequences, which were determined in a single run and in one direction only (Adams et al 1991, Adams et al. 1992, Kahn et al. 1992), ranged from 71 to 562 basepairs in length. Clones with sequences which did not align with *petB* or *D* were eliminated from the analysis.

Sequence comparisons. Fractional sequence mismatch values were computed using the computer program of Olsen (1988, 1990). The phylogenetic tree was inferred using the Phylip version 3.4 neighbor-joining program (Felsenstein 1991) and Kimura genetic distance calculations (2:1 transition:transversion ratio) (Kimura 1980). To minimize effects of sequence determination and PCR errors, alignment gaps were omitted from sequence comparisons.

Population genetics calculations. Values for $E(S_n)$ were calculated using COMPAH90 (E.D. Gallagher, University of Massachusetts/Boston).

RESULTS

Sequence diversity in the combined dataset. The combined dataset for all eight sorted samples contained 191 clones with inserts homologous to *petB* and *D*. Among the 191 clone sequences, 71 alleles were identified as deriving from genetically distinct individuals in the sampled populations (Table 1).

The 71 alleles were identified according to a two step process, employing both qualitative and quantitative criteria to distinguish true genetic differences from sequence mismatches arising from PCR and sequence determination errors. In the first, qualitative step, sequence comparisons were used to identify 59 sets of clones among which intergenic regions varied in length from 19 to 91 basepairs, and in sequence to such a degree that homologous nucleotides could not be identified for use in sequence alignment. Differences among these 59 sets were deemed qualitatively different from PCR and sequence determination errors (which were nonetheless undoubtedly present). Most of the 59 sets of clones having unalignable intergenic regions contained single or identical clones, or clones exhibiting minimal differences attributable to error, but three sets contained clones having sufficient mismatch (up to 19.1%) to suggest that they had been amplified from genetically distinct, though related individuals. In the second, quantitative step, a criterion of 3.5% mismatch was applied as an upper bound for sequence differences considered indistinguishable from PCR plus sequence determination errors in order to subdivide the three sets of clones and add an additional

Table 1: Frequency of genetic alleles among cloned amplification products from flow cytometrically sorted samples.

Allele / Sample	G50	G85Br	G85D	G135Br	G135D	S40	S70	S120	TOTAL
P. marinus									
1 Allele1	17		3	2		12	14	1	49
2 Allele2	1								1
3 Allele3	1								1
4 Allele4			1						1
5 Allele5		5	17	16	5	1			44
6 Allele6		5	1		1			4	11
7 Allele7		2							2
8 Allele8		1							1
9 Allele9		1							1
10 Allele11		1							1
11 Allele12		1				1			2
12 Allele13		1							1
13 Allele14		1							1
14 Allele15		1							1
15 Allele16		1					1	1	3
16 Allele17		1					1		2
17 Allele19		1							1
18 Allele21			1						1
19 Allele22				2				1	3
20 Allele23				1					1
21 Allele25				1					1
22 Allele26					3				3
23 Allele27					2				2
24 Allele28					1				1
25 Allele29					1				1
26 Allele30					1				1
27 Allele31					1			1	2
28 Allele32					1				1
29 Allele33					1				1
30 Allele35					1				1
31 Allele36					1				1
32 Allele37					1				1
33 Allele38					1				1
34 Allele39						1			1
35 Allele41						1			1
36 Allele42							1		1
37 Allele43							1		1
38 Allele44							1		1
39 Allele45							1		1
40 Allele46							1		1
41 Allele47							1		1
42 Allele48							1	1	2
43 Allele50							1		1
44 Allele51							1		1
45 Allele52								1	1
46 Allele54								1	1
47 Allele56								1	1
48 Allele57								1	1
49 Allele58								1	1
50 Allele59								1	1
51 Allele60					1			1	2
52 Allele61								1	1
53 Allele62								1	1
54 Allele63								1	1
55 Allele64								1	1
56 Allele65		1						1	2
57 Allele66								1	1
58 Allele67								1	1
59 Allele68								1	1
60 Allele69				1					1
61 Allele71			1		1	1	1		4
62 Allele72	1								1
63 Allele73							1		1
64 Allele74							1		1
65 Allele75	2								2
66 Allele76	1								1
67 Allele77						1			1
68 Allele78						1			1
P. marinus Total	23	23	24	23	23	19	28	24	187
No. different Alleles	6	14	6	6	16	8	15	21	67
$E(S_{19})$	5.3	12.1	5.2	5.4	13.7	8.0	10.5	16.8	10.9
Chloroplast-like									
69 Allele20			1						1
70 Allele24				1					1
71 Allele40						2			2
TOTAL	23	23	25	24	23	21	28	24	191

Abbreviations: G50, 50m; G85Br and G85D, Bright and Dim clouds of 85m double population; G135Br and G135D, Bright and Dim clouds of 135m double population; all from the Gulf Stream depth profile. S40, 40m; S70, 70m; S120, 120m; all from the Sargasso Sea depth profile. $E(S_{19})$, estimated number of different alleles for a constant sample size of 19 clones.

12 alleles to the dataset[2]. The subdivision of these three sets of clones according to the quantitative criterion resulted in the total of 71 alleles (Table 2).

Pairwise comparisons among alleles identified by subdivision of sets of clones having similar intergenic regions show a mean fraction of sequence mismatch at third codon positions (fractional mismatch at third codon positions/3 x fractional mismatch at all codon positions) of 0.82, significantly different from the expectations of random PCR and sequence determination errors (this thesis Appendix II).

A prototype sequence (the longest and least ambiguous) was chosen to represent each allele in sequence comparisons and aligned with other sequences in the database according to conserved regions in amino acid translations (Figure 3, this thesis Appendix IV). Translations of prototype sequences were highly homologous to *petB* and *D* sequences already in the database, with all translations containing phylogenetically conserved amino acid residues, including histidines at positions 186 and 201 in the *Synechococcus* PCC7002 sequence, hypothesized to participate in heme binding (Carter et al. 1981, Widger and Cramer 1991, Hope 1993), when sequence was available for these positions. The most striking differences among alleles were in the intergenic region, which differered in length and sequence among alleles. In coding regions pairwise nucleotide comparisons among the 71 alleles showed that they differed by an average of 24.1% ±4.9% (Figure 4a), with most differences occurring at third codon positions which differed by 51.4% ±9.5% (Figure 4b). The more highly conserved first two codon postions differed by only 10.8% ±6.8% (Figure 4c). Since the overwhelming

[2]The 3.5% upper limit criterion for sequence mismatch among clones assigned to a single Allele was arrived at by an optimization protocol which considers the fraction of sequence mismatch at third (silent) codon positions for sequence comparisons along a scale of total sequence mismatch. Within the bounds of theoretical limits, the criterion divides the set of pairwise sequence comparisons within the three sets of alleles having similar intergenic regions into the two subsets having the greatest possible difference in their mean third codon position to all codon position mismatch ratios (this thesis Appendix II).

Table 2a. Fractional sequence mismatch at all sequence positions, including intergenic regions (x1000, below the diagonal) and fraction of sequence mismatch located at third codon positions (fractional mismatch at third positions/3 x fractional mismatch at all codon positions, for coding regions only) (x100, above the diagonal) for pairwise comparisons of prototype sequences for Alleles having intergenic regions alignable with Allele 1.

	1	2	3	4	5	6	7	8	9	10
1. Alle1	-	082	089	094	085	100	090	090	081	080
2. Alle2	126	-	092	083	079	080	078	080	078	073
3. Alle3	158	124	-	091	085	086	084	079	089	080
4. Alle69	057	133	160	-	084	100	083	085	083	078
5. Alle71	080	167	173	095	-	100	079	080	071	072
6. Alle72	059	148	134	051	057	-	077	080	071	059
7. Alle73	073	156	169	070	077	076	-	085	060	078
8. Alle74	148	166	190	153	181	154	143	-	071	084
9. Alle75	081	157	155	077	072	048	086	155	-	057
10. Alle76	119	154	191	129	180	147	142	142	164	-

Table 2b. Fractional sequence mismatch at all sequence positions, including intergenic regions (x1000, below the diagonal) and fraction of sequence mismatch located at third codon positions (fractional mismatch at third positions/3 x fractional mismatch at all codon positions, for coding regions only) (x100, above the diagonal) for pairwise comparisons of prototype sequences for Alleles having intergenic regions alignable with Allele 4.

	1	2	3	4
1. Alle4	-	102	101	079
2. Alle41	132	-	102	091
3. Alle77	042	138	-	079
4. Alle78	045	160	060	-

Table 2c. Fractional sequence mismatch at all sequence positions, including intergenic regions (x1000, below the diagonal) and fraction of sequence mismatch located at third codon positions (fractional mismatch at third positions/3 x fractional mismatch at all codon positions, for coding regions only) (x100, above the diagonal) for pairwise comparisons of prototype sequences for Alleles having intergenic regions alignable with Allele 17.

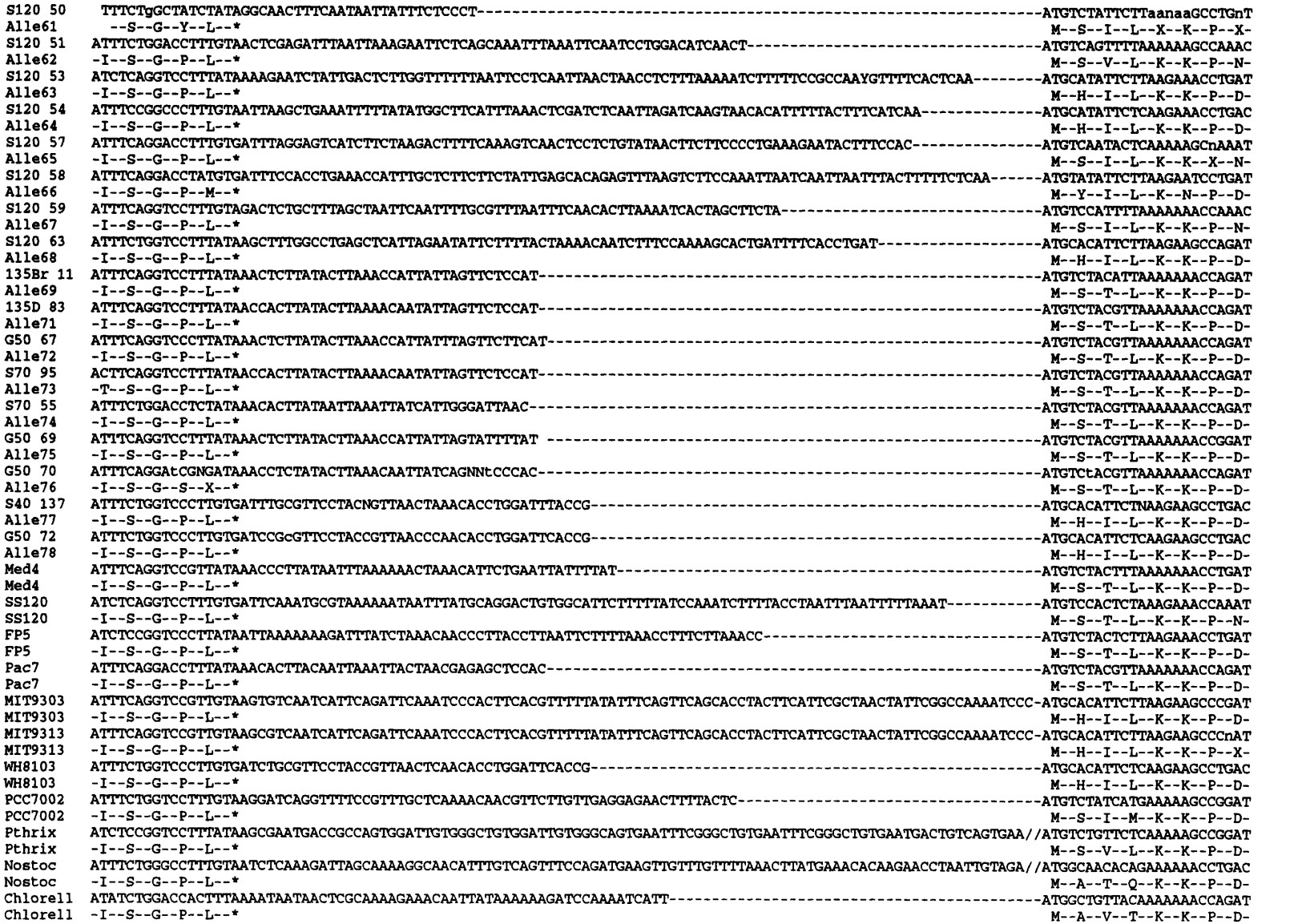
	1	2
1. ALLE17	-	071
2. ALLE31	075	-

Figure 3. Intergenic region sequences for 71 *petB/D* alleles recovered from the Sargasso Sea and Gulf Stream field samples (this work, Genbank accession numbers), cultured *P. marinus* strains Med4, SS120, FP5, MIT9107 (this thesis Chapter Two) and MIT9313 (this work, Genbank accession number), *Synechococcus* strains WH8103 (this thesis Chapter II) and PCC7002 (Brand et al. 1992), cyanobacteria *Prochlorothrix hollandica* (Greer and Golden 1991) and *Nostoc* PCC7906 (Kallas et al. 1988) and chloroplasts from *Chlorella protothecoides* (Reimann and Kueck 1989), *Marchantia polymorpha* (Ohyama et al. 1986) and *Zea maize* (Rock et al. 1987). A potion of the sequence alignment is shown, with the final six codons of *petB* at left and the initial eight codons of *petD* at right. Intergenic region sequences are arbitrarily represented as contiguous with *petB*, with alignment gaps inserted between the end of most intergenic region sequences and the beginning of *petD*. Intergenic region sequences for *P. hollandica*, *Nostoc* PCC7906, *C. protothecoides*, *M. polymorpha Z. maize* are truncated. Designations on nucleotide lines are clone names. Allele designations are indicated on the translation lines below.

Figure 3. Intergenic region sequences for *petB/D* alleles cloned from Sargasso Sea and Gulf Stream field samples. Designations on nucleotide lines are clone names. Allele designations are indicated on the translation lines below.

```
S40 96    ATTTCAGGCCCTTTATAAACTCTTATACTTAAACCAATACTAGTTCTCCAT----------ATGTCTACTTTAAAAAAACCAGAT
Alle1     -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
G50 90    ATTTCAGGTCCTCTATAAACTCCTATACTTTAATTATTAATTAGATTCAA----------ATGTCTACGTTAAAAAAACCAGAT
Alle2     -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
G50 107   ATTTCAGGtCCCTTGTAAACTTTTATACTTAAATCAATACAAGACTCTCTC----------ATGTCAACGTTAAAGAAACCAGAT
Alle3     -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
S40 128   ATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGGACTCACCG----------ATGCACATTCTCAAGAAGCCTGAC
Alle4     -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85Br 13   ATTTCTGGCCCCTTGTGATTTTCTTGATCGTTTTCTTTAGCTTGTCCCACTAAACTTAATTTCTAATCAATTCGCTATTTTTATCAA----------ATGCATATTTTAAAGAAACCTgaT
Alle5     -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85Br 100  ATTTCAGGTCCGTTGTAAGCGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTTCAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC-ATGCACATTCTTAAGAAGCTCGAT
Alle6     -I--S--G--P--L--*                                           M--H--I--L--K--K--L--D-
85Br 21   ATATCAGGACCTTTGTGATTGTTGTCACCTTTCTATAAGTTGTTTGAGCAACTTCTTTATTTAATCCTCACTTTTTAAATAAGACTAAGTTTCCT----------ATGTCAACACTCAAGAAGCCAAAT
Alle7     -I--S--G--P--L--*                                           M--S--T--L--K--K--P--N-
85Br 31   ATCTCTGGTCCCTTGTAGTTGTATTTATTTTCTAAAAACTAATTTTAAATTACGTTAAAAATTTTTCTTAAACAC----------ATGTCTACTCTTAAGAAACCAGAT
Alle8     -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
85Br 66   ATTTCAGGCCCTTTATGATTTAAAACTTTTATTCCCAGCCTTTTTCTTTAAAAGAGGTTTTTCTCCTTAATCACAATTCCCACTAGGCGCTGATTCAA----------ATGCACATTCTTAAAAAGCCAGaT
Alle9     -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85Br 106  ATTTCAGGACCTTTGTAAATTAAATTTGAGTCAGCATCTACTAGTTAGGGCTCTAATTCTCATCAAATTGTTTTATTCCTAATTTCTCTT----------ATGTCCACACTTAAAAAGCCAGAT
Alle11    -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
85Br111S  ATCTCAGGGCCTTTGTAGCCTGTGCTTTAACCTGAAAAATTCAATTTTCGAATTTGACTTCAATTTTTAAAATCGCTCCCTTGAA----------ATGTCCATTTTAAAAAAACCAAAC
Alle12    -I--S--G--P--L--*                                           M--S--I--L--K--K--P--N-
85Br 118  ATTTCCGGACCTTTATAAGAAAAATCTTTTAAAGCCTTTCCTTTTaATTAAAACTAATTTTGTTTCAGCTTCACCTACATACCCCAACCAACTT----------ATGTCAACACTAAAGAAGCCTGnT
Alle13    -I--S--G--P--L--*                                           M--S--T--L--K--K--P--X-
85Br 119  ATTTCAGGCCCTTTGTGAGCTGCCACAATTCAGCTTTCTTAATTCGATCTGATCTTTCCCCCAAGATTAATTTCTTTTCCA----------ATGCATATTCTTAAAAAGCCTGAT
Alle14    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85Br 122  ATATCTGGcCCCTTATGATTTCCGAATAGTTCTACTTGGTTAATTTTTTACAAACTCACTTTTTACAAAACTCTCTA----------ATGCACATTCTCAAGAAGCCAAAT
Alle15    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--N-
85Br 127  ATTTCCGGACCCTTATGAGCTTTAACTGTTCTTTTAAATTAAGTTGATCTGATTATTTCAATAGTCCAGTCCYGGTTCCCTACTT----------ATGCACATTCTCAAGAAGCCTGAT
Alle16    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85Br 129  ATATCAGGACCCCTGTAATTCATATCTCCAACAAAAAATTGCAACTTTTAATCCTCTGTT----------ATGTCGACTCTTAAAAAACCTGAT
Alle17    -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
85Br 135  ATATCAGGACCcTTATGAATTCGAAAAGGTTTTTTGCATCACAATCTTAATCATTTCTTGATGCCATCTTTTCCNNNTTAATTCACTTCAATAACCTTCTCTCCAC---ATGcACATACTTAAAAAGCCTGAT
Alle19    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
85D 68    ATTTCCGGTCCACTATAAGATTAAATAATAAAAAATT----------ATGTCTGTTATTAAGAAACCAGAT
Alle20    -I--S--G--P--L--*                                           M--S--V--I--K--K--P--D-
85D 90    ACTTCTGGTCCATTATAAAAATAAAAAAAAATCATTTTTCCTTATCTCATTAAATAA----------ATGTCCACTCTTAAGAAACCTGAT
Alle21    -T--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
S120 26   ATCTCAGGTCCATTGTGATATGAATTAATCAAGACAAGtTTAGAAATTTTTTTCATAGGTTTTTATTAACTCTTGCTCGACTTTTTATCCA----------ATGCATATTCTCAAGAAGCCCGAT
Alle22    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
135Br 52  ATATCAGGCCCCTTGTAATTCATATTTTCAAAAATGTATTACAACTTTTAATCCCCTATT----------ATGTCGACTCTCAAAAAACCTGAT
Alle23    -I--S--G--P--L--*                                           M--S--T--L--K--K--P--D-
135Br 53  ATCTCAGGTCCTCTTTAATTTAAAAGTTAAGAAATTAGACTCAATATTAATATCTGAATAAGGAATAGTT----------ATGTCAGTAATCAAAAAACCCGAT
Alle24    -I--S--G--P--L--*                                           M--S--V--I--K--K--P--D-
135Br 59  aTTTCTGGcccTTTGTAAGTASAATCACATTGCCTTGAATCATTCACtttTCtGTTGAgNNNNNNNNNANNNANNN----------ANNNNNANNNNAAAAAACCAGAT
Alle25    -I--S--G--P--L--*                                           X--X--X--X--K--K--P--D-
135D 80   ATTTCGGGACCCTTGTAGGCCTGTCACAATTCTCCGAACCTTTTTCTTTTAGTTCGGATTAATTACTTTTTCACAACGTTTTTAGTTCTCATCTG----------ATGCATATTCTTAAGAAGCCTGAT
Alle26    -I--S--G--P--L--*                                           M--H--I--L--K--K--P--D-
135D 95   ATTTCCGGCCCTTTATAGATCCTTTAGCTGCTTAATCTTTTATAGGATTTGGCTGGTCAATTTTCCTTAATCAATTTGATTAATTAAAAAAAAATTCGTCTT--------ATGTCCATCTTAAAGAAACCTGAT
Alle27    -I--S--G--P--L--*                                           M--S--I--L--K--K--P--D-
135D 81         GGACCTTTGTaAtTCtTGTCTGATCTCAAGAAATTTCTCCTCAAATCTCAATTAAATTCTTAAACAACTTAACCCT----------ATGTCCATTTTAAAAAAGCCAAAT
Alle28           G--P--L--*                                           M--S--I--L--K--K--P--N-
```

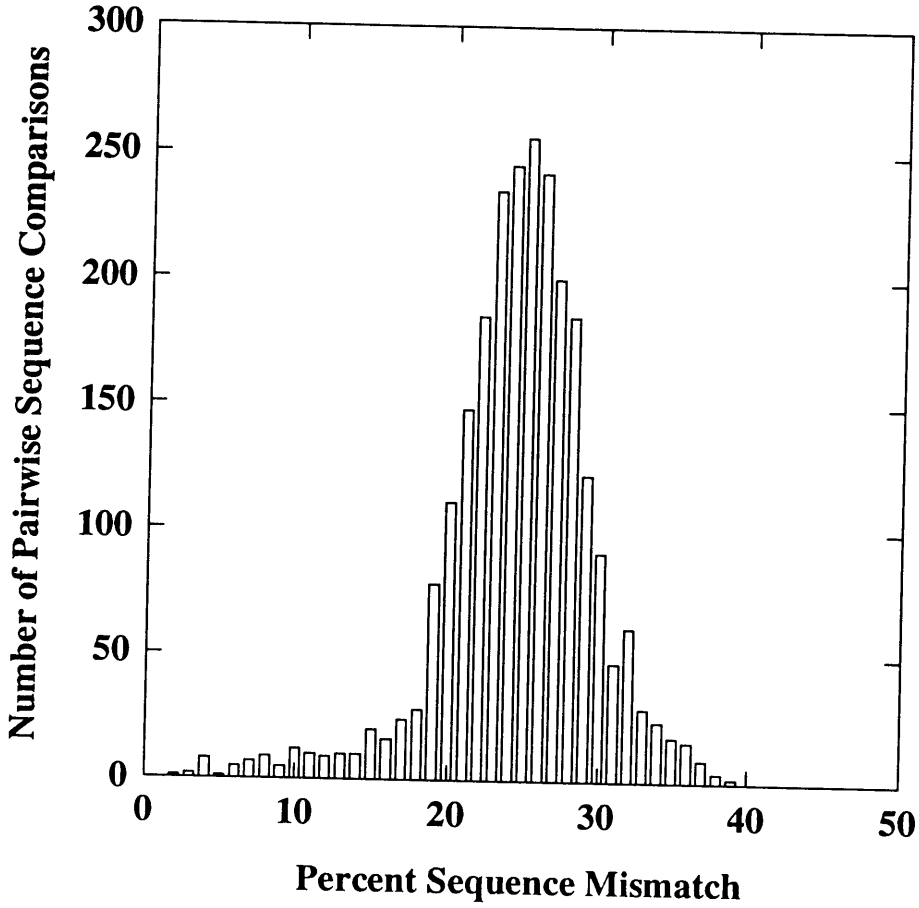
```
135D 84   AtTTCTGGTCCCTTATAGATTGTTTTTTATTTAGACTATTAATCTTCTTTATCTTTTTAATCGTCCCAAACCCTTTTTCTAACGGTTCCTGATCTA----------ATGCACATTCTCAAGAAGCCAGAT
Alle29    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
135D 87   ATTTCAGGACCTTTATAAATTTATTTAATTTTTCATATTTAGACCTTTTCCTCCAAAGCTTTTTAGTTTCAGTAATT------------------------------ATGTCAACTCTCAAAAAACCAGAT
Alle30    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
135D 91   ATATCAGGCCCTCTGTGACTCATATCTCAAACAATAAATTGCAACTTTTAATCCTCTGTT-----------------------------------------------ATGTCGACTCTTAAAAAACCnGAT
Alle31    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
135D 92   ATTTCAGGCCCTTTGTAATGCTTTATTAAATCCTCAATAATTTATTGTTAACCCTTCGACTCGCCTCAATCAACAGCTCCTTACAAGCCTCTCAAA----------ATGTCAGTTTTAAAAAAACCAGAT
Alle32    -I--S--G--P--L--*                                                                                          M--S--V--L--K--K--P--D-
135D 93   ATTTCGGGACCTTTGTAGCCCTCGACTAACCTGCTAATAGAGTTCACATCCAACCTAAATCCTTTAAAGAGACTCTTCACTT-------------------------ATGTCCATTTTAAGAAAGCCAAAT
Alle33    -I--S--G--P--L--*                                                                                          M--S--I--L--R--K--P--N-
135D 97   ATCTCTGGACCACTATGGGCTTTTATCTAAGTTTGTTTTCTAGGTGAGATCTTTGATCCCTTCCATTTTCTCTAATTTCTTCACTCTGACTTCAATTAAATCTC-----ATGCACATTCTTAAAAAGCCAGAT
Alle35    -I--S--G--P--L--W--A--F--I--*                                                                              M--H--I--L--K--K--P--D-
135D 110  ATTTCAGGACCTTTGTAAATTAAATTTGAGTCAGCATCTACTAGTTAGGGCTCTAATTCTCATCAAATTGTTTTATTCCTAATTTCTCTT-----------------ATGTCCACACTTAAAAAGCCAGAT
Alle36    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
135D 119  ATATCAGGTCCCTTGTAATTCATATTTCAAAAATAAATTACAACTCTTAATTCCTCTCTGTT-----------------------------------------ATGTCGACTCTTAAAAAACCTGAT
Alle37    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
135D 122  ATTTCAGGCCCATTATGAGNATTGGGTTAAGGCTTCTTAATCCCCCCTCTcGCTTTCTGTTAATTCATCCTTAAACAAAATCTAATTATTTAGAGTT----------ATGCACATTCTTAAGAAGCCCGAT
Alle38    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
S40 78    aTCTCTGGtCCTTTGTAATTTCTCTAAAAGCTATGCTTTCTGGCCTATCTATAGGCAACTTTCAATAATTATTTCTCCCT----------------------------ATGTCTATTCTTAAAAAGCCTGAT
Alle39    -I--S--G--P--L--*                                                                                          M--S--I--L--K--K--P--D-
S40 121   ATTTCCGGTCCTCTGTAACTTTAACTTCACAAATTTTTGTACCTTATTAACT---------------------------------------------------ATGTCTATTTTAAAAAAGCCCGAT
Alle40    -I--S--G--P--L--*                                                                                          M--S--I--L--K--K--P--D-
S40 117   ATTTCCGGTCCCTTGTGATCGTTCTCTCTACTCTTCAACCATCACCCGGAGTTTCTG-----------------------------------------------ATGCACATTCTCAAGAAGCCTGAT
Alle41    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
S70 43    ATTTCCGGACCTTTATAAGAAAATCTTTTAAAGCCTTTCCTTTTAATTAAAACTAATTTTGTTTCAGCTTCACCTACATACCCCAACCAACTT---------------ATGTCAACACTAAAGAAGCCTGaT
Alle42    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
S70 44    ACTTCAGGGCCTTTGTAGCCTGTGCTTTAACCTAATATTCCATTTGCGAATTTAAAACTTAAATTTTTAAAATCACTCCCTTCCA----------------------ATGTCCATTTTAAAAAAACCAAAC
Alle43    -T--S--G--P--L--*                                                                                          M--S--I--L--K--K--P--N-
S70 45    ATCTCAGGGCCTTTGTAAACAACACTAATTCAATAAGGATTTTTTCCTTTTCAAGCTTCATTAAAGCTTTGTTTTTAGTCAAAATCTAAATTTTCACCA---------ATGCACATTCTCAAGAAGCCAGAT
Alle44    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
S70 50    ATTTCTGGCCCTCTTTAAGTAAAATCAATAAGCTTGCTTGATTTATCGAGATTCTTTGATTCTTAACCCAAGATCCAAATCAAATT----------------------ATGACGACTCTAAAAAAACCTGAT
Alle45    -I--S--G--P--L--*                                                                                          M--T--T--L--K--K--P--D-
S70 53              TTTGTAACTTCATTATGAGGTGACTCATATATCTTCCTCACCCTTCTTTTACCCTTCGATTTTATTTCA---------------------------ATGAGTACTCTTAAAAAGCCTGAC
Alle46                L--*                                                                                           M--S--T--L--K--K--P--D-
S70 59    ATTTCAGGCCCTTTGTGAGCTGCAACAATTCAGCTTTCTTAATTCGATCTGATTTTTCCCCCAAGATTAATTTCTTTTCCA----------------------ATGCATATTCTTAAAAAGCCTGAT
Alle47    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
S70 63    ATTTCCGGTCCTTTATGAGTAACAATCCTTTCATCCAATAACTAAAACTGGTTTTTTAGAGCTTTAACCCAAGATCCCAAAACACTT----------------------ATGACGACACTAAAGAAACCTGAT
Alle48    -I--S--G--P--L--*                                                                                          M--T--T--L--K--K--P--D-
S70 66    ATTTCTGGNNCTTTATAACCGTTTTGATTTTTTAGGGACGAATTAAACCTCTGTTGATTCCATCCCAGATTCAATTCTCAAGAAACAATTACCTCTT-----------ATGGATGCTCTTAAAAAGCCTGAT
Alle50    -I--S--G--X--L--*                                                                                          M--D--A--L--K--K--P--D-
S70 72    ATTTCAGGAcCtTTGTAGGTTCTTCTTTGGCTTGAAACTTTCTTTCAATTTCAATTTTCGCTCTTTAAATCACTTACTCCTCAT----------------------ATGTCCATTTTAAAAAA
Alle51    -I--S--G--P--L--*                                                                                          M--S--I--L--K--
S120 4    ATATCTGGNNCTTTATGATTTGCGAATAGGTCTACTTGGTTAATTTTTACAAACTCACTTTTTACAAAACTCACTA----------------------------ATGCACATTCTCAAGAAGCCAAAT
Alle52    -I--S--G--X--L--*                                                                                          M--H--I--L--K--K--P--N-
S120 28   ATTTCTGGTCCATTGTGATTACTGTGAGATTCCACAATAAACAGGCACTTTGTGAATAACAATTTGCCTTAAACCTTTCCCAACCCAAAACTAGTTTCTACTTAT----ATGTCTACTCTCAAGAAACCTGAC
Alle54    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
S120 30   ATTTCAGGCCCGTTATAAGCCTCTCAGTTGGGTTTATTTTTGATTTCATTGACACTAGTTCCTAAAAGAATCTCTTTTTAGTT----------------------ATGACCACTCTCAAAAAGCCCGAC
Alle56    -I--S--G--P--L--*                                                                                          M--T--T--L--K--K--P--D-
S120 34   ATTTCAGGTCCGTTATAAACTTTTATAATTAAAACATTATTCGATTTCTAT-------------------------------------------------------ATGTCTACCTTAAAAAAACCAGAT
Alle57    -I--S--G--P--L--*                                                                                          M--S--T--L--K--K--P--D-
S120 45   CTTTCAGGACCTTTGTAAACAGAGGCAAAAGATTCCCTGTTTGCCAAGCcttttTTTGATTTTCTATTGTTTCAATAATCTTTTTTTTCTA----------------ATGAGCATTCCTAAGAAGCCTGAC
Alle58    -L--S--G--P--L--*                                                                                          M--S--I--P--K--K--P--D-
S120 46   ATTTCAGGTCCTTTATGACTTGAATTCGTTACCTAATTCTCCTTTTGTATCAGGATTGCTTCAAAAGGGTTTCCTTTGACTTTTCTCTCTTCCTTTAGGCTCTACTCAAATGCACATTCTTAAGAAGCCAGAT
Alle59    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
S120 49   ATCTCAGGGCCTTTGTGATACTTGAGAATATTTTTAAATTTTCTGGATTAATTCAAACATTACAAACAACTATCCTTTCATTCCCAA----------------------ATGCATATCCTCAAAAAGCCTGAT
Alle60    -I--S--G--P--L--*                                                                                          M--H--I--L--K--K--P--D-
```

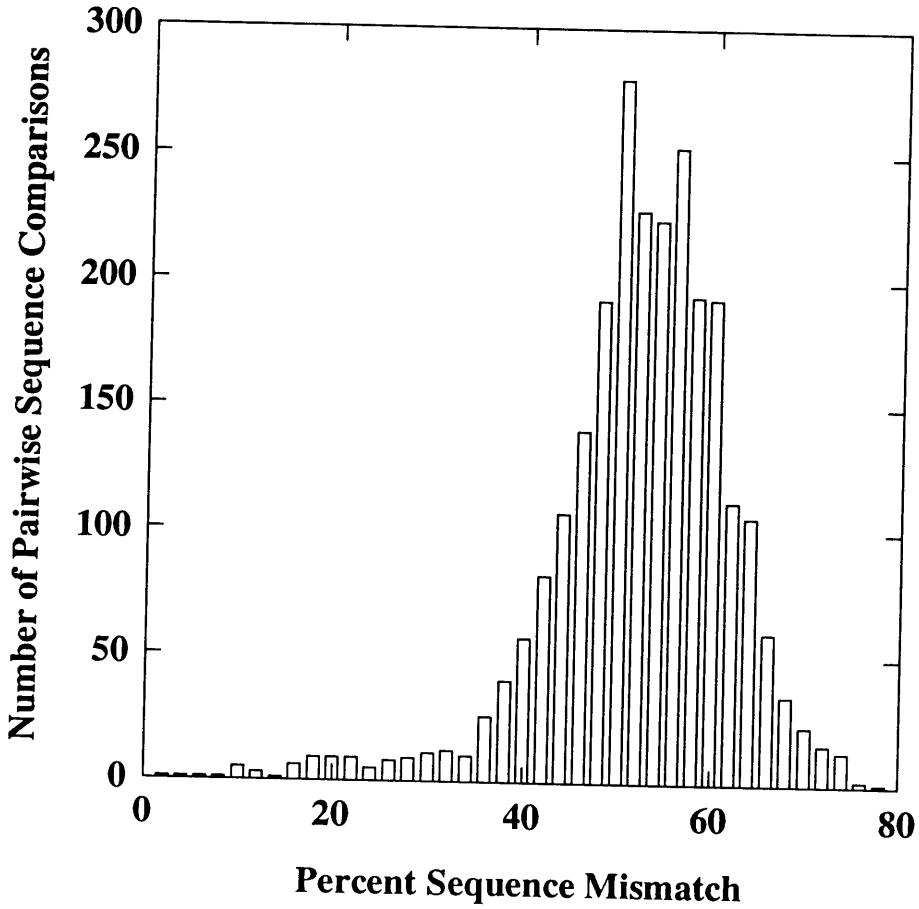

S120 50 TTTCTgGCTATCTATAGGCAACTTTCAATAATTATTTCTCCCT----------ATGTCTATTCTTaanaaGCCTGnT
Alle61 --S--G--Y--L--* M--S--I--L--X--K--P--X-
S120 51 ATTTCTGGACCTTTGTAACTCGAGATTTAATTAAAGAATTCTCAGCAAATTTAAATTCAATCCTGGACATCAACT----------ATGTCAGTTTTAAAAAAGCCAAAC
Alle62 -I--S--G--P--L--* M--S--V--L--K--K--P--N-
S120 53 ATCTCAGGTCCTTTATAAAAGAATCTATTGACTCTTGGTTTTTTAATTCCTCAATTAACTAACCTCTTTAAAAATCTTTTTCCGCCAAYGTTTTCACTCAA---------ATGCATATTCTTAAGAAACCTGAT
Alle63 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
S120 54 ATTTCCGGCCCTTTGTAATTAAGCTGAAATTTTTATATGGCTTCATTTAAACTCGATCTCAATTAGATCAAGTAACACATTTTTACTTTCATCAA----------ATGCATATTCTCAAGAAACCTGAC
Alle64 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
S120 57 ATTTCAGGACCTTTGTGATTTAGGAGTCATCTTCTAAGACTTTTCAAAGTCAACTCCTCTGTATAACTTCTTCCCCTGAAAGAATACTTTCCAC----------ATGTCAATACTCAAAAAGCnAAAT
Alle65 -I--S--G--P--L--* M--S--I--L--K--K--X--N-
S120 58 ATTTCAGGACCTATGTGATTTCCACCTGAAACCATTTGCTCTTCTTCTATTGAGCACAGAGTTTAAGTCTTCCAAATTAATCAATTAATTTACTTTTTCTCAA------ATGTATATTCTTAAGAATCCTGAT
Alle66 -I--S--G--P--M--* M--Y--I--L--K--N--P--D-
S120 59 ATTTCAGGTCCTTTGTAGACTCTGCTTTAGCTAATTCAATTTTGCGTTTAATTTCAACACTTAAAATCACTAGCTTCTA----------ATGTCCATTTTAAAAAAACCAAAC
Alle67 -I--S--G--P--L--* M--S--I--L--K--K--P--N-
S120 63 ATTTCTGGTCCTTTATAAGCTTTGGCCTGAGCTCATTAGAATATTCTTTTACTAAAACAATCTTTCCAAAAGCACTGATTTTCACCTGAT----------ATGCACATTCTTAAGAAGCCAGAT
Alle68 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
135Br 11 ATTTCAGGTCCTTTATAAACTCTTATACTTAAACCATTATTAGTTCTCCAT----------ATGTCTACATTAAAAAAACCAGAT
Alle69 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
135D 83 ATTTCAGGTCCTTTATAACCACTTATACTTAAAACAATATTAGTTCTCCAT----------ATGTCTACGTTAAAAAAACCAGAT
Alle71 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
G50 67 ATTTCAGGTCCCTTATAAACTCTTATACTTAAACCATTATTTAGTTCTTCAT----------ATGTCTACGTTAAAAAAACCAGAT
Alle72 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
S70 95 ACTTCAGGTCCTTTATAACCACTTATACTTAAAACAATATTAGTTCTCCAT----------ATGTCTACGTTAAAAAAACCAGAT
Alle73 -T--S--G--P--L--* M--S--T--L--K--K--P--D-
S70 55 ATTTCTGGACCTCTATAAACACTTATAATTAAATTATCATTGGGATTAAC----------ATGTCTACGTTAAAAAAACCAGAT
Alle74 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
G50 69 ATTTCAGGTCCTTTATAAACTCTTATACTTAAACCATTATTAGTATTTTAT ----------ATGTCTACGTTAAAAAAACCGGAT
Alle75 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
G50 70 ATTTCAGGAtCGNGATAAACCTCTATACTTAAACAATTATCAGNNtCCCAC----------ATGTCtACGTTAAAAAAACCAGAT
Alle76 -I--S--G--S--X--* M--S--T--L--K--K--P--D-
S40 137 ATTTCTGGTCCCTTGTGATTTGCGTTCCTACNGTTAACTAAACACCTGGATTTACCG----------ATGCACATTCTNAAGAAGCCTGAC
Alle77 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
G50 72 ATTTCTGGTCCCTTGTGATCCGcGTTCCTACCGTTAACCCAACACCTGGATTCACCG----------ATGCACATTCTCAAGAAGCCTGAC
Alle78 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
Med4 ATTTCAGGTCCGTTATAAACCCTTATAATTTAAAAAACTAAACATTCTGAATTATTTTTAT----------ATGTCTACTTTAAAAAAACCTGAT
Med4 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
SS120 ATCTCAGGTCCTTTGTGATTCAAATGCGTAAAAAATAATTTATGCAGGACTGTGGCATTCTTTTTATCCAAATCTTTTACCTAATTTAATTTTTAAAT----------ATGTCCACTCTAAAGAAACCAAAT
SS120 -I--S--G--P--L--* M--S--T--L--K--K--P--N-
FP5 ATCTCCGGTCCCTTATAATTAAAAAAAGATTTATCTAAACAACCCTTACCTTAATTCTTTTAAACCTTTCTTAAACC----------ATGTCTACTCTTAAGAAACCTGAT
FP5 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
Pac7 ATTTCAGGACCTTTATAAACACTTACAATTAAATTACTAACGAGAGCTCCAC----------ATGTCTACGTTAAAAAAACCAGAT
Pac7 -I--S--G--P--L--* M--S--T--L--K--K--P--D-
MIT9303 ATTTCAGGTCCGTTGTAAGTGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTTCAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC-ATGCACATTCTTAAGAAGCCCGAT
MIT9303 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
MIT9313 ATTTCAGGTCCGTTGTAAGCGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTTCAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC-ATGCACATTCTTAAGAAGCCCnAT
MIT9313 -I--S--G--P--L--* M--H--I--L--K--K--P--X-
WH8103 ATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGGATTCACCG----------ATGCACATTCTCAAGAAGCCTGAC
WH8103 -I--S--G--P--L--* M--H--I--L--K--K--P--D-
PCC7002 ATTTCTGGTCCTTTGTAAGGATCAGGTTTTCCGTTTGCTCAAAACAACGTTCTTGTTGAGGAGAACTTTTACTC----------ATGTCTATCATGAAAAAGCCGGAT
PCC7002 -I--S--G--P--L--* M--S--I--M--K--K--P--D-
Pthrix ATCTCCGGTCCTTTATAAGCGAATGACCGCCAGTGGATTGTGGGCTGTGGATTGTGGGCAGTGAATTTCGGGCTGTGAATTTCGGGCTGTGAATGACTGTCAGTGAA//ATGTCTGTTCTCAAAAAGCCGGAT
Pthrix -I--S--G--P--L--* M--S--V--L--K--K--P--D-
Nostoc ATTTCTGGGCCTTTGTAATCTCAAAGATTAGCAAAAGGCAACATTTGTCAGTTTCCAGATGAAGTTGTTTGTTTTAAACTTATGAAACACAAGAACCTAATTGTAGA//ATGGCAACACAGAAAAAACCTGAC
Nostoc -I--S--G--P--L--* M--A--T--Q--K--K--P--D-
Chlorell ATATCTGGACCACTTTAAAATAATAACTCGCAAAAGAAACAATTATAAAAAAGATCCAAAATCATT----------ATGGCTGTTACAAAAAAACCAGAT
Chlorell -I--S--G--P--L--* M--A--V--T--K--K--P--D-

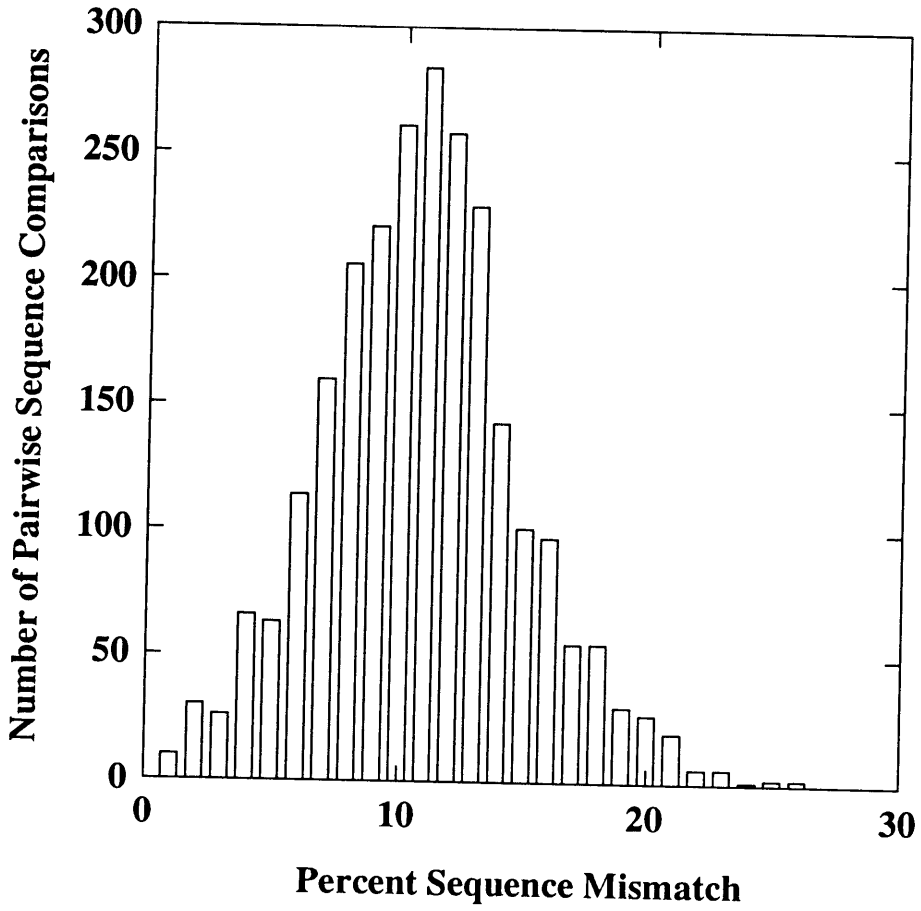
```
Liverwor  ATTTCAGGTCCGTTATAAATTACGTAAATTTATTACAAAATAAAAAAGTTTAAATACTAATTTCATTGCCATATTTTATGGATTTTTTTTTTCTATTTGAAAGTTCT//ATGGGAGTAACAAAAAAACCTGAT
Liverwor  -I--S--G--P--L--*                                                                                                                  M--G--V--T--K--K--P--D-
Maize     ATTTCGGGTCCGTTATAGGGAAGGCATAGCATAGAGAATTCTAATTCTCATATATCATATCGGGTAGGTTGTGGTATTTCATTGCTACAAACATGGGTTATTGTAAA//ATGGGAGTAACAAAGAAACCTGAC
Maize     -I--S--G--P--L--*                                                                                                                  M--G--V--T--K--K--P--D-
```

Abbreviations: Med4, MIT9107, FP5, SS120, MIT9303, MIT9313: *P. marinus* strains; WH8103, PCC7002: *Synechococcus* strains; Pthrix: *Prochlorothrix hollandica*; Nostoc: *Nostoc* PCC7906; Chlorell: *Chlorella protothecoides*; Liverwor: *Marchantia polymorpha*; Maize: *Zea maize*.

Figure 4. Frequency distributions for nucleotide mismatch in all pairwise comparisons of allele prototype sequences. The number of nucleotides in each comparison varies according to the lengths of the different sequences. a) all codon positions, mean 24.1%, std. dev. 4.9%. b) third codon positions, mean 51.4%, std. dev. 9.5%. c) first two codon positions, mean 10.8%, std dev. 6.8%.


Number of Pairwise Sequence Comparisons
300
250
200
150
100
50
0
0
10
20
30
40
50
Percent Sequence Mismatch


Number of Pairwise Sequence Comparisons
300
250
200
150
100
50
0
0
20
40
60
80
Percent Sequence Mismatch

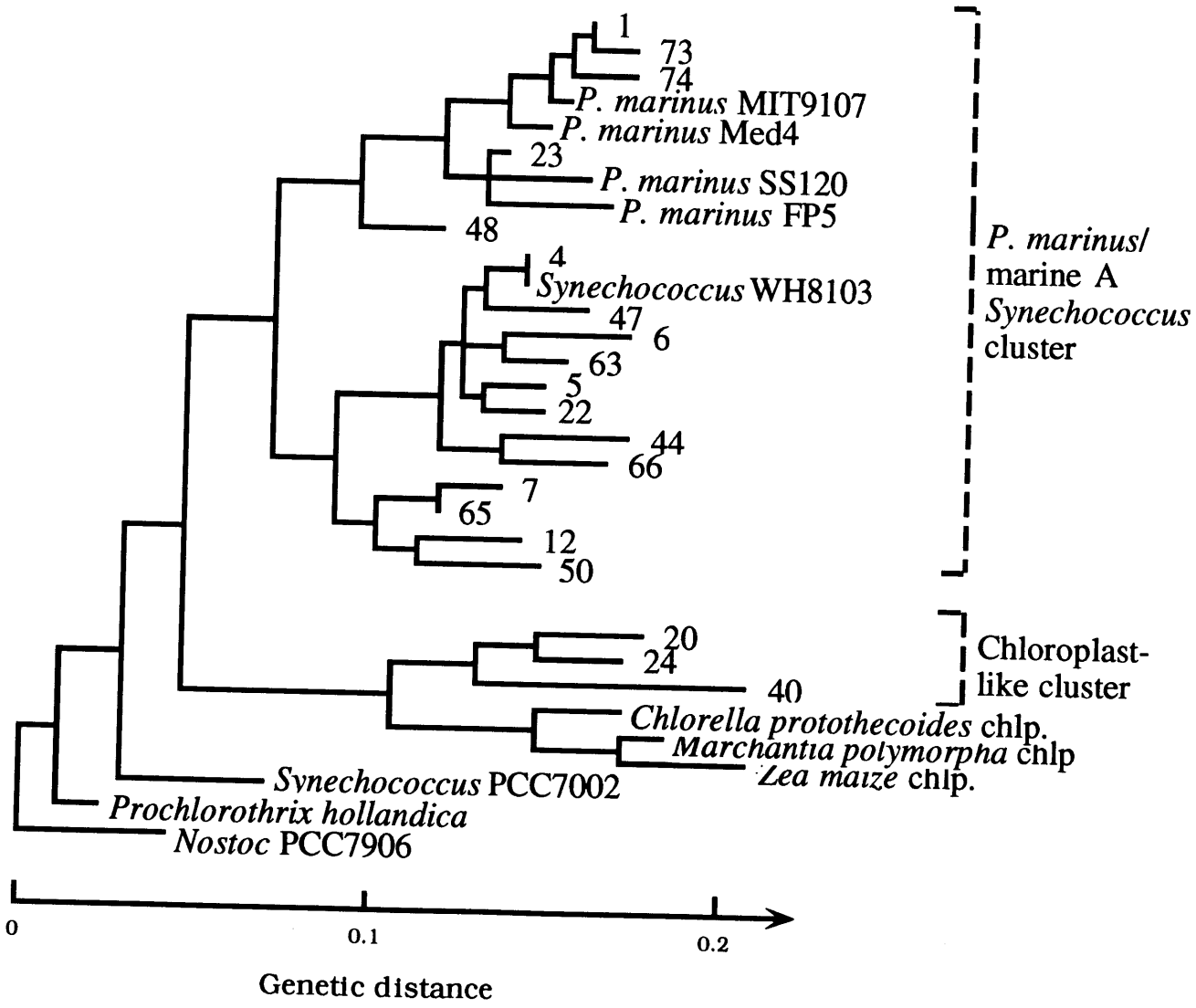

Number of Pairwise Sequence Comparisons
300
250
200
150
100
50
0
0
10
20
30
Percent Sequence Mismatch

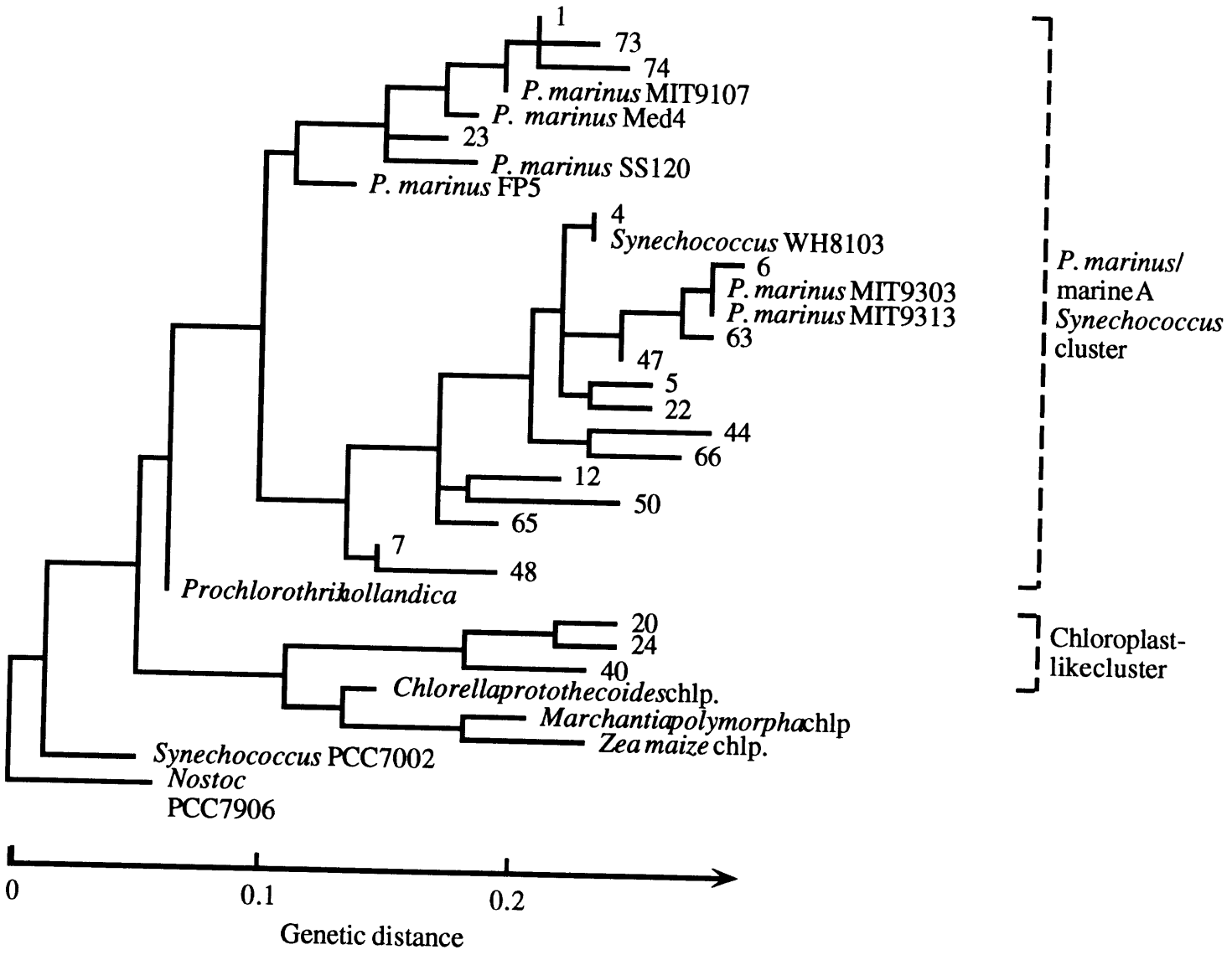
majority of sequence differences among alleles are located at evolutionarily unconstrained sites, most mismatches among alleles are likely to be due to evolutionary processes. Thus, even though some prototype sequences undoubtedly contain errors, the alleles named in this study serve the purpose of identifying genotypes present in the natural populations.

Phylogenetic affinities of the cloned sequences. Neighbor-joining analysis including twenty of the longest prototype sequences revealed that most alleles belong to the *P. marinus*/marine A *Synechococcus* cluster, within which *P. marinus* and marine A *Synechococcus* lineages are not well resolved (Figure 5) (this thesis Chapter Two). The presumption that alleles branching with *Synechococcus* WH8103 in the tree actually derive from correctly sorted *P. marinus* cells, and not from contaminating *Synechococcus*, is reinforced by the observation that Type 6, an allele which branches close to *Synechococcus* WH8103, is highly similar to *petB/D* sequences from cultured *P. marinus* MIT9313 (99.7% sequence identity over 260 bp) and *P. marinus* MIT9303 (98.3% identity over 272 bp) (Figure 5b). The possibility remains, however, that some sequences recovered in this study do, in fact, come from mis-sorted *Synechococcus*. To reflect this possibility, alleles which fall into the *P. marinus*/marine A *Synechococcus* lineage should be considered as presumptive *P. marinus* alleles.

Three alleles, numbers 20, 24 and 40, do not cluster with *P. marinus*/marine A *Synechococcus*, but instead form a phylogenetic group branching below the divergence of chlorophyte and land plant chloroplasts (Figure 5). The position of this cluster is reminiscent of the phylogenetic branch point for rhodophyte, pheophyte and some green algal chloroplasts in 5S rRNA, 16S rRNA and *psbA* phylogenies (Van den Eynde et al. 1988, Maid et al. 1989, Douglas and Turner 1991, Markowicz and Loiseaux-de Goer 1991, Oyaizu et al. 1993). Thus, these alleles may have originated in plastids of

Figure 5. Phylogenetic relationships among 20 *petB/D* alleles cloned from flow cytometrically sorted populations in the Sargasso Sea and the Gulf Stream, cultured strains of *P. marinus* and other members of the oxygenic phototroph radiation. (a) tree inferred by neighbor joining using 163 nucleotides at the first two codon positions of *petB* and *petD*. Alleles 20, 24 and 40 form a cluster allied with green plant and algal chloroplasts. The remaining clones cluster with *P. marinus* and marine A *Synechococcus* WH8103. (b) tree inferred from 77 nucleotides at first two codon positions in *petB* and *D* to include short sequences from *P. marinus* cultured strains MIT9303 and MIT9313. Allele 6 is 99.7% identical to *P. marinus* MIT9313 and 98.3% identical to *P. marinus* MIT9303 for all sequence positions.


1
73
74
P. marinus MIT9107
P. marinus Med4
23
P. marinus SS120
P. marinus FP5
48
4
Synechococcus WH8103
47
6
63
5
22
44
66
7
65
12
50
P. marinus/ marine A Synechococcus cluster
20
24
40
Chloroplast-like cluster
Chlorella protothecoides chlp.
Marchantia polymorpha chlp
Zea maize chlp.
Synechococcus PCC7002
Prochlorothrix hollandica
Nostoc PCC7906
0
0.1
0.2
Genetic distance


1
73
74
P. marinus MIT9107
P. marinus Med4
23
P. marinus SS120
P. marinus FP5
4
Synechococcus WH8103
6
P. marinus MIT9303
P. marinus MIT9313
63
47
5
22
44
66
12
50
65
7
48
Prochlorothrix hollandica
20
24
40
Chlorella prototheocoides chlp.
Marchantia polymorpha chlp
Zea maize chlp.
Synechococcus PCC7002
Nostoc
PCC7906
0
0.1
0.2
Genetic distance
P. marinus/
marine A
Synechococcus
cluster
Chloroplast-
like cluster

cryptophytes or coccolithophorids, both of which were present at the Gulf Stream and Sargasso Sea stations (R. Olsen, pers. comm.), or perhaps in plastids of the newly discovered marine chlorophyte *Ostreococcus tauri* (Courties et al. 1994), which may have been liberated during sample concentration. Liberated choroplasts lacking phycoerythrin would match the flow cytometric definition of *P. marinus* used for cell sorting and could therefore have been included in the sorted samples without machine error. The three chloroplast-like alleles were omitted from the population genetic analysis, leaving 68 *P. marinus* alleles in the dataset..

Phylogenetic affiliations of alleles not included in the neighbor-joining tree (because of short sequences) were assessed by comparison of mismatch frequencies between these sequences and selected sequences used to infer the tree (Table 3). All alleles not included in the tree were more similar to at least one member of the *P. marinus*/marine A *Synechococcus* cluster than to sequences outside the cluster and so were included in the population genetic analysis.

Allele counts for *P. marinus* in sorted samples. Clones recovered from each of the eight sorted samples contained between 6 and 21 *P. marinus petB/D* alleles (Table 1). Correcting for unequal sample size by the method of Hurlbert (1971), the estimated number of alleles present in a sample size of 19 clones for each sorted sample varied from 5.16 to 16.83 (Table 1).

To assess the completeness of sampling for genetic diversity, rarefaction curves were generated for each sorted sample and for lumped samples containing both members of flow cytometric double populations, constituents of entire water columns or the entire dataset (Figure 6). For a sample which contains a good representation of the genetic diversity in its parent population, a plot of $E(S_n)$ (the estimated number of genetic

Table 3. Fractional mismatch (x1000) at first two codon positions for all alleles not included in the phylogenetic tree (Figure 5) with sequences from cultured organisms, chloroplasts and selected alleles included in the tree.

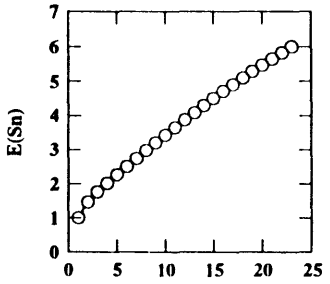
Allele		Med4		FP5		Alle12		WH8103		Pthrix		Chlorel		Maize
	#pos		MIT9107		SS120		Alle20		PCC7002		Nostoc		Liver	
2	187	032	027	075	086	069	176	118	134	123	139	139	160	166
3	157	032	025	083	070	069	159	127	121	102	121	127	140	159
8	175	057	063	023	086	078	223	080	126	131	177	177	200	183
9	97	124	134	113	134	104	227	052	103	093	134	186	175	175
11	118	051	059	068	085	081	203	093	119	102	144	178	186	169
13	52	096	058	115	135	074	135	135	115	115	096	173	192	231
14	84	107	143	095	119	098	238	036	143	107	179	179	190	155
15	198	091	081	096	066	086	157	056	086	081	121	141	136	141
16	194	088	098	088	082	081	160	046	108	088	134	139	119	129
17	80	063	100	075	075	125	200	125	175	112	162	175	250	225
19	64	125	125	156	141	098	203	031	141	125	172	172	188	172
21	141	092	092	099	071	097	213	092	121	113	163	184	206	199
25	46	022	022	087	087	097	130	065	087	065	065	043	065	109
26	98	122	082	122	112	103	184	071	112	082	122	153	173	204
27	116	052	078	086	112	083	198	095	103	095	147	147	164	147
28	32	156	125	188	188	020	188	219	156	188	250	281	313	313
29	38	184	158	184	211	107	132	079	132	158	158	237	211	184
30	118	093	085	068	093	081	178	102	110	110	153	186	186	178
31	152	099	099	092	053	109	224	092	158	145	198	198	237	231
32	112	080	098	107	107	077	188	098	125	088	179	170	179	152
33	106	113	113	132	132	024	217	104	113	113	179	189	198	179
35	92	109	120	109	109	136	217	065	174	120	185	152	207	196
36	102	078	098	118	098	100	176	127	147	118	157	196	225	206
37	93	054	086	065	065	117	204	108	151	097	161	161	204	183
38	47	128	149	128	106	125	170	043	149	106	170	170	170	170
39	167	060	060	078	079	043	126	030	066	054	102	108	120	114
41	179	089	089	089	102	092	184	006	101	084	128	140	162	145
42	167	054	054	072	079	054	132	066	108	078	078	114	114	138
43	64	094	094	141	141	012	156	125	109	109	172	172	188	172
45	145	117	090	103	090	107	200	090	131	145	159	179	214	214
46	116	103	112	086	105	105	155	138	172	138	147	172	164	164
51	58	086	103	086	138	000	224	103	086	086	155	207	207	190

Table 3, Continued

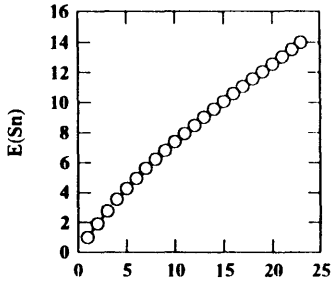
Allele		Med4		FP5		Alle12		WH8103		Pthrix		Chlorel		Maize
	#pos		MIT9107		SS120		Alle20		PCC7002		Nostoc		Liver	
52	82	134	146	159	122	110	207	098	134	110	159	195	183	183
54	78	077	090	090	103	094	231	077	141	103	128	205	244	205
56	97	052	093	093	082	104	196	113	144	093	144	165	196	175
57	100	040	020	120	110	093	190	120	110	110	140	170	190	200
58	42	190	167	190	214	102	190	143	167	167	167	190	167	167
59	217	088	092	088	084	090	180	074	120	120	129	157	157	157
60	178	107	079	101	097	103	169	045	107	090	112	135	140	157
61	34	206	176	206	235	118	147	206	176	176	235	235	324	324
62	73	123	123	110	123	044	192	123	151	096	178	192	233	219
64	92	098	109	076	141	082	228	033	109	076	141	174	185	152
67	108	102	102	120	120	024	204	093	102	102	167	176	185	167
68	94	106	096	117	138	081	202	053	128	096	138	160	181	170
69	217	037	023	074	088	068	194	115	138	143	152	157	180	180
71	175	040	034	086	087	089	154	114	131	120	137	154	154	171
72	116	034	017	095	103	105	181	112	112	121	147	147	164	190
75	99	040	040	091	111	111	182	131	121	101	152	141	162	192
76	187	032	011	059	054	047	160	086	102	107	128	128	166	150
77	195	077	077	082	088	076	174	005	087	072	123	133	149	138
78	76	145	145	105	145	128	263	013	171	132	171	211	250	224

Alleles not included in the neighbor-joining tree were compared on a pairwise basis with *P. marinus* cultured isolates Med4 (Med4), MIT9107 (MIT9107), FP5 (FP5) and SS120 (SS120), alleles 20 (Alle20) and 12 (Alle12), *Synechococcus* WH8103 (WH8103) and PCC7002 (PCC7002), *Nostoc* PCC7906 (Nostoc) and chloroplasts of *Chlorella protothecoides* (Chlorel), *Marchantia polymorpha* (Liver) and *Zea maize* (Maize). All comparisons in a row were based on the same set of aligned nucleotide positions, the number of which appear in column 2 (#pos).

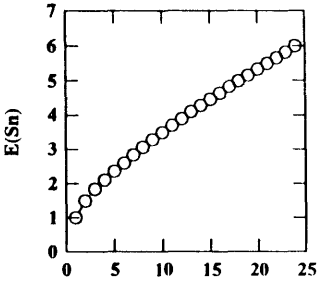
Figure 6. Rarefaction curves for clones recovered from a) individual sorted samples, b) lumped samples containing both members of flow cytometric double populations or constituents of entire water columns and c) a lumped sample containing the entire dataset. The upward trend of all curves indicates that the number of alleles in all populations are incompletely sampled. $E(S_n)$ (the estimated number of alleles for n, the number of clones in a random subsample (Hurlbert 1971).



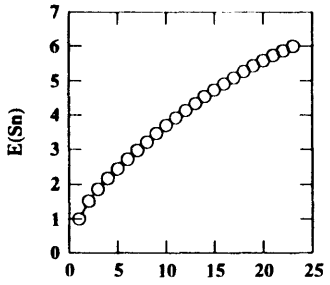
Gulf Stream 50m



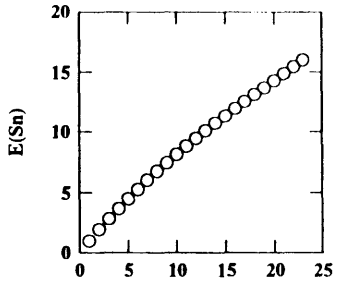
Gulf Stream 85m Bright



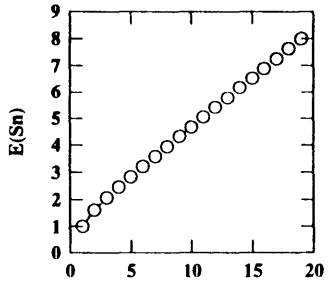
Gulf Stream 85m Dim



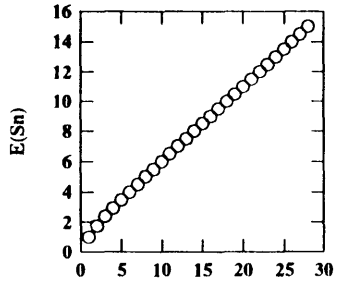
Gulf Stream 135m Bright



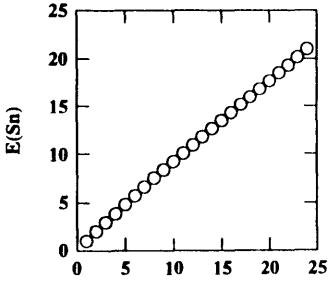
Gulf Stream 135m Dim



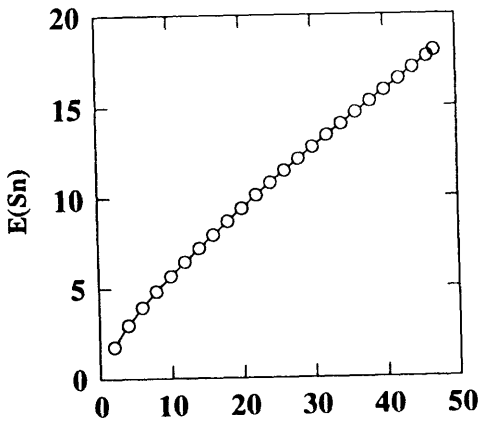
Sargasso Sea 40m



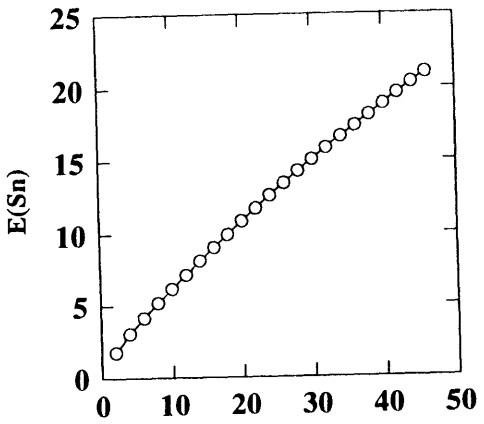
Sargasso Sea 70m



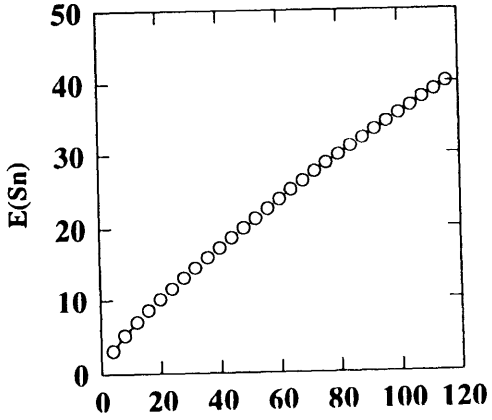
Sargasso Sea 120m



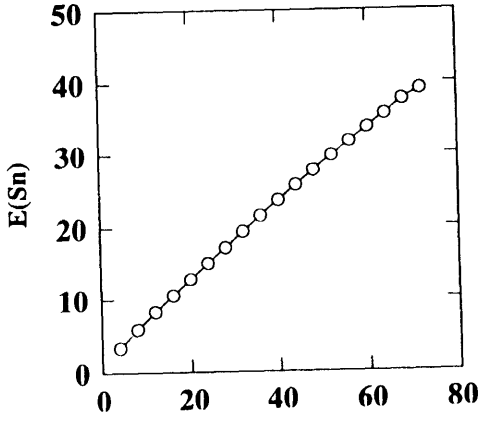
Combined Gulf Stream
85m Bright + Dim



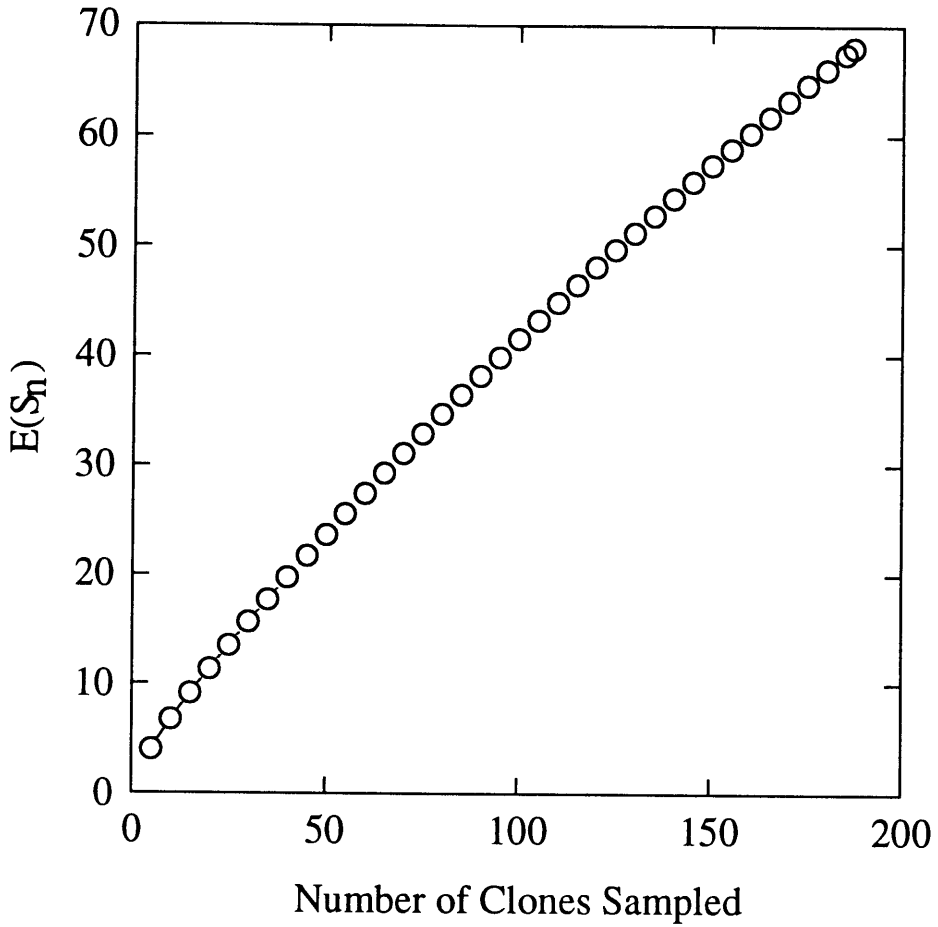
Combined Gulf Stream
135m Bright + Dim



Combined Gulf Stream
Samples from all Depths



Combined Sargasso Sea
Samples from all Depths


70
60
50
40
30
20
10
0
E(Sn)
0
50
100
150
200
Number of Clones Sampled

variants for subsample size n) versus n approaches horizontal as n approaches the sample size (Hurlbert 1971)[3]. Curves for each of the eight sorted samples, and for lumped samples show upward trends of $E(S_n)$ with n, and no indication of curve flattening. It is therefore apparent that many more genetic variants were present in the Sargasso Sea and Gulf Stream populations than were recovered in this study.

Distribution of alleles among sorted samples. Of the 68 *P. marinus* alleles in the dataset, only 16 were found more than once, with 12 appearing in more than one sample. Since rarefaction curves indicate that populations in the different sorted samples were not exhaustively sampled, the absence of shared alleles is not considered evidence of dissimilarity between samples. However, the shared presence of alleles is positive evidence of similarity. Scoring alleles according to their presence in the different samples reveals no consistent pattern of shared genotypes, but instead suggests that all of the sorted populations drew membership from a common gene pool (Table 4). Of particular interest is the finding that eleven alleles were shared between the Gulf Stream and Sargasso Sea depth profiles and that two alleles were shared by members of the double population from 85 m in the Gulf Stream, and one by the pair recovered from 135 m.

[3]This assumes a large parent population with a relatively small number of genotypes. It is possible to imagine a parent population in which every individual has a different genotype, in which case rarefaction curves would never flatten out, even for samples approaching the size of the parent population. However, even for this imagined situation an upward trend in the rarefaction curve for a finite sample would indicate that unsampled variants were present in the parent population.

Table 4. Presence of alleles found in multiple samples. Each sample is scored for the presence of 12 alleles which were recovered in more than one sorted sample.

	SargassoSea			Gulf Stream				
Allele	40m	70m	120m	50m	85m Bright	85m Dim	135m Bright	135m Dim
1	*	*	*	*		*	*	
5	*				*	*	*	*
6			*		*	*		*
12	*				*			
16		*	*		*			
17		*			*			
22			*				*	
31			*					*
48		*	*					
60			*					*
65			*		*			
71	*	*				*		*

DISCUSSION

Field populations of *P. marinus* are genetically heterogeneous. We find that field populations of *P. marinus* are genetically heterogeneous, according to the distribution of "presumptive" *P. marinus petB/D* alleles cloned out of flow cytometrically sorted field samples. The *P. marinus* alleles must be considered only presumptive because they fall into a lineage containing both *P. marinus* and marine A *Synechococcus* cultured isolates, which fail to form distinct phylogenetic clusters (Figure 5, this thesis Chapter Two). Thus the *P. marinus* identity of these alleles, presumed because of the flow cytometric purification, cannot be confirmed by phylogenetic analysis, and it is possible that some may have originated in marine A *Synechococcus* mistakenly included in the *P. marinus* sorted samples.

Between 6 and 21 alleles were present in the set of clones recovered from each of eight flow cytometrically sorted samples, with the presence of large numbers of additional alleles implied by rarefaction analysis (Figure 6). The set of alleles recovered within a water column is phylogenetically diverse within the *P. marinus*/marine A *Synechococcus* cluster, with both Sargasso Sea and Gulf Stream water columns containing alleles which form phylogenetic clusters with distantly related cultures isolated from the Pacific and the Sargasso Sea (compare Table 1 and Figure 5). The structure of *P. marinus* populations implied by these results is consistent with the detailed findings of Selander and colleagues for the population genetics of *Escherichia coli*, in which most of the genetic diversity of the species is found within local populations, and in which cell lineages are globally distributed (Selander et al. 1987).

Oceanographic implications of genetic diversity within *P. marinus* populations are twofold. First, the physiologic activities of *P. marinus* in the water column are not

simply described by the properties of one or a small number of cultures isolated from the same geographic area, but by the combined activities of a phylogenetically diverse set of cells, some of which are similar to isolates from distant parts of the world. Like SS120 and Med4, these genetic variants may exhibit physiological differences which would give each a selective advantage under different environmental conditions. The second oceanographic implication is that adaptive changes in *P. marinus* populations over the annual cycle and with depth are as likely to be due to shifts in their genetic composition, due to differences in growth rates of indigenous genetic variants under varying environmental conditions, as to intracellular adaptive responses (Wood 1988, Wood and Leatham 1992).

P. marinus populations recovered from the Gulf Stream and the Sargasso Sea water columns share eleven out of the 12 alleles which were found in more than one sorted sample (Table 4), despite the large scale hydrographic differences which distinguish these two water bodies. The large number of these shared alleles suggests, again, a similarity to the population structures of *E. coli*, in which populations in distant locations share alleles.

Flow cytometrically defined subpopulations from two "double populations" recovered from the Gulf Stream exhibited overlapping sets of *petB/D* alleles in our analysis, with the pair of subpopulations at 85 m sharing two alleles and the pair at 135 m sharing one. This result implies that these flow cytometric subpopulations are not genetically different, but may owe their distinct properties to processes such as synchronized cell division (Vaulot et al 1994) or mixing of cells acclimated to conditions in different water masses. Since the subpopulations of the two double populations were incompletely resolved, however, it is also possible that the shared alleles could result from the presence in either sort window of cells in the "tail" of the other subpopulation's

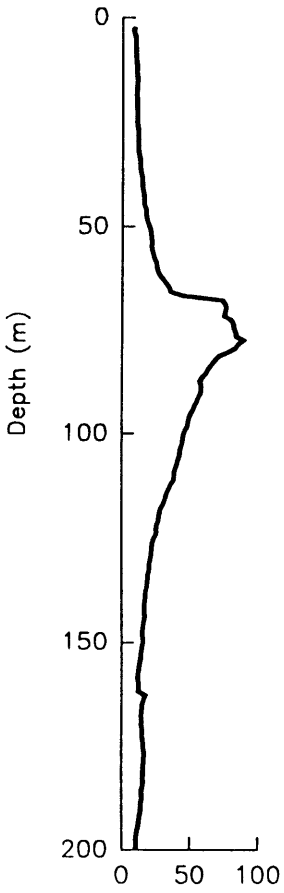
distribution. Also, it should be noted that multiple populations observed at DCM's in the Atlantic are transient phenomena (unpublished observations of R. Olsen), whereas multiple populations at Pacific station ALOHA are a stable feature of a permanent DCM (Cambell and Vaulot 1993). Multiple populations at ALOHA may therefore have a different origin from those examined in our study, and may indeed contain genetically different subpopulations.

Genotype frequencies among PCR clones derived from natural populations using degenerate, inosine-containing primers reflect both gene frequencies in the sample population and PCR amplification bias for or against specific alleles. It is therefore speculative to draw inferences from gene frequencies in datasets such as the one at hand. With this caveat in mind, it nonetheless bears mentioning that Allele 1 and Allele 1-like sequences predominate in clones recovered from the shallow mixed layer at both sampling sites (Figure 7), while at DCM's Allele 1-like sequences are present but do not predominate. Since HPLC and flow cytometric studies find that surface waters under stratified conditions contain low chl b_2/a_2 ratios (Goericke and Repeta 1993, Cambell and Vaulot 1993) similar to those found in *P. marinus* Med4 (Goericke and Repeta 1993), and since Allele 1 belongs to the phylogenetic cluster containing MIT9107 and Med4, these results may be correlated. The results of this study are therefore consistent with the hypothesis that populations at different depths in stratified watercolumns draw their membership from a single gene pool, but that gene frequencies vary at different depths, resulting in populations exhibiting different chl b/a_2 pigment ratios.

Figure 7. Frequency of Allele 1-like alleles in sorted samples. The frequency of alleles having intergenic regions similar to Allele 1 (Alleles 1, 2, 3, 69, 71, 72, 73, 74, 75 and 76) is scored for samples at different depths in the two water columns, lumping Bright and Dim flow cytometric subpopulations at two depths in the Gulf Stream profile. Although this inference must be considered speculative, the high frequency of recovery of Allele 1-like sequences in surface water samples is consistent with a predominance of these alleles in *P. marinus* populations in the surface mixed layer.

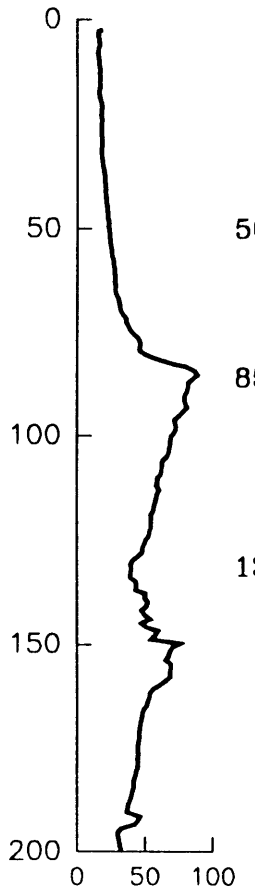
Frequency of Allele 1–Like Alleles in Sorted Samples

Sargasso Sea



	#Allele 1 –like / #Clones sampled	Frequency
40m	13/19	0.68
70m	17/28	0.61
120m	1/24	0.04

Gulf Stream



	#Allele 1 –like / #Clones sampled	Frequency
50m	23/23	1.00
85m	4/47	0.09
135m	4/46	0.09

REFERENCES

Brand, S.N., Tan, X. and Widger, W.R. (1992). Cloning and sequencing of the *petBD* operon from the cyanobacterium *Synechococcus sp.* PCC7002. Plant Mol. Biol. 20:481-491.

Britschgi, T.B. and Giovannoni, S.J. (1991) Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. 57:1707-1713.

Campbell, L., and Vaulot, D. (1993). Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawii (station ALOHA). Deep-Sea Res. 40:2043-2060.

Carter, K.R., Tsai, A. and Palmer, G. (1981). FEBS Lett. 132:243-246.

Chisholm, S.W., R.J. Olson, E.R. Zettler, R. Goericke, J. Waterbury, and N. Welschmeyer. (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature*, 334(6180):340-343.

Chisholm, S.W., Frankel, S.L., Goericke, R., Olson, R.J., Palenik, B., Waterbury, J.B., West-Johnsrud, L. and Zettler, E.R. (1992). *Prochlorococcus marinus nov.* gen *nov.* sp.: a marine prokaryote containing divinyl chlorophyll *a* and *b*. Arch. Microbiol. 157:297-300.

DeLong, E.F., Franks, D.G. and Alldredge, A.L. (1993) Phylogenetic diversity of aggregate-attatched vs. free-living marine bacterial assemblages. Limnol. Oceanogr. 38:924-934.

Geiskes, W.W.C., G.W. Kraay, A. Nontji, D. Setiapermana, and Sutomo. 1988. Monsoonal alternation of a mixed and a layered structure in the phytoplankton of the euphotic zone of the Banda Sea (Indonesia): A mathematical analysis of algal pigment fingerprints. Neth. J. Sea Res. 22:(2):123-137.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and K.G. Field. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-62.

Goericke, R. and Repeta, D.J. (1993). Chromatographic analysis of divinyl-chlorophylls *a* and *b* in samples from the subtropical north Atlantic Ocean. Mar. Ecol. Prog. Ser.

Goericke, R. and Welschmeyer, N.A. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep-Sea Res. 40:2283-2294.

Greer, K.L. and Golden, S.S. (1992). Conserved relationship between *psbH* and *petBD* genes: presence of a shared upstream element in *Prochlorothrix hollandica.* Plant Mol. Biol. 19:355-365.

Hope, A.B. (1993). The chloroplast cytochrome *bf* complex: a critical focus on function. Biochim. Biophys. Acta 1143:1-22.

Hurlbert, S.M (1971). The non-concept of species diversity: a critique and alternative parameters. Ecology 52:577-586.

Kallas, T., Spiller, S. and Malkin, R.C. (1988). Characterization of two operons encoding the cytochrome b6-f complex of the cyanobacterium *Nostoc* PCC7906 and highly conserved sequences but different gene organization than in chloroplasts. J. Biol. Chem. 263:14334-14342.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

Knoth, K., Roberds, S., Poteet, C. and Tamkun, M. (1988). Highly degenerate, inosine-containing primers specifically amplify rare cDNA using the polymerase chain reaction. Nuleeic Acids Res. 16:10371.

Lewin, R.A. (1981). Prochlorophytes. In *The Prokaryotes.* Vol 1. (eds. M.P. Starr, H. Stolp, H.G. Truper, A. Balows, and H. G. Schlegel). 256-266. Springer, Berlin.

Li, H., Cui, X. and Arnheim, A. (1991). Analysis of DNA sequence variation in single cells. Methods 2:49-59.

Li, W.K.W., Dickie, P.M., Irwin, B.D., Wood, A.M. (1992). Biomass of bacteria, cyanobacteria, prochlorophytes and photosynthetic eukaryotes in the Sargasso Sea. Deep-Sea Res. 39:501-519.

Li, W.K.W., and M. Wood. 1988. Vertical distribution of North Atlantic ultraphytoplankton: analysis by flow cytometry and epifluorescence microscopy. Deep Sea Res. 35:1615-1638.

Moore, L.R., Goericke, R., and Chisholm, S.W. (1994). The comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, flourescence and absorptive properties. Mar. Ecol. Prog. Ser., *in press.*

Morel, A., Ahn, Y.-H., Partensky, F., Vaulot, D. and Claustre, H. (1993). *Prochlorococcus* and *Synechococcus*: a comparative study of their optical properties in relation to their size and pigmentation. J. Mar. Res. 51:617-649.

Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S., Inokuchi, H. and Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572-574.

Olson, R.J., Chisholm, S.W., Zettler, E.R., Altabet, M.A., and Dusenberry, J.A. (1990) Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. Deep Sea Res. **37**, 1033-1051.

Partensky, R., Hoepffner, N., Li, W.K.W., Ulloa, O. and Vaulot, D. (1993). Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains isolated from the north Atlantic and the Mediterranean Sea. Plant Physiol. 101:285-296.

Reimann, A. and Kueck, U. (1989). Nucleotide sequence of the plastid genes for apocytochrome b_6 (*petB*) and subunit IV of the cytochrome b_6/f complex (*petD*) from the green alga *Chlorella protothecoides*: lack of introns. Plant Mol. Biol. 13:255-256.

Rock, C.D., Barkan, A. and Taylor, W.C. (1987). The maize plastid *psbB-psbF-petB-petD* RNAs encode alternative products. Curr. Genet 12:69-77.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning: a laboratory manual.* Cold Spring Harbor Laboratory Press, New York.

Selander, R.K., Caugant, D.A. and Whittam, T.S. (1987). Genetic structure and variation in natural populations of *Escherichia coli.* In Escherichia coli *and* Salmonella typhimurium*: cellular and molecular biology* (F.C. Neidhart, ed.). Am. Soc. Microbiol. pp. 1625-1648.

Swift, H. and Palenik, B. (1992). Prochlorophyte evolution and the origin of chloroplasts: morphological and molecular evidence. In *Origins of Plastids* (R.A. Lewin, ed.) Chapman and Hall, pp. 123-139.

Vaulot, D., Marie, D., Olson, R.J. and Chisholm, S.W. (1994). *Prochlorococcus* is highly synchronized to the diel cycle in the equatorial Pacific and divides up to once a day. submitted.

Vaulot, D., and Partensky, F. (1992). Cell cycle distributions of prochlorophytes in the northwestern Mediterranean Sea. Deep-Sea Res. 39:727-742.

Vaulot, D., Partensky, F., Neveux, J., Mantoura, R.F.C. and C. Llewellyn. 1990. Wintertime presence of prochlorophytes in surface waters of the North-Western Mediterranean Sea. Limnol. Oceanogr. 35:1156-1164.

Vermaas, W.F.J., and Ikeuchi, M. (1991). Photosystem II. In *The Photosynthetic Apparatus: molecular biology and operation* (L. Bogorad and I.K. Vasil, eds.). Academic Press pp. 26-111.

Widger, W.R. and Cramer, W.A. (1991). The b_6f complex. in *The Photosynthetic Apparatus: molecular biology and operation* (L. Bogorad and I.K. Vasil, eds.). Academic Press pp. 149-224.

Wood, A.M. (1988). Molecular biology, single cell analysis and quantitative genetics: new evolutionary genetic approaches in phytoplankton ecology. In *Immunochemical Approaches to Coastal, Extuarine and Oceanographic Questions* (C.M. Yentsch, F.C. Mague and P.K. Horan, eds.) Springer-Verlag pp 41-73.

Wood, A.M. and Leatham, T. (1992). The species concept in phytoplankton ecology. J. Phycol. 28:723-729.

Chapter 4

EPILOGUE: RESPONSES IN THE LITERATURE TO THE PUBLICATION OF CHAPTER ONE: MULTIPLE ORIGINS OF PROCHLOROPHYTES WITHIN THE CYANOBACTERIAL RADIATION

Since its publication, "Multiple Origins of Prochlorophytes within the Cyanobacterial Radiation" (Urbach et al. 1992) has generated numerous citations and has contributed to the formulation of new theories about the evolution of pigments in cyanobacteria and chloroplasts (Bryant 1992, Bullerjahn and Post 1992). This Epilogue is a brief review of published responses to this paper. The conclusions of Chapter One relevant to the discussion are (1) according to 16S ribosomal RNA (rRNA) phylogenetic analysis the three known prochlorophytes and chloroplasts each fall into separate lineages dispersed among the cyanobacteria, and (2) this polyphyletic distribution suggests chl *b* photosynthesis was "invented" several times during the evolution of cyanobacteria and chloroplasts (Urbach et al. 1992).

Most authors citing this work have no argument with the first of these conclusions (e.g. Bryant 1992, Bullerjahn and Post 1992, Cavalier-Smith 1992, Palenik and Haselkorn 1992, Swift and Palenik 1992, Cretiennot-Dinet et al. 1993) which echoes an inference drawn by Turner et al. (1989) and by Morden and Golden (1989a, b) from analyses of the evolution of *Prochlorothrix hollandica*, and which agrees with the analyses of Palenik and Haselkorn (1992) and Lockhart et al. (1992c). However, several workers have pointed out that the independent invention of chlorophyll *b* in each of several lineages is not the only evolutionary scenario to fit the data, and that the distribution of chl *b* among cyanobacterial and chloroplast lineages could also result from lateral gene transfer (Palenik and Haselkorn 1992) or by descent of the entire cyanobacterial lineage from an ancestor containing both chlorophyll *b* and phycobilisomes, with subsequent loss of phycobilisomes in some lineages (prochlorophytes and green chloroplasts) and chl *b* in others (most cyanobacteria) (Bryant 1992, Cavalier-Smith 1992). I wholly concur with these observations, and with the suggestion that information on prochlorophyte chlorophyll biosynthesis and chlorophyll *a/b* binding proteins will help to clarify this issue. To this end, Bullerjahn

and Post (1992) cite findings that the major chl *a/b* binding proteins of *Prochloron sp.* and *P. hollandica* are of similar molecular weight and are immunologically cross reactive. They combine these observations with sequence data indicating that *P. hollandica* antenna proteins are different from chloroplast light harvesting complexes, and form the hypothesis that photosynthetic antennae in *P. hollandica* and *Prochloron sp.* share a common origin separate from that of green chloroplasts. Future work in this direction, including analysis of the *P. marinus* photosynthetic system, will surely reveal a fascinating evolutionary story.

One research group, including P. Lockhart, A.W.D Larkum, D. Penny and coworkers does, however, take issue with the phylogenetic branching pattern inferred from 16S rRNA sequences in Chapter One (Lockhart and Penny 1992, Lockhart et al. 1992b, Larkum 1992, Lockhart et al. 1993). Their skepticism comes as an extension of their work showing that branching patterns inferred from protein-encoding genes are confounded by nucleotide substitution biases (detected as differences in G+C composition at third codon positions, see Chapter Two). Lockhart et al. (1992b) argue that phylogenetic trees inferred from cyanelle, prochlorophyte, cyanobacterial and chloroplast protein-encoding sequences are invalid because patterns of similarity in third codon position G+C content are congruent with the inferred phylogenetic branching patterns. They argue, in addition, that 16S rRNA trees, even those inferred from sequences without detectable G+C bias, are also invalid, based on the following premise:

> Although the effects of substitutional bias upon inference are best demonstrated by analysis of protein-coding nucleotide sequences, bias (a genomic effect in eubacterial lineages) must also affect variable sites within rRNA genes and may also have an effect on amino acid substitution (Jukes and Bhushan 1986; but see Prager and Wilson 1988). Inference from such datasets (particularly where long edges link taxa: e.g., Giovannoni et al. 1988; Morden and Golden 1989; Evrard et al. 1990; Kishino et al. 1990) may therefore also be misled.
>
> (Lockhart et al. 1992a)

Larkum (1992) proposes that, since phylogenetic trees linking the *Cyanophora paradoxa* cyanelle to chloroplasts and showing prochlorophytes dispersed among the cyanobacteria are invalid, it is reasonable to propose that prochlorophytes are monophyletic and and share an ancestry with green chloroplasts.

Lockhart and coworkers make a valid point in saying that phylogenetic trees, inferred by methods which assume constant substitution biases across taxa, may be in error when calculated from data having varying third codon position G+C content. Their extension to 16S rRNA sequences quoted above is a subtle argument, but also may hold true. Resolution of these phylogenetic problems must await the development of methods demonstrated to be insensitive to nucleotide substitution biases. However, the failure of *P. marinus* strains to group with chloroplasts having similar third codon position G+C content in trees inferred with *psbB* and *petB/D* sequences (Urbach and Chisholm in preparation, this thesis Chapter Two) argues against an evolutionary link between *P. marinus* and chloroplasts.

I'll close with a quote from Bullerjahn and Post (1992) which is appropos:

> If one reviews much of the literature on the prochlorophytes, it becomes clear that taxonomic classification and the inferring of relationships between organisms can be a risky enterprise. Some microbiologists have been eager to show a direct relationship between prochlorophytes and the chloroplast, whereas others were determined to place them among the cyanobacteria. The underlying problem in such exercises lies both in the danger of overemphasizing one aspect while belittling others and in the import of externally imposed bias into the judgement of a certain property.

REFERENCES

Bryant, D.A. (1992). Puzzles of chloroplast ancestry. Curr. Biol. 5:240-242.

Bullerjahn, G.S. and Post, A.F. (1992). The prochlorophytes: are they more than just chlorophyll *a/b*-containing cyanobacteria? Crit. Rev. Microbiol.

Cavalier-Smith, T. (1992). The number of symbiotic origins of organelles. Biosystems 28:91-106.

Cretiennot-Dinet, M.-J., Sournia, A., Ricard, M. and Billard, C. (1993). A classification of the marine phytoplankton of the world from class to genus. Phycologia 32:159-179.

Everard, J.-L., Kuntz, M. and Weil, J.-H. (1990). The nucleotide sequence of five ribosomal protein genes from the cyanelles of *Cyanophora paradoxa*: implications concerning the phylogenetic relationship between cyanelles and chloroplasts. J. Mol. Evol. 30:16-25.

Giovannoni, S.J., Turner, S., Olsen, G.J., Barnes, S., Lane, D.J. and Pace, N.R. (1988). Evolutionary relationships among cyanobacteria and green chloroplasts. J. Bacteriol. 170:3584-3592.

Jukes, T.H. and Bhushan, V. (1986). Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J. Mol. Evol. 24:39-44.

Kishino, H., Takashi, T. and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. 31:151-160.

Larkum, A.W.D. (1992). Evolution of chlorophylls, light harvesting systems and photoreaction centres. In *Research in Photosynthesis, Vol. III* (N. Murata, ed.). Kluwer Academic, pp. 475-482.

Lockhart, P.J., Penny, D. (1992). The problem of GC content, evolutionary trees and the origins of chl-*a/b* photosynthetic organelles: are the prochlorophytes a eubacterial model for higher plant photosynthesis? in *Research in Photosynthesis, Vol. III* (N. Murata, ed.). Kluwer Academic, pp. 499-505.

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J. and Larkum, A.W.D. (1992a). Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol 34:153-162.

Lockhart, P.J., Penny, D., Hendy, M.D., Howe, C.J., Beanland, T.J. and Larkum, A.W.D. (1992b). Controversy on chloroplast origins. FEBS Lett. 301:127-131.

Lockhart, P.J., Beanland, Howe, C.J. and Larkum, A.W.D. (1992c). Sequence of Prochloron didemni atpBE and the inference of chloroplast origins. Proc. Natl. Acad. Sci. USA 89:2742-2746.

Lockhart, P.J., Penny, D., Hendy, M.D., and Larkum, A.D. (1993). Is *Prochlorothrix hollandica* the best choice as a prokaryotic model for higher plant chl *a/b* photosynthesis? Photosynthesis Res. 37:61-68.

Morden, C.W. and S.S. Golden. 1989a. *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts. Nature 337:382-385.

Morden, C.W. and S.S. Golden. 1989b. Corrigendum: *psb*A genes indicate common ancestry of prochlorophytes and chloroplasts. Nature 339:400.

Palenik, B. and Haselkorn, R. (1992). Multiple evolutionary origins of prochlorophytes, the chlorophyll *b*-containing prokaryotes. Nature 355:265-267.

Prager, E.M. and Wilson, A.C. (1988). Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J. Mol. Evol. 27:326-335.

Swift, H., and Palenik, B. (1992) Prochlorophyte evolution and the origin of chloroplasts: morphological and molecular evidence. In *Origins of Plastids* (R.A. Lewin, ed.) Chapman and Hall pp. 123-138.

Turner, S., Burger.Wiersma, T., Giovannoni, S.J., Mur, L.R., and N.R. Pace. (1989). The relationship of a prochlorophyte *Prochlorothrix hollandica* to green chloroplasts. Nature 337:380-382.

Urbach, E., Robertson, D. and Chisholm, S.W. (1992). Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. Nature 355:267-269.

Appendix I

PRELIMINARY ANALYSIS OF GENETIC DIVERSITY IN GULF STREAM POPULATIONS OF *PROCHLOROCOCCUS MARINUS* BY DIRECT SEQUENCING OF PCR PRODICTS AMPLIFIED FROM FLOW-CYTOMETRICALLY SORTED CELLS

Chapter Three details an investigation into the genetic diversity of natural populations of *Prochlorococcus marinus* from the the Gulf Stream and the Sargasso Sea in which *P. marinus* were sorted from field samples by flow cytometry and the diversity of *petB/D* sequences in the sorted samples assessed by cloning and sequencing individual molecules from PCR amplification products. This Appendix describes a preliminary investigation which preceded the cloning experiment in which PCR products sorted from Gulf Stream field samples were directly sequenced. Autoradiograms of the resulting sequencing gels revealed comigrating bands at numerous nucleotide positions, especially in the intergenic region between *petB* and *petD*, indicating that multiple *petB/D* sequences were likely to be present in the amplification products.

METHODS

Sample collection, flow cytometric sorting and DNA isolation. The DNA preparations used for this preliminary experiment were the same as those used for the analysis of Gulf Stream populations in Chapter Three, and originated in samples collected from 50 m, 85 m and 135 m at 37^{o} 30.68'N, 68^{o} 13.69'W during July of 1993. Samples collected from 85 m and 135 m each contained a pair of flow cytometrically defined "double populations," subpopulations of which were sorted into separate samples, resulting in a total of five sorted samples (Figure 2 in Chapter Three).

PCR amplification. Sequences at the *petB/D* locus were amplified as described for the analysis of *petB/D* sequences from cultured *P. marinus* and *Synechococcus*, using biotinylated forward primer PPETBD314 and unbiotinylated reverse primer PPETBD1160R (this thesis Chapter Two). PPETBD1160R also served as a primer for sequencing. Conditions for PCR amplification were less stringent than those used for the cloning experiment (this thesis Chapter Three) and produced sufficient amounts of PCR

products to serve directly as sequencing templates. Templates were prepared using streptavidin-coated magnetic beads (Dynal) and sequenced directly with Sequenase (United States Biochemical), both according to their manufacturers' instructions.

RESULTS

Autoradiograms from direct sequencing of *petB/D* amplified from sorted samples revealed multiple comigrating bands which differed in numbers and intensities among the different samples (Figure 1). Most dramatic were the results of direct sequencing of *petB/D* from the 85 m Bright and 135 m Dim samples in which multiple comigrating bands were frequent in the intergenic region above the *petB/D* start codon and were also visibile at some third codon positions in *petD* (Figure 1, lanes b and e). Multiple comigrating bands were less prominent, but visible, in sequencing reactions from the 50 m amplification (lane a) and could not be detected in the 85 m Dim and 135 Bright (lanes c and d, compare to reactions for a *P. marinus* culture in lane e).

"Major" sequences could be read above the relatively low or absent background of comigrating bands for the 50 m, 85 m Dim and 135 m Bright samples. All of these major sequences were similar to Allele 1, the most frequently recovered allele in the cloning experiment (49 out of 191 clones in the combined dataset for Gulf Stream and Sargasso Sea sorted samples, 22 out of 139 clones for Gulf Stream samples alone, Table 1 in Chapter 3) (Figure 1, lanes a, c and d).

DISCUSSION

The presence of multiple comigrating bands in sequencing autoradiograms can result either from sequence heterogeneity in the template DNA or from suboptimal

reaction conditions which allow premature termination of nascent DNA chains. The latter cause would not, however, be expected to produce a disproportionate number of comigrating bands at third codon position sites and intergenic regions, a result which is dramatically evident in autoradiogram lanes from the 85 m Bright and 135 m Dim sorted samples (Figure 1, lanes b and e) and, less dramatically, in lanes from the 50 m sorted sample (lane a). The results of this analysis were therefore consistent with the presence of multiple *petB/D* alleles in at least some of the flow cytometrically sorted field samples, and prompted the cloning and sequencing investigation in Chapter Three.

The investigation in Chapter Three revealed that, in fact, all of the PCR products from the Gulf Stream and similar products from the Sargasso Sea contained multiple *petB/D* sequences. Consistent with the results of direct sequencing, the 85 m Bright and 135 m Dim samples contained the largest variety of *petB/D* alleles: when corrected to a constant sample size of 19 clones these samples contained 12.1 and 13.7 alleles, respectively. The 50 m, 85 m Dim and 135 m Bright sorted samples yielded 5.3, 5.2 and 5.4 alleles for the same constant sample size (Table 1 in Chapter Three).

There was, however, one type of comparison in which the results of the direct sequencing and cloning experiments did not consistently agree. The major sequence and the most frequently recovered allele were not consistently identical for each of the sorted samples: while the 50 m sample yielded a majority of Allele 1 clones (17 out of 23 clones) consistent with its major sequence, the sets of clones recovered from the 85 m Dim and 135 m Bright samples were mostly Allele 5 (17 out of 24 and 16 out of 23 clones, respectively), with only a minority of clones belonging to Allele 1 (three out of 24 and two out of 23 clones, respectively) (Table 1 in Chapter Three). This discrepency may be attributable to two potential causes: 1) Allele 1 may have been preferentially amplified under the less stringent conditions used to produce PCR products for direct

sequencing as compared to the cloning experiment, or 2) Allele 1 may have been preferentially chosen as template during direct sequencing reactions. Discrepencies in the identities of major sequences and most frequently recovered alleles for the different sorted samples reinforce the necessity of mentioning the caveat in Chapter Three: genotype frequencies among PCR clones (and major sequences of PCR products) derived from natural populations using degenerate, inosine-containing primers reflect both gene frequencies in the sample population and PCR amplification bias (as well as sequencing reaction bias) for or against specific alleles. It is therefore speculative to draw inferences from gene frequencies (or major sequences) in datasets such as the one at hand.

Figure 1. Autoradiograms from direct sequencing of *petB/D* PCR products amplified from flow cytometrically sorted *P. marinus* from natural populations in the Gulf Stream (a-e) and from unsorted, cultured *P. marinus* MIT9313, isolated from the 135 m Gulf Stream water sample (L. Moore pers. comm.) (f). The arrow marks the initial *petD* ATG codon, above which is intergenic region sequence. Dashes mark the positions of third codon positions in *petD*. a) 50 m, b) 85 m Bright, c) 85 m Dim, d) 135 m Bright, e) 135 m Dim sorted samples (Chapter Three). f) unsorted, cultured MIT9313.

a
b
c
d
e
f

REFERENCE

Hurlbert, S.M. (1971). The non-concept of species diversity: a critique and alternative parameters. Ecology 52:577-586.

Appendix II

JUSTIFICATION FOR THE ≤3.5% SEQUENCE DIFFERENCE CRITERION FOR SEQUENCES ASSIGNED TO A SINGLE ALLELE

Introduction. In Chapter Three, genetic diversity in natural populations of *Prochlorococcus marinus* was investigated by comparing DNA sequences of cloned PCR products amplified from flow cytometrically sorted field samples. The genetic locus examined included the 3' end of the *petB* gene, an intergenic "nonsense" region and the 5' end of the *petD* gene (the amplified locus, including both gene fragments and the intergenic region, being referred to as "*petB/D*"). Clone sequences (which were inferred from single gel readings, and not confirmed by sequencing complimentary strands) were categorized into Alleles, or sets of clones within which sequence differences were indistinguishable from PCR and sequence determination (PCR+SD) error, corrresponding to genotypes present in the field samples. For ease of analysis, each Allele was characterized by a single clone designated as its "prototype." Among the 191 clones analyzed there was a great deal of sequence diversity, with 59 sets of clones exhibiting intergenic region sequences that were so different in length and sequence that it was impossible to identify homologous nucleotides for use in sequence alignment. These sequence differences were easily identified, despite the undoubted presence of PCR+SD error, as ones which originated in genetically distinct individuals in the sorted populations. Three of these sets of clones, however, contained sequences exhibiting sufficient mismatch to suggest that they were amplified from individuals containing different, though related sequences. These sets were subdivided into Alleles according to a ≤3.5% total sequence mismatch with prototype criterion for clones included within an Allele. The total number of Alleles arrived at by this process, including the sets of clones with unalignable intergenic sequences, was 71.

This Appendix presents arguments in support of the ≤3.5% total sequence mismatch criterion used to group clones into Alleles in Chapter Three. Initially, bounds are set on possible values for the criterion using a theoretical calculation based on published values for Taq polymerase error rates in PCR and an *ad hoc* estimate of the range of sequence determination error rates. Next, evidence is presented from the distribution of closely related

clones among field samples suggesting that some sequences within these bounds do differ as a result of natural variation, and not simply due to PCR+SD error. Finally, a practical method is described for optimizing the choice of criterion using the fraction of sequence mismatch occuring at third codon positions, and the method is applied to the dataset.

Theoretical confidence limits for the frequency of sequence mismatches between cloned PCR products attributable to PCR+SD error. During the course of PCR amplification, replication errors give rise to PCR products containing one or more mismatches with the original template sequence. The frequency of errors within individual PCR product molecules can be seen to follow a Poisson distribution, with variance equal to the mean, despite the fact that the frequency of errors in whole PCR reactions conforms to a Luria-Delbruck distribution (Eckert and Kunkel 1991a,b), and thus has a very high variance[1].

In the final, cloned PCR product the mean error frequency (f) is equal to the probability that polymerase error has caused a nucleotide to become mismatched with its progenitor in the original template (Eckert and Kunkel 1991a,b),

[1]The Luria-Delbruck distribution was first proposed to describe the widely varying frequency of mutant cells in replicate bacterial cultures grown under non-selective conditions in the classic fluctuation test (Luria and Delbruck 1943), but the distribution is equally applicable to the frequency of errors in replicate PCR reactions (Eckert and Kunkel 1991a,b). In a Luria-Delbruck process the rare occurrence of a mutation (PCR error) during an early generation (cycle) causes the mutant (error) frequency in a small number of cultures (reactions) to be substantially greater than the mean, giving rise to a frequency distribution having a high variance. This is because the final number of mutants (PCR products) resulting from an early mutation (error) is amplified through the process by a factor of 2^X, where X equals the number of generations (cycles) between the mutation (error) and the analysis. Thus, in a Luria-Delbruck process the number of the round in which a error occurs influences the final error frequency attributable to that error. In the case of the distribution of errors within a single PCR product molecule, however, the situation is different. In this case, each nucleotide is descended from its original template nucleotide through a number of rounds of PCR replication, each with a probability of error. Every nucleotide in the chain is subject to this process independently, and the final error frequency attributable to any error is equal to 1/L, where L is the number of nucleotides in the chain, and is independent of the number of the round in which the error occurred. If the probability of error is assumed to be equal for every nucleotide position, errors within the chain are seen to be rare, independent events with equal probabilities of occurrence. Thus the frequency of these errors conforms to a Poisson distribution.

$$f = \frac{nm}{2} \qquad (1)$$

where n is the number of PCR cycles and m the probability of error per nucleotide per cycle[2]. Since these errors are distributed as a Poisson distribution, the variance for the frequency of errors in individual PCR product molecules is equal to the mean, and confidence limits can be calculated for the frequency of PCR errors in individual product molecules using known values for the polymerase error rate. The expected frequency of differences between two cloned PCR products that can be attributed to PCR error is equal to the expected frequency of errors in a cloned molecule twice the length (l) of the region compared between the two products (neglecting coincidence of errors at homologous nucleotide sites and assuming the two cloned molecules are not amplification products of the same template molecule, which would decrease the expected frequency of their sequence differences). Accordingly, the 95% confidence upper limit (L95) for the frequency of mismatch between two cloned sequences attributable to PCR error is

$$L_{95} = f + t_{0.1[2l-1]}\sqrt{f} \qquad (2)$$

where $t_{0.1[2l-1]}$ is the critical value for student's t distribution for a one-tailed test with α = 0.05 and degrees of freedom equal to 2l - 1.

A term may be added to expression (1), the mean frequency of polymerase errors in cloned PCR products, in order to account for sequence determination errors. This gives

$$f = \frac{nm}{2} + g \qquad (3)$$

where g is the rate of sequence determination errors in errors nucleotide^{-1}. If sequence determination errors are assumed to follow a Poisson distribution, this expression for f may be used to calculate 95% confidence intervals for PCR+SD error using equation (2).

---

[2]The probability that a molecule has inherited a DNA strand containing an error introduced during the previous round is m/2, hence the 2 in the denominator. This factor applies to the product of the final PCR round as well, although in this case it reflects the probability that the progeny of a mutant strand in a cloned heteroduplex will predominate in the final plasmid preparation. This formulation neglects the effects of superimposed errors.

Sequence determination errors are estimated *ad hoc* at between 0.0 and 5.0 x 10^{-3} errors nucleotide^{-1}.

Application of the theoretical confidence limits to analysis of the data in Chapter Three. Mean, standard deviation and upper 95% confidence limits for the frequency of sequence differences expected in cloned PCR products compared over 71 basepairs (the length of the shortest sequence in the dataset) were calculated using the lowest and highest reported Taq polymerase error rates (Eckert and Kunkel 1991a) (corrected to reflect nucleotide substitutions only[3]), over the estimated range of sequence determination errors (Table 1). These upper confidence limits provide theoretical bounds for the value of the criterion to be used to declare sequences different or indistinguishable. If the lowest combined error rate were to apply, sequences in Chapter Three having alignable intergenic regions and differing by up to 3.1% of nucleotide positions would be considered indistinguishable from PCR products amplified from identical template sequences, at p = 0.05. However, applying the highest reported error rate raises this criterion value to 18.5%. Thus, the 95% upper confidence confidence limit for the amount of sequence difference between cloned PCR products attributable to error is shown to be sensitive to estimates of the polymerase and sequence determination error rates, the true values of which are unknown for this experiment (Table 1).

[3] Values for the range of Taq polymerase error rates in PCR are available from the literature (Eckert and Kunkel 1991) but require correction in order to be applicable to data analysis for Chapter 3. In Chapter 3, apparent insertions and deletions are not counted as sequence mismatches (*i.e.*, gap weights were set to zero) on the assumption that these mismatches are likely to have been caused by PCR+SD error. Accordingly, Taq polymerase error rates used to calculate theoretical upper 95% confidence limits for PCR+SD error in Chapter 3 have been corrected to omit insertion and deletion errors using the original data, where this is applicable. Specifically, the highest reported Taq polymerase PCR error rate has been corrected from 2 x 10^{-4} to 1.6 x 10^{-4} errors nucleotide^{-1} cycle^{-1}, using the original data (Dunning et al. 1988), while the lowest reported error rate is unchanged because insertions and deletions were not observed in the original experiments (Goodenow et al. 1989).

Table 1. Theoretical statistics for error in PCR product sequences for the range of reported Taq polymerase error rates and an estimated range of sequence determination errors.

Polymerase error rate (errors nucleotide^{-1} cycle^{-1})[a]	Sequence determination error rate (errors nucleotide^{-1})[b]	Mean error rate in PCR product sequence (errors nucleotide^{-1})[c]	Standard deviation (errors nucleotide^{-1})[d]	L_{95} (errors nucleotide^{-1})[e]
1.0×10^{-5}	0	3.5×10^{-4}	1.87×10^{-2}	3.10×10^{-2},[f]
1.6×10^{-4}	0	5.6×10^{-3}	7.48×10^{-2}	1.30×10^{-1}
1.6×10^{-4}	5.0×10^{-3}	1.1×10^{-2}	1.02×10^{-1}	1.85×10^{-1},[f]

[a]Range of reported reported Taq polymerase error rates (Eckert and Kunkel 1991), corrected from the original data (Dunning et al. 1988, Goodenow et al. 1989) to omit PCR insertion and deletion errors. The corrected rate is applicable to analysis of the data in Chapter 3 because alignment gaps were not considered mismatches in the analysis. The lowest reported Taq polymerase error rate did not include any insertions or deletions.

[b]Estimated range of sequence determination errors, 0 to 5.0×10^{-3} errors nucleotide^{-1}.

[c]Mean error rate for PCR product sequences produced by 70 cycles of PCR (Eckert and Kunkel 1991), including potential sequence determination error,

$$f = \frac{nm}{2} + g$$

f = mean error rate in PCR products (errors nucleotide^{-1})
n = number of cycles
m = polymerase error rate (errors nucleotide^{-1} cycle^{-1}).
g = sequence determination error rate (errors nucleotide^{-1})

[d]Standard deviation for frequency of errors in PCR product sequences,

$$\text{standard deviation} = \sqrt{f}$$

[e]Upper 95% confidence limit for frequency of errors in comparisons of cloned PCR product sequences,

$$L_{95} = f + t_{0.1[2l-1]}\sqrt{f}$$

$t_{0.1[2l-1]}$ = critical value for student's t distribution for a one-tailed test with $\alpha = 0.05$ and degrees of freedom equal to 2l - 1
l = length of the region compared between the two cloned products, set to 71 basepairs, the length of the shortest sequence in the dataset.

[f]Theoretical upper 95% confidence limits for sequence mismatch attributable to PCR+SD error for the highest and lowest combinations of reported Taq polymerase PCR error rates and estimated sequence determination error rates, illustrated in Figures 1 and 2.

Applying the calculated extreme bounds for the criterion to the task of lumping indistinguishable sequences into Alleles using the set of clones recovered from the Gulf Stream and Sargasso Sea yields a total of 74 Alleles in the combined dataset for the lowest error estimate and 59 Alleles for the highest one[4]. While the impulse to err on the side of conservatism would dictate the choice of ≤18.5% mismatch as criterion for inclusion within an allele, independent evidence suggests that this criterion would lump together a number of cloned sequences which do, in fact, derive their differences from genetic heterogeneity in *P. marinus* populations. A lower criterion, while still above the 95% upper confidence limit calculated from the lowest reported Taq polymerase error rate, can be estimated by considering the pattern of distribution of closely related sequences among field samples and the ratio of their mismatches at third codon positions to mismatches at all codon positions in order to distinguish sets of related sequences giving useful information about genetic diversity from sets in which members differ mostly due to PCR+SD error.

An upper limit for the sequence mismatch criterion suggested by the distribution of closely related sequences among field samples. In the dataset of Chapter Three four 100.0% identical sequences belonging to Allele 71 were recovered from four different field samples: the 85m Dim and 135m Dim Gulf Stream samples and the 40m and 70m Sargasso Sea samples, and therefore were cloned out of separate PCR reactions. Allele 71 is highly similar to Allele 1 (8.1% mismatch), which includes numerous identical clones distributed among several field samples. If the 18.5% mismatch criterion were to be applied, Allele 1 and Allele 71 (as well as several others) would be lumped together as members of a single Allele. While it is not unlikely that a single Allele 71 clone could have been produced from an original Allele 1 template, it is highly unlikely that the identical set of errors would

[4] Application of the upper bound criterion calculated using the highest combined PCR+SD error rate causes all sequences having alignable intergenic regions to be lumped with each other. There are 59 sets of intergenic regions which cannot be aligned with each other, therefore application of this higherst possible criterion value results in a dataset containing 59 Types.

occur in four separate PCR reactions and four separate sequence determinations. Therefore, at least some sequences in the dataset that differ by 8.1% represent true genetic variants and not the results of PCR+SD error.

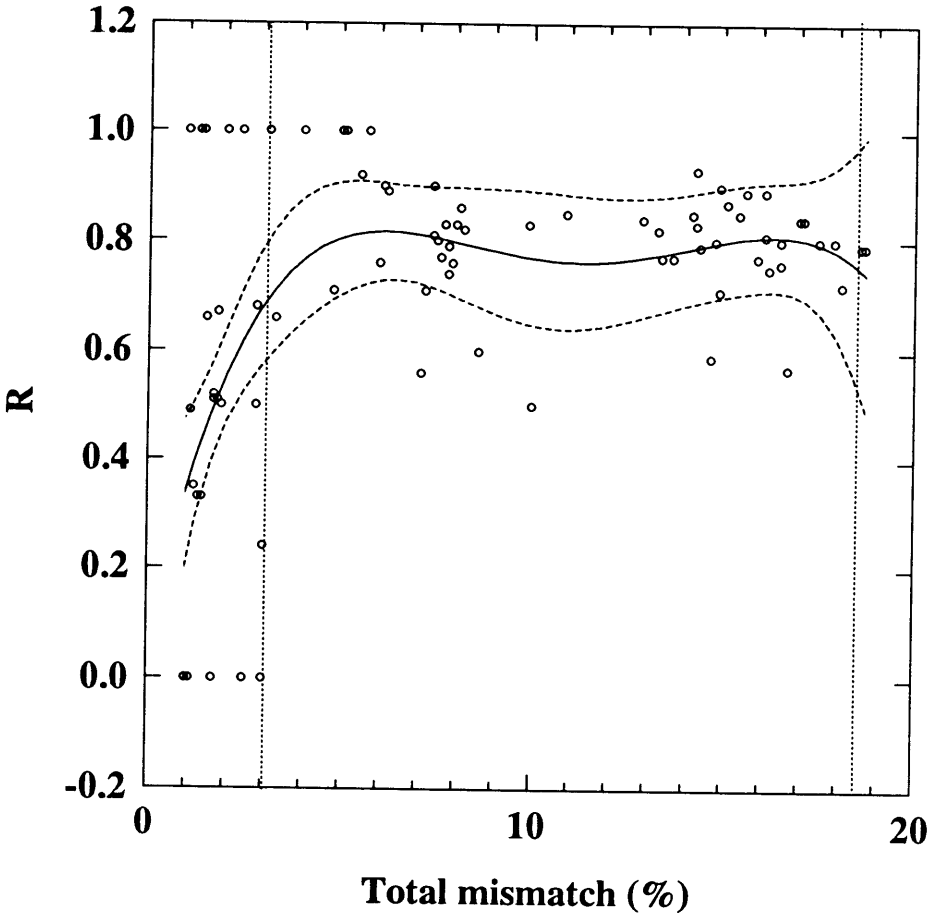
Optimization of the criterion for distinguishing sequence mismatch attributable to PCR+SD error from mismatch due to genetic diversity by considering the fraction of sequence mismatch at third codon positions: a practical method. Evolutionary change at third codon positions is generally faster than at the first two positions because degeneracy in the genetic code allows substitution at the third position of most codons without changing the encoded amino acid. PCR+SD errors, on the other hand, should be random with respect to codon postion. A high fraction of sequence mismatch at third codon positions (fraction of mismatch at third codon positions = fractional mismatch at third codon positions/3 x fractional mismatch at all codon positions) is therefore evidence that a pair of cloned sequences differ mostly due to true evolutionary divergence under selection for a conserved amino acid sequence, rather than due to PCR+SD error. Conversely, a low ratio is evidence of the activity of random processes like PCR error and inaccurate sequence determination.

In an ideal dataset, a pair of cloned sequences which differ due to PCR+SD error alone should have a fraction of sequence mismatch at third codon positions equal to approximately 0.33. In an actual dataset like the one in Chapter Three, however, there is likely to be a significant amount of scatter in values for this ratio at the low end of the total sequence mismatch scale due to sampling error associated with the small number of mismatches in highly similar cloned sequences of finite length. Additionally, in a real dataset some sequences differing by small increments of total mismatch may, in fact, differ due to evolutionary processes, but these individual comparisons cannot be distinguished from PCR+SD error because sampling error obscures differences in their fractions of mismatch at third codon positions. The effect of the presence of these small evolutionary differences will

be to increase the average fraction of mismatch at third codon positions above the 0.33 value for comparisons at the low end of the total sequence mismatch scale. In the face of this complexity, the task of setting a total mismatch criterion below which sequence differences are considered indistinguishable from PCR+SD error (for experiments in which the error rate has not been experimentally determined) becomes a process of examining the distribution of fractional mismatch at third codon positions along the scale of total mismatch and identifying a point of discontinuity at which the sphere of influence of random errors gives way to one of evolutionary divergence. At this point there will be a maximum difference between the means for fractional sequence differences at third codon positions for sequence comparisons above and below the criterion. This transition point should fall between (or perhaps slightly below) the theoretical upper 95% confidence limits for mismatch attributable to PCR+SD error, determined according to the range of probable polymerase and sequence determination error rates.

Optimization of the criterion for sequences in Chapter Three. For comparisons of clones having alignable intergenic regions in the dataset of Chapter Three, a plot of the fractional sequence mismatch at third codon positions (in coding regions only) versus total mismatch (in coding regions plus intergenic regions) exhibits a marked biphasic appearance (Figure 1). Consistent with the expectations of random error due to PCR+SD superimposed over a smaller amount of difference due to evolutionary processes, comparisons at the low end of the total mismatch scale have a lower average fraction of mismatches at third codon positions than do comparisons at the high end, but this average is greater than 0.33. A criterion chosen to distinguish domains of total sequence mismatch in which the dominant cause of sequence mismatch is PCR+SD error or true genetic diversity should divide the dataset into two subsets having the greatest possible difference in their average fraction of sequence mismatch at third codon positions. A graph plotting this difference for

Figure 1. Fraction of sequence mismatch at third codon positions (in coding regions only), plotted against total mismatch (including coding and intergenic regions) for comparisons of clones with alignable intergenic regions in the dataset of Chapter Three. Dotted vertical lines are theoretical upper 95% confidence limits for percent sequence mismatch (L_{95} x 100) attributable to PCR+SD error for the highest and lowest combined Taq polymerase and sequence determination error rates (Table 1). Datapoints below 1% total mismatch are omitted, and some symbols represent superimposed data points. Solid curve: fourth order polynomial fitted to all datapoints. Dashed curves: upper and lower 95% confidence intervals for the solid curve.


1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
R
0
10
20
Total mismatch (%)

hypothetical criteria placed along the total mismatch frequency scale at intervals of 0.1% mismatch shows a local maximum at 3.5% total mismatch, between the theoretical bounds imposed by published values for Taq polymerase error rates combined with estimates for sequence determination error rates (Figure 2). The value 3.5% total mismatch was therefore chosen as the best possible criterion to distinguish sequences differing due to PCR+SD error from those differing mostly due to evolutionary divergence, for the dataset of Chapter Three. This value is consistent with a Taq polymerase error rate near the low end of the range of reported values[5].

Reexamination of the graph plotting the fraction of sequence mismatch at third codon positions versus total mismatch frequency (Figure 1) using the 3.5% criterion to divide the data into two subsets and analyzing each subset independently for the mean and its 95% confidence intervals, illustrates the contrast between sequence differences attributed to PCR+SD error at the low end of the total sequence mismatch scale, and differences due to natural variation at the high end (Figure 3). The mean fraction of sequence mismatch at third codon positions for comparisons to the left of the criterion equals 0.48; for comparisons to the right of the criterion the mean is 0.81. 95% confidence intervals for these means do not overlap, indicating that the criterion chosen by this method separates the sequence comparisons in the dataset of Chapter Three into groups which are significantly different in their fraction of sequence difference at third codon positions, and therefore in their ratio of random to evolutionary sequence differences.

[5] It should be pointed out that this analysis would be improved by an additional sequencing effort, providing full-length sequences confirmed by sequencing in both directions for all clones having alignable intergenic regions. It is predicted that the number of sequence comparisons yielding ones and zeros for third codon position/all codon position mismatch ratios would decrease with increased sequence lengths and accuracy, reducing scatter at low total mismatch values in Figures AI-1 and AI-3 and perhaps altering Mean R - mean L values at low total mismatch values in Figure AI-2 as well. An additional sequencing effort before publication of these results is being considered.

Figure 2. Optimization of the criterion identifying the upper bound for total sequence mismatch considered indistinguishable from PCR+SD error. Datapoints are differences between mean values for the fraction of sequence mismatch at third codon positions for datapoints in Figure 1, grouped to the right and left of hypothetical criteria placed at 0.1% intervals along the total mismatch axis. Dotted vertical lines are theoretical upper 95% confidence limits for percent sequence mismatch (L_{95} x 100) attributable to PCR+SD error for the highest and lowest combined Taq polymerase and sequence determination error rates (Table 1), and therefore represent theoretical bounds for the criterion. The curve is a second order polynomial fitted to points between 2.0% and 5.9% total mismatch to aid in identifying the local maximum, which is marked with a solid vertical line at 3.5% total mismatch. The criterion is therefore set at the value of total mismatch dividing the sequence comparisons into the two groups having the most different mean fraction of sequence mismatch at third codon positions possible within the bounds of the theoretical limits.

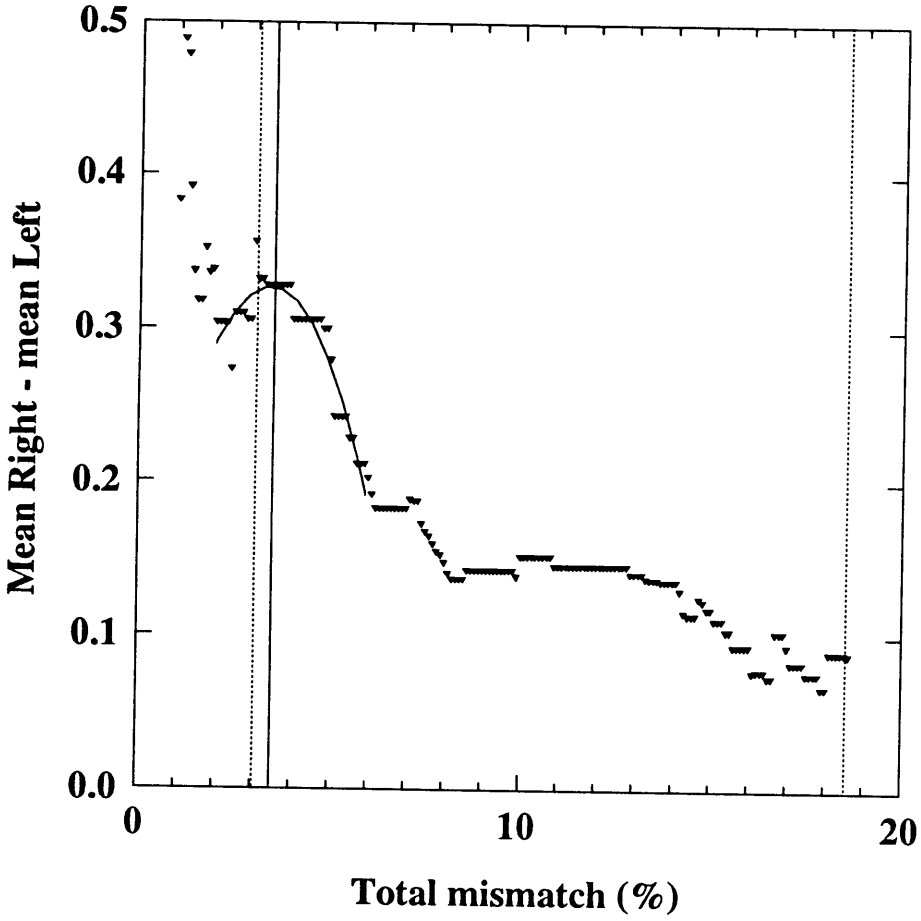
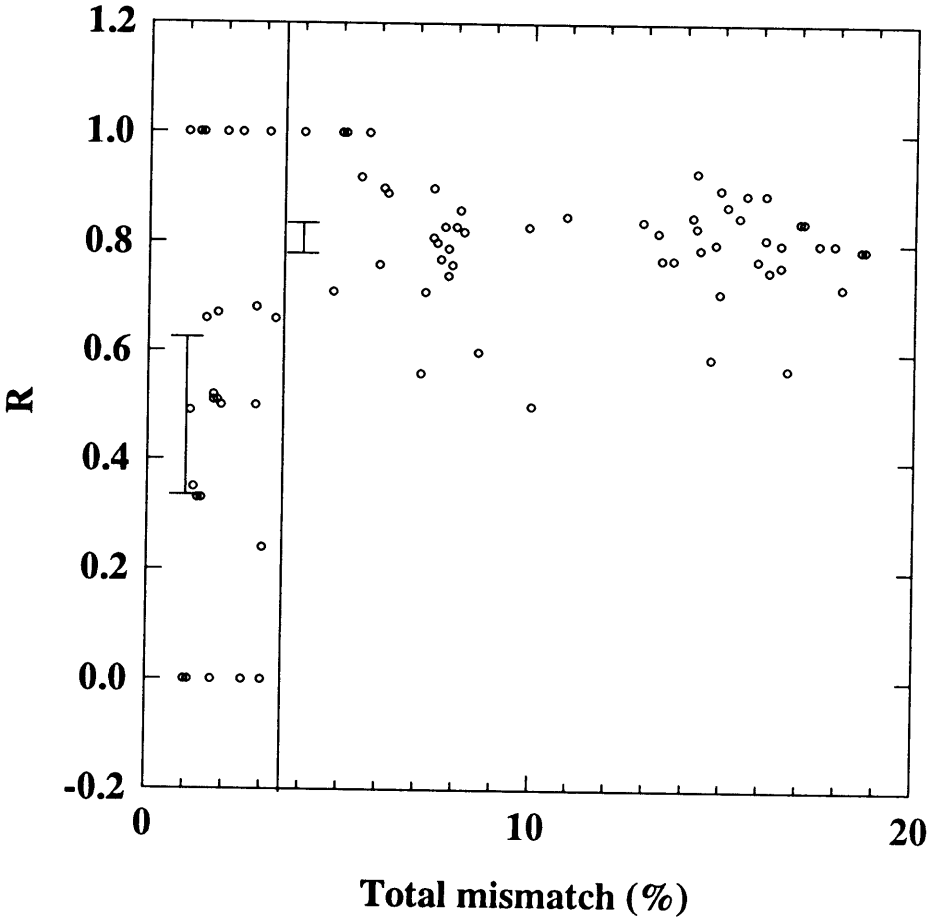

0.5
0.4
0.3
0.2
0.1
0.0
Mean Right - mean Left
0
10
20
Total mismatch (%)

Figure 3. Data from Figure 1 divided into two groups by the criterion (solid vertical line) drawn at 3.5% total mismatch. Data to the left of the criterion are comparisons between closely related sequences which, on average, exhibit low fractions of sequence mismatch at third codon positions emblematic of sequences which differ mostly due to PCR+SD error. Data to the right of the criterion are comparisons between sequences differing by greater than 3.5% total mismatch which, on average, exhibit higher fractions of sequence mismatch at third codon positions, characteristic of sequences differing due to evolutionary processes. Error bars enclose 95% confidence intervals for the mean of all datapoints to the left or right of the criterion.


1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
R
0
10
20
Total mismatch (%)

Conclusion. This Appendix presents both theoretical and practical methods for discriminating between true natural variation and PCR+SD error in the analysis of cloned sequences amplified from populations of mixed genetic composition. To my knowledge, both are currently absent from the literature. The practical method is applied to the data of Chapter Three for sorting clone sequences having alignable intergenic regions into Alleles corresponding to genotypes present in natural populations. Even without this analysis, the dataset of Chapter Three shows abundant genetic diversity, with 59 sets of *petB/D* clones having unalignable intergenic regions; differences among these sets must be attributed to natural variation and not to PCR+SD error. Application of the practical method to the analysis of clones having alignable intergenic regions adds 12 additional Alleles to the Chapter Three data used for population genetics analysis. The Alleles produced by subdivision of sets of clones having alignable intergenic regions exhibit an average fraction of sequence mismatch at third codon positions of 0.82 when compared to related Alleles, significantly different from the expectations of random error. Therefore, even if their prototype sequences contain errors, these Alleles serve to identify genotypes present in the natural populations and are legitimately included in the population genetics analysis.

REFERENCES

Dunning, A.M., Talmud, P. and Humphries, S.E. (1988). Errors in the polymerase chain reaction. Nucleic Acids Res. 16:10393.

Eckert, K.A. and Kunkel, T.A. (1991a). DNA Polymerase Fidelity and the Polymerase Chain Reaction. PCR Methods Applic. 1:17-24.

Eckert, K.A. and Kunkel, T.A. (1991b). The fidelity of DNA polymerases used in the PCR. In *Polymerase chain reaction: a practical approach* (eds. M.J. McPherson, P. Quirke and G.R. Taylor). IRL Press, Oxford, pp. .

Ennis, P.D., Zemmour, J., Salter, R.D. and Parham, P. (1990) Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: frequency and nature of errors produced in amplification. Proc. Natl. Acad. Sci. USA 87:2833-2837.

Goodenow, M., Huet, T., Saurin, W., Kwok, S., Sninsky, J. and Wain-Hobson, S. (1989). HIV-1 isolates are rapidly evovling quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. J. Acquir. Imm. Defic. Syn. 2:344-352.

Keohavong, P. and Thilly, W.G. (1989). Fidelity of DNA polymerases in DNA amplification. Proc. Natl. Acad. Sci. USA 86:9253-9257.

Luria, S.E. and Delbruck, M (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491-511.

Saiki, R.K., Gelfand, D.J., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487-491.

Sarkar, S. (1991). Haldane's solution of the Luria-Delbruck Distribution. Genetics 127:257-261.

Tindall, K.R., and Kunkel, T.A. (1988). Fidelity of DNA synthesis by the *Thermus aquaticus* DNA Polymerase. Biochemistry 27:6008-6013.

Appendix III

COMPARISON OF NUCLEOTIDE DIFFERENCES AMONG CLONES ASSIGNED TO ALLELE 1 TO KNOWN PATTERNS OF NUCLEOTIDE SUBSTITUTION ERROR BY TAQ POLYMERASE

FOR DATA OF CHAPTER THREE

This Appendix provides an analysis of the pattern of sequence mismatch among clones assigned to one "Allele," or set of clones within which sequence differences are considered indistinguishable from PCR and sequence determination (PCR+SD) error according to the 3.5% total sequence mismatch criterion (Appendix I), in order to investigate the hypothesis that their differences arose through Taq polymerase error during PCR. Results suggest that different sources of sequence variation, possibly sequence determination error or natural genetic variation, contribute to sequence mismatch within Types, in addition to Taq polymerase error.

Sequence mismatches among clones assigned to Allele 1, the allele containing the greatest number of clones in the dataset of Chapter Three, were compared to nucleotide substitutions characteristic of Taq polymerase error. Allele 1 was spawned from the subdivision of a larger set of clones having alignable intergenic regions (Appendix I) and contains a total of 44 clones, 11 of which are identical to the prototype sequence. 11 additional clones, 10 derived from the 70m Sargasso Sea sorted sample and one from the 50m Gulf Stream sample, contained the identical G->A substitution as their only mismatch with the prototype. This nucleotide substitution was present in the remaining S70 clone sequenced across the site in question, as well, and may have originated in a natural variant of Allele 1 *P. marinus* present at 70m in the Sargasso Sea[1]. If this nucleotide substitution is not attributed to PCR+SD error, then at least 22 out of 44 (50%) of the clone sequences classified as Allele 1 are error-free. Among other mismatches, T->C transitions were the most frequently encountered substitutions (14 out of 44 total substitutions), with C->T almost equally abundant (12 out of 44), followed by A->G (5 out of 44) and G->A (5 out of 44)

[1]Although it is possible that a significant fraction of *pet B/D* sequences cloned out of a pair of pooled PCR reactions, each beginning with about 10^5 sorted cells could contain the same PCR error, especially if this error occurred at a hotspot for PCR misincorporation (Keohavong and Thilly 1989), it seems unlikely that *all* of them would have the same error.

(Table 1). Detailed studies examining Taq polymerase errors have proven that most Taq polymerase errors are A->G or T->C transitions (Dunning et al. 1988, Saiki et al. 1988 Tindall and Kunkel 1988, Keohavong and Thilly 1989, Ennis et al 1990, Eckert and Kunkel 1991). Therefore, the type of sequence mismatches observed between clones assigned to Allele 1 and the Allele 1 prototype do not conform to expectations of variation due solely to Taq polymerase error. It therefore seems likely that natural genetic variation and/or sequence determination error contribute significantly to sequence variation within Alleles.

Table 1. Nucleotide mismatches between clones assigned to Allele1 and the Allele 1 prototype sequence, excluding G->A at position 1243.

Nucleotide Identities: Prototype->clone	Number of Occurrences	Number of Sites
T->C	14	13
T->A	2	2
T->G	2	2
C->T	12	7
C->A	2	2
G->A	5	5
G->T	1	1
A->G	5	3
A->T	1	1

REFERENCES

Dunning, A.M., Talmud, P. and Humphries, S.E. (1988). Errors in the polymerase chain reaction. Nucleic Acids Res. 16:10393.

Eckert, K.A. and Kunkel, T.A. (1991). The fidelity of DNA polymerases used in the PCR. In *Polymerase chain reaction: a practical approach* (eds. M.J. McPherson, P. Quirke and G.R. Taylor). IRL Press, Oxford, pp. .

Ennis, P.D., Zemmour, J., Salter, R.D. and Parham, P. (1990) Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: frequency and nature of errors produced in amplification. Proc. Natl. Acad. Sci. USA 87:2833-2837.

Keohavong, P. and Thilly, W.G. (1989). Fidelity of DNA polymerases in DNA amplification. Proc. Natl. Acad. Sci. USA 86:9253-9257.

Luria, S.E. and Delbruck, M (1943). Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491-511.

Saiki, R.K., Gelfand, D.J., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487-491.

Tindall, K.R., and Kunkel, T.A. (1988). Fidelity of DNA synthesis by the *Thermus aquaticus* DNA Polymerase. Biochemistry 27:6008-6013.

Appendix IV

ALLIGNED SEQUENCE DATA

Figure 1. 16S ribosomal RNA sequence data used to infer trees in Chapter 2, Figure 1. Mask on line 16 indicates nucleotide positions used in the analysis. Abbreviations: Agrotume: *Agrobacterium tumefaciens*; Med4, MIT9107, FP5, MIT9303: *P. marinus* cultured strains (this work); *P. marinus* SSW5 (Urbach et al. 1992); WH8103, WH7805, PCC6301: *Synechococcus* cultured strains; SAR6, SAR7, SAR139: cloned sequences from Sargasso Sea bacterioplankton; Anab7120: *Anabaena* PCC7120; Mpolyclp: *Marchantia polymorpha* chloroplast; Pchlron: *Prochloron sp.* (references in Chapter Two).

```
 Posi-   Sequence
 tion:   identity:   Data:

    1    1 Agrotume  CUCAACUUGAGAGUUUGAUCCUGGCUCAGAACGAACGCUGGCGGCAGGCUUAACACAUGCAAGUCGAACGCCCCGC-----------AAGGGGAGUGGCA  Agrotume
    1    2 Med4                                                                                                          Med4
    1    3 MIT9107                                                                                                       MIT9107
    1    4 FP5                                                                                                           FP5
    1    5 SSW5                                                                                                          SSW5
    1    6 MIT9303                                                                                                       MIT9303
    1    7 WH8103                                    --CGCUGGCGGCGUGCUUAACACAUGCAAGUCGAACGAACCUUC-----------GGGUUAGUGGCG  WH8103
    1    8 WH7805                                    ------------------------------------------------------------AGUGGCG  WH7805
    1    9 SAR6            ...AGAGUUCGAUCCUGGCUCAGGAUGAACGCUGGCGGCGUGCUUAACACAUGCAAGUCGAACGAAC-CUU---------CGG-GUUAGUGGCG  SAR6
    1   10 SAR7           ...GAGAGUUUGAUCCUGGCUCAGGAUGAACGCUGGCGGCGUGCUUAACACAUGCAAGUCGAACGAAC-CUU---------CGG-GUUAGUGGCG  SAR7
    1   11 SAR139                                                   UNANACAUGCAAGUCGAKCGAAC--UU---------CGG-GUUAGUGGCG  SAR139
    1   12 PCC6301    CAAAAUGGAGAGUUUGAUCCUGGCUCAGGAUGAACGCUGGCGGCGUGCUUAACACAUGCAAGUCGAACGGGCUCUU---------CGGAGCUAGUGGCG  PCC6301
    1   13 Anab7120    GAAACGGAGAGUUUGAUCCUGGCUCAGGAUGAACGCUGGCGGUAUGCUUAACACAUGCAAGUCGAACGGUCUCUU---------CGGAGAUAGUGGCG  Anab7120
    1   14 Mpolyclp   UCUCAUGGAGAGUUUGAUCCUGGCUCAGGAUGAACGCUGGCGGCAUGCUUAACACAUGCAAGUCGUACGGGAAGGAUCCUAGUGGUGUUUCCAGUGGCG  Mpolyclp
    1   15 Pchlron                                  GAUGAACGCUGGCGGUAUGCUUAACACAUGCAAGUCGAACGAACUCUU---------CGGAGUUAGUGGCG  Pchlron
    1   16 MASK                                                                                                         MASK

  101    1 Agrotume  GACGGGUGAGUAACGCGUGGGAAUCUACCGUGCCCUGCGGAAUAGCUCCGGGAAACUGGAAUUAAUACCGCAUACGCCCUACGGGGGAAAGAUU--UAUC  Agrotume
  101    2 Med4                        GAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCGAUAGGU-GAAAUGAA--UUUC  Med4
  101    3 MIT9107                     GAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCUACUGGU-GAAAUGAA--UUUC  MIT9107
  101    4 FP5                         GAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCGAGAGGU-GAAAUGAA--UUUC  FP5
  101    5 SSW5                        GAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCGAGAGGU-GAAAUGAA--UUUC  SSW5
  101    6 MIT9303                                 GGGAUAACGGUUGGAAACGGCCGCUAAUACCCCAUAUGCCGAGAGGU-GAAAUGAA--UUUC  MIT9303
  101    7 WH8103    GACGGGUGAGUAACGCGUGAGAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUYYCCYAUAUGCCGCGAGGU-GAAAUGAA--UUUC  WH8103
  101    8 WH7805    GACGGGUGAGUAACGCGUGGGAAUCUGCCCUCAGGAGGGGGAUAACGGCUGGAAACGGCCGCUAAUACCCUAUAUGCCGAGAGGU-GAAAUGAA--UUUC  WH7805
  101    9 SAR6      GACGGGUGAGUAACGCGUGAGAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCGAUAGGU-GAAAUGAA--UUUC  SAR6
  101   10 SAR7      GACGGGUGAGUAACGCGUGAGAAUCUGCCCUCAGGAGGGGGAUAACAGCUGGAAACGGCUGCUAAUACCCCAUAUGCCGCGAGGU-GAAAUGAA--UUUC  SAR7
  101   11 SAR139    GACGGGUGAGUAACGCGUGAGAAUCUGCCCUCAGGAGGGGGAUAACGGUUGGAAACGACCGCUAAUACCCCAUAUGCCGCGAGGU-GAAAUGAA--UUUC  SAR139
  101   12 PCC6301   GACGGGUGAGUAACGCGUGAGAAUCUGCCUACAGGACGGGGACAACAGUUGGAAACGACUGCUAAUACCCGAUGUGCCGAGAGGU-GAAACAUU--UAUG  PCC6301
  101   13 Anab7120  GACGGGUGAGUAACGCGUGAGAAUCUAGCUUCAGGUCGGGGACAACCACUGGAAACGGUGGCUAAUACCGGAUGUGCCGAAAGGU-GAAAGAUU--UAUU  Anab7120
  101   14 Mpolyclp  GACGGGUGAGUAACGCGUAAGAACCUGCCCUUGGGAGGGGGACAACAGCUGGAAACGGUUGCUAAUACCCCAUAGGCUGAGGAGC-AAAAGGAGGAAUCC  Mpolyclp
  101   15 Pchlron   GACGGGUGAGUAACGCGUGAGAAUCUACCUCAAGGACGGGGACAACAGCAGGAAACUGGUGCUAAUACCCGAUAAGCUGAAAAGU-GAAAGAUU--UAUU  Pchlron
  101   16 MASK                                    11111111111111111111111111--11-11111111111111-11111111--1111  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

  201    1 Agrotume  GGGGUAUGAUGAGCCCGCGUUGGAUUAGCUAGUUGGUG-GGGUAAAGGCCUACCAAGGCGACGAUCCAUAGCUGGUCUGAGAGGAUGAUCAGCCACAUUG  Agrotume
  201    2 Med4      GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUUGUUGGUG-AGGUAAUGGCUCACCAAGGCUUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  Med4
  201    3 MIT9107   GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUG-AGGUAAUGGCUCACCAAGGCUUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  MIT9107
  201    4 FP5       GCCUGAGGAUGAGCCCGCGUCUGAUUAGCUUGUUGGUG-AGGUAAUGGCUCACCAAGGCAUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  FP5
  201    5 SSW5      GCCUGAGGAUGAGCCCGCGUCUGAUUAGCUUGUUGGUG-AGGUAAUGGCUCACCAAGGCUUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  SSW5
  201    6 MIT9303   GCCUGAGGAUGAGCCCGCGUCUGAUUAGCUAGUUGGUG-AGGUAAUGGCUCACCAAGGCAUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  MIT9303
  201    7 WH8103    GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUG-GGGUAAAGGCUUACCAAGGCAUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  WH8103
  201    8 WH7805    GCCUGAGGAUGAGCCCGCGUCUGAUUAGCUAGUUGGUG-AGGUAAGAGCUUACYAAGGCGNCGAUCAGUAGCUNGUCUGAGAGGAUGAUCAGCCACACUG  WH7805
  201    9 SAR6      GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUUGUUGGUG-AGGUAAUGGCUCACCAAGGCUUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  SAR6
  201   10 SAR7      GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUGUAGGUAAAGGCUCACCAAGGCAUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  SAR7
  201   11 SAR139    GCCUGAGGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUG-GGGUAAAGGCCUACCAAGGCAUCGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  SAR139
  201   12 PCC6301   GCCUGUAGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUG-GGGUAAGGGCCUACCAAGGCGACGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  PCC6301
  201   13 Anab7120  GCCUGAAGAUGAGCUCGCGUCUGAUUAGCUAGUUGGUG-UGGUAAGAGCGCACCAAGGCGACGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  Anab7120
  201   14 Mpolyclp  GCCUAAGGAGGGGCUUGCGUCUGAUUAGCUAGUUGGUG-AGGUAAUAGCUUACCAAGGCGACGAUCAGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  Mpolyclp
  201   15 Pchlron   GCCGAGAGAGGAGCUCGCGUCCGAUUAGCUAGUUGGAA-GGGUAAAAGCCUACCAAGGCGACGAUCGGUAGCUGGUCUGAGAGGAUGAUCAGCCACACUG  Pchlron
  201   16 MASK      11111111111111111111111111111111111111-11111111111111-111111-111111111111-1111111111111111111111111111  MASK

  301    1 Agrotume  GGACUGAGACACGGCCCAAACUCCUACGGGAGGCAGCAGUGGGGAAUAUUGGACAAUGGGCGCAAGCCUGAUCCAGCCAUGCCGCGUGAGUGAUGAAGGC  Agrotume
  301    2 Med4      GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAACG-UGACGGAGCAACGCCGCGUGAGGGACGAAGGC  Med4
  301    3 MIT9107   GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGC-UGACGGAGCAACGCCGCGUGAGGGACGAAGGC  MIT9107
  301    4 FP5       GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCSUGACGGAGCAACGCCGCGUGAGGGACGAAGGC  FP5
  301    5 SSW5      GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCCUGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  SSW5
  301    6 MIT9303   GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGCAAGCCUGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  MIT9303
  301    7 WH8103    GGACUGAGACACGGCCCAGACUCCUUCGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCCUGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  WH8103
  301    8 WH7805    GGACUGAGACACGGCCCUGUCUCCUUUGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCCUGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  WH7805
  301    9 SAR6      GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAACG-UGACGGAGCAACGCCGCGUGAGGGACGAAGGC  SAR6
  301   10 SAR7      GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGC-UGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  SAR7
  301   11 SAR139    GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGC-UGACGGAGCAACGCCGCGUGAGGGAUGAAGGC  SAR139
  301   12 PCC6301   GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGCAAGC--GACGGAGCAACGCCGCGUGGGGGAGGAAGGU  PCC6301
  301   13 Anab7120  GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCCUGACGGAGCAAUACCGCGUGAGGGAGGAAGGC  Anab7120
  301   14 Mpolyclp  GGACUGAGACACGGCCCAGACUCUUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAA-CGUGACGGAGCAAUGCCGCGUGGAGGUAGAAGGC  Mpolyclp
  301   15 Pchlron   GGACUGAGACACGGCCCAGACUCCUACGGGAGGCAGCAGUGGGGAAUUUUCCGCAAUGGGCGAAAGCCUGACGGAGCAAUACCGCGUGAGGGAAGAAGGC  Pchlron
  301   16 MASK      11111111111111111111111111111111111111111111111111111111111111111111----11111111111111111111111111111  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

  401    1 Agrotume  CUUAGGGUUGUAAAGCUCUUUCACCGGAGAAGAU---AAUGACGGUAUCCGGAGAAGAAGCCCCGGCUAACUUCGUGCCAGCAGCCGCGGUAAUACGAAG  Agrotume
  401    2 Med4      CUCUGGGCUGUAAACCUCUUUUCUCAAGGAAGAAGA-UAUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  Med4
  401    3 MIT9107   CUCUGGGCUGUAAACCUCUUUUCUCAAGGAAGAAGA-UAUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  MIT9107
  401    4 FP5       CUCUGGGCUGUAAACCUCUUUUCUCAAGGAAGAAGA-UAUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  FP5
  401    5 SSW5      CUCUGGGCUGUAAACCUCUUUUCUCAAGGAAGAAGA-UAUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  SSW5
  401    6 MIT9303   CUCUGGGCUGUAAACCUCUUUUCUCAAGGAAGAAGA-UCUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  MIT9303
  401    7 WH8103    CUCUGGGCUGURAACCYYUUUUAUCAAGGAAGAAYGAUCUGACGGUACUUGAUGAAUAAGCCACGGCUHAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  WH8103
  401    8 WH7805    CUCUGGGCUGUUAACCUNUUUUCUCAAGGAAGAAGNAUCUGACGGUACUUGAGGAAUAAGCCACGGCUHAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  WH7805
  401    9 SAR6      CUCUGGGCUGUAAACCUCUUUUUUCAAGGAAGAAGA-UAUGACGGUACUUGAGGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  SAR6
  401   10 SAR7      CUCUGGGCUGUAAACCUCUUUUAUCAAGGAAGAAGA-UCUGACGGUACUUGAUGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  SAR7
  401   11 SAR139    CUCUGGGCUGUAAACCUCUUUUAUCAAGGAAGAAGA-UCUGACGGUACUUGAUGAAUAAGCCACGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  SAR139
  401   12 PCC6301   UUUUGGACUGUAAACCCCUUUUCUCAGGGAAGAAGAAAGUGACGGUACCUGAGGAAUAAGCCUCGGCUAAUUCCGUGCCAGCAGCCGCGGUAAUACGGGA  PCC6301
  401   13 Anab7120  UCUUGGGUUGUAAACCUCUUUUCUCAGGGAAUAAAAAAAUGAAGGUACCUGAGGAAUAAGCAUCGGCUAACUCCGUGCCAGCAGCCGCGGUAAUACGGAG  Anab7120
  401   14 Mpolyclp  UCACGGGUCGUAAACUCCUUUUCUCAGAGAAGAUG-CAAUGACGGUAUCUGAGGAAUAAGCAUCGGCUAACUCUGUGCCAGCAGCCGCGGUAAGACAGAG  Mpolyclp
  401   15 Pchlron   UUUUGGGUUGUAAACCUCUUUUGUCAGGGAAGAAGA-AUUGACGGUACCUGAAGAAAAAGCAUCGGCUAACUCCGUGCCAGCAGCCGCGGUAAUACGGAG  Pchlron
  401   16 MASK      11111111111-1111--111111111111111------11111111111111111111111111111-11111111111111111111111111111  MASK

  501    1 Agrotume  GGGGCUAGCGUUGUUCGGAAUUACUGGGCGUAAAGCGCACGUAGGCGGAUAUUUAAGUCAGGGGUGAAAUCCCAGAGCUCAACUCUGGAACUGCCUUUGA  Agrotume
  501    2 Med4      GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCUUUCAAGUCUGCUGUUAAAGCGUGGAGCUUAACUCCAUUAUGGCAGUGGA  Med4
  501    3 MIT9107   GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCUUUUCAAGUCUGCUGUUAAAACGUGGAGCUUAACUCCAUCAUGGCAGUGGA  MIT9107
  501    4 FP5       GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCUUCCAAGUCUGCUGUUAAAGCGUGGAGCUUAACUCCAUCAUGGCAGUGGA  FP5
  501    5 SSW5      GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCUUUCAAGUCUGCUGUAAAAACGUGGAGCUUAACUCCAUCAUCGCAGUGGA  SSW5
  501    6 MIT9303   GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCUUUUCAAGUCUGCUGUUAAAACGUGGAGCUUAACUCCAUCAAGGCAGUGGA  MIT9303
  501    7 WH8103    GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCCUUCAAGUCUGCUNUUAAAAAGUGGAGCUUAACUCCAUCAUGGCAGUGGA  WH8103
  501    8 WH7805    GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCCUUCAAGUCUGCUNUUAAAACGUGGAGCUUAACUCCAUCAUGGCAGUGGA  WH7805
  501    9 SAR6      GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCUUUCAAGUCUGCUGUUAAAGCGUGGAGCUUAACUCCAUUAUGGCAGUGGA  SAR6
  501   10 SAR7      GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCCAGAAAGUCUGCUGUUAAAACGUGGAGCUUAACUCCAUCAUGGCAGUGGA  SAR7
  501   11 SAR139    GUGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGUCCGCAGGCGGCCCUUCAAGUCUGCUGUUAAAAAGUGGAGCUUAACUCCAUCAUGGCAGUGGA  SAR139
  501   12 PCC6301   GAGGCAAGCGUUAUCCGGAAUUAUUGGGCGUAAAGCGCCUGCAGGCGGUUAAUCAAGUCUGUUGUCAAAGCGUGGGGCUCAACCUCAUACAGGCAAUGGA  PCC6301
  501   13 Anab7120  GAUGCAAGCGUUAUCCGGAAUGAUUGGGCGUAAAGCGUCCGCAGGUGGCACUGUAAGUCUGCUGUUAAAGAGCAAGGCUCAACCUUGUAAAGGCAGUGGA  Anab7120
  501   14 Mpolyclp  GAUGCAAGCGUUAUCCGGAAUGAUUGGGCGUAAAGCGUCUGUAGGUGGCUUUUUAAGUCCGCCGUCAAAUCCCAGGGCUCAACCCUGGACAGGCGGUGGA  Mpolyclp
  501   15 Pchlron   GAUGCAAGCGUUAUCCGGAAUCAUUGGGCGUAAAGGGUCCGCAGGUGGCGAAUAAAGUCUGCUGUUAAAGACCGGGGCUAAACUCCGGAAAGGCAGUGGA  Pchlron
  501   16 MASK      11111111111111111111111111111111111111111111111111111111111111111-11111111111111111111111111111111  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

  601    1 Agrotume  UACUGGGUAUCUUGAGUAUGGAAGAGGUAAGUGGAAUUCCGAGUGUAGAGGUGAAAUUCGUAGAUAUUCGGAGGAACGCCAGUGGCGAAGG-CGGCUUAC  Agrotume
  601    2 Med4      AACUGAUCGGCUUGAGUAUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGGgCGCUCUGC  Med4
  601    3 MIT9107   AACUGAAAGGCUUGAGUAUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  MIT9107
  601    4 FP5       AACUGAUUGGCUUGAGUAUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  FP5
  601    5 SSW5      AACUGAAGGGCUCGAGUGUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  SSW5
  601    6 MIT9303   AACUGAUUGGCUUGAGUGUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  MIT9303
  601    7 WH8103    AACUGUUGGGYYUGAGUGUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  WH8103
  601    8 WH7805    AACUGUUGGGCNCGAGUGUGGUAGGGGCAGAGGGUAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  WH7805
  601    9 SAR6      AACUGAUCGGCUUGAGUAUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  SAR6
  601   10 SAR7      AACUGUUGGGCUUGAGUGUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  SAR7
  601   11 SAR139    AACUGUUGGGCUUGAGUGUGGUAGGGGCAGAGGGAAUUCCCGGUGUAGCGGUGAAAUGCGUAGAUAUCGGGAAGAACACCAGUGGCGAAGG-CGCUCUGC  SAR139
  601   12 PCC6301   AACUGAUUGACUAGAGUAUGGUAGGGGUAGCGGGAAUUCCAGGUGUAGCGGUGAAAUGCGUAGAUAUCUGGAAGAACACCAGCGGCGAAAG-CGCGCUAC  PCC6301
  601   13 Anab7120  AACUACAGAGCUAGAGUACGUUCGGGGCAGAGGGAAUUCCUGGUGUAGCGGUGAAAUGCGUAGAGAUCAGGAAGAACACCGGUGGCGAAAG-CGCUCUGC  Anab7120
  601   14 Mpolyclp  AACUACCAAGCUGGAGUACGGUAGGGGCAGAGGGAAUUUCCGGUGGAGCGGUGAAAUGCGUAGAGAUCGGAAAGAACACCAAUGGCGAAAG-CACUCUUC  Mpolyclp
  601   15 Pchlron   AACUGAUUAGCUAGAGUACGGUAGGGGUAGAGGGAAUUCCCAGUGUAGCGGUGAAAUGCGUAGAUAUUGGGAAGAACACCGGUGGCGAAAG-CGCUCUAC  Pchlron
  601   16 MASK      1111111111--11111111111111111111111111111111111111111111111111111111111111111111111111111111-11111111  MASK

  701    1 Agrotume  UGGUCCAUUACUGACGCUGAGGUGCGAAAGCGUGGGGAGCAAACAGGAUUAGAUACCCUGGUAGUCCACGCCGUAAACGAUGAAUGUUAGCCGUCGGGCA  Agrotume
  701    2 Med4      UGGGCCAUUACUGACGCUCAU-GACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  Med4
  701    3 MIT9107   UGGGCCAUUACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  MIT9107
  701    4 FP5       UGGGCCAUUACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  FP5
  701    5 SSW5      UGGGCCAUAACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  SSW5
  701    6 MIT9303   UGGGCCAUAACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  MIT9303
  701    7 WH8103    UGGGCCAUCACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  WH8103
  701    8 WH7805    UGGGCCAUAACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUUGGGGG  WH7805
  701    9 SAR6      UGGGCCAUUACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  SAR6
  701   10 SAR7      UGGGCCAUUACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  SAR7
  701   11 SAR139    UGGGCCAUCACUGACGCUCAUGGACGAAAGCCAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUGGCCGUAAACGAUGAACACUAGGUGUCGGGGG  SAR139
  701   12 PCC6301   UGGGCCAUAACUGACGCUCAUGGACGAAAGCUAGGGGAGCGAAAGGGAUUAGAUACCCCUGUAGUCCUAGCCGUAAACGAUGAACACUAGGUGUUGCGUG  PCC6301
  701   13 Anab7120  UAGGCCGUAACUGACACUGAGGGACGAAAGCUAGGGGAGCGAAUGGGAUUAGAUACCCCAGUAGUCCUAGCCGUAAACGAUGGAUACUAGGCGUGGCUUG  Anab7120
  701   14 Mpolyclp  UGGGCCGACACUGACACUGAGAGACGAAAGCUAGGGGAGCAAAUGGGAUUAGAUACCCCAGUAGUCCUAGCCGUAAACGAUGGAUACUAAGCGCUGUGC-  Mpolyclp
  701   15 Pchlron   UAGGCCGAAACUGACACUGAGGGACGAAAGCUAGGGUAGCGAAAGGGAUUAGAUACCCCUGUAGUCUUAGCGGUAAACGAUGGAUACUAGGUGUAGCUUG  Pchlron
  701   16 MASK      1111111111111111111111-11111111111111111111111111111111111111111111111111111111111111111111111111-  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

  801    1 Agrotume  GUA--UACUGUUCGGUGGCGCAGCUAACGCAUUAAACAUUCCGCCUGGGGAGUACGGUCGCAAGAUUAAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  Agrotume
  801    2 Med4      AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  Med4
  801    3 MIT9107   AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  MIT9107
  801    4 FP5       AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGGUUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  FP5
  801    5 SSW5      AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  SSW5
  801    6 MIT9303   AAUCGACCCCCUCGGUGUCGUAGCCAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  MIT9303
  801    7 WH8103    AAUCGACCCCCUCGGUGUCGUAGCCAACGCGUUAAGUGUNCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  WH8103
  801    8 WH7805    AAUCGACCCCCUCGGUGUCGUAGCCAACGCGUUAAGUGUUCCGCYUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  WH7805
  801    9 SAR6      AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGU-UGAAACUCAAAGGAAUUGACGGGNNNCCGCACAA  SAR6
  801   10 SAR7      AAUCGACCCCUUCGGUGUCGUAGCUAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  SAR7
  801   11 SAR139    AAUCGACCCCCUCGGUGUCGUAGCCAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGCGCCGCACAA  SAR139
  801   12 PCC6301   AAUCGACCCGCGCAGUGCCGUAGCCAACGCGUUAAGUGUUCCGCCUGGGGAGUACGCACGCAAGUUGGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  PCC6301
  801   13 Anab7120  UAUCGACCCGAGCCGUGCCGGAGCCAACGCGUUAAGUAUCCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  Anab7120
  801   14 Mpolyclp  UAUCGACCCGUGCAGUGCUGUAGCUAACGCGUUAAGUAUCCCGCCUGGGGAGUACGUUCGCAAGAAUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  Mpolyclp
  801   15 Pchlron   UAUCGACCCGAGCUGUGCCGAAGCUAACGCGAUAAGUAUCCCGCCUGGGGAGUACGCACGCAAGUGUGAAACUCAAAGGAAUUGACGGGGGCCCGCACAA  Pchlron
  801   16 MASK      111--11111111111111111111111111111111-1111-1111111111111111111-11111111111111111111111---11111111  MASK

  901    1 Agrotume  GCGGUGGAGCAUGUGGUUUAAUUCGAAGCAACGCGCAGAACCUUACCAGCUCUUGACAUUCGGGGUUUGGGCAGUGGAGACAUUGUCCUUCAGUUAGGCU  Agrotume
  901    2 Med4      GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGAGAA--CCUCUUAGAAAUUAGAGG-GUGCCUUCGGGA  Med4
  901    3 MIT9107   GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGAGAA--CCUCUUAGAAAUUUGAGG-GUGCCUUCGGGA  MIT9107
  901    4 FP5       GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGCUUGACAUCCUGCGAA--CCUUUAAGAAAUUAGAGG-GUGCCUUCGGGA  FP5
  901    5 SSW5      GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGAUUUGACAUCCUGCGAA--CCUCUUAGAAAUUUGAGG-GUGCCUUCGGGA  SSW5
  901    6 MIT9303   GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGCGAA--UCUCUUAGAAACKAGAGA-GUGCCUUUGGGA  MIT9303
  901    7 WH8103    GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGCGAA--UCUCUUGGAAACGAGAGA-GUGNCUUCGGGA  WH8103
  901    8 WH7805    GCGGUGGAGUAUGUGGUUUAAUUCGAUGNAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGCGAA--CCUCUUGGAAWCGAGAGG-GUGCCUUCGGGA  WH7805
  901    9 SAR6      GCG-UGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGAGAA--CCUCUUAGAAAUUUGAGG-GUGCCUUCGGGA  SAR6
  901   10 SAR7      GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGAGAA--CCUCUUAGAAAUUUGAGG-GUGCCUUCGGGA  SAR7
  901   11 SAR139    GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCUGCGAA--UCUCUUGGAAACGAGAGA-GUGCCUUCGGGA  SAR139
  901   12 PCC6301   GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGUUUGACAUCCCCCGAA--UCUCUUGGAAACGAGAGA-GUGCCUUCGGGA  PCC6301
  901   13 Anab7120  GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAAGACUUGACAUGUCGCGAA--UCUUCUUGAAAGGGAAGA-GUGCCUUAGGGA  Anab7120
  901   14 Mpolyclp  GCGGUGGAGCAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGCUUGACAUGCCGUGAA--UCUUUUUGAAAGAAAAGA-GUGCCUUCGGGA  Mpolyclp
  901   15 Pchlron   GCGGUGGAGUAUGUGGUUUAAUUCGAUGCAACGCGAAGAACCUUACCAGGGCUUGACAUGUCGCGAA--UCUAGCGGAAACGUUAGA-GUGCCUUAGGGA  Pchlron
  901   16 MASK      111-1111111111111111111111-11111111111111111111111111111111111111--1111111111-1-11111-111-11111111  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

1001    1 Agrotume  GGCCCCAGAACAGGUGCUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCUCGCCCUUAGUUGCCAGCAU  Agrotume
1001    2 Med4      AUUCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  Med4
1001    3 MIT9107   AUUCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  MIT9107
1001    4 FP5       ACGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  FP5
1001    5 SSW5      ACGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  SSW5
1001    6 MIT9303   ACGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUU-UAGUuGCcAGCAU  MIT9303
1001    7 WH8103    ACGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGYUUUUAGUUGCCAGCAU  WH8103
1001    8 WH7805    GCGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  WH7805
1001    9 SAR6      AUUCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAGCAU  SAR6
1001   10 SAR7      AUUCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUCUUUAGUUGCCAGCAU  SAR7
1001   11 SAR139    ACGCAGUG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUCUUUAGUUGCCAGCAU  SAR139
1001   12 PCC6301   GCGGGGAG-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCACGUUUUUAGUUGCCAUCAU  PCC6301
1001   13 Anab7120  GCGCGAAC-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCUCGUUUUUAGUUGCCAGCAU  Anab7120
1001   14 Mpolyclp  ACGCGGAC-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGCCGUAAGGUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCUCUUGUUUAGUUGCCAUCAU  Mpolyclp
1001   15 Pchlron   AUGCGAAC-ACAGGUGGUGCAUGGCUGUCGUCAGCUCGUGUCGUGAGAUGUUGGGUUAAGUCCCGCAACGAGCGCAACCCUCGUCCUUAGUUGCCAAUAU  Pchlron
1001   16 MASK      11111111-11111111111111111111111111111111111111111111111111111111111111111111111111111111-11-111111111111  MASK

1101    1 Agrotume  UUAGUUGGGCACUCUAAGGGGACUGCCGGUGAUAAGCCGAGAGGAAGGUGGGGAUGACGUCAAGUCCUCAUGGCCCUUACGGGCUGGGCUACACACGUGC  Agrotume
1101    2 Med4      UUAGUUGGGCACUCUAGAAAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACACCCUGGGCUACACACGUAC  Med4
1101    3 MIT9107   UUAGUUGGGCACUCUAGAAAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACACCCUGGGCUACACACGUAC  MIT9107
1101    4 FP5       UCAGUUGGGCACUCUAGAAAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACGUCCUGGGCUACACACGUAC  FP5
1101    5 SSW5      UUAGUUGGGCACUCUAGAAAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  SSW5
1101    6 MIT9303   UCAGUUGGGCACUCUAGAGAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  MIT9303
1101    7 WH8103    UUAGUUGGGCACUCUAGAGAGACCGCCGGUGAUAAACUG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  WH8103
1101    8 WH7805    UUAGUUGGGCACUCUAGAGAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  WH7805
1101    9 SAR6      UUAGUUGGGCACUCUAGAAAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACACCCUGGGCUACACACGUAC  SAR6
1101   10 SAR7      UUAGUUGGGCACUCUAGAGAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  SAR7
1101   11 SAR139    UUAGUUGGGCACUCUAGAGAGACCGCCGGUGAUAAACCG-GAGGAAGGUGUGGAUGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  SAR139
1101   12 PCC6301   UCAGUUGGGCACUCUAGAGAAACUGCCGGUGACAAACCG-GAGGAAGGUGUGGACGACGUCAAGUCAUCAUGCCCCUUACAUCCUGGGCUACACACGUAC  PCC6301
1101   13 Anab7120  UAAGUUGGGCACUCUAGAGAGACUGCCGGUGACAAACCG-GAGGAAGGUGGGGAUGACGUCAAGUCAGCAUGCCCCUUACGUCUUGGGCUACACACGUAC  Anab7120
1101   14 Mpolyclp  UAAGUUUGGAACCCUAAACAGACUGCCGGUGAUAAGCCG-GAGGAAGGUGAGGAUGACGUCAAGUCAGCAUGCCCCUUACGCCCUGGGCGACACACGUGC  Mpolyclp
1101   15 Pchlron   UAAGUUAGGCACUCUAGGGAGACUGCCGGGGACAACUCG-GAGGAAGGUGGGGAUGACGUCAAGUCAGCAUGCCCCUUACGUCCUGGGCGACACACGUAC  Pchlron
1101   16 MASK      1111111111111111111111111111111111111111-11111111111111111111111111111111111111111111111111111111111111  MASK
```

```
Posi-   Sequence
tion:   identity:   Data:

 1201    1 Agrotume  UACAAUGGUGGUGACAGUGGGCAGCGAGACAGCGAUGUCGAGCUAAAUCUC-C-AAAAGCCAUCUCAGUUCGGAUUGCACUCUGCAACUCGAGUGCAUGAA  Agrotume
 1201    2 Med4      UACAAUGCUACG-ACAAAGGGCAGCAAACUCGCGAGAGCUAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAG                      Med4
 1201    3 MIT9107   UACAAUGCUACGGACAAAGGGCAGCAAACUCGCGAGAGCUAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAG                      MIT9107
 1201    4 FP5       UACAAUGCUACGGACAAAGGGCAGCAAACUCGCGAGAGCUAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAG                      FP5
 1201    5 SSW5      UACAAUGCUACGGACAAAGGGCAGCAAACUCGCGAGAGCAAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAG                      SSW5
 1201    6 MIT9303   UACAAUGCUACGGACAAAGGGCAGCAAGYUCGCGAGGACAAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAG                      MIT9303
 1201    7 WH8103    UACAAUGCUACGGACAWAGGGCAGCAAGUUCGCGAGGACAAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAGGCUGCAACUCGCCUACGUGAA  WH8103
 1201    8 WH7805    UACAAUGCUACGGACAAAGAGCAGCAAGUUCGCGAGGACAAGCAAAUCUCAU-AAACCGUGGCUCAGUUCAGAUCGUAGGCUGCAACUCGCCUACGUGAA  WH7805
 1201    9 SAR6      UACAAUGCUACGGACAAAGGGCAGCACACUCGCGAGAGCUAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAGGCUGCAACUCGCCUACGUGAA  SAR6
 1201   10 SAR7      UACAAUGCUACGGACAAGGGGCAGCAAGUUCGCGAGGACAAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAGGCUGCAACUCGCCUACGUGAA  SAR7
 1201   11 SAR139    UACAAUGCUACGGACAAAGGGCAGCAACUUCGCGAGGACAAGCAAAUCCCAU-AAACCGUGGCUCAGUUCAGAUCGUAGGCUGCAACUCGCCUACGUGAA  SAR139
 1201   12 PCC6301   UACAAUGCUCCGGACAGCGAGACGCGAAGCCGCGAGGUGAAGCAAAUCUCCC-AAACCGGGGCUCAGUUCAGAUUGCAGGCUGCAACUCGCCUGCAUGAA  PCC6301
 1201   13 Anab7120  UACAAUGCUACGGACAGAGGGCAGCAAGCUAGCGAUAGCAAGCAAAUCCCGU-AAACCGUAGCUCAGUUCAGAUCGCAGGCUGCAACUCGCCUGCGUGAA  Anab7120
 1201   14 Mpolyclp  UACAAUGGCCGGGACAAAGGGUCGCGACCUCGCGAGAGAAAGCUAACCUCAA-AAACCCGGCCUCAGUUCGGAUUGCAGGCUGCAACUCGCCUGCAUGAA  Mpolyclp
 1201   15 Pchlron   UACAAUGGUUGGGACAAAGGGCAGCAAACUCGCGAGAGUAAGCGAAUCUCAUCAAACCCAGCCACAGUUCAGAUUGCAGGCUGCAACUCGCCUGCAUGAA  Pchlron
 1201   16 MASK      111111111111-111-11111111111-11111111111111111111---11111111111111111111111111                    MASK

 1301    1 Agrotume  GUUGGAAUCGCUAGUAAUCGCAGAUCAGC-AUGCUGCGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAUGGGAGUUGGUUUUACCCG  Agrotume
 1301    2 Med4                                                                                                          Med4
 1301    3 MIT9107                                                                                                       MIT9107
 1301    4 FP5                                                                                                           FP5
 1301    5 SSW5                                                                                                          SSW5
 1301    6 MIT9303                                                                                                       MIT9303
 1301    7 WH8103    GGAGGAAUCGCUAGUAAUCGCAGGUCAGC-AUAC----------------                                                   WH8103
 1301    8 WH7805    GGAGGAAUCGCUAGUAAUCGCAGGUCAGC-AUACUGCGG----------                                                   WH7805
 1301    9 SAR6      GUAGGAAUCGCUAGUAAUCGCAGGUCAGC-AUACUGCGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCCCACCAUGGAAGUUGGCCAUGCCCG  SAR6
 1301   10 SAR7      GGAGGAAUCGCUAGUAAUCACAGGUCAGC-AUACUGCGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAUGGAAGUUGGCCACGCCCG  SAR7
 1301   11 SAR139    GGAGGAAUCGCUAGUAAUCGCAGGUCAGC-AUACUGCGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGU                        SAR139
 1301   12 PCC6301   GGCGGAAUCGCUAGUAAUCGCAGGUCAGC-AUACUGCGGUGAAUACGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAUGGAAGUUGGCCAUGCCCG  PCC6301
 1301   13 Anab7120  GGAGGAAUCGCUAGUAAUUGCAGGUCAGC-AUACUGCAGUGAAUUCGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAUGGAAGCUGGCAACGCCCG  Anab7120
 1301   14 Mpolyclp  GCCGGAAUCGCUAGUAAUCGCCGGUCAGCCAUACGGCGGUGAAUCCGUUCCCGGGCCUUGUACACACCGCCCGUCACACUAUGGGAGCUGGCCAUGCCCG  Mpolyclp
 1301   15 Pchlron   GUAGGAAUCGCUAGUAAUCGCAGGUCAGC-AUACUGCGGUGAAUUCGUUCCCGGGCCUUGUACACACCGCCCGUCACACCAUGGAAGCUGGCUACGCCCG  Pchlron
 1301   16 MASK                                                                                                          MASK
```

Position:	Sequence identity:	Data:	
1401	1 Agrotume	AAGGUAGUGCGCUAACCG-C--AAGGAGGCAGCUAACCACGGUAGGGUCAGCGACUGGGGUGAAGUCGUAACAAGGUAGCCGUAGGGGAACCUGCGGCUG	Agrotume
1401	2 Med4		Med4
1401	3 MIT9107		MIT9107
1401	4 FP5		FP5
1401	5 SSW5		SSW5
1401	6 MIT9303		MIT9303
1401	7 WH8103		WH8103
1401	8 WH7805		WH7805
1401	9 SAR6	AAGUCGUUACUCCAACCC-UU-GUGGAGGAGGACGCCGAA-GUGGGGCUAAUGACUGGGGUGAAGUCGUAACAAGGUAGCCGUACCGGAAGGUGGGGCUG	SAR6
1401	10 SAR7	AAGCCGUUACUCNAACCC-U-CG-GGAGGAGGAAGUN-NAGGUGGGGCUGAUGACUGGGGUNAAGUCGUAACAAGGUAUCCGUACCGGAAGGUGCGGUUG	SAR7
1401	11 SAR139		SAR139
1401	12 PCC6301	AAGUCGUUACCCUAACCG-UUCGCGGAGGGGGGCGCCGAAGGUAGGGCUGAUGACUGGGGUGAAGUCGUAACAAGGUAGCCGUACCGGAAGGUGUGGCUG	PCC6301
1401	13 Anab7120	AAGUCAUUACUCCAACUU-UUAGGAGAGGAGGAUGCCUAAGGCAGUGCUGGUGACUGGGGUGAAGUCGUAACAAGGUAGCCGUACCGGAAGGUGUGGCUG	Anab7120
1401	14 Mpolyclp	AAGUCGUUACUCUAACCG-UA--AGGAGGGGGGUGCCGAA-CAGGGGCUAGUGACUGGAGUGAAGUCGUAACAAGGUAGCCGUACUGGAAGGUGCGGCUG	Mpolyclp
1401	15 Pchlron	AAGUCGUUACCCUAACCGGUUUUGGGA	Pchlron
1401	16 MASK		MASK

1501	1 Agrotume	GAUCACCUCCUUUCC	Agrotume
1501	2 Med4		Med4
1501	3 MIT9107		MIT9107
1501	4 FP5		FP5
1501	5 SSW5		SSW5
1501	6 MIT9303		MIT9303
1501	7 WH8103		WH8103
1501	8 WH7805		WH7805
1501	9 SAR6	GAUCACCUCCUU	SAR6
1501	10 SAR7	GAUCACCUCCUUU	SAR7
1501	11 SAR139		SAR139
1501	12 PCC6301	GAUCACCUCCUUU	PCC6301
1501	13 Anab7120	GAUCACCUCCUUU	Anab7120
1501	14 Mpolyclp	GAUCACCUCCUUU	Mpolyclp
1501	15 Pchlron		Pchlron
1501	16 MASK		MASK

Figure 2. *psbB* sequence data used to infer trees in Chapter 2, Figure 2. Abbreviations: Med4, MIT9107, FP5, SS120: *P. marinus* cultured strains (this work); WH8103, PCC7942: *Synechococcus* cultured strains; Scystis: *Synechocystis* PCC6803, Pthrix: *Prochlorothrix hollandica*; Anab7120: *Anabaena* PCC7120; Maize: *Zea maize* chloroplast. (references in Chapter Two).

Position:	Sequence identity:	Data:	
1	1 Med4		Med4
1	2 Med4		Med4
1	3 MIT9107		MIT9107
1	4 MIT9107		MIT9107
1	5 FP5		FP5
1	6 FP5		FP5
1	7 SS120		SS120
1	8 SS120		SS120
1	9 WH8103		WH8103
1	10 WH8103		WH8103
1	11 Scystis	ATGGGACTACCTTGGTATCGCGTTCATACAGTTGTCCTGAATGATCCAGGGCGACTCATCTCTGTCCATTTAATGCACACTGCCCTAGTGGCTGGCTGGG	Scystis
1	12 Scystis	M--G--L--P--W--Y--R--V--H--T--V--V--L--N--D--P--G--R--L--I--S--V--H--L--M--H--T--A--L--V--A--G--W--	Scystis
1	13 Pthrix	ATGGGACTACCCTGGTATCGCGTACACACCGCCGTCATCAATGACCCCGGTCGGCTGCTAGCTGTACACCTAATGCACACAGCCCTGGTCTTGGGCTGGG	Pthrix
1	14 Pthrix	M--G--L--P--W--Y--R--V--H--T--A--V--I--N--D--P--G--R--L--L--A--V--H--L--M--H--T--A--L--V--L--G--W--	Pthrix
1	15 PCC7942	ATGGGACTACCCTGGTACCGTGTCCACACGGTCGTCCTCAATGATCCGGGACGACTGATTGCAGTGCACTTGATGCACACTGCCCTCGTGGCAGGTTGGG	PCC7942
1	16 PCC7942	M--G--L--P--W--Y--R--V--H--T--V--V--L--N--D--P--G--R--L--I--A--V--H--L--M--H--T--A--L--V--A--G--W--	PCC7942
1	17 Anab7120	ATGGGACTACCCTGGTACCGAGTACATACAGTAGTTCTGAATGACCCAGGGCGACTGATTTCTGTACACTTGATGCACACAGCCCTGGTGGCAGGTTGGG	Anab7120
1	18 Anab7120	M--G--L--P--W--Y--R--V--H--T--V--V--L--N--D--P--G--R--L--I--S--V--H--L--M--H--T--A--L--V--A--G--W--	Anab7120
1	19 Maize	ATGGGTTTGCCTTGGTATCGTGTTCATACTGTCGTATTGAATGATCCGGGTCGATTGCTTTCGGTGCATATAATGCACACAGCTCTAGTTTCTGGTTGGG	Maize
1	20 Maize	M--G--L--P--W--Y--R--V--H--T--V--V--L--N--D--P--G--R--L--L--S--V--H--I--M--H--T--A--L--V--S--G--W--	Maize
101	1 Med4		Med4
101	2 Med4		Med4
101	3 MIT9107		MIT9107
101	4 MIT9107		MIT9107
101	5 FP5		FP5
101	6 FP5		FP5
101	7 SS120		SS120
101	8 SS120		SS120
101	9 WH8103		WH8103
101	10 WH8103		WH8103
101	11 Scystis	CTGGATCTATGGCTCTCTATGAGTTGGCCATTTTTGACTCCAGCGATGCTGTGCTTAACCCCATGTGGCGGCAAGGCATGTTTGTTTTGCCCTTCATGGC	Scystis
101	12 Scystis	A--G--S--M--A--L--Y--E--L--A--I--F--D--S--S--D--A--V--L--N--P--M--W--R--Q--G--M--F--V--L--P--F--M--A	Scystis
101	13 Pthrix	CCGGTTCAATGGCGCTCTATGAGCTAGCTGTTTTTGATCCCAGTGATGCAGTTCTCAATCCCATGTGGCGGCAAGGCATGTTTGTGCTGCCCTTCATGAC	Pthrix
101	14 Pthrix	A--G--S--M--A--L--Y--E--L--A--V--F--D--P--S--D--A--V--L--N--P--M--W--R--Q--G--M--F--V--L--P--F--M--T	Pthrix
101	15 PCC7942	CAGGTTCGATGGCTTTATACGAACTTGCCATTTTTGATCCTTCTGATGCCGTGCTGAACCCCATGTGGCGGCAAGGCATGTTCGTGTTGCCGTTCATGGC	PCC7942
101	16 PCC7942	A--G--S--M--A--L--Y--E--L--A--I--F--D--P--S--D--A--V--L--N--P--M--W--R--Q--G--M--F--V--L--P--F--M--A	PCC7942
101	17 Anab7120	CTGGTTCGATGGCACTATACGAACTAGCTATTTATGACCCCAGCGATCCGGTTCTCAACCCGATGTGGCGGCAAGGGATGTTCGTGCTACCCTTCATGGC	Anab7120
101	18 Anab7120	A--G--S--M--A--L--Y--E--L--A--I--Y--D--P--S--D--P--V--L--N--P--M--W--R--Q--G--M--F--V--L--P--F--M--A	Anab7120
101	19 Maize	CTGGTTCGATGGCTTTATACGAATTAGCGGTTTTTGATCCCTCTGATCCTGTTCTGGATCCAATGTGGAGACAAGGTATGTTCGTCATCCCCTTCATGAC	Maize
101	20 Maize	A--G--S--M--A--L--Y--E--L--A--V--F--D--P--S--D--P--V--L--D--P--M--W--R--Q--G--M--F--V--I--P--F--M--T	Maize

Posi-tion:	Sequence identity:	Data:	
201	1 Med4		Med4
201	2 Med4		Med4
201	3 MIT9107		MIT9107
201	4 MIT9107		MIT9107
201	5 FP5		FP5
201	6 FP5		FP5
201	7 SS120		SS120
201	8 SS120		SS120
201	9 WH8103		WH8103
201	10 WH8103		WH8103
201	11 Scystis	CCGCCTCGGTGTCACCAGTTCCTGGAATGGCTGGAGCGTCACCGGAGAAACT------------------------GGTTTGGATCCCGGTTTCTGGTCC	Scystis
201	12 Scystis	--R--L--G--V--T--S--S--W--N--G--W--S--V--T--G--E--T--------------------------G--L--D--P--G--F--W--S-	Scystis
201	13 Pthrix	CCGCTTGGGAGTCACCCACTCCTGGTCTGGTTGGACCGTGACTGGAGAGCCTTGGATTAATGAACCGGGCTTTTTAAATGCACACTTTAACTTCTGGAGC	Pthrix
201	14 Pthrix	--R--L--G--V--T--H--S--W--S--G--W--T--V--T--G--E--P--W--I--N--E--P--G--F--L--N--A--H--F--N--F--W--S-	Pthrix
201	15 PCC7942	GCGTTTGGGCGTCACCCAATCTTGGGGTGGCTGGAGCATCACCGGCGAAACC------------------------GCCGTGGATCCTGGCTATTGGAGC	PCC7942
201	16 PCC7942	--R--L--G--V--T--Q--S--W--G--G--W--S--I--T--G--E--T--------------------------A--V--D--P--G--Y--W--S-	PCC7942
201	17 Anab7120	ACGGTTAGGTGTTACCCAATCTTGGGGCGGTTGGAGCGTTACTGGCGGTACA------------------------GCAACTGACCCTGGTTTCTGGTCA	Anab7120
201	18 Anab7120	--R--L--G--V--T--Q--S--W--G--G--W--S--V--T--G--G--T--------------------------A--T--D--P--G--F--W--S-	Anab7120
201	19 Maize	TCGTTTAGGAATAACGAATTCGTGGGGTGGTTGGAGTATTTCAGGAGGAACT------------------------GTAACGAATCCGGGTATTTGGAGT	Maize
201	20 Maize	--R--L--G--I--T--N--S--W--G--G--W--S--I--S--G--G--T--------------------------V--T--N--P--G--I--W--S-	Maize
301	1 Med4		Med4
301	2 Med4		Med4
301	3 MIT9107		MIT9107
301	4 MIT9107		MIT9107
301	5 FP5		FP5
301	6 FP5		FP5
301	7 SS120		SS120
301	8 SS120		SS120
301	9 WH8103		WH8103
301	10 WH8103		WH8103
301	11 Scystis	TTTGAAGGGGTAGCTGCTGCCCACATCGTTCTATCTGGTCTGTTGTTCCTAGCCGCCGTATGGCACTGGGTATTTTGGGACCTGGAATTATTTGTTGACC	Scystis
301	12 Scystis	-F--E--G--V--A--A--A--H--I--V--L--S--G--L--L--F--L--A--A--V--W--H--W--V--F--W--D--L--E--L--F--V--D--	Scystis
301	13 Pthrix	TACGAAGGCGTAGCCCTGATGCACATTGTTCTGTCCGGTCTCTTCTTCCTGGCTGCCGTGTGGCACTGGGTTTACTGGGATTTGGATCTGTTTGAGGATC	Pthrix
301	14 Pthrix	-Y--E--G--V--A--L--M--H--I--V--L--S--G--L--F--F--L--A--A--V--W--H--W--V--Y--W--D--L--D--L--F--E--D--	Pthrix
301	15 PCC7942	TTTGAAGGCGTCGCGATCGCCCACATCGTACTGTCGGGTCTGCTGTTCCTCGCAGCAGTCTGGCACTGGGTCTACTGGGACCTCGAACTCTTTACCGATC	PCC7942
301	16 PCC7942	-F--E--G--V--A--I--A--H--I--V--L--S--G--L--L--F--L--A--A--V--W--H--W--V--Y--W--D--L--E--L--F--T--D--	PCC7942
301	17 Anab7120	TTTGAAGGGGTTGCCGCAGCTCACATTGTGCTTTCTGGTTTATTGTTCCTAGCTGCCGTTTGGCACTGGGTTTACTGGGATTTGGAACTCTTTAGAGATC	Anab7120
301	18 Anab7120	-F--E--G--V--A--A--A--H--I--V--L--S--G--L--L--F--L--A--A--V--W--H--W--V--Y--W--D--L--E--L--F--R--D--	Anab7120
301	19 Maize	TATGAAGGTGTGGCAGGTGCGCATATTGTGTTTTCTGGCTTGTGTTTCTTGGCAGCTATCTGGCATTGGGTATATTGGGACCTAGAAATATTCTGTGATG	Maize
301	20 Maize	-Y--E--G--V--A--G--A--H--I--V--F--S--G--L--C--F--L--A--A--I--W--H--W--V--Y--W--D--L--E--I--F--C--D--	Maize

```
Posi-   Sequence
tion:   identity:   Data:

 401     1 Med4                                                                                                              Med4
 401     2 Med4                                                                                                              Med4
 401     3 MIT9107                                                                                                           MIT9107
 401     4 MIT9107                                                                                                           MIT9107
 401     5 FP5                                                                                                               FP5
 401     6 FP5                                                                                                               FP5
 401     7 SS120                                                                                                             SS120
 401     8 SS120                                                                                                             SS120
 401     9 WH8103                                                                                                            WH8103
 401    10 WH8103                                                                                                            WH8103
 401    11 Scystis    CCCGTACCGGTGAATCCGCATTGGATTTGCCCAAAATGTTTGGTATTCACCTCTTCTTGTCTGGTTTGCTCTGCTTCGGTTTTGGTGCTTTCCACCTCAC  Scystis
 401    12 Scystis    P--R--T--G--E--S--A--L--D--L--P--K--M--F--G--I--H--L--F--L--S--G--L--L--C--F--G--F--G--A--F--H--L--T  Scystis
 401    13 Pthrix     CCCGTACCGGCGAGCCTGCCTTAGACCTACCCAAAATCTTTGGCGGTCACCTCTTCCTGCTAGGCTTCCTTTGCTTTAACTTCGGCACCTTCCACCTGAC  Pthrix
 401    14 Pthrix     P--R--T--G--E--P--A--L--D--L--P--K--I--F--G--G--H--L--F--L--L--G--F--L--C--F--N--F--G--T--F--H--L--T  Pthrix
 401    15 PCC7942    CCCGCACCGGCGAACCGGCCTTGGACCTGCCAAAAATGTTTGGCATCCACCTGTTCCTCTCCGGTCTTCTTTGCTTCGGCTTCGGTGCTTTCCACCTGTC  PCC7942
 401    16 PCC7942    P--R--T--G--E--P--A--L--D--L--P--K--M--F--G--I--H--L--F--L--S--G--L--L--C--F--G--F--G--A--F--H--L--S  PCC7942
 401    17 Anab7120   CTCGAACCGGTGAACCTGCTTTAGATTTGCCAAAAATGTTTGGCATTCACCTGTTCTTATCCGGTTTACTCTGTTTCGGCTTTGGTGCTTTCCATCTCAC  Anab7120
 401    18 Anab7120   P--R--T--G--E--P--A--L--D--L--P--K--M--F--G--I--H--L--F--L--S--G--L--L--C--F--G--F--G--A--F--H--L--T  Anab7120
 401    19 Maize      AGCGGACGGGAAAACCTTCTTTGGATTTGCCCAAGATCTTTGGAATTCATTTATTTCTTGCAGGGGTGGCTTGTTTTGGCTTTGGTGCATTTCATGTAAC  Maize
 401    20 Maize      E--R--T--G--K--P--S--L--D--L--P--K--I--F--G--I--H--L--F--L--A--G--V--A--C--F--G--F--G--A--F--H--V--T  Maize

 501     1 Med4                                                                                                              Med4
 501     2 Med4                                                                                                              Med4
 501     3 MIT9107                                                                                                           MIT9107
 501     4 MIT9107                                                                                                           MIT9107
 501     5 FP5                                                                                                               FP5
 501     6 FP5                                                                                                               FP5
 501     7 SS120                                                                                                             SS120
 501     8 SS120                                                                                                             SS120
 501     9 WH8103                                                                                                            WH8103
 501    10 WH8103                                                                                                            WH8103
 501    11 Scystis    CGGCGTTTGGGGCCCTGGCATGTGGGTATCCGACCCCTATGGTCTGACGGGCCACGTACAGCCAGTGGCACCGGAATGGGGACCAGCTGGCTTTAACCCG  Scystis
 501    12 Scystis    --G--V--W--G--P--G--M--W--V--S--D--P--Y--G--L--T--G--H--V--Q--P--V--A--P--E--W--G--P--A--G--F--N--P-  Scystis
 501    13 Pthrix     CGGTATCTTCGGACCCGGTATGTGGGTTTCCGATGCCTATGGCTTAACGGGCCACATCGAGCATATTGCTCCTGAATGGGGTCCGGCTGGGTTTAACCCC  Pthrix
 501    14 Pthrix     --G--I--F--G--P--G--M--W--V--S--D--A--Y--G--L--T--G--H--I--E--H--I--A--P--E--W--G--P--A--G--F--N--P-  Pthrix
 501    15 PCC7942    TGGCCTCTGGGGCCCGGGGATGTGGGTCTCCGATCCCTACGGCTTGACCGGCCATGTCCAACCAGTTGCCCCGGCCTGGGGTCCTGAAGGCTTCAACCCC  PCC7942
 501    16 PCC7942    --G--L--W--G--P--G--M--W--V--S--D--P--Y--G--L--T--G--H--V--Q--P--V--A--P--A--W--G--P--E--G--F--N--P-  PCC7942
 501    17 Anab7120   AGGCTTATTCGGCCCTGGAATGTGGATTTCTGACCCCTATGGAGTCACCGGCAGCGTCCAGCCAGTAGCACCCGAATGGGGTCCAGATGGATTTAACCCA  Anab7120
 501    18 Anab7120   --G--L--F--G--P--G--M--W--I--S--D--P--Y--G--V--T--G--S--V--Q--P--V--A--P--E--W--G--P--D--G--F--N--P-  Anab7120
 501    19 Maize      GGGTTTATATGGCCCTGGGATATGGGTGTCCGATCCTTACGGACTCACTGGAAAAGTACAAGCTGTAAATCCTGCGTGGGGTGCAGAAGGTTTTGATCCT  Maize
 501    20 Maize      --G--L--Y--G--P--G--I--W--V--S--D--P--Y--G--L--T--G--K--V--Q--A--V--N--P--A--W--G--A--E--G--F--D--P-  Maize
```

Position:	Sequence identity:	Data:	
601	1 Med4		Med4
601	2 Med4		Med4
601	3 MIT9107		MIT9107
601	4 MIT9107		MIT9107
601	5 FP5		FP5
601	6 FP5		FP5
601	7 SS120		SS120
601	8 SS120		SS120
601	9 WH8103		WH8103
601	10 WH8103		WH8103
601	11 Scystis	TTTAACCCCGGTGGGGTAGTGGCTCACCACATTGCTGCTGGTATTGTCGGTATCATTGCCGGTCTTTTCCACCTTACGGTACGGCCCCCTGAGCGCTTGT	Scystis
601	12 Scystis	-F--N--P--G--G--V--V--A--H--H--I--A--A--G--I--V--G--I--I--A--G--L--F--H--L--T--V--R--P--P--E--R--L--	Scystis
601	13 Pthrix	TTTAACCCCGGTGGTGTTGTGGCTCACCACATCGCAGCCGGTATCACCCTGATGATTGGGGGGTCTTCCACCTGACGGCTCGTCCCCCTGAGCGCCTTT	Pthrix
601	14 Pthrix	-F--N--P--G--G--V--V--A--H--H--I--A--A--G--I--T--L--M--I--G--G--V--F--H--L--T--A--R--P--P--E--R--L--	Pthrix
601	15 PCC7942	TTCAATCCGGGTGGCATTGTGGCTCACCACATTGCAGCCGGTGTCGTTGGCATCGTCGCAGGCCTCTTCCACCTGACGGTTCGTCCCCCCGAGCGCCTCT	PCC7942
601	16 PCC7942	-F--N--P--G--G--I--V--A--H--H--I--A--A--G--V--V--G--I--V--A--G--L--F--H--L--T--V--R--P--P--E--R--L--	PCC7942
601	17 Anab7120	TTTAACCCTGGTGGCGTAGTAGCTCACCATATTGCAGCTGGTATTGTCGGTATCATTGCAGGTTTATTCCACCTCACTGTTAGACCCCCCGAAAGGCTCT	Anab7120
601	18 Anab7120	-F--N--P--G--G--V--V--A--H--H--I--A--A--G--I--V--G--I--I--A--G--L--F--H--L--T--V--R--P--P--E--R--L--	Anab7120
601	19 Maize	TTCGTTCCGGGAGGAATAGCTTCGCATCATATTGCTGCGGGTACTTTGGGCATATTAGCGGGCCTATTCCATCTAAGTGTCCGTCCGCCTCAACGTCTAT	Maize
601	20 Maize	-F--V--P--G--G--I--A--S--H--H--I--A--A--G--T--L--G--I--L--A--G--L--F--H--L--S--V--R--P--P--Q--R--L--	Maize
701	1 Med4		Med4
701	2 Med4		Med4
701	3 MIT9107		MIT9107
701	4 MIT9107		MIT9107
701	5 FP5	TGG	FP5
701	6 FP5	G	FP5
701	7 SS120	TGG	SS120
701	8 SS120	G	SS120
701	9 WH8103	TACGG	WH8103
701	10 WH8103	-Y--G	WH8103
701	11 Scystis	ATAAAGCCCTCCGCATGGGGAATATTGAGACCGTATTGTCCAGTAGTATTGCCGCTGTATTTTTTGCCGCTTTCGTCGTGGCCGGTACCATGTGGTACGG	Scystis
701	12 Scystis	Y--K--A--L--R--M--G--N--I--E--T--V--L--S--S--S--I--A--A--V--F--F--A--A--F--V--V--A--G--T--M--W--Y--G	Scystis
701	13 Pthrix	ACACCGCCCTGCGTATGGGTAATGTGGAAACGGCTCTGGCTAGCGCGATCGCCGCTGTCTTTGGGGCTGCTTTCGTGGTGGCTGGAACCATGTGGTATGG	Pthrix
701	14 Pthrix	Y--T--A--L--R--M--G--N--V--E--T--A--L--A--S--A--I--A--A--V--F--G--A--A--F--V--V--A--G--T--M--W--Y--G	Pthrix
701	15 PCC7942	ACAAAGCGCTGCGGATGGGCAACATCGAAACCGTCTTGTCGAGCTCCTTGGCAGCAGTCTTCTTCGCTGCTTTTGTGGTCGCTGGCACGATGTGGTACGG	PCC7942
701	16 PCC7942	Y--K--A--L--R--M--G--N--I--E--T--V--L--S--S--S--L--A--A--V--F--F--A--A--F--V--V--A--G--T--M--W--Y--G	PCC7942
701	17 Anab7120	ACAAAGCTCTACGGATGGGTAACATTGAAACCGTACTTTCCAGCAGTATTGCGGCAGTATTCTTCGCAGCTTTCGTAGTTGCAGGAACCATGTGGTACGG	Anab7120
701	18 Anab7120	Y--K--A--L--R--M--G--N--I--E--T--V--L--S--S--S--I--A--A--V--F--F--A--A--F--V--V--A--G--T--M--W--Y--G	Anab7120
701	19 Maize	ACAAAGGGTTACGTATGGGCAATATTGAAACTGTACTTTCCAGTAGTATCGCTGCTGTTTTTTTGCTGCTTTCGTAGTTGCCGGAACTATGTGGTATGG	Maize
701	20 Maize	Y--K--G--L--R--M--G--N--I--E--T--V--L--S--S--S--I--A--A--V--F--F--A--A--F--V--V--A--G--T--M--W--Y--G	Maize

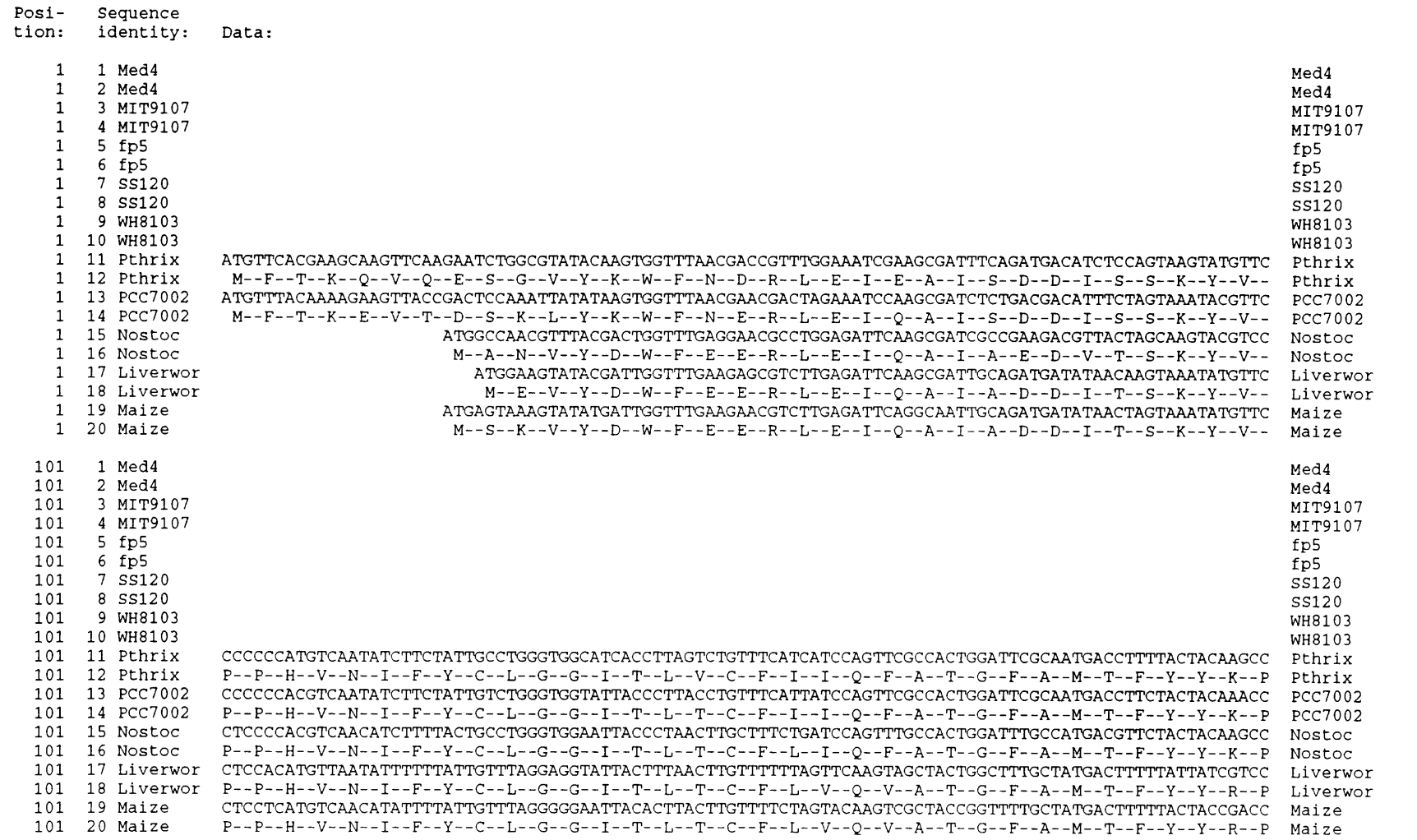
```
Posi-   Sequence
tion:   identity:   Data:

  801    1 Med4                                ATGTTTGGACCAACAAGATATCAATGGGACTCAGGTTACTTCAAAACTGAAATAAACAGGAGAGTTCAAACAGCTATT  Med4
  801    2 Med4                                 M--F--G--P--T--R--Y--Q--W--D--S--G--Y--F--K--T--E--I--N--R--R--V--Q--T--A--I-  Med4
  801    3 MIT9107                        CTCAGTTGAATTATTTGGTCCTACCAGATATCAATGGGATTCTGGATATTTCAAAACTGAAATTAATAGAAGAGTGCAAGTGGCTATC  MIT9107
  801    4 MIT9107                         S--V--E--L--F--G--P--T--R--Y--Q--W--D--S--G--Y--F--K--T--E--I--N--R--R--V--Q--V--A--I-  MIT9107
  801    5 FP5          ATCAGCAACAACTCCAGTGGAATTATTTGGTCCTACTAGATATCAATGGGATTCTGGTTATTTCAAAACTGAAATCAACCGCAGAGTGCAAGAATCTATG  FP5
  801    6 FP5          --S--A--T--T--P--V--E--L--F--G--P--T--R--Y--Q--W--D--S--G--Y--F--K--T--E--I--N--R--R--V--Q--E--S--M-  FP5
  801    7 SS120        ATCAGCAACAACTCCAATTGAGTTGTTTGGTCCAACCCGATATCAATGGGATTCCAATTACTTCCAAAATGAAATCAATAGAAGAGTTGAAGCTTCTATA  SS120
  801    8 SS120        --S--A--T--T--P--I--E--L--F--G--P--T--R--Y--Q--W--D--S--N--Y--F--Q--N--E--I--N--R--R--V--E--A--S--I-  SS120
  801    9 WH8103       AGCAGCTCGTACGCCTGTTGAGCTGTTTGGTCCGACCCGCTACCAGTgGGATCAGAGCTACTTCAAGACCGAGATCAACCGTCGTGTCCAGACCGCCATG  WH8103
  801   10 WH8103       --A--A--R--T--P--V--E--L--F--G--P--T--R--Y--Q--W--D--Q--S--Y--F--K--T--E--I--N--R--R--V--Q--T--A--M-  WH8103
  801   11 Scystis      TAATGCCACCACCCCCATCGAACTCTTTGGCCCCACCCGTTATCAGTGGGATAAGGGTTATTTTCAAGAAGAAATCCAACGCCGGGTTGATAGCCAGTTA  Scystis
  801   12 Scystis      --N--A--T--T--P--I--E--L--F--G--P--T--R--Y--Q--W--D--K--G--Y--F--Q--E--E--I--Q--R--R--V--D--S--Q--L-  Scystis
  801   13 Pthrix       CCACGTGACCACACCTATCGAACTCTTTGGACCTACCCGCTACCAGTGGGATCAGGGCTACTTCACCCAGGAAATTCAGCGTCGTGTGGATTCCCAACTG  Pthrix
  801   14 Pthrix       --H--V--T--T--P--I--E--L--F--G--P--T--R--Y--Q--W--D--Q--G--Y--F--T--Q--E--I--Q--R--R--V--D--S--Q--L-  Pthrix
  801   15 PCC7942      CAACGCTGCCACGCCAGTCGAACTGTTTGGCCCAACTCGCTACCAGTGGGACCAAGGCTACTTCCGTCAGGAAATTGCCCGCCGGGTTGATACGGCTGTC  PCC7942
  801   16 PCC7942      --N--A--A--T--P--V--E--L--F--G--P--T--R--Y--Q--W--D--Q--G--Y--F--R--Q--E--I--A--R--R--V--D--T--A--V-  PCC7942
  801   17 Anab7120     CAACGCTACTACACCCATCGAACTGTTTGGGCCCACACGTTATCAATGGGATCAAGGCTATTTCCATCAAGAAATTGAGCGCCGTGTGCAATCAAGCGTA  Anab7120
  801   18 Anab7120     --N--A--T--T--P--I--E--L--F--G--P--T--R--Y--Q--W--D--Q--G--Y--F--H--Q--E--I--E--R--R--V--Q--S--S--V-  Anab7120
  801   19 Maize        ATCAGCAACTACCCCAATCGAATTATTTGGGCCTACTCGTTATCAGTGGGATCAGGGATACTTTCAGCAAGAAATATATCGAAGAGTTAGCGATGGGTTA  Maize
  801   20 Maize        --S--A--T--T--P--I--E--L--F--G--P--T--R--Y--Q--W--D--Q--G--Y--F--Q--Q--E--I--Y--R--R--V--S--D--G--L-  Maize

  901    1 Med4         GATAATGGAGCAACTAAGGAAGAAGCATATGCTTCAATTCCAGAAAAGCTTGCATTCTACGATTATGTAGGGAACAGCCCTGCTAAAGGAGGATTGTTTA  Med4
  901    2 Med4         -D--N--G--A--T--K--E--E--A--Y--A--S--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  Med4
  901    3 MIT9107      GACAATGGTGCCACAAAAGAAGAGGCATATGATTCTATCCCAGAAAAATTAGCCTTCTACGATTATGTTGGGAATAGCCCTGCTAAAGGAGGATTATTCC  MIT9107
  901    4 MIT9107      -D--N--G--A--T--K--E--E--A--Y--D--S--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  MIT9107
  901    5 FP5          AATGATGGTGCATCGAAAGCTGAGGCTTATGCATCTATTCCTGAGCAACTTGCTTTCTACgNcTATGTAGGTAATAGTCCCGCCAAAGGAGGACTATTTA  FP5
  901    6 FP5          -N--D--G--A--S--K--A--E--A--Y--A--S--I--P--E--Q--L--A--F--Y--X--Y--V--G--N--S--P--A--K--G--G--L--F--  FP5
  901    7 SS120        GCAGATGGTGCAAGCAAGGAAGAAGCTTATTCAGCTATTCCAGaNNAGCTAGCCTTCTATGACTATGTAGGTAATAGCCCTGCCCCAGGAGGTCTGTTCA  SS120
  901    8 SS120        -A--D--G--A--S--K--E--E--A--Y--S--A--I--P--X--X--L--A--F--Y--D--Y--V--G--N--S--P--A--P--G--G--L--F--  SS120
  901    9 WH8103       GATGACGGTGCGACGCGTCAGGAGGCTTATGAGTCCATCCCTGAAAAGCTTGCTTTCTACGATTACGTCGGCAACAGCCcTGCCAAgGGTGGTTTGTTCC  WH8103
  901   10 WH8103       -D--D--G--A--T--R--Q--E--A--Y--E--S--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  WH8103
  901   11 Scystis      GCAGAAGGAGCTTCCCTATCTGAAGCCTGGTCCACCATCCCTGAAAAGTTGGCTTTCTACGATTACGTCGGTAACAGCCCCGCTAAAGGTGGTTTGTTTC  Scystis
  901   12 Scystis      -A--E--G--A--S--L--S--E--A--W--S--T--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  Scystis
  901   13 Pthrix       GCTGAAGGGGCTTCCCTGTCTGAAGCTTGGTCTTCCATCCCTGAGAAGCTAGCCTTCTACGACTATGTTGGTAACAGCCCCGCTAAAGGCGGTCTGTTCC  Pthrix
  901   14 Pthrix       -A--E--G--A--S--L--S--E--A--W--S--S--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  Pthrix
  901   15 PCC7942      GCCAGTGGCGCTTCTCTAGAGGAAGCTTGGAGCTCCATTCCTGAAAAACTGGCCTTCTATGACTACGTCGGCAACAGCCCTGCTAAAGGTGGCTTGTTCC  PCC7942
  901   16 PCC7942      -A--S--G--A--S--L--E--E--A--W--S--S--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  PCC7942
  901   17 Anab7120     GCTCAAGGTGCAAGCCTTTCCGAAGCTTGGTCACAGATTCCTGAAAAACTAGCTTTCTATGATTACGTCGGTAACAGCCCCGCCAAAGGTGGTTTGTTCC  Anab7120
  901   18 Anab7120     -A--Q--G--A--S--L--S--E--A--W--S--Q--I--P--E--K--L--A--F--Y--D--Y--V--G--N--S--P--A--K--G--G--L--F--  Anab7120
  901   19 Maize        GCCGAAAATCTTAGTTTATCAGAAGCTTGGTCTAAAATTCCCGAAAAATTAGCCTTTTATGATTATATTGGTAATAATCCGGCAAAGGGGGGATTATTCA  Maize
  901   20 Maize        -A--E--N--L--S--L--S--E--A--W--S--K--I--P--E--K--L--A--F--Y--D--Y--I--G--N--N--P--A--K--G--G--L--F--  Maize
```

Position:	Sequence identity:	Data:	
1001	1 Med4	GGGTTGGTGCTCTTGTAAATGGAGATGGTTTACCAACAGGCTGGCAAGGTCATATTGCCTTTACTGATAAAGAGGGCAATGATTTAGAAGTTAGAAGAAT	Med4
1001	2 Med4	R--V--G--A--L--V--N--G--D--G--L--P--T--G--W--Q--G--H--I--A--F--T--D--K--E--G--N--D--L--E--V--R--R--I	Med4
1001	3 MIT9107	GAGTTGGTGCTCTCGTTAATGGTGATGGTTTACCAACTGGTTGGCAAGGTCACATCGCTTTCCAAGACAAGGAAGGAAATGATTTGGAAGTCAGAAGAAT	MIT9107
1001	4 MIT9107	R--V--G--A--L--V--N--G--D--G--L--P--T--G--W--Q--G--H--I--A--F--Q--D--K--E--G--N--D--L--E--V--R--R--I	MIT9107
1001	5 FP5	GAGTTGGTGCAATGGTTAATGGAGATGGAGTTCCGAGTGGATGGCAAGGTCACATCTCATTTACTGATAGCGATGGCAATGACTTAGAAGTAAGAAGAAT	FP5
1001	6 FP5	R--V--G--A--M--V--N--G--D--G--V--P--S--G--W--Q--G--H--I--S--F--T--D--S--D--G--N--D--L--E--V--R--R--I	FP5
1001	7 SS120	GAGCAGGTGCCTTAGTTAATGGCGACGGTGTCCCCACTGGCTGGCAAGGGCATGTTTCCTTCAAAGATAAAGAAGGAAATGACCTTGAAGTAAGAAGAAT	SS120
1001	8 SS120	R--A--G--A--L--V--N--G--D--G--V--P--T--G--W--Q--G--H--V--S--F--K--D--K--E--G--N--D--L--E--V--R--R--M	SS120
1001	9 WH8103	GTGTTGGCCCGATGGTGAATGGTGATGGTCTGGCCACCTCCTGGGTTGGTCACATCGTCTTCACCGACAGAGAGGGTCGTGAACTTGAGGTTCGCCGCCT	WH8103
1001	10 WH8103	R--V--G--P--M--V--N--G--D--G--L--A--T--S--W--V--G--H--I--V--F--T--D--R--E--G--R--E--L--E--V--R--R--L	WH8103
1001	11 Scystis	GTACCGGTGCTATGAACAGTGGTGATGGCATTGCCCAGGAATGGATTGGTCACCCCATATTCAAAGACAAAGAAGGTCGGGAACTGGAGGTACGGCGTAT	Scystis
1001	12 Scystis	R--T--G--A--M--N--S--G--D--G--I--A--Q--E--W--I--G--H--P--I--F--K--D--K--E--G--R--E--L--E--V--R--R--M	Scystis
1001	13 Pthrix	GCGTGGGTGCCATGGATAGTGGTGATGGTATTGCCGAAGAGTGGCTGGGTCACCCCGTCTTCCAAGATGGTGCAGGCCGTGCCCTGTCGGTGCGTCGTTT	Pthrix
1001	14 Pthrix	R--V--G--A--M--D--S--G--D--G--I--A--E--E--W--L--G--H--P--V--F--Q--D--G--A--G--R--A--L--S--V--R--R--L	Pthrix
1001	15 PCC7942	GTACCGGTCAGATGAACAAAGGTGACGGGATTGCCCAAGGCTGGCTCGGCCACGCTGTCTTCAAGGACAAAAATGGCGATGTGCTCGACGTCCGTCGCTT	PCC7942
1001	16 PCC7942	R--T--G--Q--M--N--K--G--D--G--I--A--Q--G--W--L--G--H--A--V--F--K--D--K--N--G--D--V--L--D--V--R--R--L	PCC7942
1001	17 Anab7120	GTACAGGGCCAATGGTTAAGGGTGATGGTATTGCCCAATCTTGGCAAGGTCACGGCGTATTCAAAGATGCTGAAGGCCGGGAATTGACAGTACGTCGTCT	Anab7120
1001	18 Anab7120	R--T--G--P--M--V--K--G--D--G--I--A--Q--S--W--Q--G--H--G--V--F--K--D--A--E--G--R--E--L--T--V--R--R--L	Anab7120
1001	19 Maize	GAGCAGGCTCAATGGACAATGGGGATGGCATAGCTGTTGGATGGTTAGGACATCCCGTCTTTAGAGATAAAGAAGGACGCGAGCTTTTTGTACGTCGTAT	Maize
1001	20 Maize	R--A--G--S--M--D--N--G--D--G--I--A--V--G--W--L--G--H--P--V--F--R--D--K--E--G--R--E--L--F--V--R--R--M	Maize
1101	1 Med4	TCCAAACTTCTTTGAAAACTTCCCTGTTATTTTGGAAGACAAGGAAGGGAATGTGAGAGCTGACATCCCATTTAGAAGGGCTGAAGCAAAGTATTCATTT	Med4
1101	2 Med4	--P--N--F--F--E--N--F--P--V--I--L--E--D--K--E--G--N--V--R--A--D--I--P--F--R--R--A--E--A--K--Y--S--F-	Med4
1101	3 MIT9107	CCCAAACTTCTTTGAGAACTTCCCAGTCATACTTGAAGATAAAGAGGGGAATGTTAGGGCTGATATTCCATTTAGAAGAGCTGAAGCTAAGTATTCATTT	MIT9107
1101	4 MIT9107	--P--N--F--F--E--N--F--P--V--I--L--E--D--K--E--G--N--V--R--A--D--I--P--F--R--R--A--E--A--K--Y--S--F-	MIT9107
1101	5 FP5	TCCAAATTTCTTTGAGAACTTCCCTGTCATTCTTGAAGACAAGGATGGGAATGTCAAAGCTGACATTCnTTTCCGTCGAGCCGAAGCTAAGTACTCCATT	FP5
1101	6 FP5	--P--N--F--F--E--N--F--P--V--I--L--E--D--K--D--G--N--V--K--A--D--I--X--F--R--R--A--E--A--K--Y--S--I-	FP5
1101	7 SS120	GCCTAACTTCTTCGAMAACTTCCCTGTAATCCTTGAAGACAAAGATGGCAATGTTCGTGCTGACATTCCTTTCCGCCGCGCAGAAGCCAAATACTCGTTT	SS120
1101	8 SS120	--P--N--F--F--X--N--F--P--V--I--L--E--D--K--D--G--N--V--R--A--D--I--P--F--R--R--A--E--A--K--Y--S--F-	SS120
1101	9 WH8103	GCCCAACTTCTTCGAGAACTTCCCTGTCGTCCTCCAGGACGAGCAAGGCATCGTGCGCGCCGACATTCCCTACCGTCGCGCAGAAGCCAAGTATTCCTTC	WH8103
1101	10 WH8103	--P--N--F--F--E--N--F--P--V--V--L--Q--D--E--Q--G--I--V--R--A--D--I--P--Y--R--R--A--E--A--K--Y--S--F-	WH8103
1101	11 Scystis	GCCTAACTTCTTTGAAACTTTCCCCGTCATCATGACCGATGCGGATGGTGTAGTCCGGGCGGATATTCCCTTCCGTCGTTCCGAGTCTAAATTCAGTGTG	Scystis
1101	12 Scystis	--P--N--F--F--E--T--F--P--V--I--M--T--D--A--D--G--V--V--R--A--D--I--P--F--R--R--S--E--S--K--F--S--V-	Scystis
1101	13 Pthrix	GCCCAACTTCTTTGAGAACTTCCCCGTGATCCTCACGGATGGCGATGGTGTAGTTCGTGCTGATATTCCCTTCCGCCGCTCTGAGTCCCAGTACAGCTTC	Pthrix
1101	14 Pthrix	--P--N--F--F--E--N--F--P--V--I--L--T--D--G--D--G--V--V--R--A--D--I--P--F--R--R--S--E--S--Q--Y--S--F-	Pthrix
1101	15 PCC7942	GCCGAACTTCTTCGAGAACTTCCCGATCGTCTTGACTGACAGCAAAGGTGCTGTGCGGGCAGACATTCCTTTCCGTCGTGCTGAAGCGAAATTCAGCTTC	PCC7942
1101	16 PCC7942	--P--N--F--F--E--N--F--P--I--V--L--T--D--S--K--G--A--V--R--A--D--I--P--F--R--R--A--E--A--K--F--S--F-	PCC7942
1101	17 Anab7120	CCCCAACTTCTTTGAAACCTTCCCAGTAATCTTGACAGATGCTGATGGTGTTGTCCGCGCTGACATCCCTTTCCGTCGAGCAGAATCCAAGTATAGCTTT	Anab7120
1101	18 Anab7120	--P--N--F--F--E--T--F--P--V--I--L--T--D--A--D--G--V--V--R--A--D--I--P--F--R--R--A--E--S--K--Y--S--F-	Anab7120
1101	19 Maize	GCCTACTTTTTTTGAAACATTTCCGGTCGTTTTGGTAGATGAAGAGGGAATTGTGAGAGCGGACGTTCCTTTTAGAAGAGCAGAATCCAAATATAGTGTT	Maize
1101	20 Maize	--P--T--F--F--E--T--F--P--V--V--L--V--D--E--E--G--I--V--R--A--D--V--P--F--R--R--A--E--S--K--Y--S--V-	Maize

```
Posi-   Sequence
tion:   identity:   Data:

1201     1 Med4       GAACAAACTGGCATCACAGCGACTATTTATGGAGGTGATCTTAATGGTCAAACATTCACAGATCCTGCTGTAGTTAAAAGATTAGCTAGGAAAGCACAAC  Med4
1201     2 Med4       -E--Q--T--G--I--T--A--T--I--Y--G--G--D--L--N--G--Q--T--F--T--D--P--A--V--V--K--R--L--A--R--K--A--Q--  Med4
1201     3 MIT9107    GAGCAAACTGGTATCACCGCTACTATCTATGGAGGAGATTTAAATGGGCAAACATTTACTGATCCTGCAGTAGTTAAAAGGTTAGCTAGAAAAGCTCAGC  MIT9107
1201     4 MIT9107    -E--Q--T--G--I--T--A--T--I--Y--G--G--D--L--N--G--Q--T--F--T--D--P--A--V--V--K--R--L--A--R--K--A--Q--  MIT9107
1201     5 FP5        GAACAAACTGGAGTCACAGCTACTGTTTATGGTGGTGAATTAAATGGACAAacatttaCTGATCCAGTAATAGTTAAACGTCTAGCTAGAAAATCTCAGC  FP5
1201     6 FP5        -E--Q--T--G--V--T--A--T--V--Y--G--G--E--L--N--G--Q--T--F--T--D--P--V--I--V--K--R--L--A--R--K--S--Q--  FP5
1201     7 SS120      GAGCAAACTGGTATTACTGCAACMGTTTATGGTGGTGAGTTAAGTGGACAGACTTTCTCTGACCCAGTTGTTGTCAAGCGTCTAGCACGTAAGGCtcAAC  SS120
1201     8 SS120      -E--Q--T--G--I--T--A--T--V--Y--G--G--E--L--S--G--Q--T--F--S--D--P--V--V--V--K--R--L--A--R--K--A--Q--  SS120
1201     9 WH8103     GAACAGCAGGGCGTCACTGCTGAGGTGTTCGGTGGAGCTCTGGACGGTCAGAGATTCACCGATCCTGCTGACGTGAAGCGTCTGGCTCGTAAGGCTCAGC  WH8103
1201    10 WH8103     -E--Q--Q--G--V--T--A--E--V--F--G--G--A--L--D--G--Q--R--F--T--D--P--A--D--V--K--R--L--A--R--K--A--Q--  WH8103
1201    11 Scystis    GAACAAACCGGTGTTACCGTCAGCTTCTACGGTGGTGCTTTAGACGGCCAGACCTTCAGCAATCCCAGTGATGTGAAGAAGTTTGCCCGGAAAGCTCAGT  Scystis
1201    12 Scystis    -E--Q--T--G--V--T--V--S--F--Y--G--G--A--L--D--G--Q--T--F--S--N--P--S--D--V--K--K--F--A--R--K--A--Q--  Scystis
1201    13 Pthrix     GAGCAAACTGGAGTAACTGTCAGCTTCTACGGCGGTGCTCTGGATGGTCAAACCTTCACCAATCCTTCGGACGTGAAGAAGTTTGCCCGCCGTGCCCAAC  Pthrix
1201    14 Pthrix     -E--Q--T--G--V--T--V--S--F--Y--G--G--A--L--D--G--Q--T--F--T--N--P--S--D--V--K--K--F--A--R--R--A--Q--  Pthrix
1201    15 PCC7942    GAGGAAACCGGAATTACGGCTAGCTTCTACGGCGGTTCTCTGAATGGCCAAACCATCACTGATCCGGCGCAGGTGAAGAAATACGCCCGTAAGGCTCAGT  PCC7942
1201    16 PCC7942    -E--E--T--G--I--T--A--S--F--Y--G--G--S--L--N--G--Q--T--I--T--D--P--A--Q--V--K--K--Y--A--R--K--A--Q--  PCC7942
1201    17 Anab7120   GAACAATCAGGCGTAACAGTTAGCTTCTACGGTGGCGATTTGGACGGTAAAACCTTTACTGATCCCGCCGATGTGAAGAAATATGCCCGTAAAGCTCAAG  Anab7120
1201    18 Anab7120   -E--Q--S--G--V--T--V--S--F--Y--G--G--D--L--D--G--K--T--F--T--D--P--A--D--V--K--K--Y--A--R--K--A--Q--  Anab7120
1201    19 Maize      GAACAAGTAGGCGTAACGGTGGAGTTCTATGGTGGCGAACTTAATGGAGTAAGTTATTCTGATCCTGCTACTGTAAAAAAATATGCGCGGCGTGCTCAAT  Maize
1201    20 Maize      -E--Q--V--G--V--T--V--E--F--Y--G--G--E--L--N--G--V--S--Y--S--D--P--A--T--V--K--K--Y--A--R--R--A--Q--  Maize

1301     1 Med4       TTGGCAAAGCATTCAAGTTTGATAGAGAAA                                                                    Med4
1301     2 Med4       L--G--K--A--F--K--F--D--R--E-                                                                     Med4
1301     3 MIT9107    TAGGAGAAGCATTTAAGTTTGACAGAGAAACTTATAAAT                                                           MIT9107
1301     4 MIT9107    L--G--E--A--F--K--F--D--R--E--T--Y--K-                                                            MIT9107
1301     5 FP5        TTGGAGAGGCCTTCAAGTTTGACAGAGATAGATACAAGTCAGATGG                                                    FP5
1301     6 FP5        L--G--E--A--F--K--F--D--R--D--R--Y--K--S--D-                                                      FP5
1301     7 SS120      TTGgTGAGTCCTTCAAGTTTGATCGTGAGCGTTACAAGTCCGATGGTGTATTTAGAA                                         SS120
1301     8 SS120      L--G--E--S--F--K--F--D--R--E--R--Y--K--S--D--G--V--F--R-                                          SS120
1301     9 WH8103     TTGGAGAGGGCTTTGACTTCGACCGCGAGACCTATGTTCC                                                          WH8103
1301    10 WH8103     L--G--E--G--F--D--F--D--R--E--T--Y--V-                                                            WH8103
1301    11 Scystis    TGGGTGAAGGCTTCGACTTCGATACGGAAACCTTCAACTCTGATGGTGTATTCCGCACCAGTCCCCGGGGTTGGTTTACCTTTGGCCACGCTGTCTTTGC  Scystis
1301    12 Scystis    L--G--E--G--F--D--F--D--T--E--T--F--N--S--D--G--V--F--R--T--S--P--R--G--W--F--T--F--G--H--A--V--F--A  Scystis
1301    13 Pthrix     TGGGTGAAGCCTTCGACTTCGACACCGAAACCCTCGGATCTGATGGTGTGTTCCGCACCAGCACCCGGGGTTGGTTCACCTTTGGCCATGCTTGCTTTGC  Pthrix
1301    14 Pthrix     L--G--E--A--F--D--F--D--T--E--T--L--G--S--D--G--V--F--R--T--S--T--R--G--W--F--T--F--G--H--A--C--F--A  Pthrix
1301    15 PCC7942    TGGGTGAAGCGTTCGAATTCGACACCGAAACCCTTAACTCGGACGGTGTGTTCCGGACTTCGCCGCGTGGCTGGTTCACCTTTGGTCACGCCAGCTTTGC  PCC7942
1301    16 PCC7942    L--G--E--A--F--E--F--D--T--E--T--L--N--S--D--G--V--F--R--T--S--P--R--G--W--F--T--F--G--H--A--S--F--A  PCC7942
1301    17 Anab7120   GTGGAGAAATATTTGAATTCGACCGCGAAACCTTAAACTCTGACGGTGTATTCCGTACATCCCCCAGAGGTTGGTTTACCTTTGGTCACGCCGTATTTGC  Anab7120
1301    18 Anab7120   G--G--E--I--F--E--F--D--R--E--T--L--N--S--D--G--V--F--R--T--S--P--R--G--W--F--T--F--G--H--A--V--F--A  Anab7120
1301    19 Maize      TAGGGGAAATTTTTGAATTAGATCGAGCTACTTTGAAATCAGATGGTGTTTTTCGCAGCAGTCCAAGGGGTTGGTTCACTTTTGGTCATGCTACCTTTGC  Maize
1301    20 Maize      L--G--E--I--F--E--L--D--R--A--T--L--K--S--D--G--V--F--R--S--S--P--R--G--W--F--T--F--G--H--A--T--F--A  Maize
```

Position:	Sequence identity:	Data:	
1401	1 Med4		Med4
1401	2 Med4		Med4
1401	3 MIT9107		MIT9107
1401	4 MIT9107		MIT9107
1401	5 FP5		FP5
1401	6 FP5		FP5
1401	7 SS120		SS120
1401	8 SS120		SS120
1401	9 WH8103		WH8103
1401	10 WH8103		WH8103
1401	11 Scystis	CTTGCTCTTCTTCTTTGGTCACATCTGGCATGGTTCCCGGACTCTGTTCCGTGACGTATTTGCTGGGGTTGATCCTGGTTTGGAAGAACAAGTGGAATTT	Scystis
1401	12 Scystis	--L--L--F--F--F--G--H--I--W--H--G--S--R--T--L--F--R--D--V--F--A--G--V--D--P--G--L--E--E--Q--V--E--F-	Scystis
1401	13 Pthrix	CCTGCTCTTCTTCTTCGGCCACATCTGGCACGGTGCTCGCACCCTGTTCCGTGATGTGTTCGCCGGGATCGACGCTGACCTGGGTGAGCAAATCGAATTC	Pthrix
1401	14 Pthrix	--L--L--F--F--F--G--H--I--W--H--G--A--R--T--L--F--R--D--V--F--A--G--I--D--A--D--L--G--E--Q--I--E--F-	Pthrix
1401	15 PCC7942	TCTGCTCTTCTTCTTTGGCCATATCTGGCACGGCTCTCGGACGCTGTTCCGCGATGTCTTTGCTGGGATTGAAGCCGACTTGGGCGAGCAGATTGAATTC	PCC7942
1401	16 PCC7942	--L--L--F--F--F--G--H--I--W--H--G--S--R--T--L--F--R--D--V--F--A--G--I--E--A--D--L--G--E--Q--I--E--F-	PCC7942
1401	17 Anab7120	TCTGTTATTCTTCTTCGGTCACCTCTGGCACGGCGCTCGGACAATATACCGAGACGTATTCGCCGGTGTGGAAGCGGATCTAGAAGAACAAGTAGAGTGG	Anab7120
1401	18 Anab7120	--L--L--F--F--F--G--H--L--W--H--G--A--R--T--I--Y--R--D--V--F--A--G--V--E--A--D--L--E--E--Q--V--E--W-	Anab7120
1401	19 Maize	TTTGCTCTTCTTTTTCGGACACATTTGGCATGGCGCTCGAACCTTGTTCCGAGATGTTTTTGCTGGTATTGATCCAGACTTGGATGCTCAAGTGGAATTT	Maize
1401	20 Maize	--L--L--F--F--F--G--H--I--W--H--G--A--R--T--L--F--R--D--V--F--A--G--I--D--P--D--L--D--A--Q--V--E--F-	Maize
1501	1 Med4		Med4
1501	2 Med4		Med4
1501	3 MIT9107		MIT9107
1501	4 MIT9107		MIT9107
1501	5 FP5		FP5
1501	6 FP5		FP5
1501	7 SS120		SS120
1501	8 SS120		SS120
1501	9 WH8103		WH8103
1501	10 WH8103		WH8103
1501	11 Scystis	GGTGTGTTTGCTAAGGTTGGTGACCTTTCCACCCGGAAAGAAGCCTAGGTGCTTCTTGCACAGCTTTTAACCACAGCTTAAGAGCGTGTTTGAAAAGCCT	Scystis
1501	12 Scystis	-G--V--F--A--K--V--G--D--L--S--T--R--K--E--A--*	Scystis
1501	13 Pthrix	GGTGCGTTCCAGAAGCTGGGTGACCTGACCACCCGCAAGAGCTAAGACCCGAATCGCTCCCTTGGGAGTCTCTAGGCTAGGGTTTTAGAGCGCATCATCG	Pthrix
1501	14 Pthrix	-G--A--F--Q--K--L--G--D--L--T--T--R--K--S--*	Pthrix
1501	15 PCC7942	GGGGCCTTCCAGAAATTGGGTGACCCGACCACTCGGAAAACAGCCGCTTAAGGTTGTGAATGGTTAGAGCGGTTTGGGTGATGCTCATTCCGCTCAGCTT	PCC7942
1501	16 PCC7942	-G--A--F--Q--K--L--G--D--P--T--T--R--K--T--A--A--*	PCC7942
1501	17 Anab7120	GGTCTATTCCAGAAAGTGGGTGACAAATCAACCCGCGTTCGGAAGGAAGCTTAATTTTAGGGACTGGGGACTGGGGACTGGGAAATATACCGAGTACCGA	Anab7120
1501	18 Anab7120	-G--L--F--Q--K--V--G--D--K--S--T--R--V--R--K--E--A--*	Anab7120
1501	19 Maize	GGAACATTCCAAAAAGTTGGAGATCCAACTACAAGGAGACAGGCAGCCTGATACCACATTGCTATGGTATCTTTCACCTCCATTTTTTGTTTGATTTGAC	Maize
1501	20 Maize	-G--T--F--Q--K--V--G--D--P--T--T--R--R--Q--A--A--*	Maize

Figure 3. *petB/D* sequence data used to infer trees in Chapter 2, Figure 3. Abbreviations: Med4, MIT9107, FP5, SS120: *P. marinus* cultured strains (this work); WH8103, PCC7002: *Synechococcus* cultured strain; Pthrix: *Prochlorothrix hollandica*; Nostoc: *Nostoc* PCC7906;Liverwor: *Marchantia polymorpha* chloroplast Maize: *Zea maize* chloroplast (references in Chapter Two).



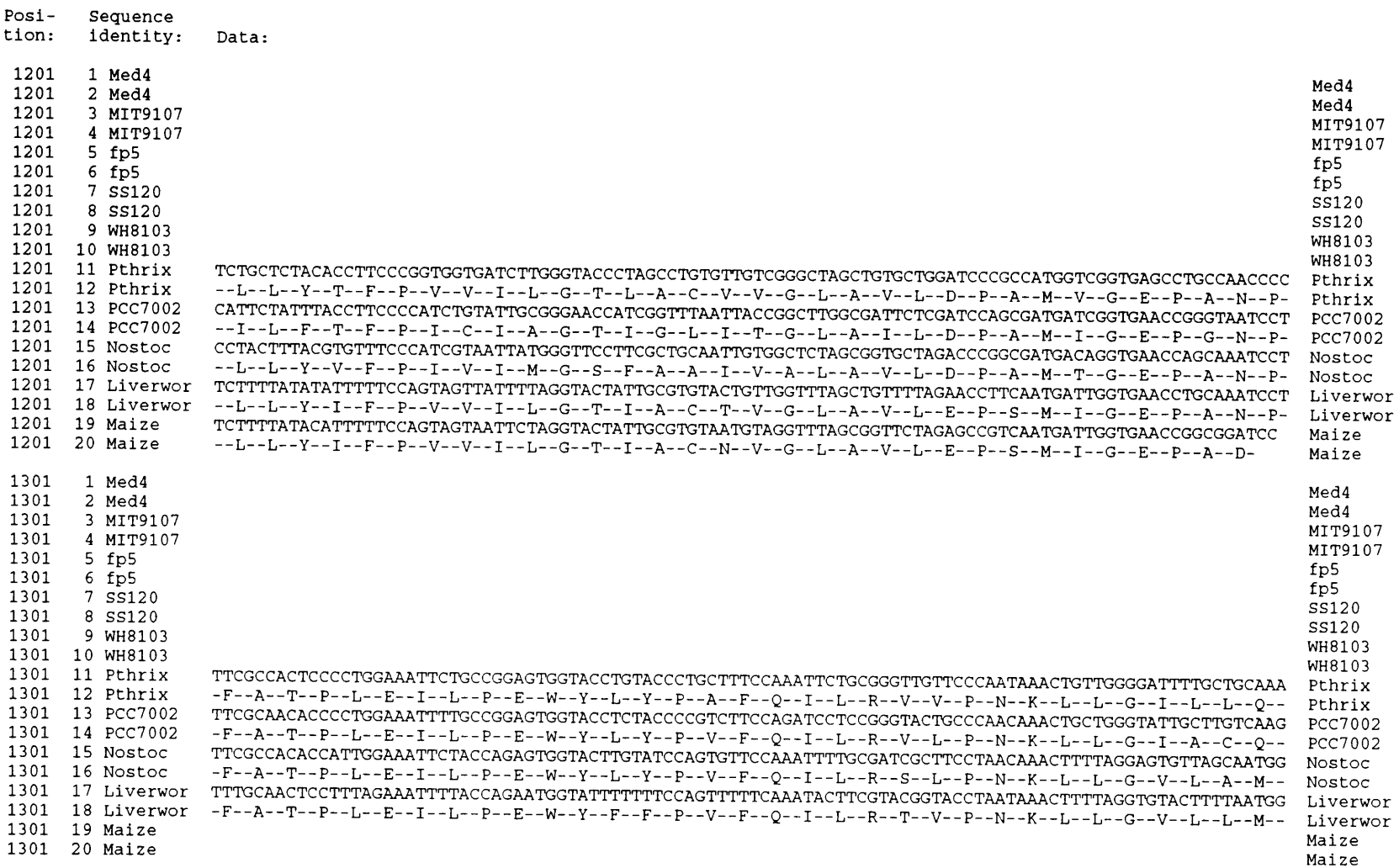
```
Posi-  Sequence
tion:  identity:   Data:

201    1 Med4                                                                                                                        Med4
201    2 Med4                                                                                                                        Med4
201    3 MIT9107                                                                                                                     MIT9107
201    4 MIT9107                                                                                                                     MIT9107
201    5 fp5                                                                                                                         fp5
201    6 fp5                                                                                                                         fp5
201    7 SS120                                                                                                                       SS120
201    8 SS120                                                                                                                       SS120
201    9 WH8103                                                                                                                      WH8103
201   10 WH8103                                                                                                                      WH8103
201   11 Pthrix     TTCGGTTACCGAGGCTTTCACCTCCGTGCAGTACCTCATGAATGAGGTCAGTTTTGGCTGGTTAATTCGCTCCATCCACCGCTGGTCTGCCAGCATGATG  Pthrix
201   12 Pthrix     --S--V--T--E--A--F--T--S--V--Q--Y--L--M--N--E--V--S--F--G--W--L--I--R--S--I--H--R--W--S--A--S--M--M-  Pthrix
201   13 PCC7002    CACCGTAGCAGAGGCATTTACTTCCGTTCAATACATCATGAATGAAGTGAACTTCGGGTGGTTGATCCGTTCGATCCACCGTTGGTCTGCCAGCATGATG  PCC7002
201   14 PCC7002    --T--V--A--E--A--F--T--S--V--Q--Y--I--M--N--E--V--N--F--G--W--L--I--R--S--I--H--R--W--S--A--S--M--M-  PCC7002
201   15 Nostoc     AACAGTCGCTGAAGCTTTCTCCTCAGTTGAGTACATCATGAATGAAGTGAACTTCGGTTGGCTAATTCGCTCCATCCATCGCTGGTCTGCCAGCATGATG  Nostoc
201   16 Nostoc     --T--V--A--E--A--F--S--S--V--E--Y--I--M--N--E--V--N--F--G--W--L--I--R--S--I--H--R--W--S--A--S--M--M-  Nostoc
201   17 Liverwor   TACTGTAACTGAAGCTTTTTCATCTGTTCAATACATTATGACTGAAGTAAATTTTGGATGGCTTATTCGCTCAGTTCATCGCTGGTCAGCAAGTATGATG  Liverwor
201   18 Liverwor   --T--V--T--E--A--F--S--S--V--Q--Y--I--M--T--E--V--N--F--G--W--L--I--R--S--V--H--R--W--S--A--S--M--M-  Liverwor
201   19 Maize      GACCGTTACAGAGGCTTTTTCCTCCGTTCAATATATAATGACGGAGGCCAACTTTGGTTGGTTAATCCGATCGGTTCATCGATGGTCAGCAAGTATGATG  Maize
201   20 Maize      --T--V--T--E--A--F--S--S--V--Q--Y--I--M--T--E--A--N--F--G--W--L--I--R--S--V--H--R--W--S--A--S--M--M-  Maize

301    1 Med4                     TCTTCATGTATTTAGGGTATATCTTACTGGAGGATTTAAAAGACCAAGAGAATTAACATGGGTAACGGGGGTTGTTATGGCAGTTA  Med4
301    2 Med4                       L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--  Med4
301    3 MIT9107                  CCTTCACGTTTTTAGAGTTTATCTTACGGGTGGTTTTAAAAGGCCCAGAGAACTAACGTGGGTTACAGGTGTAGTAATGGCAGTAA  MIT9107
301    4 MIT9107                    L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--  MIT9107
301    5 fp5                      TCTGCATGTTTTCCGCGTTTATTTGACTGGTGGATTTAAAAGACCTAGAGAATTGACATGGATAACTGGAGTGGTTATGGCAGTAA  fp5
301    6 fp5                        L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--I--T--G--V--V--M--A--V--  fp5
301    7 SS120                    TNNGCATGTTTTTAGAGTTTATTTAACGGGCGGATTTAAAAGACCTAGGGAGCTTACNTGGGTTACAGGAGTTGTAATGGCAGTTA  SS120
301    8 SS120                      X--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--  SS120
301    9 WH8103                   TCTTTACGTGTTCCGGGTCTACCTCACCGGTGGTTTCAAGCGTCCCCGTGAGCTCACCTGGGTCACCGGCGTGACCATGGCCGTGA  WH8103
301   10 WH8103                     L--Y--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--  WH8103
301   11 Pthrix     GTGTTGATGATGATCCTCCATGTGTTCCGGGTGTACCTCACCGGTGGTTTCAAAAATCCCCGCGAACTGACCTGGATTACGGGGGTTATTTTGGCGGTGA  Pthrix
301   12 Pthrix     -V--L--M--M--I--L--H--V--F--R--V--Y--L--T--G--G--F--K--N--P--R--E--L--T--W--I--T--G--V--I--L--A--V--  Pthrix
301   13 PCC7002    GTCTTGATGATGATTCTTCACATTTTCCGGGTGTACCTCACCGGTGGCTTTAAACGTCCCCGTGAGCTGACTTGGATTACTGGGGTCATCATGGCGACGA  PCC7002
301   14 PCC7002    -V--L--M--M--I--L--H--I--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--I--T--G--V--I--M--A--T--  PCC7002
301   15 Nostoc     GTATTGATGATGATTTTGCACGTCTTCCGGGTTTATCTGACTGGTGGTTTTAAAAAGCCCCGCGAATTGACCTGGGTGAGTGGTGTGATTTTGGCTGTAA  Nostoc
301   16 Nostoc     -V--L--M--M--I--L--H--V--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--S--G--V--I--L--A--V--  Nostoc
301   17 Liverwor   GTTTTAATGATGATTTTACATATTTTTCGTGTTTATCTAACAGGAGGTTTTAAAAAACCTCGGGAATTAACTTGGGTTACTGGTGTTATTTTAGCAGTTT  Liverwor
301   18 Liverwor   -V--L--M--M--I--L--H--I--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--I--L--A--V--  Liverwor
301   19 Maize      GTTCTAATGATGATCCTGCACGTATTTCGTGTGTATCTCACAGGTGGATTTAAAAAACCCCGTGAATTAACTTGGGTCACAGGCGTGGTTTTGGCTGTAT  Maize
301   20 Maize      -V--L--M--M--I--L--H--V--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--V--L--A--V--  Maize
```

Position:	Sequence identity:	Data:	
401	1 Med4	TAACAGTTGCTTTCGGAGTTACAGGTTATTCACTTCCATGGGACCAGGTTGGATACTGGGCAGTTAAAATAGTTTCAGGTGTTCCTGCAGCAATACCTAT	Med4
401	2 Med4	I--T--V--A--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--I	Med4
401	3 MIT9107	TAACTGTCGCTTTTGGAGTTACAGGATATTCTTTGCCTTGGGATCAAGTCGGATATTGGGCTGTTAAGATTGTCTCAGGTGTTCCTGCCGCAATTCCAGT	MIT9107
401	4 MIT9107	I--T--V--A--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V	MIT9107
401	5 fp5	TCACAGTTGCTTTTGGAGTTACAGGATATTCATTGCCTTGGGATCAAGTTGGTTATTGGGCAGTTAAAATTGTCTCTGGAGTCCCAGCAGCTATCCCAGT	fp5
401	6 fp5	I--T--V--A--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V	fp5
401	7 SS120	TAACAGTTGCTTtTGGCGTTACAGGCTATTCTTTGCCTTGGGATCAAGTCGGATATTGGGCTGTAAAGATTGTTTCAGGCGTTCCTGCTGCCATCCCAGT	SS120
401	8 SS120	I--T--V--A--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V	SS120
401	9 WH8103	TCACAGTTTCcTTCGGTGTCACCGGTTACTCCCTGCCCTGGGACCAGGTTGGTTATTGGGCCGTCAAGATTGTTTCCGGCGTCCCAGCAGCCATCCCAGT	WH8103
401	10 WH8103	I--T--V--S--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V	WH8103
401	11 Pthrix	TCACCGTATCCTTCGGCGTGACCGGCTACTCCTTGCCCTGGGATCAAGTGGGTTACTGGGCCGTGAAAATTGTGTCCGGTGTCCCTGAGGCCATTCCCCT	Pthrix
401	12 Pthrix	I--T--V--S--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--E--A--I--P--L	Pthrix
401	13 PCC7002	TCACCGTTTCCTTCGGTGTAACTGGTTACTCCTTGCCTTGGGACCAAGTTGGTTACTGGGCAGTCAAGATTGTGTCTGGTGTACCTGCGGCGATTCCTGT	PCC7002
401	14 PCC7002	I--T--V--S--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V	PCC7002
401	15 Nostoc	TTACCGTTTCCTTTGGAGTTACCGGCTATTCCTTACCTTGGGACCAAGTTGGCTACTGGGCTGTGAAAATCGTTAGCGGTGTACCAGAAGCAATTCCCGT	Nostoc
401	16 Nostoc	I--T--V--S--F--G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--E--A--I--P--V	Nostoc
401	17 Liverwor	TAACTGTATCTTTTGGTGTTACAGGTTATTCTTTACCTTGGGATCAAATTGGTTATTGGGCAGTTAAAATTGTAACTGGTGTACCAGAAGCAATTCCAAT	Liverwor
401	18 Liverwor	L--T--V--S--F--G--V--T--G--Y--S--L--P--W--D--Q--I--G--Y--W--A--V--K--I--V--T--G--V--P--E--A--I--P--I	Liverwor
401	19 Maize	TGACTGCATCATTTGGTGTAACTGGTTATTCTTTGCCTTGGGATCAAATTGGTTATTGGGCAGTCAAAATTGTTACAGGTGTACCCGAAGCTATTCCGGT	Maize
401	20 Maize	L--T--A--S--F--G--V--T--G--Y--S--L--P--W--D--Q--I--G--Y--W--A--V--K--I--V--T--G--V--P--E--A--I--P--V	Maize
501	1 Med4	CATTGGTGATTTTATGGTTGAACTTCTTCGAGGAGGAGAAAGTGTTGGTCAATCTACTTTGACTAGATTTTATAGCCTCCATACTTTTGTTCTACCTTGG	Med4
501	2 Med4	--I--G--D--F--M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W-	Med4
501	3 MIT9107	AATAGGAGATTTTATGGTTGAACTACTGAGGGGAGGAGAAAGTGTTGGACAGTCTACTCTTACAAGATTTTATAGTCTTCATACTTTTGTTTTGCCATGG	MIT9107
501	4 MIT9107	--I--G--D--F--M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W-	MIT9107
501	5 fp5	TATTGGAGACTTTATGGTCGAACTTCTTAGAGGTGGTGAGAGCGTTGGACAGTCAACTTTGACAAGGTTTTATAGTCTTCACACTTTTGTGATGCCATGG	fp5
501	6 fp5	--I--G--D--F--M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--M--P--W-	fp5
501	7 SS120	TGTAGGCGACTTCATGGTTGAATTGCTTCGAGGAGGTGAAAGTGTTGGCCAGACAaCACTTACCCGTTTTTACAGTCTTCATACTTTTGTATTGCCTTGG	SS120
501	8 SS120	--V--G--D--F--M--V--E--L--L--R--G--G--E--S--V--G--Q--T--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W-	SS120
501	9 WH8103	TGTGGGTGACTTCATGGTGGAGCTGCTCCGCGGTGGCGAAAGTGTCGGTCAGTCCACACTCACTCGCTTCTACAGCCTCCACACCTTTGTGATGCCATGG	WH8103
501	10 WH8103	--V--G--D--F--M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--M--P--W-	WH8103
501	11 Pthrix	GGTTGGTCCCCTGATGGTGGAACTGATTCGCGGTAGTGCCAGTGTGGGTCAAGCGACCCTGACCCGCTTCTATAGCCTGCACACCTTTGTGTTGCCCTGG	Pthrix
501	12 Pthrix	--V--G--P--L--M--V--E--L--I--R--G--S--A--S--V--G--Q--A--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W-	Pthrix
501	13 PCC7002	CGTGGGCGACCAAATGGTTGAGTTGCTCCGGGGTGGCGCAAGCGTTGGCCAAGCAACCCTAACCCGTTTCTACAGTCTTCATACTTTTGTTTTACCTTGG	PCC7002
501	14 PCC7002	--V--G--D--Q--M--V--E--L--L--R--G--G--A--S--V--G--Q--A--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W-	PCC7002
501	15 Nostoc	GGTTGGTGTTCTGATCTCCGACCTACTGCGTGGTGGTTCTAGTGTTGGCCAAGCAACACTGACTCGTTACTACAGCGCACACACCTTCGTTTTGCCTTGG	Nostoc
501	16 Nostoc	--V--G--V--L--I--S--D--L--L--R--G--G--S--S--V--G--Q--A--T--L--T--R--Y--Y--S--A--H--T--F--V--L--P--W-	Nostoc
501	17 Liverwor	AATTGGATCTCCTTTAGTTGAGTTATTACGCGGAAGTGTAAGTGTTGGTCAATCGACATTAACTCGATTTTATAGTTTACATACTTTTGTATTGCCTCTT	Liverwor
501	18 Liverwor	--I--G--S--P--L--V--E--L--L--R--G--S--V--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--L-	Liverwor
501	19 Maize	AATAGGGTCCCCTTTAGTGGAATTATTACGTGGAAGTGCTAGTGTGGGCCAATCCACTTTGACTCGTTTTTATAGTTTACATACCTTTGTACTACCTCTG	Maize
501	20 Maize	--I--G--S--P--L--V--E--L--L--R--G--S--A--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--L-	Maize

```
Posi-   Sequence
tion:   identity:   Data:

 601    1 Med4      TCTTTAGCAGTATTTATGCTGATGCATTTTCTAATGATTCGTAAACAAGGTATTTCAGGTCCGTTATAAACCCTTATAATTTAAAAAACTAAACATTCTG  Med4
 601    2 Med4      -S--L--A--V--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  Med4
 601    3 MIT9107   TCATTAGCAGTATTTATGcTGATGCATTTTTTAATGATTCGTAAACAGGGTATTTCAGGACCTTTATAAACACTTACAATTAAATTACTAACGAGAGCTC  MIT9107
 601    4 MIT9107   -S--L--A--V--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  MIT9107
 601    5 fp5       CTGCTAGCAGTTTTTATGcTGATGCatTTtCTAATGATTAGAAAACAAGGAATCTCCGGTCCCTTATAATTAAAAAAAGATTTATCTAAACAACCCTTAC  fp5
 601    6 fp5       -L--L--A--V--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  fp5
 601    7 SS120     ACATTGGCAATCTTTATGCTCATGCACTTTCTTATGATTAGAAAACAGGGTATCTCAGGTCCTTTGTGATTCAAATGCGTAAAAAATAATTTATGCAGGA  SS120
 601    8 SS120     -T--L--A--I--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  SS120
 601    9 WH8103    CTGCTCGCCGTATTCATGCTCATGCACTTCCTGATGATTCGGAAGCAGGGCATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGG  WH8103
 601   10 WH8103    -L--L--A--V--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  WH8103
 601   11 Pthrix    TTCATTGCGGTGTTCATGCTGATGCACTTCCTGATGATTCGCAAGCAAGGGATCTCCGGTCCTTTATAAGCGAATGACCGCCAGTGGATTGTGGGCTGTG  Pthrix
 601   12 Pthrix    -F--I--A--V--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  Pthrix
 601   13 PCC7002   TTGATTGCGGTCTTCATGTTGGCCCACTTCCTCATGATCCGTAAGCAAGGTATTTCTGGTCCTTTGTAAGGATCAGGTTTTCCGTTTGCTCAAAACAACG  PCC7002
 601   14 PCC7002   -L--I--A--V--F--M--L--A--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  PCC7002
 601   15 Nostoc    TTAATAGCAGTCTTCATGCTGTTTCACTTCTTGATGATCCGCAAACAAGGTATTTCTGGGCCTTTGTAATCTCAAAGATTAGCAAAAGGCAACATTTGTC  Nostoc
 601   16 Nostoc    -L--I--A--V--F--M--L--F--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  Nostoc
 601   17 Liverwor  TTAACTGCAATATTTATGTTAATGCACTTTTTAATGATTCGTAAACAAGGTATTTCAGGTCCGTTATAAATTACGTAAATTTATTACAAAATAAAAAAGT  Liverwor
 601   18 Liverwor  -L--T--A--I--F--M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  Liverwor
 601   19 Maize     CTTACTGCCGTATTTATGTTAATGCACTTTCCAATGATACGTAAGCAAGGTATTTCGGGTCCGTTATAGGGAAGGCATAGCATAGAGAATTCTAATTCTC  Maize
 601   20 Maize     -L--T--A--V--F--M--L--M--H--F--P--M--I--R--K--Q--G--I--S--G--P--L--*-------------------------------  Maize

 701    1 Med4      AATTATTTTAT-------------------------------------------------------------------------------------  Med4
 701    2 Med4      -------------------------------------------------------------------------------------------------  Med4
 701    3 MIT9107   CAC---------------------------------------------------------------------------------------------  MIT9107
 701    4 MIT9107   -------------------------------------------------------------------------------------------------  MIT9107
 701    5 fp5       CTTAATTCTTTTAAACCTTTCTTAAACC---------------------------------------------------------------------  fp5
 701    6 fp5       -------------------------------------------------------------------------------------------------  fp5
 701    7 SS120     CTGTGGCATTCTTTTTATCCAAATCTTTTACCTAATTTAATTTTTAAAT-------------------------------------------------  SS120
 701    8 SS120     -------------------------------------------------------------------------------------------------  SS120
 701    9 WH8103    ATTCACCG-----------------------------------------------------------------------------------------  WH8103
 701   10 WH8103    -------------------------------------------------------------------------------------------------  WH8103
 701   11 Pthrix    GATTGTGGGCAGTGAATTTCGGGCTGTGAATTTCGGGCTGTGAATGACTGTCAGTGAATTACCGCCAGTGAATAACCACCAGTGAATGACCGCCTGTAGT  Pthrix
 701   12 Pthrix    -------------------------------------------------------------------------------------------------  Pthrix
 701   13 PCC7002   TTCTTGTTGAGGAGAACTTTTACTC---------------------------------------------------------------------------  PCC7002
 701   14 PCC7002   -------------------------------------------------------------------------------------------------  PCC7002
 701   15 Nostoc    AGTTTCCAGATGAAGTTGTTTGTTTTAAACTTATGAAACACAAGAACCTAATTGTAGATAAGTTACAAAAAACAAACACTTGCATCCCAAGCCGAAAGCT  Nostoc
 701   16 Nostoc    -------------------------------------------------------------------------------------------------  Nostoc
 701   17 Liverwor  TTAAATACTAATTTCATTGCCATATTTTATGGATTTTTTTTTTCTATTTGAAAGTTCTTTTTAGAGAAATGCTAAAAAAATTTTTTTTAATAGATATTTT  Liverwor
 701   18 Liverwor  -------------------------------------------------------------------------------------------------  Liverwor
 701   19 Maize     ATATATCATATCGGGTAGGTTGTGGTATTTCATTGCTACAAACATGGGTTATTGTAAAATAAGACATGTCATTTGGATACTTCTCTTCAACTTCGAAGTA  Maize
 701   20 Maize     -------------------------------------------------------------------------------------------------  Maize
```

```
Posi-   Sequence
tion:   identity:   Data:

  801    1 Med4      ----------------------------------------------------------------------------------------------------  Med4
  801    2 Med4      ----------------------------------------------------------------------------------------------------  Med4
  801    3 MIT9107   ----------------------------------------------------------------------------------------------------  MIT9107
  801    4 MIT9107   ----------------------------------------------------------------------------------------------------  MIT9107
  801    5 fp5       ----------------------------------------------------------------------------------------------------  fp5
  801    6 fp5       ----------------------------------------------------------------------------------------------------  fp5
  801    7 SS120     ----------------------------------------------------------------------------------------------------  SS120
  801    8 SS120     ----------------------------------------------------------------------------------------------------  SS120
  801    9 WH8103    ----------------------------------------------------------------------------------------------------  WH8103
  801   10 WH8103    ----------------------------------------------------------------------------------------------------  WH8103
  801   11 Pthrix    TTGCTGCAATCCCTAAGCGGGTGTGTTTGTCACTCCCTATCGTTCCGACTTGGGGGTGAGAGACCCGTTCTTATTCTTGTTCAGGCCAACCTGATTAAGG  Pthrix
  801   12 Pthrix    ----------------------------------------------------------------------------------------------------  Pthrix
  801   13 PCC7002   ----------------------------------------------------------------------------------------------------  PCC7002
  801   14 PCC7002   ----------------------------------------------------------------------------------------------------  PCC7002
  801   15 Nostoc    TTAGACGCCCTCTCAGCGGCTGACGTAGACAGAGAAAACGCTGCTTGCCGCTTGTTAGGGCTGCCTACTAGCCGATAGCTTTCACAAATGCTTTAATTCA  Nostoc
  801   16 Nostoc    ----------------------------------------------------------------------------------------------------  Nostoc
  801   17 Liverwor  ATAAAGGAAAATGGATT----------------------------------------------------------------------------------  Liverwor
  801   18 Liverwor  ----------------------------------------------------------------------------------------------------  Liverwor
  801   19 Maize     TTTTTATACAAATAGTTGAAGTGAATTTTACGAAAGAAAATAAGGCGGATT----------------------------------------------  Maize
  801   20 Maize     ----------------------------------------------------------------------------------------------------  Maize

  901    1 Med4      ----------------------------------------------------------------------------------------------------  Med4
  901    2 Med4      ----------------------------------------------------------------------------------------------------  Med4
  901    3 MIT9107   ----------------------------------------------------------------------------------------------------  MIT9107
  901    4 MIT9107   ----------------------------------------------------------------------------------------------------  MIT9107
  901    5 fp5       ----------------------------------------------------------------------------------------------------  fp5
  901    6 fp5       ----------------------------------------------------------------------------------------------------  fp5
  901    7 SS120     ----------------------------------------------------------------------------------------------------  SS120
  901    8 SS120     ----------------------------------------------------------------------------------------------------  SS120
  901    9 WH8103    ----------------------------------------------------------------------------------------------------  WH8103
  901   10 WH8103    ----------------------------------------------------------------------------------------------------  WH8103
  901   11 Pthrix    CCAACTGATTAAGGTTAGCTGAGACCCCTGCAAGGTTGATGTTTCCTTGGTAGCTTAGACCTACCGCCTTGGTTATCATCGTTAATGGGTGCCATACCCT  Pthrix
  901   12 Pthrix    ----------------------------------------------------------------------------------------------------  Pthrix
  901   13 PCC7002   ----------------------------------------------------------------------------------------------------  PCC7002
  901   14 PCC7002   ----------------------------------------------------------------------------------------------------  PCC7002
  901   15 Nostoc    ACTTAAGCAGGAGAGCACATTTAAAA------------------------------------------------------------------------  Nostoc
  901   16 Nostoc    ----------------------------------------------------------------------------------------------------  Nostoc
  901   17 Liverwor  ----------------------------------------------------------------------------------------------------  Liverwor
  901   18 Liverwor  ----------------------------------------------------------------------------------------------------  Liverwor
  901   19 Maize     ----------------------------------------------------------------------------------------------------  Maize
  901   20 Maize     ----------------------------------------------------------------------------------------------------  Maize
```

```
Posi-   Sequence
tion:   identity:   Data:

1001    1 Med4      ----------------------------------------------------------------------------------------ATGT  Med4
1001    2 Med4      ----------------------------------------------------------------------------------------M--   Med4
1001    3 MIT9107   ----------------------------------------------------------------------------------------ATGT  MIT9107
1001    4 MIT9107   ----------------------------------------------------------------------------------------M--   MIT9107
1001    5 fp5       ----------------------------------------------------------------------------------------ATGT  fp5
1001    6 fp5       ----------------------------------------------------------------------------------------M--   fp5
1001    7 SS120     ----------------------------------------------------------------------------------------ATGT  SS120
1001    8 SS120     ----------------------------------------------------------------------------------------M--   SS120
1001    9 WH8103    ----------------------------------------------------------------------------------------ATGC  WH8103
1001   10 WH8103    ----------------------------------------------------------------------------------------M--   WH8103
1001   11 Pthrix    GATTGGTGCCATAACCTGTGCAATGCCCCTGGCTAGGGATTATCCCGTTGTAACCTTCCCCAGGTGACTGCTCTCTTACGGAGATCATTGAAAACTATGT  Pthrix
1001   12 Pthrix    ----------------------------------------------------------------------------------------M--   Pthrix
1001   13 PCC7002   ----------------------------------------------------------------------------------------ATGT  PCC7002
1001   14 PCC7002   ----------------------------------------------------------------------------------------M--   PCC7002
1001   15 Nostoc    ----------------------------------------------------------------------------------------ATGG  Nostoc
1001   16 Nostoc    ----------------------------------------------------------------------------------------M--   Nostoc
1001   17 Liverwor  ----------------------------------------------------------------------------------------ATGG  Liverwor
1001   18 Liverwor  ----------------------------------------------------------------------------------------M--   Liverwor
1001   19 Maize     ----------------------------------------------------------------------------------------ATGG  Maize
1001   20 Maize     ----------------------------------------------------------------------------------------M--   Maize

1101    1 Med4      CTACTTTAAAAAAACCTGATTTATCTGATCCTAAATTAAGAGCTAAGTTGGCTAAA                                      Med4
1101    2 Med4      S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K-                                      Med4
1101    3 MIT9107   CTACGTTAAAAAAACCAGATCTATCTGATCCAAAATTAAGAGCAAAATTAGCTatt                                      MIT9107
1101    4 MIT9107   S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--I-                                      MIT9107
1101    5 fp5       CTACTCTTAAGAAACCTGATTTAACCGATACAAAATTAAGAGCAAAACTTGCTaaa                                      fp5
1101    6 fp5       S--T--L--K--K--P--D--L--T--D--T--K--L--R--A--K--L--A--K-                                      fp5
1101    7 SS120     CCACTCTAAAGAAACCAAATTTATCTGATCCAAAGCTAAGGGCTAAGCTTTCTAAT                                      SS120
1101    8 SS120     S--T--L--K--K--P--N--L--S--D--P--K--L--R--A--K--L--S--N-                                      SS120
1101    9 WH8103    ACATTCTCAAGAAGCCTGACCTCTCCGATCCCAAGAYGGYGAAAGCCAGGTAAG                                        WH8103
1101   10 WH8103    H--I--L--K--K--P--D--L--S--D--P--K--X--X--K--A--R--*                                          WH8103
1101   11 Pthrix    CTGTTCTCAAAAAGCCGGATTTAACCGATCCCGTCCTATTGGAAAAGCTGGCCCAAAATATGGGTCATAACTACTATGGGGAACCCGCTTGGCCCAACGA  Pthrix
1101   12 Pthrix    S--V--L--K--K--P--D--L--T--D--P--V--L--L--E--K--L--A--Q--N--M--G--H--N--Y--Y--G--E--P--A--W--P--N--D  Pthrix
1101   13 PCC7002   CTATCATGAAAAAGCCGGATCTTAGCGATCCAAAACTCCGGGCAAAACTGGCTCAAAACATGGGACACAATTACTATGGTGAGCCGGCTTGGCCCAATGA  PCC7002
1101   14 PCC7002   S--I--M--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--Q--N--M--G--H--N--Y--Y--G--E--P--A--W--P--N--D  PCC7002
1101   15 Nostoc    CAACACAGAAAAAAACCTGACCTCAGCGACCCCCAGTTAAGAGCCAAACTGGCTAAAGGTATGGGTCACAACTACTATGGTGAACCAGCTTGGCCTAATGA  Nostoc
1101   16 Nostoc    A--T--Q--K--K--P--D--L--S--D--P--Q--L--R--A--K--L--A--K--G--M--G--H--N--Y--Y--G--E--P--A--W--P--N--D  Nostoc
1101   17 Liverwor  GAGTAACAAAAAAACCTGATTTAAGTGATCCTATATTACGAGCTAAATTAGCAAAAGGTATGGGACATAATTATTATGGTGAGCCTGCTTGGCCAAACGA  Liverwor
1101   18 Liverwor  G--V--T--K--K--P--D--L--S--D--P--I--L--R--A--K--L--A--K--G--M--G--H--N--Y--Y--G--E--P--A--W--P--N--D  Liverwor
1101   19 Maize     GAGTAACAAAGAAACCTGACTTAAATGATCCTGTATTAAGAGCAAAATTAGCTAAAGGGATGGGACATAATTATTATGGGGAACCCGCGTGGCCCAACGA  Maize
1101   20 Maize     G--V--T--K--K--P--D--L--N--D--P--V--L--R--A--K--L--A--K--G--M--G--H--N--Y--Y--G--E--P--A--W--P--N--D  Maize
```


Posi-
tion:
Sequence
identity:
Data:
1201 1 Med4
1201 2 Med4
1201 3 MIT9107
1201 4 MIT9107
1201 5 fp5
1201 6 fp5
1201 7 SS120
1201 8 SS120
1201 9 WH8103
1201 10 WH8103
1201 11 Pthrix TCTGCTCTACACCTTCCCGGTGGTGATCTTGGGTACCCTAGCCTGTGTTGTCGGGCTAGCTGTGCTGGATCCCGCCATGGTCGGTGAGCCTGCCAACCCC
1201 12 Pthrix --L--L--Y--T--F--P--V--V--I--L--G--T--L--A--C--V--V--G--L--A--V--L--D--P--A--M--V--G--E--P--A--N--P-
1201 13 PCC7002 CATTCTATTTACCTTCCCCATCTGTATTGCGGGAACCATCGGTTTAATTACCGGCTTGGCGATTCTCGATCCAGCGATGATCGGTGAACCGGGTAATCCT
1201 14 PCC7002 --I--L--F--T--F--P--I--C--I--A--G--T--I--G--L--I--T--G--L--A--I--L--D--P--A--M--I--G--E--P--G--N--P-
1201 15 Nostoc CCTACTTTACGTGTTTCCCATCGTAATTATGGGTTCCTTCGCTGCAATTGTGGCTCTAGCGGTGCTAGACCCGGCGATGACAGGTGAACCAGCAAATCCT
1201 16 Nostoc --L--L--Y--V--F--P--I--V--I--M--G--S--F--A--A--I--V--A--L--A--V--L--D--P--A--M--T--G--E--P--A--N--P-
1201 17 Liverwor TCTTTTATATATTTTTCCAGTAGTTATTTTAGGTACTATTGCGTGTACTGTTGGTTTAGCTGTTTTAGAACCTTCAATGATTGGTGAACCTGCAAATCCT
1201 18 Liverwor --L--L--Y--I--F--P--V--V--I--L--G--T--I--A--C--T--V--G--L--A--V--L--E--P--S--M--I--G--E--P--A--N--P-
1201 19 Maize TCTTTTATACATTTTTCCAGTAGTAATTCTAGGTACTATTGCGTGTAATGTAGGTTTAGCGGTTCTAGAGCCGTCAATGATTGGTGAACCGGCGGATCC
1201 20 Maize --L--L--Y--I--F--P--V--V--I--L--G--T--I--A--C--N--V--G--L--A--V--L--E--P--S--M--I--G--E--P--A--D-
1301 1 Med4
1301 2 Med4
1301 3 MIT9107
1301 4 MIT9107
1301 5 fp5
1301 6 fp5
1301 7 SS120
1301 8 SS120
1301 9 WH8103
1301 10 WH8103
1301 11 Pthrix TTCGCCACTCCCCTGGAAATTCTGCCGGAGTGGTACCTGTACCCTGCTTTCCAAATTCTGCGGGTTGTTCCCAATAAACTGTTGGGGATTTTGCTGCAAA
1301 12 Pthrix -F--A--T--P--L--E--I--L--P--E--W--Y--L--Y--P--A--F--Q--I--L--R--V--V--P--N--K--L--L--G--I--L--L--Q--
1301 13 PCC7002 TTCGCAACACCCCTGGAAATTTTGCCGGAGTGGTACCTCTACCCCGTCTTCCAGATCCTCCGGGTACTGCCCAACAAACTGCTGGGTATTGCTTGTCAAG
1301 14 PCC7002 -F--A--T--P--L--E--I--L--P--E--W--Y--L--Y--P--V--F--Q--I--L--R--V--L--P--N--K--L--L--G--I--A--C--Q--
1301 15 Nostoc TTCGCCACACCATTGGAAATTCTACCAGAGTGGTACTTGTATCCAGTGTTCCAAATTTTGCGATCGCTTCCTAACAAACTTTTAGGAGTGTTAGCAATGG
1301 16 Nostoc -F--A--T--P--L--E--I--L--P--E--W--Y--L--Y--P--V--F--Q--I--L--R--S--L--P--N--K--L--L--G--V--L--A--M--
1301 17 Liverwor TTTGCAACTCCTTTAGAAATTTTACCAGAATGGTATTTTTTTCCAGTTTTTCAAATACTTCGTACGGTACCTAATAAACTTTTAGGTGTACTTTTAATGG
1301 18 Liverwor -F--A--T--P--L--E--I--L--P--E--W--Y--F--F--P--V--F--Q--I--L--R--T--V--P--N--K--L--L--G--V--L--L--M--
1301 19 Maize
1301 20 Maize
Med4
Med4
MIT9107
MIT9107
fp5
fp5
SS120
SS120
WH8103
WH8103
Pthrix
Pthrix
PCC7002
PCC7002
Nostoc
Liverwor
Liverwor
Maize
Maize
Med4
Med4
MIT9107
MIT9107
fp5
fp5
SS120
SS120
WH8103
WH8103
Pthrix
Pthrix
PCC7002
PCC7002
Nostoc
Nostoc
Liverwor
Liverwor
Maize
Maize

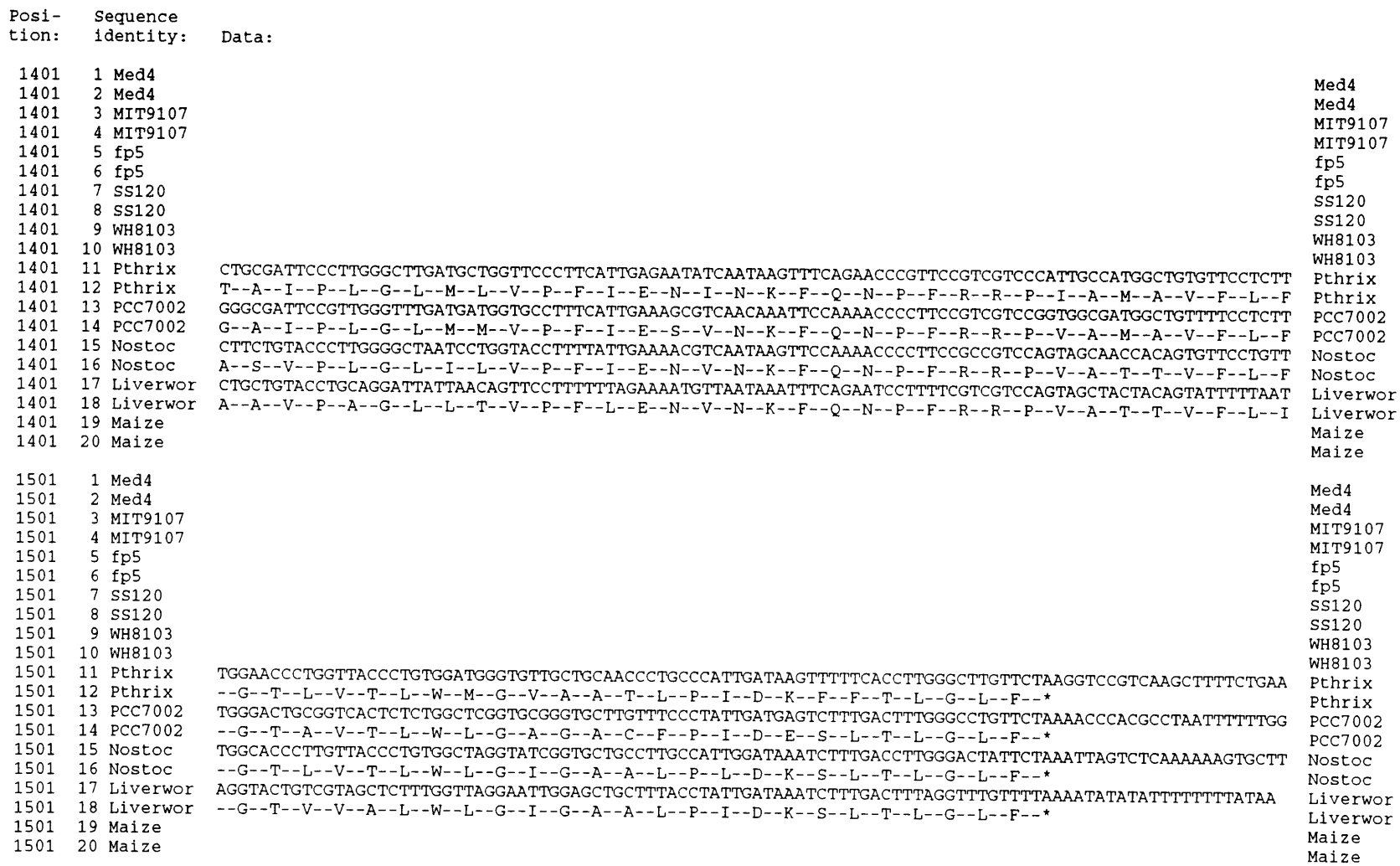
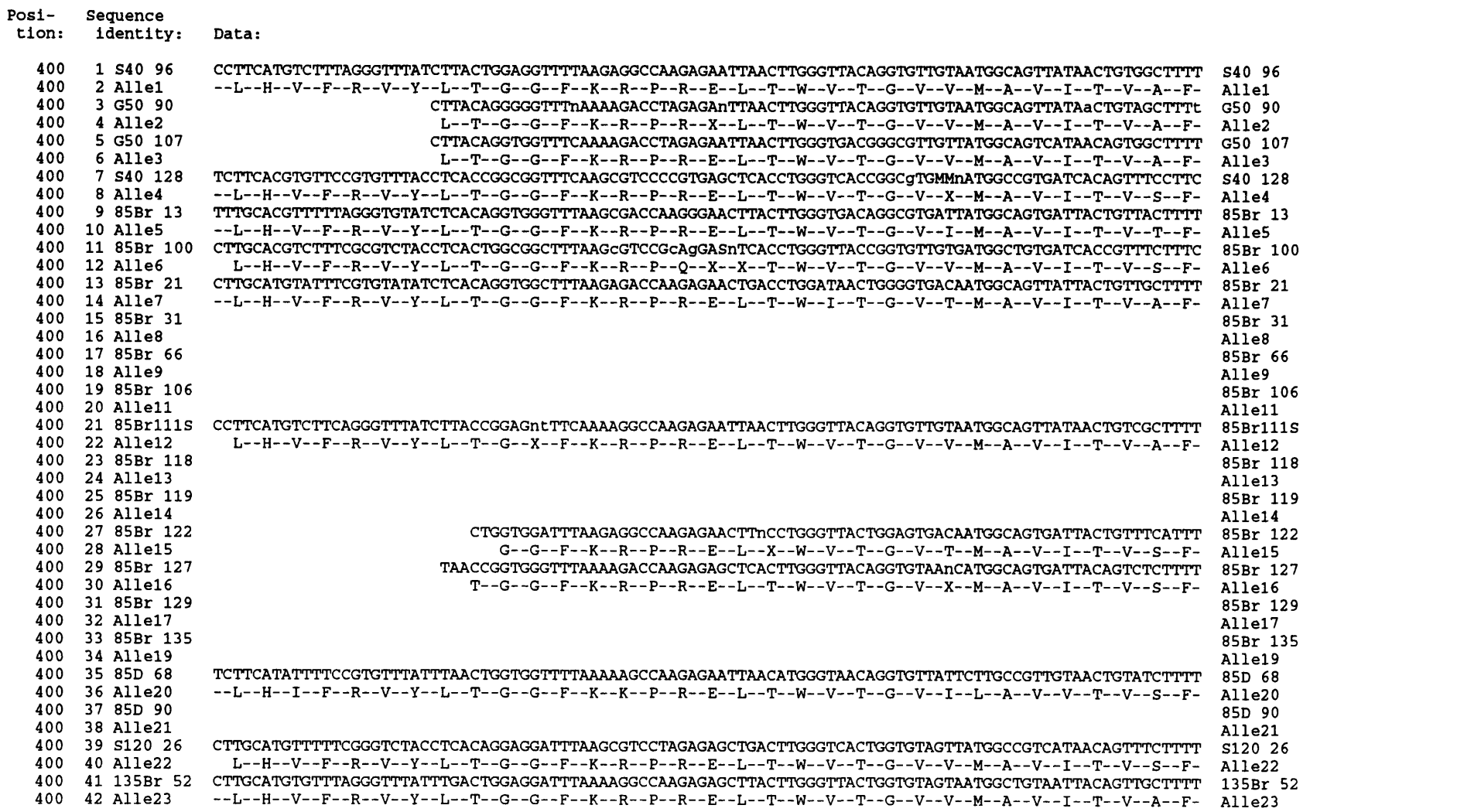
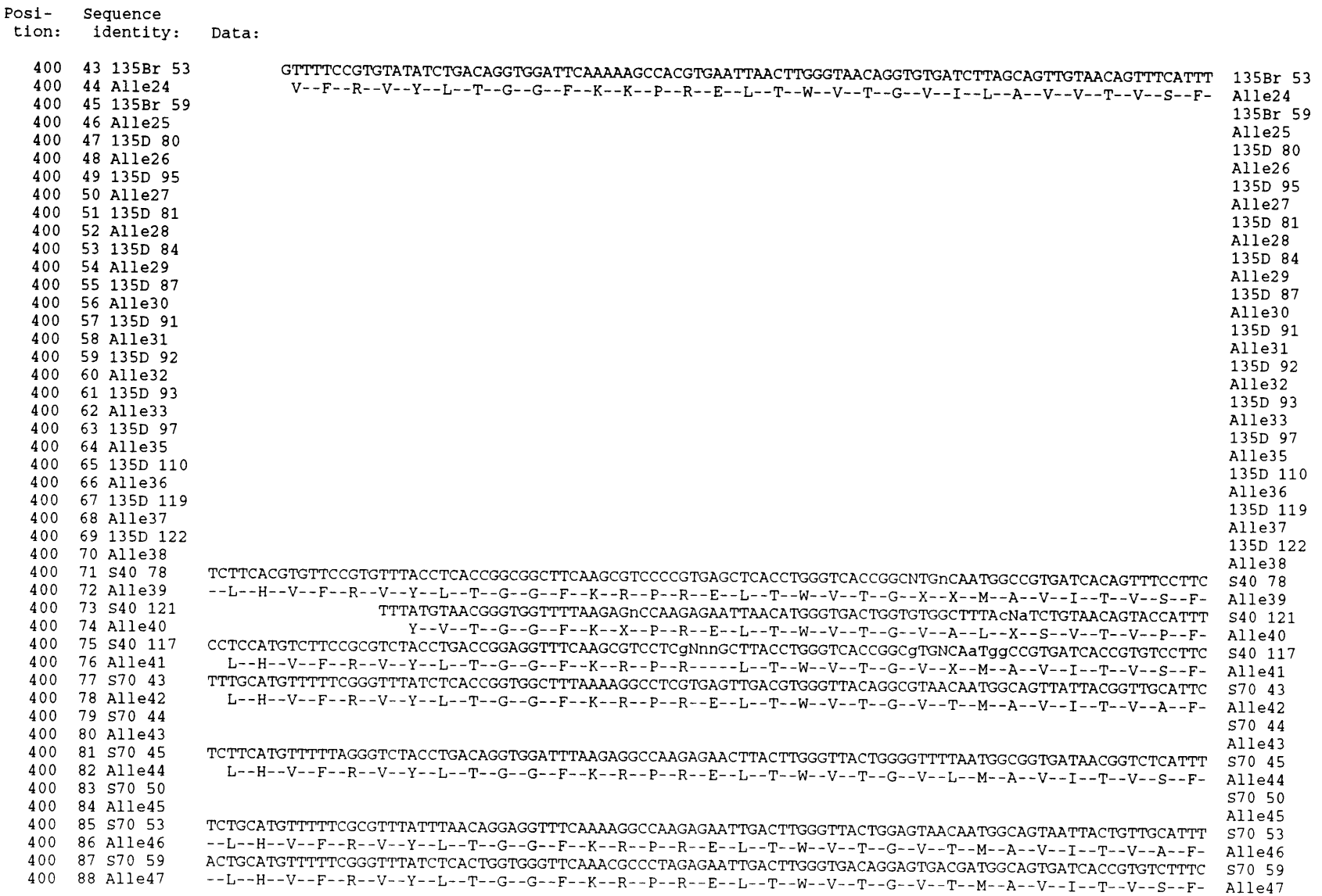
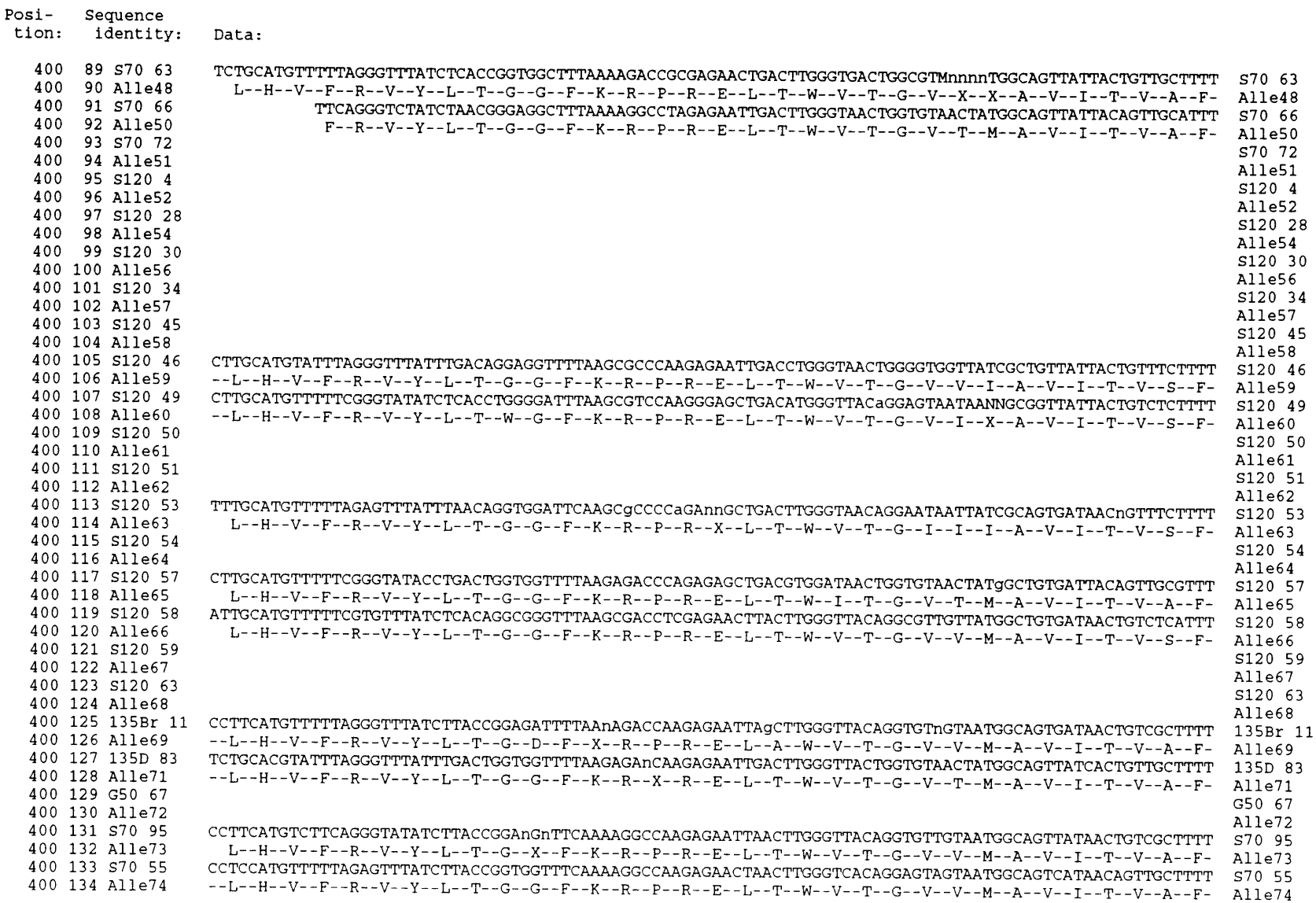

Posi-
tion:
Sequence
identity:
Data:
1401 1 Med4
1401 2 Med4
1401 3 MIT9107
1401 4 MIT9107
1401 5 fp5
1401 6 fp5
1401 7 SS120
1401 8 SS120
1401 9 WH8103
1401 10 WH8103
1401 11 Pthrix CTGCGATTCCCTTGGGCTTGATGCTGGTTCCCTTCATTGAGAATATCAATAAGTTTCAGAACCCGTTCCGTCGTCCCATTGCCATGGCTGTGTTCCTCTT
1401 12 Pthrix T--A--I--P--L--G--L--M--L--V--P--F--I--E--N--I--N--K--F--Q--N--P--F--R--R--P--I--A--M--A--V--F--L--F
1401 13 PCC7002 GGGCGATTCCGTTGGGTTTGATGATGGTGCCTTTCATTGAAAGCGTCAACAAATTCCAAAACCCCTTCCGTCGTCCGGTGGCGATGGCTGTTTTCCTCTT
1401 14 PCC7002 G--A--I--P--L--G--L--M--M--V--P--F--I--E--S--V--N--K--F--Q--N--P--F--R--R--P--V--A--M--A--V--F--L--F
1401 15 Nostoc CTTCTGTACCCTTGGGGCTAATCCTGGTACCTTTTATTGAAAACGTCAATAAGTTCCAAAACCCCTTCCGCCGTCCAGTAGCAACCACAGTGTTCCTGTT
1401 16 Nostoc A--S--V--P--L--G--L--I--L--V--P--F--I--E--N--V--N--K--F--Q--N--P--F--R--R--P--V--A--T--T--V--F--L--F
1401 17 Liverwor CTGCTGTACCTGCAGGATTATTAACAGTTCCTTTTTTAGAAAATGTTAATAAATTTCAGAATCCTTTTCGTCGTCCAGTAGCTACTACAGTATTTTTAAT
1401 18 Liverwor A--A--V--P--A--G--L--L--T--V--P--F--L--E--N--V--N--K--F--Q--N--P--F--R--R--P--V--A--T--T--V--F--L--I
1401 19 Maize
1401 20 Maize
Med4
Med4
MIT9107
MIT9107
fp5
fp5
SS120
SS120
WH8103
WH8103
Pthrix
Pthrix
PCC7002
PCC7002
Nostoc
Nostoc
Liverwor
Liverwor
Maize
Maize
1501 1 Med4
1501 2 Med4
1501 3 MIT9107
1501 4 MIT9107
1501 5 fp5
1501 6 fp5
1501 7 SS120
1501 8 SS120
1501 9 WH8103
1501 10 WH8103
1501 11 Pthrix TGGAACCCTGGTTACCCTGTGGATGGGTGTTGCTGCAACCCTGCCCATTGATAAGTTTTTCACCTTGGGCTTGTTCTAAGGTCCGTCAAGCTTTTCTGAA
1501 12 Pthrix --G--T--L--V--T--L--W--M--G--V--A--A--T--L--P--I--D--K--F--F--T--L--G--L--F--*
1501 13 PCC7002 TGGGACTGCGGTCACTCTCTGGCTCGGTGCGGGTGCTTGTTTCCCTATTGATGAGTCTTTGACTTTGGGCCTGTTCTAAAACCCACGCCTAATTTTTTGG
1501 14 PCC7002 --G--T--A--V--T--L--W--L--G--A--G--A--C--F--P--I--D--E--S--L--T--L--G--L--F--*
1501 15 Nostoc TGGCACCCTTGTTACCCTGTGGCTAGGTATCGGTGCTGCCTTGCCATTGGATAAATCTTTGACCTTGGGACTATTCTAAATTAGTCTCAAAAAAGTGCTT
1501 16 Nostoc --G--T--L--V--T--L--W--L--G--I--G--A--A--L--P--L--D--K--S--L--T--L--G--L--F--*
1501 17 Liverwor AGGTACTGTCGTAGCTCTTTGGTTAGGAATTGGAGCTGCTTTACCTATTGATAAATCTTTGACTTTAGGTTTGTTTTAAAATATATATTTTTTTTATAA
1501 18 Liverwor --G--T--V--V--A--L--W--L--G--I--G--A--A--L--P--I--D--K--S--L--T--L--G--L--F--*
1501 19 Maize
1501 20 Maize
Med4
Med4
MIT9107
MIT9107
fp5
fp5
SS120
SS120
WH8103
WH8103
Pthrix
Pthrix
PCC7002
PCC7002
Nostoc
Nostoc
Liverwor
Liverwor
Maize
Maize

Figure 4. *petB/D* sequence data for alleles cloned from flow cytometrically sorted cells from the Gulf Stream and the Sargasso Sea and from cultured cells, used for analyses in Chapter Three. Labels on nucleotide sequence lines are clone designations, labels on amino acid translation lines are allele designations.




Posi-
tion:
Sequence
identity:
Data:
400 43 135Br 53
400 44 Alle24
400 45 135Br 59
400 46 Alle25
400 47 135D 80
400 48 Alle26
400 49 135D 95
400 50 Alle27
400 51 135D 81
400 52 Alle28
400 53 135D 84
400 54 Alle29
400 55 135D 87
400 56 Alle30
400 57 135D 91
400 58 Alle31
400 59 135D 92
400 60 Alle32
400 61 135D 93
400 62 Alle33
400 63 135D 97
400 64 Alle35
400 65 135D 110
400 66 Alle36
400 67 135D 119
400 68 Alle37
400 69 135D 122
400 70 Alle38
400 71 S40 78
400 72 Alle39
400 73 S40 121
400 74 Alle40
400 75 S40 117
400 76 Alle41
400 77 S70 43
400 78 Alle42
400 79 S70 44
400 80 Alle43
400 81 S70 45
400 82 Alle44
400 83 S70 50
400 84 Alle45
400 85 S70 53
400 86 Alle46
400 87 S70 59
400 88 Alle47
GTTTTCCGTGTATATCTGACAGGTGGATTCAAAAAGCCACGTGAATTAACTTGGGTAACAGGTGTGATCTTAGCAGTTGTAACAGTTTCATTT
V--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--I--L--A--V--V--T--V--S--F-
TCTTCACGTGTTCCGTGTTTACCTCACCGGCGGCTTCAAGCGTCCCCGTGAGCTCACCTGGGTCACCGGCNTGnCAATGGCCGTGATCACAGTTTCCTTC
--L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--X--X--M--A--V--I--T--V--S--F-
TTTATGTAACGGGTGGTTTTAAGAGnCCAAGAGAATTAACATGGGTGACTGGTGTGGCTTTAcNaTCTGTAACAGTACCATTT
Y--V--T--G--G--F--K--X--P--R--E--L--T--W--V--T--G--V--A--L--X--S--V--T--V--P--F-
CCTCCATGTCTTCCGCGTCTACCTGACCGGAGGTTTCAAGCGTCCTCgNnnGCTTACCTGGGTCACCGGCgTGNCAaTggCCGTGATCACCGTGTCCTTC
L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R-----L--T--W--V--T--G--V--X--M--A--V--I--T--V--S--F-
TTTGCATGTTTTTCGGGTTTATCTCACCGGTGGCTTTAAAAGGCCTCGTGAGTTGACGTGGGTTACAGGCGTAACAATGGCAGTTATTACGGTTGCATTC
L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--A--F-
TCTTCATGTTTTTAGGGTCTACCTGACAGGTGGATTTAAGAGGCCAAGAGAACTTACTTGGGTTACTGGGGTTTTAATGGCGGTGATAACGGTCTCATTT
L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--L--M--A--V--I--T--V--S--F-
TCTGCATGTTTTTCGCGTTTATTTAACAGGAGGTTTCAAAAGGCCAAGAGAATTGACTTGGGTTACTGGAGTAACAATGGCAGTAATTACTGTTGCATTT
--L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--A--F-
ACTGCATGTTTTTCGGGTTTATCTCACTGGTGGGTTCAAACGCCCTAGAGAATTGACTTGGGTGACAGGAGTGACGATGGCAGTGATCACCGTGTCTTTC
--L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--S--F-
135Br 53
Alle24
135Br 59
Alle25
135D 80
Alle26
135D 95
Alle27
135D 81
Alle28
135D 84
Alle29
135D 87
Alle30
135D 91
Alle31
135D 92
Alle32
135D 93
Alle33
135D 97
Alle35
135D 110
Alle36
135D 119
Alle37
135D 122
Alle38
S40 78
Alle39
S40 121
Alle40
S40 117
Alle41
S70 43
Alle42
S70 44
Alle43
S70 45
Alle44
S70 50
Alle45
S70 53
Alle46
S70 59
Alle47


Posi-
tion:
Sequence
identity:
Data:
400 89 S70 63 TCTGCATGTTTTTTAGGGTTTATCTCACCGGTGGCTTTAAAAGACCGCGAGAACTGACTTGGGTGACTGGCGTMnnnnTGGCAGTTATTACTGTTGCTTTT S70 63
400 90 Alle48 L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--X--X--A--V--I--T--V--A--F- Alle48
400 91 S70 66 TTCAGGGTCTATCTAACGGGAGGCTTTAAAAGGCCTAGAGAATTGACTTGGGTAACTGGTGTAACTATGGCAGTTATTACAGTTGCATTT S70 66
400 92 Alle50 F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--A--F- Alle50
400 93 S70 72 S70 72
400 94 Alle51 Alle51
400 95 S120 4 S120 4
400 96 Alle52 Alle52
400 97 S120 28 S120 28
400 98 Alle54 Alle54
400 99 S120 30 S120 30
400 100 Alle56 Alle56
400 101 S120 34 S120 34
400 102 Alle57 Alle57
400 103 S120 45 S120 45
400 104 Alle58 Alle58
400 105 S120 46 CTTGCATGTATTTAGGGTTTATTTGACAGGAGGTTTTAAGCGCCCAAGAGAATTGACCTGGGTAACTGGGGTGGTTATCGCTGTTATTACTGTTTCTTTT S120 46
400 106 Alle59 --L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--I--A--V--I--T--V--S--F- Alle59
400 107 S120 49 CTTGCATGTTTTTCGGGTATATCTCACCTGGGGATTTAAGCGTCCAAGGGAGCTGACATGGGTTACaGGAGTAATAANNGCGGTTATTACTGTCTCTTTT S120 49
400 108 Alle60 --L--H--V--F--R--V--Y--L--T--W--G--F--K--R--P--R--E--L--T--W--V--T--G--V--I--X--A--V--I--T--V--S--F- Alle60
400 109 S120 50 S120 50
400 110 Alle61 Alle61
400 111 S120 51 S120 51
400 112 Alle62 Alle62
400 113 S120 53 TTTGCATGTTTTTAGAGTTTATTTAACAGGTGGATTCAAGCgCCCCaGAnnGCTGACTTGGGTAACAGGAATAATTATCGCAGTGATAACnGTTTCTTTT S120 53
400 114 Alle63 L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--X--L--T--W--V--T--G--I--I--I--A--V--I--T--V--S--F- Alle63
400 115 S120 54 S120 54
400 116 Alle64 Alle64
400 117 S120 57 CTTGCATGTTTTTCGGGTATACCTGACTGGTGGTTTTAAGAGACCCAGAGAGCTGACGTGGATAACTGGTGTAACTATgGCTGTGATTACAGTTGCGTTT S120 57
400 118 Alle65 L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--I--T--G--V--T--M--A--V--I--T--V--A--F- Alle65
400 119 S120 58 ATTGCATGTTTTTCGTGTTTATCTCACAGGCGGGTTTAAGCGACCTCGAGAACTTACTTGGGTTACAGGCGTTGTTATGGCTGTGATAACTGTCTCATTT S120 58
400 120 Alle66 L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--S--F- Alle66
400 121 S120 59 S120 59
400 122 Alle67 Alle67
400 123 S120 63 S120 63
400 124 Alle68 Alle68
400 125 135Br 11 CCTTCATGTTTTTAGGGTTTATCTTACCGGAGATTTTAAnAGACCAAGAGAATTAgCTTGGGTTACAGGTGTnGTAATGGCAGTGATAACTGTCGCTTTT 135Br 11
400 126 Alle69 --L--H--V--F--R--V--Y--L--T--G--D--F--X--R--P--R--E--L--A--W--V--T--G--V--V--M--A--V--I--T--V--A--F- Alle69
400 127 135D 83 TCTGCACGTATTTAGGGTTTATTTGACTGGTGGTTTTAAGAGAnCAAGAGAATTGACTTGGGTTACTGGTGTAACTATGGCAGTTATCACTGTTGCTTTT 135D 83
400 128 Alle71 --L--H--V--F--R--V--Y--L--T--G--G--F--K--R--X--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--A--F- Alle71
400 129 G50 67 G50 67
400 130 Alle72 Alle72
400 131 S70 95 CCTTCATGTCTTCAGGGTATATCTTACCGGAnGnTTCAAAAGGCCAAGAGAATTAACTTGGGTTACAGGTGTTGTAATGGCAGTTATAACTGTCGCTTTT S70 95
400 132 Alle73 L--H--V--F--R--V--Y--L--T--G--X--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F- Alle73
400 133 S70 55 CCTCCATGTTTTTAGAGTTTATCTTACCGGTGGTTTCAAAAGGCCAAGAGAACTAACTTGGGTCACAGGAGTAGTAATGGCAGTCATAACAGTTGCTTTT S70 55
400 134 Alle74 --L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F- Alle74

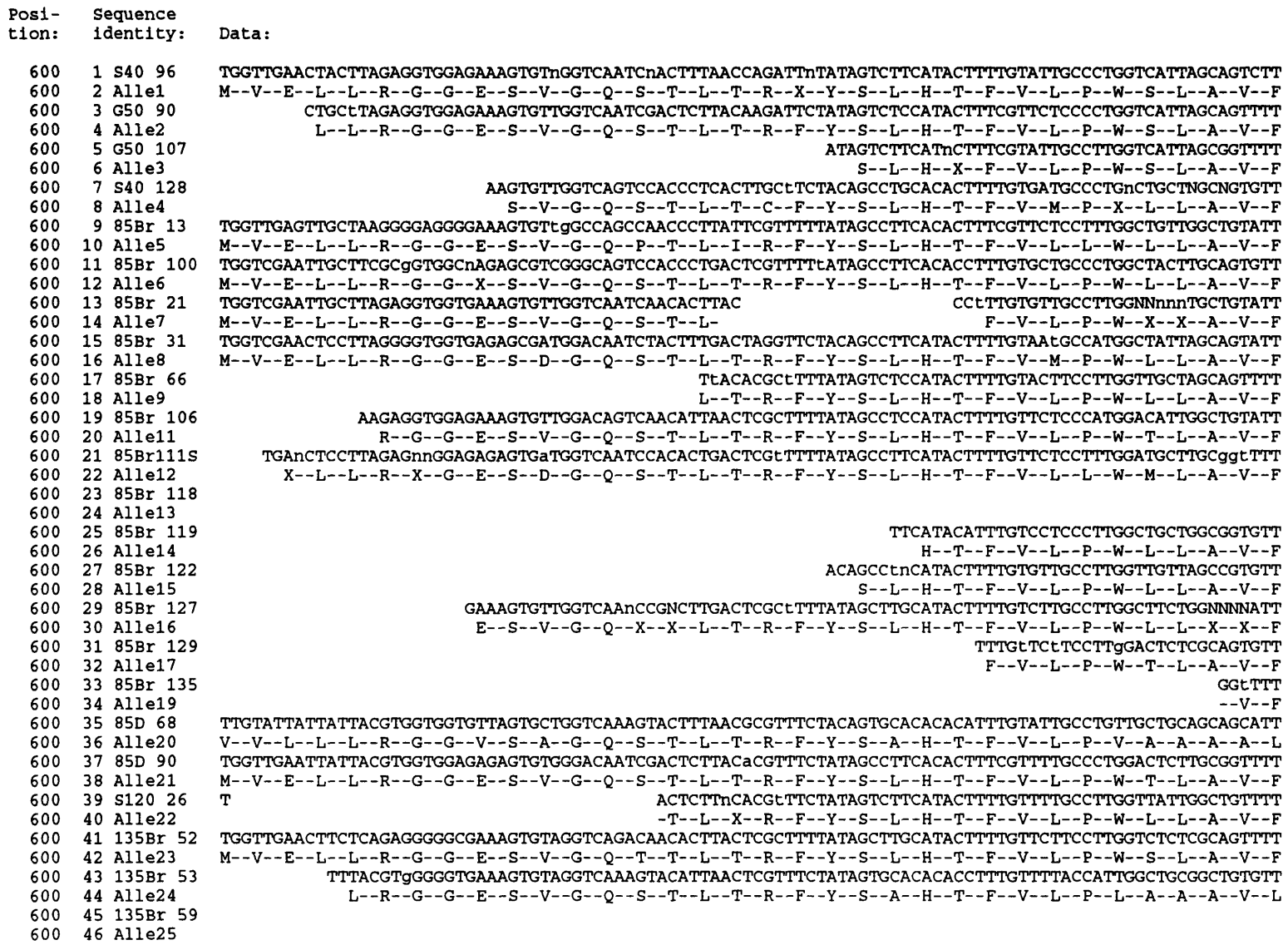
Position:	Sequence identity:	Data:	
400 135	G50 69		G50 69
400 136	Alle75		Alle75
400 137	G50 70	CCTTCATGTTTTTAGGGTTTATCTTACCGGTGGGTTTAAAAGGCCAAGAGAACTAACCTGGGTTACAGGTGTTGTAATGGCAGTTATAACAGTTGCTTTT	G50 70
400 138	Alle76	L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F-	Alle76
400 139	S40 137	TCTTCACGTGTTCCGGGTTTACCTCACCGGAGnnTTCAAGCGTCCCCGTGAGCTCACCTGGGTCACTGGCGTGanCATgGCCGTGATCACAGTTTCTTTC	S40 137
400 140	Alle77	--L--H--V--F--R--V--Y--L--T--G--X--F--K--R--P--R--E--L--T--W--V--T--G--V--X--M--A--V--I--T--V--S--F-	Alle77
400 141	G50 72		G50 72
400 142	Alle78		Alle78
400 143	Med4	TCTTCATGTATTTAGGGTATATCTTACTGGAGGATTTAAAAGACCAAGAGAATTAACATGGGTAACGGGGGTTGTTATGGCAGTTATAACAGTTGCTTTC	Med4
400 144	Med4	L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F-	Med4
400 145	SS2	TNNGCATGTTTTTAGAGTTTATTTAACGGGCGGATTTAAAAGACCTAGGGAGCTTACNTGGGTTACAGGAGTTGTAATGGCAGTTATAACAGTTGCTTtT	SS2
400 146	SS2	X--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F-	SS2
400 147	FP5	TCTGCATGTTTTCCGCGTTTATTTGACTGGTGGATTTAAAAGACCTAGAGAATTGACATGGATAACTGGAGTGGTTATGGCAGTAATCACAGTTGCTTTT	FP5
400 148	FP5	L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--I--T--G--V--V--M--A--V--I--T--V--A--F-	FP5
400 149	Pac7	CCTTCACGTTTTTAGAGTTTATCTTACGGGTGGTTTTAAAAGGCCCAGAGAACTAACGTGGGTTACAGGTGTAGTAATGGCAGTAATAACTGTCGCTTTT	Pac7
400 150	Pac7	L--H--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--V--M--A--V--I--T--V--A--F-	Pac7
400 151	MIT9303		MIT9303
400 152	MIT9303		MIT9303
400 153	MIT9313		MIT9313
400 154	MIT9313		MIT9313
400 155	WH8103	TCTTTACGTGTTCCGGGTCTACCTCACCGGTGGTTTCAAGCGTCCCCGTGAGCTCACCTGGGTCACCGGCGTGACCATGGCCGTGATCACAGTTTCcTTC	WH8103
400 156	WH8103	L--Y--V--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--V--T--G--V--T--M--A--V--I--T--V--S--F-	WH8103
400 157	PCC7002	TCTTCACATTTTCCGGGTGTACCTCACCGGTGGCTTTAAACGTCCCCGTGAGCTGACTTGGATTACTGGGGTCATCATGGCGACGATCACCGTTTCCTTC	PCC7002
400 158	PCC7002	--L--H--I--F--R--V--Y--L--T--G--G--F--K--R--P--R--E--L--T--W--I--T--G--V--I--M--A--T--I--T--V--S--F-	PCC7002
400 159	Pthrix	CCTCCATGTGTTCCGGGTGTACCTCACCGGTGGTTTCAAAAATCCCCGCGAACTGACCTGGATTACGGGGGTTATTTTGGCGGTGATCACCGTATCCTTC	Pthrix
400 160	Pthrix	--L--H--V--F--R--V--Y--L--T--G--G--F--K--N--P--R--E--L--T--W--I--T--G--V--I--L--A--V--I--T--V--S--F-	Pthrix
400 161	Nostoc	TTTGCACGTCTTCCGGGTTTATCTGACTGGTGGTTTTAAAAAGCCCCGCGAATTGACCTGGGTGAGTGGTGTGATTTTGGCTGTAATTACCGTTTCCTTT	Nostoc
400 162	Nostoc	--L--H--V--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--S--G--V--I--L--A--V--I--T--V--S--F-	Nostoc
400 163	Chlorell	TCTTCATATTTTTCGTGTTTATCTAACTGGTGGCTTTAAAAAACCACGCGAATTAACTTGGGTAACTGGTGTATTAATGGCCGTATGTACTGTTTCATTT	Chlorell
400 164	Chlorell	--L--H--I--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--L--M--A--V--C--T--V--S--F-	Chlorell
400 165	Liverwor	TTTACATATTTTTCGTGTTTATCTAACAGGAGGTTTTAAAAAACCTCGGGAATTAACTTGGGTTACTGGTGTTATTTTAGCAGTTTTAACTGTATCTTTT	Liverwor
400 166	Liverwor	--L--H--I--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--I--L--A--V--L--T--V--S--F-	Liverwor
400 167	Maize	CCTGCACGTATTTCGTGTGTATCTCACAGGTGGATTTAAAAAACCCCGTGAATTAACTTGGGTCACAGGCGTGGTTTTGGCTGTATTGACTGCATCATTT	Maize
400 168	Maize	--L--H--V--F--R--V--Y--L--T--G--G--F--K--K--P--R--E--L--T--W--V--T--G--V--V--L--A--V--L--T--A--S--F-	Maize

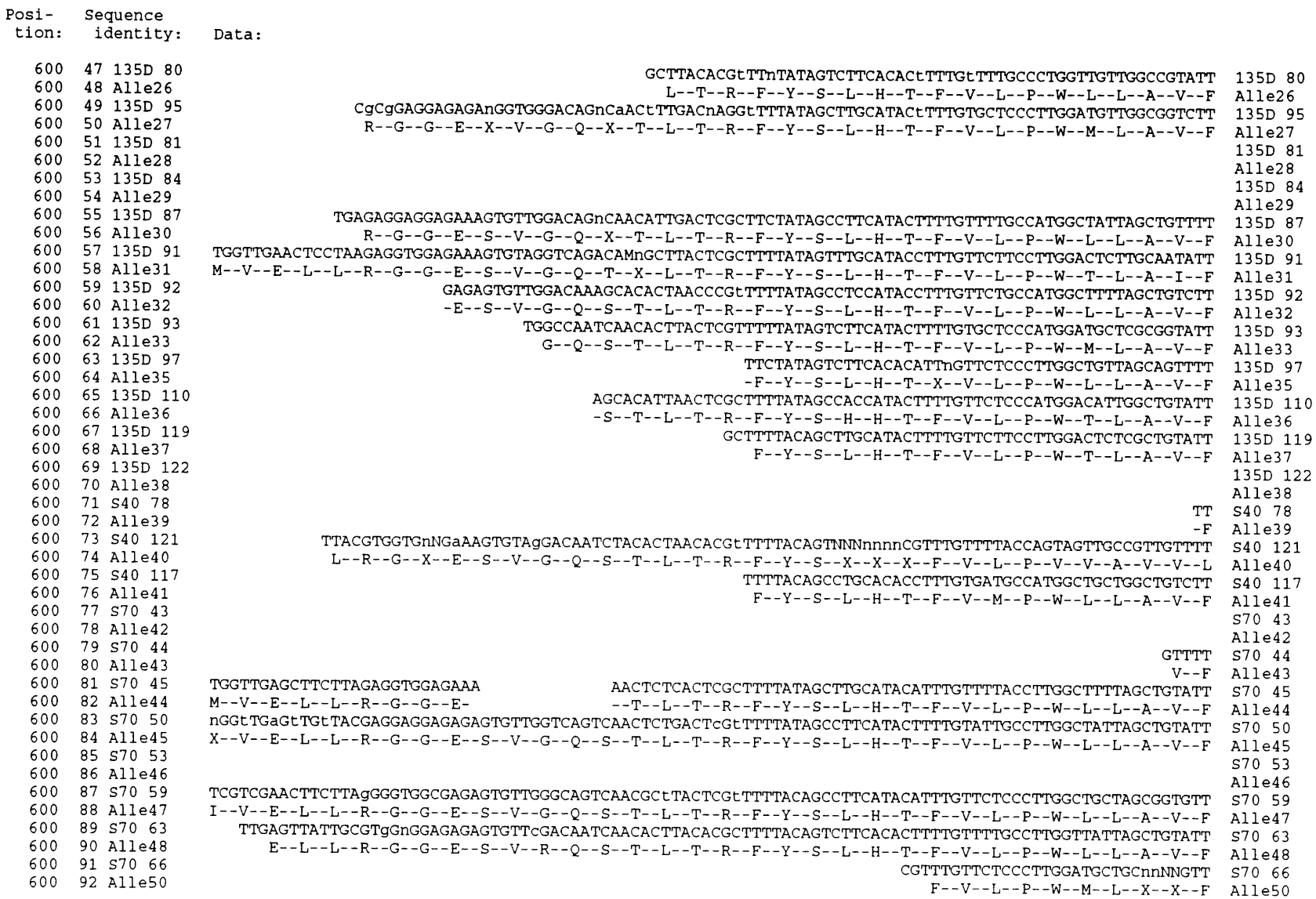

Posi-
tion:
Sequence
identity:
Data:
500 1 S40 96 GGAGTGACTGGATACTCCCTACCTTGGGATCAGGTCGGTTATTGGGCAGTCAAGATTGTTTCAGGTGTTCCTGCTGCAATACCAGTAATAGGTGACTTTA S40 96
500 2 Alle1 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F-- Alle1
500 3 G50 90 GGTGTTACAGGATATTCTT G50 90
500 4 Alle2 -G--V--T--G--Y--S- Alle2
500 5 G50 107 GGAGTTACAGGATATTCTTTGCCTG G50 107
500 6 Alle3 -G--V--T--G--Y--S--L--P- Alle3
500 7 S40 128 GGTGTCACCGGTTACTCCCTGCCCTGGGACCAGnTTGGTTATTGGGCCGTCAAGATTGTTTC S40 128
500 8 Alle4 -G--V--T--G--Y--S--L--P--W--D--Q--X--G--Y--W--A--V--K--I--V-- Alle4
500 9 85Br 13 GGAGTTACTGGATATTCTCTTCCCTGGGATCAGGTTGGTTATTGGGCCGTAAAAATTGTCTCAGGTGTTCCTGCCGCTGTCCCAGTTGTTGGTGATTTCA 85Br 13
500 10 Alle5 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--V--P--V--V--G--D--F-- Alle5
500 11 85Br 100 GAGGTAACCGGCTATTCCCTCCCATGGGATCAAGTCGGGTACTGGGCTGTGAAGATCGTTTCAGCGGTGCCTGCTGCTATCCCCGTGGTTGGAGACTCTA 85Br 100
500 12 Alle6 -E--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--A--V--P--A--A--I--P--V--V--G--D--S-- Alle6
500 13 85Br 21 GGCGTTACTGGTTACTCTTTGCCTTGGGACCAAGTTGGATATTGGGCTGTCAAAATTGTTTCTGGGGTCCCAGCTGCAATTCCCGTGATAGGAGACTTCA 85Br 21
500 14 Alle7 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F-- Alle7
500 15 85Br 31 GGATCAGGTTGGCTATTGGGCgGTGAnGATTGTTtCAgGAGTCCCAGCAGCTATTCCAGTAATTGGTGACTTTA 85Br 31
500 16 Alle8 D--Q--V--G--Y--W--A--V--X--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F-- Alle8
500 17 85Br 66 85Br 66
500 18 Alle9 Alle9
500 19 85Br 106 85Br 106
500 20 Alle11 Alle11
500 21 85Br111S GGAGTGACAGGGTACTCCCTTCCTTGGGATCAnGTAGGTTATTGGGCAGTTAAGATTGTnTCAGGTGTnCCCGCTGCAATACnAGTA 85Br111S
500 22 Alle12 -G--V--T--G--Y--S--L--P--W--D--X--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--X--V- Alle12
500 23 85Br 118 85Br 118
500 24 Alle13 Alle13
500 25 85Br 119 85Br 119
500 26 Alle14 Alle14
500 27 85Br 122 GGTGTCACTGGATATTCTTTGCCTTGGGACCAAGTTGGTTATTGGGCAGTAAAAATTGTTTCAGGTGTCCnTGCAGCCATTCCTGTTGTTGG 85Br 122
500 28 Alle15 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--X--A--A--I--P--V--V- Alle15
500 29 85Br 127 GGGGTTACTGGTTATTCACTTCCTTGGGACCAnGTTGGTTATTGGGCTGTGAAAA 85Br 127
500 30 Alle16 -G--V--T--G--Y--S--L--P--W--D--X--V--G--Y--W--A--V--K- Alle16
500 31 85Br 129 85Br 129
500 32 Alle17 Alle17
500 33 85Br 135 85Br 135
500 34 Alle19 Alle19
500 35 85D 68 GGTGTAGCGGGTTACTCTTTACCATGGGATCAAGTAGGCTACTGGGCTTGTAAAATTGTAACGGGTGTTCCTGAAGCTGTACCTGTAGTTGGAACATTAG 85D 68
500 36 Alle20 -G--V--A--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--C--K--I--V--T--G--V--P--E--A--V--P--V--V--G--T--L-- Alle20
500 37 85D 90 TnCCAGTCGTnGGTGACATTA 85D 90
500 38 Alle21 P--V--V--G--D--I-- Alle21
500 39 S120 26 GGTGTTACTGGTTACTCTCTCCCATGGGATCAAGTTGGCTATTGGGCGGTGAAAATTGTTTCAGGTGTCCcTGCTGCTATCCCTGTTGTAGGTGACTTTA S120 26
500 40 Alle22 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--F- Alle22
500 41 135Br 52 GGGGTTACTGGTTATTCTTTGCCATGGGACCAAGTTGGTTATTGGGCTGTCAAAATTGTTTCAGGTGTACCTGCAGCTATACCAGTAGTTGGTAACTTTA 135Br 52
500 42 Alle23 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--N--F-- Alle23
500 43 135Br 53 GGTGTAACcGGTTATTCTTTACCTTGGGATCAaGTTGGATTCTGGGCATGTAAAATTGTAACAGGTGTTCCTGCCGCAGtACCAATTG 135Br 53
500 44 Alle24 -G--V--T--G--Y--S--L--P--W--D--Q--V--G--F--W--A--C--K--I--V--T--G--V--P--A--A--V--P--I- Alle24
500 45 135Br 59 135Br 59
500 46 Alle25 Alle25

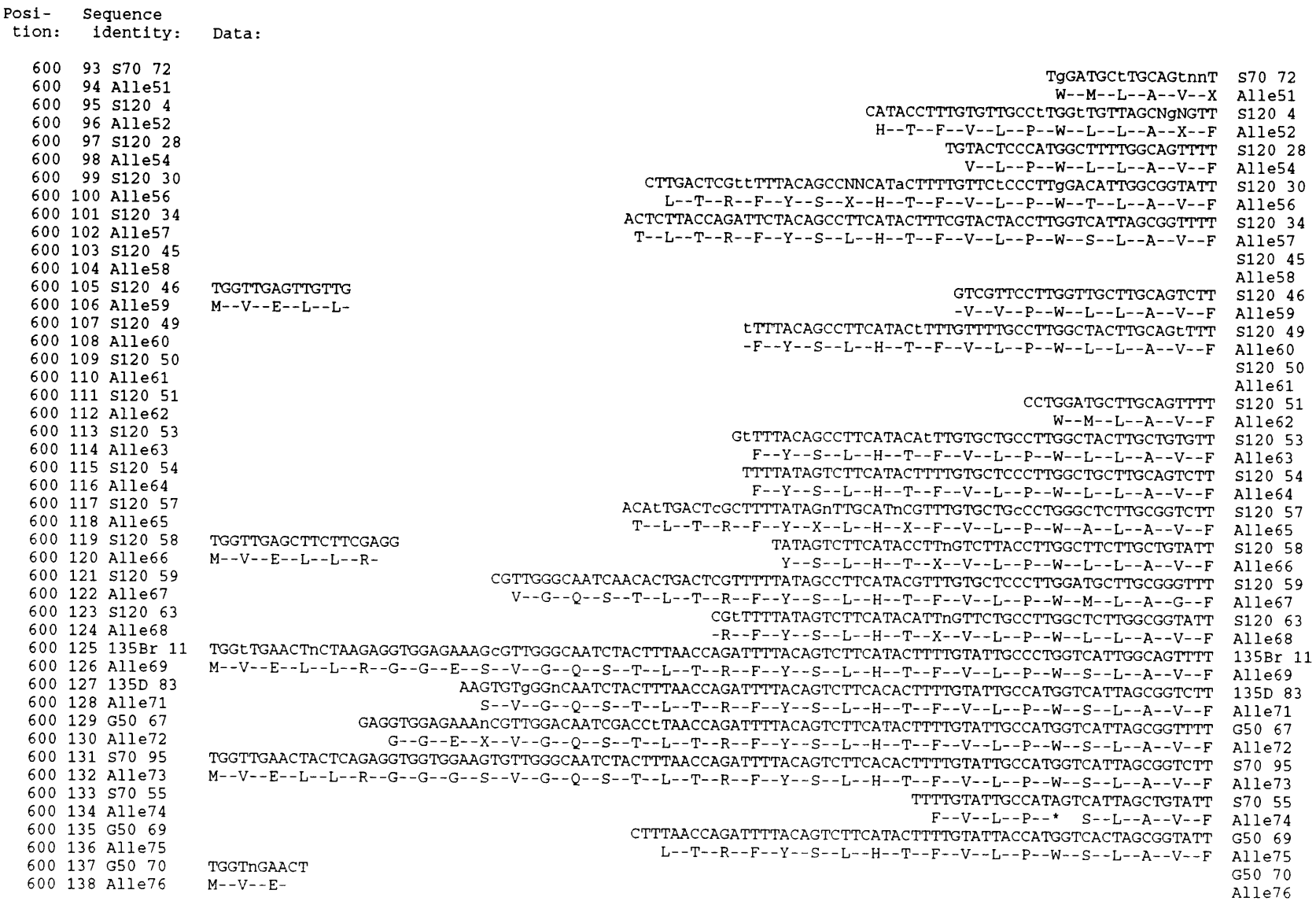

Posi-
tion:
Sequence
identity:
Data:
500 47 135D 80
500 48 Alle26
500 49 135D 95
500 50 Alle27
500 51 135D 81
500 52 Alle28
500 53 135D 84
500 54 Alle29
500 55 135D 87
500 56 Alle30
500 57 135D 91
500 58 Alle31
500 59 135D 92
500 60 Alle32
500 61 135D 93
500 62 Alle33
500 63 135D 97
500 64 Alle35
500 65 135D 110
500 66 Alle36
500 67 135D 119
500 68 Alle37
500 69 135D 122
500 70 Alle38
500 71 S40 78
500 72 Alle39
500 73 S40 121
500 74 Alle40
500 75 S40 117
500 76 Alle41
500 77 S70 43
500 78 Alle42
500 79 S70 44
500 80 Alle43
500 81 S70 45
500 82 Alle44
500 83 S70 50
500 84 Alle45
500 85 S70 53
500 86 Alle46
500 87 S70 59
500 88 Alle47
500 89 S70 63
500 90 Alle48
500 91 S70 66
500 92 Alle50
TCAGnTGtTCCTGCTGCTATTCCAGTAGTCGGTAACTTTA
S--X--V--P--A--A--I--P--V--V--G--N--F--
GGTGTCACCGGTTACTCCCTGCCCTGGGACCAnGTTGGTTATTGGGCCGTCAAGATTGTTTC
-G--V--T--G--Y--S--L--P--W--D--X--V--G--Y--W--A--V--K--I--V-
GGTGTAACTGGATATTCCTTACCTTGGGATCAGTtAGGTTATT
-G--V--T--G--Y--S--L--P--W--D--Q--L--G--Y-
GGCGTnACCGGTTATTCCCTTCCCTGGGATCAnGT
-G--V--T--G--Y--S--L--P--W--D--X--V
GGAGTGACAGGTTATTCACTGCCTTGGGACCAGGTtGGCTATTGGGCAGTAAAGATTGTTTCG
-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S-
GGCGTTACTGGATATTCGCTCCCTTGGnATCAGnTTGGTTATTGGGCTGTGAAAATTGTGTCTGGAGTTCCAGCTGCAATTCCAGTTGTAGGCGACTTTA
-G--V--T--G--Y--S--L--P--W--X--Q--X--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--F--
GCAGCnATACCAGTgGtTgTAGATTTTA
A--A--I--P--V--V--V--D--F--
GGAGTAACAGGTTATTCATTGCCTTGGGnTCAG
-G--V--T--G--Y--S--L--P--W--X-
GGTGTTACGGGGTAnTCCTTGCCATGGGATCAGTnGGGCTATTGGGCAGTGAAGATTGTTTcAGGAGTCCCTGCTGCAATTCcAGTGGTTGGTGAnTTCA
-G--V--T--G--X--S--L--P--W--D--Q--W--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--X--F--
GGAGTAACAGGTTATTCATTGCCTTGGGATCAGGTTGGTTACTGGGCAGTGAAA
-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K-
GGGGTTACTGGATATTCACTCCCTTGGGATCAGGTTGGgTACTGGgCAGTGAAAaTtGTttCTGgAGTACcAGCcGCTATt
-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I-
135D 80
Alle26
135D 95
Alle27
135D 81
Alle28
135D 84
Alle29
135D 87
Alle30
135D 91
Alle31
135D 92
Alle32
135D 93
Alle33
135D 97
Alle35
135D 110
Alle36
135D 119
Alle37
135D 122
Alle38
S40 78
Alle39
S40 121
Alle40
S40 117
Alle41
S70 43
Alle42
S70 44
Alle43
S70 45
Alle44
S70 50
Alle45
S70 53
Alle46
S70 59
Alle47
S70 63
Alle48
S70 66
Alle50

Position:	Sequence identity:	Data:	
500	93 S70 72		S70 72
500	94 Alle51		Alle51
500	95 S120 4		S120 4
500	96 Alle52		Alle52
500	97 S120 28		S120 28
500	98 Alle54		Alle54
500	99 S120 30		S120 30
500	100 Alle56		Alle56
500	101 S120 34		S120 34
500	102 Alle57		Alle57
500	103 S120 45		S120 45
500	104 Alle58		Alle58
500	105 S120 46	GGAGTTACAGGATATTCACTCCCTTGGGATCAGGTTGGTTATTGGGCAGTGAAAATTGTTTCTNGGGTTCCTGCTGCAATACCAGTAGTTNGGGATTTtA	S120 46
500	106 Alle59	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--X--V--P--A--A--I--P--V--V--X--D--F--	Alle59
500	107 S120 49	NGAGTTACAGGTTATTCATTgCCATGGGACCA	S120 49
500	108 Alle60	-X--V--T--G--Y--S--L--P--W--D-	Alle60
500	109 S120 50		S120 50
500	110 Alle61		Alle61
500	111 S120 51		S120 51
500	112 Alle62		Alle62
500	113 S120 53	NGTGTAACTGGTTACTCCCTCCCATGGGATCAAGTTGGCTATTGGGCAGTCAAAATCATCTCTGGAGTTCCTGCCGCAATTCCAGTGG	S120 53
500	114 Alle63	-X--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--I--S--G--V--P--A--A--I--P--V-	Alle63
500	115 S120 54		S120 54
500	116 Alle64		Alle64
500	117 S120 57	GGTGTTACTGGTTATTCACTCCCATGGGATCAAGTTGGTTATTGGGCAGTCAAAATTGTTTCTGGAGTTCCTGCAGCAATTCCAGTGATAGGAGATTtC	S120 57
500	118 Alle65	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F-	Alle65
500	119 S120 58	GGTGTTACCGATTATTCGCTTCCATGGGATCAGGTTGGCTATTGGGCAGTCAAGATCGTTTCAGGTGTTCCWGCGGCAATACCTGTAGTAGGCGACTTTA	S120 58
500	120 Alle66	-G--V--T--D--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--F--	Alle66
500	121 S120 59		S120 59
500	122 Alle67		Alle67
500	123 S120 63		S120 63
500	124 Alle68		Alle68
500	125 135Br 11	GGAGTGACAG GCAATTCCAGTAATTGGCGACTTTA	135Br 11
500	126 Alle69	-G--V--T- A--I--P--V--I--G--D--F--	Alle69
500	127 135D 83		135D 83
500	128 Alle71		Alle71
500	129 G50 67		G50 67
500	130 Alle72		Alle72
500	131 S70 95	GGAGTGACAGGGTACTCCCTTCCTTGGGATCAaGTAGGTTATTGGGCAGTTAAGATTGTTTcAGGTGTCCCcGCTGCAATACCAGTAATTGGCGnCtTTA	S70 95
500	132 Alle73	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--X--F--	Alle73
500	133 S70 55	GGGGTTACAGGATATTCCCTACCTTGGGATCAnGTCGGTTATTGGGCAGTTAAGATTGttTCAGGTGTACCnGCAGCAATACCAGTTATTGGTGATT	S70 55
500	134 Alle74	-G--V--T--G--Y--S--L--P--W--D--X--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D-	Alle74
500	135 G50 69		G50 69
500	136 Alle75		Alle75
500	137 G50 70	GGTGTGACGGGATACTCTTTACCTTGGGATCAAGTCGGTTATTGGGCAGTTAAGATCGTTTCAGGCGTTCCTGCGGCAATACCAGTTATAGGTGANTTTA	G50 70
500	138 Alle76	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--X--F--	Alle76

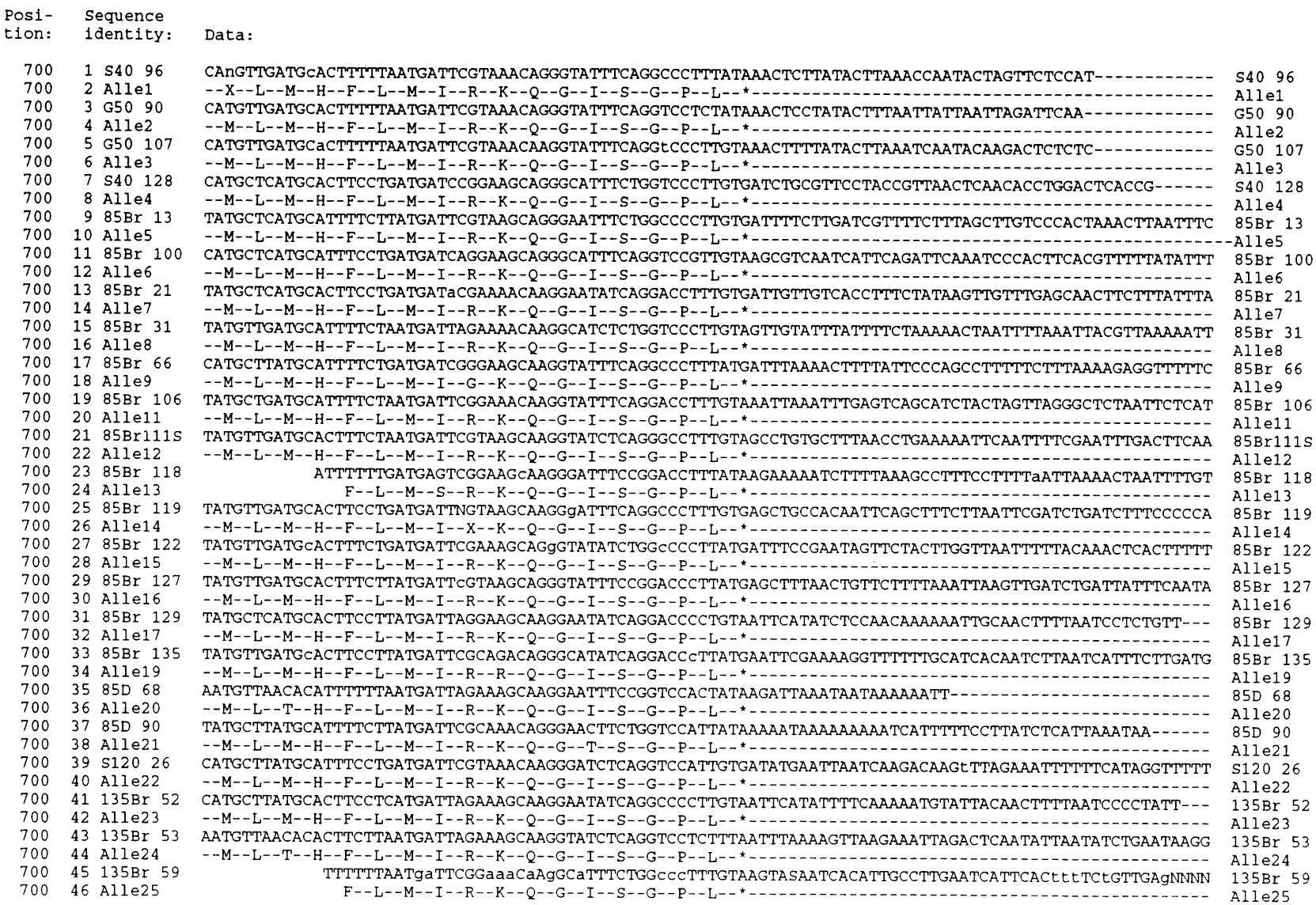
Position	Sequence identity	Data	
500 139	S40 137	GGTGTCACCGGTTACTCCCTgCCCTGGGACCAGGTTGGTTATTgGGCCGTCAAGATTGTTTC	S40 137
500 140	Alle77	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--	Alle77
500 141	G50 72		G50 72
500 142	Alle78		Alle78
500 143	Med4	GGAGTTACAGGTTATTCACTTCCATGGGACCAGGTTGGATACTGGGCAGTTAAAATAGTTTCAGGTGTTCCTGCAGCAATACCTATCATTGGTGATTTTA	Med4
500 144	Med4	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--I--I--G--D--F--	Med4
500 145	SS2	GGCGTTACAGGCTATTCTTTGCCTTGGGATCAAGTCGGATATTGGGCTGTAAAGATTGTTTCAGGCGTTCCTGCTGCCATCCCAGTTGTAGGCGACTTCA	SS2
500 146	SS2	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--F--	SS2
500 147	FP5	GGAGTTACAGGATATTCATTGCCTTGGGATCAAGTTGGTTATTGGGCAGTTAAAATTGTCTCTGGAGTCCCAGCAGCTATCCCAGTTATTGGAGACTTTA	FP5
500 148	FP5	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F--	FP5
500 149	Pac7	GGAGTTACAGGATATTCTTTGCCTTGGGATCAAGTCGGATATTGGGCTGTTAAGATTGTCTCAGGTGTTCCTGCCGCAATTCCAGTAATAGGAGATTTTA	Pac7
500 150	Pac7	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--I--G--D--F--	Pac7
500 151	MIT9303		MIT9303
500 152	MIT9303		MIT9303
500 153	MIT9313		MIT9313
500 154	MIT9313		MIT9313
500 155	WH8103	GGTGTCACCGGTTACTCCCTGCCCTGGGACCAGGTTGGTTATTGGGCCGTCAAGATTGTTTCCGGCGTCCCAGCAGCCATCCCAGTTGTGGGTGACTTCA	WH8103
500 156	WH8103	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--F--	WH8103
500 157	PCC7002	GGTGTAACTGGTTACTCCTTGCCTTGGGACCAAGTTGGTTACTGGGCAGTCAAGATTGTGTCTGGTGTACCTGCGGCGATTCCTGTCGTGGGCGACCAAA	PCC7002
500 158	PCC7002	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--A--A--I--P--V--V--G--D--Q--	PCC7002
500 159	Pthrix	GGCGTGACCGGCTACTCCTTGCCCTGGGATCAAGTGGGTTACTGGGCCGTGAAAATTGTGTCCGGTGTCCCTGAGGCCATTCCCCTGGTTGGTCCCCTGA	Pthrix
500 160	Pthrix	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--E--A--I--P--L--V--G--P--L--	Pthrix
500 161	Nostoc	GGAGTTACCGGCTATTCCTTACCTTGGGACCAAGTTGGCTACTGGGCTGTGAAAATCGTTAGCGGTGTACCAGAAGCAATTCCCGTGGTTGGTGTTCTGA	Nostoc
500 162	Nostoc	-G--V--T--G--Y--S--L--P--W--D--Q--V--G--Y--W--A--V--K--I--V--S--G--V--P--E--A--I--P--V--V--G--V--L--	Nostoc
500 163	Chlorell	GGGGTAACGGGTTATTCTTTACCGTGGGATCAGATTGGGTATTGGGCTGTTAAAATTGTAACTGGTGTTCCAGATGCTATTCCAGTAATAGGGCAGGTGT	Chlorell
500 164	Chlorell	-G--V--T--G--Y--S--L--P--W--D--Q--I--G--Y--W--A--V--K--I--V--T--G--V--P--D--A--I--P--V--I--G--Q--V--	Chlorell
500 165	Liverwor	GGTGTTACAGGTTATTCTTTACCTTGGGATCAAATTGGTTATTGGGCAGTTAAAATTGTAACTGGTGTACCAGAAGCAATTCCAATAATTGGATCTCCTT	Liverwor
500 166	Liverwor	-G--V--T--G--Y--S--L--P--W--D--Q--I--G--Y--W--A--V--K--I--V--T--G--V--P--E--A--I--P--I--I--G--S--P--	Liverwor
500 167	Maize	GGTGTAACTGGTTATTCTTTGCCTTGGGATCAAATTGGTTATTGGGCAGTCAAAATTGTTACAGGTGTACCCGAAGCTATTCCGGTAATAGGGTCCCCTT	Maize
500 168	Maize	-G--V--T--G--Y--S--L--P--W--D--Q--I--G--Y--W--A--V--K--I--V--T--G--V--P--E--A--I--P--V--I--G--S--P--	Maize

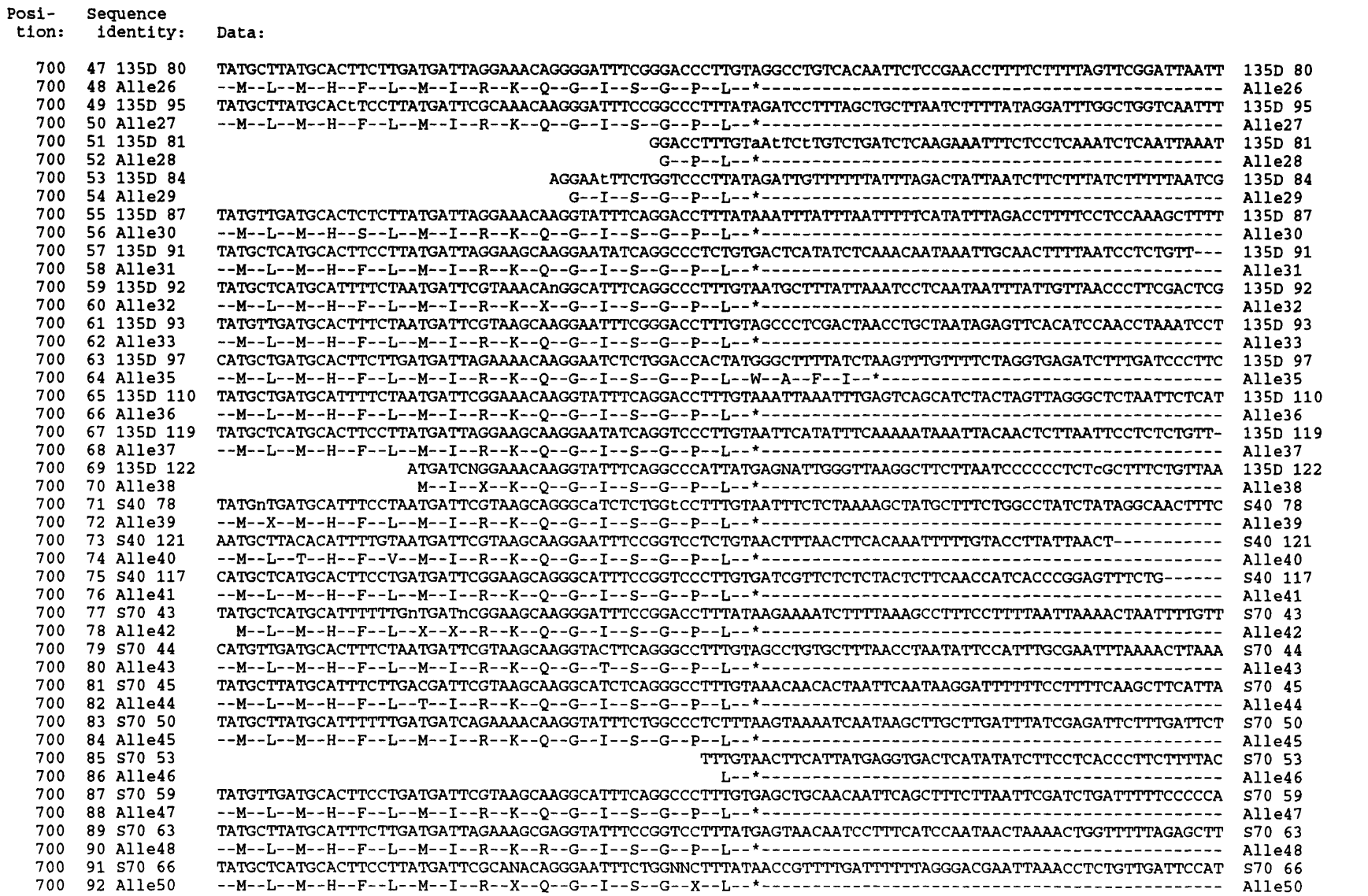

Posi-
tion:
Sequence
identity:
Data:
600 1 S40 96 TGGTTGAACTACTTAGAGGTGGAGAAAGTGTnGGTCAATCnACTTTAACCAGATTnTATAGTCTTCATACTTTTGTATTGCCCTGGTCATTAGCAGTCTT S40 96
600 2 Alle1 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--X--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle1
600 3 G50 90 CTGCtTAGAGGTGGAGAAAGTGTTGGTCAATCGACTCTTACAAGATTCTATAGTCTCCATACTTTCGTTCTCCCCTGGTCATTAGCAGTTTT G50 90
600 4 Alle2 L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle2
600 5 G50 107 ATAGTCTTCATnCTTTCGTATTGCCTTGGTCATTAGCGGTTTT G50 107
600 6 Alle3 S--L--H--X--F--V--L--P--W--S--L--A--V--F Alle3
600 7 S40 128 AAGTGTTGGTCAGTCCACCCTCACTTGCtTCTACAGCCTGCACACTTTTGTGATGCCCTGnCTGCTNGCNGTGTT S40 128
600 8 Alle4 S--V--G--Q--S--T--L--T--C--F--Y--S--L--H--T--F--V--M--P--X--L--L--A--V--F Alle4
600 9 85Br 13 TGGTTGAGTTGCTAAGGGGAGGGGAAAGTGTtgGCCAGCCAACCCTTATTCGTTTTTATAGCCTTCACACTTTCGTTCTCCTTTGGCTGTTGGCTGTATT 85Br 13
600 10 Alle5 M--V--E--L--L--R--G--G--E--S--V--G--Q--P--T--L--I--R--F--Y--S--L--H--T--F--V--L--L--W--L--L--A--V--F Alle5
600 11 85Br 100 TGGTCGAATTGCTTCGCgGTGGCnAGAGCGTCGGGCAGTCCACCCTGACTCGTTTTtATAGCCTTCACACCTTTGTGCTGCCCTGGCTACTTGCAGTGTT 85Br 100
600 12 Alle6 M--V--E--L--L--R--G--G--X--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle6
600 13 85Br 21 TGGTCGAATTGCTTAGAGGTGGTGAAAGTGTTGGTCAATCAACACTTAC CCtTTGTGTTGCCTTGGNNnnnTGCTGTATT 85Br 21
600 14 Alle7 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L- F--V--L--P--W--X--X--A--V--F Alle7
600 15 85Br 31 TGGTCGAACTCCTTAGGGGTGGTGAGAGCGATGGACAATCTACTTTGACTAGGTTCTACAGCCTTCATACTTTTGTAAtGCCATGGCTATTAGCAGTATT 85Br 31
600 16 Alle8 M--V--E--L--L--R--G--G--E--S--D--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--M--P--W--L--L--A--V--F Alle8
600 17 85Br 66 TtACACGCtTTTATAGTCTCCATACTTTTGTACTTCCTTGGTTGCTAGCAGTTTT 85Br 66
600 18 Alle9 L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle9
600 19 85Br 106 AAGAGGTGGAGAAAGTGTTGGACAGTCAACATTAACTCGCTTTTATAGCCTCCATACTTTTGTTCTCCCATGGACATTGGCTGTATT 85Br 106
600 20 Alle11 R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--T--L--A--V--F Alle11
600 21 85Br111S TGAnCTCCTTAGAGnnGGAGAGAGTGaTGGTCAATCCACACTGACTCGtTTTTATAGCCTTCATACTTTTGTTCTCCTTTGGATGCTTGCggtTTT 85Br111S
600 22 Alle12 X--L--L--R--X--G--E--S--D--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--L--W--M--L--A--V--F Alle12
600 23 85Br 118 85Br 118
600 24 Alle13 Alle13
600 25 85Br 119 TTCATACATTTGTCCTCCCTTGGCTGCTGGCGGTGTT 85Br 119
600 26 Alle14 H--T--F--V--L--P--W--L--L--A--V--F Alle14
600 27 85Br 122 ACAGCCtnCATACTTTTGTGTTGCCTTGGTTGTTAGCCGTGTT 85Br 122
600 28 Alle15 S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle15
600 29 85Br 127 GAAAGTGTTGGTCAAnCCGNCTTGACTCGCtTTTATAGCTTGCATACTTTTGTCTTGCCTTGGCTTCTGGNNNNATT 85Br 127
600 30 Alle16 E--S--V--G--Q--X--X--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--X--X--F Alle16
600 31 85Br 129 TTTGtTCtTCCTTgGACTCTCGCAGTGTT 85Br 129
600 32 Alle17 F--V--L--P--W--T--L--A--V--F Alle17
600 33 85Br 135 GGtTTT 85Br 135
600 34 Alle19 --V--F Alle19
600 35 85D 68 TTGTATTATTATTACGTGGTGGTGTTAGTGCTGGTCAAAGTACTTTAACGCGTTTCTACAGTGCACACACATTTGTATTGCCTGTTGCTGCAGCAGCATT 85D 68
600 36 Alle20 V--V--L--L--L--R--G--G--V--S--A--G--Q--S--T--L--T--R--F--Y--S--A--H--T--F--V--L--P--V--A--A--A--L Alle20
600 37 85D 90 TGGTTGAATTATTACGTGGTGGAGAGAGTGTGGGACAATCGACTCTTACaCGTTTCTATAGCCTTCACACTTTCGTTTTGCCCTGGACTCTTGCGGTTTT 85D 90
600 38 Alle21 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--T--L--A--V--F Alle21
600 39 S120 26 T ACTCTTnCACGtTTCTATAGTCTTCATACTTTTGTTTTGCCTTGGTTATTGGCTGTTTT S120 26
600 40 Alle22 -T--L--X--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle22
600 41 135Br 52 TGGTTGAACTTCTCAGAGGGGGCGAAAGTGTAGGTCAGACAACACTTACTCGCTTTTATAGCTTGCATACTTTTGTTCTTCCTTGGTCTCTCGCAGTTTT 135Br 52
600 42 Alle23 M--V--E--L--L--R--G--G--E--S--V--G--Q--T--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle23
600 43 135Br 53 TTTACGTgGGGGTGAAAGTGTAGGTCAAAGTACATTAACTCGTTTCTATAGTGCACACACCTTTGTTTTACCATTGGCTGCGGCTGTGTT 135Br 53
600 44 Alle24 L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--A--H--T--F--V--L--P--L--A--A--V--L Alle24
600 45 135Br 59 135Br 59
600 46 Alle25 Alle25

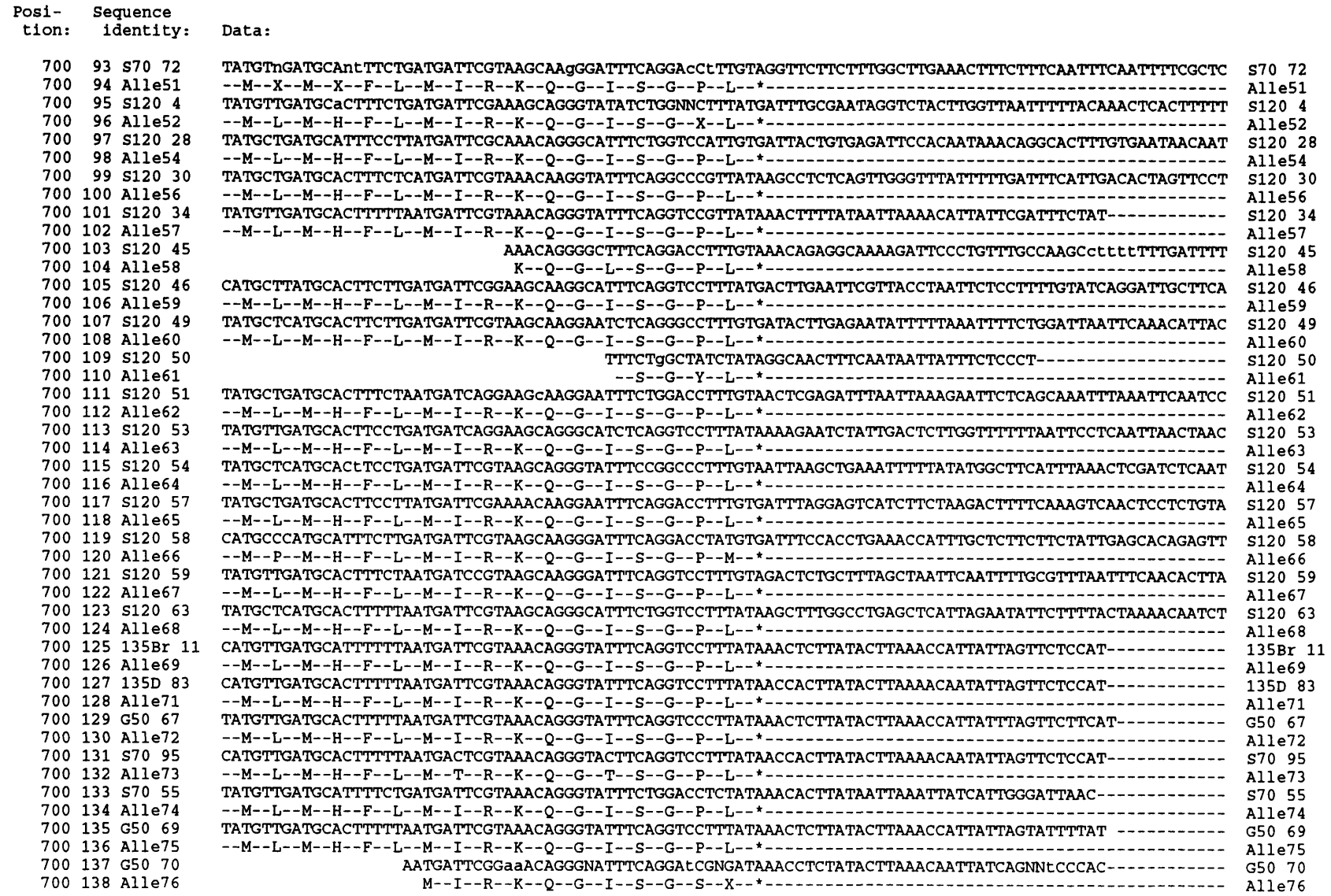

Posi-
tion:
Sequence
identity:
Data:
600 47 135D 80 GCTTACACGtTTnTATAGTCTTCACACtTTTGtTTTGCCCTGGTTGTTGGCCGTATT 135D 80
600 48 Alle26 L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle26
600 49 135D 95 CgCgGAGGAGAGAnGGTGGGACAGnCaACtTTGACnAGGtTTTATAGCTTGCATACtTTTGTGCTCCCTTGGATGTTGGCGGTCTT 135D 95
600 50 Alle27 R--G--G--E--X--V--G--Q--X--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--M--L--A--V--F Alle27
600 51 135D 81 135D 81
600 52 Alle28 Alle28
600 53 135D 84 135D 84
600 54 Alle29 Alle29
600 55 135D 87 TGAGAGGAGGAGAAAGTGTTGGACAGnCAACATTGACTCGCTTCTATAGCCTTCATACTTTTGTTTTGCCATGGCTATTAGCTGTTTT 135D 87
600 56 Alle30 R--G--G--E--S--V--G--Q--X--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle30
600 57 135D 91 TGGTTGAACTCCTAAGAGGTGGAGAAAGTGTAGGTCAGACAMnGCTTACTCGCTTTTATAGTTTGCATACCTTTGTTCTTCCTTGGACTCTTGCAATATT 135D 91
600 58 Alle31 M--V--E--L--L--R--G--G--E--S--V--G--Q--T--X--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--T--L--A--I--F Alle31
600 59 135D 92 GAGAGTGTTGGACAAAGCACACTAACCCGtTTTTATAGCCTCCATACCTTTGTTCTGCCATGGCTTTTAGCTGTCTT 135D 92
600 60 Alle32 -E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle32
600 61 135D 93 TGGCCAATCAACACTTACTCGTTTTTATAGTCTTCATACTTTTGTGCTCCCATGGATGCTCGCGGTATT 135D 93
600 62 Alle33 G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--M--L--A--V--F Alle33
600 63 135D 97 TTCTATAGTCTTCACACATTnGTTCTCCCTTGGCTGTTAGCAGTTTT 135D 97
600 64 Alle35 -F--Y--S--L--H--T--X--V--L--P--W--L--L--A--V--F Alle35
600 65 135D 110 AGCACATTAACTCGCTTTTATAGCCACCATACTTTTGTTCTCCCATGGACATTGGCTGTATT 135D 110
600 66 Alle36 -S--T--L--T--R--F--Y--S--H--H--T--F--V--L--P--W--T--L--A--V--F Alle36
600 67 135D 119 GCTTTTACAGCTTGCATACTTTTGTTCTTCCTTGGACTCTCGCTGTATT 135D 119
600 68 Alle37 F--Y--S--L--H--T--F--V--L--P--W--T--L--A--V--F Alle37
600 69 135D 122 135D 122
600 70 Alle38 Alle38
600 71 S40 78 TT S40 78
600 72 Alle39 -F Alle39
600 73 S40 121 TTACGTGGTGnNGaAAGTGTAgGACAATCTACACTAACACGtTTTTACAGTNNNnnnnCGTTTGTTTTACCAGTAGTTGCCGTTGTTTT S40 121
600 74 Alle40 L--R--G--X--E--S--V--G--Q--S--T--L--T--R--F--Y--S--X--X--X--F--V--L--P--V--V--A--V--V--L Alle40
600 75 S40 117 TTTTACAGCCTGCACACCTTTGTGATGCCATGGCTGCTGGCTGTCTT S40 117
600 76 Alle41 F--Y--S--L--H--T--F--V--M--P--W--L--L--A--V--F Alle41
600 77 S70 43 S70 43
600 78 Alle42 Alle42
600 79 S70 44 GTTTT S70 44
600 80 Alle43 V--F Alle43
600 81 S70 45 TGGTTGAGCTTCTTAGAGGTGGAGAAA AACTCTCACTCGCTTTTATAGCTTGCATACATTTGTTTTACCTTGGCTTTTAGCTGTATT S70 45
600 82 Alle44 M--V--E--L--L--R--G--G--E- --T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle44
600 83 S70 50 nGGtTGaGtTGtTACGAGGAGGAGAGAGTGTTGGTCAGTCAACTCTGACTcGtTTTTATAGCCTTCATACTTTTGTATTGCCTTGGCTATTAGCTGTATT S70 50
600 84 Alle45 X--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle45
600 85 S70 53 S70 53
600 86 Alle46 Alle46
600 87 S70 59 TCGTCGAACTTCTTAgGGGTGGCGAGAGTGTTGGGCAGTCAACGCtTACTCGtTTTTACAGCCTTCATACATTTGTTCTCCCTTGGCTGCTAGCGGTGTT S70 59
600 88 Alle47 I--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle47
600 89 S70 63 TTGAGTTATTGCGTgGnGGAGAGAGTGTTcGACAATCAACACTTACACGCTTTTACAGTCTTCACACTTTTGTTTTGCCTTGGTTATTAGCTGTATT S70 63
600 90 Alle48 E--L--L--R--G--G--E--S--V--R--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle48
600 91 S70 66 CGTTTGTTCTCCCTTGGATGCTGCnnNNGTT S70 66
600 92 Alle50 F--V--L--P--W--M--L--X--X--F Alle50


Posi-
tion:
Sequence
identity:
Data:
600 93 S70 72 TgGATGCtTGCAGtnnT S70 72
600 94 Alle51 W--M--L--A--V--X Alle51
600 95 S120 4 CATACCTTTGTGTTGCCtTGGtTGTTAGCNgNGTT S120 4
600 96 Alle52 H--T--F--V--L--P--W--L--L--A--X--F Alle52
600 97 S120 28 TGTACTCCCATGGCTTTTGGCAGTTTT S120 28
600 98 Alle54 V--L--P--W--L--L--A--V--F Alle54
600 99 S120 30 CTTGACTCGttTTTACAGCCNNCATaCTTTTGTTCtCCCTTgGACATTGGCGGTATT S120 30
600 100 Alle56 L--T--R--F--Y--S--X--H--T--F--V--L--P--W--T--L--A--V--F Alle56
600 101 S120 34 ACTCTTACCAGATTCTACAGCCTTCATACTTTCGTACTACCTTGGTCATTAGCGGTTTT S120 34
600 102 Alle57 T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle57
600 103 S120 45 S120 45
600 104 Alle58 Alle58
600 105 S120 46 TGGTTGAGTTGTTG GTCGTTCCTTGGTTGCTTGCAGTCTT S120 46
600 106 Alle59 M--V--E--L--L- -V--V--P--W--L--L--A--V--F Alle59
600 107 S120 49 tTTTACAGCCTTCATACtTTGTTTTGCCTTGGCTACTTGCAGtTTT S120 49
600 108 Alle60 -F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle60
600 109 S120 50 S120 50
600 110 Alle61 Alle61
600 111 S120 51 CCTGGATGCTTGCAGTTTT S120 51
600 112 Alle62 W--M--L--A--V--F Alle62
600 113 S120 53 GtTTTACAGCCTTCATACAtTTGTGCTGCCTTGGCTACTTGCTGTGTT S120 53
600 114 Alle63 F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle63
600 115 S120 54 TTTTATAGTCTTCATACTTTTGTGCTCCCTTGGCTGCTTGCAGTCTT S120 54
600 116 Alle64 F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F Alle64
600 117 S120 57 ACAtTGACTcGCTTTTATAGnTTGCATnCGTTTGTGCTGcCCTGGGCTCTTGCGGTCTT S120 57
600 118 Alle65 T--L--T--R--F--Y--X--L--H--X--F--V--L--P--W--A--L--A--V--F Alle65
600 119 S120 58 TGGTTGAGCTTCTTCGAGG TATAGTCTTCATACCTTnGTCTTACCTTGGCTTCTTGCTGTATT S120 58
600 120 Alle66 M--V--E--L--L--R- Y--S--L--H--T--X--V--L--P--W--L--L--A--V--F Alle66
600 121 S120 59 CGTTGGGCAATCAACACTGACTCGTTTTTATAGCCTTCATACGTTTGTGCTCCCTTGGATGCTTGCGGGTTT S120 59
600 122 Alle67 V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--M--L--A--G--F Alle67
600 123 S120 63 CGtTTTTATAGTCTTCATACATTnGTTCTGCCTTGGCTCTTGGCGGTATT S120 63
600 124 Alle68 -R--F--Y--S--L--H--T--X--V--L--P--W--L--L--A--V--F Alle68
600 125 135Br 11 TGGtTGAACTnCTAAGAGGTGGAGAAAGcGTTGGGCAATCTACTTTAACCAGATTTTACAGTCTTCATACTTTTGTATTGCCCTGGTCATTGGCAGTTTT 135Br 11
600 126 Alle69 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle69
600 127 135D 83 AAGTGTgGGnCAATCTACTTTAACCAGATTTTACAGTCTTCACACTTTTGTATTGCCATGGTCATTAGCGGTCTT 135D 83
600 128 Alle71 S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle71
600 129 G50 67 GAGGTGGAGAAAnCGTTGGACAATCGACCtTAACCAGATTTTACAGTCTTCATACTTTTGTATTGCCATGGTCATTAGCGGTTTT G50 67
600 130 Alle72 G--G--E--X--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle72
600 131 S70 95 TGGTTGAACTACTCAGAGGTGGTGGAAGTGTTGGGCAATCTACTTTAACCAGATTTTACAGTCTTCACACTTTTGTATTGCCATGGTCATTAGCGGTCTT S70 95
600 132 Alle73 M--V--E--L--L--R--G--G--G--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle73
600 133 S70 55 TTTTGTATTGCCATAGTCATTAGCTGTATT S70 55
600 134 Alle74 F--V--L--P--* S--L--A--V--F Alle74
600 135 G50 69 CTTTAACCAGATTTTACAGTCTTCATACTTTTGTATTACCATGGTCACTAGCGGTATT G50 69
600 136 Alle75 L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Alle75
600 137 G50 70 TGGTnGAACT G50 70
600 138 Alle76 M--V--E- Alle76


Posi-
tion:
Sequence
identity:
Data:
600 139 S40 137 CTTCTACAGCCTCCAcACCTTTGTGATGCNATGnCNGCTNGCNGTATT S40 137
600 140 Alle77 F--Y--S--L--H--T--F--V--M--X--X--X--L--A--V--F Alle77
600 141 G50 72 tTGTgATGnNCTGncTGctNggNGTATT G50 72
600 142 Alle78 V--M--X--X--L--L--G--V--F Alle78
600 143 Med4 TGGTTGAACTTCTTCGAGGAGGAGAAAGTGTTGGTCAATCTACTTTGACTAGATTTTATAGCCTCCATACTTTTGTTCTACCTTGGTCTTTAGCAGTATT Med4
600 144 Med4 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Med4
600 145 SS2 TGGTTGAATTGCTTCGAGGAGGTGAAAGTGTTGGCCAGACAaCACTTACCCGTTTTTACAGTCTTCATACTTTTGTATTGCCTTGGACATTGGCAATCTT SS2
600 146 SS2 M--V--E--L--L--R--G--G--E--S--V--G--Q--T--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--T--L--A--I--F SS2
600 147 FP5 TGGTCGAACTTCTTAGAGGTGGTGAGAGCGTTGGACAGTCAACTTTGACAAGGTTTTATAGTCTTCACACTTTTGTGATGCCATGGCTGCTAGCAGTTTT FP5
600 148 FP5 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--M--P--W--L--L--A--V--F FP5
600 149 Pac7 TGGTTGAACTACTGAGGGGAGGAGAAAGTGTTGGACAGTCTACTCTTACAAGATTTTATAGTCTTCATACTTTTGTTTTGCCATGGTCATTAGCAGTATT Pac7
600 150 Pac7 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--S--L--A--V--F Pac7
600 151 MIT9303 TTTTTATAGCCTTCACACCTTTGTGCTGCCCTGGCTACTTGCAGTGTT MIT9303
600 152 MIT9303 F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F MIT9303
600 153 MIT9313 GTGnCGAGAGCGTCGGGCAGTCCACCCTGACTCGTTTTTATAGCCTTCACACCTTTGTGCTGCCCTGGCTACTTGCAGTGTT MIT9313
600 154 MIT9313 -X--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--L--A--V--F MIT9313
600 155 WH8103 TGGTGGAGCTGCTCCGCGGTGGCGAAAGTGTCGGTCAGTCCACACTCACTCGCTTCTACAGCCTCCACACCTTTGTGATGCCATGGCTGCTCGCCGTATT WH8103
600 156 WH8103 M--V--E--L--L--R--G--G--E--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--M--P--W--L--L--A--V--F WH8103
600 157 PCC7002 TGGTTGAGTTGCTCCGGGGTGGCGCAAGCGTTGGCCAAGCAACCCTAACCCGTTTCTACAGTCTTCATACTTTTGTTTTACCTTGGTTGATTGCGGTCTT PCC7002
600 158 PCC7002 M--V--E--L--L--R--G--G--A--S--V--G--Q--A--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--L--I--A--V--F PCC7002
600 159 Pthrix TGGTGGAACTGATTCGCGGTAGTGCCAGTGTGGGTCAAGCGACCCTGACCCGCTTCTATAGCCTGCACACCTTTGTGTTGCCCTGGTTCATTGCGGTGTT Pthrix
600 160 Pthrix M--V--E--L--I--R--G--S--A--S--V--G--Q--A--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--W--F--I--A--V--F Pthrix
600 161 Nostoc TCTCCGACCTACTGCGTGGTGGTTCTAGTGTTGGCCAAGCAACACTGACTCGTTACTACAGCGCACACACCTTCGTTTTGCCTTGGTTAATAGCAGTCTT Nostoc
600 162 Nostoc I--S--D--L--L--R--G--G--S--S--V--G--Q--A--T--L--T--R--Y--Y--S--A--H--T--F--V--L--P--W--L--I--A--V--F Nostoc
600 163 Chlorell TGTTGGAGCTTTTACGTGGGGGTGTTGCTGTTGGTCAATCCACATTAACACGTTTTTACAGTTTACATACTTTTGTATTACCTCTTTTTACAGCCGTGTT Chlorell
600 164 Chlorell L--L--E--L--L--R--G--G--V--A--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--L--F--T--A--V--F Chlorell
600 165 Liverwor TAGTTGAGTTATTACGCGGAAGTGTAAGTGTTGGTCAATCGACATTAACTCGATTTTATAGTTTACATACTTTTGTATTGCCTCTTTTAACTGCAATATT Liverwor
600 166 Liverwor L--V--E--L--L--R--G--S--V--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--L--L--T--A--I--F Liverwor
600 167 Maize TAGTGGAATTATTACGTGGAAGTGCTAGTGTGGGCCAATCCACTTTGACTCGTTTTTATAGTTTACATACCTTTGTACTACCTCTGCTTACTGCCGTATT Maize
600 168 Maize L--V--E--L--L--R--G--S--A--S--V--G--Q--S--T--L--T--R--F--Y--S--L--H--T--F--V--L--P--L--L--T--A--V--F Maize


Posi-
tion:
Sequence
identity:
Data:
700 1 S40 96 CAnGTTGATGcACTTTTTAATGATTCGTAAACAGGGTATTTCAGGCCCTTTATAAACTCTTATACTTAAACCAATACTAGTTCTCCAT---------- S40 96
700 2 Alle1 --X--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle1
700 3 G50 90 CATGTTGATGCACTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCTCTATAAACTCCTATACTTTAATTATTAATTAGATTCAA---------- G50 90
700 4 Alle2 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle2
700 5 G50 107 CATGTTGATGCaCTTTTTAATGATTCGTAAACAAGGTATTTCAGGtCCCTTGTAAACTTTTATACTTAAATCAATACAAGACTCTCTC---------- G50 107
700 6 Alle3 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle3
700 7 S40 128 CATGCTCATGCACTTCCTGATGATCCGGAAGCAGGGCATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGGACTCACCG------ S40 128
700 8 Alle4 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle4
700 9 85Br 13 TATGCTCATGCATTTTCTTATGATTCGTAAGCAGGGAATTTCTGGCCCCTTGTGATTTTCTTGATCGTTTTCTTTAGCTTGTCCCACTAAACTTAATTTC 85Br 13
700 10 Alle5 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*----------Alle5
700 11 85Br 100 CATGCTCATGCATTTCCTGATGATCAGGAAGCAGGGCATTTCAGGTCCGTTGTAAGCGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTT 85Br 100
700 12 Alle6 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle6
700 13 85Br 21 TATGCTCATGCACTTCCTGATGATaCGAAAACAAGGAATATCAGGACCTTTGTGATTGTTGTCACCTTTCTATAAGTTGTTTGAGCAACTTCTTTATTTA 85Br 21
700 14 Alle7 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle7
700 15 85Br 31 TATGTTGATGCATTTTCTAATGATTAGAAAACAAGGCATCTCTGGTCCCTTGTAGTTGTATTTATTTTCTAAAAACTAATTTTAAATTACGTTAAAAATT 85Br 31
700 16 Alle8 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle8
700 17 85Br 66 CATGCTTATGCATTTTCTGATGATCGGGAAGCAAGGTATTTCAGGCCCTTTATGATTTAAAACTTTTATTCCCAGCCTTTTTCTTTAAAAGAGGTTTTTC 85Br 66
700 18 Alle9 --M--L--M--H--F--L--M--I--G--K--Q--G--I--S--G--P--L--*---------- Alle9
700 19 85Br 106 TATGCTGATGCATTTTCTAATGATTCGGAAACAAGGTATTTCAGGACCTTTGTAAATTAAATTTGAGTCAGCATCTACTAGTTAGGGCTCTAATTCTCAT 85Br 106
700 20 Alle11 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle11
700 21 85Br111S TATGTTGATGCACTTTCTAATGATTCGTAAGCAAGGTATCTCAGGGCCTTTGTAGCCTGTGCTTTAACCTGAAAAATTCAATTTTCGAATTTGACTTCAA 85Br111S
700 22 Alle12 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle12
700 23 85Br 118 ATTTTTTGATGAGTCGGAAGcAAGGGATTTCCGGACCTTTATAAGAAAAATCTTTTAAAGCCTTTCCTTTTaATTAAAACTAATTTTGT 85Br 118
700 24 Alle13 F--L--M--S--R--K--Q--G--I--S--G--P--L--*---------- Alle13
700 25 85Br 119 TATGTTGATGCACTTCCTGATGATTNGTAAGCAAGGgATTTCAGGCCCTTTGTGAGCTGCCACAATTCAGCTTTCTTAATTCGATCTGATCTTTCCCCCA 85Br 119
700 26 Alle14 --M--L--M--H--F--L--M--I--X--K--Q--G--I--S--G--P--L--*---------- Alle14
700 27 85Br 122 TATGTTGATGcACTTTCTGATGATTCGAAAGCAGgGTATATCTGGcCCCTTATGATTTCCGAATAGTTCTACTTGGTTAATTTTTACAAACTCACTTTTT 85Br 122
700 28 Alle15 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle15
700 29 85Br 127 TATGTTGATGCACTTTCTTATGATTcGTAAGCAGGGTATTTCCGGACCCTTATGAGCTTTAACTGTTCTTTTAAATTAAGTTGATCTGATTATTTCAATA 85Br 127
700 30 Alle16 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle16
700 31 85Br 129 TATGCTCATGCACTTCCTTATGATTAGGAAGCAAGGAATATCAGGACCCCTGTAATTCATATCTCCAACAAAAAATTGCAACTTTTAATCCTCTGTT--- 85Br 129
700 32 Alle17 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle17
700 33 85Br 135 TATGTTGATGcACTTCCTTATGATTCGCAGACAGGGCATATCAGGACCcTTATGAATTCGAAAAGGTTTTTTGCATCACAATCTTAATCATTTCTTGATG 85Br 135
700 34 Alle19 --M--L--M--H--F--L--M--I--R--R--Q--G--I--S--G--P--L--*---------- Alle19
700 35 85D 68 AATGTTAACACATTTTTTAATGATTAGAAAGCAAGGAATTTCCGGTCCACTATAAGATTAAATAATAAAAAATT---------- 85D 68
700 36 Alle20 --M--L--T--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle20
700 37 85D 90 TATGCTTATGCATTTTCTTATGATTCGCAAACAGGGAACTTCTGGTCCATTATAAAAATAAAAAAAAATCATTTTTCCTTATCTCATTAAATAA------ 85D 90
700 38 Alle21 --M--L--M--H--F--L--M--I--R--K--Q--G--T--S--G--P--L--*---------- Alle21
700 39 S120 26 CATGCTTATGCATTTCCTGATGATTCGTAAACAAGGGATCTCAGGTCCATTGTGATATGAATTAATCAAGACAAGtTTAGAAATTTTTTCATAGGTTTTT S120 26
700 40 Alle22 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle22
700 41 135Br 52 CATGCTTATGCACTTCCTCATGATTAGAAAGCAAGGAATATCAGGCCCCTTGTAATTCATATTTTCAAAAATGTATTACAACTTTTAATCCCCTATT--- 135Br 52
700 42 Alle23 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle23
700 43 135Br 53 AATGTTAACACACTTCTTAATGATTAGAAAGCAAGGTATCTCAGGTCCTCTTTAATTTAAAAGTTAAGAAATTAGACTCAATATTAATATCTGAATAAGG 135Br 53
700 44 Alle24 --M--L--T--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle24
700 45 135Br 59 TTTTTTAATgaTTCGGaaaCaAgGCaTTTCTGGcccTTTGTAAGTASAATCACATTGCCTTGAATCATTCACtttTCtGTTGAgNNNN 135Br 59
700 46 Alle25 F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle25


Posi-
tion:
Sequence
identity:
Data:
700 47 135D 80 TATGCTTATGCACTTCTTGATGATTAGGAAACAGGGGATTTCGGGACCCTTGTAGGCCTGTCACAATTCTCCGAACCTTTTCTTTTAGTTCGGATTAATT 135D 80
700 48 Alle26 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle26
700 49 135D 95 TATGCTTATGCACttCCTTATGATTCGCAAACAAGGGATTTCCGGCCCTTTATAGATCCTTTAGCTGCTTAATCTTTTATAGGATTTGGCTGGTCAATTT 135D 95
700 50 Alle27 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle27
700 51 135D 81 GGACCTTTGTaAtTCtTGTCTGATCTCAAGAAATTTCTCCTCAAATCTCAATTAAAT 135D 81
700 52 Alle28 G--P--L--* Alle28
700 53 135D 84 AGGAAtTTCTGGTCCCTTATAGATTGTTTTTTATTTAGACTATTAATCTTCTTTATCTTTTTAATCG 135D 84
700 54 Alle29 G--I--S--G--P--L--* Alle29
700 55 135D 87 TATGTTGATGCACTCTCTTATGATTAGGAAACAAGGTATTTCAGGACCTTTATAAATTTATTTAATTTTTCATATTTAGACCTTTTCCTCCAAAGCTTTT 135D 87
700 56 Alle30 --M--L--M--H--S--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle30
700 57 135D 91 TATGCTCATGCACTTCCTTATGATTAGGAAGCAAGGAATATCAGGCCCTCTGTGACTCATATCTCAAACAATAAATTGCAACTTTTAATCCTCTGTT--- 135D 91
700 58 Alle31 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle31
700 59 135D 92 TATGCTCATGCATTTTCTAATGATTCGTAAACAnGGCATTTCAGGCCCTTTGTAATGCTTTATTAAATCCTCAATAATTTATTGTTAACCCTTCGACTCG 135D 92
700 60 Alle32 --M--L--M--H--F--L--M--I--R--K--X--G--I--S--G--P--L--* Alle32
700 61 135D 93 TATGTTGATGCACTTTCTAATGATTCGTAAGCAAGGAATTTCGGGACCTTTGTAGCCCTCGACTAACCTGCTAATAGAGTTCACATCCAACCTAAATCCT 135D 93
700 62 Alle33 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle33
700 63 135D 97 CATGCTGATGCACTTCTTGATGATTAGAAAACAAGGAATCTCTGGACCACTATGGGCTTTTATCTAAGTTTGTTTTTCTAGGTGAGATCTTTGATCCCTTC 135D 97
700 64 Alle35 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--W--A--F--I--* Alle35
700 65 135D 110 TATGCTGATGCATTTTCTAATGATTCGGAAACAAGGTATTTCAGGACCTTTGTAAATTAAATTTGAGTCAGCATCTACTAGTTAGGGCTCTAATTCTCAT 135D 110
700 66 Alle36 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle36
700 67 135D 119 TATGCTCATGCACTTCCTTATGATTAGGAAGCAAGGAATATCAGGTCCCTTGTAATTCATATTTCAAAAATAAATTACAACTCTTAATTCCTCTCTGTT- 135D 119
700 68 Alle37 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle37
700 69 135D 122 ATGATCNGGAAACAAGGTATTTCAGGCCCATTATGAGNATTGGGTTAAGGCTTCTTAATCCCCCCTCTcGCTTTCTGTTAA 135D 122
700 70 Alle38 M--I--X--K--Q--G--I--S--G--P--L--* Alle38
700 71 S40 78 TATGnTGATGCATTTCCTAATGATTCGTAAGCAGGGCaTCTCTGGtCCTTTGTAATTTCTCTAAAAGCTATGCTTTCTGGCCTATCTATAGGCAACTTTC S40 78
700 72 Alle39 --M--X--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle39
700 73 S40 121 AATGCTTACACATTTTGTAATGATTCGTAAGCAAGGAATTTCCGGTCCTCTGTAACTTTAACTTCACAAATTTTTGTACCTTATTAACT---------- S40 121
700 74 Alle40 --M--L--T--H--F--V--M--I--R--K--Q--G--I--S--G--P--L--* Alle40
700 75 S40 117 CATGCTCATGCACTTCCTGATGATTCGGAAGCAGGGCATTTCCGGTCCCTTGTGATCGTTCTCTCTACTCTTCAACCATCACCCGGAGTTTCTG------ S40 117
700 76 Alle41 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle41
700 77 S70 43 TATGCTCATGCATTTTTTGnTGATnCGGAAGCAAGGGATTTCCGGACCTTTATAAGAAAATCTTTTAAAGCCTTTCCTTTTAATTAAAACTAATTTTGTT S70 43
700 78 Alle42 M--L--M--H--F--L--X--X--R--K--Q--G--I--S--G--P--L--* Alle42
700 79 S70 44 CATGTTGATGCACTTTCTAATGATTCGTAAGCAAGGTACTTCAGGGCCTTTGTAGCCTGTGCTTTAACCTAATATTCCATTTGCGAATTTAAAACTTAAA S70 44
700 80 Alle43 --M--L--M--H--F--L--M--I--R--K--Q--G--T--S--G--P--L--* Alle43
700 81 S70 45 TATGCTTATGCATTTCTTGACGATTCGTAAGCAAGGCATCTCAGGGCCTTTGTAAACAACACTAATTCAATAAGGATTTTTTCCTTTTCAAGCTTCATTA S70 45
700 82 Alle44 --M--L--M--H--F--L--T--I--R--K--Q--G--I--S--G--P--L--* Alle44
700 83 S70 50 TATGCTTATGCATTTTTTGATGATCAGAAAACAAGGTATTTCTGGCCCTCTTTAAGTAAAATCAATAAGCTTGCTTGATTTATCGAGATTCTTTGATTCT S70 50
700 84 Alle45 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle45
700 85 S70 53 TTTGTAACTTCATTATGAGGTGACTCATATATCTTCCTCACCCTTCTTTTAC S70 53
700 86 Alle46 L--* Alle46
700 87 S70 59 TATGTTGATGCACTTCCTGATGATTCGTAAGCAAGGCATTTCAGGCCCTTTGTGAGCTGCAACAATTCAGCTTTCTTAATTCGATCTGATTTTTCCCCCA S70 59
700 88 Alle47 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--* Alle47
700 89 S70 63 TATGCTTATGCATTTCTTGATGATTAGAAAGCGAGGTATTTCCGGTCCTTTATGAGTAACAATCCTTTCATCCAATAACTAAAACTGGTTTTTAGAGCTT S70 63
700 90 Alle48 --M--L--M--H--F--L--M--I--R--K--R--G--I--S--G--P--L--* Alle48
700 91 S70 66 TATGCTCATGCACTTCCTTATGATTCGCANACAGGGAATTTCTGGNNCTTTATAACCGTTTTGATTTTTTAGGGACGAATTAAACCTCTGTTGATTCCAT S70 66
700 92 Alle50 --M--L--M--H--F--L--M--I--R--X--Q--G--I--S--G--X--L--* Alle50

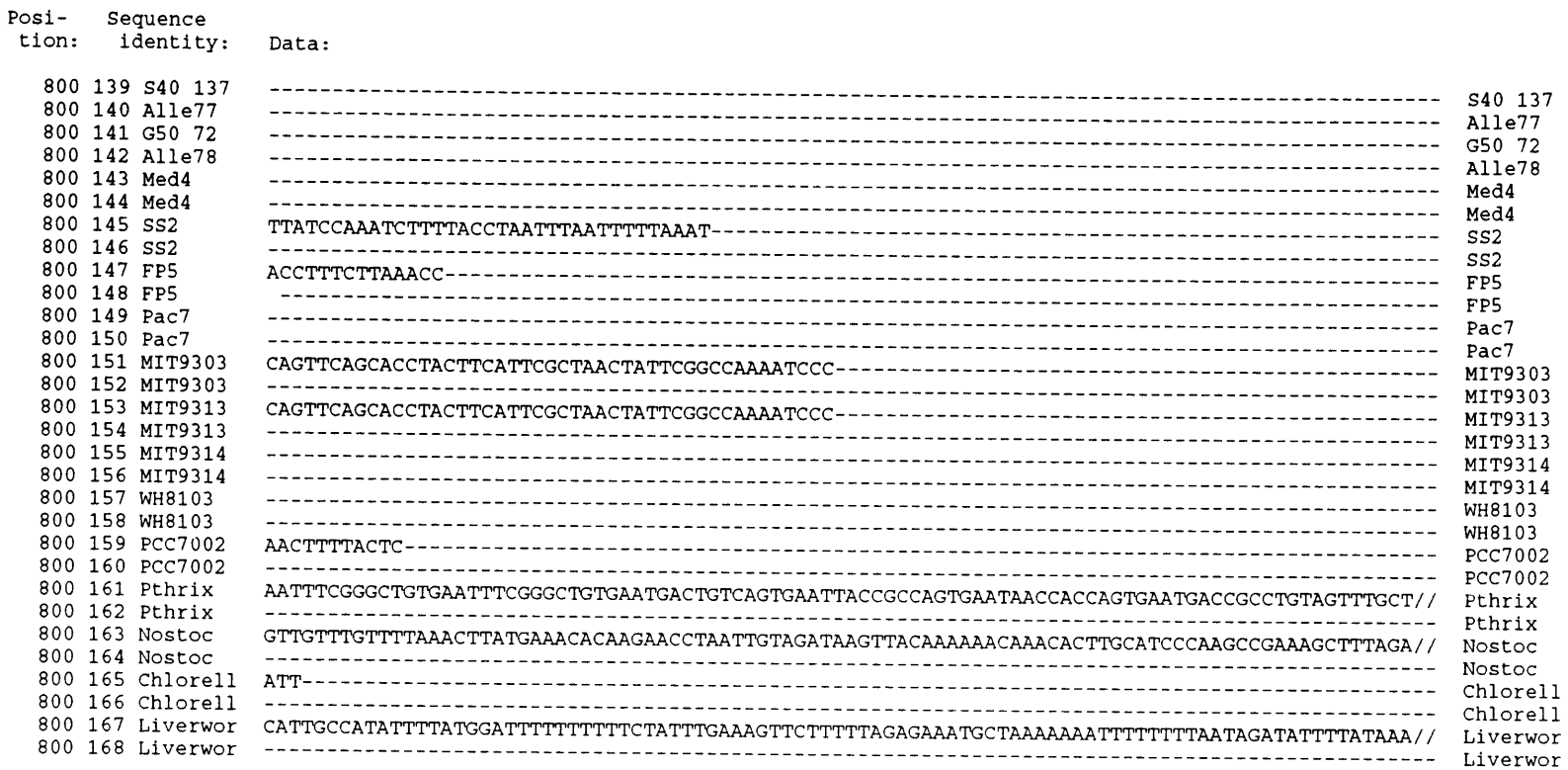

Posi-
tion:
Sequence
identity:
Data:
700 93 S70 72 TATGTnGATGCAntTTCTGATGATTCGTAAGCAAgGGATTTCAGGAcCtTTGTAGGTTCTTCTTTGGCTTGAAACTTTCTTTCAATTTCAATTTTCGCTC S70 72
700 94 Alle51 --M--X--M--X--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle51
700 95 S120 4 TATGTTGATGCaCTTTCTGATGATTCGAAAGCAGGGTATATCTGGNNCTTTATGATTTGCGAATAGGTCTACTTGGTTAATTTTTACAAACTCACTTTTT S120 4
700 96 Alle52 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--X--L--*---------- Alle52
700 97 S120 28 TATGCTGATGCATTTCCTTATGATTCGCAAACAGGGCATTTCTGGTCCATTGTGATTACTGTGAGATTCCACAATAAACAGGCACTTTGTGAATAACAAT S120 28
700 98 Alle54 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle54
700 99 S120 30 TATGCTGATGCACTTTCTCATGATTCGTAAACAAGGTATTTCAGGCCCGTTATAAGCCTCTCAGTTGGGTTTATTTTTGATTTCATTGACACTAGTTCCT S120 30
700 100 Alle56 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle56
700 101 S120 34 TATGTTGATGCACTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCGTTATAAACTTTTATAATTAAAACATTATTCGATTTCTAT---------- S120 34
700 102 Alle57 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle57
700 103 S120 45 AAACAGGGGCTTTCAGGACCTTTGTAAACAGAGGCAAAAGATTCCCTGTTTGCCAAGCcttttTTTGATTTT S120 45
700 104 Alle58 K--Q--G--L--S--G--P--L--*---------- Alle58
700 105 S120 46 CATGCTTATGCACTTCTTGATGATTCGGAAGCAAGGCATTTCAGGTCCTTTATGACTTGAATTCGTTACCTAATTCTCCTTTTGTATCAGGATTGCTTCA S120 46
700 106 Alle59 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle59
700 107 S120 49 TATGCTCATGCACTTCTTGATGATTCGTAAGCAAGGAATCTCAGGGCCTTTGTGATACTTGAGAATATTTTTAAATTTTCTGGATTAATTCAAACATTAC S120 49
700 108 Alle60 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle60
700 109 S120 50 TTTCTgGCTATCTATAGGCAACTTTCAATAATTATTTCTCCCT---------- S120 50
700 110 Alle61 --S--G--Y--L--*---------- Alle61
700 111 S120 51 TATGCTGATGCACTTTCTAATGATCAGGAAGcAAGGAATTTCTGGACCTTTGTAACTCGAGATTTAATTAAAGAATTCTCAGCAAATTTAAATTCAATCC S120 51
700 112 Alle62 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle62
700 113 S120 53 TATGTTGATGCACTTCCTGATGATCAGGAAGCAGGGCATCTCAGGTCCTTTATAAAAGAATCTATTGACTCTTGGTTTTTTAATTCCTCAATTAACTAAC S120 53
700 114 Alle63 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle63
700 115 S120 54 TATGCTCATGCACtTCCTGATGATTCGTAAGCAGGGTATTTCCGGCCCTTTGTAATTAAGCTGAAATTTTTATATGGCTTCATTTAAACTCGATCTCAAT S120 54
700 116 Alle64 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle64
700 117 S120 57 TATGCTGATGCACTTCCTTATGATTCGAAAACAAGGAATTTCAGGACCTTTGTGATTTAGGAGTCATCTTCTAAGACTTTTCAAAGTCAACTCCTCTGTA S120 57
700 118 Alle65 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle65
700 119 S120 58 CATGCCCATGCATTTCTTGATGATTCGTAAGCAAGGGATTTCAGGACCTATGTGATTTCCACCTGAAACCATTTGCTCTTCTTCTATTGAGCACAGAGTT S120 58
700 120 Alle66 --M--P--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--M--*---------- Alle66
700 121 S120 59 TATGTTGATGCACTTTCTAATGATCCGTAAGCAAGGGATTTCAGGTCCTTTGTAGACTCTGCTTTAGCTAATTCAATTTTGCGTTTAATTTCAACACTTA S120 59
700 122 Alle67 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle67
700 123 S120 63 TATGCTCATGCACTTTTTAATGATTCGTAAGCAGGGCATTTCTGGTCCTTTATAAGCTTTGGCCTGAGCTCATTAGAATATTCTTTTACTAAAACAATCT S120 63
700 124 Alle68 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle68
700 125 135Br 11 CATGTTGATGCATTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCTTTATAAACTCTTATACTTAAACCATTATTAGTTCTCCAT---------- 135Br 11
700 126 Alle69 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle69
700 127 135D 83 CATGTTGATGCACTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCTTTATAACCACTTATACTTAAAACAATATTAGTTCTCCAT---------- 135D 83
700 128 Alle71 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle71
700 129 G50 67 TATGTTGATGCACTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCCTTATAAACTCTTATACTTAAACCATTATTTAGTTCTTCAT---------- G50 67
700 130 Alle72 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle72
700 131 S70 95 CATGTTGATGCACTTTTTAATGACTCGTAAACAGGGTACTTCAGGTCCTTTATAACCACTTATACTTAAAACAATATTAGTTCTCCAT---------- S70 95
700 132 Alle73 --M--L--M--H--F--L--M--T--R--K--Q--G--T--S--G--P--L--*---------- Alle73
700 133 S70 55 TATGTTGATGCATTTTCTGATGATTCGTAAACAGGGTATTTCTGGACCTCTATAAACACTTATAATTAAATTATCATTGGGATTAAC---------- S70 55
700 134 Alle74 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle74
700 135 G50 69 TATGTTGATGCACTTTTTAATGATTCGTAAACAGGGTATTTCAGGTCCTTTATAAACTCTTATACTTAAACCATTATTAGTATTTTAT ---------- G50 69
700 136 Alle75 --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*---------- Alle75
700 137 G50 70 AATGATTCGGaaACAGGGNATTTCAGGAtCGNGATAAACCTCTATACTTAAACAATTATCAGNNtCCCAC---------- G50 70
700 138 Alle76 M--I--R--K--Q--G--I--S--G--S--X--*---------- Alle76

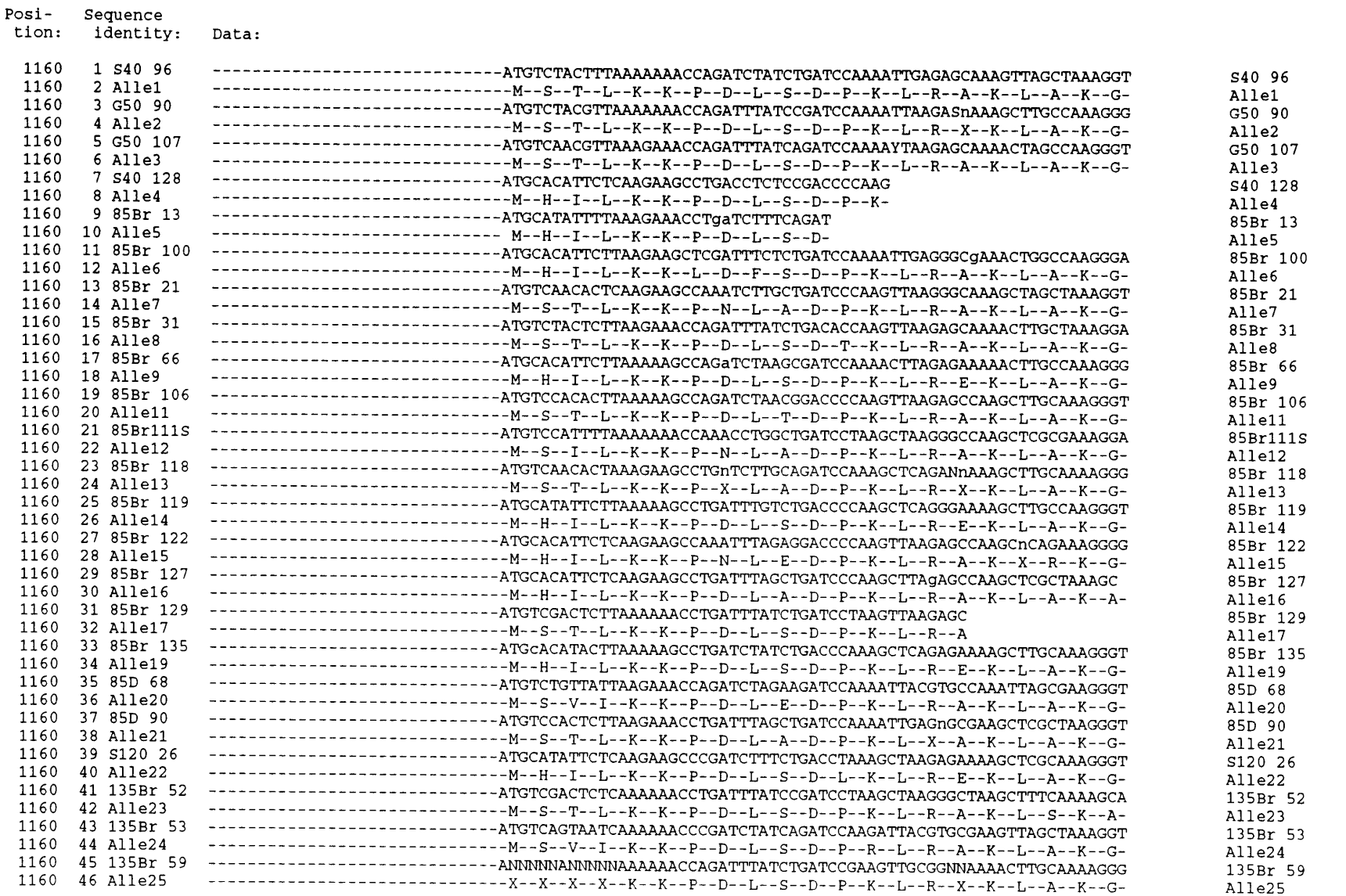
```
Posi-   Sequence
 tion:   identity:   Data:

   700 139 S40 137   CATGCTCATGcACTTCCTGATGATTCGGAAGNAGGGNATTTCTGGTCCCTTGTGATTTGCGTTCCTACNGTTAACTAAACACCTGGATTTACCG------  S40 137
   700 140 Alle77    --M--L--M--H--F--L--M--I--R--K--X--G--I--S--G--P--L--*-----------------------------------------  Alle77
   700 141 G50 72    TATGCTCATGcACTTCCTGATGATcNGGAAGCAGGGNATTTCTGGTCCCTTGTGATCCGcGTTCCTACCGTTAACCCAACACCTGGATTCACCG------  G50 72
   700 142 Alle78    --M--L--M--H--F--L--M--I--X--K--Q--G--I--S--G--P--L--*-----------------------------------------  Alle78
   700 143 Med4      TATGCTGATGCATTTTCTAATGATTCGTAAACAAGGTATTTCAGGTCCGTTATAAACCCTTATAATTTAAAAAACTAAACATTCTGAATTATTTTAT---  Med4
   700 144 Med4      --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Med4
   700 145 SS2       TATGCTCATGCACTTTCTTATGATTAGAAAACAGGGTATCTCAGGTCCTTTGTGATTCAAATGCGTAAAAAATAATTTATGCAGGACTGTGGCATTCTTT  SS2
   700 146 SS2       --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  SS2
   700 147 FP5       TATGcTGATGCatTTtCTAATGATTAGAAAACAAGGAATCTCCGGTCCCTTATAATTAAAAAAAGATTTATCTAAACAACCCTTACCTTAATTCTTTTAA  FP5
   700 148 FP5       --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  FP5
   700 149 Pac7      TATGcTGATGCATTTTTTAATGATTCGTAAACAGGGTATTTCAGGACCTTTATAAACACTTACAATTAAATTACTAACGAGAGCTCCAC----------  Pac7
   700 150 Pac7      --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Pac7
   700 151 MIT9303   CATGCTCATGCATTTCCTGATGATCAGAAAGCAGGGCATTTCAGGTCCGTTGTAAGTGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTT  MIT9303
   700 152 MIT9303   --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  MIT9303
   700 153 MIT9313   CATGCTCATGCATTTCCTGATGATCAGGAAGCAGGGCATTTCAGGTCCGTTGTAAGCGTCAATCATTCAGATTCAAATCCCACTTCACGTTTTTATATTT  MIT9313
   700 154 MIT9313   --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  MIT9313
   700 155 WH8103    CATGCTCATGCACTTCCTGATGATTCGGAAGCAGGGCATTTCTGGTCCCTTGTGATCTGCGTTCCTACCGTTAACTCAACACCTGGATTCACCG------  WH8103
   700 156 WH8103    --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  WH8103
   700 157 PCC7002   CATGTTGGCCCACTTCCTCATGATCCGTAAGCAAGGTATTTCTGGTCCTTTGTAAGGATCAGGTTTTCCGTTTGCTCAAAACAACGTTCTTGTTGAGGAG  PCC7002
   700 158 PCC7002   --M--L--A--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  PCC7002
   700 159 Pthrix    CATGCTGATGCACTTCCTGATGATTCGCAAGCAAGGGATCTCCGGTCCTTTATAAGCGAATGACCGCCAGTGGATTGTGGGCTGTGGATTGTGGGCAGTG  Pthrix
   700 160 Pthrix    --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Pthrix
   700 161 Nostoc    CATGCTGTTTCACTTCTTGATGATCCGCAAACAAGGTATTTCTGGGCCTTTGTAATCTCAAAGATTAGCAAAAGGCAACATTTGTCAGTTTCCAGATGAA  Nostoc
   700 162 Nostoc    --M--L--F--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Nostoc
   700 163 Chlorell  TATGCTAATGCATTTTTTAATGATTCGCAAACAAGGTATATCTGGACCACTTTAAAATAATAACTCGCAAAAGAAACAATTATAAAAAAGATCCAAAATC  Chlorell
   700 164 Chlorell  --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Chlorell
   700 165 Liverwor  TATGTTAATGCACTTTTTAATGATTCGTAAACAAGGTATTTCAGGTCCGTTATAAATTACGTAAATTTATTACAAAATAAAAAAGTTTAAATACTAATTT  Liverwor
   700 166 Liverwor  --M--L--M--H--F--L--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Liverwor
   700 167 Maize     TATGTTAATGCACTTTCCAATGATACGTAAGCAAGGTATTTCGGGTCCGTTATAGGGAAGGCATAGCATAGAGAATTCTAATTCTCATATATCATATCGG  Maize
   700 168 Maize     --M--L--M--H--F--P--M--I--R--K--Q--G--I--S--G--P--L--*-----------------------------------------  Maize
```

Position:	Sequence identity:	Data:	
800	1 S40 96	----------	S40 96
800	2 Alle1	----------	Alle1
800	3 G50 90	----------	G50 90
800	4 Alle2	----------	Alle2
800	5 G50 107	----------	G50 107
800	6 Alle3	----------	Alle3
800	7 S40 128	----------	S40 128
800	8 Alle4	----------	Alle4
800	9 85Br 13	TAATCAATTCGCTATTTTTATCAA----------	85Br 13
800	10 Alle5	----------	Alle5
800	11 85Br 100	CAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC----------	85Br 100
800	12 Alle6	----------	Alle6
800	13 85Br 21	ATCCTCACTTTTTAAATAAGACTAAGTTTCCT----------	85Br 21
800	14 Alle7	----------	Alle7
800	15 85Br 31	TTTCTTAAACAC----------	85Br 31
800	16 Alle8	----------	Alle8
800	17 85Br 66	TCCTTAATCACAATTCCCACTAGGCGCTGATTCAA----------	85Br 66
800	18 Alle9	----------	Alle9
800	19 85Br 106	CAAATTGTTTTATTCCTAATTTCTCTT----------	85Br 106
800	20 Alle11	----------	Alle11
800	21 85Br111S	TTTTTAAAATCGCTCCCTTGAA----------	85Br111S
800	22 Alle12	----------	Alle12
800	23 85Br 118	TTCAGCTTCACCTACATACCCCAACCAACTT----------	85Br 118
800	24 Alle13	----------	Alle13
800	25 85Br 119	AGATTAATTTCTTTTCCA----------	85Br 119
800	26 Alle14	----------	Alle14
800	27 85Br 122	ACAAAACTCTCTA----------	85Br 122
800	28 Alle15	----------	Alle15
800	29 85Br 127	GTCCAGTCCYGGTTCCCTACTT----------	85Br 127
800	30 Alle16	----------	Alle16
800	31 85Br 129	----------	85Br 129
800	32 Alle17	----------	Alle17
800	33 85Br 135	CCATCTTTTCCNNNTTAATTCACTTCAATAACCTTCTCTCCAC----------	85Br 135
800	34 Alle19	----------	Alle19
800	35 85D 68	----------	85D 68
800	36 Alle20	----------	Alle20
800	37 85D 90	----------	85D 90
800	38 Alle21		Alle21
800	39 S120 26	ATTAACTCTTGCTCGACTTTTTATCCA----------	S120 26
800	40 Alle22	----------	Alle22
800	41 135Br 52	----------	135Br 52
800	42 Alle23	----------	Alle23
800	43 135Br 53	AATAGTT----------	135Br 53
800	44 Alle24	----------	Alle24
800	45 135Br 59	NNNNNANNNANNN----------	135Br 59
800	46 Alle25	----------	Alle25

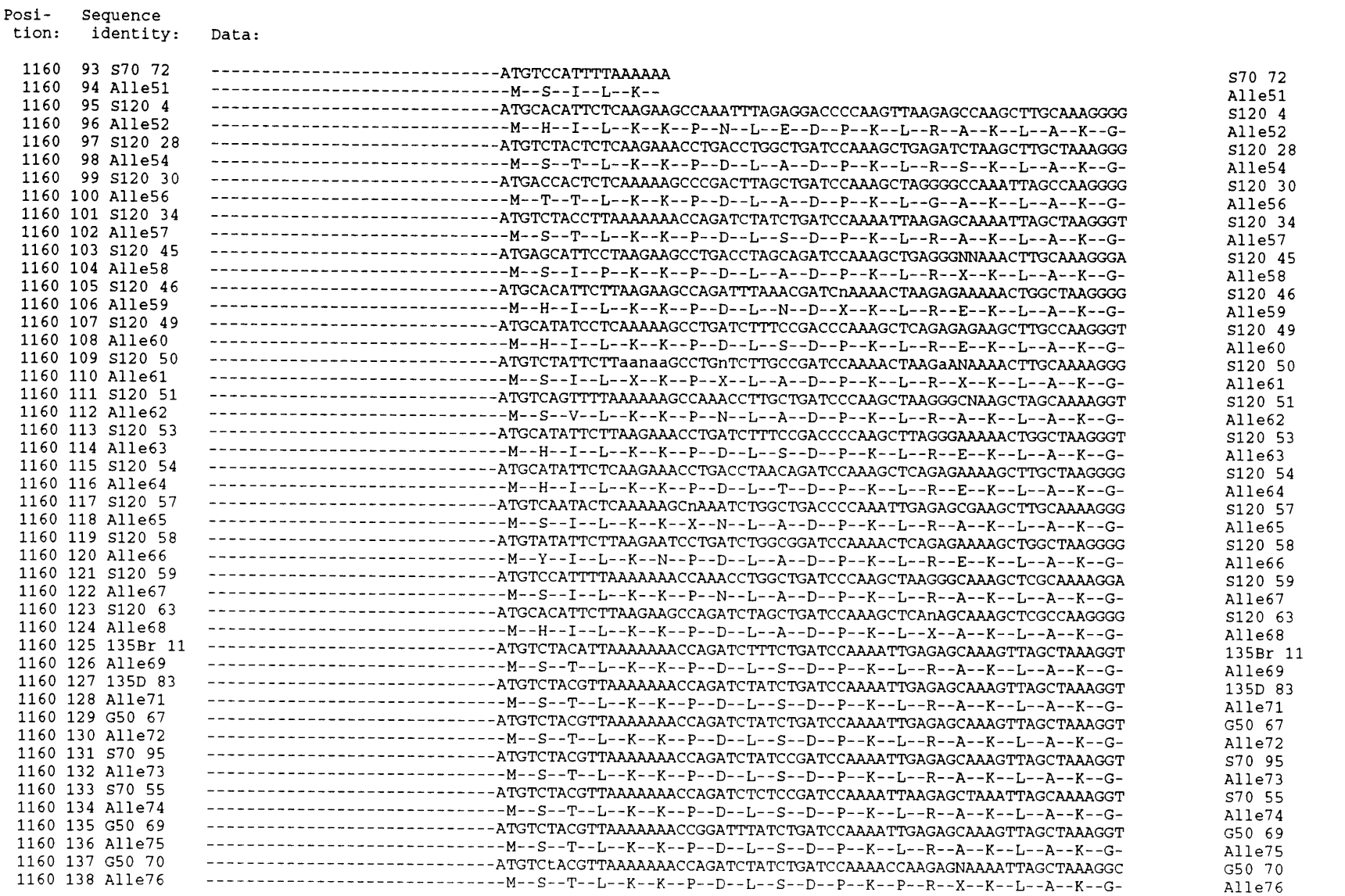
Position:	Sequence identity:	Data:	
800	47 135D 80	ACTTTTTCACAACGTTTTTAGTTCTCATCTG----------	135D 80
800	48 Alle26	----------	Alle26
800	49 135D 95	TCCTTAATCAATTTGATTAATTAAAAAAAATTCGTCTT----------	135D 95
800	50 Alle27	----------	Alle27
800	51 135D 81	TCTTAAACAACTTAACCCT----------	135D 81
800	52 Alle28	----------	Alle28
800	53 135D 84	TCCCAAACCCTTTTTCTAACGGTTCCTGATCTA----------	135D 84
800	54 Alle29	----------	Alle29
800	55 135D 87	TAGTTTCAGTAATT----------	135D 87
800	56 Alle30	----------	Alle30
800	57 135D 91	----------	135D 91
800	58 Alle31	----------	Alle31
800	59 135D 92	CCTCAATCAACAGCTCCTTACAAGCCTCTCAAA----------	135D 92
800	60 Alle32	----------	Alle32
800	61 135D 93	TTAAAGAGACTCTTCACTT----------	135D 93
800	62 Alle33	----------	Alle33
800	63 135D 97	CATTTTCTCTAATTTCTTCACTCTGACTTCAATTAAATCTC----------	135D 97
800	64 Alle35	----------	Alle35
800	65 135D 110	CAAATTGTTTTATTCCTAATTTCTCTT----------	135D 110
800	66 Alle36	----------	Alle36
800	67 135D 119	----------	135D 119
800	68 Alle37	----------	Alle37
800	69 135D 122	TTCATCCTTAAACAAAATCTAATTATTTAGAGTT----------	135D 122
800	70 Alle38	----------	Alle38
800	71 S40 78	AATAATTATTTCTCCCT----------	S40 78
800	72 Alle39	----------	Alle39
800	73 S40 121	----------	S40 121
800	74 Alle40	----------	Alle40
800	75 S40 117	----------	S40 117
800	76 Alle41	----------	Alle41
800	77 S70 43	TCAGCTTCACCTACATACCCCAACCAACTT----------	S70 43
800	78 Alle42	----------	Alle42
800	79 S70 44	TTTTTAAAATCACTCCCTTCCA----------	S70 44
800	80 Alle43	----------	Alle43
800	81 S70 45	AAGCTTTGTTTTTAGTCAAAATCTAAATTTTCACCA----------	S70 45
800	82 Alle44	----------	Alle44
800	83 S70 50	TAACCCAAGATCCAAATCAAATT----------	S70 50
800	84 Alle45	----------	Alle45
800	85 S70 53	CCTTCGATTTTATTTCA----------	S70 53
800	86 Alle46	----------	Alle46
800	87 S70 59	AGATTAATTTCTTTTCCA----------	S70 59
800	88 Alle47	----------	Alle47
800	89 S70 63	TAACCCAAGATCCCAAAACACTT----------	S70 63
800	90 Alle48	----------	Alle48
800	91 S70 66	CCCAGATTCAATTCTCAAGAAACAATTACCTCTT----------	S70 66
800	92 Alle50	----------	Alle50

```
Posi-   Sequence
 tion:   identity:   Data:

   800   93 S70 72    TTTAAATCACTTACTCCTCAT----------                                                  S70 72
   800   94 Alle51    ----------                                                                       Alle51
   800   95 S120 4    ACAAAACTCACTA----------                                                          S120 4
   800   96 Alle52    ----------                                                                       Alle52
   800   97 S120 28   TTGCCTTAAACCTTTCCCAACCCAAAACTAGTTTCTACTTAT----------                            S120 28
   800   98 Alle54    ----------                                                                       Alle54
   800   99 S120 30   AAAAGAATCTCTTTTTAGTT----------                                                 S120 30
   800  100 Alle56    ----------                                                                       Alle56
   800  101 S120 34   ----------                                                                       S120 34
   800  102 Alle57    ----------                                                                       Alle57
   800  103 S120 45   CTATTGTTTCAATAATCTTTTTTTCTA----------                                           S120 45
   800  104 Alle58    ----------                                                                       Alle58
   800  105 S120 46   AAAGGGTTTCCTTTGACTTTTCTCTCTTCCTTTAGGCTCTACTCAA----------                        S120 46
   800  106 Alle59    ----------                                                                       Alle59
   800  107 S120 49   AAACAACTATCCTTTCATTCCCAA----------                                             S120 49
   800  108 Alle60    ----------                                                                       Alle60
   800  109 S120 50   ----------                                                                       S120 50
   800  110 Alle61    ----------                                                                       Alle61
   800  111 S120 51   TGGACATCAACT----------                                                          S120 51
   800  112 Alle62    ----------                                                                       Alle62
   800  113 S120 53   CTCTTTAAAAATCTTTTTCCGCCAAYGTTTTCACTCAA----------                               S120 53
   800  114 Alle63    ----------                                                                       Alle63
   800  115 S120 54   TAGATCAAGTAACACATTTTTACTTTCATCAA----------                                     S120 54
   800  116 Alle64    ----------                                                                       Alle64
   800  117 S120 57   TAACTTCTTCCCCTGAAAGAATACTTTCCAC----------                                      S120 57
   800  118 Alle65    ----------                                                                       Alle65
   800  119 S120 58   TAAGTCTTCCAAATTAATCAATTAATTTACTTTTTCTCAA----------                            S120 58
   800  120 Alle66    ----------                                                                       Alle66
   800  121 S120 59   AAATCACTAGCTTCTA----------                                                     S120 59
   800  122 Alle67    ----------                                                                       Alle67
   800  123 S120 63   TTCCAAAAGCACTGATTTTCACCTGAT----------                                          S120 63
   800  124 Alle68    ----------                                                                       Alle68
   800  125 135Br 11  ----------                                                                       135Br 11
   800  126 Alle69    ----------                                                                       Alle69
   800  127 135D 83   ----------                                                                       135D 83
   800  128 Alle71    ----------                                                                       Alle71
   800  129 G50 67    ----------                                                                       G50 67
   800  130 Alle72    ----------                                                                       Alle72
   800  131 S70 95    ----------                                                                       S70 95
   800  132 Alle73    ----------                                                                       Alle73
   800  133 S70 55    ----------                                                                       S70 55
   800  134 Alle74    ----------                                                                       Alle74
   800  135 G50 69    ----------                                                                       G50 69
   800  136 Alle75    ----------                                                                       Alle75
   800  137 G50 70    ----------                                                                       G50 70
   800  138 Alle76    ----------                                                                       Alle76
```


Posi-
tion:
Sequence
identity:
Data:
800 139 S40 137
800 140 Alle77
800 141 G50 72
800 142 Alle78
800 143 Med4
800 144 Med4
800 145 SS2
800 146 SS2
800 147 FP5
800 148 FP5
800 149 Pac7
800 150 Pac7
800 151 MIT9303
800 152 MIT9303
800 153 MIT9313
800 154 MIT9313
800 155 MIT9314
800 156 MIT9314
800 157 WH8103
800 158 WH8103
800 159 PCC7002
800 160 PCC7002
800 161 Pthrix
800 162 Pthrix
800 163 Nostoc
800 164 Nostoc
800 165 Chlorell
800 166 Chlorell
800 167 Liverwor
800 168 Liverwor
TTATCCAAATCTTTTACCTAATTTAATTTTTAAAT
ACCTTTCTTAAACC
CAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC
CAGTTCAGCACCTACTTCATTCGCTAACTATTCGGCCAAAATCCC
AACTTTTACTC
AATTTCGGGCTGTGAATTTCGGGCTGTGAATGACTGTCAGTGAATTACCGCCAGTGAATAACCACCAGTGAATGACCGCCTGTAGTTTGCT//
GTTGTTTGTTTTAAACTTATGAAACACAAGAACCTAATTGTAGATAAGTTACAAAAAACAAACACTTGCATCCCAAGCCGAAAGCTTTAGA//
ATT
CATTGCCATATTTTATGGATTTTTTTTTTCTATTTGAAAGTTCTTTTTAGAGAAATGCTAAAAAAATTTTTTTTAATAGATATTTTATAAA//
S40 137
Alle77
G50 72
Alle78
Med4
Med4
SS2
SS2
FP5
FP5
Pac7
Pac7
MIT9303
MIT9303
MIT9313
MIT9313
MIT9314
MIT9314
WH8103
WH8103
PCC7002
PCC7002
Pthrix
Pthrix
Nostoc
Nostoc
Chlorell
Chlorell
Liverwor
Liverwor


Posi-
tion:
Sequence
identity:
Data:
1160 1 S40 96 ----------ATGTCTACTTTAAAAAAACCAGATCTATCTGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT S40 96
1160 2 Alle1 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle1
1160 3 G50 90 ----------ATGTCTACGTTAAAAAAACCAGATTTATCCGATCCAAAATTAAGASnAAAGCTTGCCAAAGGG G50 90
1160 4 Alle2 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--X--K--L--A--K--G- Alle2
1160 5 G50 107 ----------ATGTCAACGTTAAAGAAACCAGATTTATCAGATCCAAAAYTAAGAGCAAAACTAGCCAAGGGT G50 107
1160 6 Alle3 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle3
1160 7 S40 128 ----------ATGCACATTCTCAAGAAGCCTGACCTCTCCGACCCCAAG S40 128
1160 8 Alle4 ----------M--H--I--L--K--K--P--D--L--S--D--P--K- Alle4
1160 9 85Br 13 ----------ATGCATATTTTAAAGAAACCTgaTCTTTCAGAT 85Br 13
1160 10 Alle5 ---------- M--H--I--L--K--K--P--D--L--S--D- Alle5
1160 11 85Br 100 ----------ATGCACATTCTTAAGAAGCTCGATTTCTCTGATCCAAAATTGAGGGCgAAACTGGCCAAGGGA 85Br 100
1160 12 Alle6 ----------M--H--I--L--K--K--L--D--F--S--D--P--K--L--R--A--K--L--A--K--G- Alle6
1160 13 85Br 21 ----------ATGTCAACACTCAAGAAGCCAAATCTTGCTGATCCCAAGTTAAGGGCAAAGCTAGCTAAAGGT 85Br 21
1160 14 Alle7 ----------M--S--T--L--K--K--P--N--L--A--D--P--K--L--R--A--K--L--A--K--G- Alle7
1160 15 85Br 31 ----------ATGTCTACTCTTAAGAAACCAGATTTATCTGACACCAAGTTAAGAGCAAAACTTGCTAAAGGA 85Br 31
1160 16 Alle8 ----------M--S--T--L--K--K--P--D--L--S--D--T--K--L--R--A--K--L--A--K--G- Alle8
1160 17 85Br 66 ----------ATGCACATTCTTAAAAAGCCAGaTCTAAGCGATCCAAAACTTAGAGAAAAACTTGCCAAAGGG 85Br 66
1160 18 Alle9 ----------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G- Alle9
1160 19 85Br 106 ----------ATGTCCACACTTAAAAAGCCAGATCTAACGGACCCCAAGTTAAGAGCCAAGCTTGCAAAGGGT 85Br 106
1160 20 Alle11 ----------M--S--T--L--K--K--P--D--L--T--D--P--K--L--R--A--K--L--A--K--G- Alle11
1160 21 85Br111S ----------ATGTCCATTTTAAAAAAACCAAACCTGGCTGATCCTAAGCTAAGGGCCAAGCTCGCGAAAGGA 85Br111S
1160 22 Alle12 ----------M--S--I--L--K--K--P--N--L--A--D--P--K--L--R--A--K--L--A--K--G- Alle12
1160 23 85Br 118 ----------ATGTCAACACTAAAGAAGCCTGnTCTTGCAGATCCAAAGCTCAGANnAAAGCTTGCAAAAGGG 85Br 118
1160 24 Alle13 ----------M--S--T--L--K--K--P--X--L--A--D--P--K--L--R--X--K--L--A--K--G- Alle13
1160 25 85Br 119 ----------ATGCATATTCTTAAAAAGCCTGATTTGTCTGACCCCAAGCTCAGGGAAAAGCTTGCCAAGGGT 85Br 119
1160 26 Alle14 ----------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G- Alle14
1160 27 85Br 122 ----------ATGCACATTCTCAAGAAGCCAAATTTAGAGGACCCCAAGTTAAGAGCCAAGCnCAGAAAGGGG 85Br 122
1160 28 Alle15 ----------M--H--I--L--K--K--P--N--L--E--D--P--K--L--R--A--K--X--R--K--G- Alle15
1160 29 85Br 127 ----------ATGCACATTCTCAAGAAGCCTGATTTAGCTGATCCCAAGCTTAgAGCCAAGCTCGCTAAAGC 85Br 127
1160 30 Alle16 ----------M--H--I--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--A- Alle16
1160 31 85Br 129 ----------ATGTCGACTCTTAAAAAACCTGATTTATCTGATCCTAAGTTAAGAGC 85Br 129
1160 32 Alle17 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A Alle17
1160 33 85Br 135 ----------ATGcACATACTTAAAAAGCCTGATCTATCTGACCCAAAGCTCAGAGAAAAGCTTGCAAAGGGT 85Br 135
1160 34 Alle19 ----------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G- Alle19
1160 35 85D 68 ----------ATGTCTGTTATTAAGAAACCAGATCTAGAAGATCCAAAATTACGTGCCAAATTAGCGAAGGGT 85D 68
1160 36 Alle20 ----------M--S--V--I--K--K--P--D--L--E--D--P--K--L--R--A--K--L--A--K--G- Alle20
1160 37 85D 90 ----------ATGTCCACTCTTAAGAAACCTGATTTAGCTGATCCAAAATTGAGnGCGAAGCTCGCTAAGGGT 85D 90
1160 38 Alle21 ----------M--S--T--L--K--K--P--D--L--A--D--P--K--L--X--A--K--L--A--K--G- Alle21
1160 39 S120 26 ----------ATGCATATTCTCAAGAAGCCCGATCTTTCTGACCTAAAGCTAAGAGAAAAGCTCGCAAAGGGT S120 26
1160 40 Alle22 ----------M--H--I--L--K--K--P--D--L--S--D--L--K--L--R--E--K--L--A--K--G- Alle22
1160 41 135Br 52 ----------ATGTCGACTCTCAAAAAACCTGATTTATCCGATCCTAAGCTAAGGGCTAAGCTTTCAAAAGCA 135Br 52
1160 42 Alle23 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--S--K--A- Alle23
1160 43 135Br 53 ----------ATGTCAGTAATCAAAAAACCCGATCTATCAGATCCAAGATTACGTGCGAAGTTAGCTAAAGGT 135Br 53
1160 44 Alle24 ----------M--S--V--I--K--K--P--D--L--S--D--P--R--L--R--A--K--L--A--K--G- Alle24
1160 45 135Br 59 ----------ANNNNNANNNNAAAAAACCAGATTTATCTGATCCGAAGTTGCGGNNAAAACTTGCAAAAGGG 135Br 59
1160 46 Alle25 ----------X--X--X--X--K--K--P--D--L--S--D--P--K--L--R--X--K--L--A--K--G- Alle25

```
Posi-   Sequence
 tion:   identity:   Data:

 1160  47 135D 80   ----------------------------ATGCATATTCTTAAGAAGCCTGATCTTGCTGATCCCAAACTTAGGGAAAAACTGGCAAAGGGA          135D 80
 1160  48 Alle26    ----------------------------M--H--I--L--K--K--P--D--L--A--D--P--K--L--R--E--K--L--A--K--G-          Alle26
 1160  49 135D 95   ----------------------------ATGTCCATCTTAAAGAAACCTGATCTAACTGATCCAAAACTTAGATCTAAACTCGCCAAGGGA          135D 95
 1160  50 Alle27    ----------------------------M--S--I--L--K--K--P--D--L--T--D--P--K--L--R--S--K--L--A--K--G-          Alle27
 1160  51 135D 81   ----------------------------ATGTCCATTTTAAAAAAGCCAAATCTTGCTGATTCCAAGCTAAGGGcAAAGCTTGCTAAAGGA          135D 81
 1160  52 Alle28    ----------------------------M--S--I--L--K--K--P--N--L--A--D--S--K--L--R--A--K--L--A--K--G-          Alle28
 1160  53 135D 84   ----------------------------ATGCACATTCTCAAGAAGCCAGATCTCGATGATCCCAAGCTAAGGGCTAAATTGGCTAAAGGT          135D 84
 1160  54 Alle29    ----------------------------M--H--I--L--K--K--P--D--L--D--D--P--K--L--R--A--K--L--A--K--G-          Alle29
 1160  55 135D 87   ----------------------------ATGTCAACTCTCAAAAAACCAGATCTTGCGGACCCAAAGCTTAGGGCCAAGCTAGCTAAAGGG          135D 87
 1160  56 Alle30    ----------------------------M--S--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle30
 1160  57 135D 91   ----------------------------ATGTCGACTCTTAAAAAACCnGATTTATCCGACCCTAAGCTGAGAGCTAAGCTTTCAAAAGCA          135D 91
 1160  58 Alle31    ----------------------------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--S--K--A-          Alle31
 1160  59 135D 92   ----------------------------ATGTCAGTTTTAAAAAAACCAGATTTAGCTGATCCAAAACTTAGAGCAAAGCTTGCGAAAGGT          135D 92
 1160  60 Alle32    ----------------------------M--S--V--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle32
 1160  61 135D 93   ----------------------------ATGTCCATTTTAAGAAAGCCAAATCTTGCTGATCTCAAGCTCAGGGCAAAGCTTGCAAAGGGA          135D 93
 1160  62 Alle33    ----------------------------M--S--I--L--R--K--P--N--L--A--D--L--K--L--R--A--K--L--A--K--G-          Alle33
 1160  63 135D 97   ----------------------------ATGCACATTCTTAAAAAGCCAGATTTGTCAGATCCAAAGCTGAGGGAAAAGCTTGCTAAAGGG          135D 97
 1160  64 Alle35    ----------------------------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G-          Alle35
 1160  65 135D 110  ----------------------------ATGTCCACACTTAAAAAGCCAGATCTAGCGGACCCCAAGTTAAGAGCCAAGCTTGCAAAGGGT          135D 110
 1160  66 Alle36    ----------------------------M--S--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle36
 1160  67 135D 119  ----------------------------ATGTCGACTCTTAAAAAACCTGATTTATCCGATCCTAAGTTAAGAGCCAAGCTTTCAAAAGCA          135D 119
 1160  68 Alle37    ----------------------------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--S--K--A          Alle37
 1160  69 135D 122  ----------------------------ATGCACATTCTTAAGAAGCCCGATTTAGCTGACCCGAAGTTAAGGNAnAAGCTCGCAAAGGGA          135D 122
 1160  70 Alle38    ----------------------------M--H--I--L--K--K--P--D--L--A--D--P--K--L--R--X--K--L--A--K--G-          Alle38
 1160  71 S40 78    ----------------------------ATGTCTATTCTTAAAAAGCCTGATCTTGCCGATCCAAAACTAAGAGcAAAACTTGCAAAAGGG          S40 78
 1160  72 Alle39    ----------------------------M--S--I--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle39
 1160  73 S40 121   ----------------------------ATGTCTATTTTAAAAAAGCCCGATCTAGCAGATCCTAAACTAAGAGCTAAACTAGCTAAAGGA          S40 121
 1160  74 Alle40    ----------------------------M--S--I--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle40
 1160  75 S40 117   ----------------------------ATGCACATTCTCAAGAAGCCTGATCTGTCCGACCCCAAGATGCGCGnGAAGCTCGCCAAGGGG          S40 117
 1160  76 Alle41    ----------------------------M--H--I--L--K--K--P--D--L--S--D--P--K--M--R--X--K--L--A--K--G-          Alle41
 1160  77 S70 43    ----------------------------ATGTCAACACTAAAGAAGCCTGaTCTTGCAGATCCAAAGCTCAGAGCGAAGCTTGCAAAAGGG          S70 43
 1160  78 Alle42    ----------------------------M--S--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle42
 1160  79 S70 44    ----------------------------ATGTCCATTTTAAAAAAACCAAACCTGGCTGACCCGAAGCTAAGGGCAAAGCTTGCAAAAGGA          S70 44
 1160  80 Alle43    ----------------------------M--S--I--L--K--K--P--N--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle43
 1160  81 S70 45    ----------------------------ATGCACATTCTCAAGAAGCCAGATTTAGCTGATCCCAAGTTAAGAGAAAAACTCGCAAAAgGC          S70 45
 1160  82 Alle44    ----------------------------M--H--I--L--K--K--P--D--L--A--D--P--K--L--R--E--K--L--A--K--G-          Alle44
 1160  83 S70 50    ----------------------------ATGACGACTCTAAAAAAACCTGATCTTGCTGATCCCAAACTTAGAGCCAAGCTTGCTAAAGGA          S70 50
 1160  84 Alle45    ----------------------------M--T--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle45
 1160  85 S70 53    ----------------------------ATGAGTACTCTTAAAAAGCCTGACTTAGCTGATCCAAAGCTTAGGGCGAAACTTGCTAAAGGG          S70 53
 1160  86 Alle46    ----------------------------M--S--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle46
 1160  87 S70 59    ----------------------------ATGCATATTCTTAAAAAGCCTGATTTGTCTGACCCCAAGCTCAGGGAAAAGCTTGCCGAGGGT          S70 59
 1160  88 Alle47    ----------------------------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--E--G-          Alle47
 1160  89 S70 63    ----------------------------ATGACGACACTAAAGAAACCTGATCTTGCAGATCCAAAGCTTAGAGCAAAGCTTGCAAAAGGG          S70 63
 1160  90 Alle48    ----------------------------M--T--T--L--K--K--P--D--L--A--D--P--K--L--R--A--K--L--A--K--G-          Alle48
 1160  91 S70 66    ----------------------------ATGGATGCTCTTAAAAAGCCTGATCTAGCTGACCAAAAACTGAGAGCTAAGCTTGCCAAGGGA          S70 66
 1160  92 Alle50    ----------------------------M--D--A--L--K--K--P--D--L--A--D--Q--K--L--R--A--K--L--A--K--G-          Alle50
```


Posi-
tion:
Sequence
identity:
Data:
1160 93 S70 72 ----------ATGTCCATTTTAAAAAA S70 72
1160 94 Alle51 ----------M--S--I--L--K-- Alle51
1160 95 S120 4 ----------ATGCACATTCTCAAGAAGCCAAATTTAGAGGACCCCAAGTTAAGAGCCAAGCTTGCAAAGGGG S120 4
1160 96 Alle52 ----------M--H--I--L--K--K--P--N--L--E--D--P--K--L--R--A--K--L--A--K--G- Alle52
1160 97 S120 28 ----------ATGTCTACTCTCAAGAAACCTGACCTGGCTGATCCAAAGCTGAGATCTAAGCTTGCTAAAGGG S120 28
1160 98 Alle54 ----------M--S--T--L--K--K--P--D--L--A--D--P--K--L--R--S--K--L--A--K--G- Alle54
1160 99 S120 30 ----------ATGACCACTCTCAAAAAGCCCGACTTAGCTGATCCAAAGCTAGGGGCCAAATTAGCCAAGGGG S120 30
1160 100 Alle56 ----------M--T--T--L--K--K--P--D--L--A--D--P--K--L--G--A--K--L--A--K--G- Alle56
1160 101 S120 34 ----------ATGTCTACCTTAAAAAAACCAGATCTATCTGATCCAAAATTAAGAGCAAAATTAGCTAAGGGT S120 34
1160 102 Alle57 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle57
1160 103 S120 45 ----------ATGAGCATTCCTAAGAAGCCTGACCTAGCAGATCCAAAGCTGAGGGNNAAACTTGCAAAGGGA S120 45
1160 104 Alle58 ----------M--S--I--P--K--K--P--D--L--A--D--P--K--L--R--X--K--L--A--K--G- Alle58
1160 105 S120 46 ----------ATGCACATTCTTAAGAAGCCAGATTTAAACGATCnAAAACTAAGAGAAAAACTGGCTAAGGGG S120 46
1160 106 Alle59 ----------M--H--I--L--K--K--P--D--L--N--D--X--K--L--R--E--K--L--A--K--G- Alle59
1160 107 S120 49 ----------ATGCATATCCTCAAAAAGCCTGATCTTTCCGACCCAAAGCTCAGAGAGAAGCTTGCCAAGGGT S120 49
1160 108 Alle60 ----------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G- Alle60
1160 109 S120 50 ----------ATGTCTATTCTTaanaaGCCTGnTCTTGCCGATCCAAAACTAAGaANAAAACTTGCAAAAGGG S120 50
1160 110 Alle61 ----------M--S--I--L--X--K--P--X--L--A--D--P--K--L--R--X--K--L--A--K--G- Alle61
1160 111 S120 51 ----------ATGTCAGTTTTAAAAAAGCCAAACCTTGCTGATCCCAAGCTAAGGGCNAAGCTAGCAAAAGGT S120 51
1160 112 Alle62 ----------M--S--V--L--K--K--P--N--L--A--D--P--K--L--R--A--K--L--A--K--G- Alle62
1160 113 S120 53 ----------ATGCATATTCTTAAGAAACCTGATCTTTCCGACCCCAAGCTTAGGGAAAAACTGGCTAAGGGT S120 53
1160 114 Alle63 ----------M--H--I--L--K--K--P--D--L--S--D--P--K--L--R--E--K--L--A--K--G- Alle63
1160 115 S120 54 ----------ATGCATATTCTCAAGAAACCTGACCTAACAGATCCAAAGCTCAGAGAAAAGCTTGCTAAGGGG S120 54
1160 116 Alle64 ----------M--H--I--L--K--K--P--D--L--T--D--P--K--L--R--E--K--L--A--K--G- Alle64
1160 117 S120 57 ----------ATGTCAATACTCAAAAAGCnAAATCTGGCTGACCCCAAATTGAGAGCGAAGCTTGCAAAAGGG S120 57
1160 118 Alle65 ----------M--S--I--L--K--K--X--N--L--A--D--P--K--L--R--A--K--L--A--K--G- Alle65
1160 119 S120 58 ----------ATGTATATTCTTAAGAATCCTGATCTGGCGGATCCAAAACTCAGAGAAAAGCTGGCTAAGGGG S120 58
1160 120 Alle66 ----------M--Y--I--L--K--N--P--D--L--A--D--P--K--L--R--E--K--L--A--K--G- Alle66
1160 121 S120 59 ----------ATGTCCATTTTAAAAAAACCAAACCTGGCTGATCCCAAGCTAAGGGCAAAGCTCGCAAAAGGA S120 59
1160 122 Alle67 ----------M--S--I--L--K--K--P--N--L--A--D--P--K--L--R--A--K--L--A--K--G- Alle67
1160 123 S120 63 ----------ATGCACATTCTTAAGAAGCCAGATCTAGCTGATCCAAAGCTCAnAGCAAAGCTCGCCAAGGGG S120 63
1160 124 Alle68 ----------M--H--I--L--K--K--P--D--L--A--D--P--K--L--X--A--K--L--A--K--G- Alle68
1160 125 135Br 11 ----------ATGTCTACATTAAAAAAACCAGATCTTTCTGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT 135Br 11
1160 126 Alle69 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle69
1160 127 135D 83 ----------ATGTCTACGTTAAAAAAACCAGATCTATCTGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT 135D 83
1160 128 Alle71 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle71
1160 129 G50 67 ----------ATGTCTACGTTAAAAAAACCAGATCTATCTGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT G50 67
1160 130 Alle72 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle72
1160 131 S70 95 ----------ATGTCTACGTTAAAAAAACCAGATCTATCCGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT S70 95
1160 132 Alle73 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle73
1160 133 S70 55 ----------ATGTCTACGTTAAAAAAACCAGATCTCTCCGATCCAAAATTAAGAGCTAAATTAGCAAAAGGT S70 55
1160 134 Alle74 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle74
1160 135 G50 69 ----------ATGTCTACGTTAAAAAAACCGGATTTATCTGATCCAAAATTGAGAGCAAAGTTAGCTAAAGGT G50 69
1160 136 Alle75 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K--G- Alle75
1160 137 G50 70 ----------ATGTCtACGTTAAAAAAACCAGATCTATCTGATCCAAAACCAAGAGNAAAATTAGCTAAAGGC G50 70
1160 138 Alle76 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--P--R--X--K--L--A--K--G- Alle76


Posi-
tion:
Sequence
identity:
Data:
1160 139 S40 137 ----------ATGCACATTCTNAAGAAGCCTGACCTCTCCGATCCCAAG S40 137
1160 140 Alle77 ----------M--H--I--L--K--K--P--D--L--S--D--P--K- Alle77
1160 141 G50 72 ----------ATGCACATTCTCAAGAAGCCTGACCTCTCCGATCCCAAG G50 72
1160 142 Alle78 ----------M--H--I--L--K--K--P--D--L--S--D--P--K- Alle78
1160 143 Med4 ----------ATGTCTACTTTAAAAAAACCTGATTTATCTGATCCTAAATTAAGAGCTAAGTTGGCTAAA Med4
1160 144 Med4 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--K- Med4
1160 145 SS2 ----------ATGTCCACTCTAAAGAAACCAAATTTATCTGATCCAAAGCTAAGGGCTAAGCTT SS2
1160 146 SS2 ----------M--S--T--L--K--K--P--N--L--S--D--P--K--L--R--A--K--L- SS2
1160 147 FP5 ----------ATGTCTACTCTTAAGAAACCTGATTTAACCGATACAAAATTAAGAGCAAAACTTGCTaaa FP5
1160 148 FP5 ----------M--S--T--L--K--K--P--D--L--T--D--T--K--L--R--A--K--L--A--K- FP5
1160 149 Pac7 ----------ATGTCTACGTTAAAAAAACCAGATCTATCTGATCCAAAATTAAGAGCAAAATTAGCT Pac7
1160 150 Pac7 ----------M--S--T--L--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A- Pac7
1160 151 MIT9303 ----------ATGCACATTCTTAAGAAGCCCGATTNcTcTGATCNAAAATTGAGA MIT9303
1160 152 MIT9303 ----------M--H--I--L--K--K--P--D--X--S--D--X--K--L--R- MIT9303
1160 153 MIT9313 ----------ATGCACATTCTTAAGAAGCCCnATTTCTCTGAT MIT9313
1160 154 MIT9313 ----------M--H--I--L--K--K--P--X--F--S--D- MIT9313
1160 155 MIT9314 ----------ATGTCTACATTNAAAAAACCAGATCNATCAGATCC MIT9314
1160 156 MIT9314 ----------M--S--T--X--K--K--P--D--X--S--D- MIT9314
1160 157 WH8103 ----------ATGCACATTCTCAAGAAGCCTGACCTCTCCGATCCCAAG WH8103
1160 158 WH8103 ----------M--H--I--L--K--K--P--D--L--S--D--P--K- WH8103
1160 159 PCC7002 ----------ATGTCTATCATGAAAAAGCCGGATCTTAGCGATCCAAAACTCCGGGCAAAACTGGCTCAAAACATGGG... PCC7002
1160 160 PCC7002 ----------M--S--I--M--K--K--P--D--L--S--D--P--K--L--R--A--K--L--A--Q--N--M--G... PCC7002
1160 161 Pthrix CTGCTCTCTTACGGAGATCATTGAAAACTATGTCTGTTCTCAAAAAGCCGGATTTAACCGATCCCGTCCTATTGGAAAAGCTGGCCCAAAATATGGG... Pthrix
1160 162 Pthrix ----------M--S--V--L--K--K--P--D--L--T--D--P--V--L--L--E--K--L--A--Q--N--M--G... Pthrix
1160 163 Nostoc ----------ATGGCAACACAGAAAAAACCTGACCTCAGCGACCCCCAGTTAAGAGCCAAACTGGCTAAAGGTATGGG... Nostoc
1160 164 Nostoc ----------M--A--T--Q--K--K--P--D--L--S--D--P--Q--L--R--A--K--L--A--K--G--M--G... Nostoc
1160 165 Chlorell ----------ATGGCTGTTACAAAAAAACCAGATTTATCAGACCCTCAATTACGTGCTAAATTAGCAAAAGGTATGGG... Chlorell
1160 166 Chlorell ----------M--A--V--T--K--K--P--D--L--S--D--P--Q--L--R--A--K--L--A--K--G--M--G... Chlorell
1160 167 Liverwor ----------ATGGGAGTAACAAAAAAACCTGATTTAAGTGATCCTATATTACGAGCTAAATTAGCAAAAGGTATGGG... Liverwor
1160 168 Liverwor ----------M--G--V--T--K--K--P--D--L--S--D--P--I--L--R--A--K--L--A--K--G--M--G... Liverwor